Analyzing Data Science Salary R Notebook

Rusiru Pabasara

28/03/2024

Table of Contents

In this notebook, I will embark on a journey through a comprehensive data set of data science salaries, exploring various facets such as job titles, experience levels, and salary trends across different employment types. Our goal is to uncover valuable insights that can shed light on the current landscape of the data science job market.

OBJECTIVES

1. Identify the association between avg salary(USD) and most common jobs
2. Identify the impact of experience level to the salary
3. Identify association between remote ratio and salary(USD)
4. Identify association between salary and employee type
5. Identify company location and employee residence have a association

- First,Load the R libraries and load our data set from a CSV file named 'ds_salaries.csv' into a data frame called 'df'

```
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.0     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

```
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors

df <- read.csv("ds_salaries.csv", header = TRUE, sep = ",")
#load the CSV File

dim(df)

## [1] 3755   11

head(df)

##   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
##   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
##   company_location company_size
## 1               ES            L
## 2               US            S
## 3               US            S
## 4               CA            M
## 5               CA            M
## 6               US            L

sum(is.na(df))  #checking the missing values

## [1] 0

str(df)

## 'data.frame':    3755 obs. of  11 variables:
##  $ work_year         : int  2023 2023 2023 2023 2023 2023 2023 2023 2023
2023 ...
##  $ experience_level  : chr  "SE" "MI" "MI" "SE" ...
##  $ employment_type   : chr  "FT" "CT" "CT" "FT" ...
```

```
## $ job_title : chr "Principal Data Scientist" "ML Engineer" "ML
Engineer" "Data Scientist" ...
## $ salary : int 80000 30000 25500 175000 120000 222200 136000
219000 141000 147100 ...
## $ salary_currency : chr "EUR" "USD" "USD" "USD" ...
## $ salary_in_usd : int 85847 30000 25500 175000 120000 222200 136000
219000 141000 147100 ...
## $ employee_residence: chr "ES" "US" "US" "CA" ...
## $ remote_ratio : int 100 100 100 100 100 0 0 0 0 0 ...
## $ company_location : chr "ES" "US" "US" "CA" ...
## $ company_size : chr "L" "S" "S" "M" ...
```

DATA PREPROSSESING

- Then we have to categorized the job titles to specific main job titles and convert to factors.

eg:- Applied Data Analyst —> Data Analyst

```
#Categorized job titles to Main job title
categorize_title <-function(job_title){
  if (grepl('Data Scientist',job_title,ignore.case = TRUE)){
    return('Data Scientist')
  }else if(grepl('Data Analyst',job_title,ignore.case = TRUE)){
    return('Data Analyst')
  }else if(grepl('Applied Data Analyst',job_title,ignore.case = TRUE)){
    return('Data Analyst')
  }else if(grepl('Business Data Analyst',job_title,ignore.case = TRUE)){
    return('Data Analyst')
  }else if(grepl('BI Data Analyst',job_title,ignore.case = TRUE)){
    return('Data Analyst')
  }else if(grepl('Lead Data Analyst',job_title,ignore.case = TRUE)){
    return('Data Analyst')
  }else if(grepl('Applied Data Scientist',job_title,ignore.case = TRUE)){
    return('Data Scientist')
  }else if(grepl('Principal Data Scientist',job_title,ignore.case = TRUE)){
    return('Data Scientist')
  }else if(grepl('Lead Data Scientist',job_title,ignore.case = TRUE)){
    return('Data Scientist')
  }else if(grepl('Data Engineer',job_title,ignore.case = TRUE)){
    return('Data Engineer')
  }else if(grepl('ML Engineer',job_title,ignore.case = TRUE)){
    return('ML Engineer')
  }else if(grepl('Data Architect',job_title,ignore.case = TRUE)){
    return('Data Architect')
  }else if(grepl('Machine Learning Engineer',job_title,ignore.case = TRUE)){
    return('ML Engineer')
  }else if(grepl('Machine Learning Software Engineer',job_title,ignore.case =
TRUE)){
    return('ML Engineer')
  }else if(grepl('Applied Machine Learning Engineer',job_title,ignore.case =
```

```
TRUE)){
    return('ML Engineer')
  }else if(grepl('Analytics Engineer',job_title,ignore.case = TRUE)){
    return('Analytics Engineer')
  }else if(grepl('Research scientists',job_title,ignore.case = TRUE)){
    return('Reasearch scientists')
  }else{
    return('Other related jobs')
  }
}

df$Main_Title <- sapply(df$job_title,categorize_title)

