Case study

Analyzing the Data Science Salaries

Submitted by:
Rufina Gafiiatullina, Arina Goncharova, B22-DS-02

11.03.2024

Contents

Introduction

Problem context

The data science industry is rapidly developing, leading to a high demand for skilled professionals. Therefore, understanding factors that influence their salaries is crucial for individuals making career choices and for organizations offering employment.

Scientific question

We want to analyse what factors influence the salaries of IT specialists, and what is the impact of different factors on salaries compared between machine learning (ML) specialists, data scientists (DS), and data analysts (DA). We focus on the following questions:

- Does experience level impact the salary?
- Does remoting of worker impact the salary?
- Does size of company impact the salary?
- Does employment type impact the salary?
- Does country impact the salary?
- Does job (DS, DA, ML) impact the salary?
- Does experience level impact the salary of job (DS, DA, ML)?

Motivation

Investigating the factors affecting IT salaries is crucial to understanding the dynamic employment market in information technology sphere. This study becomes important for several reasons.

First of all, the data science industry is full of competitiveness, and understanding the factors affecting salaries allows both companies to effectively adapt to the requirements of the market. Understanding the relevant factors gives a competitive advantage in attracting and retaining highly qualified employees. At the same time, this knowledge is a valuable tool for professionals to help them choose a job with optimal conditions and salary expectations.

Secondly, salary difference between machine learning specialists, data scientists and data analysts, as well as income variation within each speciality depending on experience level can be significant. Understanding these differences can help companies create attractive offers and retain workers.

Finally, the results of this study can be useful for both employees and companies in the process of salary negotiations. It allows the specialists to adequately assess their value and successfully negotiate with an employer while companies will be able to effectively distribute the salary budget based on the results of this case study.

Data

To address the scientific question, the dataset "Data Science Salaries" was chosen from kaggle. It contains 607 salary records for various IT professionals with the following columns:

Column name	**Value**
work year	The year the salary was paid: 2020 2021 2022
experience level	EN Entry-level Junior MI Mid-level Intermediate SE Senior-level Expert EX Executive-level / Director
employment type	Part-time (PT) Full-time (FT) Contract (CT) Freelance (FL)
job title	The role worked in during the year
salary	The total gross salary amount paid
salary currency	The currency of the salary paid as an ISO 4217 currency code.
salary in USD	The salary in USD
employee residence	Employee's primary country of residence during the work year as an ISO 3166 country code
remote ratio	The overall amount of work done remotely: 0 - No remote work (0, less than 20%) 50 - Partially remote 100 - Fully remote (more than 80%)
company location	The country of the employer's main office or contracting branch as an ISO 3166 country code
company size	S - less than 50 employees M - from 50 to 250 employees L - more than 250 employees

From dataset it is possible to distinguish 3 most populated (represented) groups: machine learning specialists, data scientists and data analysts. Therefore, it was decided to first test the

hypotheses without dividing by specialization, and then to compare the differences in salary based on specific factors between these three types of specialists.

Link to dataset: `https://www.kaggle.com/datasets/zain280/data-science-salaries`

Study protocol

1. **Data loading and preparation** (check the dataset for null values, delete columns salary and salary currency because of redundancy)
2. **Comparing salary in 2020, 2021 and 2022 years**
   - compare mean salary in 2020, 2021, 2022 using bar plots visualization
   - Plot Q-Q plot to check whether the sample was drawn from the population with normal distribution.
   - Remove outliers if needed.
   - Plot Q-Q plot for the data without outliers.
   - Conduct Shapiro–Wilk test to check whether the sample without outliers was drawn from the population with normal distribution.
3. **Preliminary steps for testing the following hypotheses:**
   - Ensure that the samples satisfy the assumptions of the Kruskal-Wallis test (check of the assumption about same shape distribution can be found in the code, the others are explained in the Statistical techniques section).
   - Perform the Kruskal-Wallis test.
   - If the null hypothesis is rejected, conduct the Dunn's test.
4. **Analysis without splitting into specializations:**
   - check whether there is no significant difference in mean salary of junior, middle, senior, executive specialists
   - check whether there is no significant difference in salary depending on remote ratio
   - check whether there is no significant difference in salary in small, medium and large companies
   - check whether there is no significant difference in salary between different employment types
   - check whether there is no significant difference in salary between different workers from different countries

5. **Analysis with splitting into specializations:**
    - splitting data into 3 groups based on the job type: ml specialists, data scientists and data analysts
    - check whether there is no significant in salary of ml specialists, data scientists and data analysts
    - check whether there is no significant in data scientists salary of different experience level
    - check whether there is no significant in machine learning specialists salary of different experience level
    - check whether there is no significant difference in the data analysts salary of different experience level
    - check whether there is no significant difference in the salary of middle specialists between different job titles
    - check whether there is no significant difference in the salary of senior specialists between different job titles
    - check whether there is no significant difference in the salary of junior specialists between different job titles

Theory, Statistical techniques

Statistical Techniques:

- Descriptive Statistics: Measures like mean, median, mode and visualizations (histograms) were used to explore the salary within each category (experience level, job title, country, employment type, year).

