Final Project

PSTAT100: Data Science Concepts and Analysis

Ali Abuzaid

STUDENT NAME

- Sophie Lian (sophielian)
- Veronica Strempfer (vstrempfer)
- Paridhi Jay Singh (paridhijsingh)
- Bahaar Ahuja (bahaar)

Due Date

The deadline for this step is **December 3, 2024**.

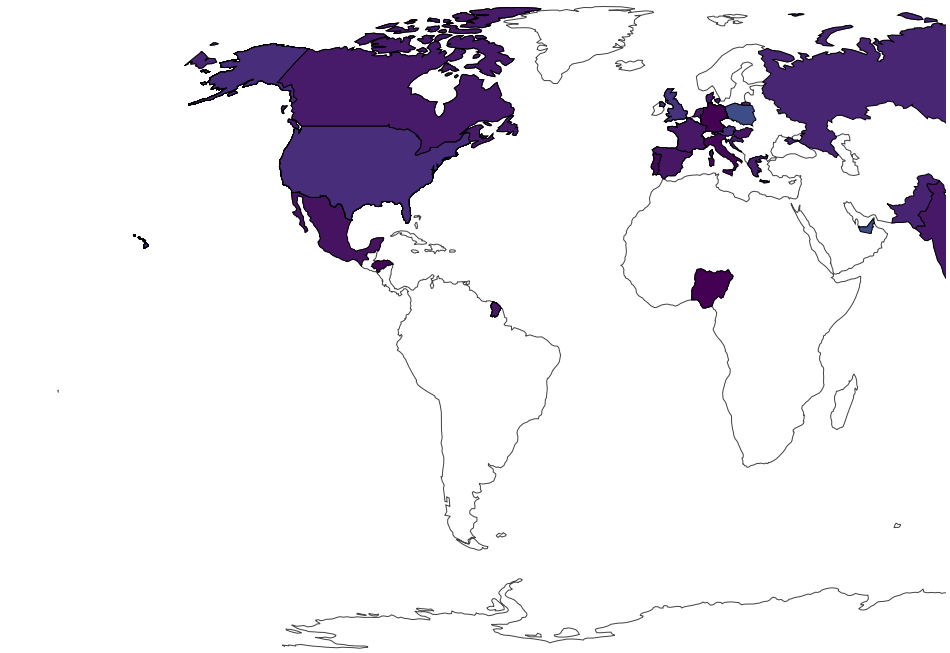
Instructions

In this step, you will conduct a thorough data science analysis based on the foundational work you've completed in Step 2. This analysis should be comprehensive, addressing your research questions and testing your hypotheses with suitable models. Additionally, you will prepare a professional report and presentation to showcase your findings, interpretations, and conclusions.

1 Introduction

The data set that we decided to analyze for this project is DataScience_salaries_2024. The objective for our project focuses on understanding key factors that influence data science salaries and how geographical distributions affect various Data Science roles across the U.S. This data set explores different data science roles and their respective salaries as well aspects such as experience level, employee residence, and remote work ratio. This dataset can be utilized for analyzing compensation patterns and trends in data science roles across various demographics.

Data Science Salary Distribution by Country



Play

2020 2021 2022

The map above demonstrates the range of values that the salary variable takes as well as the countries represented, with most of the data coming from US and GB across different years, with there being minimal changes in the average salary over the years.

1.1 Initial Research Questions

1. How do different factors of an employee's working conditions affect their overall salary?
2. How do different geographic locations affect the distribution of different salaries (USD)?

Initially, during Step 2, we decided to explore how various factors influenced the variable salary_in_usd. However, when performing the linear regression step, we realized that only numeric variables were applicable, which led us to focus on experience_level (categorized as Junior, Mid-level, Senior, and Expert) and an employee's remote_ratio. Unfortunately, our dataset had limited numeric variables, restricting our analysis to the relationship between salary_in_usd and these two factors.

For the first question, we decided to visualize using a boxplot, which displayed a positive relationship between salary_in_usd and experience_level, with the salary steadily increasing as employees progress from Entry Level to Mid-Level, Senior, and Expert roles. This trend highlights how career advancement and accumulated expertise are rewarded with higher compensation. The widening salary range at the Expert level reflects greater variability in top-tier roles, emphasizing the value of experience in achieving higher earnings.

