Data Science Jobs Salary Analysis

Ri-on Kim

2024-03-30

Intro

Data science-related jobs have become increasingly prominent across all industries. Analyzing these roles by experience level, specific job titles, and salaries can provide valuable insights, particularly for individuals studying computer science, statistics, and data science. This project will focus primarily on salary, as it represents a clear and significant index for differentiation.

Questions

- How do salaries differ by specific jobs?
- How is salary distributed across company sizes?
- How does salary change based on experience level?
- Overall, which specific job offers the highest salary?

Importing the Data & Libraries

```r
data <- read.csv("ds_salaries.csv")
library(ggplot2)
library(dplyr)
library(scales)
```

For Data Science job salary analysis, importing data set is essential. The 'ds_salaries.csv' data is sourced from Kaggle (https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023/data). This project will use the library ggplot2 and dplyr mainly.

Understanding the Data

```r
head(data, 7)
```

```
##   work_year experience_level employment_type                job_title salary
## 1      2023               SE              FT Principal Data Scientist  80000
## 2      2023               MI              CT              ML Engineer  30000
## 3      2023               MI              CT              ML Engineer  25500
## 4      2023               SE              FT           Data Scientist 175000
## 5      2023               SE              FT           Data Scientist 120000
## 6      2023               SE              FT        Applied Scientist 222200
## 7      2023               SE              FT        Applied Scientist 136000
##   salary_currency salary_in_usd employee_residence remote_ratio
## 1             EUR         85847                 ES          100
## 2             USD         30000                 US          100
## 3             USD         25500                 US          100
## 4             USD        175000                 CA          100
```

```
## 5              USD        120000                  CA          100
## 6              USD        222200                  US            0
## 7              USD        136000                  US            0
##   company_location company_size
## 1               ES            L
## 2               US            S
## 3               US            S
## 4               CA            M
## 5               CA            M
## 6               US            L
## 7               US            L
```

The imported dataset comprises 11 columns and 3,755 rows. It includes crucial variables like experience_level, job_title, salary_in_usd, and company_size. Initial inspection reveals several issues that need addressing through data cleaning:

- The abbreviations in the experience_level column (e.g., "MI", "SE") are not intuitive.
- Since salaries can vary significantly based on employee residence, analyzing data without filtering by location may lead to confusion.
- The job_title column contains an overly broad array of titles.(Beside the data shown in the head, there are too many different job titles like AI engineer, data science consultant.)

Cleaning the Data

```r
data_cleaned <- data %>%
  filter(employee_residence == "US") %>%
  mutate(
    job_title = case_when(
    grepl("Data Scientist", job_title, ignore.case = TRUE) ~ "Data Scientist",
    grepl("Data Engineer", job_title, ignore.case = TRUE) ~ "Data Engineer",
    grepl("Analyst", job_title, ignore.case = TRUE) ~ "Data Analyst",
    grepl("Machine Learning", job_title, ignore.case = TRUE) ~ "Machine Learning Engineer",
    grepl("Manager", job_title, ignore.case = TRUE) ~ "Data Science Manager",
    grepl("Director", job_title, ignore.case = TRUE) ~ "Data Science Manager",
    grepl("Architect", job_title, ignore.case = TRUE) ~ "Data Architect",
    TRUE ~ "Other"
  ))

data_cleaned <- data_cleaned %>%
  mutate(
    experience_level = case_when(
      experience_level == "EN" ~ "Entry",
      experience_level == "MI" ~ "Midium",
      experience_level == "SE" ~ "Senior",
      experience_level == "EX" ~ "Executive",
      TRUE ~ as.character(experience_level)
    ),
    experience_level = factor(experience_level, levels =
                              c("Entry", "Midium", "Senior", "Executive"))
  )

head(data_cleaned, 7)
```

```
##   work_year experience_level employment_type      job_title salary
```

```
## 1      2023           Midium              CT          Other  30000
## 2      2023           Midium              CT          Other  25500
## 3      2023           Senior              FT          Other 222200
## 4      2023           Senior              FT          Other 136000
## 5      2023           Senior              FT Data Scientist 147100
## 6      2023           Senior              FT Data Scientist  90700
## 7      2023           Senior              FT   Data Analyst 130000
##   salary_currency salary_in_usd employee_residence remote_ratio
## 1             USD         30000                 US          100
## 2             USD         25500                 US          100
## 3             USD        222200                 US            0
## 4             USD        136000                 US            0
## 5             USD        147100                 US            0
## 6             USD         90700                 US            0
## 7             USD        130000                 US          100
##   company_location company_size
## 1               US            S
## 2               US            S
## 3               US            L
## 4               US            L
## 5               US            M
## 6               US            M
## 7               US            M
```

```r
table(data_cleaned$job_title)
```

```
##
##              Data Analyst            Data Architect             Data Engineer
##                       554                        98                       908
##      Data Science Manager            Data Scientist Machine Learning Engineer
##                       102                       679                       266
##                     Other
##                       397
```

