Faculty of Mathematics and Informatics

VILNIUS UNIVERSITY
FACULTY OF MATHEMATICS AND INFORMATICS
MODELLING AND DATA ANALYSIS
MASTER'S STUDY PROGRAMME

The differentiation of women's and men's earnings and its modelling

Master's thesis

Author: Ugnė Samulevičiūtė
VU email address: ugne.samuleviciute@mif.stud.vu.lt
Supervisor: Prof. habil. dr. Vydas Čekanavičius

Vilnius

2025

Contents

Abstract

This thesis focuses on the average hourly wage situation in Lithuania for different genders and employs statistical modelling techniques to analyze whether globally defined wage disparity problems are present in Lithuania. Occupational segregation as well as human capital theory are explored measuring the impact of education, work experience and different occupational spheres impact on the average wage for men and women. The analysis confirms a persistent gender wage gap in Lithuania, with men consistently earning higher wages than women.

Keywords:

gender wage difference, statistical modelling, SEM, OLS, HML, men wage, women wage, structural equation modelling, ordinary least squares regression, multilevel (hierarchical) modelling

1 Literature Review

The gender pay gap, the disparity in earnings between men and women has been a persistent issue across various sectors and regions. Despite significant progress in gender equality, the wage gap remains a pressing concern. Research consistently shows that women earn less than men on average. The Organisation for Economic Co-operation and Development (OECD) reported that women earn approximately 11.9% less than men in OECD countries, meaning that the median full-time working woman earns about 88 cents to every euro earned by the median full-time working man [14]. The gap varies by industry, age and education level, but can be noticed accross most of the work sectors.

Various theories have been proposed to explain earnings differentiation, including human capital theory, occupational segregation and discrimination.

One of the primary causes of the gender pay gap is occupational segregation, where men and women tend to work in different occupations and industries. Male-dominated fields such as engineering, technology and finance generally offer higher salaries compared to female-dominated fields like education, healthcare and social services [2]. This segregation can be across different fields or within the same field, but in different positions of hierarchy. While occupational segregation explains part of the pay gap, it does not account for all the disparities observed. Some studies suggest that even within the same job, women are paid less than men, indicating that other factors also play a significant role [13].

The human capital theory suggests that differences in education, work experience and job tenure contribute to the gender pay gap. Historically, men have had greater access to higher education and more continuous work experiences. However, recent studies have shown that even when controlling for these variables significant pay gaps persist [11]. This indicates that human capital differences alone cannot fully explain the gender pay gap, because women with similar educational backgrounds and work experiences as men still earn less, highlighting the limitations of human capital theory in explaining gender wage disparities.

Discrimination, both overt and covert, plays a critical role in perpetuating the gender pay gap. Women often face biases in hiring, promotions and salary negotiations. Experimental studies have demonstrated that identical resumes receive different responses based on the gender implied by the name [17], highlighting the presence of gender bias in hiring practices. Additionally, women are often penalized for negotiating salaries [7], a behavior that is typically expected and rewarded in men. This contributes to lower starting salaries for women and slower wage growth over their careers.

Another factor that impacts women's earnings potential is family responsibilities which women are more likely to take part in resulting in career breaks or working part-time. The "motherhood penalty" refers to the negative impact on wages experienced by women who take time off for household commitments. Conversely, men often receive a "fatherhood bonus"

in wages, reflecting societal expectations and norms around gender roles [19]. This dynamic deepens the problem of wage differences for women with children in particular.

Despite the economic benefits of childbearing for society, women face financial penalties for having children. A study by Census Bureau researchers found that from two years before the birth of a couple's first child to one year after, the earnings gap between opposite-sex spouses doubles and continues to widen until the child reaches the age of 10 [18]. Although it narrows later on, it never fully closes.

This disparity is not merely a matter of equity but also has profound implications for economic growth, social cohesion and family welfare. Lower earnings for women contribute to higher poverty rates among single mothers and elderly women [1]. Economic dependency on partners or the state can limit women's financial autonomy and decision-making power [4]. The problem is persistent thorughout most of the countries around the world and Lithuania is not an exception [13]. In this case, Luxembourg could be taken as an example that should be followed as differences between the earnings of men and women no longer exist there [16].

Addressing the gender wage gap is critical for achieving inclusive economic development and ensuring that all individuals have equal opportunities to contribute to and benefit from economic activities.

Descriptive statistics and regression analysis have been employed to identify the key determinants of the gender wage gap. The use of regression models helps isolate the effect of individual factors such as education, experience and occupation. However, the approach is limited in addressing the interaction between multiple variables simultaneously [10].

Longitudinal data analysis has also been used to explore the impact of education on gender wage disparities by Claudia Goldin. By tracking changes over time, study of this kind provides valuable insights into the persistent nature of the gender pay gap despite increasing educational attainment among women. However, this analysis primarily focuses on educational factors and may overlook other significant economic variables [6].

And while there are many cross-country analysis done on the gender pay gap which offers a broad perspective, nuances of individual countries might become unnoticed.

This thesis focuses on the situation in Lithuania only and employs statistical modelling techniques to analyze whether global problems stand out in our country as well. Occupational segregation as well as human capital theory will be explored measuring the impact of education, work experience and different occupational spheres impact on the average wage for men and women.

2 Exploratory Data Analysis

The following section defines data that is used in this work. Data has been taken from State Data Agency of Lithuania. Two datasets - one for year 2014 and another for year 2018 are analyzed. Datasets consist of 24 columns, but only the following ones are used: bdu_val (hourly wage), stazas (work experience), im_dydzio_kodas (company size, indicated by a 0-2 code, where 0 is less than 50, 1 is 50 to 249 and 2 is more than 250 workers), lytis (gender, F for females, M for males), issilavinimas (education, where 1 is primary education, 2 is high school diploma, 3 is bachelors degree and 4 is masters degree and everything above that), profesija (profession, indicated by the employee's profession code (according to the Lithuanian profession classifier LPK-2012 – 3 characters)).

2.1 2014 dataset

The 2014 dataset provides a breakdown of wage statistics by gender, highlighting key measures for women and men across a variety of wage percentiles. As the dataset contains information about the wages in 2014, originally it was using Litas currency, but the data was transformed to match 2018 dataset and current national currency - euro. It includes 17 793 observations of females and 19 512 of males that are full time employees.

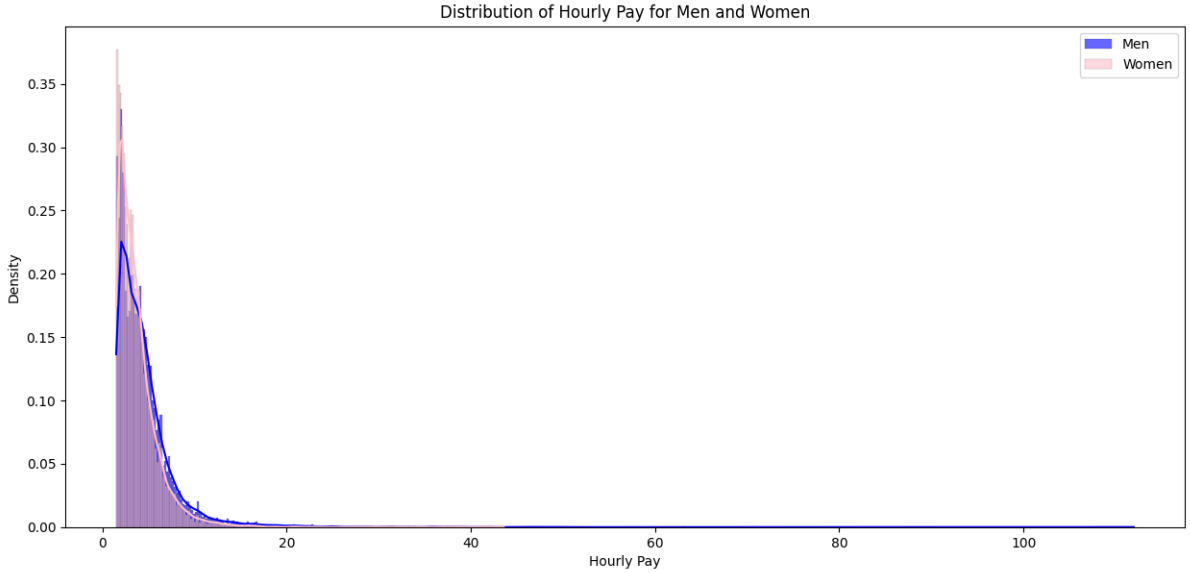
On average, females earn an hourly wage of 3.82 euros, with a standard deviation of 2.39 euros, indicating a relatively narrow distribution around their mean wage. In contrast, males have a higher average hourly wage of 4.54 euros with a standard deviation of 3.55 euros, reflecting greater wage variability among men compared to women. While the minimum wage for both genders is similar at approximately 1.50 euros per hour, the maximum hourly wage for men reaches 111.97 euros, significantly surpassing the highest hourly wage for women which is 43.58 euros. This disparity in maximum wages suggests the presence of extremely high-earning individuals among men, a trend not as prevalent among women.

A closer look at the quartile distribution further highlights these differences. For women, the 25th percentile wage is 2.22 euros, the median (50th percentile) wage is 3.20 euros and the 75th percentile wage is 4.62 euros. For men, the 25th percentile wage is slightly higher at 2.37 euros, with a median hourly wage of 3.76 euros and a 75th percentile wage of 5.42 euros. This distribution shows a consistent pattern of higher earnings for men across each level, further underscoring the wage gap.

2014 data set description								
Gender	Records	Mean	Standard deviation	Min	25%	50%	75%	Max
F	17793	3.818	2.394	1.497	2.215	3.203	4.619	43.582
M	19512	4.545	3.547	1.497	2.374	3.756	5.416	111.968

The illustration of a dataset below shows a clear gender wage disparity, with men not only earning higher hourly wages on average, but also displaying wider wage variability and significantly higher maximum earnings.

Figure 1: Distribution of Hourly Pay for Men and Women, year 2014



2.2 2018 dataset

The 2018 dataset of average hourly wages measured in euros highlights continued differences between male and female earnings across multiple percentiles. Sample size is 16 811 women and 19 974 men working full time only.

In terms of average hourly wages, women earned 5.10 euros with a standard deviation of 3.06 euros, showing a slightly broader distribution than in prior years. Men's average hourly wage in 2018 was higher at 5.90 euros with a standard deviation of 4.35 euros, indicating a greater spread in male earnings compared to women. While minimum hourly wages were nearly identical for both genders—at 2.14 euros for women and 2.13 euros for men—men's maximum hourly wage (116.85 euros) was substantially higher than that for women (71.90 euros). This substantial difference in the upper wage limits implies the presence of very high-earning individuals among men, a pattern not as pronounced among women once again.

Examining the quartile distribution also reveals disparities at different earning levels. For women, the 25th percentile wage is 3.13 euros, the median hourly wage is 4.38 euros and the 75th percentile wage is 6.06 euros. Men's wages are higher at each corresponding percentile, with the 25th percentile wage at 3.22 euros, the median at 4.89 euros and the 75th percentile at 7.13 euros. These figures indicate a consistent pattern of men earning higher wages than women at every level of the wage distribution.

2018 data set description								
Gender	Records	Mean	Standard deviation	Min	25%	50%	75%	Max
F	16811	5.098	3.06	2.14	3.13	4.38	6.06	71.90
M	19974	5.904	4.35	2.13	3.22	4.89	7.13	116.85

In summary as it is visible in the chart below, the 2018 data underscores a persistent gender wage gap with men continuing to earn more than women on average and displaying both a wider variability in wages and significantly higher maximum earnings. This pattern suggests an ongoing concentration of high-paying roles among men contributing to the observed wage disparity between genders.

Figure 2: Distribution of Hourly Pay for Men and Women, year 2018

2.3 Comparing the datasets

The box plots provide a comparison of male and female wages in 2014 and 2018, with each gender represented in separate plots. The center line within each box represents the median hourly pay, showing that both male and female median wages increased slightly from 2014 to 2018, indicating overall wage growth during this period.

For both genders, the interquartile range (IQR)—represented by the height of each box—reveals the spread of wages within the middle 50% of earners. Male wages demonstrate a larger IQR, suggesting greater variability in hourly pay among men compared to women, whose earnings are more concentrated around the median. Tighter IQR for women indicates that female wages are less dispersed and tend to cluster around the median more than male wages.