- Non-Parametric Tests:
  - **Kruskal-Wallis test** was used because it assumes no particular distribution of the data and allows to compare several groups.
    Hypotheses:
    * H_0: population medians are equal.
    * H_1: population medians are not equal.

    Checking the assumptions for Kruskal-Wallis test:
    * Salary is independent variable with 3 or 4 independent groups.
    * In the chosen dataset salary is considered an interval scale variable.
    * The observations are independent.
    * Groups in each test have the same shape distributions (checked using histograms in code part). We decided to choose non-parametric test instead of parametric because the used data was from unknown distribution.
  - **The Shapiro-Wilk test** was used to check the null hypothesis that the data was drawn from a normal distribution.
    Hypotheses:
    * H_0: data was drawn from a normally distributed population.
    * H_1: data was drawn from a not normally distributed population.

- Post-Hoc Tests: Dunn test is used if Kruskal-Wallis test reject the null hypothesis and comparisons between groups are required to determine which groups are different.
  - H_0: there is no difference between groups .
  - H_1: there is a difference between groups.

Statistical tools

Our code is written in Python, including the following libraries:

- Pandas - to import, clean, and organize our dataset;
- Numpy - to perform basic numerical operations and statistical analyses;
- Matplotlib - to create visualizations of data using bars and pies;
- Scipy - to perform hypothesis testing using statistical analyses, namely the Kruskal-Wallis test;
- Scikit_posthocs - to use post hoc tests after conducting the Kruskal-Wallis test;

Results

General analysis

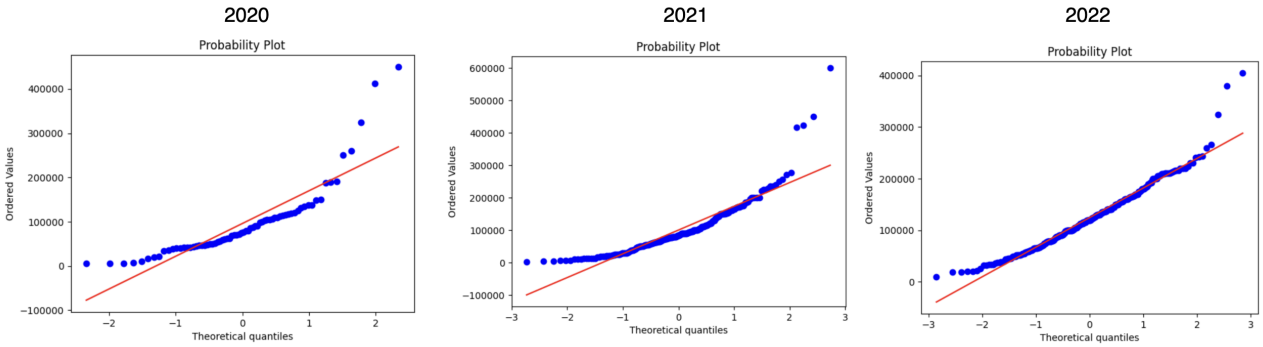
1. Mean analysis for each year

We wanted to apply t-test to analyze the mean values of salary depending on years. Therefore, our samples were tested for normal distribution. The check included analyzing the Q-Q plot, removing outliers, and using the Shapiro-Wilk test.
Hypotheses for Shapiro-Wilk test:

H_0: the sample comes from normal distribution;

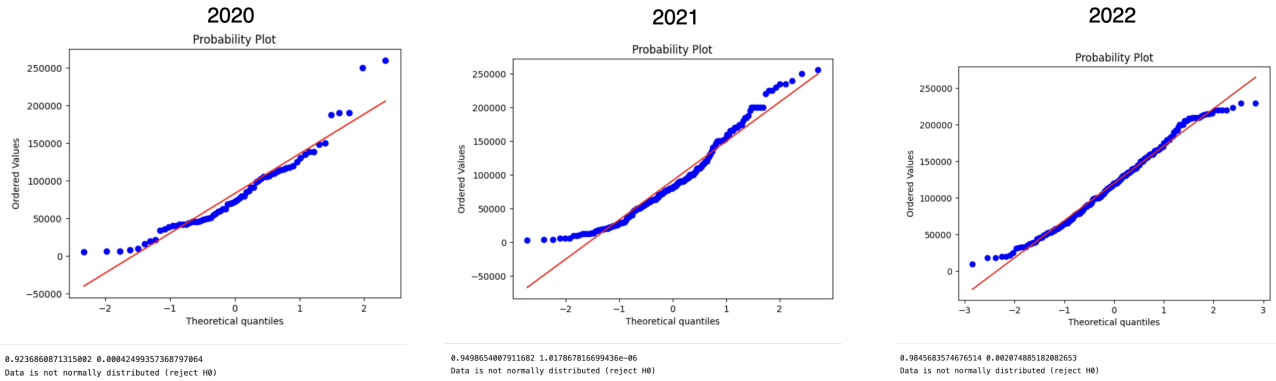
H_1: the sample is not normally distributed.