#Convert to Factors
df$work_year <- as.factor(df$work_year)
df$job_title <- as.factor(df$job_title)
df$Main_Title <- as.factor(df$Main_Title)
df <- mutate(df, Main_Title = factor(Main_Title, levels =
names(sort(table(Main_Title))))) #reorder levels of factor ascending order of
frequency count
df$employment_type <- as.factor(df$employment_type)
df$experience_level <- as.factor(df$experience_level)
df$remote_ratio <- as.factor(df$remote_ratio)
df$company_size <- as.factor(df$company_size)
df$employee_residence<-as.factor(df$employee_residence)
df$company_location<-as.factor(df$company_location)
summary(df)

##  work_year   experience_level employment_type
job_title
##  2020:  76   EN: 320          CT:  10         Data Engineer
:1040
##  2021: 230   EX: 114          FL:  10         Data Scientist           :
840
##  2022:1664   MI: 805          FT:3718         Data Analyst             :
612
##  2023:1785   SE:2516          PT:  17         Machine Learning Engineer:
289
##                                                Analytics Engineer       :
103
##                                                Data Architect           :
101
##                                                (Other)                  :
770
##      salary          salary_currency    salary_in_usd    employee_residence
##  Min.   :    6000   Length:3755        Min.   :  5132   US      :3004
##  1st Qu.:  100000   Class :character   1st Qu.: 95000   GB      : 167
##  Median :  138000   Mode  :character   Median :135000   CA      :  85
##  Mean   :  190696                      Mean   :137570   ES      :  80
##  3rd Qu.:  180000                      3rd Qu.:175000   IN      :  71
```

