MSDS 597 Final Project

Jongsoo Han and Heon Park

2022-05-04

MSDS 597 Final Project

Analysis of Data Science Career

Introduction

With the increased demand on career opportunity for Data Scientist, we would explore this dataset in order to provide the readers with the understanding/overview of the data science career market.

Dataset

The dataset is available at Kaggle, and this dataset was extracted by scrapping the job postings related to the position of ‘Data Scientist’ from www.glassdoor.com.

Reading the dataset

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
df = read.csv(file="data_cleaned_2021.csv", header=TRUE)
df = as.tibble(df)
```

```
## Warning: ‘as.tibble()‘ was deprecated in tibble 2.0.0.
## Please use ‘as_tibble()‘ instead.
## The signature and semantics have changed, see ‘?as_tibble‘.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was generated.
```

```
head(df)
```

```
## # A tibble: 6 x 42
##   index Job.Title   Salary.Estimate Job.Description Rating Company.Name Location
##   <int> <chr>       <chr>           <chr>            <dbl> <chr>        <chr>
## 1     0 Data Scien~ $53K-$91K (Gla~ "Data Scientis~    3.8 "Tecolote R~ Albuque~
## 2     1 Healthcare~ $63K-$112K (Gl~ "What You Will~    3.4 "University~ Linthic~
## 3     2 Data Scien~ $80K-$90K (Gla~ "KnowBe4, Inc.~    4.8 "KnowBe4\n4~ Clearwa~
## 4     3 Data Scien~ $56K-$97K (Gla~ "*Organization~    3.8 "PNNL\n3.8"  Richlan~
## 5     4 Data Scien~ $86K-$143K (Gl~ "Data Scientis~    2.9 "Affinity S~ New Yor~
## 6     5 Data Scien~ $71K-$119K (Gl~ "CyrusOne is s~    3.4 "CyrusOne\n~ Dallas,~
## # ... with 35 more variables: Headquarters <chr>, Size <chr>, Founded <int>,
## #   Type.of.ownership <chr>, Industry <chr>, Sector <chr>, Revenue <chr>,
## #   Competitors <chr>, Hourly <int>, Employer.provided <int>,
## #   Lower.Salary <int>, Upper.Salary <int>, Avg.Salary.K. <dbl>,
## #   company_txt <chr>, Job.Location <chr>, Age <int>, Python <int>,
## #   spark <int>, aws <int>, excel <int>, sql <int>, sas <int>, keras <int>,
## #   pytorch <int>, scikit <int>, tensor <int>, hadoop <int>, tableau <int>, ...
```

Checking there are Null/NA value in the dataset

```
sum(is.null(df))
```

```
## [1] 0
```

Exploratory Data Analysis

1. Summary of the dataset

```
df_summary = df %>% select(Job.Title, Company.Name, Location, Size, Industry,
                            Lower.Salary, Upper.Salary, Avg.Salary.K., Job.Location)

df_summary$Job.Title = as.factor(df_summary$Job.Title)
df_summary$Company.Name = as.factor(df_summary$Company.Name)
df_summary$Location = as.factor(df_summary$Location)
df_summary$Size = as.factor(df_summary$Size)
df_summary$Industry  = as.factor(df_summary$Industry )
df_summary$Job.Location  = as.factor(df_summary$Job.Location)

summary(df_summary)
```

```
##                  Job.Title                                  Company.Name
##  Data Scientist       :131   MassMutual\n3.6                     : 14
##  Data Engineer        : 53   Reynolds American\n3.1              : 14
##  Senior Data Scientist: 34   Takeda Pharmaceuticals\n3.7         : 14
##  Data Analyst         : 15   Software Engineering Institute\n2.6: 11
##  Senior Data Engineer : 14   Liberty Mutual Insurance\n3.3       : 10
##  Senior Data Analyst  : 12   PNNL\n3.8                           : 10
##  (Other)              :483   (Other)                             :669
##              Location              Size
##  New York, NY     : 55   1001 - 5000  :150
##  San Francisco, CA: 49   501 - 1000   :134
##  Cambridge, MA    : 47   10000+       :130
##  Chicago, IL      : 32   201 - 500    :117
##  Boston, MA       : 23   51 - 200     : 94
##  San Jose, CA     : 13   5001 - 10000 : 76
##  (Other)          :523   (Other)      : 41
##                                        Industry    Lower.Salary     Upper.Salary
##  Biotech & Pharmaceuticals                 :112   Min.   : 15.00   Min.   : 16.0
##  Insurance Carriers                        : 63   1st Qu.: 52.00   1st Qu.: 96.0
##  Computer Hardware & Software              : 59   Median : 69.50   Median :124.0
##  IT Services                               : 50   Mean   : 74.75   Mean   :128.2
##  Health Care Services & Hospitals          : 49   3rd Qu.: 91.00   3rd Qu.:155.0
##  Enterprise Software & Network Solutions: 42   Max.   :202.00   Max.   :306.0
##  (Other)                                   :367
##  Avg.Salary.K.    Job.Location
##  Min.   : 15.5   CA     :152
##  1st Qu.: 73.5   MA     :103
##  Median : 97.5   NY     : 72
##  Mean   :101.5   VA     : 41
##  3rd Qu.:122.5   IL     : 40
##  Max.   :254.0   MD     : 35
##                  (Other):299
```

-There 742 job postings related to Data Science.

-Average Salary is $101.5K.

2. States with Most Number of Job