1.2 New Research Questions

1. How does an employee's experience level affect their salary (USD)?
2. How does an employee's remote ratio affect their salary (USD)?

1.3 Hypothesis

1. The higher an employee's experience level, the higher their salary.
2. The higher an employee's remote ratio, the higher their salary.

2 Question 1: How does an employee's experience level affect their salary (USD)?

3 Simple Linear Regression Model

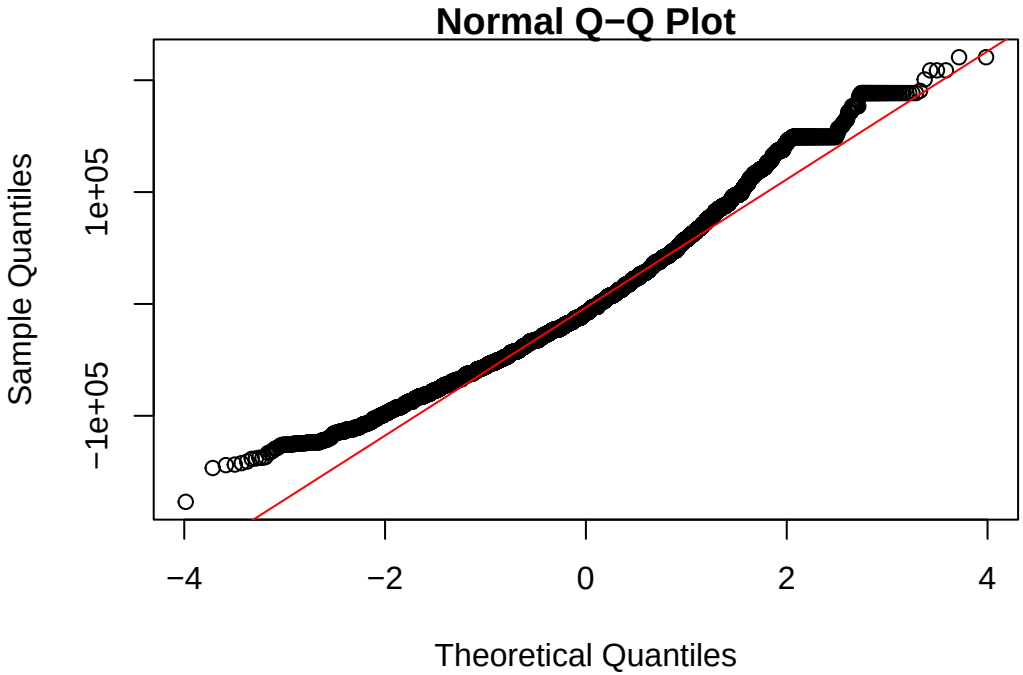
To analyze the first question we chose to do a simple linear regression model, using an employee's experience level to predict the changes in employee salaries.

```
Call:
lm(formula = salary_in_usd ~ experience_level, data = data_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-176902  -41761   -7930   35479  220529

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)           91221       1715   53.20   <2e-16 ***
experience_levelMI    32040       1972   16.25   <2e-16 ***
experience_levelSE    71254       1813   39.29   <2e-16 ***
experience_levelEX   100681       3255   30.93   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 58100 on 14834 degrees of freedom
Multiple R-squared:  0.1488,    Adjusted R-squared:  0.1486
F-statistic: 864.5 on 3 and 14834 DF,  p-value: < 2.2e-16
```



As we can see from the generated plot above, the residuals of the model follow a linear pattern, alongside the red line, which implies that the residuals are normally distributed and thus a linear regression model is a valid model to be used on our data. Additionally, as we can see from the coefficients produced when creating the model, the p-values for all variables are less than 0.001. This reveals that these coefficients are significant values at a 0.001 significance level.

3.1 Interpretation of Results

The model generated above shows that as expected, the projected salary increases as an employee's experience level increases. Entry-Level → Mid-Level: Salary increases by $32040 Mid-Level → Senior: Salary increases by $71254 Senior → Expert: Salary increases by $100681

Thus, we can conclude that there is significant evidence that an employee's experience level is positively correlated with salary, which is consistent with our original hypothesis.

However, the R-Squared value for this model is only 0.1486, meaning that only a small portion of the variation of salary is accounted for in this model.

3.2 Updated Model

We decided to update our simple linear regression model to a multiple linear regression model using both experience level and work year to try and predict an employee's salary.

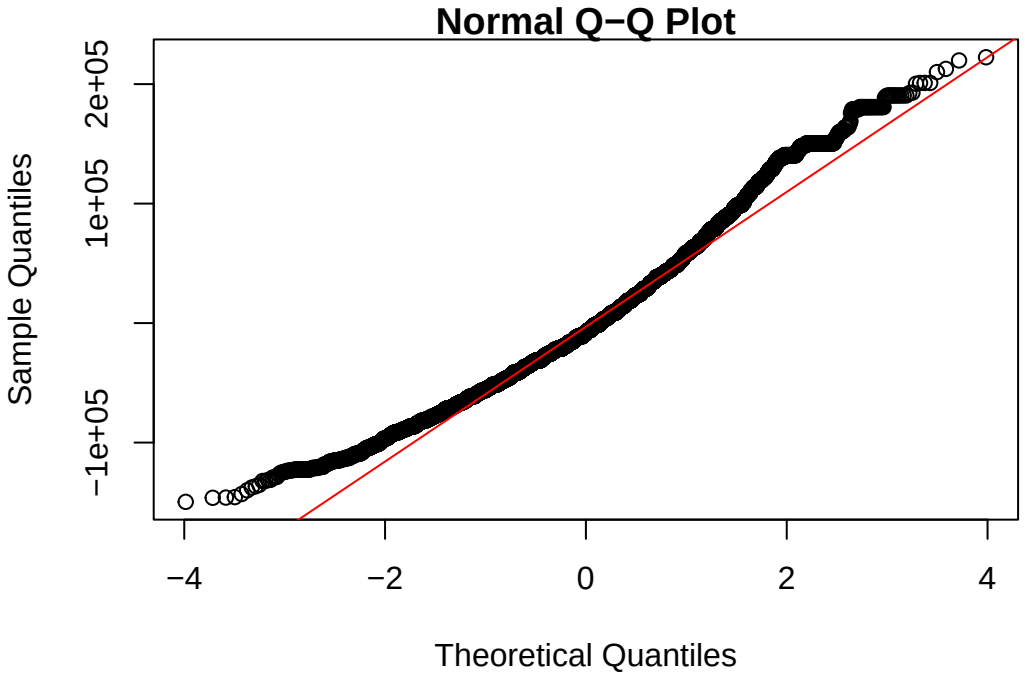
```
Call:
lm(formula = salary_in_usd ~ experience_level + work_year, data = data_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-149608  -41093   -7843   34997  222486

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -1.967e+07  1.370e+06  -14.35   <2e-16 ***
experience_levelMI  3.197e+04  1.959e+03   16.32   <2e-16 ***
experience_levelSE  7.218e+04  1.802e+03   40.05   <2e-16 ***
experience_levelEX  1.011e+05  3.232e+03   31.27   <2e-16 ***
work_year           9.765e+03  6.774e+02   14.42   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57700 on 14833 degrees of freedom
Multiple R-squared:  0.1606,    Adjusted R-squared:  0.1604
F-statistic: 709.4 on 4 and 14833 DF,  p-value: < 2.2e-16
```

We completed the same steps as above, checking the plot of the residuals to confirm that they are approximately normally distributed and that we can use a linear regression model for this data.



As we can see from the coefficients in our generated model above,

Entry-Level → Mid-Level: Salary increases by $31970 Mid-Level → Senior: Salary increases by $72180 Senior → Expert: Salary increases by $101100 Year increases by one unit → Salary increases by $9765

Additionally, the p-values for all of the coefficients stated above are less than 0.001, implying that these coefficients are significant at the 0.001 level of significance.

The value of R-Squared for this updated model is 0.1604, which is an improvement from our last model, meaning that this model using both experience level and work year explains a larger proportion of the variability contained in the salary data than our last model. However, the consistently low R-Squared values for our model suggests that there are many more variables not included that our dataset that highly correlate with salary

4 Question 2: How does an employee's remote ratio affect their salary (USD)?

4.1 Multiple Linear Regression Model

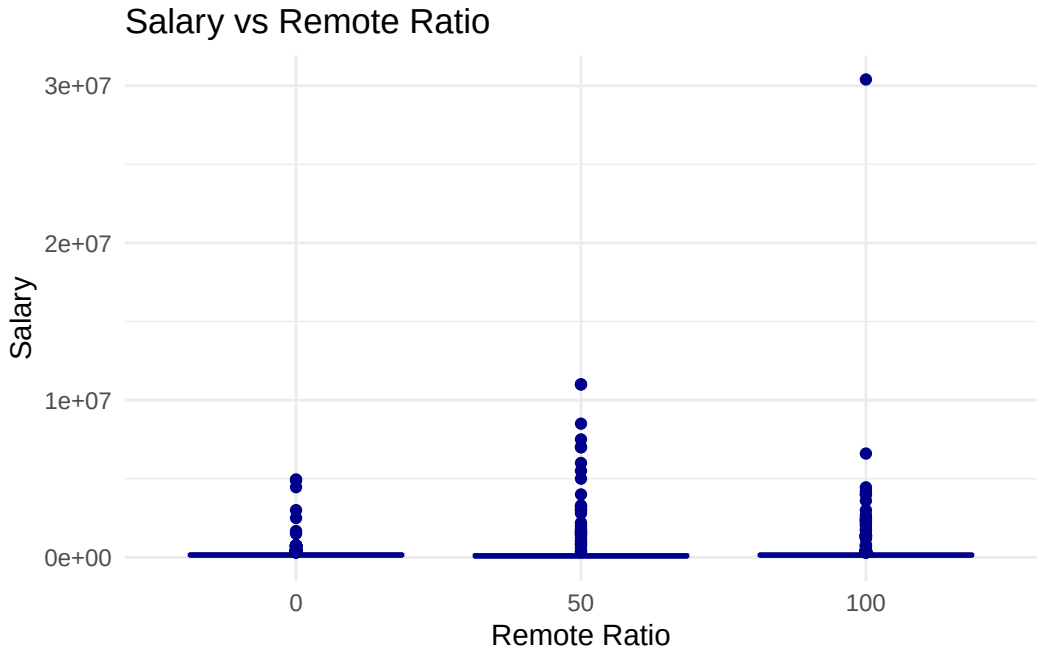
```
Regression Model MSE: 23320965999
```

```
Multiple R-squared: 0.9288203
```

```
Adjusted R-squared: 0.927069
```

```
Residual standard error: 104456.4
```

```
F-statistic: 530.3822 on 285 DF
```



4.2 Interpretation of Results

We notice that most salaries are clustered near lower values across all remote ratio categories – whether it's 0% (on-site), 50% (hybrid), or 100% (fully remote). This indicates that a majority of roles fall within a similar range of lower salaries. However, the dataset also shows extreme outliers. These could correspond to rare job titles or location-specific roles where salaries are exceptionally higher. The extreme outliers in the salary dataset cause there to seem like there's a correlation between salary and remote ratio. However, if we were to create a future model that filtered out these outliers we might see that with the average salary value there is not a significant change of salary as the remote ratio increases. Across all types of remote ratios there is generally the same salary. Interestingly, when we look at the distributions across different remote ratios, there's a significant amount of overlap. This suggests that the remote work ratio might not have a strong impact on salaries overall, at least within this dataset. The R-squared value is around 0.928, meaning 92.8% of the variance in salaries is explained by the model. However, the Mean Squared Error (MSE) of over 233 million reflects the impact of extreme outliers on model accuracy.

The residual standard error of 104,500 indicates the average error in predicting salaries. This error might seem high due to extreme outliers in the dataset. The degrees of freedom of 11,584 highlight that the analysis is based on a large dataset, making the model estimates more robust.

5 Conclusion

The different steps of this project had us evaluate our data set with different visualizations. We were able to see which variables had the most influence on salary as well as understand the limitations that came with this data set. The lack of numerical variables made regression analysis difficult to perform and going forward in projects this is something to consider. This limitation also led us to think of alternative ways to convey ideas and gain a deeper understanding of relationships between variables. Additionally, the bias within this dataset caused us to rethink how to analyze our data such as the salary having a large range of values that made certain correlations difficult to see as well as the amount of data collected for each country. Overall, this data set helped us learn more about a variety of visualizations and understand what visualizations work best for certain data.