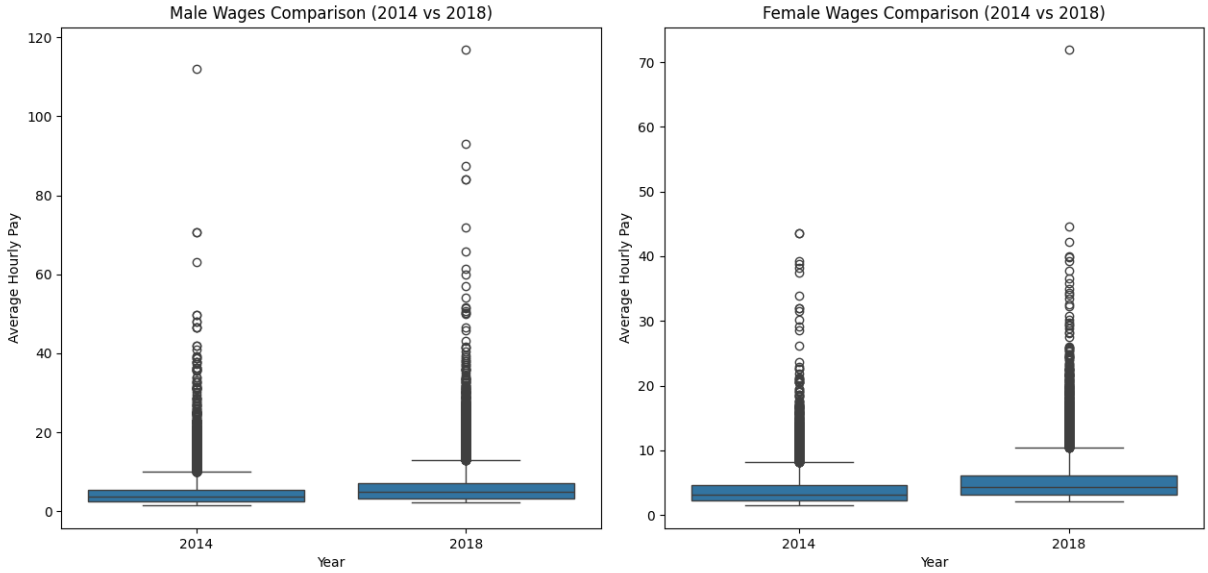
The whiskers extending from each box represent the wage range within 1.5 times the IQR, while individual points beyond the whiskers are considered outliers. Both the male and female plots display a substantial number of outliers, especially in the upper range, signifying that some individuals earn significantly higher wages than the majority of workers. The male plot has a larger number of extreme outliers, indicating that high-earning men have a more dispersed range of wages, whereas high-earning women are less spread out in the upper ranges of income.

When comparing male and female wages, it is evident that men have higher median wages and a broader distribution of high-end earnings in both 2014 and 2018. This consistent discrepancy between genders points to a wage gap, where men generally earn more than women, particularly in the upper tiers of income.

Examining changes over time, both genders show wage growth from 2014 to 2018. However, the increase in wage variability is more pronounced for men especially among high earners suggesting that men in higher-paying positions experienced more substantial wage growth compared to women in similar roles.

Overall, these box plots highlight a persistent gender wage gap, with men not only earning higher median wages but also showing greater dispersion in higher incomes.

Figure 3: Comparing 2014 vs 2018 wages for men and women



2.3.1 Wage ratio

The gender pay ratio is a measure of the average earnings of women compared to men. In this case, the gender pay ratio is calculated by dividing the mean hourly wage of women by the mean hourly wage of men for each respective year.

For 2014, the gender pay ratio is approximately 0.84, meaning that, on average, women earned 84% of what men earned per hour. By 2018, the gender pay ratio had increased slightly

to 0.86, indicating that women earned 86% of men's average hourly wage.

This small increase in the gender pay ratio from 2014 to 2018 suggests a slight narrowing of the gender wage gap, with women's mean earnings per hour growing closer to those of men. However a gap remains still.

2.4 Removing outliers

In order to enhance the integrity of the data analysis an outlier removal function was implemented. This function aims to eliminate extreme values that could skew the results and lead to misleading interpretations, particularly in the context of wage data, where outliers can significantly affect average earnings calculations.

The outlier removal function uses the Interquartile Range (IQR) method [9]. The function is designed to accept a DataFrame (df) and a specific column name (column) as parameters. It proceeds through the following steps:

1. Calculate the First and Third Quartiles: The function begins by calculating the first quartile (Q1) and the third quartile (Q3) of the specified column. Q1 represents the 25th percentile of the data, while Q3 represents the 75th percentile.

2. Determine the Interquartile Range (IQR): The IQR is then computed as the difference between Q3 and Q1 (IQR = Q3 - Q1). This range captures the central 50% of the data and serves as the basis for identifying outliers.

3. Establish Lower and Upper Bounds: Using the IQR, the function calculates the lower bound and upper bound for acceptable values. Specifically, the lower bound is defined as Q1 minus 1.5 times the IQR and the upper bound is defined as Q3 plus 1.5 times the IQR. These bounds help identify data points that are significantly lower or higher than the majority of the dataset.

4. Filter the DataFrame: Finally, the function filters the original DataFrame to include only those rows where the values in the specified column fall within the calculated bounds. This effectively removes the outliers from the dataset.

By employing this outlier removal technique, the analysis aims to provide a more accurate representation of wage distributions and minimize the impact of extreme values. This approach enhances the reliability of subsequent statistical analyses, including mean calculations and regression modeling, ensuring that the conclusions drawn from the data are more reflective of the typical earning patterns.

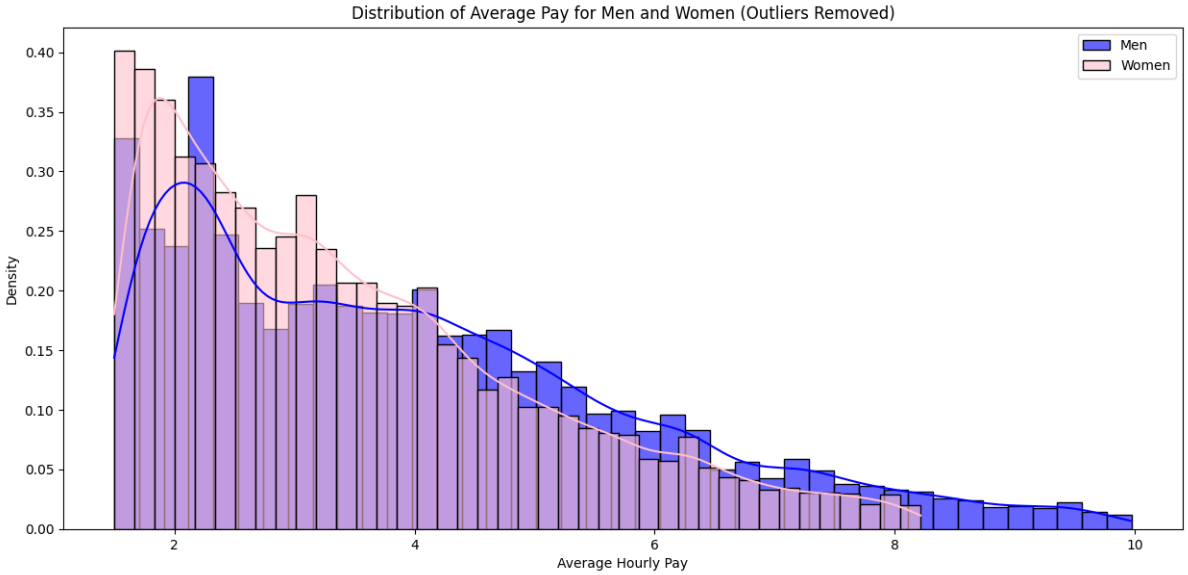
2.4.1 Removing outliers in 2014 dataset

The histogram below shows the distribution of average hourly wages for men and women in 2014 with outliers removed to focus on typical wage patterns. The wage data for men is shown in blue, while the data for women is shown in pink. Additionally, the distribution is overlaid with kernel density estimates (KDE) for each gender giving a smooth curve that helps visualize the wage distribution's general shape.

The graph reveals several key patterns in wage distribution by gender for 2014. Women's wage distribution appears slightly skewed toward lower hourly wages with the highest density around the €2-€3 range. Men's distribution is broader and extends further into higher wage categories, though most of their wages are still concentrated in lower pay ranges, with a peak density also around €2-€3 per hour. As the hourly wage increases beyond this peak, both distributions gradually decline, but men appear slightly more represented in the higher pay brackets compared to women.

The KDE lines highlight that women's wages decrease more sharply than men's as the pay rate increases, suggesting a potential gender wage gap in higher-paying positions.

Figure 4: Distribution of Hourly Pay for Men and Women, year 2014



2.4.2 Removing outliers in 2018 dataset

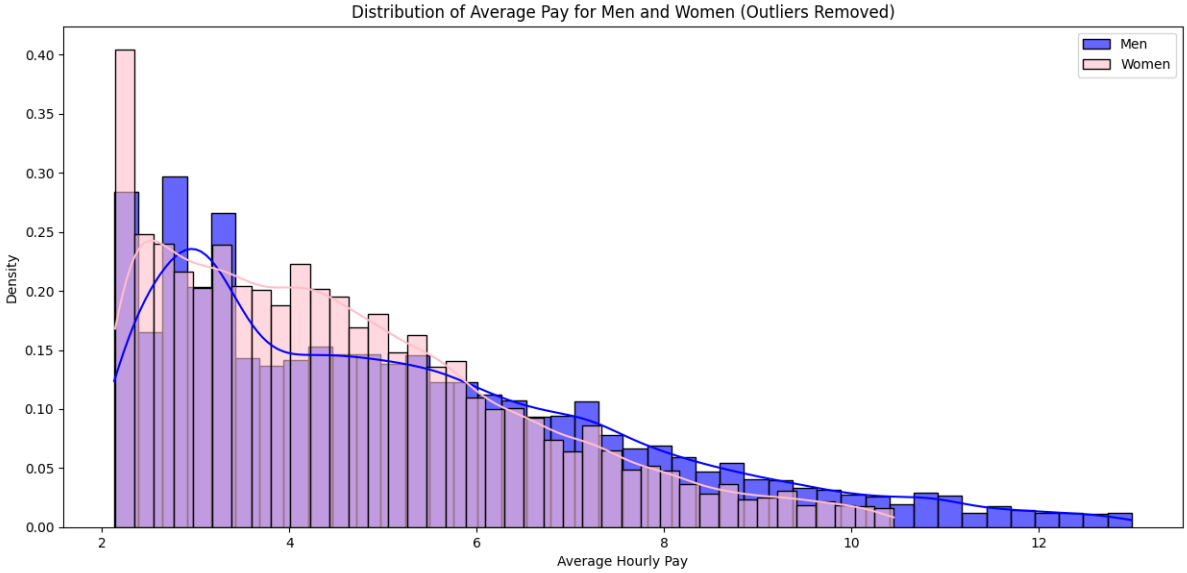
The histogram below shows the distribution of average hourly wages for men and women in 2018 with outliers removed. In the same matter as before, wages of men are displayed in the blue color while women wage data is marked in pink.

Women are more prevalent in the lower wage range, particularly around the €2 mark and they continue to outnumber men in the €4-€6 pay range. As observed in the 2014 dataset, the upper end of the wage distribution is exclusively populated by men, indicating that women are

still largely absent from higher-paying positions. The peak hourly rate for men is around €3, but from €6 onward men consistently earn higher wages than women.

Similar to the pattern observed in the 2014 data, the KDE lines show that women's wages decline more steeply than men's as the pay rate rises.

Figure 5: Distribution of Hourly Pay for Men and Women, year 2018



2.5 Importance of education

The correlation matrices for 2014 and 2018 reveal distinct patterns in the relationships between education level, work experience and hourly pay rates for men and women.

In both years, education level shows a moderate positive correlation with hourly pay for both genders, indicating that higher education levels are generally associated with higher hourly wages. For women, this relationship remains consistent between 2014 and 2018, with a correlation of 0.49, suggesting that education level has a relatively stable impact on their pay rates across the years. For men, this relationship strengthens slightly from 0.38 in 2014 to 0.42 in 2018, implying that education might have become a slightly more significant factor for men's wages over time. Overall, the coeffiecients suggest that education has a stronger influence on women's pay than on men's in both years as indicated by consistently higher correlation values for women.

The relationship between work experience and hourly pay, on the other hand, is weaker for both genders. In 2014, work experience shows a correlation of 0.16 with hourly pay for women and 0.26 for men. This suggests that while work experience does contribute to higher wages, its influence is notably less pronounced than that of education, especially for women. In 2018, the correlation between work experience and hourly pay decreases for both genders, dropping to 0.12 for women and 0.19 for men. This reduction suggests that over time, the impact of work

experience on wages may have diminished for both men and women.

These findings indicate that education level appears to be a more important predictor of hourly pay than work experience, especially for women. This consistent impact of education on women's pay suggests that formal qualifications may play a crucial role in addressing pay gaps. The slightly increasing importance of education for men from 2014 to 2018 might indicate a shift in labor market trends, where educational qualifications are increasingly valued. Meanwhile, the reduced impact of work experience on wages over time could reflect changing labor dynamics, where experience alone may no longer command the same wage premiums as it once did, possibly due to changes in job structures, market demand or employer priorities.