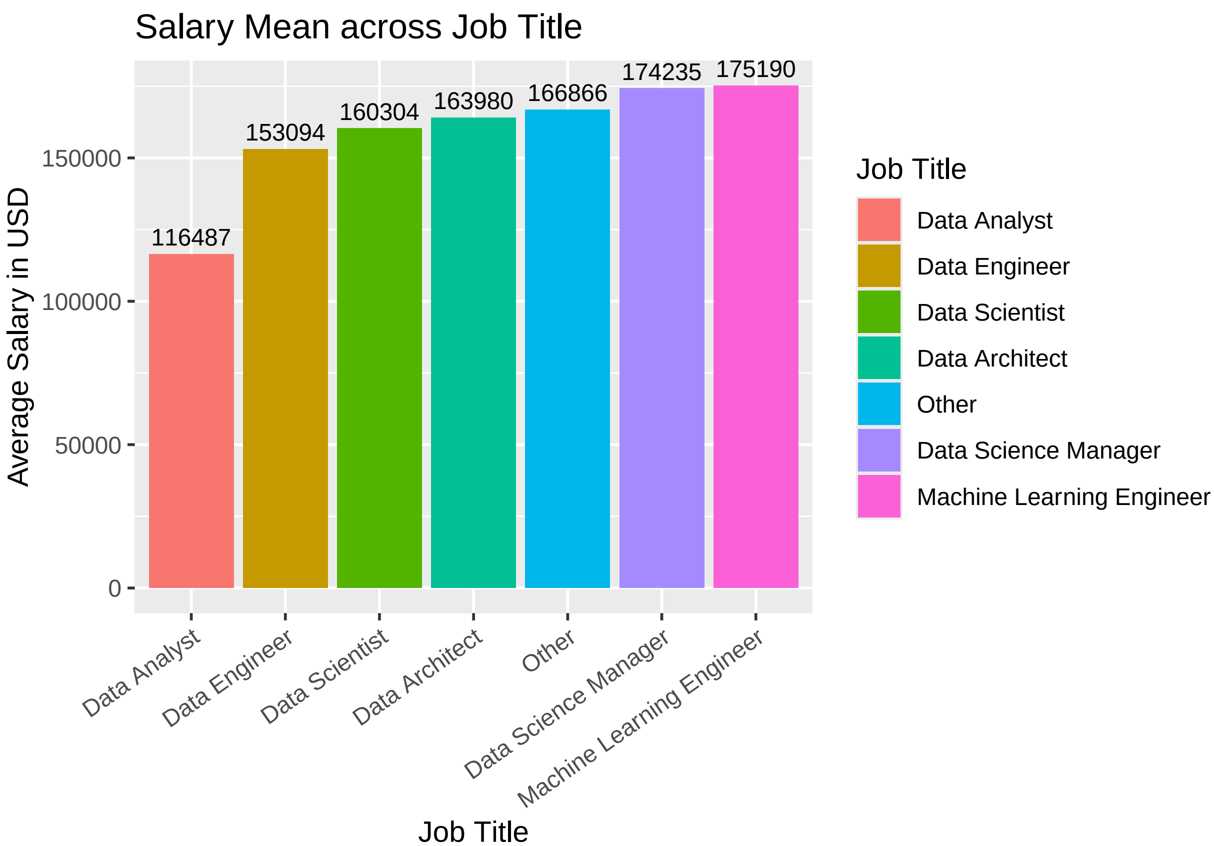
After cleaning, notable adjustments include:

- For clarity, the experience_level column now uses intuitive labels: "Entry", "Mid", "Senior", and "Executive".
- To avoid confusion, the dataset has been filtered to only include employees residing in the US.
- To make the graph clear, similar elements in job_title is merged and minor elements are classified as "Other". The "Other" variable contains the job like "Applied Scientist", "Data Modeler", "AI Engineer", "Computer Vision Engineer", etc. Finally, there are only 7 elements in job_title.

