Stat 1261 Project Part 2

Noah Blayney, Ian Keller, Emma Hubbard, Philip Kim

2023-12-12

Abstract

In this project, we used a multiple linear regression model to explore the relationship between various explanatory variables and the response variable of annual salary in the data science field. The initial model used all variables in the dataset to analyze the relationship, while the final model prioritized using certain variables that provided a more effective analysis. Using a linear regression model allows us to explore the data and eventually make predictions about salary. The distribution of annual salaries itself was heavily right-skewed, which was important to consider when analyzing the relationship with other variables. While this model gives us a great insight into what impacts a data scientist's annual salary, there are some factors not considered that could be helpful for future research.

Introduction

As data science and data-based policies continue to progress, there are more opportunities for employment across a variety of industries such as healthcare, technology, finance, education, and more. The data science job field has continued to grow and expand in recent years, as this is a broad field that allows for many different career types and skill sets. Careers can range from data analysts to data architects, with many important positions in between. This project and paper hope to evaluate the factors that potentially affect a data scientist's annual salary, in United States dollars. Through this data evaluation and exploration, we hope to develop a predictive model that is able to estimate a data scientist's salary given certain variables. The data science field has so many opportunities for employment across different sectors and for individuals with varying skills. The main question that we wanted to answer with this analysis was what factors have the greatest impact on data scientists' annual salaries.

Statement of Purpose

The purpose of this paper is to gain insight into the professional realm of data science, and to understand the factors that influence annual salaries in the field. As we are data science students in a world that continues to advance technologically, this is especially important information to have access to. This project allows us to do our own research and work with data, and ultimately learn more about the industry and what variables matter most.

Methodology

This section explains where the data set we used came from. What type of variables are in the data set. How will these variables help answer our research question. And the plan to apply modeling and answer the question statistically.

Data Set

The Data Science Job Salaries published by Ruchi Bhatia and hosted on Kaggle at https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries was our data set for this project. The data set has 11 variables and 606 entries. The data set was collected using data scraping on ai-jobs.net. Kaggle's usability rating scores it at 100%, meaning it is easy to get started modeling with the data set.

Variables

The target variable in our project is Salary in USD. This was chosen so that we can see the other variables' impact on the expected salary of a Data Scientist. Explanatory values included company size, job title, year, employee residence, company location, remote ratio, and employment type. A majority of entries were based in the US which limits the scope of our findings.

Modeling and Analysis Plan

First we plan to prepare the dataset for modeling by creating train test splits and encoding categorical variables. The model chosen will be multiple linear regression to explore how each explanatory value impacts a data scientist's salary. This will be part of our resulting analysis where each variable has estimates attached.

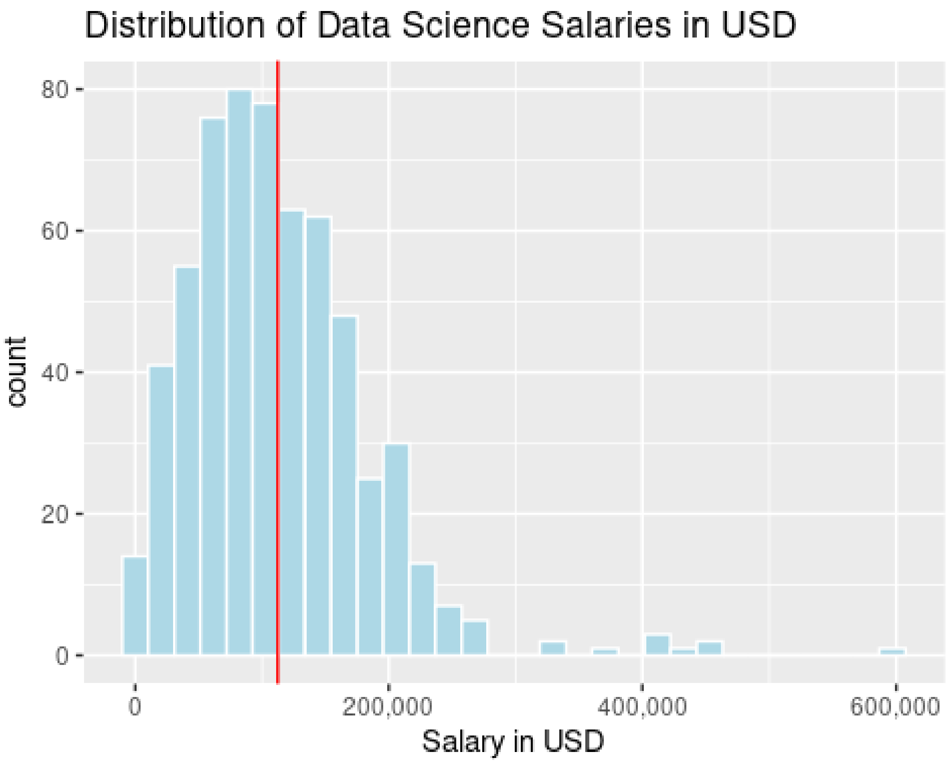
Results

EDA

Distribution of the Response Variable

Data science salaries are skewed to the right, with a mean (shown by the red line) of $112,298 and median (blue line) of $101,570. The mean of salaries is greater than the median due to the outliers with a high salary value. For example, there are observations in the data set that have salaries as high as ~$600,000, which will heavily skew the distribution of the data. These outliers are most likely values that correspond with employees with very high standing in the company, or someone with usually impressive qualifications, which should be considered when modeling. A shapiro test results in a p-value that is almost zero, indicating that the distribution is non-normal.

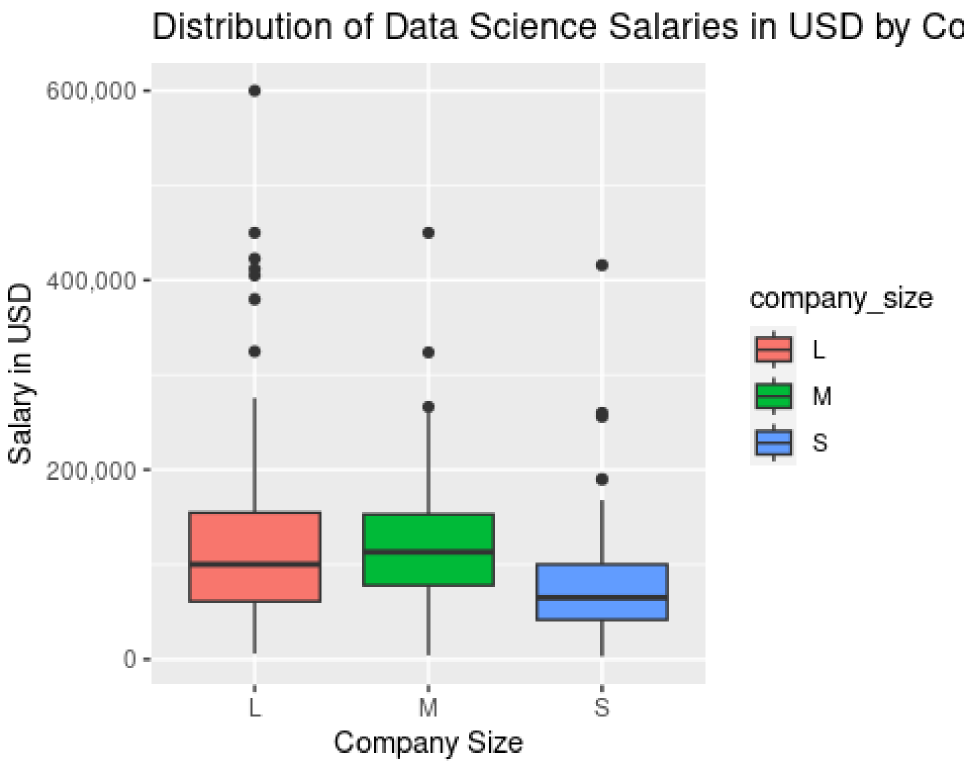
```r
# By Philip Kim
data %>%
  ggplot(aes(x = (salary_in_usd))) +
  geom_histogram(fill = 'light blue', color = 'white') +
  scale_x_continuous(labels = comma) +
  labs(title = 'Distribution of Data Science Salaries in USD', x = 'Salary in
USD') +
  geom_vline(xintercept = mean(data$salary_in_usd), color = 'red')

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Company Size

Large companies have the greatest range and spread. There doesn't seem to be a significant difference between the medians for company size, as the boxes all overlap. There are quite a few large outliers for all 3 company sizes, indicating that the distribution of this variable is skewed right. It is interesting to see that there are high outliers in each company size, but no low outliers. Moreover, it seems that in the data science field, it is very unlikely for someone to have an unusually low salary across all company sizes, but it is pretty common for someone to have an usually high salary. In addition, it seems more likely that someone would have an unusually high salary at a large company compared to a medium or small company. Further analysis is required to determine whether to include company size as a predictor.

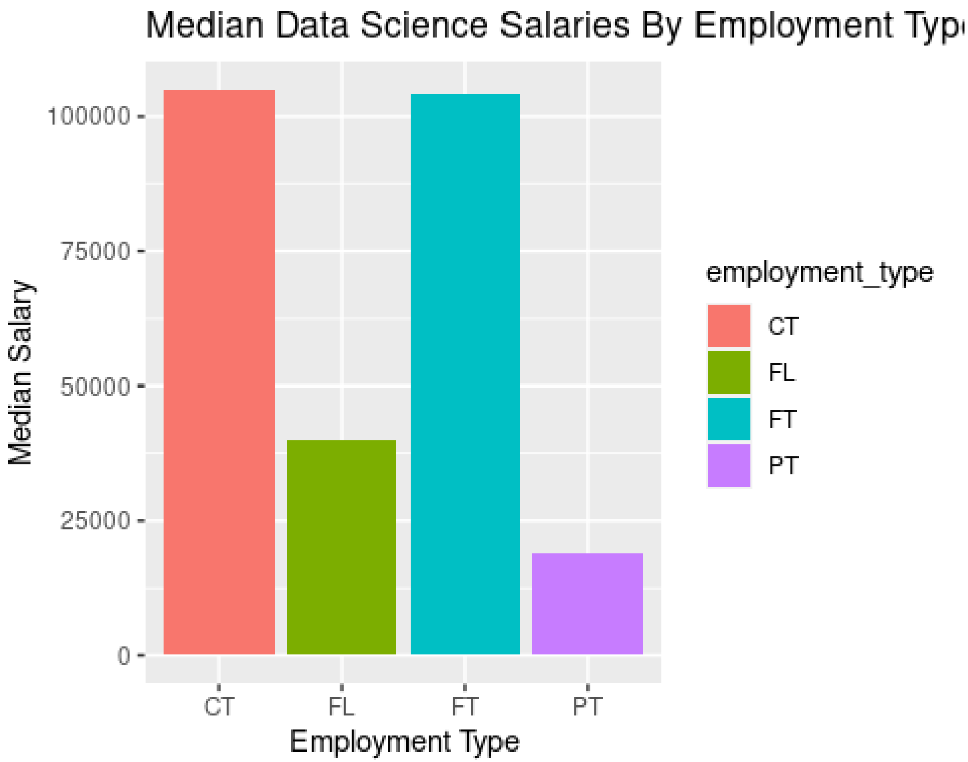
```
# By Philip Kim
data %>%
  ggplot(aes(x = company_size, y = salary_in_usd, fill = company_size)) +
  geom_boxplot() +
  scale_y_continuous(labels = comma) +
  labs(title = 'Distribution of Data Science Salaries in USD by Company
Size',
      x = 'Company Size', y = 'Salary in USD')
```



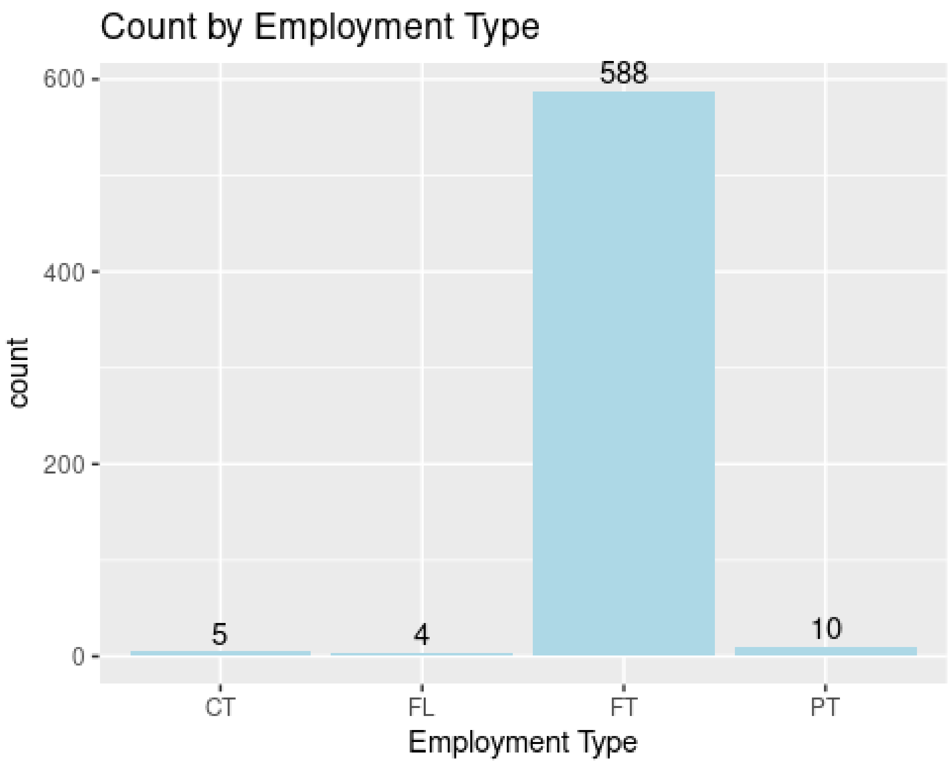
Employment Type

Due to the outliers, we take the median salary by employment type rather than the mean. We can see that contract time (CT) and full time (FT) have similar median salaries. Freelance (FL) and part time (PT) are significantly lower compared to the 2 previously mentioned employment types. This is logical, as part time workers in general make less money due to working less hours. Freelancers tend to work as a “side-hustle”, so their annual salaries are most likely lower as well. However, taking a look at the count by employment type, there are almost no data points that are under CT, FL, or PT, indicating that most of the data scientists in the dataset are full time employees.