Figure 6: Correlation matrix 2014

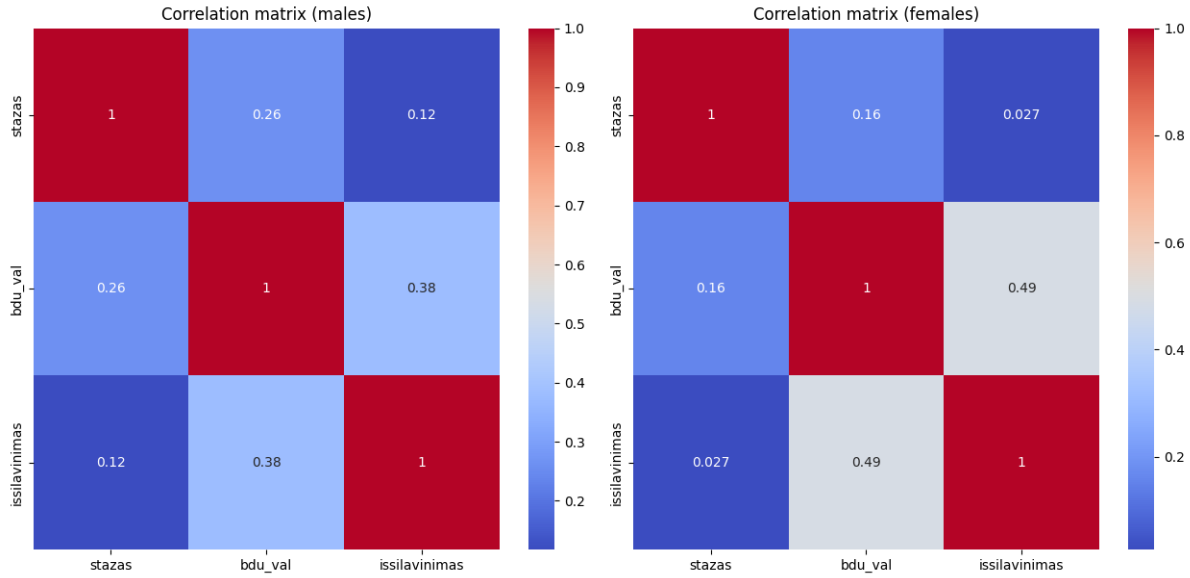
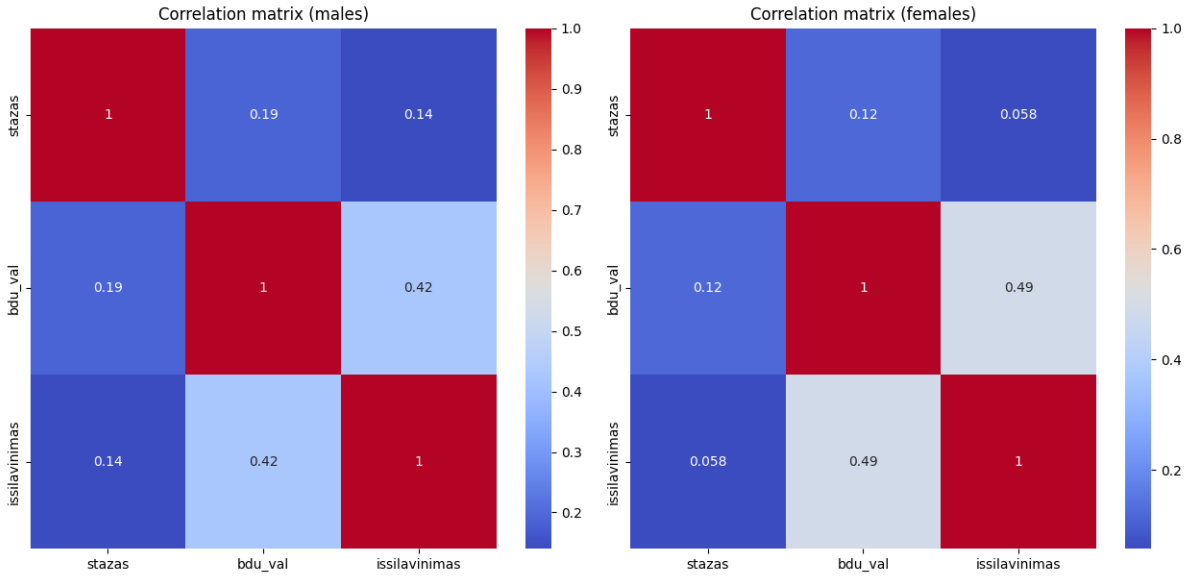


Figure 7: Correlation matrix 2018



2.6 Mean wage based on education level

Explored dataset categorizes education into four distinct levels, each representing progressively advanced qualifications. The first group includes individuals with basic education, which typically provides foundational literacy and numeracy skills but lacks specialized training. The second group represents those who have completed high school or equivalent, often enabling entry-level employment but without higher-level professional qualifications. The third category includes individuals with an undergraduate degree, usually signifying specialized knowledge and skills suited for professional roles. Finally the fourth group encompasses advanced qualifications such as master's or doctoral degrees, often required for specialized or senior-level positions.

The upcoming histograms illustrate how average hourly wages vary across these educational levels, offering insight into how increased education correlates with higher pay rates. By observing wage distribution within each group, we can better understand the wage premium associated with educational attainment and how it may differ across gender.

The histograms display the distribution of average hourly pay by education level for men and women in 2014 and 2018 with four educational categories: primary, secondary, bachelor's degree and master's degree or higher. The graphs show a clear association between education level and wage distribution with higher education generally linked to higher hourly pay.

For both years, individuals with primary and secondary education (levels 1 and 2) tend to cluster in the lower hourly wage range, particularly between €2 and €4. This trend is more pronounced among women, as higher-paid positions are less represented in the lower education levels for them. Men on the other hand exhibit a broader spread across the wage distribution, even at lower education levels, suggesting a slight advantage in wage opportunities even at these

levels.

Bachelor's and master's degree holders (levels 3 and 4) show greater representation in higher hourly wage brackets. In 2014 and 2018, men with higher education levels are more prevalent in higher pay ranges, especially above €6 per hour. Women with higher education levels also see wage increases but tend to cap at a slightly lower range compared to men. This difference is evident from the relatively steeper decline in women's density curves for higher pay, especially in the 2014 dataset.

Comparing the two years, the 2018 data shows a slight increase in the wage distribution for higher-educated groups across both genders. This may indicate some improvement in pay for educated individuals over time although the gender gap persists particularly at higher pay levels. Overall, the graphs show the significant impact of education level on wage distribution, while also highlighting the ongoing gender disparity, especially at the higher end of the pay scale.

Figure 8: Distribution of Average Pay by Education Level 2014

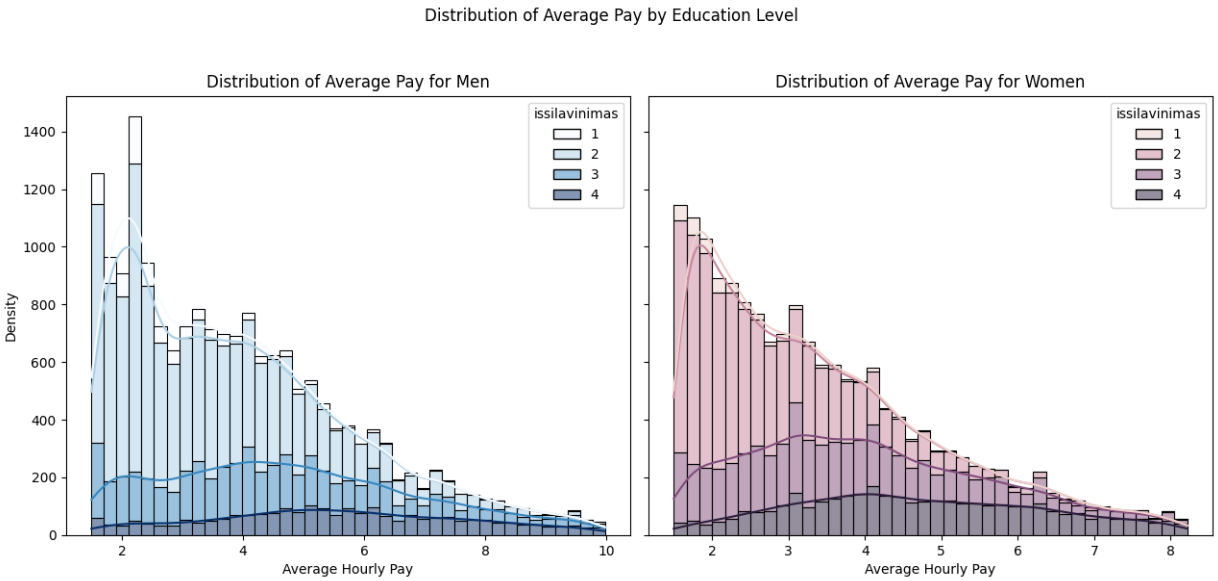
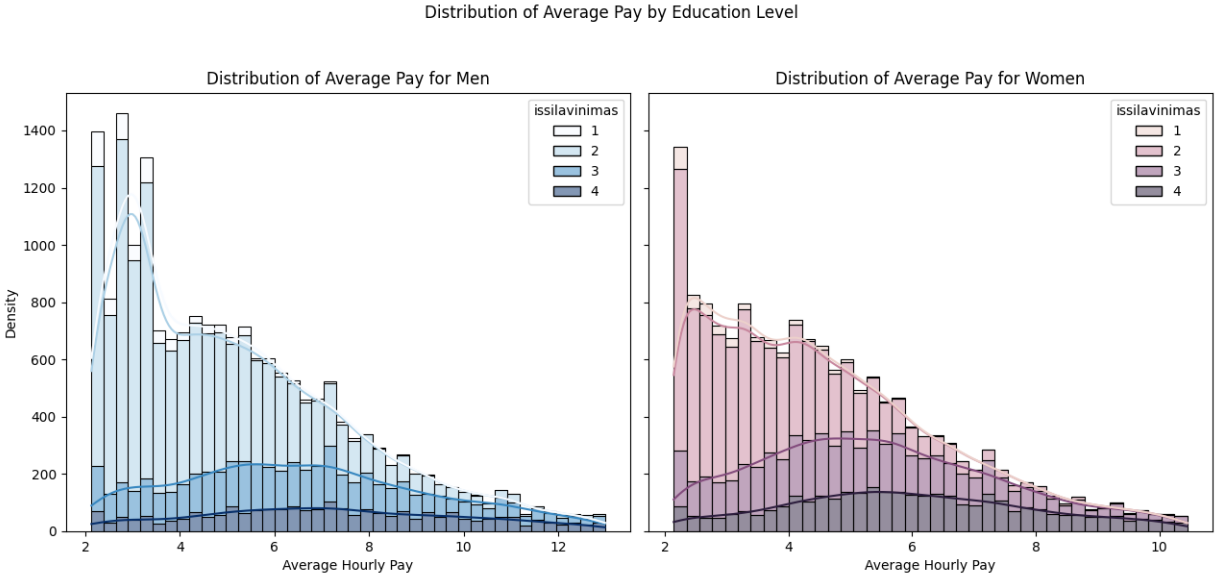


Figure 9: Distribution of Average Pay by Education Level 2018



2.7 Most popular professions

The most popular professions of men and women listed with their codes and the number of workers in each role.

The most popular professions among men:

1. Transport Drivers (Code: 833). Total Workers: 4,044. Includes roles such as bus drivers, truck drivers and other vehicle operators primarily engaged in passenger and cargo transport.

2. Construction Workers (Code: 711). Total Workers: 1,716. Encompasses various roles like masons, carpenters and builders specializing in traditional and non-traditional construction techniques.

3. Firefighters and Security Personnel (Code: 541). Total Workers: 1,670. Comprises roles in firefighting, police work and general security services.

4. Mechanics and Vehicle Repair Technicians (Code: 723). Total Workers: 1,384. Includes roles such as automobile mechanics, truck repair technicians and industrial equipment mechanics.

5. Production and Maintenance Managers (Code: 132). Total Workers: 1,326. Positions include plant managers, engineering heads and supervisors overseeing industrial operations.

The most popular professions among women:

1. Administrative and Office Workers (Code: 241). Total Workers: 2,023. Encompasses roles such as administrative assistants, clerks and other office support staff.

2. Teachers and Education Professionals (Code: 242). Total Workers: 1,985. Primarily includes roles in primary and secondary education.

3. Sales and Retail Workers (Code: 522). Total Workers: 1,502. Includes roles in customer-facing retail environments.

4. Health and Social Care Workers (Code: 222). Total Workers: 1,419. Roles include caregiving, nursing assistants and other healthcare-related support functions.

5. Personal Services Workers (Code: 911). Total Workers: 1,330. Includes occupations like housekeepers, cleaners and other service-related jobs.

