Story04:How much do we get paid?

Mahmud Hasan Al Raji

2023-10-22

Introduction

We have used the term "Data Practitioner" as a generic job descriptor because we have so many different job role titles for individuals whose work activities overlap including Data Scientist, Data Engineer, Data Analyst, Business Analyst, Data Architect, etc.In this assignment we will answer the question, "How much do we get paid?" The analysis and data visualizations will address the variation in average salary based on role descriptor and state.

It is noted that the annual salary data (up to October 2023) for Data Scientists, Data Engineers, Data Analysts, and Business Analysts role by US states are collected from ZipRecruiter, which is an online employment marketplace and job search engine that connects employers and job seekers. The data, originally presented in a tabular format on ZipRecruiter, were converted into CSV file format and subsequently uploaded to my GitHub repository for further analysis. Data links are given in Data Reference section.

Load library

```
#library(readxl)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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
```

```
#library(rvest)
#library(xml2)
```

Get Data

```
data_scientist<-read.csv("https://raw.githubusercontent.com/Raji030/story4_data/main/data_scientist.csv
data_engineer<-read.csv("https://raw.githubusercontent.com/Raji030/story4_data/main/data_engineer.csv")
data_analyst<-read.csv("https://raw.githubusercontent.com/Raji030/story4_data/main/data_analyst.csv")
business_analyst<-read.csv("https://raw.githubusercontent.com/Raji030/story4_data/main/data_business_ana
```