Q-Q plots before removing outliers

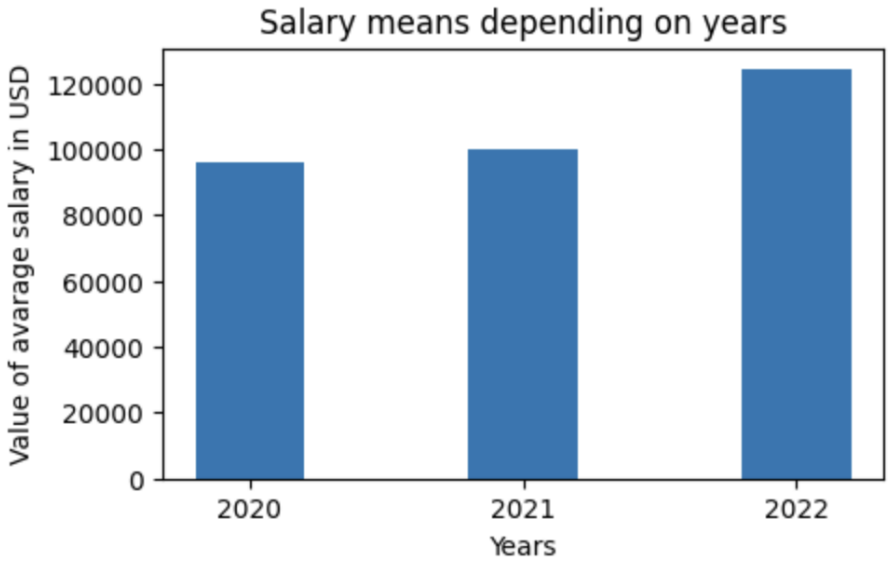


However, after removing some values, the samples were still non-normally distributed.

Q-Q plots and results of p-values after removing outliers



Therefore, mean salaries in 2020, 2021 and 2022 comparison were conducted based on the bar plot because samples from the dataset are not normally distributed (Q-Q plots and results of Shapiro–Wilk test are demonstrated in the code).



Based on the bar plot average salary in 2022 increased in comparison with 2020 and 2021.

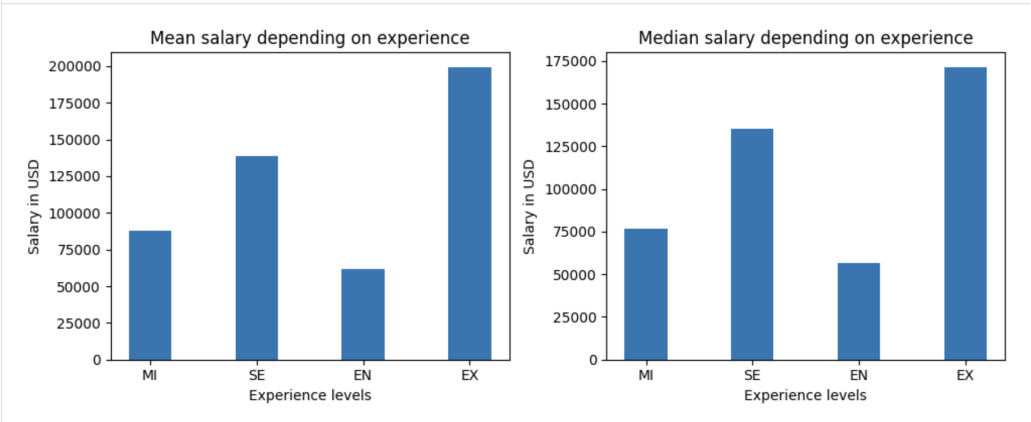
2. The difference in salary depending on experience level

Hypotheses:

H_0 : there is no significant difference between salary median of junior, middle, senior, executive specialists taking into account all jobs;

H_1 : there is significant difference.

Kruskal-Wallis test: p-value = 5.777441885479927e-41 (reject H_0)



According to the results of the Dunn's test and using the 0.05 significance level, we can conclude that all groups can be pairwise different. Bar plot also demonstrates difference in salary for all job types.