```
# By Philip Kim
data %>%
  group_by(employment_type) %>%
  summarise(avg_sal_by_emptype = median(salary_in_usd)) %>%
  ggplot(aes(x = employment_type, y = avg_sal_by_emptype, fill =
employment_type)) +
  geom_bar(stat = 'identity') +
  labs(title = 'Median Data Science Salaries By Employment Type', x =
'Employment Type', y = 'Median Salary')
```



```
data %>%
  ggplot(aes(x = employment_type)) +
  geom_bar(fill = 'light blue') +
  geom_text(stat = 'count', aes(label=..count..), vjust = -.3) +
  labs(title = "Count by Employment Type", x = 'Employment Type')

## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2
3.4.0.
## ℹ Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

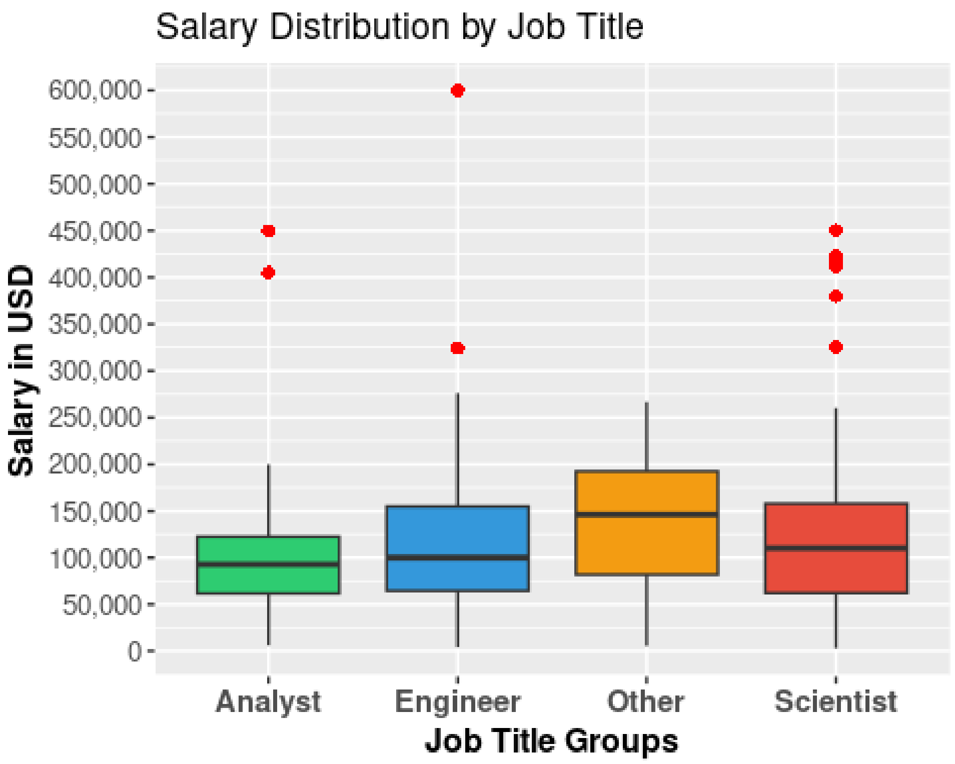


Visualizing Job's Salaries Based on the Job Title

Here I wanted to take a look at if the keywords "Analyst / Analytics" , "Science / Scientist" , "Architect" and "Engineer" seemed to have a relationship with the salary of the given position. The box plot below shows some minor differences. We see that jobs with the title of analyst tend to be more on the entry level, resulting in slightly less pay and less variation than other groups. Additionally, the term "Scientist" or "Science" in job titles appears to have a higher median and greater spread with a few outliers of high salary counts. The term "Engineer" in the job description has similar characteristics to the Scientist column with a larger variability and a few high outliers. Overall, the architect column has the highest median pay but as shown in the bar plot below it should be taken with some consideration as it has the fewest amount of data points. Lastly, the "Other" column seems to be in the middle of things, with the lowest median but some room to grow towards the maximum of the plot, some keywords in the "Other" category include "researcher" and "developer" but there were very few data points for these terms individually. Overall, the dispersions between different job titles seems very close between all groups and further analysis should be done before deciding if these are important to include in our analysis.

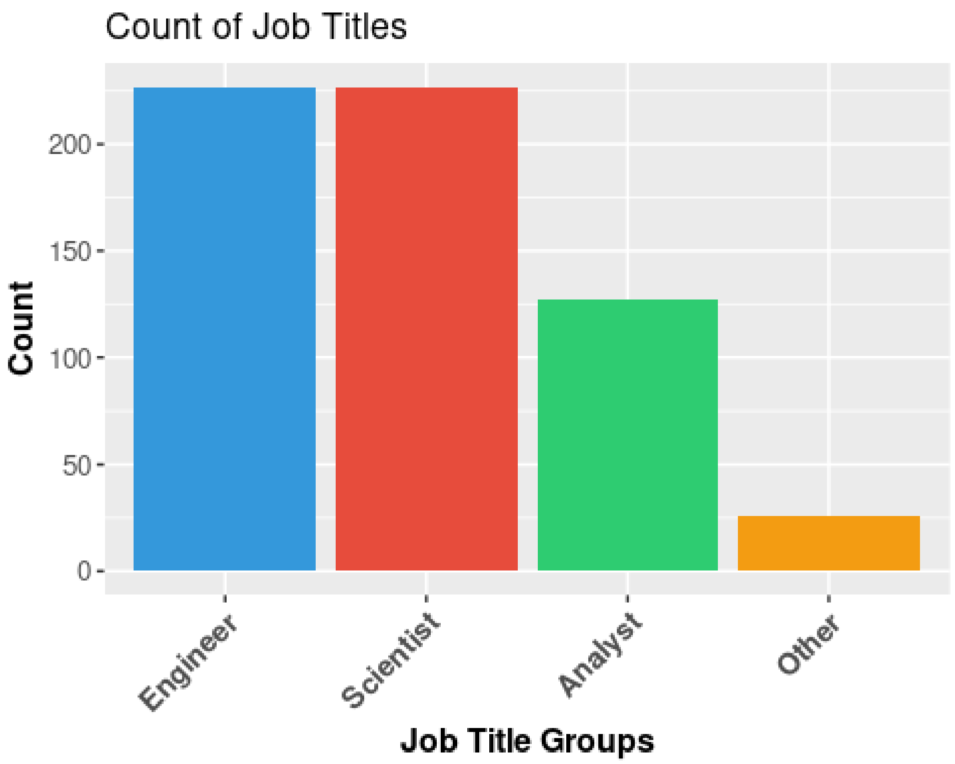
```
# By Ian Keller

# Setup Custome columns
data$job_grouping <- NA
# Use Grepl function to assign new variables based on string text
data$job_grouping[grepl("scientist", data$job_title, ignore.case = TRUE)] <-
```

```
"Scientist"
data$job_grouping[is.na(data$job_grouping) & grepl("science", data$job_title,
ignore.case = TRUE)] <- "Scientist"
data$job_grouping[is.na(data$job_grouping) & grepl("engineer",
data$job_title, ignore.case = TRUE)] <- "Engineer"
data$job_grouping[is.na(data$job_grouping) & grepl("analyst|analytics",
data$job_title, ignore.case = TRUE)] <- "Analyst"
# Other column
data$job_grouping[is.na(data$job_grouping)] <- "Other"

custom_colors <- c("Engineer" = "#3498db", "Scientist" = "#e74c3c", "Analyst"
= "#2ecc71", "Other" = "#f39c12")

# Get the sorted count of occurrences for each job grouping
sorted_counts <- sort(table(data$job_grouping), decreasing = TRUE)

ggplot(data, aes(x=job_grouping, y=salary_in_usd, fill=job_grouping)) +
  geom_boxplot(outlier.colour="red", outlier.shape=16, outlier.size=2) +
  scale_y_continuous(labels = scales::comma, breaks = scales::pretty_breaks(n
= 10)) +
  scale_fill_manual(values = custom_colors) +
  theme(
    axis.title.x = element_text(face="bold", size=12),
    axis.title.y = element_text(face="bold", size=12),
    axis.text.x = element_text(face="bold", size=11),
    axis.text.y = element_text(size=10),
    legend.position="none"
  ) +
  labs(title="Salary Distribution by Job Title", x="Job Title Groups",
y="Salary in USD")
```