```
library(dplyr)
library(ggplot2)
library(scales)
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

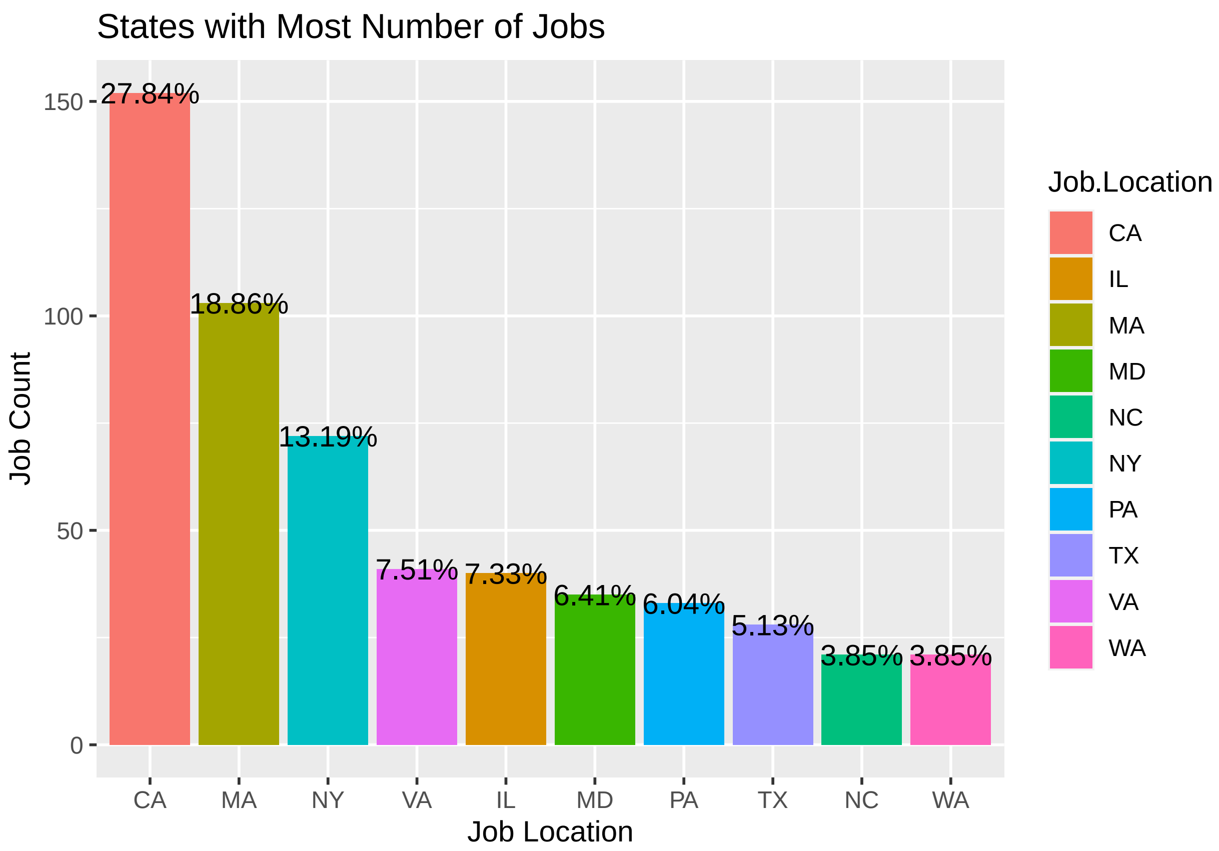
```
df %>% count(Job.Location)%>% summarise(Total_State=n())
```

```
## # A tibble: 1 x 1
##   Total_State
##         <int>
## 1          37
```

```
df_state = df %>% select(Job.Location) %>%
  group_by(Job.Location) %>%
  summarise(count = n()) %>%
  arrange(desc(count))%>%
  head(10)

df_state
```

```
## # A tibble: 10 x 2
##    Job.Location count
##    <chr>        <int>
##  1 CA             152
##  2 MA             103
##  3 NY              72
##  4 VA              41
##  5 IL              40
##  6 MD              35
##  7 PA              33
##  8 TX              28
##  9 NC              21
## 10 WA              21
```

```
df_state %>%
  ggplot(aes(reorder(Job.Location, -count), count, fill=Job.Location)) +
  geom_bar(stat="identity") +
  geom_text(aes(label = percent(count/sum(count)))) +
  ggtitle("States with Most Number of Jobs") +
  xlab("Job Location") +
  ylab("Job Count")
```



-The Data shows California has the most number of jobs.

-Evidently, California to has the most number of jobs as it is a hub for Tech. companies and has silicon valley.