Men's professions often emphasize technical skills, heavy physical labor and leadership in industrial or technical domains. Women's professions, by contrast, lean toward interpersonal roles, organizational tasks and caregiving responsibilities. The data highlights traditional gender segmentation in job roles within the labor market.

2.8 Mean wage amongst most popular professions

The chart below presents the average hourly wage (bdu_val) for various professions, differentiated by gender and displayed for two distinct years: 2018 and 2014. The X-axis lists professions represented by numeric codes explained in the subsection above, while the Y-axis reflects the average hourly wage. In each graph, blue columns illustrate the average hourly wage for men and pink columns represent that for women. The left panel focuses on data from 2018 while the right panel compares the same metrics for 2014.

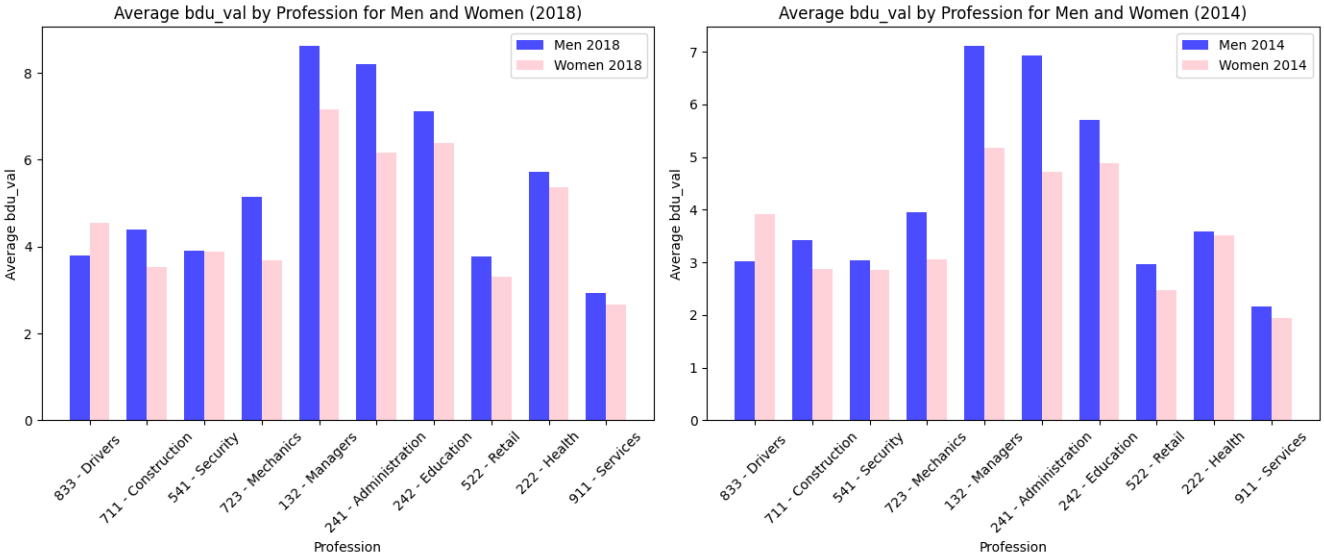
An interesting insight emerges when examining profession 833, which corresponds to drivers of passenger and cargo transport and is one of the most popular professions among men. In this specific profession, women earn a higher average wage than men, making it the only profession among the ten analyzed where this is observed. In all other professions, men consistently earn more than women, although the magnitude of the wage gap varies significantly across different occupations and between the two years.

In professions such as 132 (managers of organizations), the wage disparity is particularly noticable, with men earning significantly more than women in both 2018 and 2014. This suggests a persistent pay gap in higher-paid professions requiring managerial responsibilities. Similarly, for professions like 242 (education professionals) and 222 (health professionals), men also consistently earn more than women, highlighting broader gender disparities in earnings.

However, it is notable that the overall wage levels for both genders appear to improve when moving from 2014 to 2018. This trend may suggest economic shifts (such as minimum wage set

by the state or changing the national currency to euro) or differing labor market conditions over time. Additionally, in some professions, such as 911 (personal service workers), the wage gap between men and women appears narrower compared to other professions, possibly reflecting more standardized pay scales in these roles.

Figure 10: Distribution of Average Pay by Most Popular Professions



For more detailed explanation of what roles the profession codes displayed in the chart include, refer to section 2.7.

3 Methodology

This section defines the methods employed to analyze gender earnings differentiation in Lithuania using data from 2014 and 2018. It details statistical techniques that were used - OLS (Ordinary least squares) regression, SEM (structural equation modelling) and Multilevel modelling (MLM).

3.1 Ordinary Least Squares (OLS)

The analysis includes using Ordinary Least Squares (OLS) regression to model and examine the relationship between employees earnings and various predictor variables. The primary focus is to assess the factors that influence wages with a special emphasis on gender as a key variable of interest in order to understand the wage disparity between men and women in datasets from two distinct time periods 2014 and 2018. The following methodology section is written based on [3] and [20].

Ordinary Least Squares (OLS) is a linear regression technique used to estimate the relationship between a dependent variable (Y) and one or more independent variables ($X_1, X_2, \ldots, X_k$). The basic form of an OLS regression model can be represented as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon, \tag{1}$$

where:

- Y is the dependent variable (response),
- $X_1, X_2, \ldots, X_k$ are the independent variables (predictors),
- β_0 is the intercept of the model,
- $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients of the independent variables,
- ϵ is the error term, accounting for the variation in Y that cannot be explained by the predictors.

The assumptions of OLS include:

1. **Linearity**: The relationship between the dependent and independent variables is linear.
2. **Independence**: The observations are independent of each other.
3. **Homoscedasticity**: The variance of the error term (ϵ) is constant across all levels of the independent variables.
4. **Normality of Errors**: The error term (ϵ) is normally distributed.
5. **No Multicollinearity**: The independent variables are not highly correlated with each other.

These assumptions are tested through diagnostic procedures of residual analysis, variance inflation factors and normality tests. The estimated coefficients provide insights into the magnitude and direction of the effects of each independent variable on earnings.

3.2 Structural Equation Modelling (SEM)

SEM is utilized to estimate complex relationships involving latent variables and multiple indicators. This model combines factor analysis and path analysis in a unified framework [5]. The SEM process involves the following steps:

1. **Model Specification:** A theoretical model is developed, including measurement models for latent constructs (e.g., gender bias, skill level) and structural models for relationships among constructs.

2. **Model Estimation:** Parameters are estimated using maximum likelihood estimation (MLE) or other estimation techniques.

3. **Model Evaluation:** Goodness-of-fit indices such as the Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA) and Chi-square statistic are used to assess the model's fit.

4. **Model Modification:** If necessary, the model is refined by adding or removing paths based on theoretical and empirical considerations.

The general SEM equation is:

$$\eta = B\eta + \Gamma\xi + \zeta \tag{2}$$

where:

- η represents endogenous latent variables.
- ξ represents exogenous latent variables.
- B and Γ are coefficient matrices.
- ζ represents error terms.

Latent constructs are measured using observed indicators and factor loadings quantify the relationship between each latent variable and its indicators. SEM enables the evaluation of complex interdependencies, providing a holistic understanding of earnings differentiation [21].

3.3 Multilevel Modelling

Multilevel modeling (MLM) is particularly suited for situations where data is structured in a hierarchical or nested manner, such as when observations are grouped into higher-level units (students within schools, patients within hospitals, etc. In this case - employees across different professions) [15]. MLM allows for the modeling of both individual-level and group-level variations and it can handle dependencies between observations within groups.

The general form of the multilevel model can be represented as:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + \epsilon_{ij}$$

Where:

- y_{ij} is the outcome for the ith observation in the jth group.
- β_0 is the overall intercept (fixed effect) across all groups.

- β_1 is the slope for the predictor variable x_{ij}, representing the effect of the predictor at the individual level.
- u_j is the random effect for the jth group, which captures the group-specific deviation from the overall intercept.
- ϵ_{ij} is the residual error term for the ith observation in group j.

At the group level, the random effect u_j is assumed to vary according to:

$$u_j = \gamma_0 + v_j$$

Where:

- γ_0 is the average random effect across all groups (fixed effect).
- v_j is the error term at the group level, assumed to be normally distributed with mean 0 and variance σ_u^2.

The errors ϵ_{ij} and v_j are assumed to be independent of each other and normally distributed.

The multilevel model is estimated using maximum likelihood estimation (MLE) or restricted maximum likelihood estimation (REML). These methods account for the hierarchical structure of the data, providing unbiased estimates for both fixed and random effects [12].

Multilevel models rely on several assumptions, including:

- The random effects (u_j and v_j) are normally distributed with mean 0 and some variance.
- The residual errors (ϵ_{ij}) are independent and identically distributed with mean 0 and constant variance.
- The individual-level predictors x_{ij} are not perfectly collinear.

The results of the multilevel model provide estimates for both the fixed effects (e.g., the overall intercept and slope) and the random effects (e.g., the variation between groups). These results allow to examine both individual-level and group-level predictors of the outcome variable.

The fixed effects provide the average relationship between predictors and the outcome, while the random effects offer insight into how much variation exists between groups [8].

4 Modelling

4.1 Ordinary Least Squares analysis

The results of the Ordinary Least Squares (OLS) regression models for 2014 and 2018 highlight significant factors influencing hourly wages (bdu_val) for both men and women. These

models reveal the importance of education level (issilavinimas), company size (im_dydzio_kodas) and work experience (stazas) to earnings.

In the 2014 male model, the intercept value of 1.87 euros represents the predicted hourly wage for men with primary education working in the smallest company size, without experience. The coefficient for stazas is 0.044, indicating that each additional year of experience is associated with a 4.4 cent increase in hourly wages. Education levels show significant positive effects on wages. Secondary education (issilavinimas_2) increases wages by 0.53 euros compared to primary education, while a bachelor's degree (issilavinimas_3) contributes an additional 1.57 euros and a master's degree or higher (issilavinimas_4) adds 2.36 euros. Company size also has a significant impact. Working in a medium-sized company (im_dydzio_kodas_1) increases wages by 0.84 euros, while being in a large company (im_dydzio_kodas_2) adds 1.23 euros to hourly pay. These results suggest that higher education levels and employment in larger organizations are strongly associated with better wages for men in 2014.

Table 1: Male Regression Model (2014)

Variable	**Coefficient**	**Std. Error**	**t-value**	**P>|t|**
Intercept	1.8666	0.058	32.317	0.000
Stazas	0.0444	0.002	27.879	0.000
Issilavinimas_2	0.5329	0.055	9.712	0.000
Issilavinimas_3	1.5738	0.058	27.098	0.000
Issilavinimas_4	2.3558	0.063	37.234	0.000
Im_dydzio_kodas_1	0.8368	0.034	24.964	0.000
Im_dydzio_kodas_2	1.2331	0.032	38.847	0.000
R-squared	0.253	Adj. R-squared: 0.253		
F-statistic	1046.0	Prob(F-statistic): 0.000		

The 2014 female model presents a similar pattern, but with some notable differences. The intercept value is slightly lower at 1.79 euros, reflecting a smaller base wage for women with primary education and no experience in the smallest companies. The coefficient for stazas is smaller at 0.020, indicating that each additional year of experience increases wages by only 2 cents for women, half the increase observed for men. Education still plays a significant role, but wage increases are slightly less pronounced. Secondary education (issilavinimas_2) raises wages by 0.24 euros, while a bachelor's degree (issilavinimas_3) adds 1.32 euros and a master's degree or higher (issilavinimas_4) contributes 2.14 euros. The effects of company size are also present, though smaller compared to men. Medium-sized companies (im_dydzio_kodas_1) increase wages by 0.53 euros, while large companies (im_dydzio_kodas_2) add 0.74 euros. These differences highlight a gender wage gap, with women receiving lower returns from education, experience and company size compared to men in 2014.

Table 2: Female Regression Model (2014)

Variable	**Coefficient**	**Std. Error**	**t-value**	**P>|t|**
Intercept	1.7914	0.067	26.919	0.000
Stazas	0.0199	0.001	18.365	0.000
Issilavinimas_2	0.2443	0.064	3.809	0.000
Issilavinimas_3	1.3219	0.065	20.275	0.000
Issilavinimas_4	2.1370	0.066	32.314	0.000
Im_dydzio_kodas_1	0.5272	0.030	17.689	0.000
Im_dydzio_kodas_2	0.7359	0.027	26.918	0.000
R-squared	0.296	Adj. R-squared: 0.296		
F-statistic	1189.0	Prob(F-statistic): 0.000		

In 2018, the male model demonstrates an increase in the base wage, with the intercept increasing to 2.72 euros. Experience continues to have a significant effect, with a coefficient of 0.037, indicating a 3.7 cent increase in hourly pay per additional year of experience. Education levels show greater returns compared to 2014. Secondary education adds 0.56 euros to wages, while a bachelor's degree adds 2.48 euros and a master's degree or higher contributes 3.13 euros. The effect of company size remains strong, with medium-sized companies increasing wages by 1.29 euros and large companies adding 1.35 euros. These findings suggest that the labor market for men in 2018 rewards higher education and larger company sizes more significantly than in 2014, indicating upward trends in wage premiums associated with these factors.