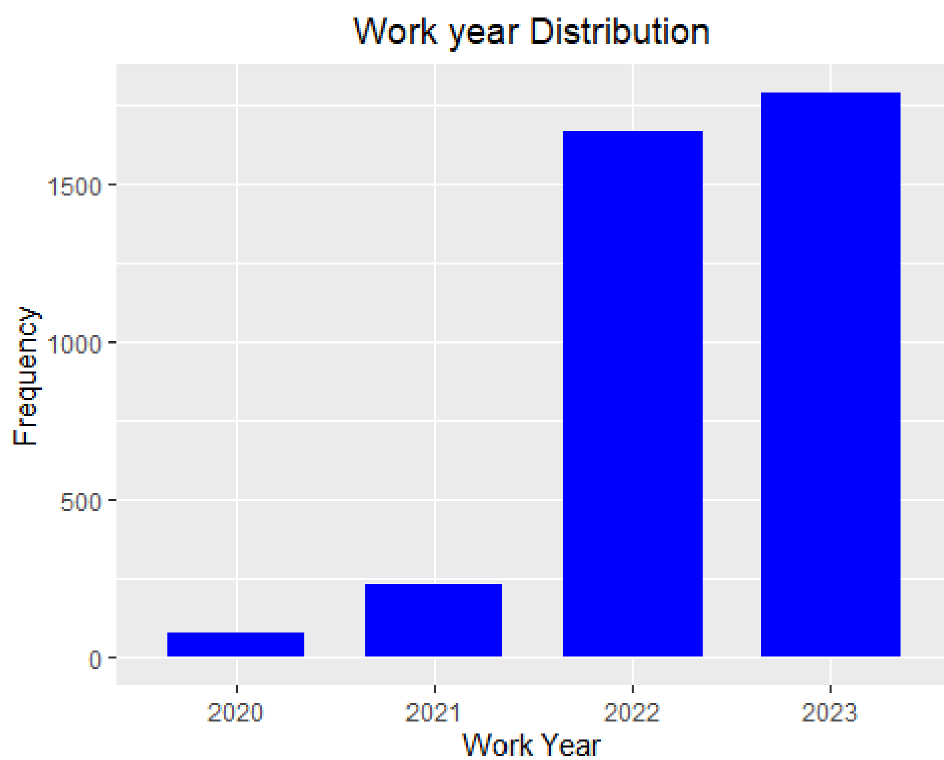
```
##  Max.   :30400000                      Max.   :450000   DE      :  48  
##                                                          (Other): 300  
##  remote_ratio company_location company_size              Main_Title  
##  0  :1923     US     :3040     L: 454       Data Architect     : 105  
##  50 : 189     GB     : 172     M:3153       Analytics Engineer: 109  
##  100:1643     CA     :  87     S: 148       ML Engineer        : 339  
##               ES     :  77                  Other related jobs: 602  
##               IN     :  58                  Data Analyst       : 662  
##               DE     :  56                  Data Scientist     : 871  
##               (Other): 265                  Data Engineer      :1067
```

- We checked missing values and converted to factors and did pre-processing part of our data set.Now let's look at the exploratory data analysis part.

EXPLORATARY ANALYSIS

UNIVARIATE ANALYSIS

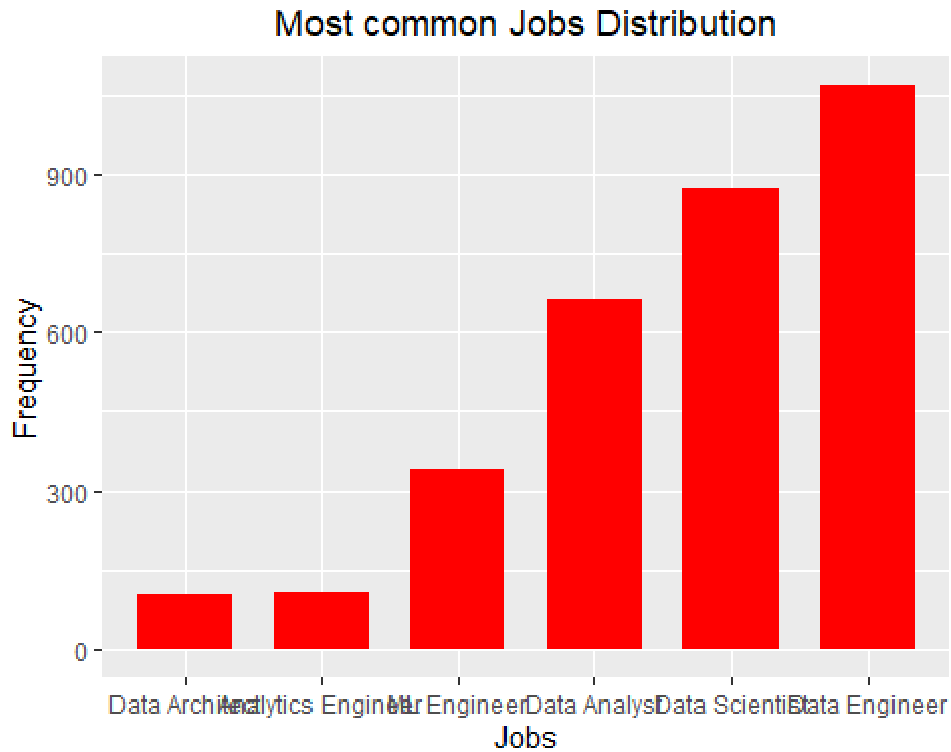
```
ggplot(df, aes(x=work_year))+
  geom_bar(fill="blue", width=0.7)+
  labs(x="Work Year" ,y="Frequency")+ggtitle("Work year Distribution")+
  theme(plot.title = element_text(hjust = 0.5))
```



* 2023 year has the highest employee responses and according to this graph there is a positive trend and after 2021 employees are highly increased

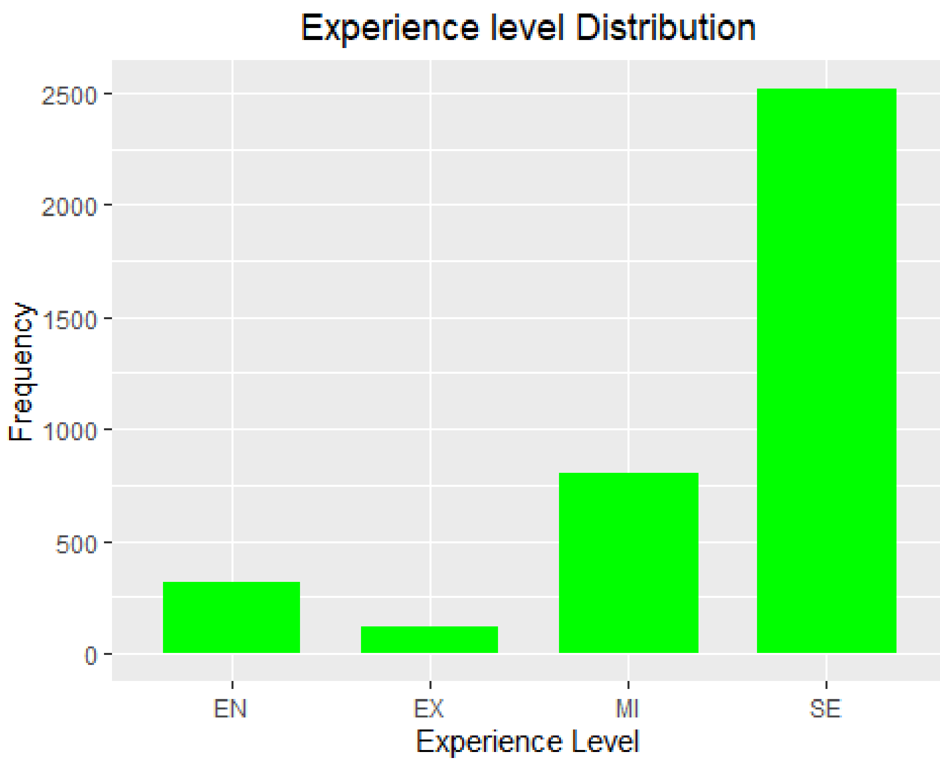
```
df1 <- filter(df, Main_Title != "Other related jobs")#filter out other jobs
and create new data frame with most common six job titles
```