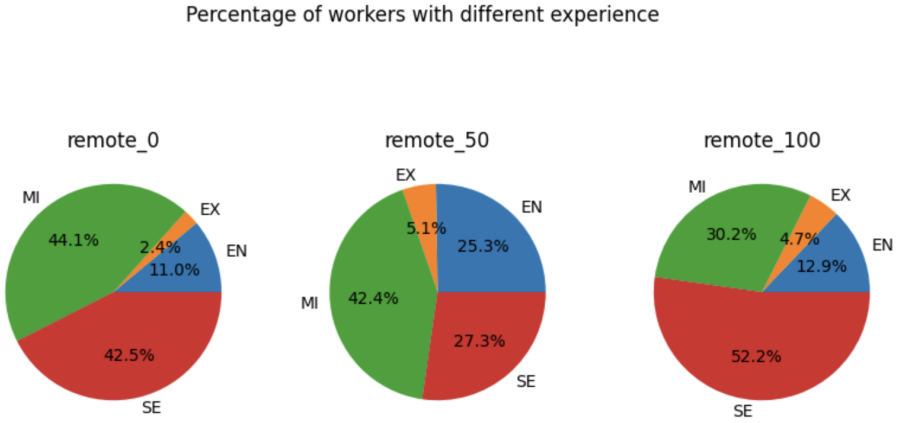
3. The difference in salary depending on remote ratio

Hypotheses:

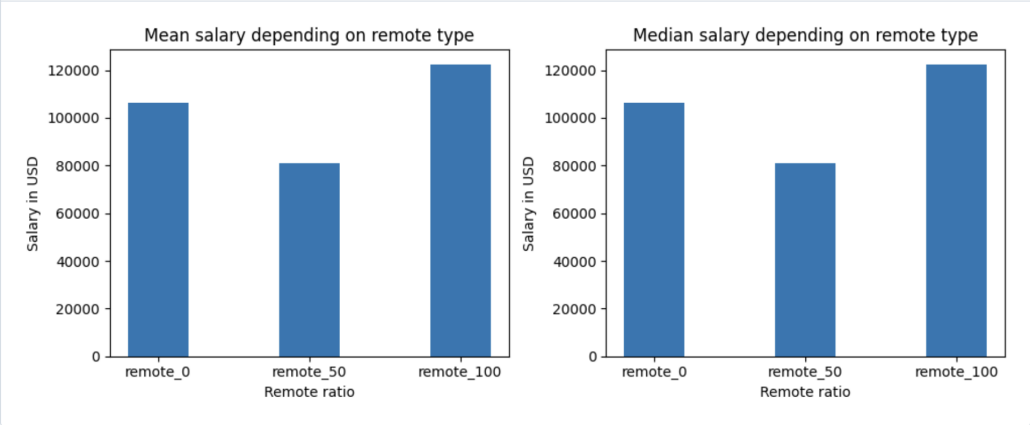
H_0 : there is no significant difference between salary median of workers who have different remote ratio (0, 50, 100) taking into account all jobs;

H_1 : there is significant difference.

Kruskal-Wallis test: p-value = 1.950760665277205e-09 (reject H_0)



We can see the increase of salary in remote ratio equal 100.



The bar plots for mean and median of salary for each type of remote ratio confirm the difference between all types of remote ratios, which was noticed by the performed tests.

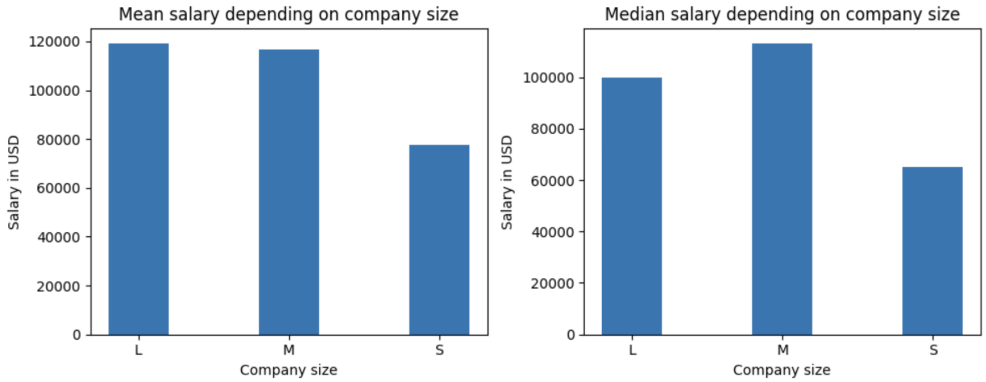
4. The difference in salary between companies with different size