Table 3: Male Regression Model (2018)

Variable	**Coefficient**	**Std. Error**	**t-value**	**P>|t|**
Intercept	2.7209	0.080	33.969	0.000
Stazas	0.0374	0.002	19.057	0.000
Issilavinimas_2	0.5626	0.078	7.230	0.000
Issilavinimas_3	2.4793	0.082	30.226	0.000
Issilavinimas_4	3.1261	0.088	35.553	0.000
Im_dydzio_kodas_1	1.2927	0.040	32.054	0.000
Im_dydzio_kodas_2	1.3532	0.038	35.409	0.000
R-squared	0.266	Adj. R-squared: 0.265		
F-statistic	1145.0	Prob(F-statistic): 0.000		

The 2018 female model similarly reflects an increase in base wages, with an intercept of 2.50 euros. However, the returns from experience are even lower than in 2014, with a coefficient of 0.016, representing a 1.6 cent increase in wages per additional year of experience. Education continues to significantly impact wages, with secondary education adding 0.42 euros, a bachelor's

degree contributing 1.97 euros and a master's degree or higher raising wages by 2.60 euros. The effects of company size are also pronounced, with medium-sized companies increasing wages by 0.68 euros and large companies adding 1.02 euros. While women's wages have improved over time, the returns from experience remain disproportionately low compared to men and education and company size contribute slightly less to wage increases for women.

Table 4: Female Regression Model (2018)

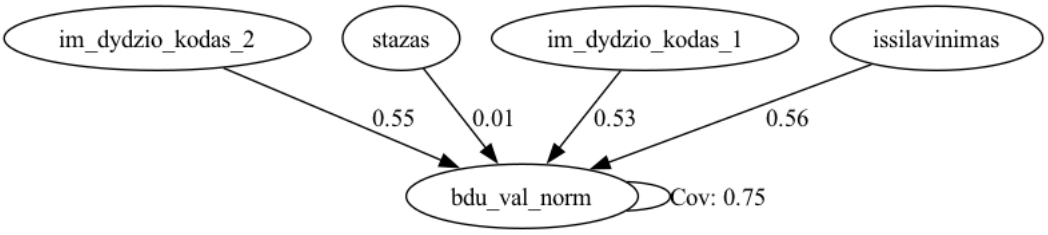
Variable	**Coefficient**	**Std. Error**	**t-value**	**P>|t|**
Intercept	2.4959	0.083	30.068	0.000
Stazas	0.0159	0.001	11.523	0.000
Issilavinimas_2	0.4180	0.082	5.125	0.000
Issilavinimas_3	1.9726	0.083	23.806	0.000
Issilavinimas_4	2.6039	0.084	30.941	0.000
Im_dydzio_kodas_1	0.6758	0.034	19.854	0.000
Im_dydzio_kodas_2	1.0167	0.032	31.970	0.000
R-squared	0.300	Adj. R-squared: 0.300		
F-statistic	1143.0	Prob(F-statistic): 0.000		

Across both years, the models show that men consistently see higher base wages, larger returns from experience and slightly greater wage increase from education and company size. This suggests structural differences in how the labor market values men and women's qualifications and work environments. The 2018 results indicate improvements in wages for both genders, likely reflecting broader economic growth, but the gender wage gap persists. The findings also highlight the importance of education and the size of the company as significant factors in determining wages.

4.2 SEM analysis

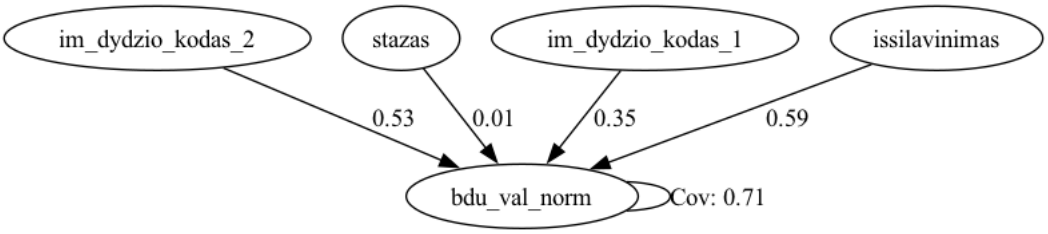
For 2018 male data, the model-implied covariance matrix yielded a chi-squared statistic of 8321706.27, a chi-squared statistic per observation of 438.03 and an RMSEA of 6.62, indicating a substantial lack of model fit. Education had a significant positive effect on average hourly wage with estimate of 0.555, $p < 0.001$, while job experience also positively influenced the outcome estimating at 0.014, $p < 0.001$). Company size contributed significantly and the variance of average hourly wage was estimated at 0.748.

Figure 11: SEM Results for 2018 Male Data



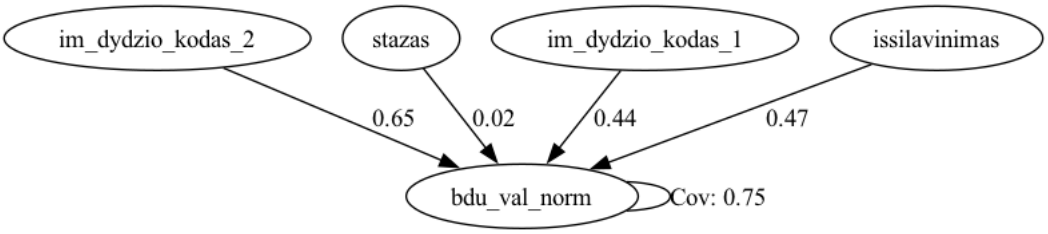
For 2018 female data, the model fit was worse, with a chi-squared statistic of 9757948.67, a chi-squared statistic per observation of 610.48 and an RMSEA of 7.81. Education had a slightly higher impact on average hourly wage compared to males with estimate at 0.589, while job experience showed a reduced influence estimating at 0.006. The effects of company size were also smaller than for males and the variance of average hourly wage was estimated at 0.713.

Figure 12: SEM Results for 2018 Female Data



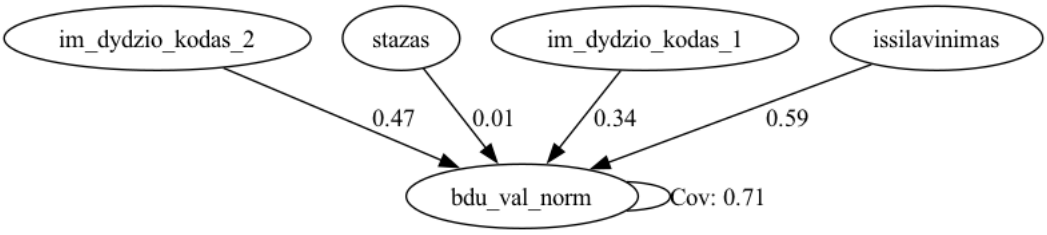
For 2014 male data, the chi-squared statistic was 8785247.89, with a chi-squared per observation of 474.70 and an RMSEA of 6.89, suggesting an improved fit relative to the 2018 female data but still inadequate. Influence of education on average hourly wage decreased compared to 2018 with estimate at 0.471, while the impact of job experience increased with coefficient of 0.023. Company size also had substantial positive effects.

Figure 13: SEM Results for 2014 Male Data



For 2014 female data exhibited the poorest fit, with a chi-squared statistic of 12191108.92, a chi-squared per observation of 717.84 and an RMSEA of 8.47. Despite this, education remained the strongest predictor of average hourly wage with estimate of 0.594, while effect of job experience was modest estimating at 0.012. Company size had a diminished influence relative to males.

Figure 14: SEM Results for 2014 Female Data



These results underscore temporal and gender-based differences in the predictors' effects on average hourly wage with variations in model fit and parameter estimates highlighting the complexity of the relationships.

4.3 Multilevel Modelling

Multilevel modelling has been applied to the group of individuals that work in the most popular professions across men and women. 10 professions were chosen and the manner of choosing them was defined in the section 2.7.

Starting with the 2018 male data, the model converges successfully with 5,808 observations across 10 groups. The intercept is significantly negative at -0.795, indicating a baseline lower level of the dependent variable when all predictors are at their reference categories. Education shows a clear positive effect, with the coefficients for higher levels of education ($issilavinimas\_3$ and $issilavinimas\_4$) being significant and progressively larger, suggesting that higher education levels are associated with higher wages. Work experience (stazas) has a small but highly significant positive effect (0.029). Company size also exhibits a significant impact, with larger companies ($im\_dydzio\_kodas\_1$ and $im\_dydzio\_kodas\_2$) associated with higher (bdu_val_norm) values. The group variance is moderate (0.226), indicating some variability across 10 profession groups.

Table 5: Multilevel Model Results for 2018 Male Data

Variables	**Coef.**	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-0.795	0.164	-4.858	0.000	-1.116	-0.474
issilavinimas_2	0.173	0.053	3.251	0.001	0.069	0.277
issilavinimas_3	0.573	0.064	8.907	0.000	0.447	0.699
issilavinimas_4	0.593	0.079	7.507	0.000	0.439	0.748
stazas	0.029	0.001	19.670	0.000	0.026	0.032
im_dydzio_kodas_1	0.643	0.032	20.284	0.000	0.581	0.705
im_dydzio_kodas_2	0.713	0.029	24.282	0.000	0.655	0.770
Group Var	0.226	0.134				

For females in 2018, the model includes 4,116 observations and also converges successfully. The intercept is slightly less negative than for males at -0.666. Unlike males, the coefficient for education level *issilavinimas_2* is not significant, suggesting no notable effect for this category. However, higher education levels *issilavinimas_3* and *issilavinimas_4* maintain significant positive effects. Work experience (stazas) does not have a significant effect, contrasting with the male model. Company size remains a significant factor, though the coefficients are smaller compared to males, implying a less pronounced influence. Group variance is slightly lower at 0.193.

Table 6: Multilevel Model Results for 2018 Female Data

Variables	**Coef.**	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-0.666	0.163	-4.086	0.000	-0.986	-0.347
issilavinimas_2	0.059	0.074	0.796	0.426	-0.087	0.205
issilavinimas_3	0.452	0.082	5.515	0.000	0.291	0.613
issilavinimas_4	0.673	0.085	7.944	0.000	0.507	0.839
stazas	0.002	0.001	1.247	0.213	-0.001	0.004
im_dydzio_kodas_1	0.325	0.032	10.030	0.000	0.262	0.389
im_dydzio_kodas_2	0.370	0.030	12.429	0.000	0.312	0.428
Group Var	0.193	0.131				

In the 2014 male data, the model with 5,240 observations shows a larger negative intercept at -1.080, suggesting a lower baseline compared to 2018. The effects of education are again significant, with *issilavinimas_4* showing the largest positive impact with coefficient of 0.745, $p < 0.001$. Work experience maintains a small positive and significant impact with coefficient 0.027. Company size has significant positive effects, with the coefficient for *im_dydzio_kodas_2* being particularly large at 0.885, reflecting a strong association between larger company size and higher wage. The group variance is higher at 0.300.

Table 7: Multilevel Model Results for 2014 Male Data

Variables	**Coef.**	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-1.080	0.183	-5.899	0.000	-1.439	-0.721
issilavinimas_2	0.361	0.044	8.222	0.000	0.275	0.447
issilavinimas_3	0.517	0.055	9.461	0.000	0.410	0.625
issilavinimas_4	0.745	0.066	11.266	0.000	0.615	0.874
stazas	0.027	0.002	17.151	0.000	0.024	0.030
im_dydzio_kodas_1	0.498	0.032	15.473	0.000	0.435	0.562
im_dydzio_kodas_2	0.885	0.031	28.516	0.000	0.824	0.945
Group Var	0.300	0.184				

For females in 2014, the model with 4,581 observations also shows a significant intercept of -0.703 . Similar to the 2018 female model, the effect of *issilavinimas_2* is not significant. Higher education levels (*issilavinimas_3* and *issilavinimas_4*) continue to exhibit strong positive effects, with *issilavinimas_4* being particularly pronounced, coeficcient at 0.682, $p < 0.001$. Work experience has a small but significant positive effect with 0.005, $p < 0.001$, which is notable given its insignificance in 2018. Company size has significant effects, with *im_dydzio_kodas_2* at 0.460 showing a stronger association than *im_dydzio_kodas_1*. Group variance is moderate at 0.208.