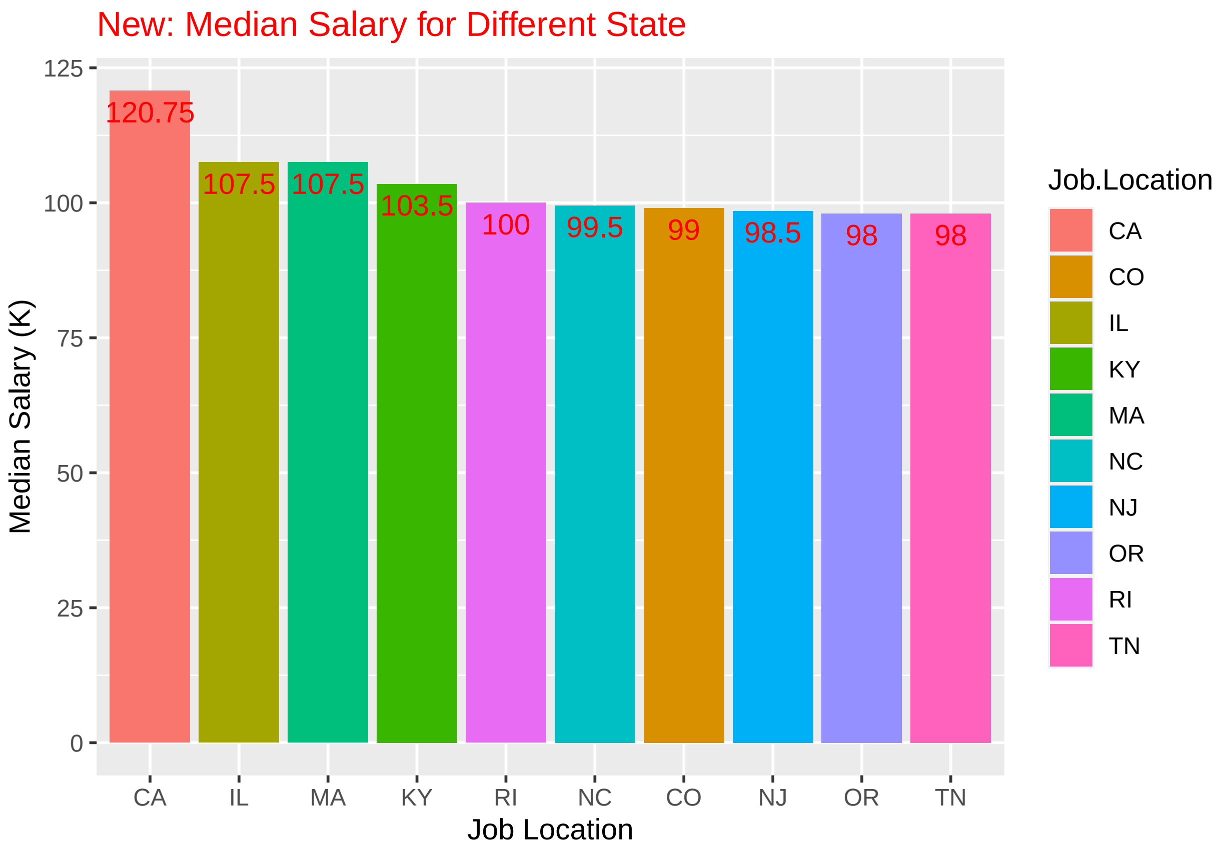
-California, Massachusetts, New York, Virginia together has around 50% jobs.

-Interesting fact is that despite having the largest number of Fortune 500 companies HQ in New York, it is still on 3rd position.

3. Median Salary in Different States

New Section: We added following section (2.5) having been suggested by the professor during the presentation.

```
df_Salary_med = df %>% select(Job.Location, Avg.Salary.K.)

df_Salary_med = df_Salary_med %>%
  group_by(Job.Location) %>%
  summarise(MedSalary = median(Avg.Salary.K.)) %>%
  arrange(desc(MedSalary)) %>%
  head(10)

df_Salary_med %>%
  ggplot(aes(reorder(Job.Location, -MedSalary), MedSalary, fill=Job.Location)) +
  geom_bar(stat="identity") +
  ggtitle("New: Median Salary for Different State") +
  theme(plot.title = element_text(color = "Red")) +
  geom_text(aes(label = MedSalary), vjust = 1.5, colour = "Red") +
  xlab("Job Location") +
  ylab("Median Salary (K)")
```

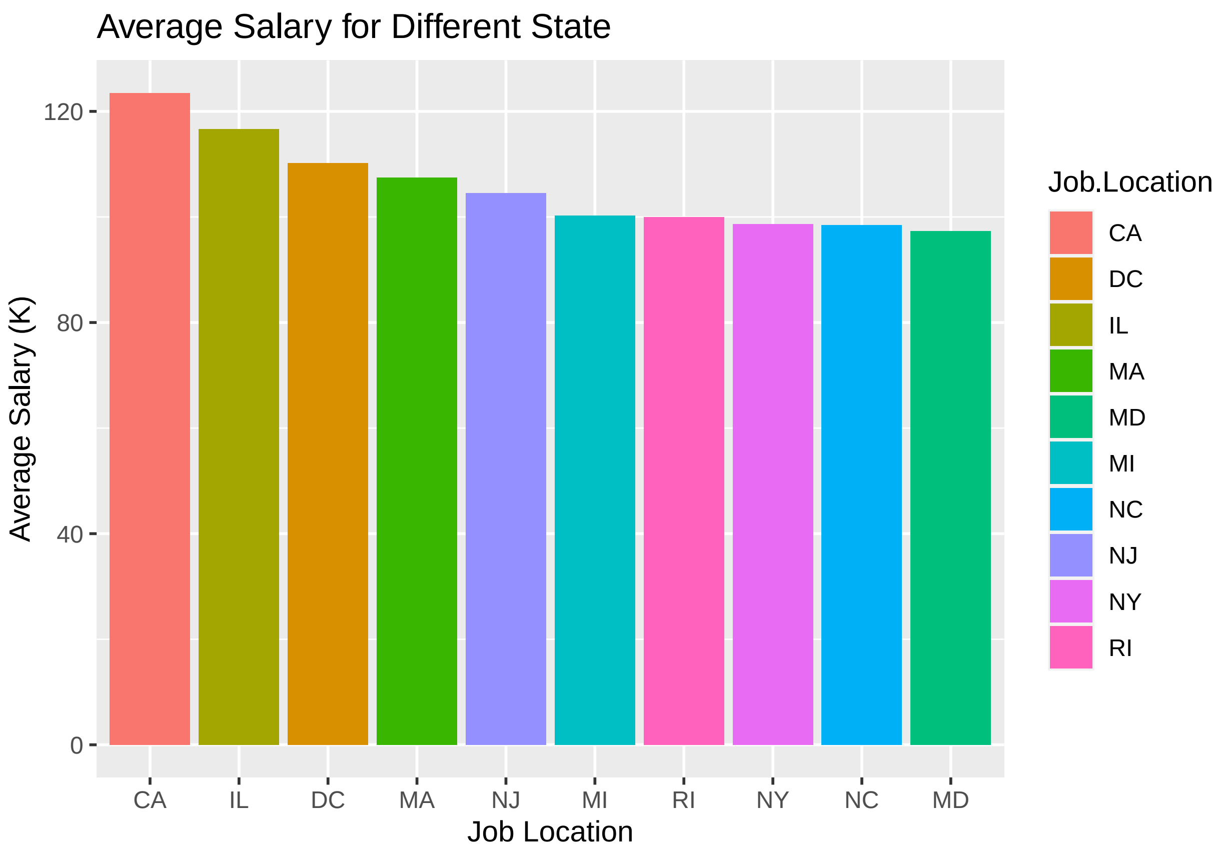


-The Data shows California has the highest Median Salary.

-IL and MA have the second and third highest Median Salary.

4. Average Salary in Different States

```
df_avg_state = df %>% select(Job.Location, Avg.Salary.K.)

df_avg_state = df_avg_state %>%
  group_by(Job.Location) %>%
  summarise(AvgSalary = mean(Avg.Salary.K.)) %>%
  arrange(desc(AvgSalary)) %>%
  head(10)

df_avg_state %>%
  ggplot(aes(reorder(Job.Location, -AvgSalary), AvgSalary, fill=Job.Location)) +
  geom_bar(stat="identity") +
  ggtitle("Average Salary for Different State") +
  xlab("Job Location") +
  ylab("Average Salary (K)")
```



-The graph shows average annual salary for different states.

-State with highest number of job, California also offers the highest average annual salary, followed by Illinois.

-Maryland has the lowest average annual salary.

-It is interesting to find that NJ has higher average salary than that of NY. (NJ: 5th & NY: 8th)

5. Average Minimal and Maximal Salaries in Different State

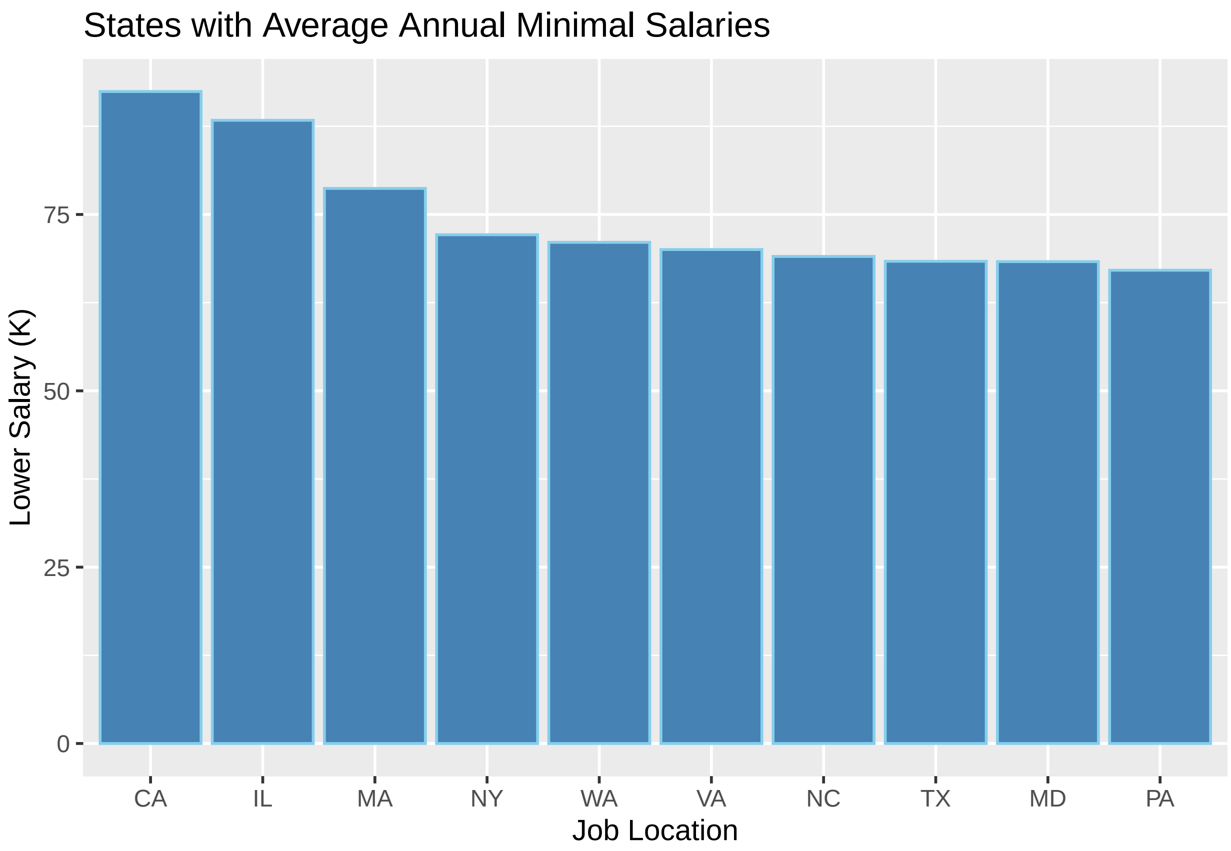
```
df_avg_min_max = df %>% select(Job.Location, Lower.Salary, Upper.Salary)

df_avg_min_max = df_avg_min_max %>%
  group_by(Job.Location) %>%
  summarise(LowSalary = mean(Lower.Salary), UpSalary = mean(Upper.Salary))

df_avg_min_max = df_avg_min_max %>%
  inner_join(df_state) %>%
  arrange(desc(count))
```

```
## Joining, by = "Job.Location"
```

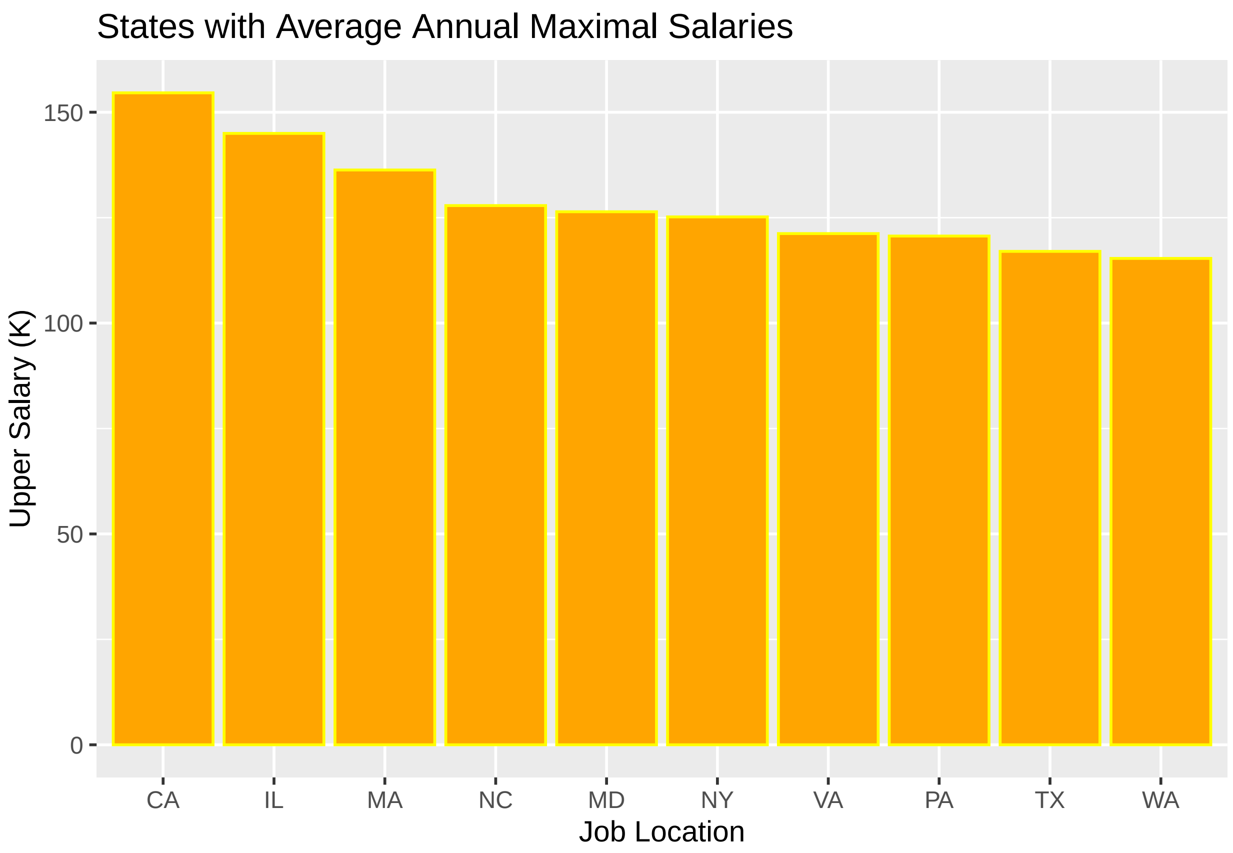
```
df_avg_min_max
```

```
## # A tibble: 10 x 4
##    Job.Location LowSalary UpSalary count
##    <chr>              <dbl>    <dbl> <int>
##  1 CA                  92.4     155.   152
##  2 MA                  78.7     136.   103
##  3 NY                  72.1     125.    72
##  4 VA                  70.0     121.    41
##  5 IL                  88.4     145.    40
##  6 MD                  68.3     126.    35
##  7 PA                  67.1     121.    33
##  8 TX                  68.4     117     28
##  9 NC                  69.0     128.    21
## 10 WA                  71.0     115.    21
```

```
df_avg_min_max %>%
  ggplot(aes(reorder(Job.Location, -LowSalary), LowSalary)) +
  geom_bar(stat="identity", color='skyblue', fill='steelblue') +
  ggtitle("States with Average Annual Minimal Salaries") +
  xlab("Job Location") +
  ylab("Lower Salary (K)")
```



```
df_avg_min_max %>%
  ggplot(aes(reorder(Job.Location, -UpSalary), UpSalary)) +
  geom_bar(stat="identity", color='yellow', fill='orange') +
  ggtitle("States with Average Annual Maximal Salaries") +
  xlab("Job Location") +
  ylab("Upper Salary (K)")
```



-State with highest number of job, California also offers the highest average maximal annual salary, followed by Illinois.

-Washington has the lowest average maximum annual salary among the top 10 states.

-We find that both California and Illinois has almost the same average minimal annual salary.

-We find that Pennsylvania has the lowest average minimum annual salary among the top 10 states.

6. Top 5 Industries with Maximum Number of Data Science Related Job Postings

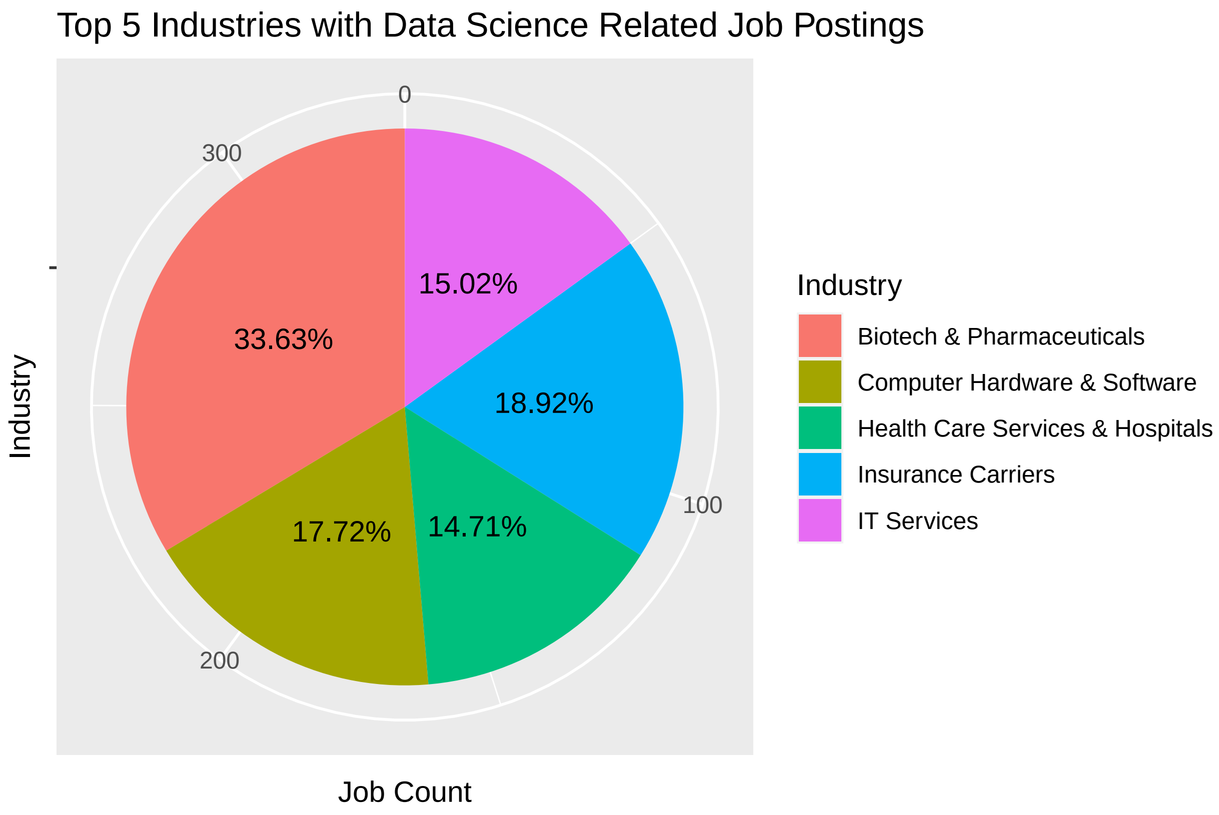
```
df %>% count(Industry)%>% summarise(Total_Industry=n())
```

```
## # A tibble: 1 x 1
##   Total_Industry
##            <int>
## 1             60
```

```
df_top5 = df %>% select(Industry) %>%
  group_by(Industry) %>%
  summarise(count = n()) %>%
  arrange(desc(count))%>%
  head(5)

df_top5
```

```
## # A tibble: 5 x 2
##   Industry                             count
##   <chr>                                <int>
## 1 Biotech & Pharmaceuticals              112
## 2 Insurance Carriers                      63
## 3 Computer Hardware & Software            59
## 4 IT Services                             50
## 5 Health Care Services & Hospitals        49
```

```
df_top5 %>%
  ggplot(aes("", count, fill=Industry)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start = 0) +
  geom_text(aes(label = percent(count/sum(count))), position = position_stack(vjust=0.5)) +
  ggtitle("Top 5 Industries with Data Science Related Job Postings") +
  xlab("Industry") +
  ylab("Job Count")
```



-Biotech & Pharmaceuticals Industry has maximum number of jobs followed by Insurance carriers.

