Final Project DSC 520

Nick Blackford

2024-03-01

Part 1

Predicting the Salary of Various Data Science Roles Using Linear Regression

Introduction

The field of data science has seen exponential growth over the past decade, driven by the increasing availability of data and the need for sophisticated analysis to drive decision-making in businesses and research. A key aspect of this growth is the demand for skilled professionals capable of translating data into actionable insights. Consequently, understanding the factors that influence the salary of data science roles is crucial for both employers and employees. This research aims to delve into the intricacies of salary determinants in the data science field, employing linear regression to model these relationships. Such an analysis is not only of interest to those directly involved in the field but also contributes to broader discussions on education, skill development, and career planning in the burgeoning era of data.

Research Questions

What are the primary factors that influence the salary of data science roles? How does work experience correlate with salary in the data science field? Do educational qualifications significantly impact the salary of data science professionals? Is there a geographical variance in the salary of data scientists across different regions? How do industry and company size affect data science salaries? What role do programming skills and tool proficiency play in determining salary? Can linear regression accurately predict salaries in the data science field based on identified factors?

Approach

To tackle these questions, linear regression models will be developed, focusing on the relationship between salary and various predictors such as experience, education, location, industry, and skill set. This approach allows for the quantification of the impact of each factor on salary, offering insights into the most influential determinants.

How the Approach Addresses the Problem

By modeling the salary determinants, this research will provide valuable guidelines for data science professionals seeking to enhance their earning potential and for employers aiming to establish competitive salary structures. It will also highlight areas for further skill development for aspiring data scientists.

Data

Three datasets have been identified for this analysis:

Stack Overflow Developer Survey Data

Description: The Stack Overflow Annual Developer Survey includes responses from thousands of developers worldwide. It covers various aspects, including programming languages used, education levels, years of professional coding experience, and salary information.

Potential Use: This dataset can help analyze how different factors, like education level, coding experience, or familiarity with specific technologies, relate to salaries in the data science field.

Access: https://insights.stackoverflow.com/survey

Kaggle Data Science & ML Salary Survey Description: Kaggle conducts an annual machine learning and data science survey that collects detailed information on the state of data science and machine learning. Among many other things, the survey includes questions about salary, education, employment, and the tools used.

Potential Use: You can use this dataset to explore the relationship between tool usage, education, job roles, and salary in the data science community.

Access: https://www.kaggle.com/competitions/kaggle-survey-2022

Kaggle Data Science Salaries 2023 Description: Salaries of various data science job scraped from various job boards

```
Potential use: various qualitative and quantitative variables for each job posting
```

Access: https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023

Required Packages

tidyverse for data manipulation and visualization

ggplot2 for creating plots

lmtest and car for diagnostic tests on linear regression models

dplyr for data manipulation

readr for reading CSV data files

Plots and Table Needs

Scatter plots to visualize the relationship between salary and continuous predictors (e.g., years of experience).

Box plots to compare salary distributions across categorical variables (e.g., education level, region).

Correlation matrices to identify potential multicollinearity among predictors. Regression diagnostics plots to assess model assumptions.

Questions for Future Steps

How to incorporate categorical variables with many levels (e.g., country or state) into the regression model without overfitting?

What are the best practices for handling missing data in salary datasets?

Step 2

We will be focusing on the Kaggle 2023 Survey Results to perform regression and predict salary of a data science role based on several categorical variables. For importing and cleaning the data, we will start by removing null and duplicate values. Then, we will filter for only US salaries as that as is what I'm most interested in predicting. Finally, we will examine the shape of the dataset and ensure the data types for each column make sense.

Examining Nulls, Duplicates, and Data Types

```r
# Load the necessary packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(readr)

# Load the dataset
ds_salaries <- read_csv("/Users/nickblackford/Desktop/R/ds_salaries.csv")
```

```
## Rows: 3755 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): experience_level, employment_type, job_title, salary_currency, empl...
## dbl (4): work_year, salary, salary_in_usd, remote_ratio
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Check for missing values
missing_values <- sum(is.na(ds_salaries))

# Check for duplicates
duplicates <- sum(duplicated(ds_salaries))

# Remove duplicates
ds_salaries <- ds_salaries %>% distinct()
```

```r
# Filter the dataset for employee residence in the US
ds_salaries_us <- ds_salaries %>%
  filter(employee_residence == "US")

# Check the shape of the dataset
dataset_shape <- dim(ds_salaries)

# Check the data types of each column
data_types <- sapply(ds_salaries, class)

# Output the results
cat("Missing Values:\n", missing_values, "\n")
```

```
## Missing Values:
##  0
```

```r
cat("Duplicates:\n", duplicates, "\n")
```

```
## Duplicates:
##  1171
```

```r
cat("Dataset Shape:\n", dataset_shape, "\n")
```

```
## Dataset Shape:
##  2584 11
```

```r
cat("Data Types:\n")
```

```
## Data Types:
```

