Portfolio 1 - Data Science job salary Analysis

Ayushi Mishra

2/19/2022

Introduction

Approaching my graduation, I am intrigued to explore different types of Data Science jobs out there in the market; especially, how the skill-set differs across different data science jobs and what does the pay-scale looks like. Therefore, I chose to analyze the `Data Science job salary` dataset as it would me help and other people who are looking for Data Science related jobs.

For this portfolio, I modified the row values for `job_title_sim` column, which generalizes the different data-related jobs in a few category. I will be mainly focusing on `Data Scientist`, `Data Engineer` and `Analyst` for my analysis.

```
url <- "https://raw.githubusercontent.com/mishra37/Portfolio-1/main/data_cleaned_2021.csv"
uncleaned <- read_csv(url, show_col_types = FALSE)

cleaned_data1 <- uncleaned[-c(1, 4)]
cleaned_data1[cleaned_data1$job_title_sim == "data scientist","job_title_sim" ] <- "Data Scientist"
cleaned_data1[cleaned_data1$job_title_sim == "analyst","job_title_sim" ] <- "Analyst"
cleaned_data1[cleaned_data1$job_title_sim == "data engineer","job_title_sim" ] <- "Data Engineer"
```

Plot - 1: Skills required by different Data-related jobs

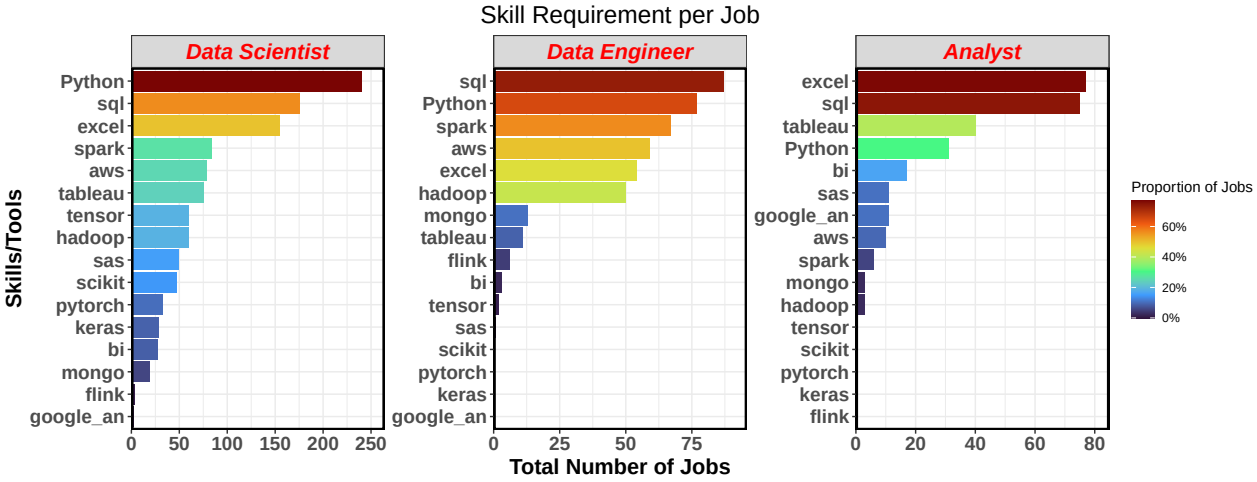
I performed `pivot_longer` on columns with different skills/tools and added them in a single column of `Skills`. Furthermore, I summarized the total number and proportion of jobs, in each job category, that required a particular skill. After creating the bar plot, I reordered the bars within each panel based on total number of jobs using `reorder_within`.

```
skills_data <- cleaned_data1 %>%
              pivot_longer(Python:google_an, names_to = "Tools", values_to = "Required")

skills_data %>%
  group_by(job_title_sim, Tools) %>%
  summarise(total = sum(Required),
            prop = mean(Required)) %>%
  filter(job_title_sim == "Data Scientist" || job_title_sim == "Data Engineer"
         || job_title_sim == "Analyst") %>%
  ggplot(aes(x = total, y = reorder_within(Tools, total, job_title_sim), fill = prop)) +
  geom_col() +
  labs(x = "Total Number of Jobs",
       y = "Skills/Tools",
       title = "Skill Requirement per Job") +
  facet_wrap(.~ reorder(job_title_sim, -total), scales = "free", ncol = 4) +
  scale_y_reordered() +
  scale_x_continuous(expand = c(0, 0, 0.1, 0.1)) +
  scale_fill_viridis_c("Proportion of Jobs", option = "turbo", label = percent) +
  theme_bw() +
```

```
  theme(
    plot.title = element_text(hjust = 0.5, size = 18),
    strip.text.x = element_text(size = 16, color = "red", face = "bold.italic"),
    axis.text.x = element_text(size = 14, face = "bold"),
    axis.title.x = element_text(size = 16, face = "bold"),
    axis.title.y = element_text(size = 16, face = "bold"),
    axis.text.y = element_text(size = 14, face = "bold"),
    panel.border = element_rect(color = "black", size = 1.5))
```