Table 8: Multilevel Model Results for 2014 Female Data

Variables	**Coef.**	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-0.703	0.168	-4.193	0.000	-1.032	-0.374
issilavinimas_2	0.075	0.075	1.003	0.316	-0.072	0.222
issilavinimas_3	0.341	0.080	4.267	0.000	0.184	0.498
issilavinimas_4	0.682	0.082	8.319	0.000	0.522	0.843
stazas	0.005	0.001	3.761	0.000	0.002	0.008
im_dydzio_kodas_1	0.290	0.035	8.194	0.000	0.221	0.360
im_dydzio_kodas_2	0.460	0.032	14.556	0.000	0.398	0.521
Group Var	0.208	0.138				

4.4 Multilevel Modelling without categorical education levels

For the 2018 male data, the model included 5808 observations across 10 groups, with a scale of 0.6830 and a log-likelihood of -7168.86. The fixed effects revealed that education, job experience and company size were all significant predictors, with positive coefficients. The group variance was estimated at 0.233.

Table 9: Multilevel Model Results for 2018 Male Data

Variables	**Coef.**	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-1.084	0.169	-6.425	0.000	-1.414	-0.753
issilavinimas	0.245	0.024	10.064	0.000	0.197	0.293
stazas	0.028	0.001	19.309	0.000	0.025	0.031
im_dydzio_kodas_1	0.652	0.032	20.566	0.000	0.590	0.714
im_dydzio_kodas_2	0.721	0.029	24.536	0.000	0.663	0.778
Group Var	0.233	0.137				

Similarly, for 2018 female data, with 4116 observations and a scale of 0.5237, the log-likelihood was -4542.88. Education and company size remained strong predictors, while job experience had no significant effect. The group variance was 0.212, slightly lower than for males.

Table 10: Multilevel Model Results for 2018 Female Data

Variables	**Coef.**	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-1.117	0.160	-6.984	0.000	-1.431	-0.804
issilavinimas	0.279	0.020	13.630	0.000	0.238	0.319
stazas	0.001	0.001	0.452	0.651	-0.002	0.003
im_dydzio_kodas_1	0.325	0.032	9.991	0.000	0.261	0.388
im_dydzio_kodas_2	0.374	0.030	12.582	0.000	0.316	0.432
Group Var	0.212	0.143				

In the 2014 datasets, the male data included 5240 observations with a scale of 0.6283 and log-likelihood of -6253.78. Education, job experience and company size were again significant predictors, with a notable increase in the coefficient for bigger companies, indicating a stronger effect of larger company sizes. The group variance was estimated at 0.286, higher than the 2018 male data.

Table 11: Multilevel Model Results for 2014 Male Data

Variables	**Coef.**	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-1.210	0.181	-6.673	0.000	-1.565	-0.855
issilavinimas	0.228	0.021	10.954	0.000	0.187	0.269
stazas	0.027	0.002	17.625	0.000	0.024	0.030
im_dydzio_kodas_1	0.495	0.032	15.376	0.000	0.432	0.559
im_dydzio_kodas_2	0.875	0.031	28.420	0.000	0.815	0.936
Group Var	0.286	0.176				

For 2014 female data, based on 4581 observations with a scale of 0.5302, education, job experience and company size categories were significant, with experience showing a modest but positive effect. The group variance was 0.205, consistent with the 2018 female data.

Table 12: Multilevel Model Results for 2014 Female Data

Variables	**Coef.**	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-1.206	0.156	-7.722	0.000	-1.513	-0.900
issilavinimas	0.288	0.018	16.039	0.000	0.253	0.323
stazas	0.005	0.001	4.093	0.000	0.003	0.008
im_dydzio_kodas_1	0.290	0.035	8.192	0.000	0.221	0.360
im_dydzio_kodas_2	0.469	0.031	14.956	0.000	0.408	0.531
Group Var	0.205	0.136				

Overall, education and comapny size consistently emerged as strong predictors across all datasets, while the effects of job experience varied by gender and year.

5 Results

The analysis of wage determining factors across multilevel, OLS and SEM models for the years 2014 and 2018 reveals consistent trends but also highlights significant gender-based differences in the variables influencing wages.

The OLS results for 2014 indicated that education level, work experience and company size all had a positive and statistically significant impact on wages for both men and women. However, the effect of these factors was more pronounced for men. In the male model, each additional year of experience contributed a 4.4 cent increase in hourly wages and higher educational attainment led to substantial wage increases, with a master's degree or higher contributing 2.36 euros to hourly wages. For women, the coefficient for experience was much smaller (0.02) and while education continued to boost wages, the impact was smaller compared to men. Company size also played a role, with medium-sized and large companies offering higher wages to both men and women, though the effect was again stronger for men.

By 2018, the OLS model showed an upward shift in base wages, with the male model's intercept rising to 2.72 euros. Experience continued to contribute to higher wages, though at a slightly reduced rate (3.7 cents per year). The return on education increased for men, with a bachelor's degree adding 2.48 euros and a master's degree adding 3.13 euros to hourly wages, marking a significant increase from 2014. Women also saw increased wages in 2018, with a base wage of 2.50 euros and a slightly higher return from education, though the gap between male and female wages persisted. For women, the effect of experience continued to be weaker than for

men and although company size still influenced wages, the effect was less pronounced compared to men.

The SEM analysis, even though with poor fit, confirmed these findings, particularly regarding the strong positive impact of education and company size on normalized hourly wages. In both years, education showed a significant positive effect on wages for both genders, with men benefiting more than women. The influence of experience was positive but weaker for both genders, with a notable decrease in its impact from 2014 to 2018, especially for men. Large company size, consistently showed a positive impact on wages for both genders, with women benefiting more from working in larger firms.

For the 2018 data, the multilevel model for males revealed a significantly negative intercept of -0.795, indicating a baseline lower level of wages across inspected professions when all predictors were set at their reference categories. The positive effect of education on wages was clear, with higher levels of education ($issilavinimas\_3$ and $issilavinimas\_4$) associated with progressively higher wages. Company size also played a key role in determining wages, with larger companies correlating with higher wages. Interestingly, while work experience had a positive effect, its impact was relatively small, suggesting that other factors, such as education and company size, were more influential in this context. The group variance of 0.226 showed moderate variability across professions, indicating that while profession-related factors did matter, they were not overwhelming other predictors.

The female multilevel model for 2018 displayed similar patterns, though there were notable differences. The intercept for females was slightly less negative at -0.666, indicating a slightly higher baseline wage compared to males. Education continued to have a positive and significant effect on wages, but the coefficient for $issilavinimas\_2$ was not significant, suggesting that this level of education did not have the same impact on female wages as it did for males. Unlike the male model, work experience was not a significant predictor of wages for females, highlighting potential differences in how experience is valued in female-dominated professions. The group variance for females was lower at 0.193, suggesting less variability across professions compared to males.

For the 2014 data, the patterns were largely consistent with those from 2018, although the magnitude of coefficients and group variances differed slightly. The male 2014 model showed a larger negative intercept of -1.080, indicating lower baseline wages compared to 2018. Education played an even more significant role in 2014, with the highest level of education ($issilavinimas\_4$) showing a particularly large positive coefficient of 0.745, which was highly significant. Company size also had more impact in 2014, with the coefficient for $im\_dydzio\_kodas\_2$ (largest companies) at 0.885, emphasizing the strong association between larger company sizes and higher wages. The group variance for the male 2014 model was larger at 0.300, suggesting more substantial variability across profession groups compared to the 2018 model.

For females in 2014, the results mirrored those of 2018 in many respects, though with some differences. The intercept was significant at -0.703 and higher education levels continued to show strong positive effects on wages. Work experience, unlike in 2018, had a small but significant positive effect, suggesting a possible shift over time in how experience is valued in female-dominated professions. Company size had significant positive effects as well, with the coefficient for $im\_dydzio\_kodas\_2$ (largest companies) being 0.460, underscoring the importance of company size for female wages. The group variance for the 2014 female model was slightly higher at 0.208 compared to the 2018 female model, suggesting some decrease in variability across professions over time.

Overall, while the three models revealed consistent findings regarding the importance of education, company size and work experience in determining wages, they also highlighted key gender differences. Education and company size were universally important wage determinants, though their impact was often more pronounced for men. Experience played a significant role in the male wage models but had less influence for women. The comparison revealed a shift in wages between 2014 and 2018, with increased returns to education and more pronounced wage differences between men and women.

6 Conclusions

The analysis confirms a persistent gender wage gap in Lithuania, with men consistently earning higher wages than women. Although the SEM and MLM models did not achieve a strong fit, both models point to the same conclusion: men get greater benefits from higher education levels, longer work experience and working in larger companies, even within sectors that are mostly occupied by women and could be considered female-dominated. The findings also align with global studies as it has been confirmed that within the same job, women are paid less than men in Lithuania, suggesting underlying discrimination in the labor market.

Ordinary Least Squares (OLS) model reinforced the finding that women earn less than men considering key factors such as education level, company size and work experience. While wages improved for both genders between 2014 and 2018 which provides a foundation for positive change, structural disparities in wage outcomes remain evident. Higher education levels positively influence wages for both men and women, but the returns are consistently greater for men, widening the wage gap. Similarly, work experience plays a significant role in determining wages, with men getting higher benefits, indicating potential differences in career advancement opportunities or access to higher-paying roles.

The findings highlight significant gender disparities in wages, but the limited fit of the SEM and MLM models suggests room for improvement in modeling approaches. Future research could benefit from exploring additional variables, such as family responsibilities or career interruptions, which may contribute to wage differences. Improved data collection, such as deeper

insights in occupational sector defining not only the sphere, but the role the employee has, and refinement of analytical methods could also lead to more precise models and a better understanding of the dynamics shaping wage inequalities in Lithuania. The need for continued efforts addressing wage disparities is crucial, understanding the root causes of wage inequality could open a possibility to envision a future where the gender wage gap is effectively eliminated, ensuring equal opportunities and outcomes for all workers.

References

[1] Damaske S, Bratter JL, Frech A. Single mother families and employment, race, and poverty in changing economic times. 2017.

[2] Tera Allas. "Female" jobs pay significantly less than "male" ones for a given level of education. 2022.

[3] Valentina Alto. Understanding Ordinary Least Squares OLS Regression. 2023.

[4] European Institue for Gender Equality. Financial independence and gender equality. Joining the dots between income, wealth, and power. 2024.

[5] James C. Anderson, David W. Gerbing. Structural equation modeling in practice: A review and recommended two-step approach. 1988.

[6] Claudia Goldin. A grand gender convergence: Its last chapter. *American Economic Review*, 2014.

[7] Michelle Gyimah. How women are penalised for negotiating. 2022.

[8] Ronald Heck and Scott L. Thomas. An introduction to multilevel modeling techniques. 2020.

[9] Sergio Hleap. Unmasking the outliers: Exploring the interquartile range method for reliable data analysis.

[10] Francine D. Blau, Lawrence M. Kahn. The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 2017.

[11] Rakesh Kochhar. The enduring grip of the gender pay gap. 2023.

[12] Sunil Kumar. A brief introduction to multilevel modelling. 2024.

[13] Edvinas Kučinskas. 'Women are still less self-confident than men' - gender pay gap persists in Lithuania. 2022.

[14] OECD. Gender pay gap reporting and equal pay audits lessons learned across oecd countries. 2023.

[15] Rasbash Jon, University of Bristol. What are multilevel models and why should I use them? 2024.

[16] European Parliament. Moterų ir vyrų darbo užmokesčio skirtumai ES. 2020.

[17] 1 Katie A. Anders Rhea E. Steinpreis and Dawn Ritzke. The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. 1999.

[18] Barbara Downs YoonKyung Chung Robert Sienkiewicz, Danielle H. Sandler. The parental gender earnings gap in the united states. 2017.

[19] Christine Siegwarth Meyer, Swati Mukerjee, Ann Sestero. Work-family benefits: Which ones maximize profits? *Journal of Managerial Issues*, 13, 2001.

[20] Departament of Economics Simon Fraser University. Ordinary least squares. 2011.

[21] Dragan Dejan Topolšek Darja. Introduction to structural equation modeling: Review, methodology and practical applications. 2014.

Appendix A

All images used in this work were generated by the author using Python coding language.

Appendix B

In the writing process of this work, no generative artificial intelligence (AI) or AI-assisted technologies were utilized to generate content. AI was used solely for the purpose of enhancing readability, refining language and debugging code. This use was under strict human oversight and control. The author carefully reviewed and edited this thesis to ensure its accuracy and coherence after the application of AI technologies. The code was also thoroughly revised by human eyes and judgment.

Appendix C

The Python code created while working on the thesis is presented here.

