Project: Public sector wage differentials

Team members: Kertu Jõgi, Anette Kuusk, Karl Ruul

Repository: https://github.com/karlru/rahaahned-ametnikud

Business understanding

1. **Identifying your business goals**
    - Background

Public sector is important part of the economy, because it helps in providing essential service to the citizens. Roughly one-third of people work in public sector, therefore public sector wages have a huge impact on pay settings in other fields. Public sector offers more compensations in general, is relatively larger employer of women and offers fairer wages compared to men. However, this can be a problem, if private sector employers can not keep up with public sector compensations, then people will rather work in public sector and this leads to skill shortage in the private sector. Our goal is to visualize how much the salaries differ, so that it would be a great guide for people who are looking for new jobs. Heads of institutions could also benefit from this research, because they have comparison and therefore know which direction to move towards. It is important for them to know other institutions salaries so they could keep up and compete in the job market.

    - Business goals

Goal 1: the goal of this project is to get an overview of wage differences in public sector based on gender and institution.

Goal 2: the goal of this project is to get an overview of how much the salaries differ and have changed compared to the general averages.

Goal 3: train a gender classification model based on first names.

    - Business success criteria

The outcome should be clear visuals about the data, where all differences are visible. The project has been successful, if we have graphs about gender and institution differences in public sector. In addition, we want to get graphs about salary differences in public and private

sectors over the years. Gender classification model is successful, if it classifies 80% of the names correctly.

2. **Assessing your situation**
    - Inventory of resources

Human power: Kertu, Anette and Karl, Pille, Jupyter notebook, Statistics Estonia data, personal laptops, supervisors for questions, lecture materials.

    - Requirements, assumptions, and constraints

The whole project has to be finished by 12th December. We have to make a poster by then and finish our analysis part in Jupyter notebook.

    - Risks and contingencies

One of our team members not being able to do the required work, solution would be that others cover for him/her. Not knowing how to solve our problems, we can get help from course supervisors. A risk of running late: we have to start early and plan our process.

    - Terminology

Private sector – the part of a country's economic system that is run by individuals and companies, rather than a government entity. The main purpose is to make profit.

Public sector – a part of the economy that comprises all organizations that are owned and operated by the government. The main purpose of the public sector is to provide services that are considered essential for the well-being of society.

Data visualisation – a way to represent information graphically, highlighting patterns and trends in data and helping the reader to achieve quick insights

Differential (in this context) – a wage differentials refers to the difference in wages between people with similar skills within differing localities or industries

    - 0 costs and 0 benefits

3. **Defining your data-mining goals**
    - We are planning to visualise the data and not to predict or build models. If possible and we have enough time, we will try to do a name classifier based on gender.

Data understanding

Our two main data mining goals are to have an overview of wage differences in the public sector based on gender and institution, and to see how much wages differ and have changed compared to the general averages. To achieve that we need a list of data which summarizes Estonian public sector salaries based on different years. In these datasets we are relying on numerical data (annual income), also text data to determine institutions and positions.

All of our data sets are for public usage provided by Statistics Estonia. Therefore, it is guaranteed that our chosen data exists. In addition, we have already made some tests to be sure if the files are usable. We are using seven different xlsx and csv files from Statistics Estonia: “ametnike_palgad_2018.xlsx”, “ametnike_palgad_2019.xlsx”, “ametnike_palgad_2020.xlsx”, “ametnike_palgad_2021.xlsx”, “keskmine_brutopalk_sektor.csv”, “keskmine_brutopalk.csv”, “sooline_palgalohe.csv”. In the first file (“ametnike_palgad_2018.xlsx”) we will mainly focus on the sheet called “Riik_põhipalk 01.04.2018”, where field “Põhipalk” provides us with 2018 public sector incomes. To understand how wages differ within institutions and positions the fields called ‘Asutus’ and ‘Ametikoht’ will also be used. To achieve our third goal (training gender classification model) we will attempt to put in use the ‘Eesnimi’ field. With all the other xlsx files (“ametnike_palgad_2019.xlsx”, “ametnike_palgad_2020.xlsx”, “ametnike_palgad_2021.xlsx”), the same logic is used. Meaning, we mainly focus on the sheet “Riik_põhipalk” fields “Põhipalk”, “Ametikoht”, “Asutus” and “Eesnimi”. File named “keskmine_brutopalk_sektor.csv” provides us with knowledge about average gross salaries in the public sector and in the private sector. Meaning field “Eesti eraõiguslik isik” shows average gross salaries in Estonian public sector, field “Kohalik omavalitsus” average gross salaries in Estonian local municipal government and field “Riik” shows average gross salaries in Estonian private sector.

In “keskmine_brutopalk.csv ” we have info about average gross salaries in different institutions and fields of activity (e.g. agriculture, education, finance and insurance etc). Similar to the last csv file, there are also fields which show how average gross salary has changed compared to the same period last year. In these files we have data from years between 2018-2021, which is in correlation with our chosen xlsx files. In our last dataset “sooline_palgalohe.csv” we have info about gender pay gap. In this file we have data

between years 2018-2020 for average gross salary of male employees (“Meestöötajate keskmine brutopalk”) and average gross salary of female employees (“Naistöötajate keskmine brutopalk”), based on different fields of activity (social work, construction, finance and insurance, education etc) and also in general.

As our datasets have many fields which will not be used in our project, data cleaning is necessary. After getting familiar with our data, we can state hypotheses. Firstly, there are few fields of areas where the percentage of gender pay gap is negative (meaning female’s pay is higher than male’s). Secondly, the average gross salary has been raising between the period 2018-2021.

Planning your project

List of tasks:

1. Setting up - 1h each
    a. Making a repository
    b. Getting data
2. Business understanding - 5h Anette
    a. Identifying the business goals
    b. Assessing the situation
    c. Defining data-mining goals
    d. Writing a report
3. Data understanding - 5h Kertu
    a. Reading in the data
    b. Gathering, describing and exploring data
    c. Verifying data quality
    d. Writing a report
4. Planning the project - 2h Karl
    a. Sharing the responsibilities
    b. Estimating time consumption
    c. Constructing the plan
5. Implementing name based gender classification - 3h Karl
    a. Either training a model or making script to extract the data online
6. Working with the data - 8h each
    a. Structuring and cleaning the data
    b. Adding gender attribute to public sector workers
    c. Possibly deriving or generating more attributes
    d. Integrating and unifying datasets
    e. Saving the ready-to-use dataset for modeling
7. Modeling - 8h each
    a. Comparing wages and gender pay gaps between public institutions
    b. Comparing the averages and trends between public and private sectors
    c. Comparing the wage gaps between public and private sectors
8. Visualizing results - 3h each
    a. Visualizing results for each comparisons

9. Wrapping up - 5h each
    a. Describing what we found
    b. Making a poster

Some parts of the project are done separately (like business- and data understanding and planning), whereas the others will involve each team member and the task will be split up into equal parts for the members to work on.