```r
ggplot(df1,aes(x=Main_Title))+
  geom_bar(fill="red", width=0.7)+
  labs(x="Jobs", y="Frequency")+ggtitle("Most common Jobs Distribution")+
  theme(plot.title = element_text(hjust = 0.5))
```

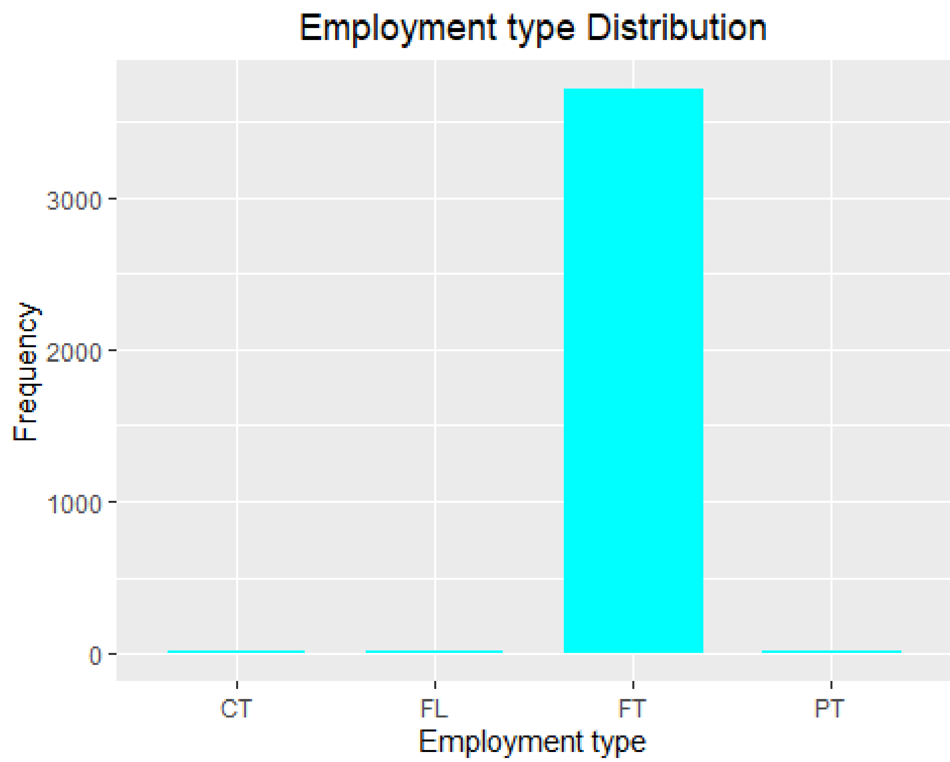


* These 6 are the most commonly used job vacancies. Data Engineering job field has highest responses since 2020 to 2023.according to this graph we can rank first 5 most common jobs

