607 Project 2 Glassdoor

Lu Beyer

Overview

This is a dataset on job posting on Glassdoor. I would like to see which locations have the most job postings and which jobs within those location are most in demand.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## 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     1.3.1
## 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
```

```
data <- read.csv("Uncleaned_DS_jobs.csv")
```

I wanted to consolidate the data so similar jobs are grouped together. Anything that did not fit within the broad title categories, I chose to list as "other". I removed text from the salary column for potential salary analysis, and changed the rating column to only show values between 0-5. Because there is no guide to this data, I don't know what a negative value means, nor what the scale is for these values.

```
a <- data %>%
  select(-c(index, Headquarters, Founded, Competitors, Job.Description)) %>%
  rename(Title = Job.Title,
         Salary = Salary.Estimate,
         Company = Company.Name,
         Type = Type.of.ownership) %>%
  mutate(Company = str_sub(Company, end = -5),
         Salary = str_sub(Salary, end = -18),
         Rating = as.character(Rating)) %>%
  mutate(across(where(is.character), ~na_if(., "-1"))) %>%
  mutate(Title = ifelse(str_detect(tolower(Title), "manager"), "Manager",
           ifelse(str_detect(tolower(Title), "data science|data scientist"), "Data Scientist",
           ifelse(str_detect(tolower(Title), "analyst"), "Data Analyst",
           ifelse(str_detect(tolower(Title), "machine learning"), "Machine Learning",
           ifelse(str_detect(tolower(Title), "big data"), "Big Data",
           ifelse(str_detect(tolower(Title), "data engineer"), "Data Engineer",
           ifelse(str_detect(tolower(Title), "computer scientist"), "Computer Scientist",
                  "Other")))))))
```

I wanted to organize the data by location to see the counts of each location within this set.

```
popular_locations <- a %>%
  group_by(Location) %>%
  summarise(counts = n())
```

I created a table listing the top ten locations within this dataset.

```
most_popular_locations <- popular_locations %>%
  arrange(., counts) %>%
  tail(10)
```

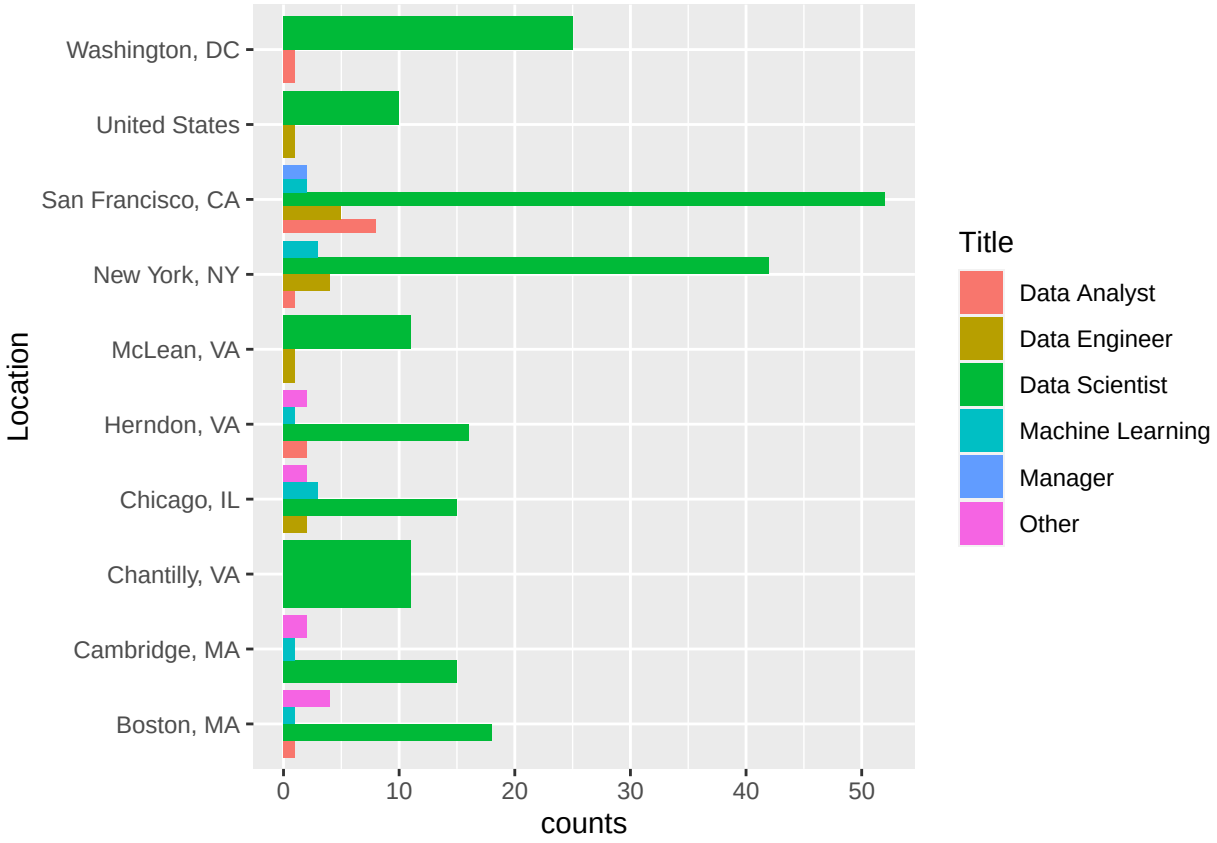
Using my dataset with the top ten locations, I filtered the data to only show jobs listed in those location.

```
b <- a %>%
  filter(Location %in% most_popular_locations$Location) %>%
  group_by(Location, Title) %>%
  summarise(counts = n())
```

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

I created this visualization to show which job titles appear most within the top ten most popular locations.

```
ggplot(data = b, aes(x = counts, y = Location, fill = Title)) +
  geom_bar(position = "dodge", stat = "identity")
```

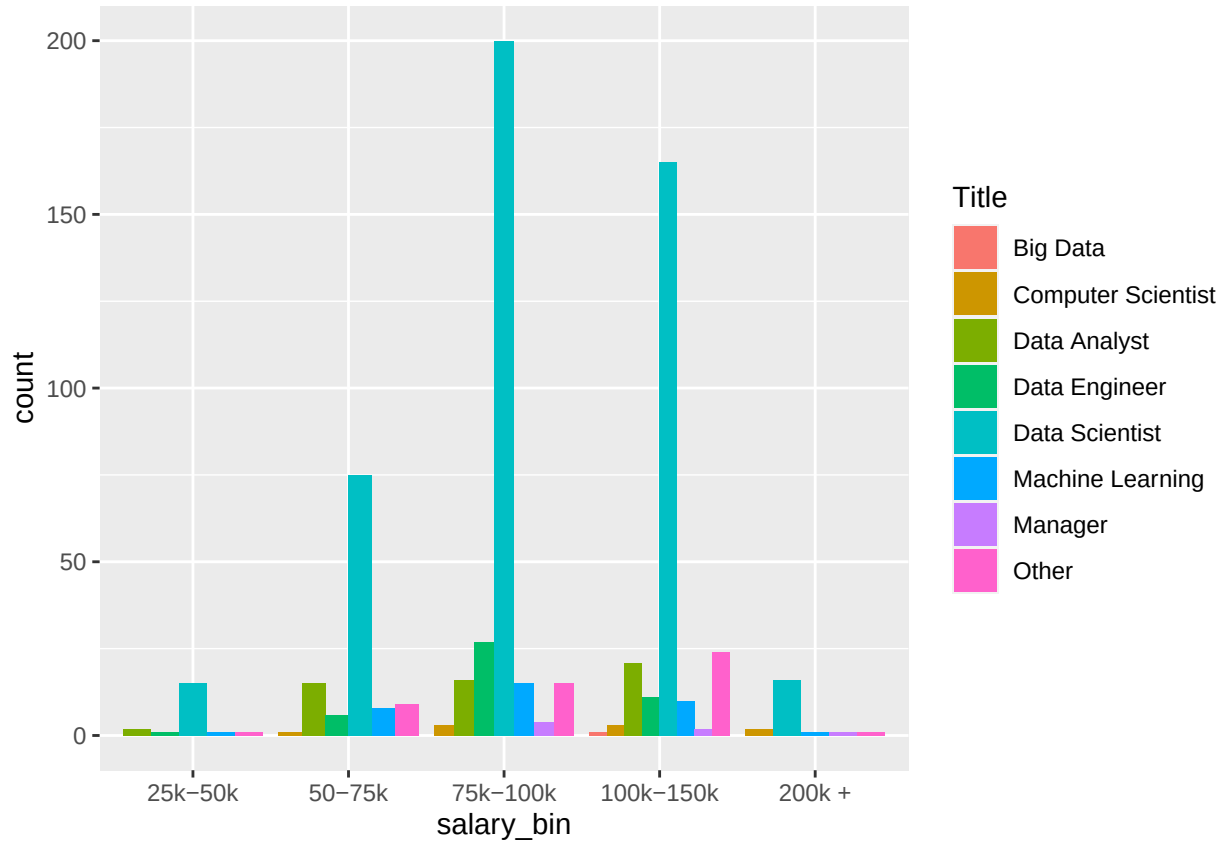


I organized the data to more easily see the minimum salary each position was offering and the frequency of those positions within the ranges I set and created a chart to visualize the results.

```
salary <- a %>%
  mutate(Salary = str_sub(Salary, end = -6),
         Salary = str_replace_all(Salary, "[^0-9]",  "")) %>%
  mutate(Salary = as.numeric(Salary)) %>%
  mutate(salary_bin = cut(Salary, breaks = c(0, 25, 50, 75, 100, 150, 200, Inf), labels = c("Below 25k"
  group_by(Title, salary_bin) %>%
  summarise(count = n())
```

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

```
ggplot(data = salary, aes(x = salary_bin, y = count, fill = Title)) +
  geom_bar(position = "dodge", stat = "identity")
```




```
table(a$Title)
```

```
##
##           Big Data Computer Scientist       Data Analyst      Data Engineer
##                  1                  9                 54                45
##     Data Scientist   Machine Learning            Manager              Other
##                471                 35                  7                50
```

Conclusion

The most in-demand position within this data set is for data scientists. The range of salary for this position ranges, but the majority of positions in this set offered a minimum salary between $75k-150k with 365 positions offered within that range. Data science positions did, however, also have the most positions offered with a minimum salary set between 25k-50k with 15 positions offered within that range. It's important to be aware that data science positions are expected to have higher numbers within any given category because "data scientist" accounted for 70% of the total positions listed, and it is important to remember there was some variance to the specific title for each position because I organized the data to group together similar positions. Overall, it feels like a good time to be in a data science program.

If I were to expand on this analysis, I would be interested in how job postings looked over the last year to see if data scientist has consistently been in demand, or if it is trending to be a more in demand position. I would also be interested in data looking at the amount of people applying to these positions over the last year, to see how many data scientists are looking for employment vs the amount of organizations hiring for this position, and see if it is shifting to larger or smaller applicant pools for positions like these.