Salary Mean and Median across Job Title

```r
avg_salary_by_title <- data_cleaned %>%
  group_by(job_title) %>%
  summarise(average_salary = mean(salary_in_usd, na.rm = TRUE)) %>%
  arrange(desc(average_salary))

ggplot(avg_salary_by_title, aes(x=reorder(job_title, average_salary), y=average_salary, fill = reorder(
  geom_bar(stat="identity") +
  geom_text(aes(label = round(average_salary, 0)),
            vjust = -0.5,
```

```
            size = 3) +
  theme(axis.text.x = element_text(angle = 35, hjust = 1)) +
  labs(title="Salary Mean across Job Title",
       x="Job Title", y="Average Salary in USD", fill = "Job Title")
```

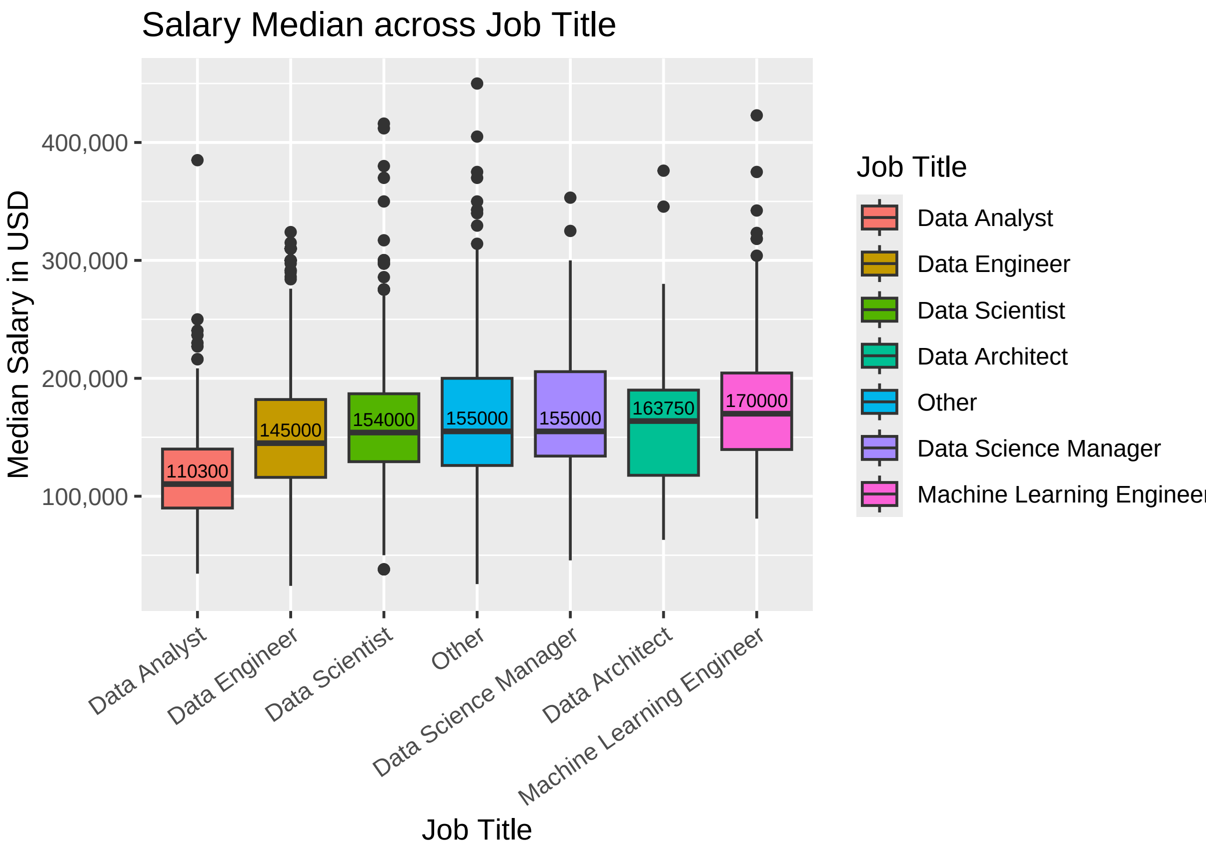


Analysis reveals:

- Data Analysts have the lowest average salary among the roles examined, with a significant gap to Data Engineers.
- Machine Learning Engineers and Data Science Managers emerge as the highest earners, showcasing the narrow salary gap between these two positions.

For consistency, the same color scheme used in the average salary graph will be applied to the median salary graph for clear comparison.

```
data_cleaned$job_title <- factor(data_cleaned$job_title, levels = rev(avg_salary_by_title$job_title))

ggplot(data_cleaned, aes(x = reorder(job_title, salary_in_usd, FUN = median), y = salary_in_usd, fill =
  geom_boxplot() +
  scale_y_continuous(labels = scales::comma) +
  theme(axis.text.x = element_text(angle = 35, hjust = 1)) +
  labs(title = "Salary Median across Job Title",
       x = "Job Title", y = "Median Salary in USD", fill = "Job Title") +
  stat_summary(fun=median, geom="text", aes(label=..y..), vjust=-0.5, color="black", size = 2.5)
```



Analysis reveals:

- Although there are some differences, the median salary across job titles is quite similar to the average salary.
- One point to note is that the Data Science Manager role shows a significant difference between the average and median salaries. The box plot displays a large range in the Q1 to Q2 quartile, indicating that high 25% of Data Science Managers are paid way more than the median salary data science manager compared to other jobs.

Salary Distribution across Job Titles and Company Size

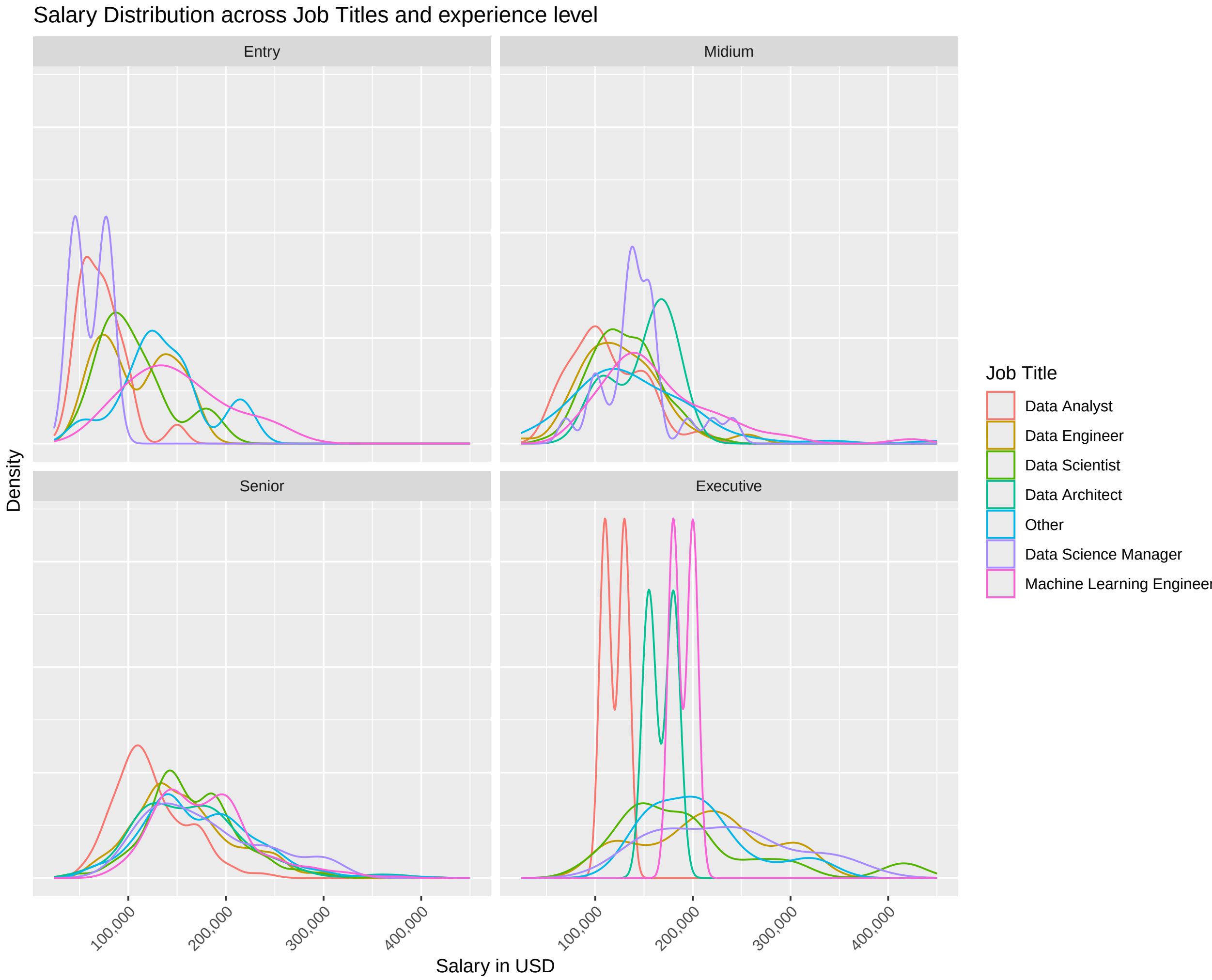
```
data_cleaned$company_size <- factor(data_cleaned$company_size, levels = c("S", "M", "L"))

ggplot(data_cleaned, aes(x = salary_in_usd, col = job_title)) +
  geom_density(alpha = 0.8) +
  scale_x_continuous(labels = label_comma()) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank()) +
  labs(title = "Salary Distribution across Job Titles and Company Size",
       x = "Salary in USD", y = "Density", col = "Job Title") +
  facet_wrap(~company_size)
```