```r
ggplot(df,aes(experience_level))+
  geom_bar(fill="green",width = 0.7)+
  labs(x="Experience Level", y="Frequency")+ggtitle("Experience level
Distribution")+
  theme(plot.title = element_text(hjust = 0.5))
```

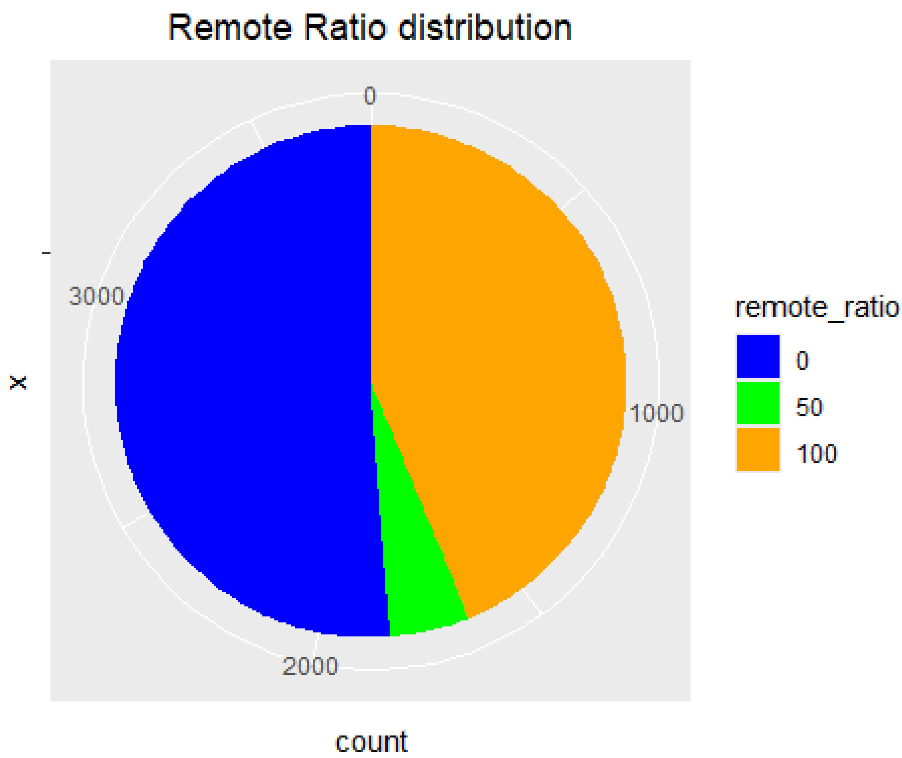



```r
ggplot(df,aes(employment_type))+
  geom_bar(fill="cyan",width = 0.7)+
  labs(x="Employment type", y="Frequency")+ggtitle("Employment type
Distribution")+
  theme(plot.title = element_text(hjust = 0.5))
```



* plot displays the frequency of responses by experience level. Clearly, Senior level of experience has the highest count and other plot displays the frequency of employee esponses by employment type. Clearly,employment_type of Full Time has the highest count.

```
ggplot(df,aes(x="",fill=remote_ratio))+geom_bar(width =
0.7)+coord_polar(theta = "y")+
  ggtitle("Remote Ratio distribution")+
  theme(plot.title = element_text(hjust=0.5))+ scale_fill_manual(values =
c("50"= "green","100"="orange","0"="blue"))
```

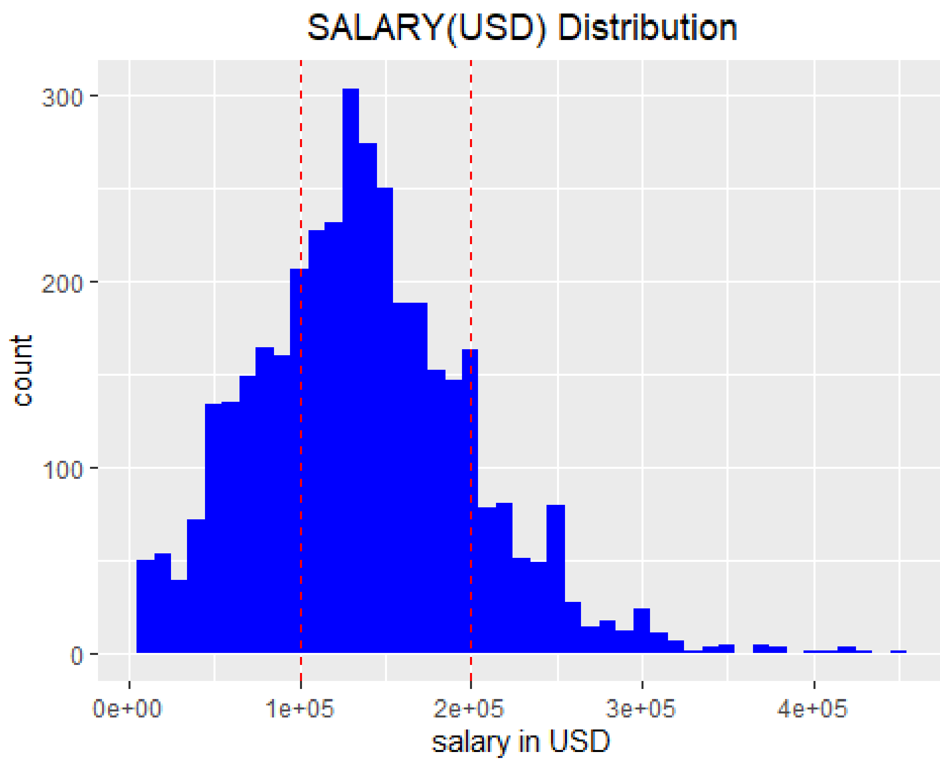


* This plot shows remote ratio of the employees responses

- 0 - Onsite
- 50 - Hybrid
- 100 - Online

Highest value is in onsite data science related jobs since 2020.