Data Preparation

```
## For Data Scientist:

# Remove "$" and ","
data_scientist$Annual.Salary <- as.numeric(gsub("[$,]", "", data_scientist$Annual.Salary))
data_scientist$Hourly.Wage <- as.numeric(gsub("[$]", "", data_scientist$Hourly.Wage))

# Rename Annual.Salary column to Annual_Salary, and Hourly.Wage to Hourly_Wage
colnames(data_scientist)[colnames(data_scientist) == "Annual.Salary"] <- "Annual_Salary"
colnames(data_scientist)[colnames(data_scientist) == "Hourly.Wage"] <- "Hourly_Wage"

# Remove Monthly.Pay and Weekly.Pay columns
data_scientist <- data_scientist[, !(colnames(data_scientist) %in% c("Monthly.Pay", "Weekly.Pay"))]

# Arrange "State" column in ascending order
data_scientist <- data_scientist[order(data_scientist$State), ]

# Include Role column
data_scientist<-data_scientist%>%mutate(Role="Data Scientist")

head(data_scientist)
```

```
##         State Annual_Salary Hourly_Wage           Role
## 44    Alabama        103437       49.73 Data Scientist
## 17     Alaska        121680       58.50 Data Scientist
## 16    Arizona        122400       58.85 Data Scientist
## 50   Arkansas         96232       46.27 Data Scientist
## 2  California        143099       68.80 Data Scientist
## 35   Colorado        110256       53.01 Data Scientist
```

```
## For Data Engineer:

# Remove "$" and ","
data_engineer$Annual.Salary <- as.numeric(gsub("[$,]", "", data_engineer$Annual.Salary))
data_engineer$Hourly.Wage <- as.numeric(gsub("[$]", "", data_engineer$Hourly.Wage))

# Rename Annual.Salary column to Annual_Salary, and Hourly.Wage to Hourly_Wage
colnames(data_engineer)[colnames(data_engineer) == "Annual.Salary"] <- "Annual_Salary"
colnames(data_engineer)[colnames(data_engineer) == "Hourly.Wage"] <- "Hourly_Wage"

# Remove Monthly.Pay and Weekly.Pay columns
data_engineer<- data_engineer[, !(colnames(data_engineer) %in% c("Monthly.Pay", "Weekly.Pay"))]

# Arrange "State" column in ascending order
data_engineer <- data_engineer[order(data_engineer$State), ]

# Include Role column
data_engineer<-data_engineer%>%mutate(Role="Data Engineer")

head(data_engineer)
```

```
##         State Annual_Salary Hourly_Wage          Role
```

```
## 49    Alabama         96843       46.56 Data Engineer
## 5      Alaska        141854       68.20 Data Engineer
## 28    Arizona        114596       55.09 Data Engineer
## 42   Arkansas        109017       52.41 Data Engineer
## 12 California        128004       61.54 Data Engineer
## 15   Colorado        124464       59.84 Data Engineer
```

```
## For Data Analyst:

# Remove "$" and ","
data_analyst$Annual.Salary <- as.numeric(gsub("[$,]", "", data_analyst$Annual.Salary))
data_analyst$Hourly.Wage <- as.numeric(gsub("[$]", "", data_analyst$Hourly.Wage))

# Rename Annual.Salary column to Annual_Salary, and Hourly.Wage to Hourly_Wage
colnames(data_analyst)[colnames(data_analyst) == "Annual.Salary"] <- "Annual_Salary"
colnames(data_analyst)[colnames(data_analyst) == "Hourly.Wage"] <- "Hourly_Wage"

# Remove Monthly.Pay and Weekly.Pay columns
data_analyst<- data_analyst[, !(colnames(data_analyst) %in% c("Monthly.Pay", "Weekly.Pay"))]

# Arrange "State" column in ascending order
data_analyst <- data_analyst[order(data_analyst$State), ]

# Include Role column
data_analyst<-data_analyst%>%mutate(Role="Data Analyst")
head(data_analyst)
```

```
##         State Annual_Salary Hourly_Wage         Role
## 36    Alabama         69689       33.50 Data Analyst
## 17     Alaska         79092       38.03 Data Analyst
## 8     Arizona         82463       39.65 Data Analyst
## 50   Arkansas         62193       29.90 Data Analyst
## 24 California         75874       36.48 Data Analyst
## 34   Colorado         71206       34.23 Data Analyst
```

```
## For Business Analyst:

# Remove "$" and ","
business_analyst$Annual.Salary <- as.numeric(gsub("[$,]", "", business_analyst$Annual.Salary))
business_analyst$Hourly.Wage <- as.numeric(gsub("[$]", "", business_analyst$Hourly.Wage))

# Rename Annual.Salary column to Annual_Salary, and Hourly.Wage to Hourly_Wage
colnames(business_analyst)[colnames(business_analyst) == "Annual.Salary"] <- "Annual_Salary"
colnames(business_analyst)[colnames(business_analyst) == "Hourly.Wage"] <- "Hourly_Wage"

# Remove Monthly.Pay and Weekly.Pay columns
business_analyst<- business_analyst[, !(colnames(business_analyst) %in% c("Monthly.Pay", "Weekly.Pay"))]

# Arrange "State" column in ascending order
business_analyst <- business_analyst[order(business_analyst$State), ]

# Include Role column
business_analyst<-business_analyst%>%mutate(Role="Business Analyst")
head(business_analyst)
```

```
##          State Annual_Salary Hourly_Wage             Role
## 46    Alabama         74375       35.76 Business Analyst
## 17     Alaska         89739       43.14 Business Analyst
## 22    Arizona         88010       42.31 Business Analyst
## 50   Arkansas         71249       34.25 Business Analyst
## 5  California        100388       48.26 Business Analyst
## 37   Colorado         81671       39.27 Business Analyst
```

