A green letter and black text

Description automatically generated

**Analysis on Data Science Salaries based on related features**

**NAME:**

**INSTRUCTOR:**

Data Visualization **INFO 5709**

**Semester Project**

**Introduction**

Over the last few years, the domain of data science has proved to be one of the most promising and developing areas of the IT market. Since the generation of data is rapidly increasing in different sectors to make competitive advantage, organizations are now focusing on data-driven approaches and thus the need for data science professionals is increasing. This report explores the dataset namely Data Science Salaries 2024, which offers the salary of the data science roles based on regions, experience, positions, and types of employment from the year 2020 to 2024.

Consequently, the goal of this report is to analyze frequencies and temporal changes of data scientists’ remuneration, to determine factors, which affect their wages, and to describe the dispersion of positions based on experience, remote work, and organizational company characteristics. In Tableau, we plan on using maps and filters to test various hypotheses and in Python’s data preprocessing/EDA to find patterns that may help job seekers, recruiters, and other industry participants.

Key Questions Addressed:

1. How does the salary increase or decrease depending on experience expiry, job position, and organization type?
2. How does remote work and employment type affect employee’s salary levels?
3. In what way can company scale be divided in terms of the availability of jobs and pay scales?

**Dataset Description**

The Data Science Salaries 2024 is a data set containing important data on salaries and associated attributes of data science professionals across industries at different organizations, locations, and experience levels. It contains 14838 records and 11 columns. Thus, the database is as from the year 2020 to 2024 containing essential data which helps in conducting a thorough analysis of the salary structure, job opportunities and determinants of pay in the data science field.

Key Features

1. work\_year:
   * Description: The year the salary was paid.
   * Example: 2023, 2024.
   * Data Type: Integer.
2. experience\_level:
   * Description: Level of experience required for the role.
   * Categories:
     + EN: Entry-level / Junior
     + MI: Mid-level / Intermediate
     + SE: Senior-level / Expert
     + EX: Executive-level / Director
   * Data Type: Categorical.
3. employment\_type:
   * Description: Type of employment contract.
   * Categories:
     + FT: Full-time
     + PT: Part-time
     + CT: Contract
     + FL: Freelance
   * Data Type: Categorical.
4. job\_title:
   * Description: The specific job title (e.g., Data Scientist, Machine Learning Engineer).
   * Data Type: Text.
5. salary:
   * Description: Total gross salary amount paid.
   * Data Type: Numerical (Float/Integer).
6. salary\_currency:
   * Description: The currency in which the salary is paid (e.g., USD, EUR).
   * Data Type: Categorical.
7. salary\_in\_usd:
   * Description: The salary amount converted to USD for standardization.
   * Data Type: Numerical (Float).
8. employee\_residence:
   * Description: The country where the employee resides during the work year.
   * Data Type: Text (ISO 3166 country codes).
9. remote\_ratio:
   * Description: The amount of work done remotely.
   * Categories:
     + 0: No remote work
     + 50: Partially remote
     + 100: Fully remote
   * Data Type: Categorical.
10. company\_location:
    * Description: The country where the employer's main office or contracting branch is located.
    * Data Type: Text (ISO 3166 country codes).
11. company\_size:
    * Description: The size of the company based on the number of employees.
    * Categories:
      + S: Small (less than 50 employees)
      + M: Medium (50 to 250 employees)
      + L: Large (more than 250 employees)
    * Data Type: Categorical.

**Dataset Source:** [**Data Science Salaries 2024 - Kaggle**](https://www.kaggle.com/datasets/yusufdelikkaya/datascience-salaries-2024)

**Tools Used:** Python, Tableau

**Data Preparation**

Using Python in Google Colab we process and verify the fields and explore categorical value and correlation of the data.

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

The correlation matrix revealed that work\_year has a very weak positive coefficient with salary\_in\_usd (0.09); this means that slightly salaries have improved in the current year and a moderate negative correlation with remote\_ratio (-0.20) that means newer jobs may contain higher remote work ratio. Moreover, using correlation analysis, remote\_ratio does not have any high correlation coefficient with salary\_in\_usd; therefore, remote work does not have a large impact on salary in this sample.

A graph of a number of salary

Description automatically generated with medium confidence

The histogram of salary\_in\_usd is right skewed as the salaries most frequently occur between $50,000 and $200,000 and a few salaries are more than $300,000. The overlaid density plot indicates that the distribution of salaries is dominated by a number between a hundred thousand to a hundred and fifty thousand.

In the boxplot below it is clear that salaries rise with experience level and executives (EX) have the highest median salaries and high variability because of the high value outliers. The results also showed that the median salaries of entry-level (EN) employees are the lowest while the variability is relatively low, suggesting higher homogeneity of pay at the junior level.

A screenshot of a computer screen

Description automatically generated

A screenshot of a graph

Description automatically generated

The above bar chart shows that the number of job postings has risen in the past few years to the extent of reaching their highest in 2023. But 2024 is less than 2023 which indicates that either the data science field is becoming static or there is a low demand for data science jobs.

A screen shot of a pie chart

Description automatically generated

The above pie chart demonstrates the distribution of the level of experience: Senior level (SE) accounts for 65.3% of the dataset, while Mid-level (MI) accounts for 23.9%. Minority status is indicated by the 7.7% representation of entry level (EN) employees and 3.0% of executive level (EX) employees in contrast to the industries selectivity for professionals with experience.