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



Looking at the plot, we can see that `Data Scientist` jobs require almost every skill. For `Data Engineer` jobs, we can see that sql and Python is the highly required skill, but there seems to be some requirement of visualization based tools (Tableau) as well. For different types of `Analyst` jobs, we can conclude that excel and SQL is the most important skill.

It was surprising to see that SQL was most common and majorly required by most Data-related jobs; while I thought that Python would have been the most common and highly asked skill for any Data-related jobs.

Plot 2 - Distribution of Average Salary per sector and job title

For this plot, I created a patchwork with two plots - one of which shows how average salary varies across different data science jobs in top 10 sectors and other one shows a comparison between how average salary varies across different Job Titles.

For the first plot, I pulled out top 10 sectors based on the number of Data Science jobs. I mainly focused on private and publicly owned companies because there were significantly higher number of jobs in these categories. I created `geom_boxplot` and faceted by type of ownership, and ordered by median salary.

For the second plot, I created `geom_segment` by using minimum and maximum of average salary as x-axis ends. I highlighted average, minimum, and maximum salaries; along with reordering the segments based on the average salaries.

```
sectors <- cleaned_data1 %>%
  group_by(Sector) %>%
  count()%>%
  arrange(desc(n)) %>%
  pull(Sector)

top10sectors <- sectors[1:10]
```

```r
sector_salary <- cleaned_data1 %>%
  filter(Sector %in% top10sectors)%>%
  filter(`Type of ownership` %in% c('Company - Private','Company - Public'))


p1 <- ggplot(sector_salary, aes(x =`Avg Salary(K)`, y = reorder_within(Sector, `Avg Salary(K)`,
                                `Type of ownership`, median), fill = `Type of ownership`)) +
  geom_boxplot() +
  labs(x = "Average Salary (K)",
       y = "Job Sectors",
       title = "Average Salary of Data Science related jobs in top 10 sectors") +
  theme_bw() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16),
    panel.border = element_rect(color = "black", size = 1.25),
    strip.text.x = element_text(color = "black", face = "bold"),
    axis.text.x = element_text(size = 13, face = "bold"),
    axis.text.y = element_text(size = 13, face = "bold"),
    axis.title.y = element_text(size = 16, face = "bold"),
    axis.title.x = element_text(size = 16, face = "bold"),
    legend.position = "none") +
  facet_wrap(~`Type of ownership`, scales = "free_y") +
  scale_y_reordered()
```