EDA

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

#EXPLORATORY DATA ANALYSIS
df_2014 = pd.read_csv('/Users/samuleviciute/Downloads/magistras/2014.csv')
description =  df_2014.groupby('lytis')['bdu_val'].describe()
print(description)
df_male = df_2014[df_2014['lytis'] == 'M']
df_female = df_2014[df_2014['lytis'] == 'F']

# FIGURE 1
plt.figure(figsize=(10, 6))
sns.histplot(df_male['bdu_val'], kde=True, color='blue', label='Men',
stat="density", alpha=0.6)
sns.histplot(df_female['bdu_val'], kde=True, color='pink', label='Women',
stat="density", alpha=0.6)

plt.title('Distribution of Hourly Pay for Men and Women')
plt.xlabel('Hourly Pay')
```

```
plt.ylabel('Density')
plt.legend()
plt.tight_layout()
plt.show()


df_2018 = pd.read_csv('/Users/samuleviciute/Downloads/magistras/2018.csv')
description =  df_2018.groupby('lytis')['bdu_val'].describe()
print(description)
df_male = df_2018[df_2018['lytis'] == 'M']
df_female = df_2018[df_2018['lytis'] == 'F']

# FIGURE 2
plt.figure(figsize=(10, 6))
sns.histplot(df_male['bdu_val'], kde=True, color='blue', label='Men',
stat="density", alpha=0.6)
sns.histplot(df_female['bdu_val'], kde=True, color='pink', label='Women',
stat="density", alpha=0.6)

plt.title('Distribution of Hourly Pay for Men and Women')
plt.xlabel('Hourly Pay')
plt.ylabel('Density')
plt.legend()
plt.tight_layout()
plt.show()


#FIGURE 3
# Combine both datasets into one DataFrame
df_combined = pd.concat([df_2014, df_2018])

plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
sns.boxplot(x='year_', y='bdu_val', data=df_combined[df_combined['lytis'] == 'M'])
plt.title('Male Wages Comparison (2014 vs 2018)')
plt.xlabel('Year')
plt.ylabel('Average Hourly Pay')

plt.subplot(1, 2, 2)
sns.boxplot(x='year_', y='bdu_val', data=df_combined[df_combined['lytis'] == 'F'])
```

```
plt.title('Female Wages Comparison (2014 vs 2018)')
plt.xlabel('Year')
plt.ylabel('Average Hourly Pay')

plt.tight_layout()
plt.show()


#SECTION 2.3.1
gender_pay_ratio = df_2014[df_2014['lytis'] == 'F']['bdu_val'].mean() /
df_2014[df_2014['lytis'] == 'M']['bdu_val'].mean()
print(gender_pay_ratio)
gender_pay_ratio = df_2018[df_2018['lytis'] == 'F']['bdu_val'].mean() /
df_2018[df_2018['lytis'] == 'M']['bdu_val'].mean()
print(gender_pay_ratio)

def remove_outliers(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return df[(df[column] >= lower_bound) & (df[column] <= upper_bound)]

#REMOVING OUTLIERS 2014
df_male = df_2014[df_2014['lytis'] == 'M']
df_female = df_2014[df_2014['lytis'] == 'F']
df_male_clean = remove_outliers(df_male, 'bdu_val')
df_female_clean = remove_outliers(df_female, 'bdu_val')

# FIGURE 4
plt.figure(figsize=(10, 6))
sns.histplot(df_male_clean['bdu_val'], kde=True, color='blue', label='Men',
stat="density", alpha=0.6)
sns.histplot(df_female_clean['bdu_val'], kde=True, color='pink',
label='Women', stat="density", alpha=0.6)
plt.title('Distribution of Average Pay for Men and Women (Outliers Removed)')
plt.xlabel('Average Hourly Pay')
plt.ylabel('Density')
plt.legend()
```

```
plt.tight_layout()
plt.show()

#FIGURE 6
plt.figure(figsize=(14, 6))
plt.subplot(1, 2, 1)
plt.title('Correlation matrix (males)')
sns.heatmap(df_male_clean[['stazas', 'bdu_val', 'issilavinimas']].corr(),
annot=True, cmap='coolwarm')
plt.subplot(1, 2, 2)
plt.title('Correlation matrix (females)')
sns.heatmap(df_female_clean[['stazas', 'bdu_val', 'issilavinimas']].corr(),
annot=True, cmap='coolwarm')
plt.tight_layout()
plt.show()

#FIGURE 8
fig, axes = plt.subplots(1, 2, figsize=(14, 6), sharey=True)
sns.histplot(data=df_male_clean, x='bdu_val', hue='issilavinimas',
multiple='stack', kde=True, ax=axes[0], palette="Blues")
axes[0].set_title('Distribution of Average Pay for Men')
axes[0].set_xlabel('Average Hourly Pay')
axes[0].set_ylabel('Density')
sns.histplot(data=df_female_clean, x='bdu_val', hue='issilavinimas',
multiple='stack', kde=True, ax=axes[1])
axes[1].set_title('Distribution of Average Pay for Women')
axes[1].set_xlabel('Average Hourly Pay')
axes[1].set_ylabel('')

plt.suptitle('Distribution of Average Pay by Education Level')
plt.tight_layout(rect=[0, 0, 1, 0.95])
plt.show()


#REMOVING OUTLIERS 2018
df_male = df_2018[df_2018['lytis'] == 'M']
df_female = df_2018[df_2018['lytis'] == 'F']
df_male_clean = remove_outliers(df_male, 'bdu_val')
df_female_clean = remove_outliers(df_female, 'bdu_val')
```

```
# FIGURE 5
plt.figure(figsize=(10, 6))

sns.histplot(df_male_clean['bdu_val'], kde=True, color='blue', label='Men',
stat="density", alpha=0.6)
sns.histplot(df_female_clean['bdu_val'], kde=True, color='pink',
label='Women', stat="density", alpha=0.6)

plt.title('Distribution of Average Pay for Men and Women (Outliers Removed)')
plt.xlabel('Average Hourly Pay')
plt.ylabel('Density')
plt.legend()
plt.tight_layout()
plt.show()

#FIGURE 7
plt.figure(figsize=(14, 6))
plt.subplot(1, 2, 1)
plt.title('Correlation matrix (males)')
sns.heatmap(df_male_clean[['stazas', 'bdu_val', 'issilavinimas']].corr(),
annot=True, cmap='coolwarm')
plt.subplot(1, 2, 2)
plt.title('Correlation matrix (females)')
sns.heatmap(df_female_clean[['stazas', 'bdu_val', 'issilavinimas']].corr(),
annot=True, cmap='coolwarm')
plt.tight_layout()
plt.show()


#FIGURE 9
fig, axes = plt.subplots(1, 2, figsize=(14, 6), sharey=True)

sns.histplot(data=df_male_clean, x='bdu_val', hue='issilavinimas',
multiple='stack', kde=True, ax=axes[0], palette="Blues")
axes[0].set_title('Distribution of Average Pay for Men')
axes[0].set_xlabel('Average Hourly Pay')
axes[0].set_ylabel('Density')
sns.histplot(data=df_female_clean, x='bdu_val', hue='issilavinimas',
multiple='stack', kde=True, ax=axes[1])
axes[1].set_title('Distribution of Average Pay for Women')
```

```
axes[1].set_xlabel('Average Hourly Pay')
axes[1].set_ylabel('')
plt.suptitle('Distribution of Average Pay by Education Level')
plt.tight_layout(rect=[0, 0, 1, 0.95])
plt.show()


#THE MOST POPULAR PROFESSIONS
df_men = df_combined[df_combined['lytis'] == 'M']
df_women = df_combined[df_combined['lytis'] == 'F']
top_5_men_professions = df_men['profesija'].value_counts().head(5)
top_5_women_professions = df_women['profesija'].value_counts().head(5)
print("Top 5 Professions for Men:")
print(top_5_men_professions)
print("\nTop 5 Professions for Women:")
print(top_5_women_professions)
target_professions = ['833', '711', '541', '723', '132', '241', '242', '522', '222', '911']

df_2018['profesija'] = df_2018['profesija'].astype(str)
df_2014['profesija'] = df_2014['profesija'].astype(str)


df_2018_men = df_2018[(df_2018['lytis'] == 'M') &
(df_2018['profesija'].isin(target_professions))]
df_2014_men = df_2014[(df_2014['lytis'] == 'M') &
(df_2014['profesija'].isin(target_professions))]

df_2018_women = df_2018[(df_2018['lytis'] == 'F') &
(df_2018['profesija'].isin(target_professions))]
df_2014_women = df_2014[(df_2014['lytis'] == 'F') &
(df_2014['profesija'].isin(target_professions))]

#Calculate average bdu_val for men and women by profession
avg_bdu_men_2018 = df_2018_men.groupby('profesija')
['bdu_val'].mean().reindex(target_professions)
avg_bdu_women_2018 = df_2018_women.groupby('profesija')
['bdu_val'].mean().reindex(target_professions)

avg_bdu_men_2014 = df_2014_men.groupby('profesija')
['bdu_val'].mean().reindex(target_professions)
```

```
avg_bdu_women_2014 = df_2014_women.groupby('profesija')
['bdu_val'].mean().reindex(target_professions)

profession_labels = {
    '833': '833 - Drivers ',
    '711': '711 - Construction',
    '541': '541 - Security',
    '723': '723 - Mechanics',
    '132': '132 - Managers',
    '241': '241 - Administration',
    '242': '242 - Education',
    '522': '522 - Retail',
    '222': '222 - Health',
    '911': '911 - Services'
}

labelled_professions = [profession_labels[code] for code in target_professions]

fig, axes = plt.subplots(1, 2, figsize=(14, 6))
bar_width = 0.35
index = range(len(target_professions))

axes[0].bar([i - bar_width / 2 for i in index], avg_bdu_men_2018,
color='blue', alpha=0.7, label='Men 2018', width=bar_width)
axes[0].bar([i + bar_width / 2 for i in index], avg_bdu_women_2018,
color='pink', alpha=0.7, label='Women 2018', width=bar_width)
axes[0].set_title('Average bdu_val by Profession for Men and Women (2018)')
axes[0].set_xlabel('Profession')
axes[0].set_ylabel('Average bdu_val')
axes[0].tick_params(axis='x', rotation=45)
axes[0].set_xticks(index)
axes[0].set_xticklabels(labelled_professions)
axes[0].legend()

axes[1].bar([i - bar_width / 2 for i in index], avg_bdu_men_2014,
color='blue', alpha=0.7, label='Men 2014', width=bar_width)
axes[1].bar([i + bar_width / 2 for i in index], avg_bdu_women_2014,
color='pink', alpha=0.7, label='Women 2014', width=bar_width)
axes[1].set_title('Average bdu_val by Profession for Men and Women (2014)')
axes[1].set_xlabel('Profession')
```

```
axes[1].set_ylabel('Average bdu_val')
axes[1].tick_params(axis='x', rotation=45)
axes[1].set_xticks(index)
axes[1].set_xticklabels(labelled_professions)
axes[1].legend()

plt.tight_layout()
plt.show()
```

OLS model

```
import pandas as pd
import statsmodels.api as sm

df_2018 = pd.read_csv('/Users/samuleviciute/Downloads/magistras/2018.csv')
df_2014 = pd.read_csv('/Users/samuleviciute/Downloads/magistras/2014.csv')

def remove_outliers(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return df[(df[column] >= lower_bound) & (df[column] <= upper_bound)]

df_male = df_2018[df_2018['lytis'] == 'M']
df_female = df_2018[df_2018['lytis'] == 'F']
df_male_clean = remove_outliers(df_male, 'bdu_val')
df_female_clean = remove_outliers(df_female, 'bdu_val')

df_male_clean['issilavinimas'] = df_male_clean['issilavinimas'].astype('category')
df_female_clean['issilavinimas'] = df_female_clean['issilavinimas'].astype('category')
df_male_clean['im_dydzio_kodas'] = df_male_clean['im_dydzio_kodas'].astype('category')
df_female_clean['im_dydzio_kodas'] = df_female_clean['im_dydzio_kodas'].astype('category')

X_male = df_male_clean[['issilavinimas', 'stazas', 'im_dydzio_kodas']]
X_male = pd.get_dummies(X_male, columns=['issilavinimas',
'im_dydzio_kodas'], drop_first=True)
y_male = df_male_clean['bdu_val']
X_male = sm.add_constant(X_male)
```

```
model_male_2018 = sm.OLS(y_male, X_male.astype(float)).fit()
print("Male Regression Model (2018):")
print(model_male_2018.summary())

X_female = df_female_clean[['issilavinimas', 'stazas', 'im_dydzio_kodas']]
X_female = pd.get_dummies(X_female, columns=['issilavinimas',
'im_dydzio_kodas'], drop_first=True)
y_female = df_female_clean['bdu_val']
X_female = sm.add_constant(X_female)
model_female_2018 = sm.OLS(y_female, X_female.astype(float)).fit()
print("Female Regression Model (2018):")
print(model_female_2018.summary())

df_male = df_2014[df_2014['lytis'] == 'M']
df_female = df_2014[df_2014['lytis'] == 'F']
df_male_clean = remove_outliers(df_male, 'bdu_val')
df_female_clean = remove_outliers(df_female, 'bdu_val')

df_male_clean['issilavinimas'] = df_male_clean['issilavinimas'].astype('category')
df_female_clean['issilavinimas'] = df_female_clean['issilavinimas'].astype('category')
df_male_clean['im_dydzio_kodas'] = df_male_clean['im_dydzio_kodas'].astype('category')
df_female_clean['im_dydzio_kodas'] = df_female_clean['im_dydzio_kodas'].astype('category')