Visualizing w a bar plot

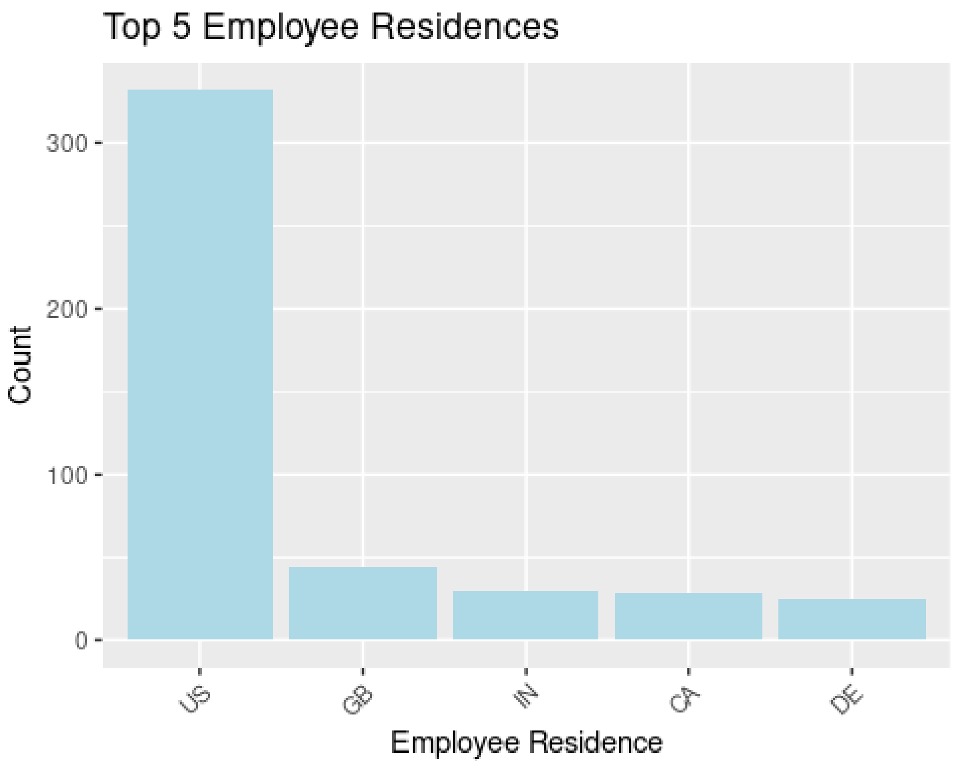
```r
# By Ian Keller
# Convert sorted_counts to a dataframe for ggplot
counts_df <- as.data.frame(sorted_counts)
colnames(counts_df) <- c("job_grouping", "count")

# Create the bar plot
ggplot(counts_df, aes(x=job_grouping, y=count, fill=job_grouping)) +
  geom_bar(stat="identity") +
  scale_fill_manual(values = custom_colors) +
  theme(
    axis.title.x = element_text(face="bold", size=12),
    axis.title.y = element_text(face="bold", size=12),
    axis.text.x = element_text(face="bold", size=11, angle=45, hjust=1),
    axis.text.y = element_text(size=10),
    legend.position="none"
  ) +
  labs(title="Count of Job Titles", x="Job Title Groups", y="Count")
```



Employee and Company Location

The dataset contains a majority of US based employees. This is important because it helps better understand other visualizations that incorporate location. Data about employees from the US may contain more outliers than those from other locations. Additionally, it means that our final analysis is primarily aimed towards US based employees and companies.

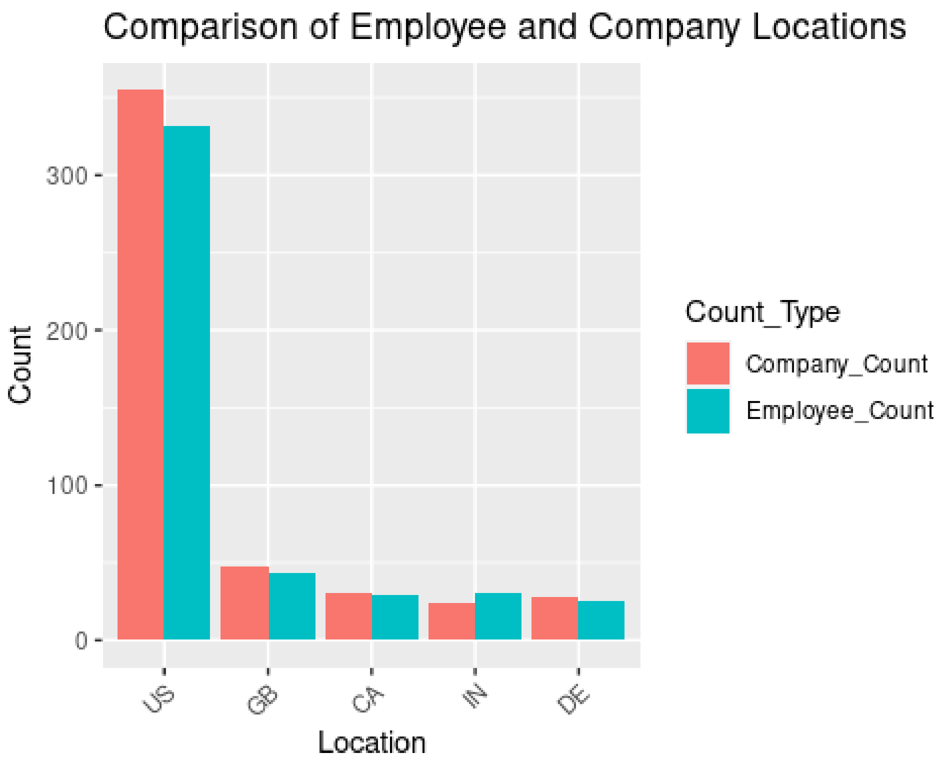
```r
# By Noah Blayney
# Count the number of employees in each location
suppressMessages(location_counts <- data %>%
  group_by(employee_residence) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  top_n(5))

# Create the bar plot
ggplot(location_counts, aes(x = reorder(employee_residence, -count), y =
count)) +
  geom_bar(stat = "identity", fill = "lightblue") +
  labs(title = "Top 5 Employee Residences", x = "Employee Residence", y =
"Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Following the previous idea, it is also worthwhile to compare the locations of the companies and the employees. Notably, there were more US companies than US employees. Meanwhile IN had more employees than companies. This shows there are some overseas workers that are likely remote.

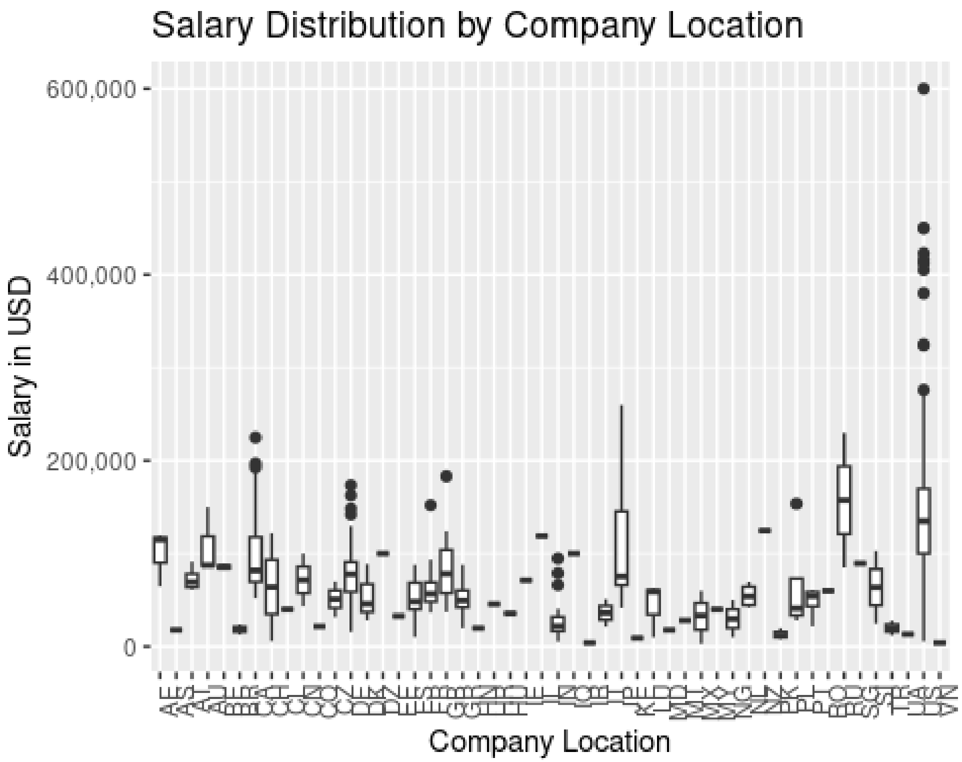
```r
# By Noah Blayney
# Count the number of employees in each location
suppressMessages(employee_location_counts <- data %>%
  group_by(employee_residence) %>%
  summarise(Employee_Count = n()) %>%
  top_n(5))

# Count the number of companies in each location
suppressMessages(company_location_counts <- data %>%
  group_by(company_location) %>%
  summarise(Company_Count = n()) %>%
  top_n(5))

# Combine the two data frames
combined_counts <- full_join(employee_location_counts,
company_location_counts, by = c("employee_residence" = "company_location"))

# Create a long format data frame for plotting
combined_counts_long <- combined_counts %>%
  pivot_longer(cols = c(Employee_Count, Company_Count), names_to =
"Count_Type", values_to = "Count")
```

```r
# Create the side-by-side bar plot
ggplot(combined_counts_long, aes(x = reorder(employee_residence,-Count), y =
Count, fill = Count_Type)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Comparison of Employee and Company Locations", x =
"Location", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Looking at the relationship between Salary in USD and Company Location provides for some interesting results. Companies located in the United States have the most amount of outliers and the largest spread overall, followed by Japan, with Canada and Russia close behind. While the United States has some extremely high outliers, the median salary in US dollars ends up being pretty close to (but less than) the median salary in US dollars in Russia. The median salary overall from the data is $101,570 (in USD). Company location definitely seems to be a factor, but it is not clear exactly how influential it is on salary in USD. Due to the United States highest outlier, the spread of all the salaries is harder to interpret because the scale is slightly skewed.

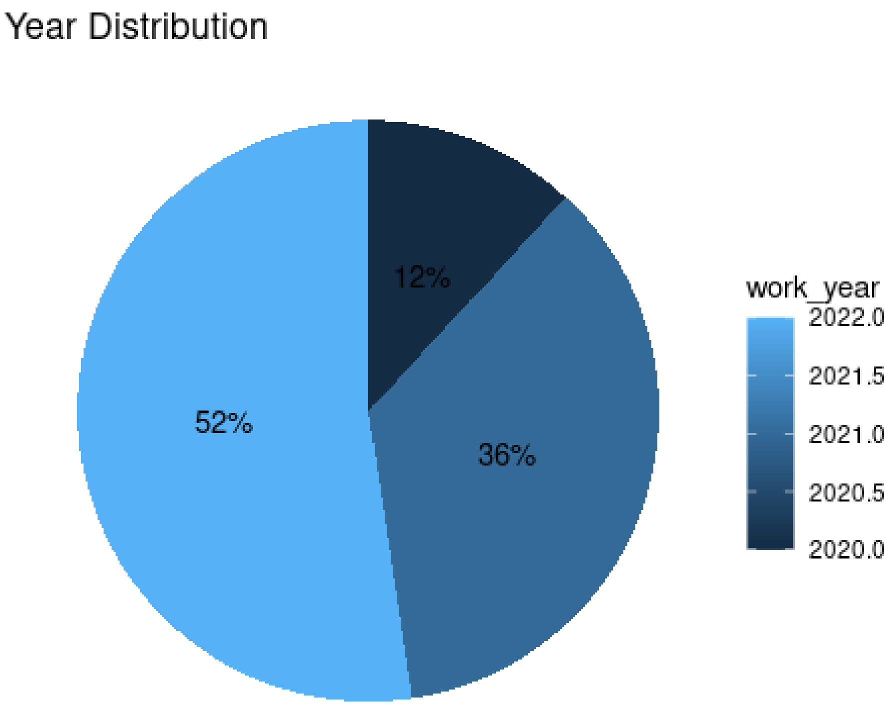
```r
# By Emma Hubbard
ggplot(data = data) +
  geom_boxplot(mapping = aes(x = company_location, y = salary_in_usd)) +
scale_y_continuous(labels = comma) + theme(axis.text.x = element_text(angle =
90, hjust = 1)) +
  labs(title = "Salary Distribution by Company Location",
       x = "Company Location",
       y = "Salary in USD")
```