-IT industry has fewer jobs for data science related roles.

-More than 65% data science related jobs lie in top 10 industries.

-For this dataset, Biotech & Pharmaceuticals Industry has twice the amount of jobs compared to IT services industry.

7. Companies with Maximum Number of Job Openings

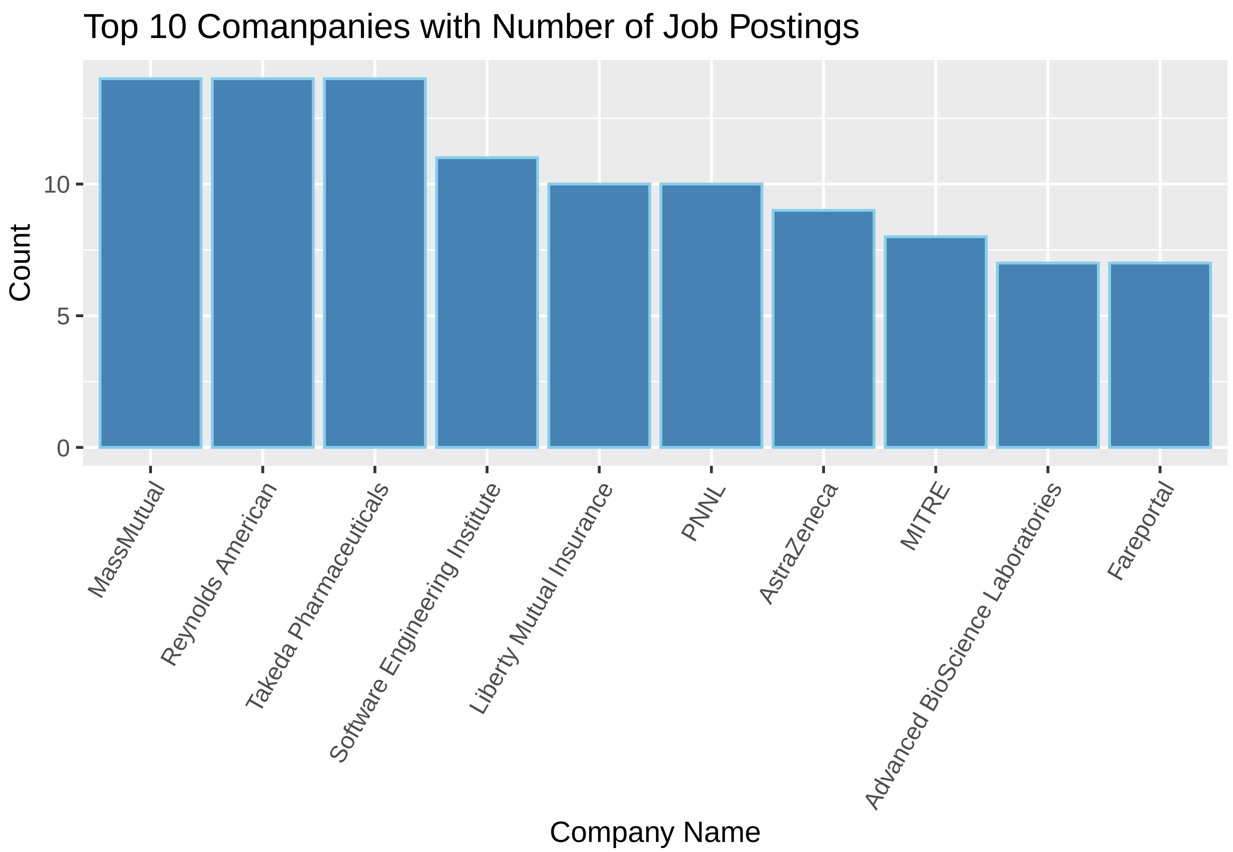
```
df %>% count(company_txt)%>% summarise(Total_Company=n())
```

```
## # A tibble: 1 x 1
##   Total_Company
##           <int>
## 1           343
```

```
df_company = df %>% select(company_txt) %>%
  group_by(company_txt) %>%
  summarise(count = n()) %>%
  arrange(desc(count))%>%
  head(10)

df_company
```

```
## # A tibble: 10 x 2
##    company_txt                      count
##    <chr>                            <int>
##  1 MassMutual                          14
##  2 Reynolds American                   14
##  3 Takeda Pharmaceuticals              14
##  4 Software Engineering Institute      11
##  5 Liberty Mutual Insurance            10
##  6 PNNL                                10
##  7 AstraZeneca                          9
##  8 MITRE                                8
##  9 Advanced BioScience Laboratories     7
## 10 Fareportal                           7
```

```
df_company %>%
  ggplot(aes(reorder(company_txt, -count), count)) +
  geom_bar(stat="identity", color='skyblue', fill='steelblue') +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  ggtitle("Top 10 Comanpanies with Number of Job Postings") +
  xlab("Company Name") +
  ylab("Count")
```



-There are total 342 companies in the dataset. This is why there is less number of job postings by each company.