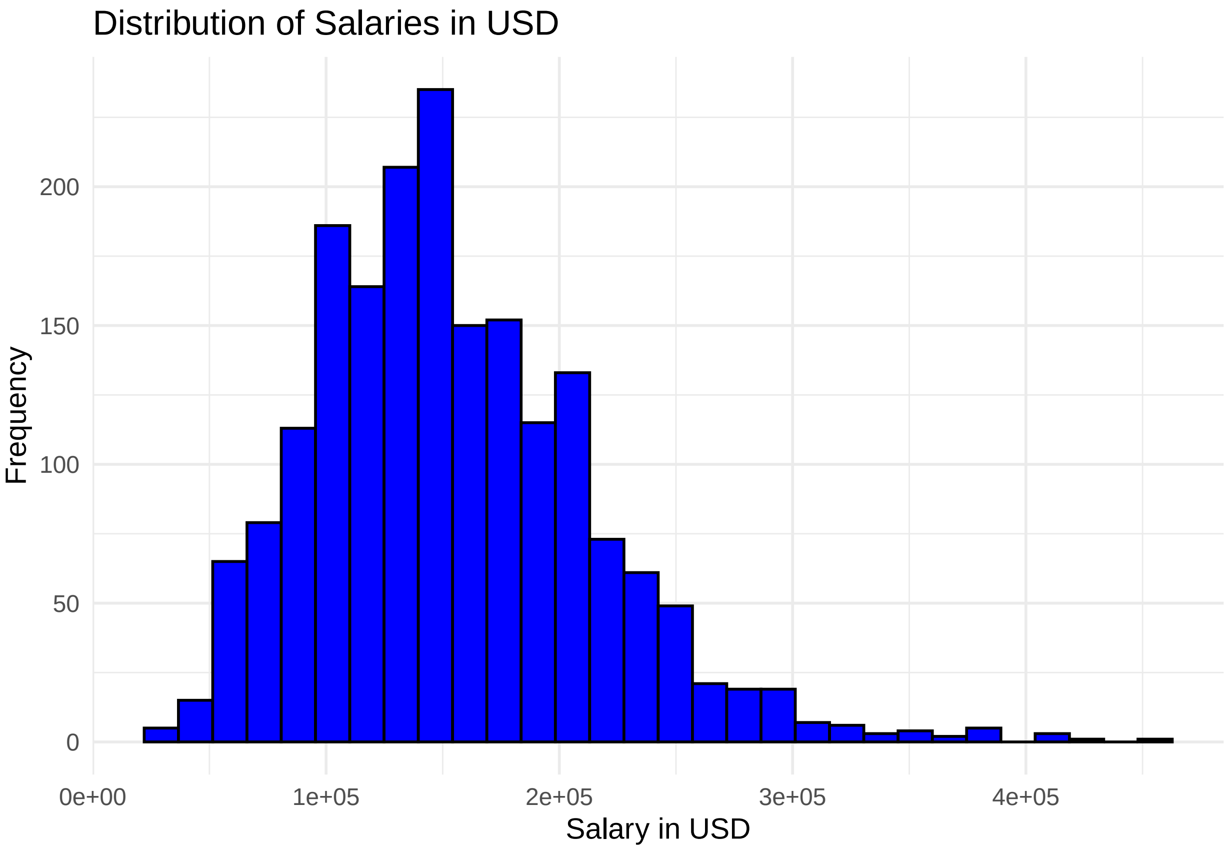
```r
print(data_types)
```

```
##          work_year   experience_level    employment_type          job_title
##          "numeric"        "character"        "character"        "character"
##             salary    salary_currency      salary_in_usd employee_residence
##          "numeric"        "character"          "numeric"        "character"
##       remote_ratio   company_location       company_size
##          "numeric"        "character"        "character"
```