```
## For Data Scientist:
# Find maximum, minimum, and average annual salary
max_salary <- max(data_scientist$Annual_Salary)
min_salary <- min(data_scientist$Annual_Salary)
avg_salary1 <- mean(data_scientist$Annual_Salary)

# Find state with maximum salary
states_max_salary <- data_scientist$State[data_scientist$Annual_Salary == max_salary]

# Find state with minimum salary
states_min_salary <- data_scientist$State[data_scientist$Annual_Salary == min_salary]

# Find number of states with a salary greater than annual average
states_above_avg <- sum(data_scientist$Annual_Salary > avg_salary1)

# Show results
cat("Maximum Annual Salary:", max_salary, "in", states_max_salary, "\n")
```

```
## Maximum Annual Salary: 145027 in New York
```

```
cat("Minimum Annual Salary:", min_salary, "in", states_min_salary, "\n")
```

```
## Minimum Annual Salary: 96232 in Arkansas
```

```
cat("Average Annual Salary:", avg_salary1, "\n")
```

```
## Average Annual Salary: 116193.8
```

```
cat("Number of states that get annual salary greater than average for Data Scientists:", states_above_a
```

```
## Number of states that get annual salary greater than average for Data Scientists: 26
```

```
## For Data Engineer:
# Find maximum, minimum, and average annual salary
max_salary <- max(data_engineer$Annual_Salary)
min_salary <- min(data_engineer$Annual_Salary)
avg_salary2 <- mean(data_engineer$Annual_Salary)

# Find state with maximum salary
states_max_salary <- data_engineer$State[data_engineer$Annual_Salary == max_salary]

# Find state with minimum salary
```

```r
states_min_salary <- data_engineer$State[data_engineer$Annual_Salary == min_salary]

# Find number of states with a salary greater than annual average
states_above_avg <- sum(data_engineer$Annual_Salary > avg_salary2)

# Show results
cat("Maximum Annual Salary:", max_salary, "in", states_max_salary, "\n")
```

```
## Maximum Annual Salary: 149976 in Nevada
```

```r
cat("Minimum Annual Salary:", min_salary, "in", states_min_salary, "\n")
```

```
## Minimum Annual Salary: 96743 in Florida
```

```r
cat("Average Annual Salary:", avg_salary2, "\n")
```

```
## Average Annual Salary: 119563.1
```

```r
cat("Number of states that get annual salary greater than average for Data Engineers:", states_above_avg
```

```
## Number of states that get annual salary greater than average for Data Engineers: 20
```

```r
# For Data Analyst:
# Find maximum, minimum, and average annual salary
max_salary <- max(data_analyst$Annual_Salary)
min_salary <- min(data_analyst$Annual_Salary)
avg_salary3 <- mean(data_analyst$Annual_Salary)

# Find state with maximum salary
states_max_salary <- data_analyst$State[data_analyst$Annual_Salary == max_salary]

# Find state with minimum salary
states_min_salary <- data_analyst$State[data_analyst$Annual_Salary == min_salary]

# Find number of states with a salary greater than annual average
states_above_avg <- sum(data_analyst$Annual_Salary > avg_salary3)

# Show results
cat("Maximum Annual Salary:", max_salary, "in", states_max_salary, "\n")
```

```
## Maximum Annual Salary: 98238 in New York
```

```r
cat("Minimum Annual Salary:", min_salary, "in", states_min_salary, "\n")
```

```
## Minimum Annual Salary: 62193 in Arkansas
```

```r
cat("Average Annual Salary:", avg_salary3, "\n")
```

```
## Average Annual Salary: 75121.02
```

```r
cat("Number of states that get annual salary greater than average for Data Analysts:", states_above_avg
```

```
## Number of states that get annual salary greater than average for Data Analysts: 25
```

```r
# For Business Analyst:

# Find maximum, minimum, and average annual salary
max_salary <- max(business_analyst$Annual_Salary)
min_salary <- min(business_analyst$Annual_Salary)
avg_salary4 <- mean(business_analyst$Annual_Salary)

# Find state with maximum salary
states_max_salary <- business_analyst$State[business_analyst$Annual_Salary == max_salary]

# Find state with minimum salary
states_min_salary <- business_analyst$State[business_analyst$Annual_Salary == min_salary]

# Find number of states with a salary greater than annual average
states_above_avg <- sum(business_analyst$Annual_Salary > avg_salary4)

# Show results
cat("Maximum Annual Salary:", max_salary, "in", states_max_salary, "\n")
```

```
## Maximum Annual Salary: 108915 in Washington
```

```r
cat("Minimum Annual Salary:", min_salary, "in", states_min_salary, "\n")
```

```
## Minimum Annual Salary: 71249 in Arkansas
```

```r
cat("Average Annual Salary:", avg_salary4, "\n")
```

```
## Average Annual Salary: 86956.96
```

```r
cat("Number of states that get annual salary greater than average for Business Analysts:", states_above
```

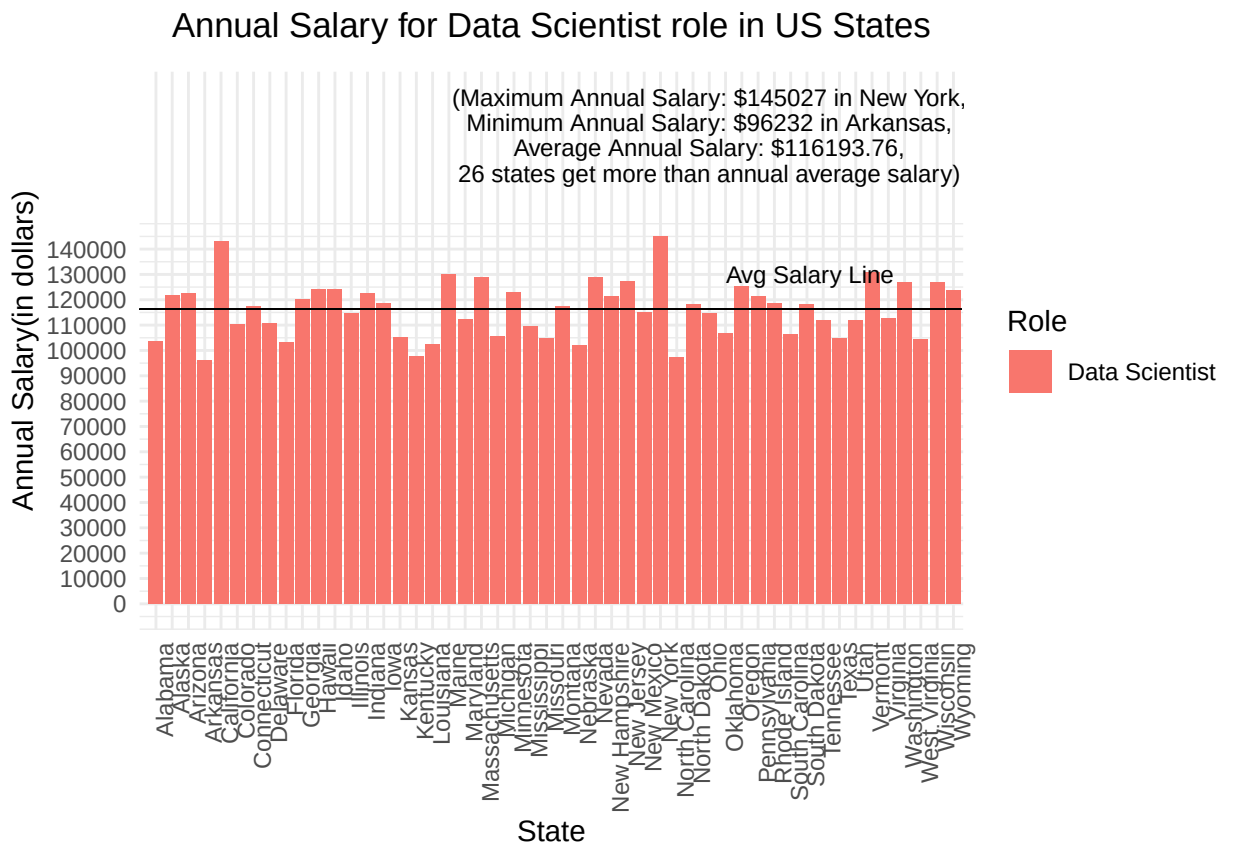
```
## Number of states that get annual salary greater than average for Business Analysts: 25
```

```r
# Create a dataframe for Data Scientist, Data Engineer, Data Analyst, and Business Analyst
role <- c("Data Scientist", "Data Engineer", "Data Analyst", "Business Analyst")
annual_avg_salary <- c(avg_salary1, avg_salary2, avg_salary3,avg_salary4)

df_new <- data.frame(Role = role, Annual_Avg_Salary = annual_avg_salary)
print(df_new)
```

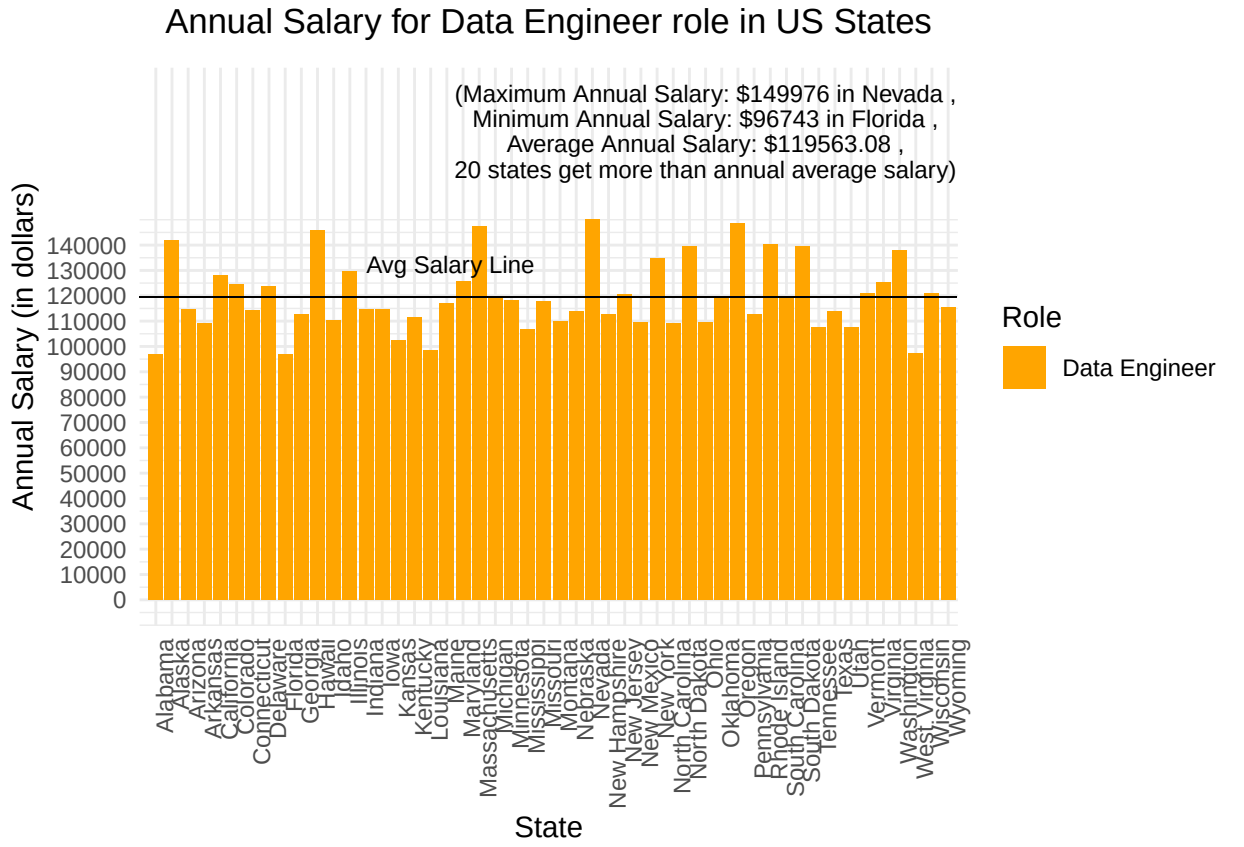
```
##               Role Annual_Avg_Salary
## 1   Data Scientist         116193.76
## 2    Data Engineer         119563.08
## 3     Data Analyst          75121.02
## 4 Business Analyst          86956.96
```

Data Visualizations:

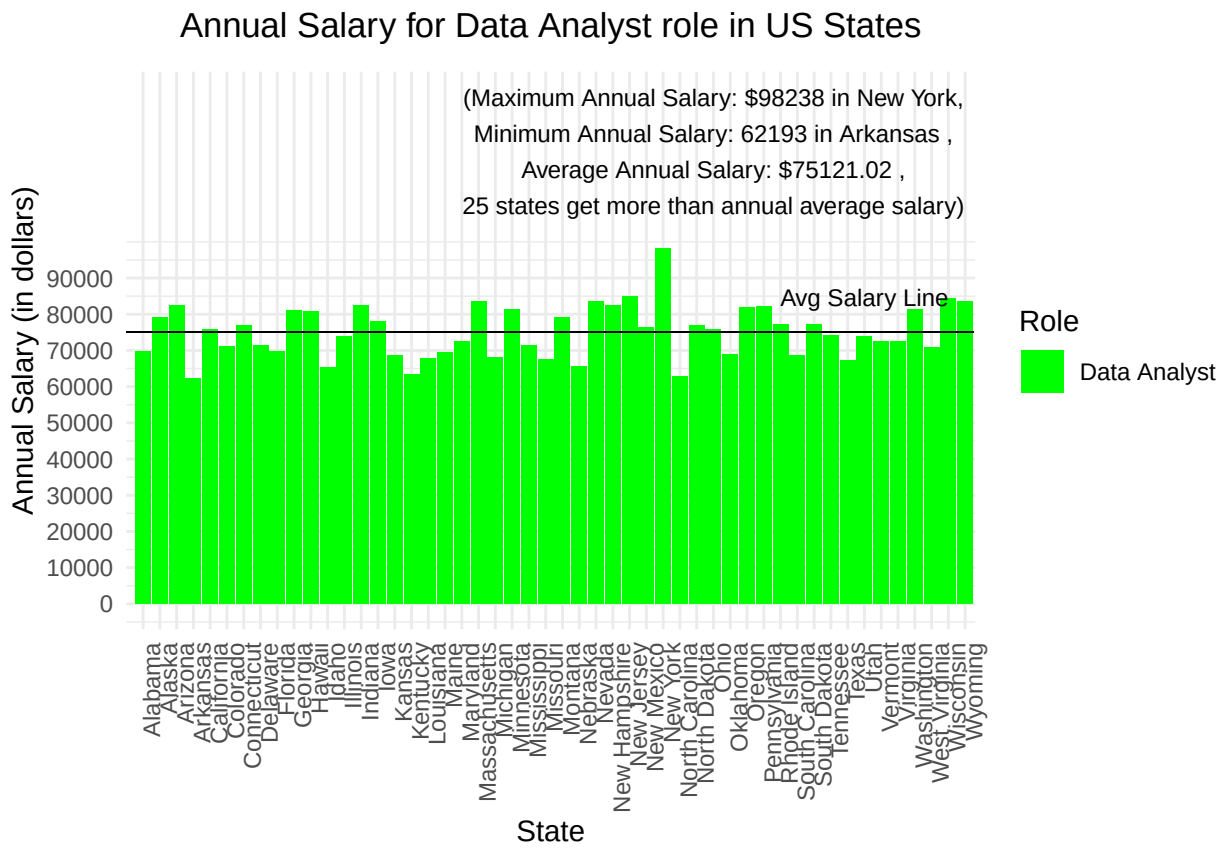
```
# Increase the height of the plot
#height <- 12  # Specify the height of the plot
#width <- 20  # Specify the width of the plot
#options(repr.plot.height=height, repr.plot.width=width)

# Create bar plot
p<-ggplot(data_scientist, aes(x = State, y = Annual_Salary, fill ="Data Scientist", label = State)) +
  geom_bar(stat = "identity") +
  #geom_text(position = position_stack(vjust =0.5),angle = 90, hjust =0.5) +
  labs(title = "Annual Salary for Data Scientist role in US States",
       x = "State", y = "Annual Salary(in dollars)", fill = "Role") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        plot.title = element_text(hjust = 0.5, margin = margin(b =10)))  # Adjust the margin)

# Set breaks for y-axis
p<-p + scale_y_continuous(breaks = seq(0, max(data_scientist$Annual_Salary), by=10000))

# Add a line for the average annual salary
p <- p + geom_segment(aes(x = 0, xend =50.5, y = avg_salary1, yend = avg_salary1), color = "black", lin

#Label average annual salary line
p <- p + annotate("text", x = 35, y = avg_salary1+3500, label = paste("Avg Salary Line"),
                  vjust = -1, hjust = -0.1, color = "black", size = 3)
# Add annotations
p + annotate("text", x=35, y = 200000, label = "(Maximum Annual Salary: $145027 in New York,", size=3.3
  annotate("text", x =35, y = 190000, label = "Minimum Annual Salary: $96232 in Arkansas,",size=3.3)+
  annotate("text",x=35,y=180000, label="Average Annual Salary: $116193.76,",size=3.3 )+
  annotate("text",x=35,y=170000, label="26 states get more than annual average salary)",size=3.3 )
```