X_male = df_male_clean[['issilavinimas', 'stazas', 'im_dydzio_kodas']]
X_male = pd.get_dummies(X_male, columns=['issilavinimas',
'im_dydzio_kodas'], drop_first=True)
y_male = df_male_clean['bdu_val']
X_male = sm.add_constant(X_male)
model_male_2014 = sm.OLS(y_male, X_male.astype(float)).fit()
print("Male Regression Model (2014):")
print(model_male_2014.summary())

X_female = df_female_clean[['issilavinimas', 'stazas', 'im_dydzio_kodas']]
X_female = pd.get_dummies(X_female, columns=['issilavinimas',
'im_dydzio_kodas'], drop_first=True)
y_female = df_female_clean['bdu_val']
X_female = sm.add_constant(X_female)
model_female_2014 = sm.OLS(y_female, X_female.astype(float)).fit()
print("Female Regression Model (2014):")
print(model_female_2014.summary())
```

SEM model

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from semopy import Model
import numpy as np
from numpy.linalg import inv
import graphviz


def remove_outliers(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return df[(df[column] >= lower_bound) & (df[column] <= upper_bound)]

df_2018 = pd.read_csv('/Users/samuleviciute/Downloads/magistras/2018.csv')
df_2014 = pd.read_csv('/Users/samuleviciute/Downloads/magistras/2014.csv')

selected_columns = ['issilavinimas', 'im_dydzio_kodas', 'bdu_val', 'lytis', 'stazas']
df_2018 = df_2018[selected_columns]
df_2014 = df_2014[selected_columns]

def prepare_data(df):
    df = remove_outliers(df, 'bdu_val')
    scaler = StandardScaler()
    df['bdu_val_norm'] = scaler.fit_transform(df[['bdu_val']])

    df['im_dydzio_kodas'] = df['im_dydzio_kodas'].astype('category')
    df = pd.get_dummies(df, columns=['im_dydzio_kodas'], drop_first=True)
    df['im_dydzio_kodas_1'] = df['im_dydzio_kodas_1'].astype(int)
    df['im_dydzio_kodas_2'] = df['im_dydzio_kodas_2'].astype(int)

    return df.reset_index(drop=True)


df_male_2018 = prepare_data(df_2018[df_2018['lytis'] == 'M'])
df_male_2018 = df_male_2018.drop('lytis', axis = 1)
df_female_2018 = prepare_data(df_2018[df_2018['lytis'] == 'F'])
```

```
df_female_2018 = df_female_2018.drop('lytis', axis = 1)
df_male_2014 = prepare_data(df_2014[df_2014['lytis'] == 'M'])
df_male_2014 = df_male_2014.drop('lytis', axis = 1)
df_female_2014 = prepare_data(df_2014[df_2014['lytis'] == 'F'])
df_female_2014 = df_female_2014.drop('lytis', axis = 1)

# Define SEM model structure
sem_model_desc = """
bdu_val_norm ~ issilavinimas + stazas + im_dydzio_kodas_1 + im_dydzio_kodas_2
"""

def run_sem(df, desc, label):
    # Ensure no missing values
    df = df.dropna()
    # Fit the SEM model
    model = Model(desc)
    model.fit(df)

    try:
        # Calculate the model-implied covariance matrix
        model_cov = model.calc_sigma()
        print("Model-implied covariance matrix:")
        #print(model_cov)
    except Exception as e:
        print("Error accessing model-implied covariance matrix:", e)

    sample_cov = np.cov(df, rowvar=False)
    model_cov_matrix = model_cov[0]
    sample_size = df.shape[0]
    variable_names = list(model.vars.get("all", []))
    aligned_data = df[variable_names]
    sample_cov = np.cov(aligned_data, rowvar=False)
    try:
        chi_squared = calculate_chi_squared(sample_cov, model_cov_matrix, sample_size)
        print("Chi-squared statistic:", chi_squared)
        print("Chi-squared statistic / Number of observations:", chi_squared/sample_size)
        num_observed_variables = len(variable_names)
        num_observed_covariances = (num_observed_variables*(num_observed_variables+1))//2
        num_parameters = num_observed_variables
        degrees_of_freedom = num_observed_covariances - num_parameters
```

```
    # RMSEA
        if degrees_of_freedom > 0:
            rmsea = np.sqrt(max(chi_squared - degrees_of_freedom, 0)/
            (degrees_of_freedom * (sample_size - 1)))
            print(f"RMSEA ({label}):", rmsea)
        else:
            print(f"RMSEA cannot be calculated for {label} due to zero degrees of freedom.")
    except Exception as e:
        print("Error calculating Chi-squared statistic or RMSEA:", e)
    print('making graph')
    inspection = model.inspect()
    df_inspect = pd.DataFrame(inspection)
    regressions = df_inspect[df_inspect["op"] == "~"]
    covariances = df_inspect[df_inspect["op"] == "~~"]
    dot = graphviz.Digraph(format="png")
    nodes = set(df_inspect["lval"]).union(set(df_inspect["rval"]))
    for node in nodes:
        dot.node(node, node)
    for _, row in regressions.iterrows():
        dot.edge(row["rval"], row["lval"], label=f"{row['Estimate']:.2f}")
    for _, row in covariances.iterrows():
        dot.edge(row["lval"], row["rval"], label=f"Cov: {row['Estimate']:.2f}", dir="none")

    output_filename = f"sem_model_{label.replace(' ', '_')}"
    dot.render(output_filename, format="png", cleanup=True)
    print(f"Graph saved as {output_filename}.png")
    dot.view()

    print(f"SEM Results for {label}:")
    print(model.inspect())

def calculate_chi_squared(sample_cov, model_cov_matrix, sample_size):
    diff = sample_cov - model_cov_matrix
    chi_squared = sample_size * np.trace(inv(model_cov_matrix) @ diff)
    return chi_squared

#SEM for each group and year
run_sem(df_male_2018, sem_model_desc, "2018 Male Data")
run_sem(df_female_2018, sem_model_desc, "2018 Female Data")
run_sem(df_male_2014, sem_model_desc, "2014 Male Data")
```

```
run_sem(df_female_2014, sem_model_desc, "2014 Female Data")
```

HML model

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import mixedlm
from sklearn.preprocessing import StandardScaler

def remove_outliers(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return df[(df[column] >= lower_bound) & (df[column] <= upper_bound)]

df_2018 = pd.read_csv('/Users/samuleviciute/Downloads/magistras/2018.csv')
df_2014 = pd.read_csv('/Users/samuleviciute/Downloads/magistras/2014.csv')

target_professions = ['833', '711', '541', '723', '132', '241', '242', '522', '222', '911']
selected_columns = ['issilavinimas', 'im_dydzio_kodas', 'profesija',
'bdu_val', 'lytis', 'stazas']
df_2018 = df_2018[selected_columns]
df_2014 = df_2014[selected_columns]

def prepare_data(df):
    df = remove_outliers(df, 'bdu_val')
    scaler = StandardScaler()
    df['bdu_val_norm'] = scaler.fit_transform(df[['bdu_val']])

    df['issilavinimas'] = df['issilavinimas'].astype('category')
    df['im_dydzio_kodas'] = df['im_dydzio_kodas'].astype('category')
    df = pd.get_dummies(df, columns=['issilavinimas', 'im_dydzio_kodas'], drop_first=True)
    df['issilavinimas_2'] = df['issilavinimas_2'].astype(int)
    df['issilavinimas_3'] = df['issilavinimas_3'].astype(int)
    df['issilavinimas_4'] = df['issilavinimas_4'].astype(int)
    df['im_dydzio_kodas_1'] = df['im_dydzio_kodas_1'].astype(int)
    df['im_dydzio_kodas_2'] = df['im_dydzio_kodas_2'].astype(int)
```

```
    return df.reset_index(drop=True)

df_2018['profesija'] = df_2018['profesija'].astype(str)
df_2014['profesija'] = df_2014['profesija'].astype(str)

df_male_2018 = prepare_data(df_2018[(df_2018['lytis'] == 'M') &
(df_2018['profesija'].isin(target_professions))])
df_male_2018 = df_male_2018.drop('lytis', axis = 1)
df_female_2018 = prepare_data(df_2018[(df_2018['lytis'] == 'F') &
(df_2018['profesija'].isin(target_professions))])
df_female_2018 = df_female_2018.drop('lytis', axis = 1)
df_male_2014 = prepare_data(df_2014[(df_2014['lytis'] == 'M') &
(df_2014['profesija'].isin(target_professions))])
df_male_2014 = df_male_2014.drop('lytis', axis = 1)
df_female_2014 = prepare_data(df_2014[(df_2014['lytis'] == 'F') &
(df_2014['profesija'].isin(target_professions))])
df_female_2014 = df_female_2014.drop('lytis', axis = 1)

# HML model formula
formula = "bdu_val_norm ~ issilavinimas_2 + issilavinimas_3 +
issilavinimas_4 + stazas + im_dydzio_kodas_1 + im_dydzio_kodas_2"


def fit_multilevel_model(df, formula, label):
    # fit a model with random intercept for 'profesija'
    model = mixedlm(formula, df, groups=df['profesija'])
    result = model.fit()
    print(f"Multilevel Model Results for {label}:")
    print(result.summary())
    return result

# Run the model for males and females in 2018 and 2014
fit_multilevel_model(df_male_2018, formula, "2018 Male Data")
fit_multilevel_model(df_female_2018, formula, "2018 Female Data")
fit_multilevel_model(df_male_2014, formula, "2014 Male Data")
fit_multilevel_model(df_female_2014, formula, "2014 Female Data")
```

HML model without categorical education variable

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import mixedlm
from sklearn.preprocessing import StandardScaler

def remove_outliers(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return df[(df[column] >= lower_bound) & (df[column] <= upper_bound)]

df_2018 = pd.read_csv('/Users/samuleviciute/Downloads/magistras/2018.csv')
df_2014 = pd.read_csv('/Users/samuleviciute/Downloads/magistras/2014.csv')

target_professions = ['833', '711', '541', '723', '132', '241', '242', '522', '222', '911']
selected_columns = ['issilavinimas', 'im_dydzio_kodas', 'profesija',
'bdu_val', 'lytis', 'stazas']
df_2018 = df_2018[selected_columns]
df_2014 = df_2014[selected_columns]

def prepare_data(df):
    df = remove_outliers(df, 'bdu_val')
    scaler = StandardScaler()
    df['bdu_val_norm'] = scaler.fit_transform(df[['bdu_val']])

    df['im_dydzio_kodas'] = df['im_dydzio_kodas'].astype('category')
    df = pd.get_dummies(df, columns=['im_dydzio_kodas'], drop_first=True)
    df['im_dydzio_kodas_1'] = df['im_dydzio_kodas_1'].astype(int)
    df['im_dydzio_kodas_2'] = df['im_dydzio_kodas_2'].astype(int)

    return df.reset_index(drop=True)

df_2018['profesija'] = df_2018['profesija'].astype(str)
df_2014['profesija'] = df_2014['profesija'].astype(str)

df_male_2018 = prepare_data(df_2018[(df_2018['lytis'] == 'M') &
(df_2018['profesija'].isin(target_professions))])
```

```
df_male_2018 = df_male_2018.drop('lytis', axis = 1)
df_female_2018 = prepare_data(df_2018[(df_2018['lytis'] == 'F') &
(df_2018['profesija'].isin(target_professions))])
df_female_2018 = df_female_2018.drop('lytis', axis = 1)
df_male_2014 = prepare_data(df_2014[(df_2014['lytis'] == 'M') &
(df_2014['profesija'].isin(target_professions))])
df_male_2014 = df_male_2014.drop('lytis', axis = 1)
df_female_2014 = prepare_data(df_2014[(df_2014['lytis'] == 'F') &
(df_2014['profesija'].isin(target_professions))])
df_female_2014 = df_female_2014.drop('lytis', axis = 1)


formula = "bdu_val_norm ~ issilavinimas + stazas + im_dydzio_kodas_1 + im_dydzio_kodas_2"

def fit_multilevel_model(df, formula, label):
    #random intercept for 'profesija'
    model = mixedlm(formula, df, groups=df['profesija'])
    result = model.fit()
    print(f"Multilevel Model Results for {label}:")
    print(result.summary())
    return result

fit_multilevel_model(df_male_2018, formula, "2018 Male Data")
fit_multilevel_model(df_female_2018, formula, "2018 Female Data")
fit_multilevel_model(df_male_2014, formula, "2014 Male Data")
fit_multilevel_model(df_female_2014, formula, "2014 Female Data")
```