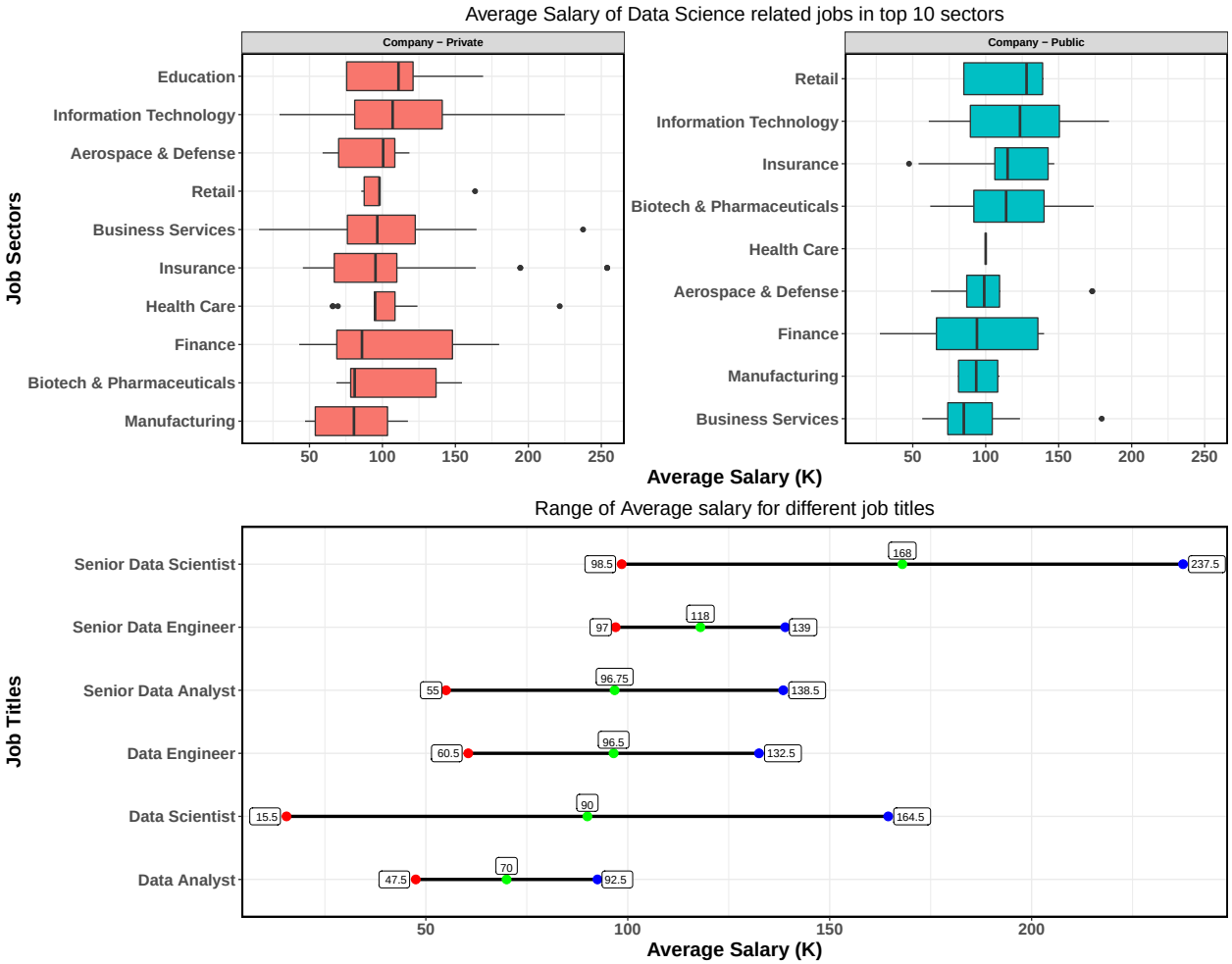
```r
jobs <- cleaned_data1 %>%
  group_by(`Job Title`) %>%
  count() %>%
  filter(n > 10) %>%
  pull(`Job Title`)

company_type <- cleaned_data1 %>%
filter(`Job Title` %in% jobs) %>%
  group_by(`Job Title`) %>%
  summarise(`min Avg Salary` = min(`Avg Salary(K)`),
            `max Avg Salary` = max(`Avg Salary(K)`)) %>%
  mutate(`Avg Salary(K)` = (`max Avg Salary` + `min Avg Salary`)/2)


p2 <- ggplot(company_type, aes(y = reorder(`Job Title`, `Avg Salary(K)`))) +
  geom_segment(aes(x = `min Avg Salary`, xend = `max Avg Salary`, yend = reorder(
    `Job Title`, `Avg Salary(K)`)), size=1.25) +
  geom_point(aes (`min Avg Salary`),  colour = "red", size = 3)+
  geom_label(aes(x = `min Avg Salary`, label = `min Avg Salary`), hjust = 1.2, size = 3) +
  geom_point (aes (`max Avg Salary`),  colour = "blue",size = 3)+
  geom_label(aes(x = `max Avg Salary`, label = `max Avg Salary`), hjust = -0.15, size = 3) +
  geom_point (aes (`Avg Salary(K)`), colour = "green", size = 3)+
  geom_label(aes(x = `Avg Salary(K)`, label = `Avg Salary(K)`), vjust = -0.3, size = 3) +
  labs(x = "Average Salary (K)",
       y = "Job Titles",
       title = "Range of Average salary for different job titles") +
  scale_x_continuous(breaks = seq(0,250,50)) +
  theme_bw() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16),
```

```
    panel.border = element_rect(color = "black", size = 1.25),
    axis.title.x = element_text(size = 16, face = "bold"),
    axis.text.x = element_text(size = 13, face = "bold"),
    axis.text.y = element_text(size = 13, face = "bold"),
    axis.title.y = element_text(size = 16, face = "bold")
  )
```

```
p1/p2
```



From the upper plot, it was surprising to see that `Education` and `Retail` sector has the highest median salary, even defeating `Information Technology`. It was also surprising to see that Business Services had lowest median salary in publicly owned sector.

From the lower plot, it can inferred that the maximum average salary is highest for `Data Scientist` jobs. It is surprising to see that an entry level Data Scientist can earn more than a Senior Data Engineer and Data Analyst. Another surprising thing is that there is a very large range for a `Data Scientist` average salary.