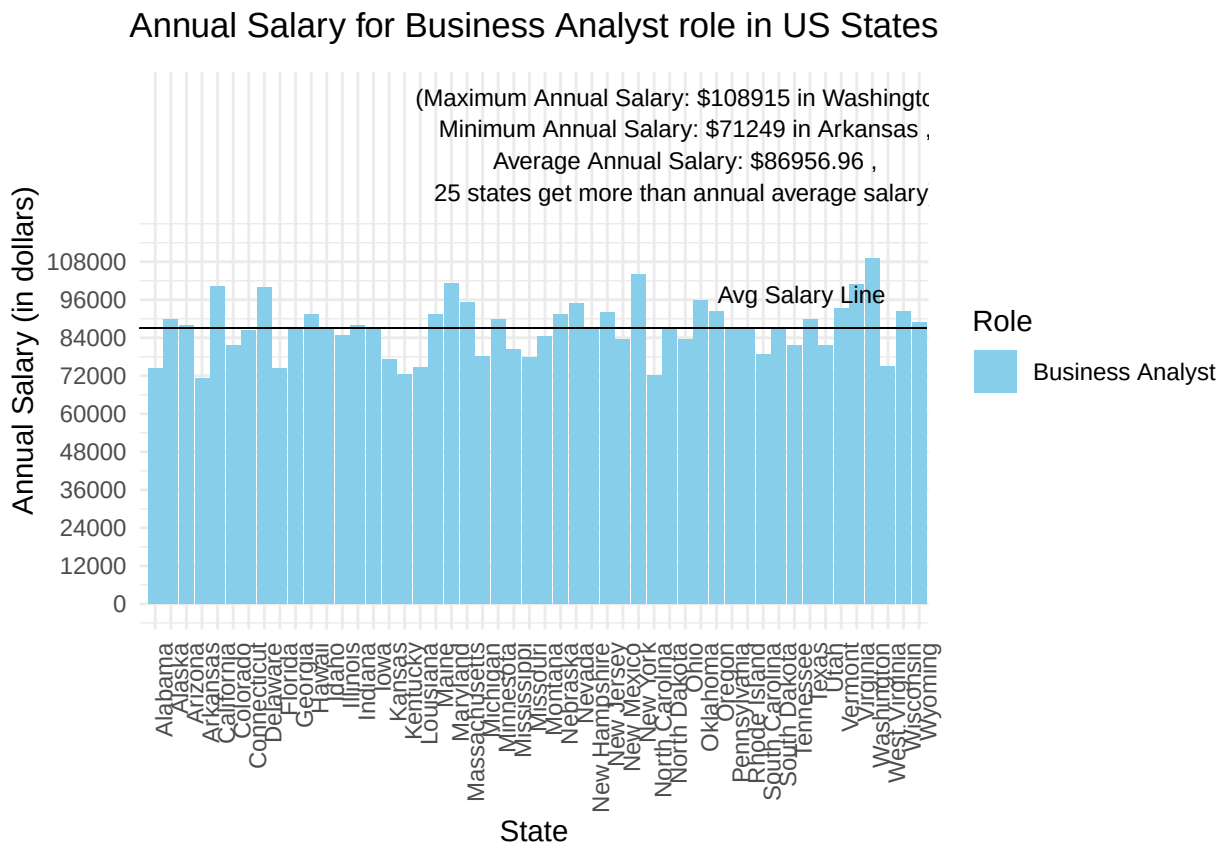

```r
# Create bar plot
p <- ggplot(data_engineer, aes(x = State, y = Annual_Salary, fill = "Data Engineer", label = State)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = "orange", name = "Role", labels = "Data Engineer") +
  labs(title = "Annual Salary for Data Engineer role in US States",
       x = "State", y = "Annual Salary (in dollars)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        plot.title = element_text(hjust = 0.5, margin = margin(b = 10)))  # Adjust margin
# Set breaks for y-axis
p<-p + scale_y_continuous(breaks = seq(0, max(data_engineer$Annual_Salary), by=10000))

# Add a line for the average annual salary
p <- p + geom_segment(aes(x = 0, xend =50.5, y = avg_salary2, yend = avg_salary2), color = "black", lin

#Label average annual salary line
p <- p + annotate("text", x =13, y = avg_salary2+2500, label = paste("Avg Salary Line"),
                  vjust = -1, hjust = -0.1, color = "black", size = 3)
# Add annotations
p + annotate("text", x=35, y = 200000, label = "(Maximum Annual Salary: $149976 in Nevada ,", size=3.3)+
  annotate("text", x =35, y = 190000, label = "Minimum Annual Salary: $96743 in Florida ,",size=3.3)+
  annotate("text",x=35,y=180000, label="Average Annual Salary: $119563.08 ,",size=3.3 )+
  annotate("text",x=35,y=170000, label="20 states get more than annual average salary)",size=3.3 )
```