Work Year

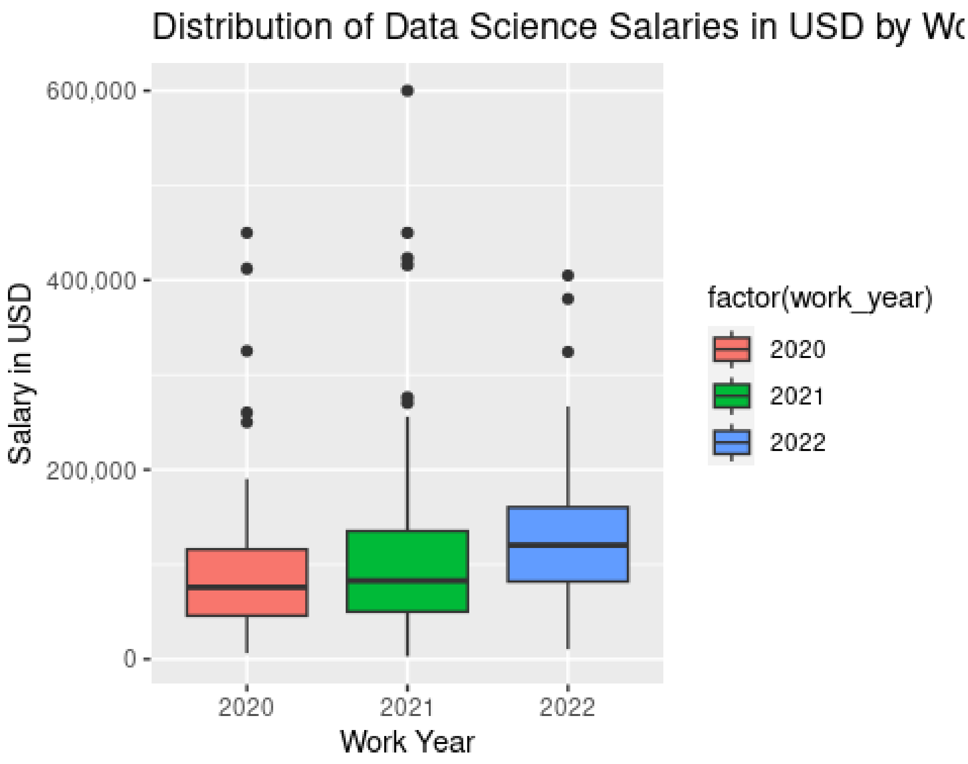
By looking at the work year we could try to see a trend over time for the salary. First, let's see how the work year is distributed throughout our dataset. From the graph below, we can see 52% of the entries are from 2022. Then 36% are from 2021 and only 12% are from 2020.

```
# By Noah Blayney
# Calculate the year distribution
year_distribution <- data %>%
  count(work_year)

# Create a pie chart with percentages
ggplot(year_distribution, aes(x = "", y = n, fill = work_year)) +
  geom_bar(stat = "identity") +
  coord_polar(theta = "y") +
  geom_text(aes(label = scales::percent(n / sum(n))), position =
position_stack(vjust = 0.5)) +
  labs(title = "Year Distribution") +
  theme_void()
```



Now looking at salaries by work year, there seems to be no difference between 2020 and 2021. However, there is a slight increase in median salary for 2022, but since all 3 boxes overlap, it is unexpected that there will be significance for the work year.

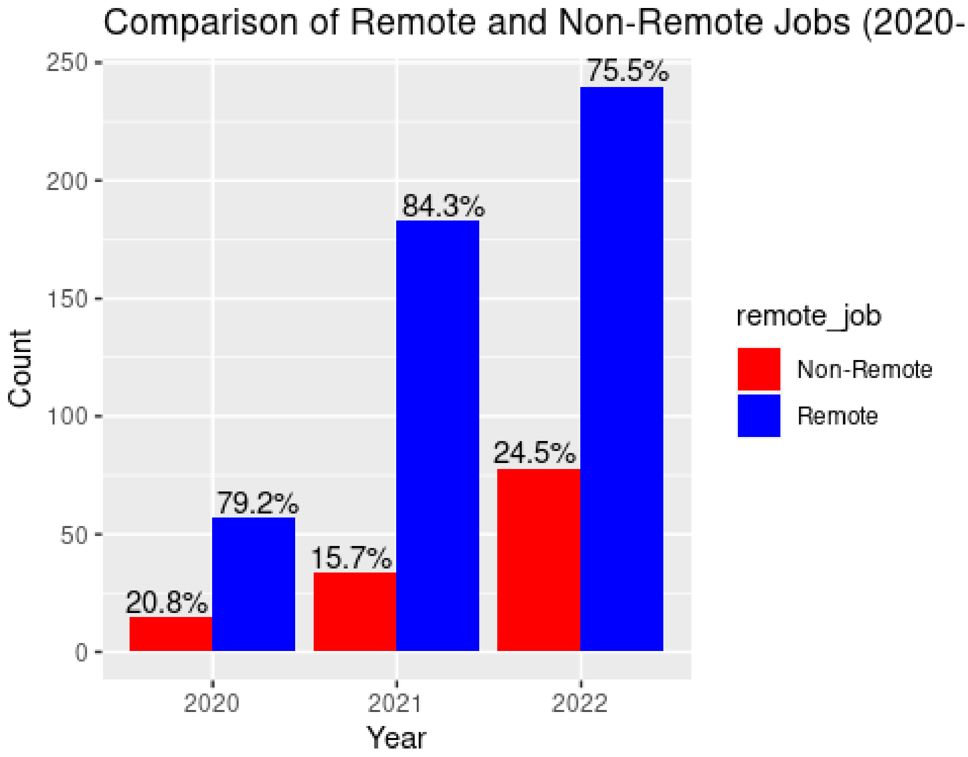
```r
# By Philip Kim
data %>%
  ggplot(aes(x = factor(work_year), y = salary_in_usd, fill =
factor(work_year))) +
  geom_boxplot() +
  scale_y_continuous(labels = comma) +
  labs(title = 'Distribution of Data Science Salaries in USD by Work Year',
       x = 'Work Year', y = 'Salary in USD')
```



Remote and Remote Ratio

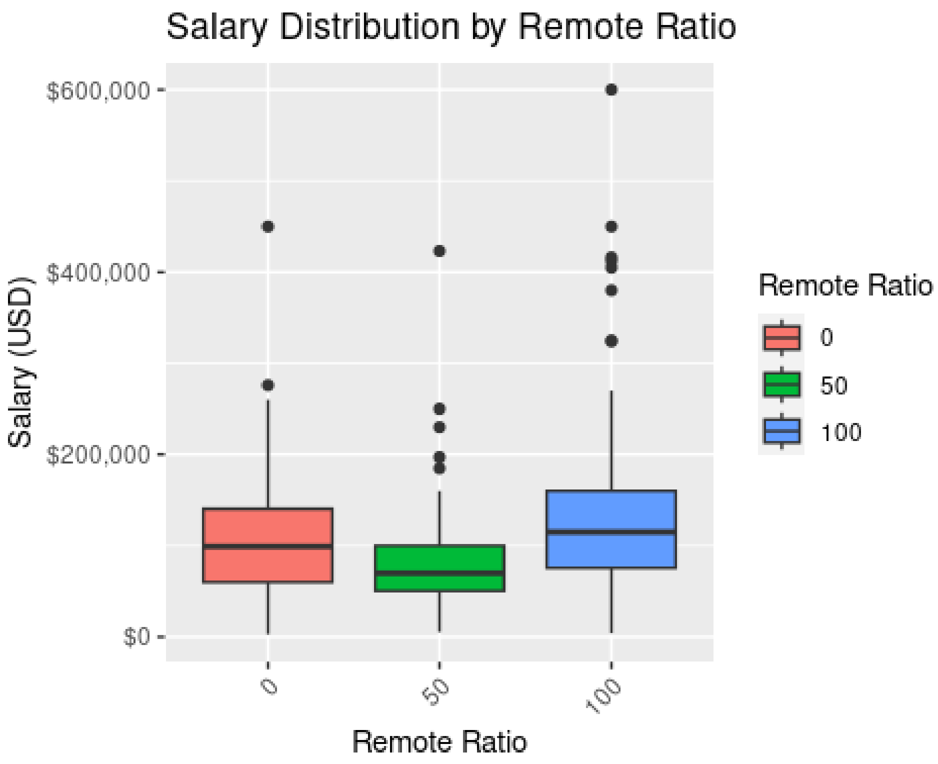
By comparing remote and non-remote jobs over (2020-2022) we can see remote jobs were most popular in 2021. We can also see that a majority of the jobs are remote (remote ratio > 0).

```
# By Noah Blayney
# Count the number of remote and non-remote jobs for each year
yearly_counts <- data %>%
  group_by(work_year, remote_job = ifelse(remote_ratio > 0, "Remote", "Non-
Remote")) %>%
  summarise(Count = n())

## `summarise()` has grouped output by 'work_year'. You can override using
the
## `.groups` argument.

# Calculate percentages for each year
yearly_counts <- yearly_counts %>%
  group_by(work_year) %>%
  mutate(Percentage = (Count / sum(Count)) * 100)

# Create a bar plot with percentages slightly above the bars
ggplot(yearly_counts, aes(x = factor(work_year), y = Count, fill =
remote_job)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = scales::percent(Percentage / 100)), position =
position_dodge(width = 1), vjust = -0.2) +
  labs(title = "Comparison of Remote and Non-Remote Jobs (2020-2022)", x =
"Year", y = "Count") +
  scale_fill_manual(values = c("Remote" = "blue", "Non-Remote" = "red")) +
  theme(axis.text.x = element_text(angle = 0))
```



To observe the discrepencies in salary among remote conditions. We can create a boxplot to show the distribution of salaries depending on the remote ratio. In this case, 0 means that it was not remote, 50 means that it is hybrid, and 100 means the jobs was fully remote. The visualization finds that 0 and 100 are similar in distribution with fully remote jobs having a slightly higher salary on average. The distribution of hybrid is far more narrow than the other two. Showing that hybrid jobs tend to have lower salaries compared to remote and non-remote jobs.