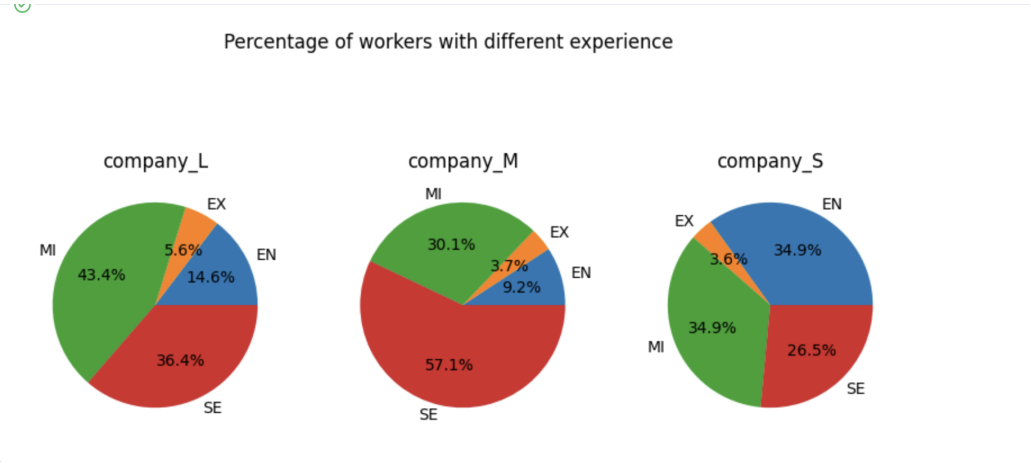
Hypotheses:

H_0 : there is no significant difference between mean salary in small, medium and large companies;

H_1 : there is significant difference.

Kruskal-Wallis test: p-value = 1.240200417426716e-08 (reject H_0)





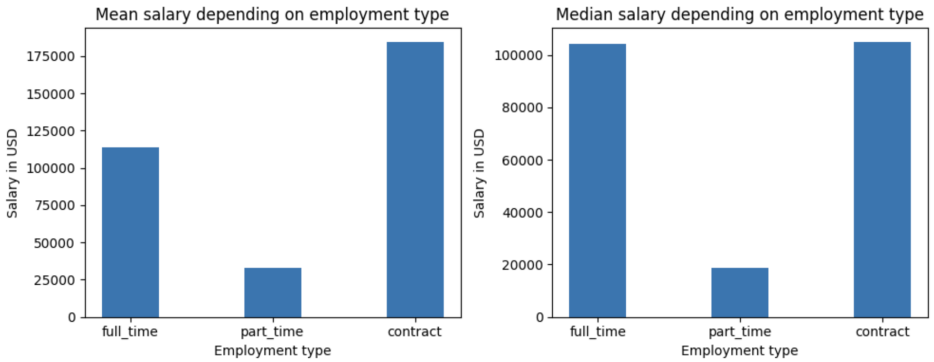
In smaller companies, salaries tend to be smaller because of limited resources. Moreover, in small companies there are much more entry-level workers, so it can be the reason why the salaries at such companies may be lower. In medium-sized companies, salaries may be higher because they usually have more opportunities for growth. In large companies, where resources and needs are greater, salaries may be larger in comparison with the previous two groups.

5. The difference in salary depending on employment type

Hypotheses:

H_0 : there is no significant difference between salary median of different types of employment (full time, part time, contract) taking into account all jobs;

H_1 : there is significant difference. Kruskal-Wallis test: p-value = 0.00012199645366253172 (reject H_0)



According to the results of the test and using the 0.05 significance level, we can conclude that salary of part time workers is significantly different from full time workers and contract workers. Part-time employment is often associated with lower salaries due to fewer hours worked. Full-time employees can expect more significant wages. Contract work can also offer high salaries because it is often associated with a huge project work and specialized skills.

6. The difference in salary depending on country

For this analysis we decided to choose the 5 countries with the highest number of employees. Then we performed tests for following countries:

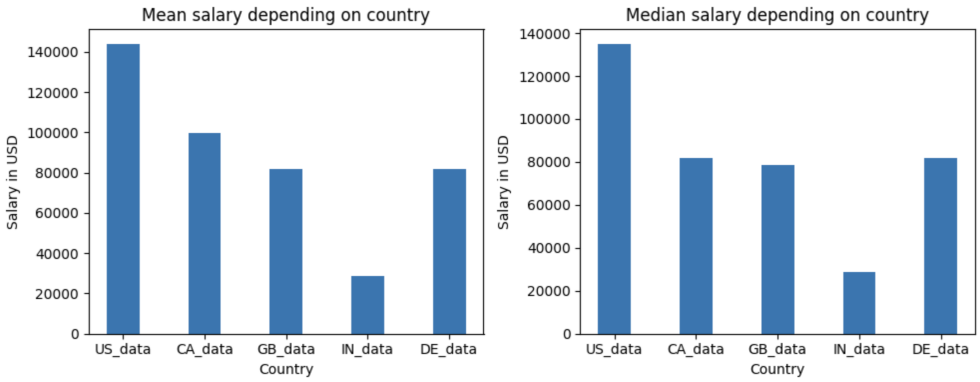
- US: the United States of America (number of workers: 355);
- GB: United Kingdom of Great Britain and Northern Ireland (number of workers: 47);
- CA: Canada (number of workers: 30);
- DE: Germany (number of workers: 28);
- IN: India (number of workers: 24).

Hypotheses:

H_0 : there is no significant difference between salary median for different countries (US, CA, GB, IN, DE) taking into account all jobs;

H_1 : there is significant difference.

Kruskal-Wallis test: p-value = 1.9574145518699078e-29 (reject H_0)



According to the results of the test and using the 0.05 significance level, we can conclude that salary of following countries (US, CA, GB, IN, DE) is significantly different: US and CA, US and GB, US and IN, US and DE, CA and IN, GB and IN, IN and DE.
This is most likely due to differences in living conditions, economic development, and the cost of labor

Analysis with splitting into specializations:

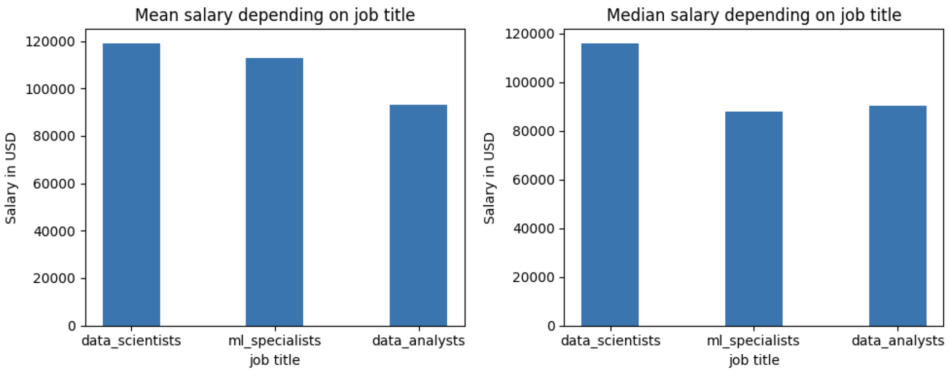
1. Salary depending on particular job

Hypotheses:

H_0 : there is no significant difference between salary median of data scientists, machine learning specialists, and data analysts;

H_1 : there is significant difference.

Kruskal-Wallis test: p-value = 0.005045255010275862 (reject H_0)



Using p-values, we can observe that there is a significant difference in median wages between data scientists and data analysts.

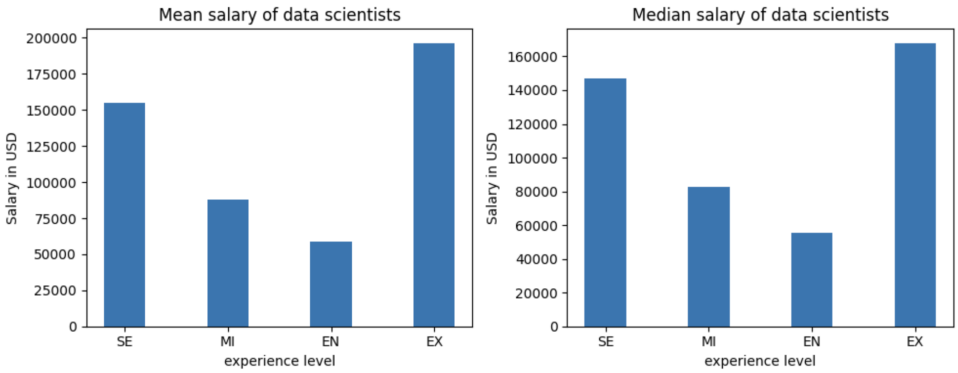
2. Salary of data scientists specialists depending on experience level

Hypotheses:

H_0 : there is no significant difference between salary median of data scientists with different experience levels;

H_1 : there is significant difference.

Kruskal-Wallis test: p-value = 1.503133336016865e-16 (reject H_0)



According to the results of the test and using the 0.05 significance level, we can conclude that salary of SE and MI, SE and EN, MI and EX, EN and EX data scientists are significantly different.