This Tableau map visualization is based on the average salaries (salary\_in\_usd) by country, which demonstrates the distinctions of the pays throughout the world. The countries with comparatively high average incomes comprise the USA and Switzerland; the countries with the small average income are the countries of Asia and Africa.

A screenshot of a computer

Description automatically generated

**Hypothesis:**

**Hypothesis 1: Senior-level employees tend to have higher salaries, and this pattern varies by company size.**

The following bar chart shows how Average salaries (Salary in USD) are distributed depending on experience level and company size. Data shows that employees at higher organization levels (e.g., SE, EX) are paid considerably higher; the highest wages are offered in large organizations (L). This pattern illustrates that experience and the size of the company that the employee works for will determine his compensation.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

This box plot represents the dispersion of the average salaries (Salary in USD) according to the experience level and it is sliced by the company size. The boxed areas (EX, SE) have relatively higher median salaries, especially among large firms (L); compared to entry-level (EN) and mid-level (MI) workers, their pay distribution is more widespread. This is why experienced, and company size should be considered when it comes to the salaries to be offered.

A screenshot of a computer

Description automatically generated

The above bubble chart displays the average salaries (Salary in USD) categorized by experience and firm size. The size of the bubble represents the salary level; EX and SE employees in M and L companies earn the highest salaries. To my mind, the chart does help to enhance the understanding of how experience and company size influence compensation graphically.

**Hypothesis 2: Remote work (high remote ratio) is associated with higher salaries, especially in executive roles.**

The below tree map shows the correlation between the remote ratio, average salary in USD and the type of employment. This implies that structures that have higher ratios of remote work are correlated to higher remunerations; especially in the executive capacity (EX). The visualization is a nice breakdown of salary distributions by employment type and the prevalence of remote work and highlights the increasing importance of remote positions.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

The above horizontal bar chart analyses the effect of Remote Ratio on Avg Salary in USD by employment types. Remote working ratios of the executives (EX) correlate with higher wages and are followed by entry-levels (EN) that have relatively lower average wages. The analysis truly demonstrates how distant relations and role hierarchy relates to pay.

A screenshot of a computer

Description automatically generated

This area chart shows how the levels of remote work (Remote Ratio), average salary in USD (Avg Salary) and experience interact with each other. They rise with higher remote ratios especially for the EX level, suggesting that there is a remote working premium for senior personnel. The trends included in the visualization are related to the employment type and experience level, and reveal

**Hypothesis 3: Large companies have more job roles offering higher salaries compared to small and medium-sized companies.**

The below bar chart shows the job title splits by experience level (EN, MI, SE, EX) on company size (L, M, S). Large organisations (L) can be seen to have varied number of positions especially the senior and middle level jobs. On the other hand, for the small companies (S), there are fewer positions and most of them are occupied by junior ranking and mid-ranking employees, best illustrating the differences of structural HRM between companies with different sizes.

A screenshot of a computer

Description automatically generated

The below bar chart that shows number of unique job title, company size and experience level Company size: Large (L), Medium (M), Small (S) Company experience level: Entry level (EN), Mid-level (MI), Senior level (SE), Expert level (EX) Larger organizations possess a wider range of senior (SE) and mid-level (MI) job titles than do small organizations (S) that offer a concentration of entry-level (EN) and mid-level positions. This demonstrates how diversity work is done in larger organizations in contrast with the smaller ones.

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

The above highlighted table below shows the number of unique job titles based on experience level, EN: Mid-level, MI: Senior and EX: Executive experience and company size L: Large, M: Medium, S: Small. Shades with higher intensity mean higher counts: thus, mid-level (MI) and senior-level (SE) positions are presented as more common in medium (M) and large (L) companies. This shows that job diversity differs depending on the level of experience and the size of the organisation.

**Dashboard**

This dashboard is an accumulation of several charts to analyze the salary distribution based on different factors. It is seen from the tree map that the high remote ratio has a strong positive correlation with salary, especially for the EX-category exclusively working remotely. The salaries are geographically distributed, areas such as North America and Europe receive more average salaries. Referring to the bar chart, the data show that employees with a senior level of experience (SE, EX) receive higher salaries, and the highest paid are the employees from large companies (L). Further, the Heatmap illustrates that mid-level (MI), and senior-level (SE) positions are more diverse across companies, especially large ones. Combined, the two visualizations give a clear insight into how elements such as experience, working remotely, size of the company in which one works, and the geographical location affect the salary.

A screenshot of a computer

Description automatically generated

**Conclusion**

Based on analysis, it has helped to obtain valuable patterns concerning the presence of relation between the level of experience, company size, remote work, geographical location and salaries in the data science industry. Thus, the SE and EX employees receive higher wages with a significant difference in large organizations (L) with a high level of remote work. Flexible work arrangements indeed are more correlated with increased salaries, especially for executives employed in the public service. On a geographic aspect, the areas like North America and Europe show higher average salaries for its employees, which follows market trends and the cost-of-living index. Large organizations provide more options when it comes to jobs than small organizations and although both could provide both junior and mid-level positions, small organizations seldomly provide senior level opportunities. The results of the current study underscore the importance of structure, remote work practices, and workers’ experience regarding compensation.

**References**

*Data Science Salaries Dataset*. (2024, October 23). Kaggle. <https://www.kaggle.com/datasets/yusufdelikkaya/datascience-salaries-2024?resource=download>