-Reynolds American, MassMutrual and Takeda Pharmaceuticals company tops the list with 14 job postings related to data science.

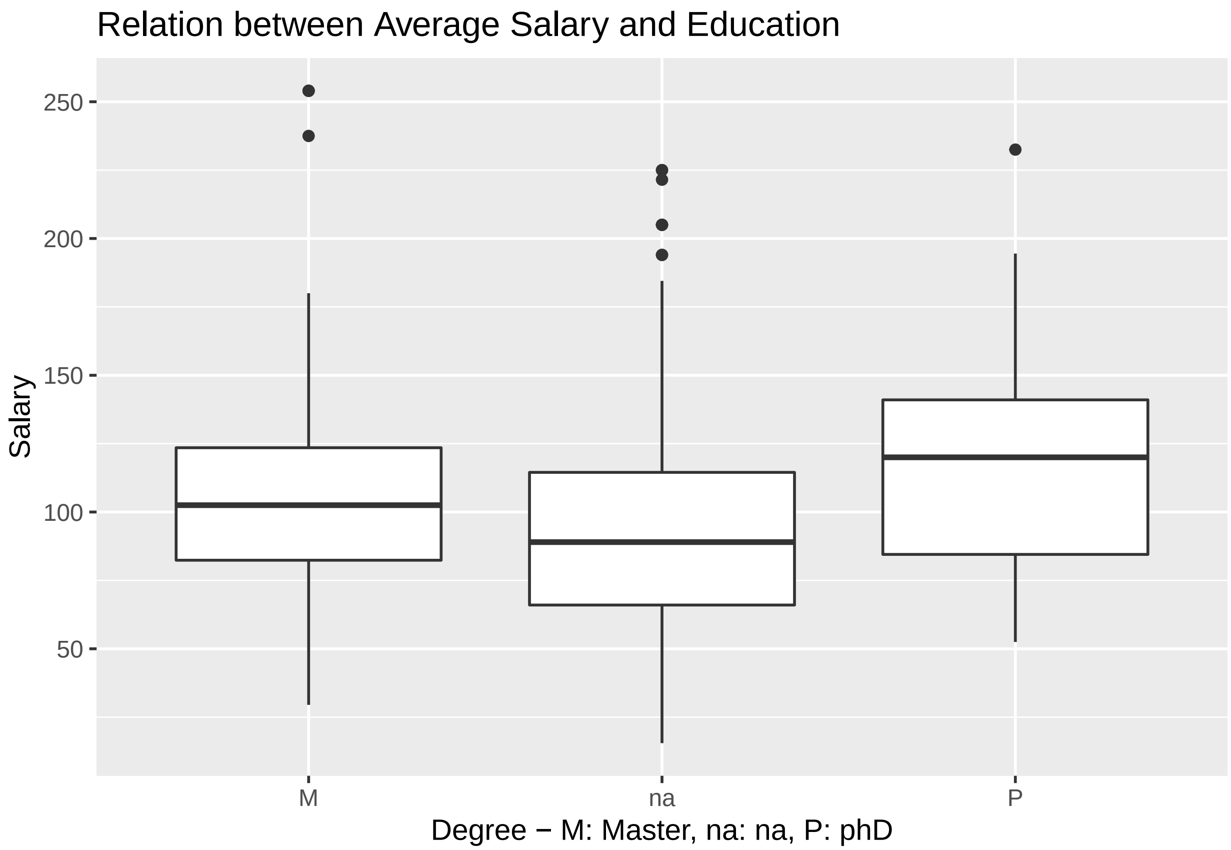
-We find that a Pharmaceutical Industry is leading with the most number of job postings, we see the same trend here as well

8. Relation between Average Salary and Education

```
df_education = df %>% select(job_title_sim, Avg.Salary.K., Degree)

df_education %>%
  group_by(Degree) %>%
  summarise(AvgSalary = mean(Avg.Salary.K.))
```

```
## # A tibble: 3 x 2
##   Degree AvgSalary
##   <chr>      <dbl>
## 1 M          106.
## 2 na          94.7
## 3 P          116.
```

```
df_education %>%
  ggplot(aes(Degree, Avg.Salary.K.)) +
  geom_boxplot() +
  ggtitle("Relation between Average Salary and Education") +
  xlab("Degree - M: Master, na: na, P: phD") +
  ylab("Salary")
```



-Most of the companies has mentioned Masters degree in their job descriptions.

-For companies that mentioned a PhD degree in their job description, they offered much highest average annual salary as compared to Masters.

Summary

-We were able to find many interesting results from this analysis such as which state's average salary is highest and lowest, and also which industry have most job posting related to Data Science.

-One interesting issue that was quite surprising is that we found no correlation between company size/revenue and salary. It seems that salary is more correlated to a specific geography and degree.

-Having conducted this analytical review of the datasets from Glassdoor, we were able to examine one of the most importing decisioning factor in job requisitions which is the salary.

-We also could correlate this analysis to the expectations of the company from a data science employee.

-As prospective candidates in Data Science industry, This project exercise has helped us to gain understanding around the various components in the current job market.