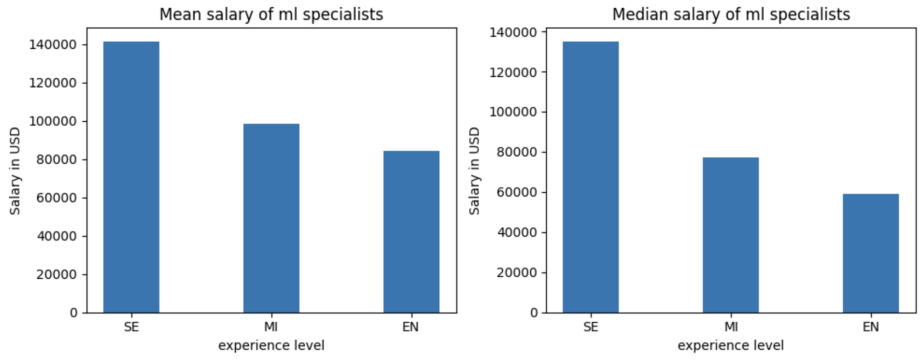
3. Salary of machine learning specialists depending on experience level

Hypotheses:

H_0 : there is no significant difference between salary median of machine learning specialists with different experience levels;

H_1 : there is significant difference.

Kruskal-Wallis test: p-value = 0.008036069152809565 (reject H_0)



According to the results of the test and using the 0.05 significance level, we can conclude that salary of SE is significantly different from MI and EN machine learning specialists.

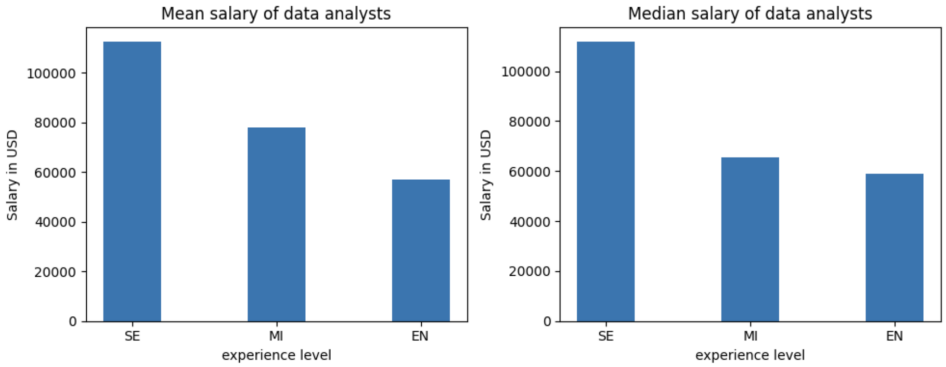
4. Salary of data analysts depending on experience level

Hypotheses:

H_0 : there is no significant difference between salary median of data analysts with different experience levels;

H_1 : there is significant difference.

Kruskal-Wallis test: p-value = 9.359981989504296e-09 (reject H_0)



According to the results of the test and using the 0.05 significance level, we can conclude that salary of SE is significantly different from MI and EN data analysts.

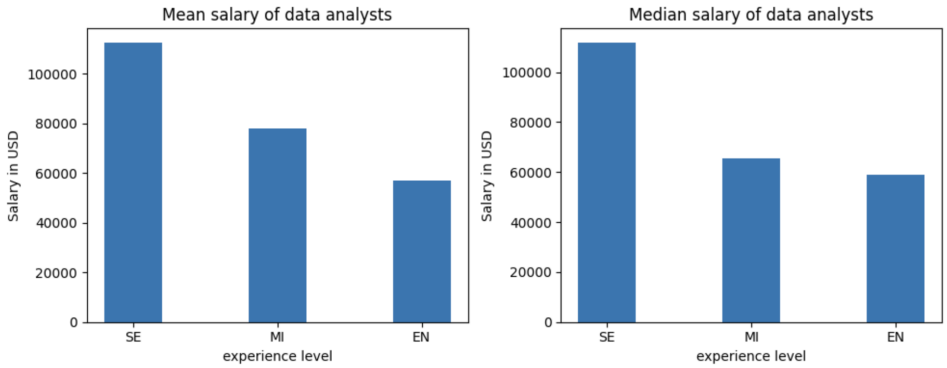
5. Salary of middle specialists depending on job title

Hypotheses:

H_0 : there is no significant difference between middle salary median of data scientists, machine learning specialists, and data analysts;

H_1 : there is significant difference.

Kruskal-Wallis test: p-value = 0.26356208409118975 (fail to reject H_0)



Test failed to reject the null hypothesis.