Now, we can examine the salary distribution based on company size. I utilized the facet_wrap function to display three company sizes simultaneously. From examining the graph, we can observe the following:

- As company size increases, the overall salary across all job titles tends to increase as well. Notably, the difference between small and mid-sized companies is larger compared to the difference between mid-sized and large companies.
- An interesting observation in the graph for small-sized companies is that some data scientists are receiving exceptionally high salaries (over $400,000). This suggests that data scientists may perform particularly well in small-sized companies compared to other jobs.
- Additionally, it's intriguing that mid-sized companies tend to offer higher salaries to data analysts than large-sized companies.

Salary Distribution across Job Titles and experience level

```
ggplot(data_cleaned, aes(x = salary_in_usd, col = job_title)) +
  geom_density(alpha = 0.8) +
  scale_x_continuous(labels = label_comma()) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank()) +
  labs(title = "Salary Distribution across Job Titles and experience level",
       x = "Salary in USD", y = "Density", col = "Job Title") +
  facet_wrap(~experience_level)
```



Given that experience level can significantly impact salary, it's crucial to examine the salary distribution by experience level. To facilitate this analysis, I utilized the facet_wrap function once more to compare across different experience levels. The following observations can be made:

- As one might expect, the overall salary distribution curves shift to the right, indicating that higher experience levels correlate with higher salaries across all job titles.
- Interestingly, Data Science Managers have the lowest salaries at the entry level. This could be attributed to the role typically requiring more extensive experience than other positions.
- Although Machine Learning Engineers and Data Architects have high average and median salaries compared to other roles, at the executive level, Data Scientists, Data Engineers, Data Science Managers, and others can get higher salaries than both Machine Learning Engineers and Data Architects.

App for Data Science Job Salary Distribution Comparisons

https://1minute99.shinyapps.io/shiny/

- The shiny app is designed for comparing salary distribution more interactively. It can help to focus on comparing two jobs' salary.

Conclusion

- Data Analysts generally earn the lowest salaries.

- Salary difference between small size company and medium size company is significant compare to the difference between mid size company and large size company for most of jobs.
- Machine Learning Engineers rank among the highest earners.
- At the executive level, several roles can surpass salary of Machine Learning Engineers and Data Architects, indicating the value of experience across job titles.