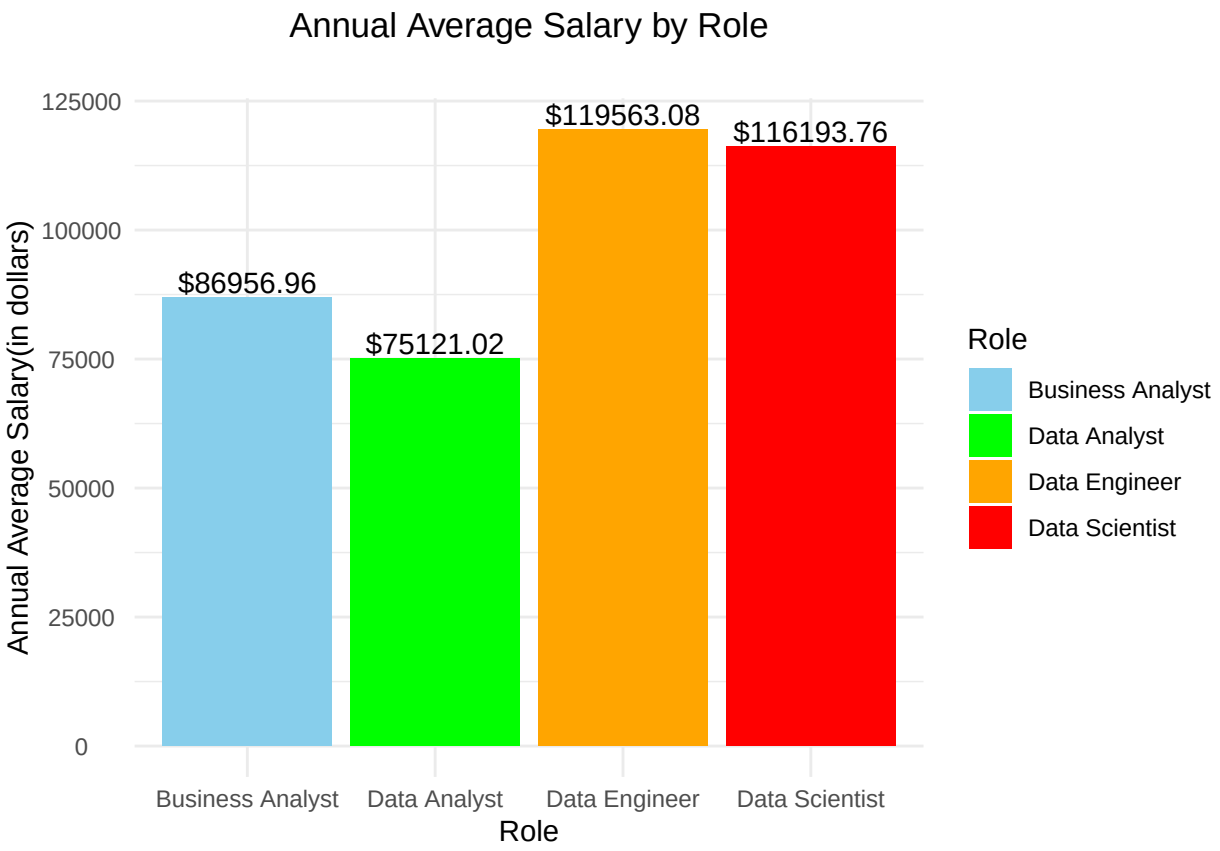

```r
# Create bar plot
p <- ggplot(data_analyst, aes(x = State, y = Annual_Salary, fill = "Data Analyst", label = State)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = "green", name = "Role", labels = "Data Analyst") +
  labs(title = "Annual Salary for Data Analyst role in US States",
       x = "State", y = "Annual Salary (in dollars)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        plot.title = element_text(hjust = 0.5, margin = margin(b = 10)))  # Adjust margin
# Set breaks for y-axis
p<-p + scale_y_continuous(breaks = seq(0, max(data_analyst$Annual_Salary), by=10000))

# Add a line for the average annual salary
p <- p + geom_segment(aes(x = 0, xend =50.5, y = avg_salary3, yend = avg_salary3), color = "black", lin

#Label average annual salary line
p <- p + annotate("text", x =38, y = avg_salary3+2500, label = paste("Avg Salary Line"),
                  vjust = -1, hjust = -0.1, color = "black", size = 3)
# Add annotations
p + annotate("text", x=35, y = 140000, label = "(Maximum Annual Salary: $98238 in New York,", size=3.3)+
  annotate("text", x =35, y = 130000, label = "Minimum Annual Salary: 62193 in Arkansas ,",size=3.3)+
  annotate("text",x=35,y=120000, label="Average Annual Salary: $75121.02 ,",size=3.3 )+
  annotate("text",x=35,y=110000, label="25 states get more than annual average salary)",size=3.3 )
```




```
# Create bar plot
p <- ggplot(business_analyst, aes(x = State, y = Annual_Salary, fill = "Business Analyst", label = State
  geom_bar(stat = "identity") +
  scale_fill_manual(values = "sky blue", name = "Role", labels = "Business Analyst") +
  labs(title = "Annual Salary for Business Analyst role in US States",
       x = "State", y = "Annual Salary (in dollars)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        plot.title = element_text(hjust = 0.5, margin = margin(b = 10)))  # Adjust margin
# Set breaks for y-axis
p<-p + scale_y_continuous(breaks = seq(0, max(business_analyst$Annual_Salary), by=12000))

# Add a line for the average annual salary
p <- p + geom_segment(aes(x = 0, xend =50.5, y = avg_salary4, yend = avg_salary4), color = "black", lin


#Label average annual salary line
p <- p + annotate("text", x =36, y = avg_salary4+2500, label = paste("Avg Salary Line"),
                  vjust = -1, hjust = -0.1, color = "black", size = 3)
# Add annotations
p + annotate("text", x=35, y = 160000, label = "(Maximum Annual Salary: $108915 in Washington,", size=3
  annotate("text", x =35, y = 150000, label = "Minimum Annual Salary: $71249 in Arkansas ,",size=3.3)+
  annotate("text",x=35,y=140000, label="Average Annual Salary: $86956.96 ,",size=3.3 )+
  annotate("text",x=35,y=130000, label="25 states get more than annual average salary)",size=3.3 )
```




```r
# Create a bar plot
p <- ggplot(df_new, aes(x = Role, y = Annual_Avg_Salary, fill = Role)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0("$", Annual_Avg_Salary)), position = position_dodge(width = 0.7), vjust =
  labs(title = "Annual Average Salary by Role",
       x = "Role", y = "Annual Average Salary(in dollars)",
       fill = "Role") +
  theme_minimal() +
  theme(axis.text.y = element_text(angle = 0, hjust = 0.5), plot.title = element_text(hjust = 0.5, margi
  scale_fill_manual(values = c("Data Scientist" = "red", "Data Engineer" = "orange", "Data Analyst" = "g

# Show plot
print(p)
```



Conclusion:

From the plots above, it is evident that: *Data Engineers receive the highest annual average salary among U.S. states.* Data Analysts receive the lowest annual average salary among U.S. states. *Data Scientists earn the highest annual salary of $145,027 in New York and the lowest of $96,232 in Arkansas.* Data Engineers earn the highest annual salary of $149,976 in Nevada and the lowest of $96,743 in Florida. *Data Analysts earn the highest annual salary of $98,238 in New York and the lowest of $62,193 in Arkansas.* Business Analysts earn the highest annual salary of $108,915 in Washington and the lowest of $71,249 in Arkansas. In this analysis the specific experience level for each role was unknown.

Data References:

1. Data Engineer: https://www.ziprecruiter.com/Salaries/What-Is-the-Average-DATA-Engineer-Salary-by-State
2. Data Analyst: https://www.ziprecruiter.com/Salaries/What-Is-the-Average-Data-Analyst-Salary-by-State
3. Data Scientist: https://www.ziprecruiter.com/Salaries/What-Is-the-Average-DATA-Scientist-Salary-by-State
4. Business Analyst: https://www.ziprecruiter.com/Salaries/What-Is-the-Average-Business-Analyst-Salary-by-State