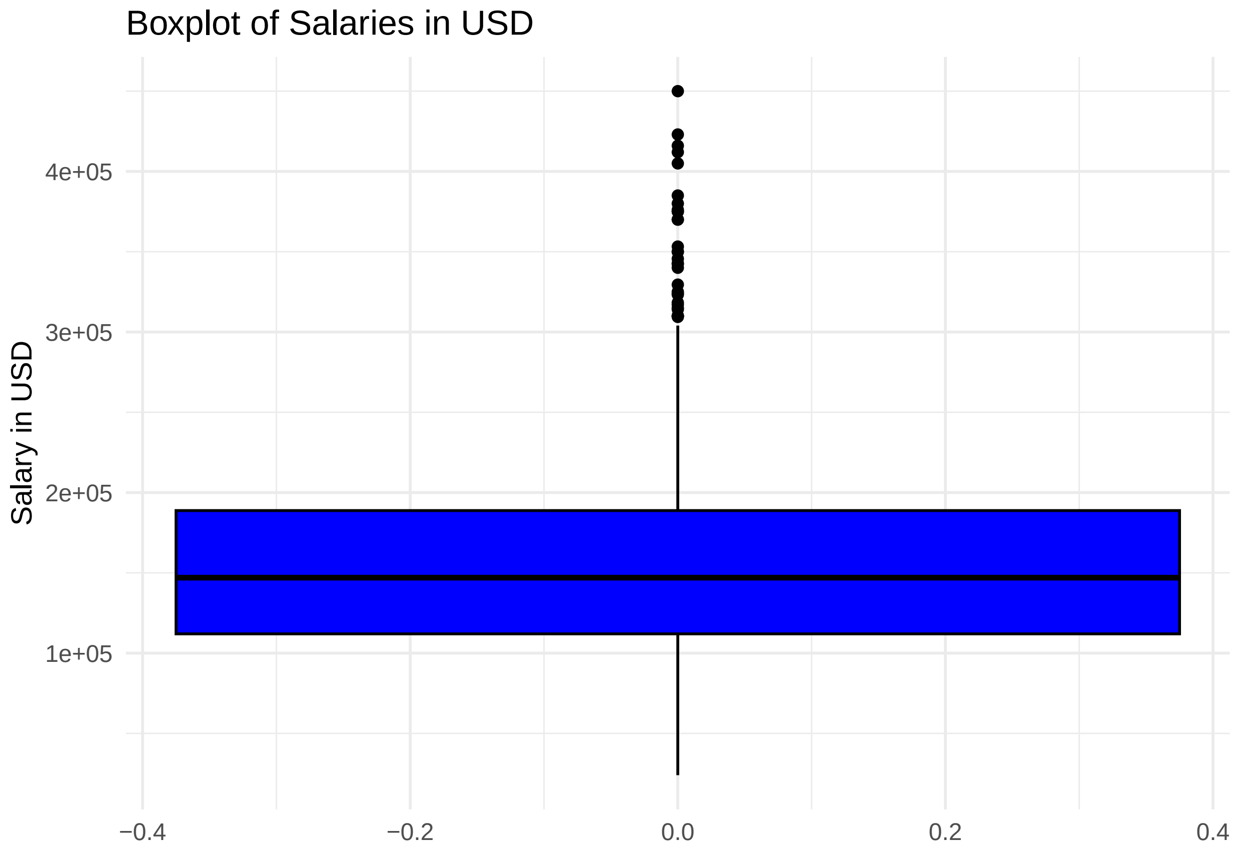
Next, we will take a look at the shape of the independent variable "salary" and check for outliers. ## Examining Outliers and Plotting the Data

```r
# Summary statistics for the salary column
salary_summary <- summary(ds_salaries_us$salary_in_usd)

# Histogram to visualize the distribution
library(ggplot2)
ggplot(ds_salaries_us, aes(x = salary_in_usd)) +
  geom_histogram(bins = 30, fill = "blue", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Salaries in USD",
       x = "Salary in USD",
       y = "Frequency")
```



```r
# Boxplot to visualize outliers
ggplot(ds_salaries_us, aes(y = salary_in_usd)) +
  geom_boxplot(fill = "blue", color = "black") +
  theme_minimal() +
  labs(title = "Boxplot of Salaries in USD",
       y = "Salary in USD")
```



```
# Outlier Detection using IQR
Q1 <- quantile(ds_salaries_us$salary_in_usd, 0.25)
Q3 <- quantile(ds_salaries_us$salary_in_usd, 0.75)
IQR <- Q3 - Q1

lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

outliers <- subset(ds_salaries_us$salary_in_usd, ds_salaries$salary_in_usd < lower_bound | ds_salaries$s

# Print the results
cat("Summary Statistics for Salary in USD:\n")
```

```
## Summary Statistics for Salary in USD:
```

```
print(salary_summary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   24000  112000  147000  153972  188800  450000
```

```
cat("\nNumber of Outliers:", length(outliers))
```

```
##
## Number of Outliers: 33
```

After examining the dataset for outliers, we will proceed with the analysis while including all outliers as salary data is often skewed to the right, and it is not unreasonable that an executive or senior director in data science could make up to $450,000 annually.

The next step to prep the data for regression is to mutate the data and create dummy variables for the categorical data in the dataset. We want to filter for only data scientist as job_title as this is the field we are interested in estimating salary for. Next, we will change all categorical variables to dummy variables so that we can perform regression on this dataset.