```
ggplot(df,aes(x=salary_in_usd))+
  geom_histogram(fill="blue",binwidth=10000)+geom_vline(xintercept = 100000,
color = "red", linetype = "dashed", linewidth = 0.5)+geom_vline(xintercept =
200000, color = "red", linetype = "dashed", linewidth = 0.5)+
  labs(x="salary in USD", Y=NULL)+ggtitle("SALARY(USD) Distribution")+
  theme(plot.title = element_text(hjust= 0.5))
```

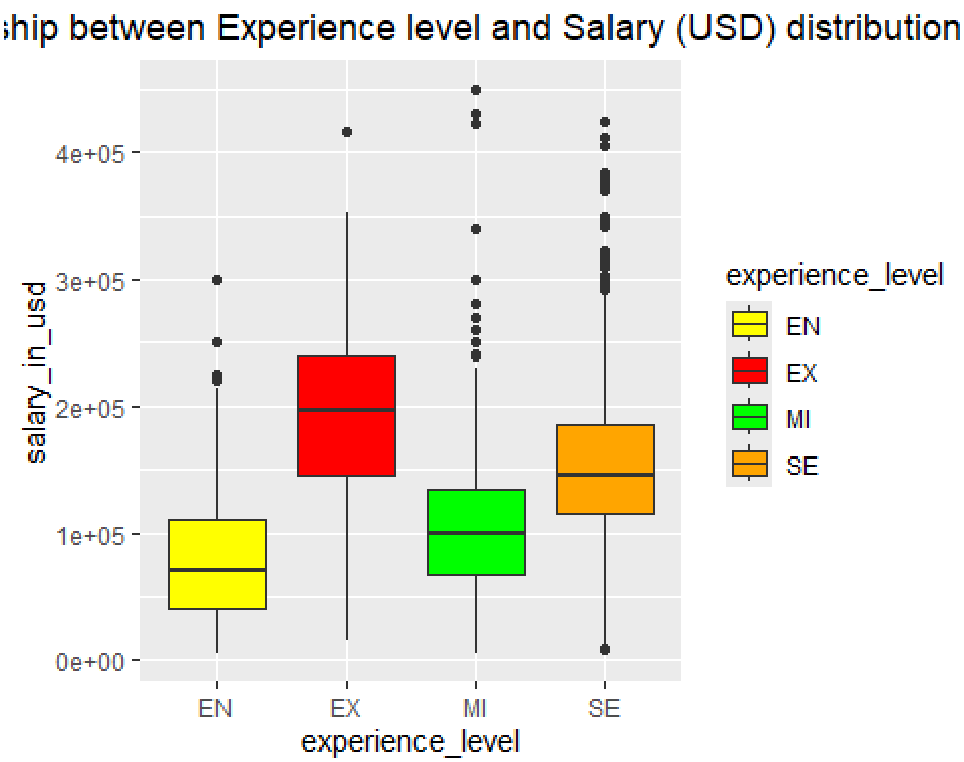


* The above histogram provides a visual representation of the salary distribution in USD, highlighting that there are more salary between 100000 and 200000 USD.

BIVARIATE ANALYSIS

Identify the impact of experience level to the salary