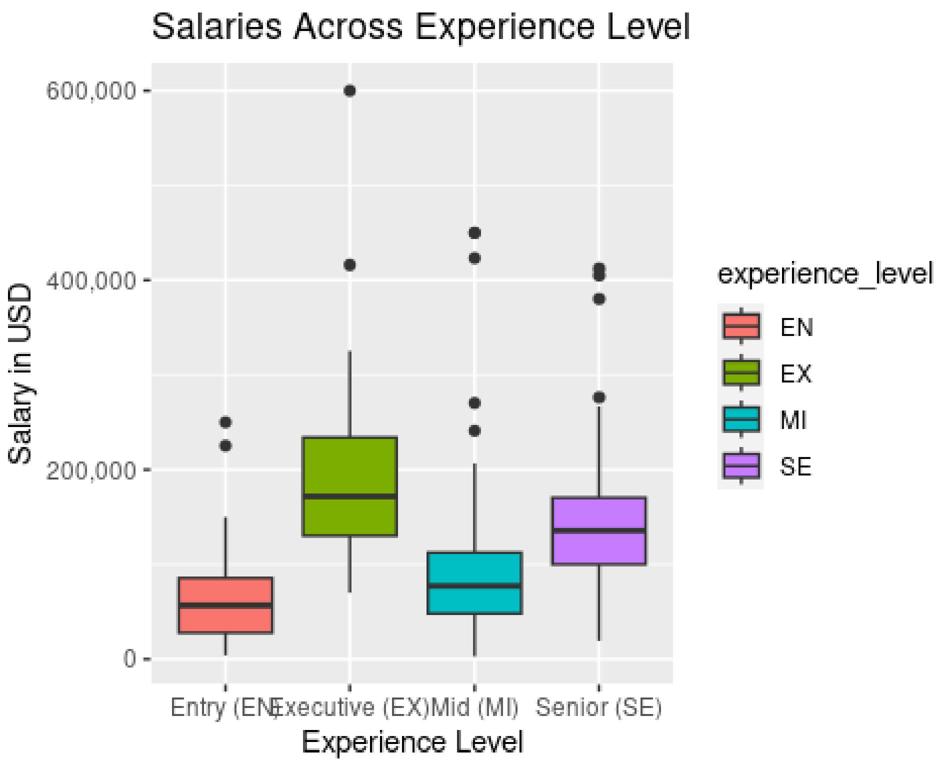
```
# By Noah Blayney
# Create a box plot to compare salary distribution by remote ratio
ggplot(data, aes(x = as.factor(remote_ratio), y = salary_in_usd, fill =
as.factor(remote_ratio))) +
  geom_boxplot() +
  labs(title = "Salary Distribution by Remote Ratio", x = "Remote Ratio", y =
"Salary (USD)") +
  scale_fill_discrete(name = "Remote Ratio") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(labels = scales::dollar_format(scale = 1, prefix = "$"))
```



Experience Level

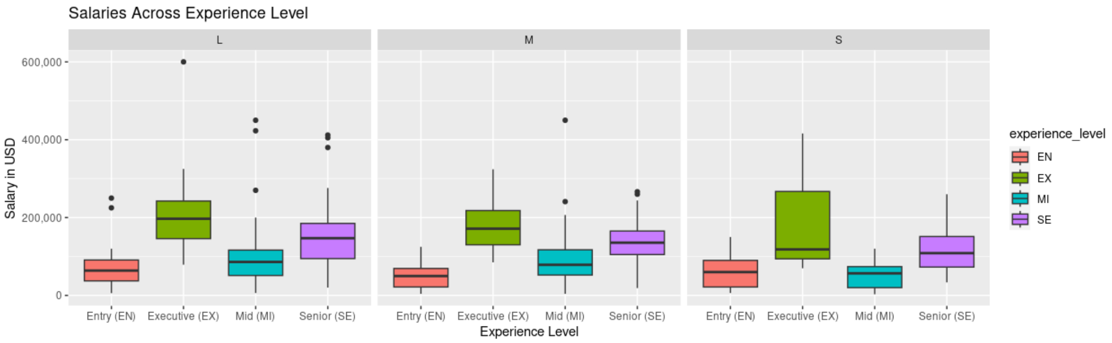
Taking a look at salaries across experience level, we can see that executives tend to make the most annually, followed by senior levels, mid levels, and then finally entry level data scientists. It seems as if the experience level is crucial in determining one's salary. Moreover, all experience levels have at least a couple of high outliers. For example, the median entry level salary falls well under $100,000, but there are a few data scientists making over $200,000. This could be explained by other variables, such as company size.

```
# By Philip Kim
# overall boxplot
data %>%
  ggplot(aes(x = experience_level, y = salary_in_usd, fill =
experience_level)) +
  geom_boxplot() +
  scale_x_discrete(labels = c('Entry (EN)', 'Executive (EX)',
                              'Mid (MI)', 'Senior (SE)')) +
  scale_y_continuous(labels = comma) +
  labs(title = 'Salaries Across Experience Level', x = "Experience Level", y
= "Salary in USD")
```



Now that the boxplot is broken out by both experience level and company size, we can see that most of the outliers belong to large firms. This is logical, as larger firms typically have higher prestige and collect more profits, and therefore they have the ability to better compensate their data scientists.

```
# By Philip Kim
# faceted boxplot
data %>%
  ggplot(aes(x = experience_level, y = salary_in_usd, fill = 
experience_level)) +
  geom_boxplot() +
  scale_x_discrete(labels = c('Entry (EN)', 'Executive (EX)',
                              'Mid (MI)', 'Senior (SE)')) +
  scale_y_continuous(labels = comma) +
  labs(title = 'Salaries Across Experience Level', x = "Experience Level", y 
= "Salary in USD") +
  facet_wrap(~ company_size)
```

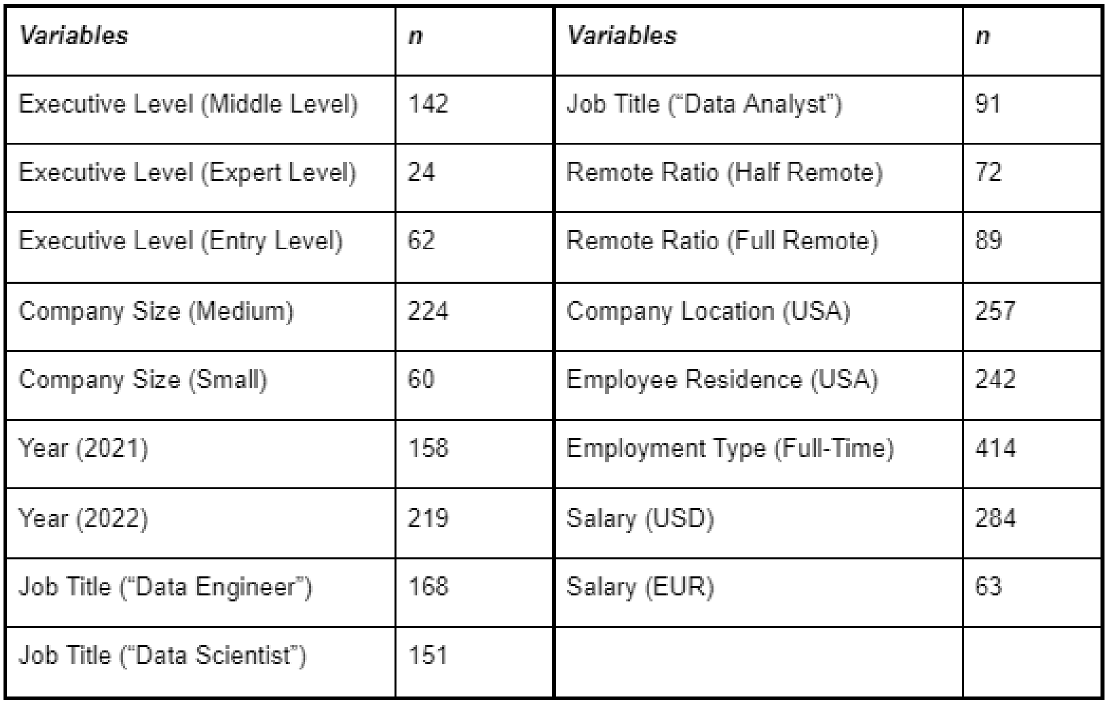


Conclusion

Through this report we have explored key determinants and features of the Data Science Salaries dataset. We aim to use our findings to strengthen our future predictive model. By doing so we hope to estimate data science salaries and provide insights into the data science field. In the report, we identified Experience Level, Employee Location, Company Location, and possibly Job Title to be helpful indicators in preparing our model.

Modeling

In our analysis we decided to implement a multiple linear regression model represented by the equation Y = 0+ 1X1 + 2X2 +…+pXp + . The equation represents a linear relationship between the dependent variable Y and the independent variables X1, X2, …, Xp, where 0 is the intercept, or baseline value of Y, and 1, 2,…, p are the coefficients quantifying the impact of each X variable on Y, and accounts for the unexplained variability or errors. We began our modeling process by fitting the full model to a linear regression with salaries in USD being our dependent variable.The full model consisted of the following independent variables:



Variables	*n*	*Variables*	*n*
Executive Level (Middle Level)	142	Job Title ("Data Analyst")	91
Executive Level (Expert Level)	24	Remote Ratio (Half Remote)	72
Executive Level (Entry Level)	62	Remote Ratio (Full Remote)	89
Company Size (Medium)	224	Company Location (USA)	257
Company Size (Small)	60	Employee Residence (USA)	242
Year (2021)	158	Employment Type (Full-Time)	414
Year (2022)	219	Salary (USD)	284
Job Title ("Data Engineer")	168	Salary (EUR)	63
Job Title ("Data Scientist")	151		

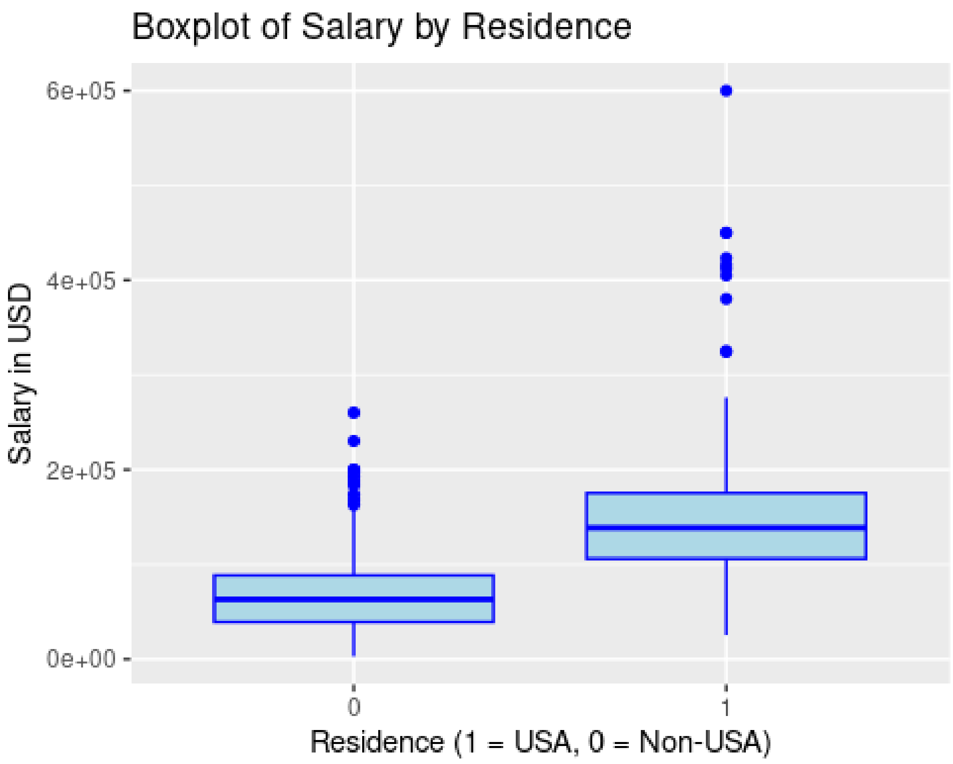
Creating Dummy Variables for Regression

```
#Creating new df to work with regression:
model_data <- data

#Dummy variables for Experience Level, with "SE" being the baseline category:
model_data$EL_MI <- ifelse(data$experience_level == "MI", 1,0)
model_data$EL_EX <- ifelse(data$experience_level == "EX", 1,0)
model_data$EL_EN <- ifelse(data$experience_level == "EN", 1,0)

#Dummy variables for company size, with "L" being the baseline category:
model_data$size_m <- ifelse(data$company_size == "M", 1,0)
model_data$size_s <- ifelse(data$company_size == "S", 1,0)

#Dummy variables for work year, with 2020 being the baseline category:
model_data$year_21 <- ifelse(data$work_year == 2021, 1, 0)
model_data$year_22 <- ifelse(data$work_year == 2022, 1, 0)

#Dummy variables for job title, with "Other" being the baseline category:
model_data$title_engineer <- ifelse(data$job_grouping == "Engineer", 1,0)
model_data$title_scientist <- ifelse(data$job_grouping == "Scientist", 1,0)
model_data$title_analyst <- ifelse(data$job_grouping == "Analyst", 1,0)

#Dummy variables for remote ratio, with 100 being the baseline category:
model_data$remote_half <- ifelse(data$remote_ratio == 50, 1,0)
model_data$remote_full <- ifelse(data$remote_ratio == 0, 1,0)

#Dummy variable for company location, 1 if US, 0 if outside of the US:
```

```r
model_data$location_USA <- ifelse(data$company_location == "US", 1,0)

#Dummy variable for employee residence, 1 if US, 0 if otherwise:
model_data$residence_USA <- ifelse(data$employee_residence == "US", 1,0)

#Dummy variable for employment type, 1 if FT, 0 if otherwise:
model_data$employee_FT <- ifelse(data$employment_type == "FT", 1,0)

#Dummy variables for salary currency, looking at US, EUR, and other being the
baseline
model_data$salary_USD <- ifelse(data$salary_currency == "USD", 1,0)
model_data$salary_EUR <- ifelse(data$salary_currency == "EUR", 1,0)
```