Mutating the Data for Linear Regression

```
# Filter the dataset for the job title "Data Scientist"
ds_salaries_ds <- ds_salaries_us %>%
  filter(job_title == "Data Scientist")

# Convert categorical variables to dummy variables
ds_salaries_ds <- ds_salaries_ds %>%
  mutate(remote_ratio = as.factor(remote_ratio),
         company_size = as.factor(company_size),
         experience_level = as.factor(experience_level)) %>%
  select(salary_in_usd, experience_level, remote_ratio, company_size) %>%
  pivot_longer(-salary_in_usd, names_to = "variable", values_to = "value") %>%
  mutate(id = row_number()) %>%
  pivot_wider(names_from = c(variable, value), values_from = value, values_fill = list(value = 0), valu
  select(-id)
```

Final Analysis of Clean Dataset

We finally have the dataset in a state where we are ready to perform regression on it. We plan to discover information that is not evident by performing regression on the dataset to see how each categorical variable contributes to the salary of a data scientist. We have sliced and diced the data in various ways - filtering for U.S. only, filtering on job title, and mutating to create dummy variables. We will not be performing any joins but could achieve a greater sample of data by merging multiple datasets with the same salary information.

Something I don't know right now is how to clean data that is even messier - user entered text data and filtering by like items. Stack overflow had a more comprehensive dataset with many more predictors but I was unable to get it into a state where regression could actually be performed.

The box plots and histograms used in this report did a great job to answer some key questions that could be asked about the dataset, particularly surrounding salary data. The summary statistics also contribute to giving one an idea of what the data looks like.

Performing Linear Regression on the Dataset

```
# Fit the linear regression model
model_ds <- lm(salary_in_usd ~ ., data = ds_salaries_ds)


#Summary of the model
summary(model_ds)
```

```
##
## Call:
## lm(formula = salary_in_usd ~ ., data = ds_salaries_ds)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
## -128679  -32671   -6172   28575  260083 
## 
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            94643      19467   4.862 1.32e-06 ***
## experience_level_SE    72036      19698   3.657 0.000267 ***
## remote_ratio_0         64994      19809   3.281 0.001065 ** 
## company_size_M         64020      19667   3.255 0.001166 ** 
## remote_ratio_100       61529      19835   3.102 0.001969 ** 
## experience_level_EX    90076      26656   3.379 0.000751 ***
## company_size_L         57274      20958   2.733 0.006375 ** 
## experience_level_EN     8182      21931   0.373 0.709146    
## experience_level_MI    34777      20520   1.695 0.090382 .  
## remote_ratio_50        10579      25956   0.408 0.683645    
## company_size_S            NA         NA      NA       NA    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 51500 on 1157 degrees of freedom
## Multiple R-squared:  0.06477,    Adjusted R-squared:  0.0575 
## F-statistic: 8.904 on 9 and 1157 DF,  p-value: 4.59e-13
```

Part 3

Introduction

The field of data science has seen exponential growth over the past decade, driven by the increasing availability of data and the need for sophisticated analysis to drive decision-making in businesses and research. Understanding the factors that influence the salary of data science roles is crucial for both employers and employees. This research aims to delve into the intricacies of salary determinants in the data science field, employing linear regression to model these relationships.

Problem Statement

The primary objective was to identify the factors that influence the salary of data science roles, including the impact of work experience, educational qualifications, geographical location, industry, company size, and programming skills.

How the Problem Was Addressed

The analysis utilized three datasets, focusing on the Kaggle 2023 Survey Results. The data was cleaned by removing nulls, duplicates, and filtering for US salaries. Linear regression models were developed to quantify the impact of various predictors on salary.

Analysis

The final analysis revealed several insights:

Experience levels, remote work ratios, and company sizes were significant determinants of salary. Senior-level professionals and those working remotely full-time had higher salaries. Medium-sized companies tended to pay more than small or large companies. The model had a Multiple R-squared value of 0.06477, indicating that the predictors explained about 6.5% of the variance in salary.

Implications

The findings provide valuable guidelines for data science professionals seeking to enhance their earning potential and for employers aiming to establish competitive salary structures. It also highlights areas for further skill development for aspiring data scientists.

Limitations

The analysis was limited to US salaries and the job title of Data Scientist. Additionally, the model's explanatory power was relatively low, suggesting that there may be other significant factors not captured in the analysis.

Concluding Remarks

This research offers a foundation for understanding salary determinants in data science. Future studies could expand the scope to include more diverse job titles and geographical locations, and employ more sophisticated modeling techniques to capture the complex nature of salary determinants.