```
ggplot(df, aes(x = experience_level, y = salary_in_usd, fill =
experience_level)) +
  geom_boxplot() +
  scale_fill_manual(values = c("EX" = "red", "EN" = "yellow", "MI" = "green",
"SE" = "orange")) +
  ggtitle("Relationship between Experience level and Salary (USD)
distribution") +
  theme(plot.title = element_text(hjust = 0.5))
```



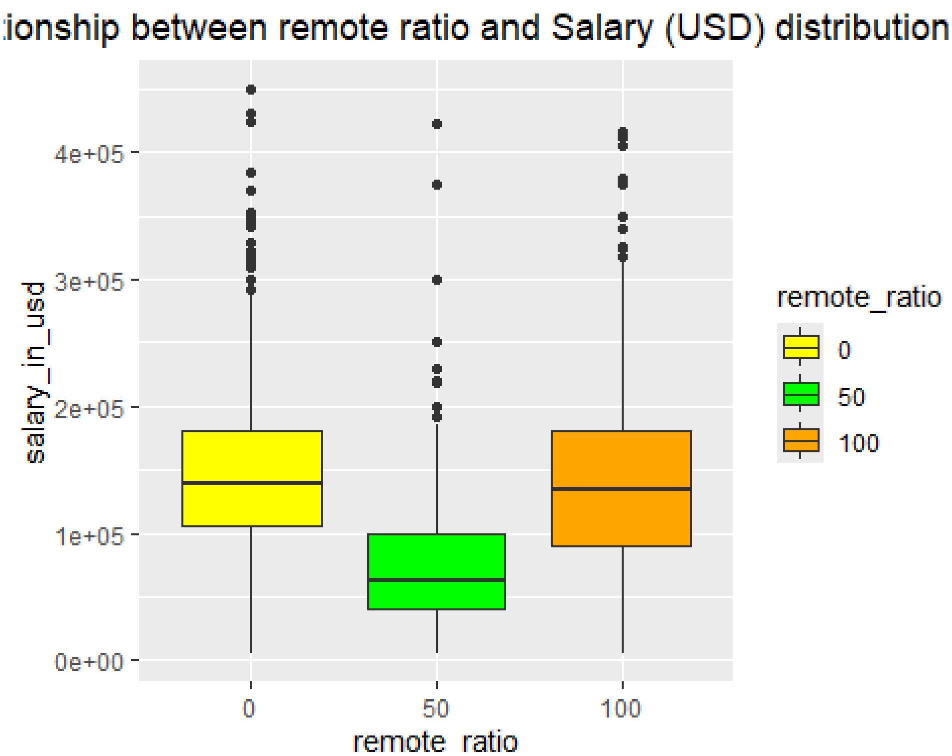
* The box plot visually contrasts salary distributions across experience levels.

- EN - Entry level
- EX - Executive level
- MI - Mid level
- SE - Senior level

The experience level with Executive level has higher salary compared to other experience levels. There are many outliers in Senior and Mid level which indicates salaries that significantly deviate from the rest of the distribution.

Identify association between remote ratio and salary(USD)

```
ggplot(df, aes(x = remote_ratio, y = salary_in_usd, fill = remote_ratio)) +
  geom_boxplot() +
  scale_fill_manual(values = c("0" = "yellow", "50" = "green", "100" =
"orange")) +
  ggtitle("Relationship between remote ratio and Salary (USD) distribution")
+
  theme(plot.title = element_text(hjust = 0.5))
```



* The box plot visually contrasts salary distributions across remote ratio.

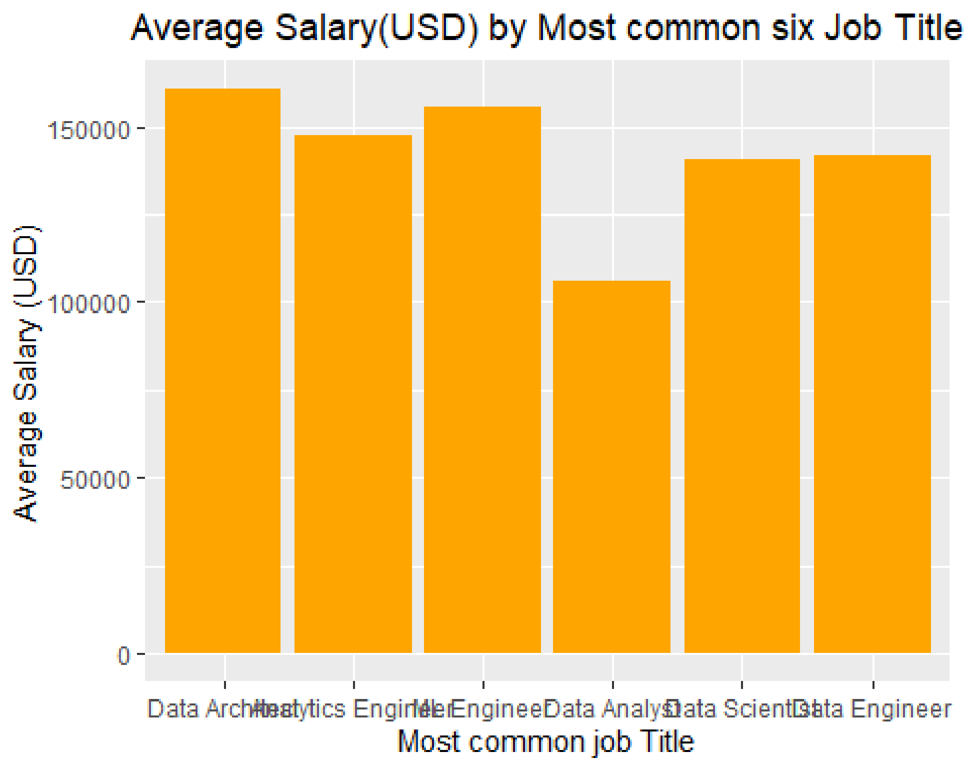
- 0 - onsite
- 50 - hybrid
- 100 - online

The onsite and online has higher salary compared to hybrid jobs. There are many outliers in onsite(0) and online(100).

Identify the association between avg salary(USD) and most common jobs

```
average_salary <- aggregate(salary_in_usd ~ Main_Title, data = df1, FUN =
mean)

ggplot(average_salary, aes(x = Main_Title, y = salary_in_usd)) +
  geom_bar(stat = "identity", fill = "orange") +
  labs(x = "Most common job Title", y = "Average Salary (USD)") +
  ggtitle("Average Salary(USD) by Most common six Job Title")+
theme(plot.title = element_text(hjust= 0.5))
```



- This bar plot shows average of salary(USD) of most common job fields,revealing Data Architect and ML Engineer has the highest average salaries while data analyst has lowest average salaries.

Identify association between salary and employee type

- H0 : No difference between mean salaries across employment type
- H1 : Difference between mean salaries across employment type

```
ANOVA1=aov(salary_in_usd~employment_type,data=df)
anova(ANOVA1)

## Analysis of Variance Table
##
## Response: salary_in_usd
##                   Df     Sum Sq    Mean Sq F value    Pr(>F)
## employment_type    3 2.4482e+11 8.1606e+10   20.85 2.151e-13 ***
## Residuals       3751 1.4681e+13 3.9139e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(ANOVA1)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = salary_in_usd ~ employment_type, data = df)
```

```
##
## $employment_type
##              diff        lwr        upr     p adj
## FL-CT -61639.10 -133548.45  10270.251 0.1225262
## FT-CT  24867.30  -26048.62  75783.223 0.5915888
## PT-CT -73913.19 -137993.97  -9832.414 0.0161228
## FT-FL  86506.40   35590.48 137422.323 0.0000764
## PT-FL -12274.09  -76354.87  51806.686 0.9608299
## PT-FT -98780.49 -137867.87 -59693.117 0.0000000
```

- The ANOVA table suggests that there is a STRONG significant difference in mean salaries across different employment types, as indicated by the very small p-value (< 0.001)
- FL-CT, FT-CT, and PT-FL pairs have adjusted p-values greater than 0.05, indicating no significant differences in mean salaries between these groups at 5% significance level.
- Therefore, we can say that employment type has a significant effect on salary, with certain pairs of employment types showing significant differences in mean salaries.

Identify company location and employee residence have an association.

- HO : No association between company location and employee residence
- H1 : Has a association between company location and employee residence

```r
contigency_table <-table(df$company_location,df$employee_residence)
fisher_exact <- fisher.test(contigency_table, simulate.p.value = TRUE)
print(fisher_exact)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  contigency_table
## p-value = 0.0004998
## alternative hypothesis: two.sided
```

- Since the p-value (0.0004998) is less than the significance level (α = 0.05), we would reject the null hypothesis and conclude that there is a significant association between residence and location.

Let’s look at the conclusions.

FINAL CONCLUSIONS

1. The distribution of salaries revealed that there are more salary paid range between 100000 and 200000 USD.
2. Distribution of salary(USD) of most common job fields,revealing Data Architect and ML Engineer has the highest average salaries while data analyst has lowest average salaries.
3. Most employees are in senior level and executive level has highest salary paid level but there are many high value outliers in senior level and mid level Therefore we cant say the relationship strong between them.
4. The onsite and online has higher salary compared to hybrid jobs. There are many outliers in onsite(0) and online(100).
5. Employment type has a significant effect on salary, with certain pairs of employment types(PT-CT, FT-FL, and PT-FT) showing significant differences in mean salaries.
6. there is a significant association between employee residence and company location.