Boxplot of Residency - Additionally eda after pt 1

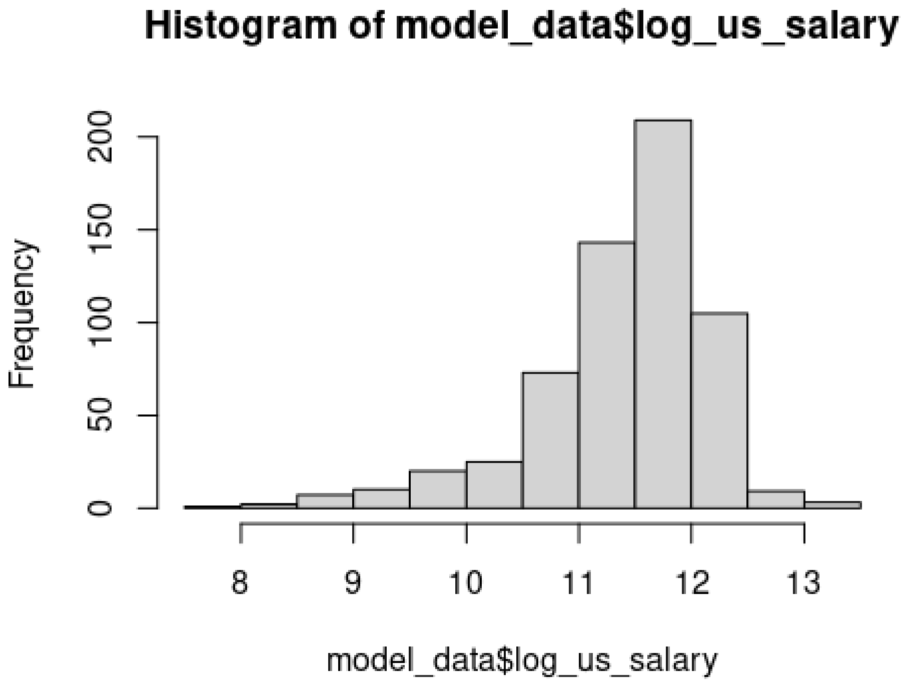
```r
# Calculate mean and median for each group
mean_usa <- mean(model_data$salary_in_usd[model_data$residence_USA == 1])
median_usa <- median(model_data$salary_in_usd[model_data$residence_USA == 1])

mean_non_usa <- mean(model_data$salary_in_usd[model_data$residence_USA == 0])
median_non_usa <- median(model_data$salary_in_usd[model_data$residence_USA ==
0])
library(ggplot2)
ggplot(data = model_data, aes(x = factor(residence_USA), y = salary_in_usd))
+
  geom_boxplot(fill = "lightblue", color = "blue") +
  labs(x = "Residence (1 = USA, 0 = Non-USA)", y = "Salary in USD",
       title = "Boxplot of Salary by Residence")
```



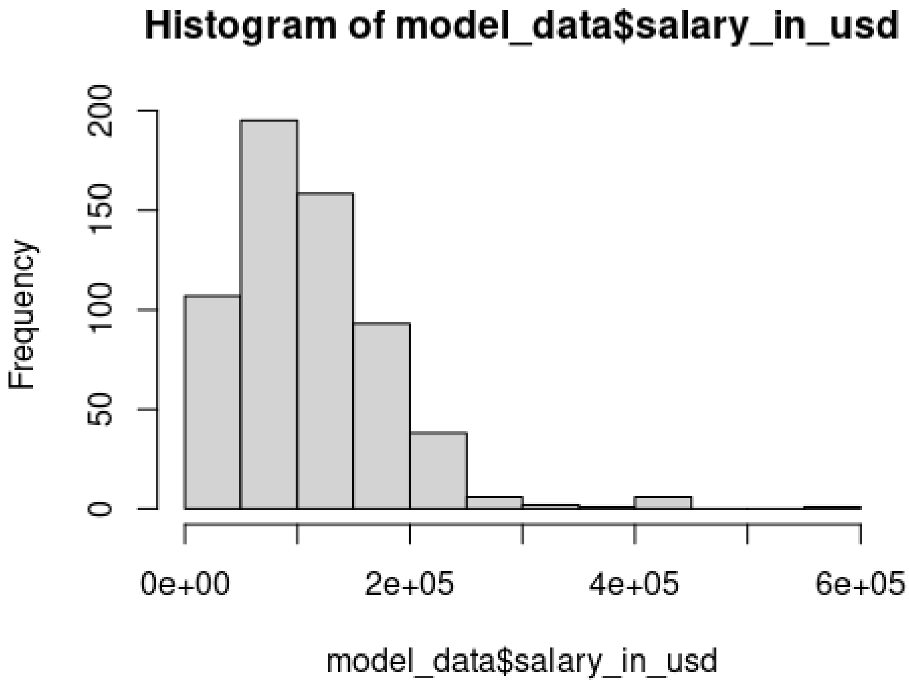
```
# Print mean and median for each group
cat("Mean Salary in USA:", mean_usa, "\n")

## Mean Salary in USA: 149194.1

cat("Median Salary in USA:", median_usa, "\n")

## Median Salary in USA: 138475

cat("Mean Salary in Non-USA:", mean_non_usa, "\n")

## Mean Salary in Non-USA: 67754.04

cat("Median Salary in Non-USA:", median_non_usa, "\n")

## Median Salary in Non-USA: 62649
```

Performing a log transformation on Salary to fix normality

```
model_data$log_us_salary <- log(model_data$salary_in_usd)
hist(model_data$log_us_salary)
```

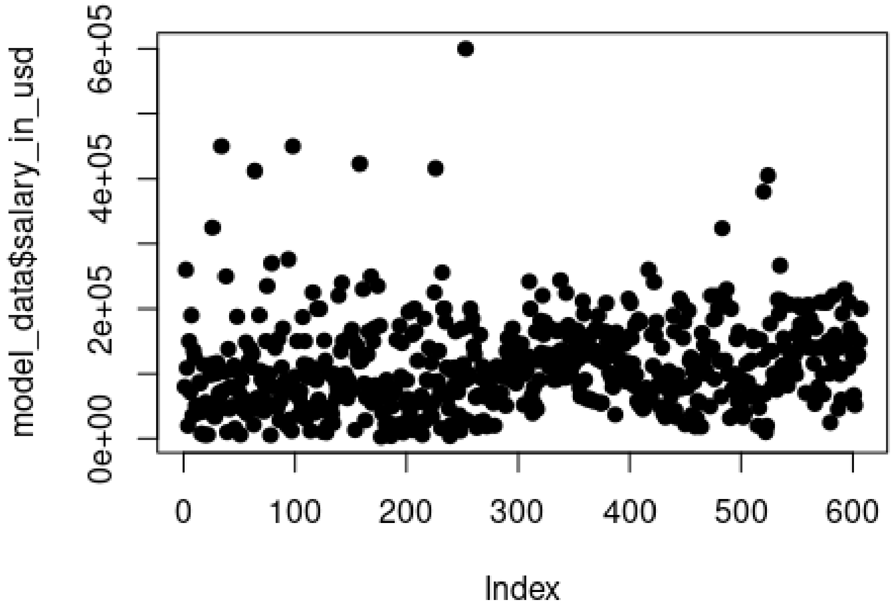


```
hist(model_data$salary_in_usd)
```



```
shapiro.test(model_data$log_us_salary)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model_data$log_us_salary
## W = 0.91931, p-value < 2.2e-16

shapiro.test(model_data$salary_in_usd)

##
##  Shapiro-Wilk normality test
##
## data:  model_data$salary_in_usd
## W = 0.89836, p-value < 2.2e-16

plot(model_data$salary_in_usd, pch =19)
```



This log transformation was not effective increasing our test stat marginally but not increasing significance therefore we will continue to use the non transformed salary to keep our results easily interpretable.

Next, we will clean the dataset up to only include variables of interest.

```
#Subsetting data for only engineered features
model_data_final <- subset(model_data, select = -c(work_year,
employment_type, X, job_title, salary, salary_currency, employee_residence,
remote_ratio, company_location, company_size, job_grouping, experience_level,
log_us_salary))
head(model_data_final)
```

```
##   salary_in_usd EL_MI EL_EX EL_EN size_m size_s year_21 year_22
title_engineer
## 1         79833     1     0     0      0      0       0       0
0
## 2        260000     0     0     0      0      1       0       0
0
## 3        109024     0     0     0      1      0       0       0
1
## 4         20000     1     0     0      0      1       0       0
0
## 5        150000     0     0     0      0      0       0       0
1
## 6         72000     0     0     1      0      0       0       0
0
##   title_scientist title_analyst remote_half remote_full location_USA
## 1               1             0           0           1            0
## 2               1             0           0           1            0
## 3               0             0           1           0            0
## 4               0             1           0           1            0
## 5               0             0           1           0            1
## 6               0             1           0           0            1
##   residence_USA employee_FT salary_USD salary_EUR
## 1             0           1          0          1
## 2             0           1          1          0
## 3             0           1          0          0
## 4             0           1          1          0
## 5             1           1          1          0
## 6             1           1          1          0
```

Removing Outliers from the dataset

```
#z_scores <- scale(model_data_final$salary_in_usd)
#model_data_final$salary_in_usd[abs(z_scores) > 3] # Typically, a Z-score
above 3 is considered an outlier.

# Filter out rows with Z-scores within the range of -3 to 3
#filtered_model_data_final <- model_data_final[abs(z_scores) <= 3, ]
```

Splitting the data into a training and testing dataset.

```
library(caret)

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```
set.seed(123)

# Create a train/test split of 70%/30%
split <- createDataPartition(model_data_final$salary_in_usd, p = 0.7, list =
FALSE)
train_set <- model_data_final[split, ]
test_set <- model_data_final[-split, ]
```

After fitting the full model to the training set, we began to use stepwise model selection, seeking to minimize the model's Akaike Information Criteria (AIC). AIC is represented by the equation: AIC = n ln(SSE) -n ln(n) +2p, where n represents the sample size, p represents the number of parameters in the model, including the intercept, and SSE which is our sum of squares error, or unexplained variability. This process removed and added parameters to reduce our prediction error, taking into consideration the simplicity of the model. The final model is reported below:

Using Stepwise Model Selection to find the best model on the training data

```
library(MASS)

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

mod1 <- lm(salary_in_usd ~ ., data=train_set)

final <- stepAIC(mod1,direction = 'both')

## Start:  AIC=9297.68
## salary_in_usd ~ EL_MI + EL_EX + EL_EN + size_m + size_s + year_21 +
##     year_22 + title_engineer + title_scientist + title_analyst +
##     remote_half + remote_full + location_USA + residence_USA +
##     employee_FT + salary_USD + salary_EUR
##
##                   Df  Sum of Sq        RSS    AIC
## - salary_EUR       1 1.6749e+08 1.1230e+12 9295.7
## - remote_half      1 5.6259e+08 1.1234e+12 9295.9
## - title_scientist  1 7.3463e+08 1.1236e+12 9296.0
## - location_USA     1 1.0292e+09 1.1239e+12 9296.1
## - employee_FT      1 1.3515e+09 1.1242e+12 9296.2
## - title_engineer   1 2.7523e+09 1.1256e+12 9296.7
## - year_21          1 2.9567e+09 1.1258e+12 9296.8
## - size_m           1 3.1862e+09 1.1261e+12 9296.9
## <none>                          1.1229e+12 9297.7
## - year_22          1 8.9143e+09 1.1318e+12 9299.1
## - remote_full      1 1.0612e+10 1.1335e+12 9299.7
## - size_s           1 1.1571e+10 1.1344e+12 9300.1
## - salary_USD       1 1.2445e+10 1.1353e+12 9300.4
## - EL_MI             1 1.9908e+10 1.1428e+12 9303.2
## - title_analyst     1 2.4056e+10 1.1469e+12 9304.7
## - residence_USA     1 5.9276e+10 1.1821e+12 9317.7
## - EL_EN             1 7.9798e+10 1.2027e+12 9325.0
## - EL_EX             1 1.1012e+11 1.2330e+12 9335.6
##
## Step:  AIC=9295.75
## salary_in_usd ~ EL_MI + EL_EX + EL_EN + size_m + size_s + year_21 +
##     year_22 + title_engineer + title_scientist + title_analyst +
##     remote_half + remote_full + location_USA + residence_USA +
##     employee_FT + salary_USD
##
##                   Df  Sum of Sq        RSS    AIC
## - remote_half      1 5.6832e+08 1.1236e+12 9294.0
## - title_scientist  1 7.4125e+08 1.1238e+12 9294.0
## - location_USA     1 1.0362e+09 1.1241e+12 9294.1
## - employee_FT      1 1.2936e+09 1.1243e+12 9294.2
## - title_engineer   1 2.7610e+09 1.1258e+12 9294.8
## - year_21          1 2.8532e+09 1.1259e+12 9294.8
## - size_m           1 3.2686e+09 1.1263e+12 9295.0
## <none>                          1.1230e+12 9295.7
## - year_22          1 8.7468e+09 1.1318e+12 9297.1
## + salary_EUR       1 1.6749e+08 1.1229e+12 9297.7
## - remote_full      1 1.0452e+10 1.1335e+12 9297.7
## - size_s           1 1.1499e+10 1.1345e+12 9298.1
## - salary_USD       1 1.5630e+10 1.1387e+12 9299.7
## - EL_MI            1 1.9804e+10 1.1428e+12 9301.2
## - title_analyst    1 2.4111e+10 1.1472e+12 9302.8
## - residence_USA    1 5.9358e+10 1.1824e+12 9315.7
## - EL_EN            1 7.9633e+10 1.2027e+12 9323.0
## - EL_EX            1 1.0997e+11 1.2330e+12 9333.6
##
## Step:  AIC=9293.96
## salary_in_usd ~ EL_MI + EL_EX + EL_EN + size_m + size_s + year_21 +
##     year_22 + title_engineer + title_scientist + title_analyst +
##     remote_full + location_USA + residence_USA + employee_FT +
##     salary_USD
##
##                   Df  Sum of Sq        RSS    AIC
## - title_scientist  1 6.9996e+08 1.1243e+12 9292.2
## - location_USA     1 9.5919e+08 1.1246e+12 9292.3
## - employee_FT      1 1.2727e+09 1.1249e+12 9292.4
## - title_engineer   1 2.5970e+09 1.1262e+12 9293.0
## - year_21          1 2.8517e+09 1.1265e+12 9293.0
## - size_m           1 2.9683e+09 1.1266e+12 9293.1
## <none>                          1.1236e+12 9294.0
## - year_22          1 8.2998e+09 1.1319e+12 9295.1
## - remote_full      1 9.8936e+09 1.1335e+12 9295.7
## + remote_half      1 5.6832e+08 1.1230e+12 9295.7
## + salary_EUR       1 1.7321e+08 1.1234e+12 9295.9
## - size_s            1 1.0992e+10 1.1346e+12 9296.1
## - salary_USD        1 1.6434e+10 1.1400e+12 9298.2
## - EL_MI             1 1.9481e+10 1.1431e+12 9299.3
## - title_analyst     1 2.3793e+10 1.1474e+12 9300.9
## - residence_USA     1 5.9673e+10 1.1833e+12 9314.1
## - EL_EN             1 8.0389e+10 1.2040e+12 9321.5
## - EL_EX             1 1.1066e+11 1.2343e+12 9332.1
## 
## Step:  AIC=9292.23
## salary_in_usd ~ EL_MI + EL_EX + EL_EN + size_m + size_s + year_21 + 
##     year_22 + title_engineer + title_analyst + remote_full + 
##     location_USA + residence_USA + employee_FT + salary_USD
## 
##                    Df  Sum of Sq        RSS    AIC
## - location_USA      1 1.0372e+09 1.1253e+12 9290.6
## - employee_FT       1 1.2354e+09 1.1255e+12 9290.7
## - year_21           1 2.7730e+09 1.1271e+12 9291.3
## - size_m            1 2.9893e+09 1.1273e+12 9291.4
## - title_engineer    1 4.0027e+09 1.1283e+12 9291.7
## <none>                          1.1243e+12 9292.2
## - year_22           1 8.1162e+09 1.1324e+12 9293.3
## + title_scientist   1 6.9996e+08 1.1236e+12 9294.0
## + remote_half       1 5.2703e+08 1.1238e+12 9294.0
## - remote_full       1 1.0201e+10 1.1345e+12 9294.1
## + salary_EUR        1 1.7954e+08 1.1241e+12 9294.2
## - size_s            1 1.1052e+10 1.1354e+12 9294.4
## - salary_USD        1 1.6637e+10 1.1409e+12 9296.5
## - EL_MI             1 1.9689e+10 1.1440e+12 9297.6
## - residence_USA     1 5.9730e+10 1.1840e+12 9312.3
## - title_analyst     1 6.8743e+10 1.1931e+12 9315.6
## - EL_EN             1 8.1500e+10 1.2058e+12 9320.1
## - EL_EX             1 1.1256e+11 1.2369e+12 9331.0
## 
## Step:  AIC=9290.62
## salary_in_usd ~ EL_MI + EL_EX + EL_EN + size_m + size_s + year_21 + 
##     year_22 + title_engineer + title_analyst + remote_full + 
##     residence_USA + employee_FT + salary_USD
## 
##                    Df  Sum of Sq        RSS    AIC
## - employee_FT       1 9.3939e+08 1.1263e+12 9289.0
## - year_21           1 2.6984e+09 1.1280e+12 9289.6
## - size_m            1 2.8461e+09 1.1282e+12 9289.7
## - title_engineer    1 4.2012e+09 1.1295e+12 9290.2
## <none>                          1.1253e+12 9290.6
## - year_22           1 7.8638e+09 1.1332e+12 9291.6
## + location_USA      1 1.0372e+09 1.1243e+12 9292.2
## + title_scientist   1 7.7793e+08 1.1246e+12 9292.3
## - remote_full       1 9.9482e+09 1.1353e+12 9292.4
## + remote_half       1 4.4809e+08 1.1249e+12 9292.5
## + salary_EUR        1 1.8672e+08 1.1252e+12 9292.6
## - size_s           1 1.1268e+10 1.1366e+12 9292.9
## - salary_USD       1 1.5674e+10 1.1410e+12 9294.5
## - EL_MI            1 1.9688e+10 1.1450e+12 9296.0
## - title_analyst    1 6.8892e+10 1.1942e+12 9314.0
## - EL_EN            1 8.0730e+10 1.2061e+12 9318.2
## - EL_EX            1 1.1310e+11 1.2384e+12 9329.5
## - residence_USA    1 1.1521e+11 1.2406e+12 9330.2
##
## Step:  AIC=9288.98
## salary_in_usd ~ EL_MI + EL_EX + EL_EN + size_m + size_s + year_21 +
##     year_22 + title_engineer + title_analyst + remote_full +
##     residence_USA + salary_USD
##
##                   Df  Sum of Sq        RSS    AIC
## - size_m           1 2.7579e+09 1.1290e+12 9288.0
## - year_21          1 2.8054e+09 1.1291e+12 9288.0
## - title_engineer   1 4.1335e+09 1.1304e+12 9288.5
## <none>                          1.1263e+12 9289.0
## - year_22          1 8.3797e+09 1.1347e+12 9290.1
## + employee_FT      1 9.3939e+08 1.1253e+12 9290.6
## + location_USA     1 7.4120e+08 1.1255e+12 9290.7
## + title_scientist  1 7.3138e+08 1.1256e+12 9290.7
## - remote_full      1 1.0150e+10 1.1364e+12 9290.8
## + remote_half      1 4.4356e+08 1.1258e+12 9290.8
## + salary_EUR       1 1.3167e+08 1.1262e+12 9290.9
## - size_s           1 1.0809e+10 1.1371e+12 9291.1
## - salary_USD       1 1.6244e+10 1.1425e+12 9293.1
## - EL_MI            1 1.9608e+10 1.1459e+12 9294.4
## - title_analyst    1 6.8768e+10 1.1951e+12 9312.3
## - EL_EN            1 7.9836e+10 1.2061e+12 9316.2
## - EL_EX            1 1.1385e+11 1.2401e+12 9328.1
## - residence_USA    1 1.1443e+11 1.2407e+12 9328.3
##
## Step:  AIC=9288.02
## salary_in_usd ~ EL_MI + EL_EX + EL_EN + size_s + year_21 + year_22 +
##     title_engineer + title_analyst + remote_full + residence_USA +
##     salary_USD
##
##                   Df  Sum of Sq        RSS    AIC
## - year_21          1 2.7407e+09 1.1318e+12 9287.1
## - title_engineer   1 4.7387e+09 1.1338e+12 9287.8
## <none>                          1.1290e+12 9288.0
## + size_m           1 2.7579e+09 1.1263e+12 9289.0
## - size_s           1 8.4646e+09 1.1375e+12 9289.2
## + employee_FT      1 8.5126e+08 1.1282e+12 9289.7
## + title_scientist  1 7.4862e+08 1.1283e+12 9289.7
## + location_USA     1 6.3544e+08 1.1284e+12 9289.8
## + salary_EUR       1 2.0023e+08 1.1288e+12 9289.9
## + remote_half      1 1.9604e+08 1.1288e+12 9290.0
## - remote_full      1 1.1373e+10 1.1404e+12 9290.3
## - year_22          1 1.3875e+10 1.1429e+12 9291.2
## - salary_USD       1 1.6278e+10 1.1453e+12 9292.1
## - EL_MI            1 1.8274e+10 1.1473e+12 9292.9
## - title_analyst    1 7.3350e+10 1.2024e+12 9312.9
## - EL_EN            1 7.9268e+10 1.2083e+12 9315.0
## - EL_EX            1 1.1531e+11 1.2444e+12 9327.5
## - residence_USA    1 1.1551e+11 1.2445e+12 9327.6
## 
## Step:  AIC=9287.06
## salary_in_usd ~ EL_MI + EL_EX + EL_EN + size_s + year_22 + title_engineer 
+ 
##     title_analyst + remote_full + residence_USA + salary_USD
## 
##                   Df  Sum of Sq        RSS    AIC
## - title_engineer   1 4.5815e+09 1.1364e+12 9286.8
## <none>                          1.1318e+12 9287.1
## - size_s           1 7.4466e+09 1.1392e+12 9287.9
## + year_21          1 2.7407e+09 1.1290e+12 9288.0
## + size_m           1 2.6933e+09 1.1291e+12 9288.0
## + employee_FT      1 9.5300e+08 1.1308e+12 9288.7
## + title_scientist  1 6.6223e+08 1.1311e+12 9288.8
## + location_USA     1 5.6647e+08 1.1312e+12 9288.8
## + remote_half      1 2.0049e+08 1.1316e+12 9289.0
## + salary_EUR       1 8.3457e+07 1.1317e+12 9289.0
## - remote_full      1 1.1460e+10 1.1432e+12 9289.4
## - year_22          1 1.4716e+10 1.1465e+12 9290.6
## - salary_USD       1 1.5778e+10 1.1476e+12 9291.0
## - EL_MI            1 1.8436e+10 1.1502e+12 9292.0
## - title_analyst    1 7.3383e+10 1.2052e+12 9311.9
## - EL_EN            1 7.8616e+10 1.2104e+12 9313.7
## - EL_EX            1 1.1428e+11 1.2461e+12 9326.1
## - residence_USA    1 1.1577e+11 1.2476e+12 9326.6
## 
## Step:  AIC=9286.78
## salary_in_usd ~ EL_MI + EL_EX + EL_EN + size_s + year_22 + title_analyst + 
##     remote_full + residence_USA + salary_USD
## 
##                   Df  Sum of Sq        RSS    AIC
## <none>                          1.1364e+12 9286.8
## + title_engineer   1 4.5815e+09 1.1318e+12 9287.1
## + size_m           1 3.2832e+09 1.1331e+12 9287.5
## - size_s           1 7.6945e+09 1.1441e+12 9287.7
## + year_21          1 2.5835e+09 1.1338e+12 9287.8
## + title_scientist  1 2.5217e+09 1.1338e+12 9287.8
## + employee_FT      1 8.6885e+08 1.1355e+12 9288.5
## + location_USA     1 7.2190e+08 1.1356e+12 9288.5
## + salary_EUR       1 8.6165e+07 1.1363e+12 9288.8
## + remote_half      1 4.2630e+07 1.1363e+12 9288.8
## - remote_full      1 1.1253e+10 1.1476e+12 9289.0
## - salary_USD       1 1.5431e+10 1.1518e+12 9290.5
## - year_22          1 1.5953e+10 1.1523e+12 9290.7
## - EL_MI            1 1.8334e+10 1.1547e+12 9291.6
## - title_analyst    1 7.1446e+10 1.2078e+12 9310.8
## - EL_EN            1 7.8871e+10 1.2152e+12 9313.4
## - residence_USA    1 1.1647e+11 1.2528e+12 9326.4
## - EL_EX            1 1.1719e+11 1.2536e+12 9326.7
```