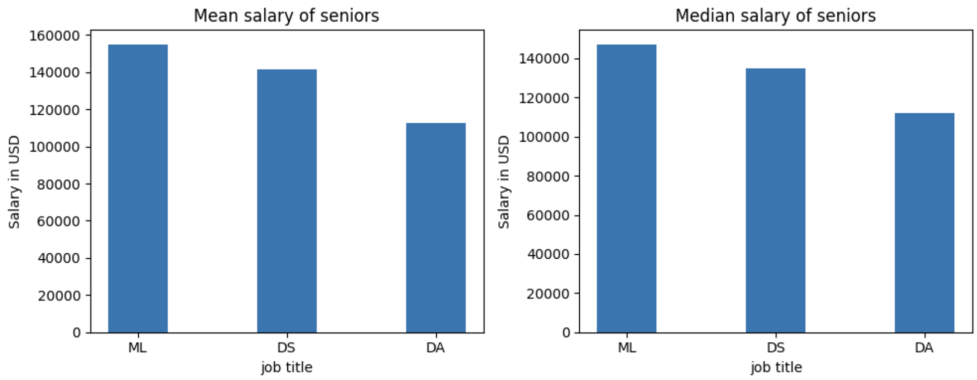
6. Salary of seniors depending on job title

Hypotheses:

H_0 : there is no significant difference between senior salary median of data scientists, machine learning specialists, and data analysts;

H_1 : there is significant difference.

Kruskal-Wallis test: p-value = 1.565313179046367e-05 (reject H_0)



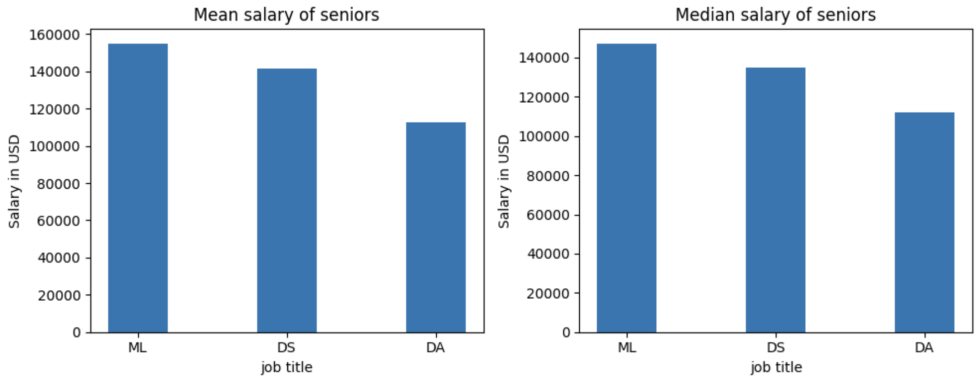
According to the results of the test and using the 0.05 significance level, we can conclude that salary of DA is significantly different from ML and DS specialists.

7. Salary of juniors depending on job title

Hypotheses:

- H_0: there is no significant difference between junior salary median of data scientists, machine learning specialists, and data analysts

- H_1: there is significant difference Kruskal-Wallis test: p-value = 0.8073987464421589 (fail to reject H_0)



Kruskal-Wallis test failed to reject H0.

Limitations

- Due to the few expert level employees, it was not possible to compare salaries between ML, DS and DA specialists.
- Due to the little number of experts among ML and DA specialists, they were not included in the comparison for these two job types.
- Comparisons of salaries within each job type by experience level in each of the three years were not made, since very small samples were available. Therefore, such a comparison can be made in the future research using other data.
- When comparing salaries of data scientists, the samples have slightly different distribution shape, so the results may be inaccurate.

Contribution of team members

- Rufina Gafiiatullina: mean analysis for each year, the difference in salary between companies with different size salary depending on particular job, the difference in salary depending on country, influence of experience level on salary inside each job types, report, code organization.
- Arina Goncharova: the difference in salary depending on experience level, the difference in salary depending on remote ratio, the difference in salary depending on employment type, influence of experience level on salary of different job types, report.

Conclusion

Taking everything into consideration, the following findings can be highlighted:

- Salaries of specialists depend on experience level. Junior receive less, middle - average, senior - higher, and experts receive the highest salaries. Therefore, improving skills and experience is important to earn more.
- Salaries of specialists depend on remote ratio. Specialists who work partially remote receive less, who work in the office - average, who work remotely - higher.
- Salary depends on the size of the company. When you apply for a job in a large company, you are more likely to get a high salary than in medium and small sized companies. This should be taken into account when applying for a job.
- Salary depends on the employment type, so it is essential to consider this aspect for planning a career, as well as in a process of hiring an employee.
- Salary depends on the country.
- The probability of getting a large salary is higher for a data scientist than for a data analyst and machine learning specialist.
- In all three fields (data science, data analysis and machine learning), seniors typically earn significantly more than middle and junior specialists.
- Only among seniors a difference between the salaries of DA professionals and ML and DS professionals was revealed.

Consequently, based on the results obtained, the highest probability of getting a high salary is found for data scientists. In addition, it is important to constantly develop skills, since experience has a significant impact on the salary. Companies can also take advantage of the above findings when organizing their budgets and hiring specialists.