Summary Statistics of the Final (Best) Model

```
summary(final)

## 
## Call:
## lm(formula = salary_in_usd ~ EL_MI + EL_EX + EL_EN + size_s + 
##     year_22 + title_analyst + remote_full + residence_USA + salary_USD, 
##     data = train_set)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -133733  -29739   -6233   19033  352839 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      90418       7281  12.419  < 2e-16 ***
## EL_MI           -16617       6406  -2.594  0.00983 ** 
## EL_EX            74795      11406   6.558 1.62e-10 ***
## EL_EN           -45118       8386  -5.380 1.25e-07 ***
## size_s          -13496       8032  -1.680  0.09364 .  
## year_22         -13633       5635  -2.419  0.01597 *  
## title_analyst   -32396       6327  -5.120 4.67e-07 ***
## remote_full     -12885       6341  -2.032  0.04278 *  
## residence_USA    59906       9163   6.538 1.83e-10 ***
## salary_USD       22041       9263   2.380  0.01778 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 52200 on 417 degrees of freedom
## Multiple R-squared:  0.4955, Adjusted R-squared:  0.4846 
## F-statistic: 45.51 on 9 and 417 DF,  p-value: < 2.2e-16
```

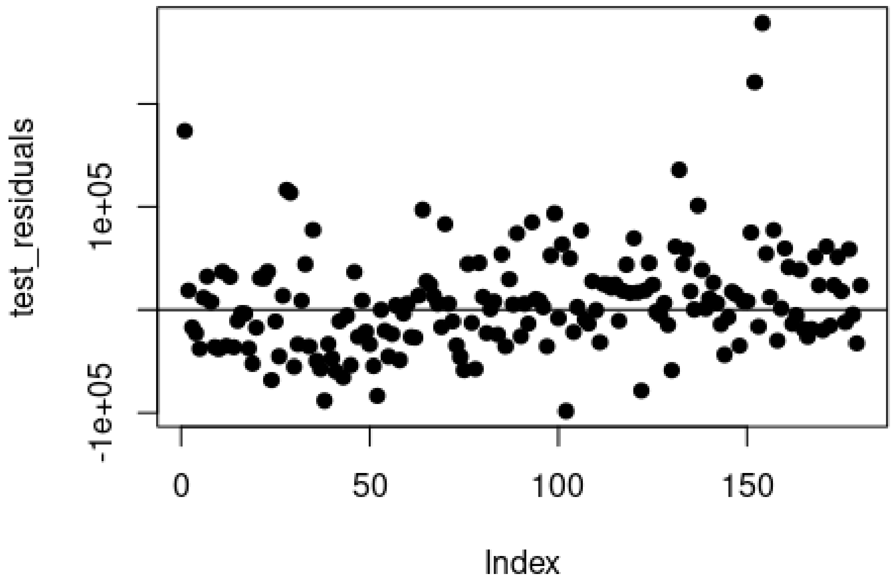
Running the Final (Best) Model on the test dataset

```
test_predictions <- predict(final, newdata = test_set)

#Looking at residual stats
mse <- mean((test_set$salary_in_usd - test_predictions)^2)
rmse <- sqrt(mse)
mae <- mean(abs(test_set$salary_in_usd - test_predictions))

#Calculating RSE to compare to Training Set
test_residuals <- test_set$salary_in_usd - test_predictions
```

```r
# Sum of squared residuals
ssr <- sum(test_residuals^2)
# Degrees of freedom
df <- nrow(test_set) - (length(coef(final)) - 1)
# Calculate RSE
rse <- sqrt(ssr / df)


paste("Testing Set MSE: ", mse)

## [1] "Testing Set MSE:  2615680219.8606"

paste("Testing Set RMSE: ", rmse)

## [1] "Testing Set RMSE:  51143.7212163976"

paste("Testing Set MAE: ", mae)

## [1] "Testing Set MAE:  36083.4346281401"

paste("Testing Set RSE: ",rse)

## [1] "Testing Set RSE:  52472.3508131021"
```

Looking at the residual plot

```r
plot(test_residuals, pch = 19)
abline(h = 0)
```



Training Set RMSE and MAE

```
# Load necessary library
library(Metrics)

##
## Attaching package: 'Metrics'

## The following objects are masked from 'package:caret':
##
##     precision, recall

# Assuming your linear model is named lm_model and your training dataset is
named train_data
# Fit the model (if not already done)
# lm_model <- lm(y ~ x1 + x2, data=train_data)

# Generate predictions
train_predictions <- predict(final, train_set)

# Calculate RMSE
train_rmse <- rmse(train_set$salary_in_usd, train_predictions)

mae_value <- mae(train_set$salary_in_usd, train_predictions)

# Print the MAE
print(paste("MAE:", mae_value))
```

```
## [1] "MAE: 34597.9177188507"

# Print the RMSE
print(paste("RMSE:", train_rmse))

## [1] "RMSE: 51587.5826265684"
```

Calculate R-Squared of the model on the test set

```
# Calculate R-squared
SSE <- sum((test_set$salary_in_usd - test_predictions)^2)
SST <- sum((test_set$salary_in_usd - mean(test_set$salary_in_usd))^2)
r_squared <- 1 - SSE/SST

# Print the R-squared
print(paste("R-squared:", r_squared))

## [1] "R-squared: 0.410335979796309"
```

Interpretation of the Results

Our model results align strongly with our hypothesis and intuitively make sense within the context of the business world. Starting with the model performance metrics, we see that the training data set and testing data set performed similarly based on residual metrics RMSE and MAE. This shows the model was not overfit on the training data and can perform strongly on testing data which gives us confidence to employ this model on more data it has yet to see before. Looking at the model's R-Squared metric we see that the model only accounts for 49.6% and 41.0% of the explained variability on the training data and testing data, respectively. This tells us that there are some variables not accounted for and can be explored and added to the model in future research.

Looking at some of our variables within our final regression model we see that Expert and Entry Level Experience Level, along with Employee Residence (USA) are the variables of highest significance, with p-values approaching 0. Digging deeper into the variables we see the coefficient for Entry Level jobs equaling -$45,118 which represents a decrease of that amount relative to our baseline category of a Senior Level position, which makes sense given that entry level jobs are given to employees with much less experience then a Senior Level worker. On the other side, for executive level employees we see an increase of $74,795 for having an executive level position. Another variable of interest due to its high significance is Employee Residence (USA). In our regression, living in the United States adds $59,906 to the worker's salary as opposed to the baseline category of living outside of the United States. This is an interesting finding from our regression model, especially considering that salaries were controlled for by converting all salaries to USD. Digging deeper into these variables means and medians we see that USA residents have substantially higher values of $149,194.10 and $138,475.00 respectively, compared to $67,754.04 and $62,649 of non-US residents, which helps explain the variables strong impact on our model. Looking at the Job Title Analyst variable we see a coefficient of -$32,396 which represents a decrease of that amount relative to our baseline category of "Other" which includes key words like "Architect", "Developer", and "Machine Learning".

Intuitively, this makes sense as many “Analyst” roles tend to be entry level in the business world. Additionally, the role of a data analyst is often making sense out of existing data through analysis, visualization and writing reports as opposed to a more demanding skill set of developing predictive models using Python, Java, and machine learning techniques often seen in other roles such as a “data scientist” 2.

Conclusion:

Taking a look at the final model, the most significant predictors for salary were Executive and Entry experience levels, having the “analyst” job title, and residing in the US.

As for the limitations of the project, the dataset that was chosen did not involve some key variables that could potentially contribute to someone’s data scientist salary. For example, an important area to examine when determining salary is education level, which could vary from a high school diploma, bachelor’s degree, masters degree, or even PhD. Thus, we would expect someone with a higher education level to have a greater earning potential compared to one with a lower education level. Another area to investigate is the industry in which the data scientist works. For instance, if a data scientist is working at a large technology firm, such as Google or Amazon, it is expected that their potential salary is greater than a data scientist working at a large bank. Therefore, looking at the industry in which the data scientist works could explain some of the variance in which our current variables were not able to cover. Additionally, the dataset included 8 large outliers which were kept in hopes these data points would explain what variables can cause such high salaries. Our interest in this question limited our model’s overall performance, increasing the RMSE and MAE, but helps provide a holistic view of data science salaries. Other considerations to improve model performance were performing a log transformation of the dependent variable (Salaries) as it was partially right skewed or using bootstrap regression methods to help deal with outliers.

In the future, this project could take various directions and answer more in depth research questions. A more complex research question that could be explored is a more US-based analysis, answering how much working in different areas of the US affects one’s earning potential. Perhaps working as a data scientist in a more technology centric area such as San Francisco increases earning potential compared to a more financial centric area such as New York City.

Sources:

https://www.mastersindatascience.org/careers/data-analyst-vs-data-scientist/#:~:text=A%20data%20analyst%20relies%20on,to%20manipulate%20and%20analyze%20data.

https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries/