Stat402- Presentation 1- EDA boxplots

Hannah Aguirre

2024-10-27

```r
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   3.5.1     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
```

```r
library(ggplot2)
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```r
df <- read.csv("~/Downloads/DataScience_salaries_2024.csv")
head(df)
```

```
##   work_year experience_level employment_type                  job_title
## 1      2021               MI              FT             Data Scientist
## 2      2021               MI              FT            BI Data Analyst
## 3      2020               MI              FT             Data Scientist
## 4      2021               MI              FT                ML Engineer
## 5      2022               SE              FT Lead Machine Learning Engineer
## 6      2021               MI              FT                ML Engineer
##     salary salary_currency salary_in_usd employee_residence remote_ratio
## 1 30400000             CLP         40038                 CL          100
```

```
## 2 11000000           HUF       36259              HU           50
## 3 11000000           HUF       35735              HU           50
## 4  8500000           JPY       77364              JP           50
## 5  7500000           INR       95386              IN           50
## 6  7000000           JPY       63711              JP           50
##   company_location company_size
## 1               CL            L
## 2               US            L
## 3               HU            L
## 4               JP            S
## 5               IN            L
## 6               JP            S
```

```
sum(complete.cases(df)) == nrow(df) #Check for NA Values TRUE means there is none
```

```
## [1] TRUE
```

```
df <- unique(df) #Get rid of duplicate observations
```

```
library(tidyverse)
table(df$employment_type)
```

```
##
##   CT   FL   FT   PT
##   26   13 9061   27
```

```
df <- filter(df,employment_type == 'FT') # We will focus on full time salaries
```

```
table(df$work_year) #Too few observations in 2020-2022 so we'll combine them into "Pandemic Era"
```

```
##
## 2020 2021 2022 2023 2024
##   69  206 1099 4616 3071
```

```
df$work_year <- ifelse(df$work_year == 2024, "2024",
                ifelse(df$work_year == 2023, "2023", "Pandemic"))
df$work_year <- factor(df$work_year,levels = c("Pandemic","2023","2024"), ordered = TRUE)
#Relevel
```

```
table(df$experience_level)
```

```
##
##   EN   EX   MI   SE
##  862  353 2445 5401
```

```
df$experience_level <- factor(df$experience_level, levels = c("EN","MI","SE","EX"), ordered = T) #Relev
```

```r
df <- select(df,-c("employment_type","salary","salary_currency","employee_residence"))
#Get rid of employment type as every observation is full time
#Get rid of salary and currency as we have the salary in USD
#Get rid of employee residene as we already have the company location and the company location
#matters more for predicting the salary
```

```r
table(df$remote_ratio)
```

```
##
##    0   50  100
## 5670  233 3158
```

```r
table(df$company_size)
```

```
##
##    L    M    S
##  608 8293  160
```

```r
df$company_size <- factor(df$company_size,levels = c("S","M","L"), ordered = TRUE)
```

```r
df$remote_ratio <- car::recode(df$remote_ratio,"0='In-Person';50='Hybrid';100='Remote'")
df$remote_ratio <- factor(df$remote_ratio, levels = c("In-Person","Hybrid","Remote"))
```

```r
library(countrycode)
df$company_location <- countrycode(df$company_location, origin = "iso2c", destination = "country.name")
table(df$company_location)
```

```
##
##           American Samoa                  Andorra                Argentina
##                        3                        1                        7
##                  Armenia                Australia                  Austria
##                        1                       48                       10
##                  Belgium       Bosnia & Herzegovina                   Brazil
##                        4                        2                       21
##                   Canada Central African Republic                    Chile
##                      348                        2                        1
##                    China                 Colombia                  Croatia
##                        1                       14                        3
##                  Czechia                  Denmark                  Ecuador
##                        2                        3                        1
##                    Egypt                  Estonia                  Finland
##                       11                       10                        4
##                   France                  Germany                    Ghana
##                       59                       93                        2
##                Gibraltar                   Greece                 Honduras
##                        1                       11                        1
##      Hong Kong SAR China                  Hungary                    India
##                        1                        3                       57
##                Indonesia                     Iraq                  Ireland
##                        2                        1                       12
```

```
##                  Israel                    Italy                    Japan
##                       3                       13                        8
##                   Kenya                   Latvia                  Lebanon
##                       2                       14                        2
##               Lithuania               Luxembourg                 Malaysia
##                      16                        2                        1
##                   Malta                Mauritius                   Mexico
##                       3                        1                       15
##                 Moldova              Netherlands              New Zealand
##                       1                       26                        5
##                 Nigeria                   Norway                     Oman
##                       7                        2                        1
##                Pakistan              Philippines                   Poland
##                       2                        6                       14
##                Portugal              Puerto Rico                    Qatar
##                      28                        4                        1
##                 Romania                   Russia             Saudi Arabia
##                       4                        7                        3
##               Singapore                 Slovenia             South Africa
##                       5                        6                       13
##             South Korea                    Spain                   Sweden
##                       2                       70                        3
##             Switzerland                 Thailand                   Turkey
##                       9                        3                        6
##                 Ukraine     United Arab Emirates           United Kingdom
##                       9                        4                      514
##           United States                  Vietnam
##                    7483                        3
```

```r
developed_countries <- c("Andorra","Australia","Austria","Belgium","Canada","Croatia","Czechia","Denmar

df$company_location <- ifelse(df$company_location == "United States", "US",
                       ifelse(df$company_location %in% developed_countries,"First World","Developing"))
#Combine all first world countries into developed label leaving the rest as the developing countries
df$company_location <- factor(df$company_location, levels = c("US","First World","Developing"))
```

From this table, we can see that our data came from all over the globe. While the diversity is great, there are way too many levels that also have too little observations which suggests we should relevel everything. We will combine all the countries with booming economies as First World Countries, all the third world countries as developing, and leaving the USA as its own category to have a baseline.

```r
titles <- names(table(df$job_title))
analyst <- c(titles[c(1,40,44:51,67,75,78,93,103:104,109,126,134,139,148)])
scientist <- c(titles[c(13,79:89,99,106,111,125,128,137,149)])
mle <- c(titles[c(14:15,107,112:123,129:131,138)])
engineer <- c(titles[c(20,27,38:39,54:56,58,69,72,76,94,110,127,136,147,150)])
BI <- c(titles[c(21:24,28:36,133)])
df$job_title <- ifelse(df$job_title %in% analyst, "DA",
                ifelse(df$job_title %in% scientist,"DS",
                ifelse(df$job_title %in% engineer, "DE",
                ifelse(df$job_title %in% mle,"ML",
                ifelse(df$job_title %in% BI,"BI","Other")))))
df$job_title <- relevel(factor(df$job_title), ref = "DA")
#Condense all titles with data science in the name to data scientist and repeat for analysts, MLE and e
```

```r
head(df)
```

```
##   work_year experience_level job_title salary_in_usd remote_ratio
## 1  Pandemic               MI        DS         40038       Remote
## 2  Pandemic               MI        BI         36259       Hybrid
## 3  Pandemic               MI        DS         35735       Hybrid
## 4  Pandemic               MI        ML         77364       Hybrid
## 5  Pandemic               SE        ML         95386       Hybrid
## 6  Pandemic               MI        ML         63711       Hybrid
##   company_location company_size
## 1       Developing            L
## 2               US            L
## 3       Developing            L
## 4      First World            S
## 5       Developing            L
## 6      First World            S
```

Codebook

With our cleaned data, let's summarize all the information in our data.

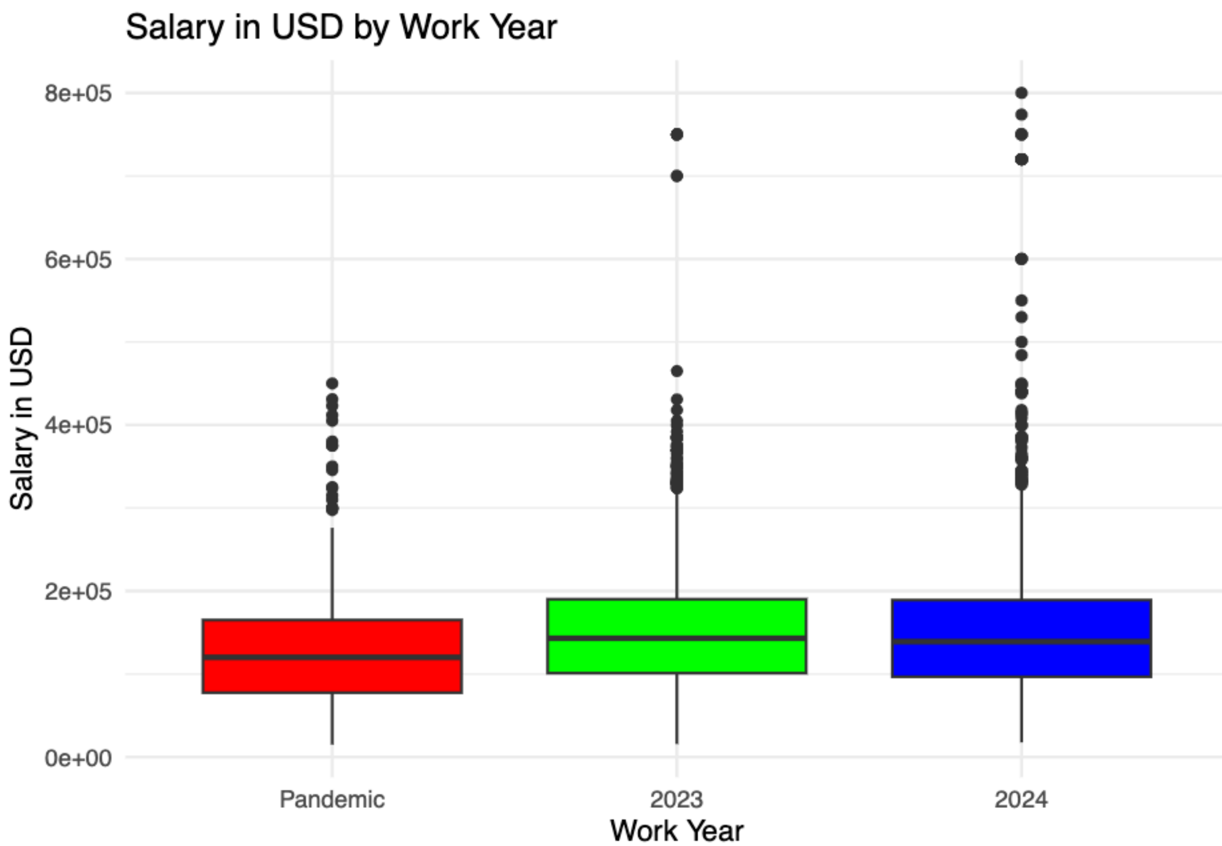
work_year:Describes what year the salary came from. Ordinal variable with 3 levels. "Pandemic", "2023", "2024"
experience_level: Describes level of experience for the role. Ordinal variable with 4 levels. "EN"=Entry Level "MI"=Mid-Level "SE"=Senior "EX"=Executive
job_title: Describes the job title and what kind of work is done. Nominal variable with 6 levels. "DA"=Data Analyst "BI"=Business Intelligence "DE"=Data Engineer "DS"=Data Science "ML"=Machine Learning "Other"
salary_in_usd:The salary in USD which is a quantitative variable
remote_ratio: Describes the format of the job. Nominal variable with 3 levels. "In-Person","Hybrid","Remote"
company_location:Describes what kind of country this job came from. Nominal variable with 3 levels. "US" "First World" "Developing"
company_size: Describes the size of the company. Ordinal variable with 3 levels. "S"=Small "M"=Medium "L" = Large

```r
#Table of values in each level of work_year
work_year_table <- table(df$work_year)
work_year_table
```

```
##
## Pandemic     2023     2024
##     1374     4616     3071
```

```r
# Create the side-by-side boxplots for work_year
ggplot(df, aes(x = work_year, y = salary_in_usd, fill = work_year)) +
  geom_boxplot() +
  labs(title = "Salary in USD by Work Year",
       x = "Work Year",
       y = "Salary in USD") +
  scale_fill_manual(values = c("Pandemic" = "red", "2023" = "green", "2024" = "blue")) +
```

```r
  theme_minimal() +
  theme(legend.position = "none") # Hide the legend as fill corresponds to x-axis categories
```

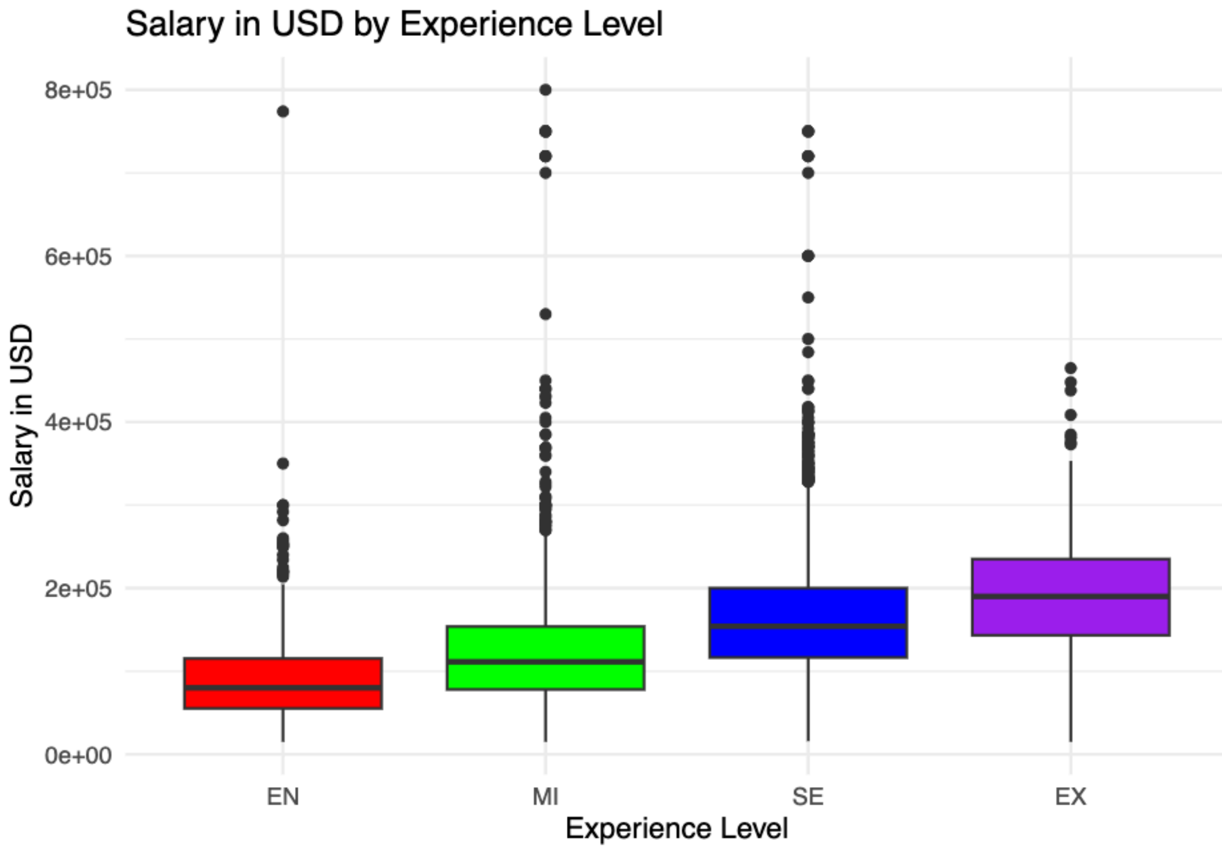


In the side-by-side boxplots of Salary by Work Year, we can see that all three factors have high outliers and are skewed right, with 2024 having the most outliers and strongest skew. Might consider tranforming (logarithmic?). There is variability (IQR) for all three factors is similar.

```r
#Table of values in each level of experience_level
experience_level_table <- table(df$experience_level)
experience_level_table
```

```
##
##   EN   MI   SE   EX
##  862 2445 5401  353
```

```r
# Create the boxplot side-by-side boxplots for experience_level
ggplot(df, aes(x = experience_level, y = salary_in_usd, fill = experience_level)) +
  geom_boxplot() +
  labs(title = "Salary in USD by Experience Level",
       x = "Experience Level",
       y = "Salary in USD") +
  scale_fill_manual(values = c("EN" = "red", "MI" = "green", "SE" = "blue", "EX" = "purple")) +
  theme_minimal() +
  theme(legend.position = "none") # Hide legend as fill is mapped to x-axis categories
```

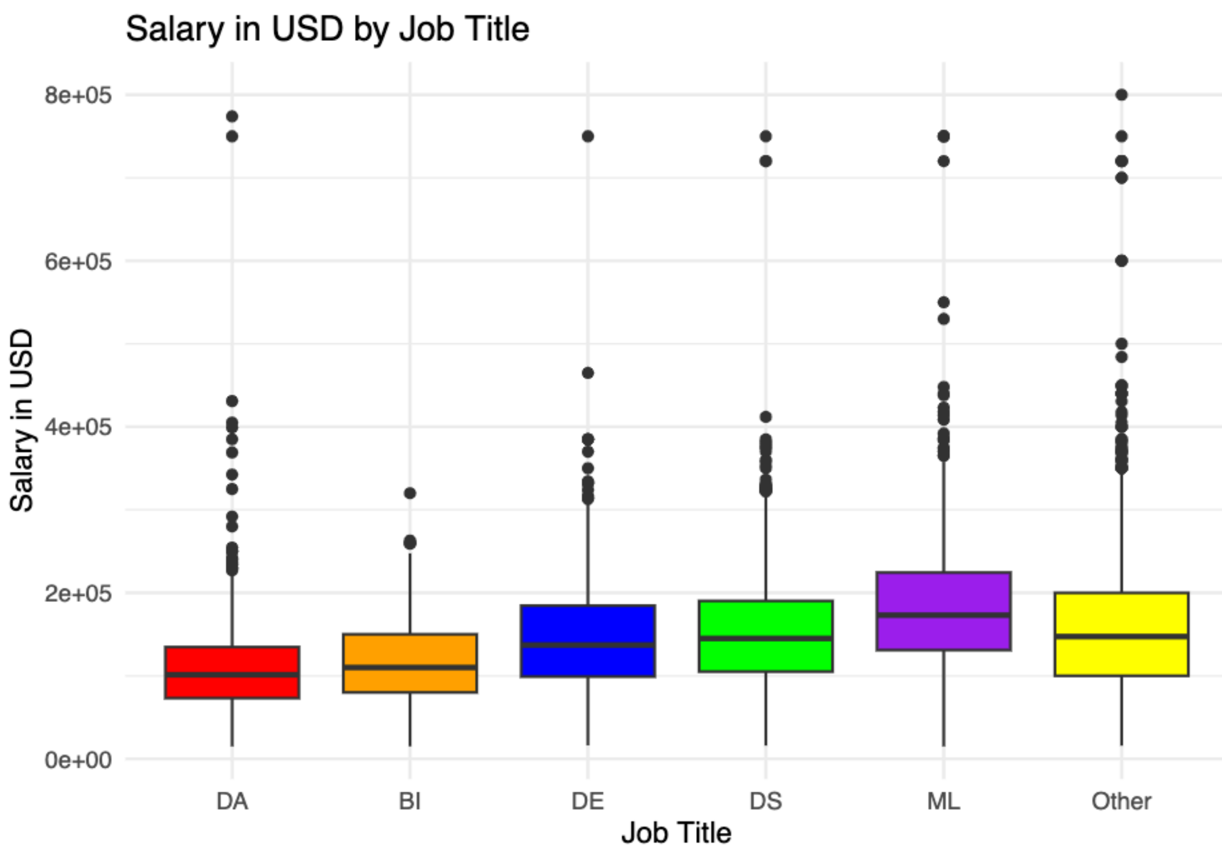


In the side-by-side boxplots of Salary by Experience Level, we can see that all factors have high outliers and are skewed right, with Mid and Senior levels having the most outliers and strongest skew. Might consider tranforming (or looking at subset of data, under 3-400K?). The pattern is predictable that as experience level increase so do the median/ majority of salaries.

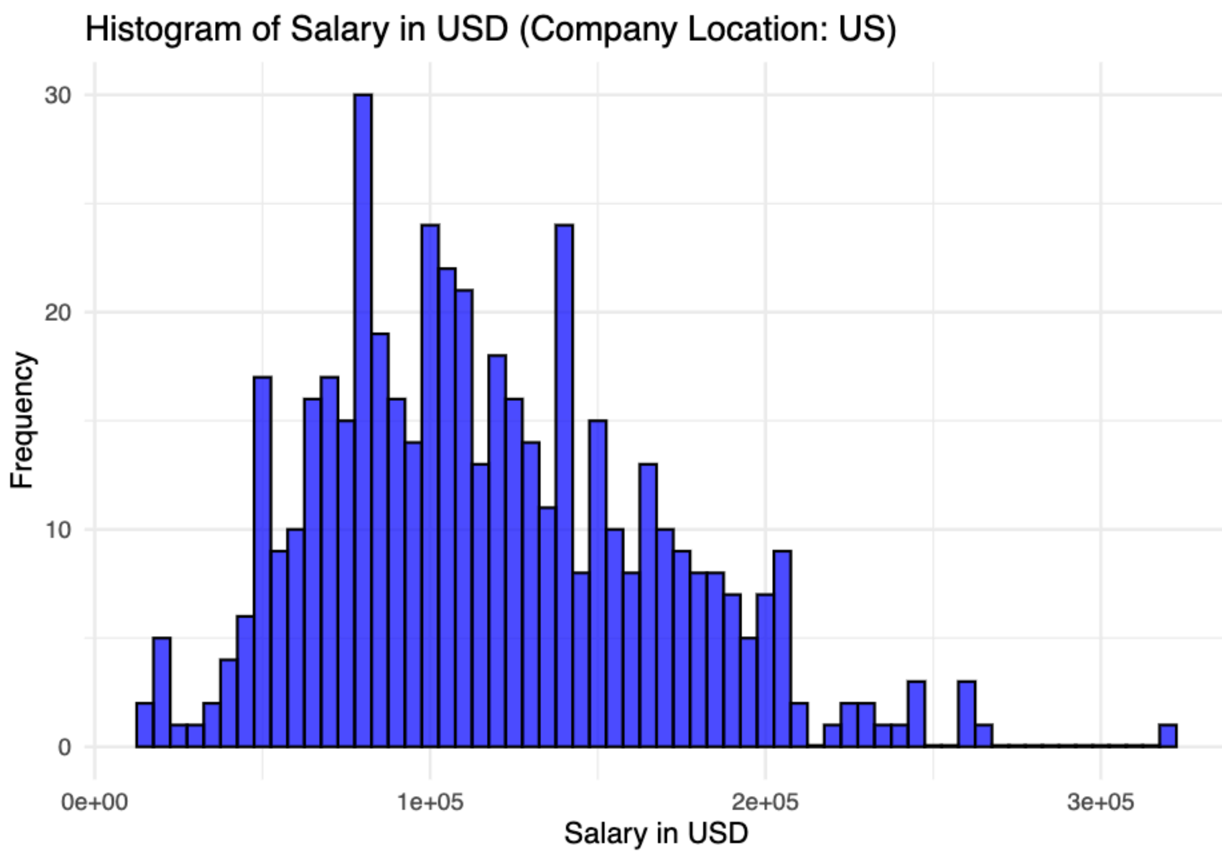
```
#Table of values in each level of Job_title
job_title_table <- table(df$job_title)
job_title_table
```

```
##
##    DA    BI    DE    DS    ML Other
##  1388   481  1802  2130  1250  2010
```

```
# Create the boxplot side-by-side boxplots for job_title
ggplot(df, aes(x = job_title, y = salary_in_usd, fill = job_title)) +
  geom_boxplot() +
  labs(title = "Salary in USD by Job Title",
       x = "Job Title",
       y = "Salary in USD") +
  scale_fill_manual(values = c("DA" = "red", "DS" = "green", "DE" = "blue", "ML" = "purple", "BI" = "or
  theme_minimal() +
  theme(legend.position = "none") # Hide legend as fill is mapped to x-axis categories
```

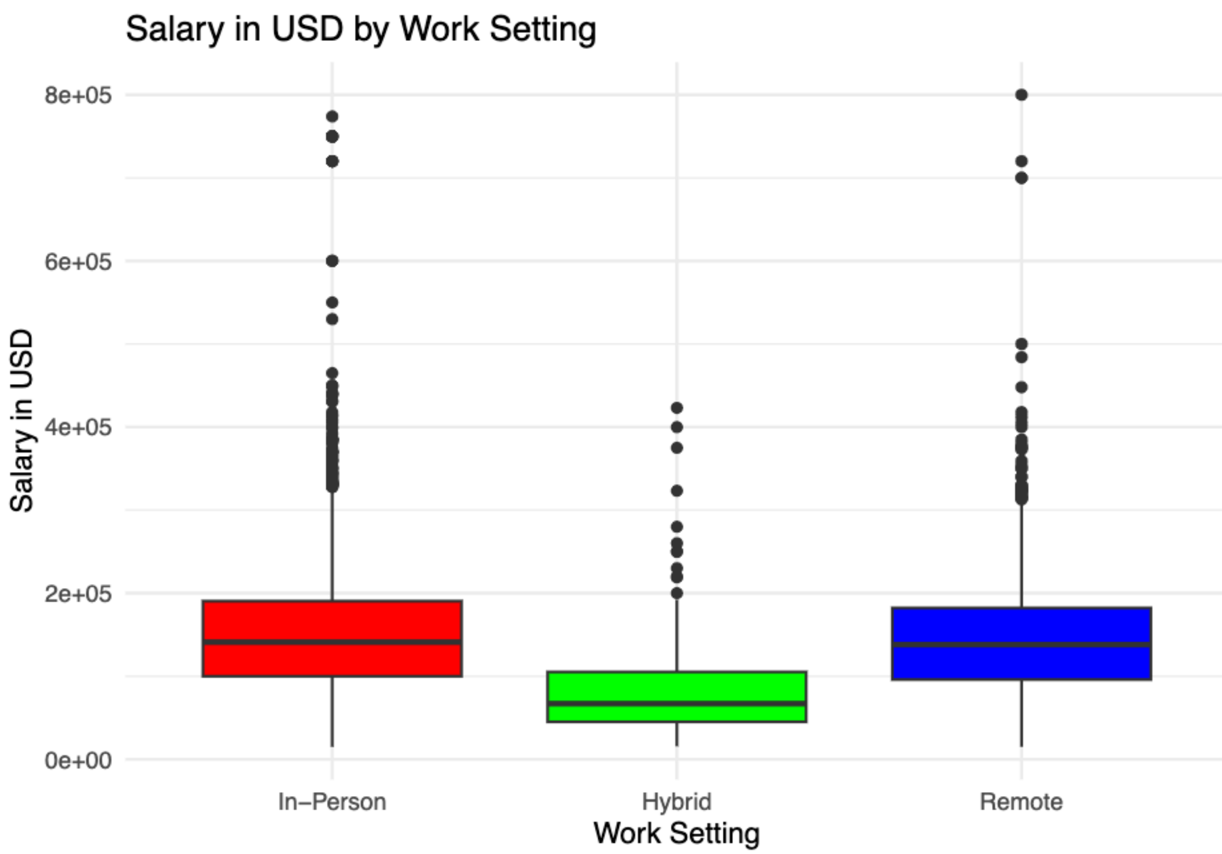



```r
# Create histogram of salary_in_usd for job_title "BI"
ggplot(df %>% filter(job_title == "BI"), aes(x = salary_in_usd)) +
  geom_histogram(binwidth = 5000, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Salary in USD (Company Location: US)",
       x = "Salary in USD",
       y = "Frequency") +
  theme_minimal()
```



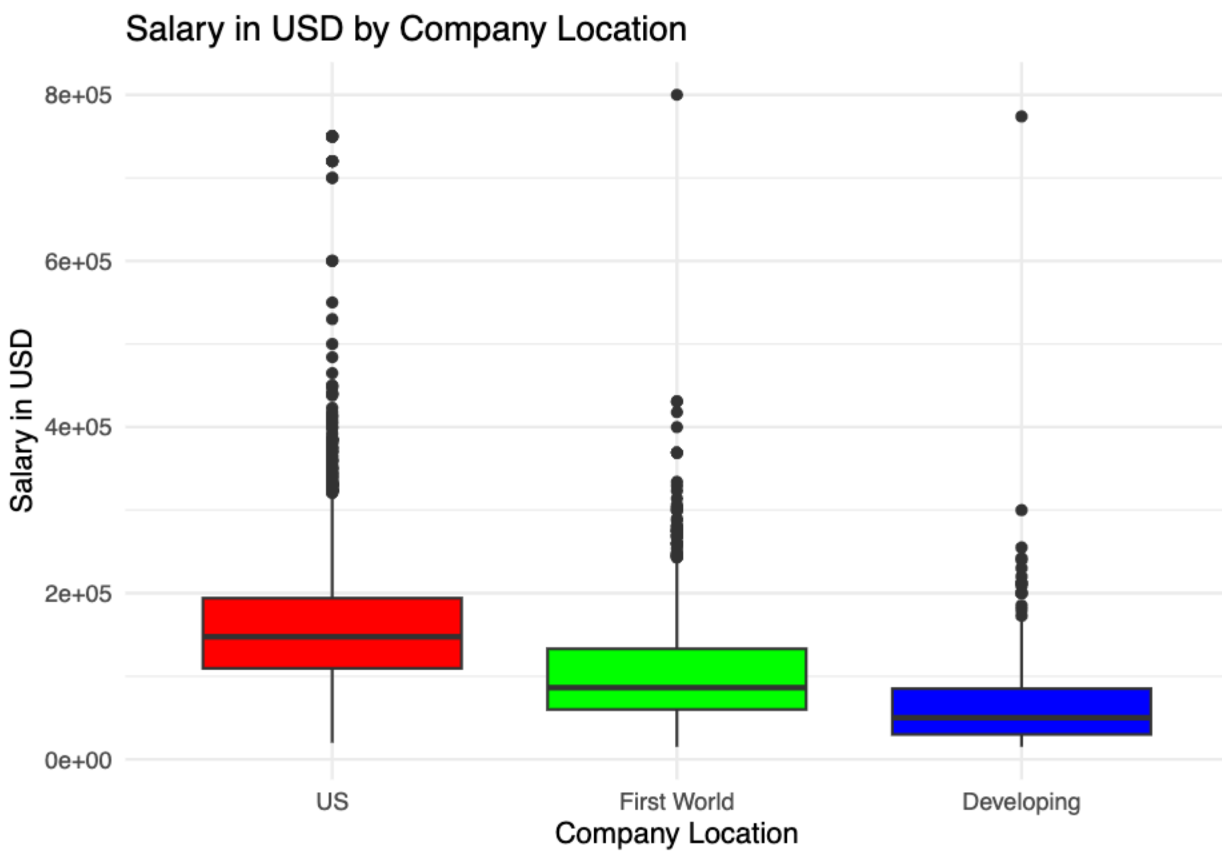
In the side-by-side boxplots of Salary by Job Title, we can see that all factors have high outliers and are skewed right, except BI which has the least amount of outliers and slightly symmetric. Might consider tranforming (or looking at subset of data, under 3-400K?).

```r
# Create the boxplot side-by-side boxplots for remote_ratio
ggplot(df, aes(x = remote_ratio, y = salary_in_usd, fill = remote_ratio)) +
  geom_boxplot() +
  labs(title = "Salary in USD by Work Setting",
       x = "Work Setting",
       y = "Salary in USD") +
  scale_fill_manual(values = c("In-Person" = "red", "Hybrid" = "green", "Remote" = "blue")) +
  theme_minimal() +
  theme(legend.position = "none") # Hide legend as fill is mapped to x-axis categories
```

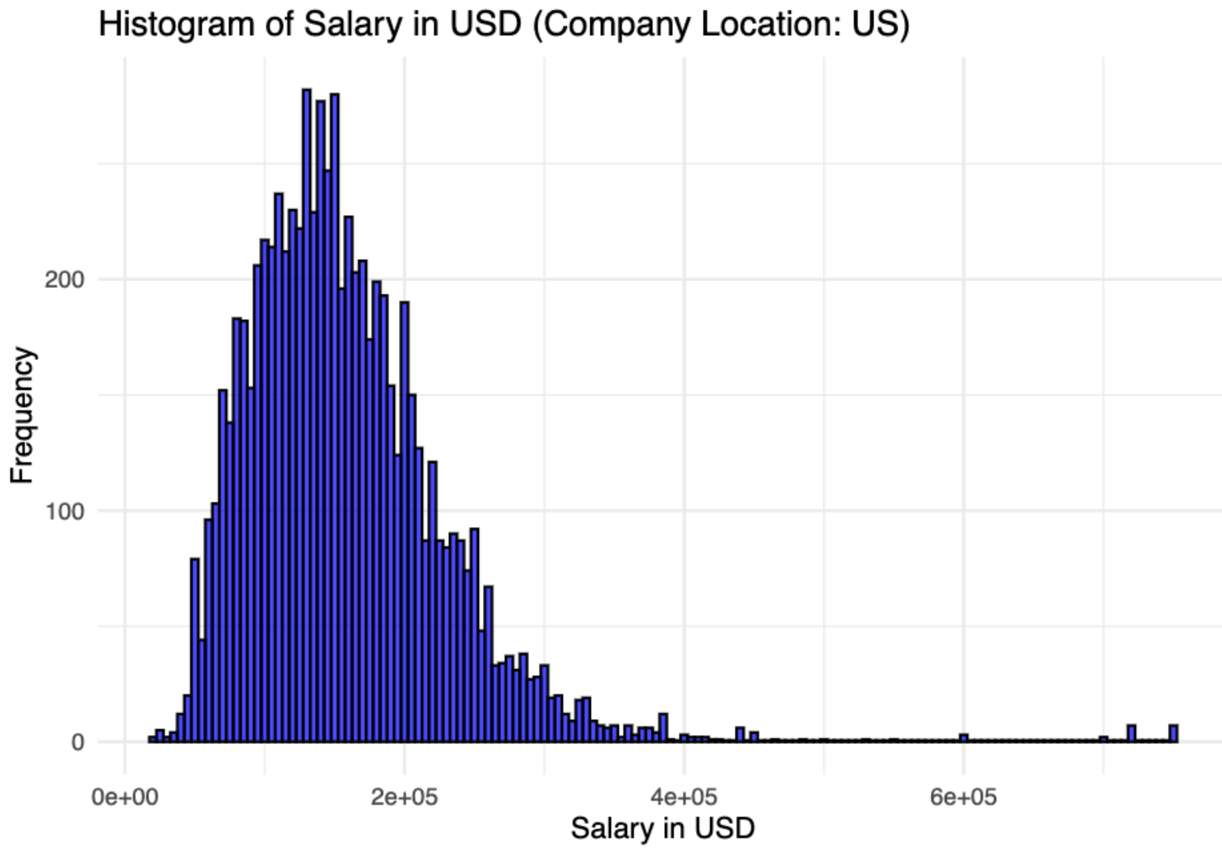


In the side-by-side boxplots of Salary by Work Setting, we can see that all factors have high outliers and are skewed right, with In-Person and Remote making significantly more than the "Hybrid"setting.

```r
# Create the boxplot side-by-side boxplots for company_location
ggplot(df, aes(x = company_location, y = salary_in_usd, fill = company_location)) +
  geom_boxplot() +
  labs(title = "Salary in USD by Company Location",
       x = "Company Location",
       y = "Salary in USD") +
  scale_fill_manual(values = c("US" = "red", "First World" = "green", "Developing" = "blue")) +
  theme_minimal() +
  theme(legend.position = "none") # Hide legend as fill is mapped to x-axis categories
```

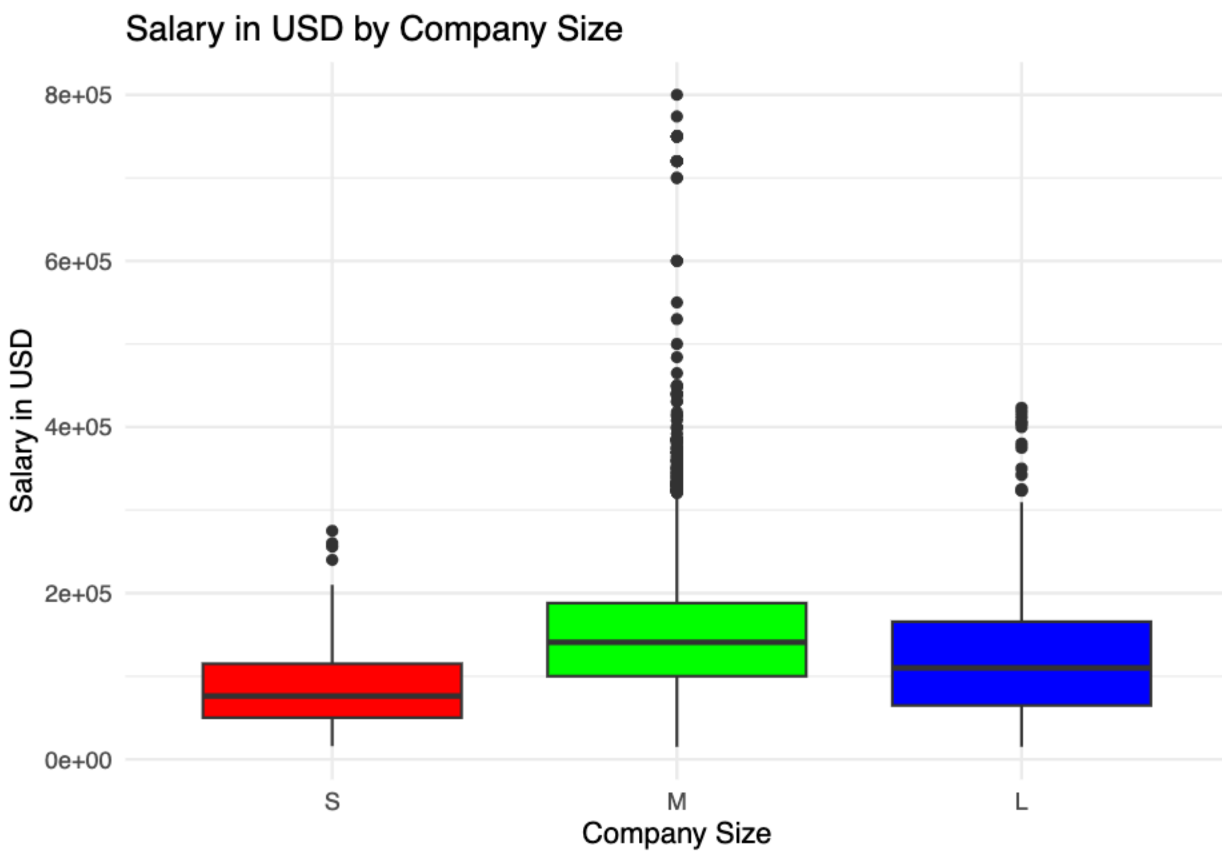



```r
# Create histogram of salary_in_usd for company_location "US"
ggplot(df %>% filter(company_location == "US"), aes(x = salary_in_usd)) +
  geom_histogram(binwidth = 5000, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Salary in USD (Company Location: US)",
       x = "Salary in USD",
       y = "Frequency") +
  theme_minimal()
```



In the side-by-side boxplots of Salary by Company Location, we can see that all factors have high outliers and are skewed right. With the US making the most on average, followed by First world, then developing countries. Might consider tranforming (or looking at subset of data, under 3-400K?). Or should we look at just US data?

```
# Create the boxplot side-by-side boxplots for company_size
ggplot(df, aes(x = company_size, y = salary_in_usd, fill = company_size)) +
  geom_boxplot() +
  labs(title = "Salary in USD by Company Size",
       x = "Company Size",
       y = "Salary in USD") +
  scale_fill_manual(values = c("S" = "red", "M" = "green", "L" = "blue")) +
  theme_minimal() +
  theme(legend.position = "none") # Hide legend as fill is mapped to x-axis categories
```



In the side-by-side boxplots of Salary by Company size, we can see that all factors have high outliers and Medium size companies have the strongest right skew and most outliers, followed by Large then small size companies. Might consider tranforming (or looking at subset of data, under 3-400K?).

```r
# Create a two-way table
two_way_table <- table(df$job_title, df$company_location)
two_way_table
```

```
##
##           US First World Developing
##   DA    1148         204         36
##   BI     394          71         16
##   DE    1478         280         44
##   DS    1746         329         55
##   ML    1017         194         39
##   Other 1700         261         49
```

```r
# Convert to proportions
proportional_table <- prop.table(two_way_table)

# Print the proportional table
print(proportional_table)
```

```
##
##                   US First World  Developing
##   DA    0.126696833 0.022514071 0.003973071
```

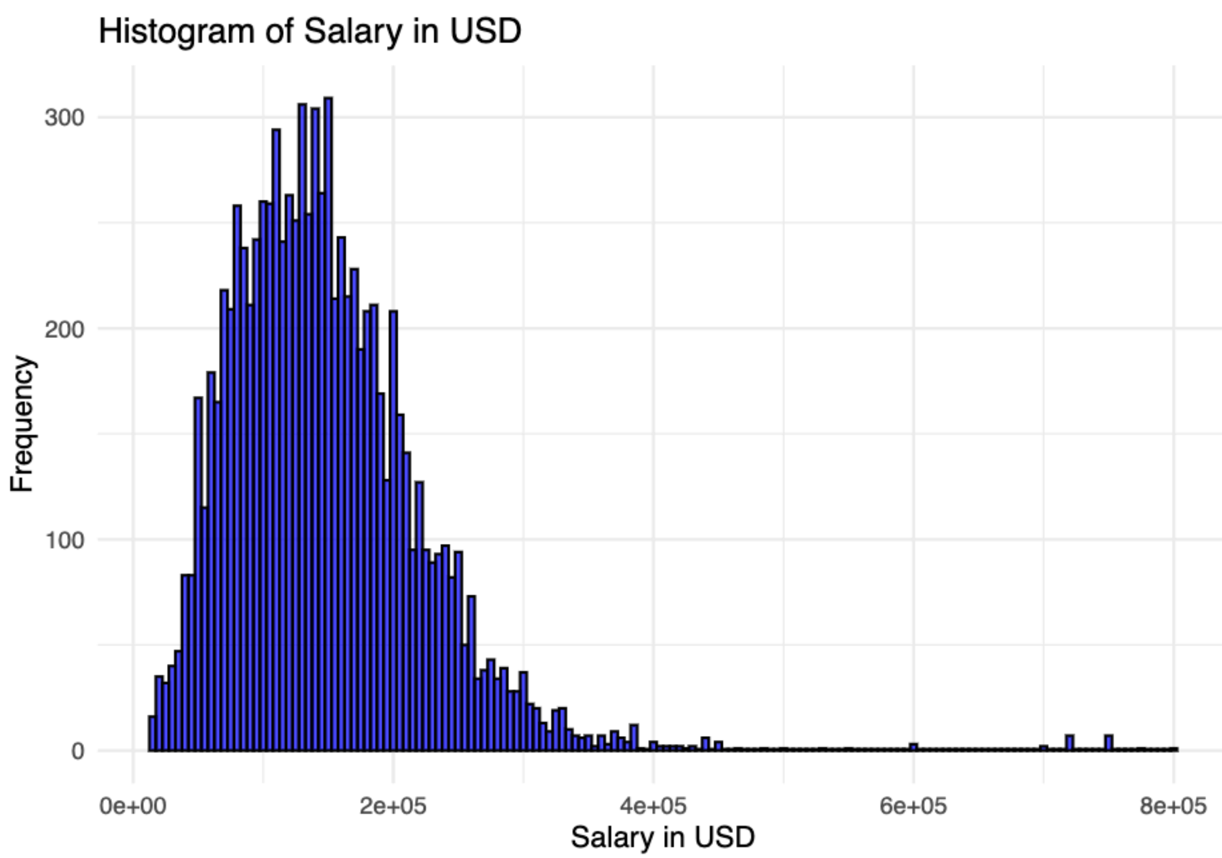
```
##   BI    0.043483059 0.007835780 0.001765810
##   DE    0.163116654 0.030901666 0.004855976
##   DS    0.192693963 0.036309458 0.006069970
##   ML    0.112239267 0.021410440 0.004304161
##   Other 0.187617261 0.028804768 0.005407792
```

```
total <- colSums(proportional_table)
total
```

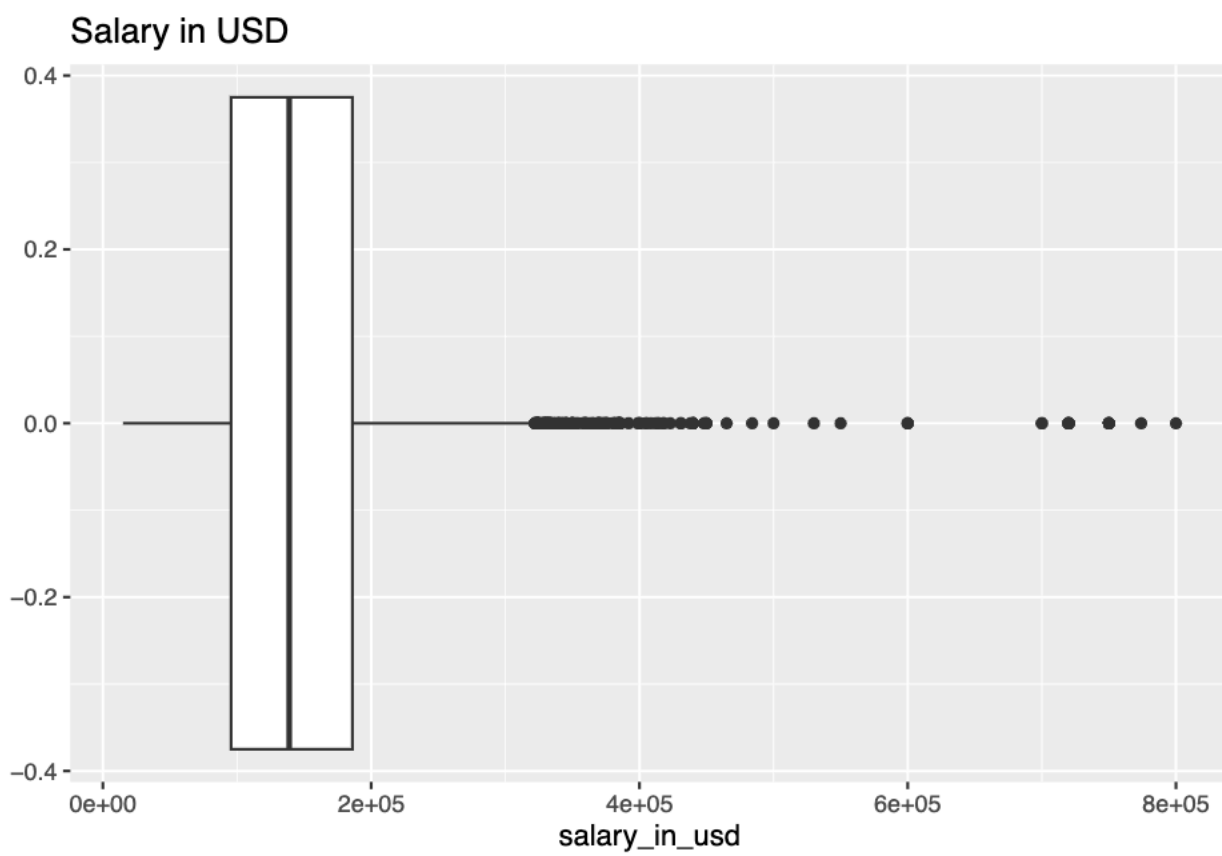
```
##          US First World  Developing
##  0.82584704  0.14777618  0.02637678
```

We can see a majority of the data comes from the US. I am wondering how much the context changes outside of the US, would it be worth it to make predictions based on data only from the US?

```
# Create histogram of salary_in_usd
ggplot(df, aes(x = salary_in_usd)) +
  geom_histogram(binwidth = 5000, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Salary in USD",
       x = "Salary in USD",
       y = "Frequency") +
  theme_minimal()
```



```r
# Create the boxplot for salary_in_usd
ggplot(df, aes(salary_in_usd)) +
  geom_boxplot() +
  labs(title = "Salary in USD")
```

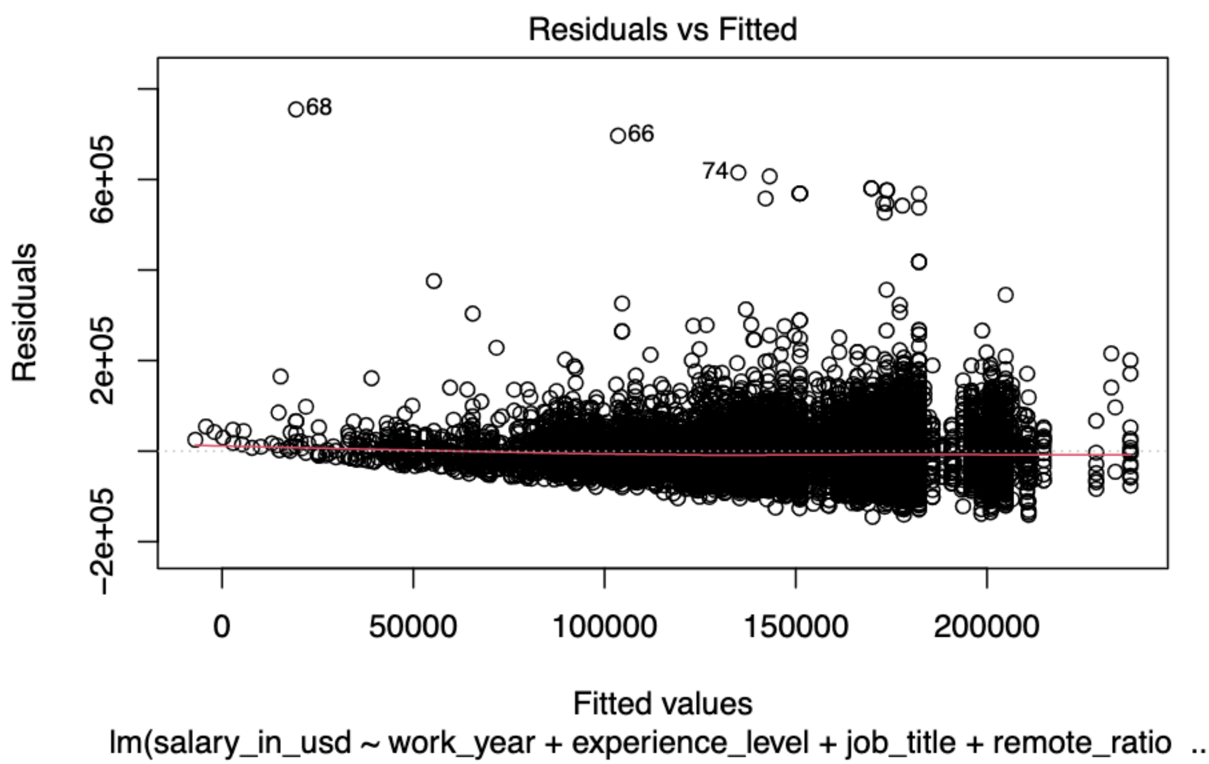


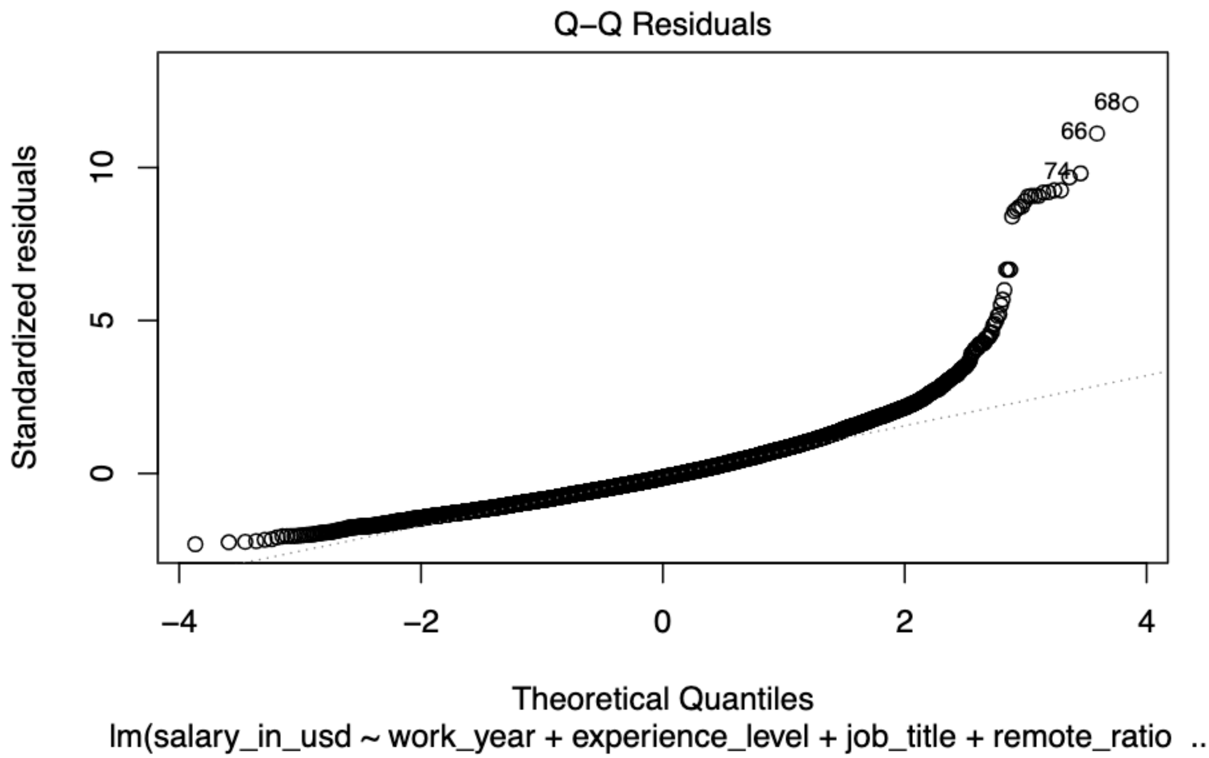

```r
#MLR model
model1 <- lm(salary_in_usd ~ work_year + experience_level + job_title + remote_ratio + company_location
summary(model1)
```

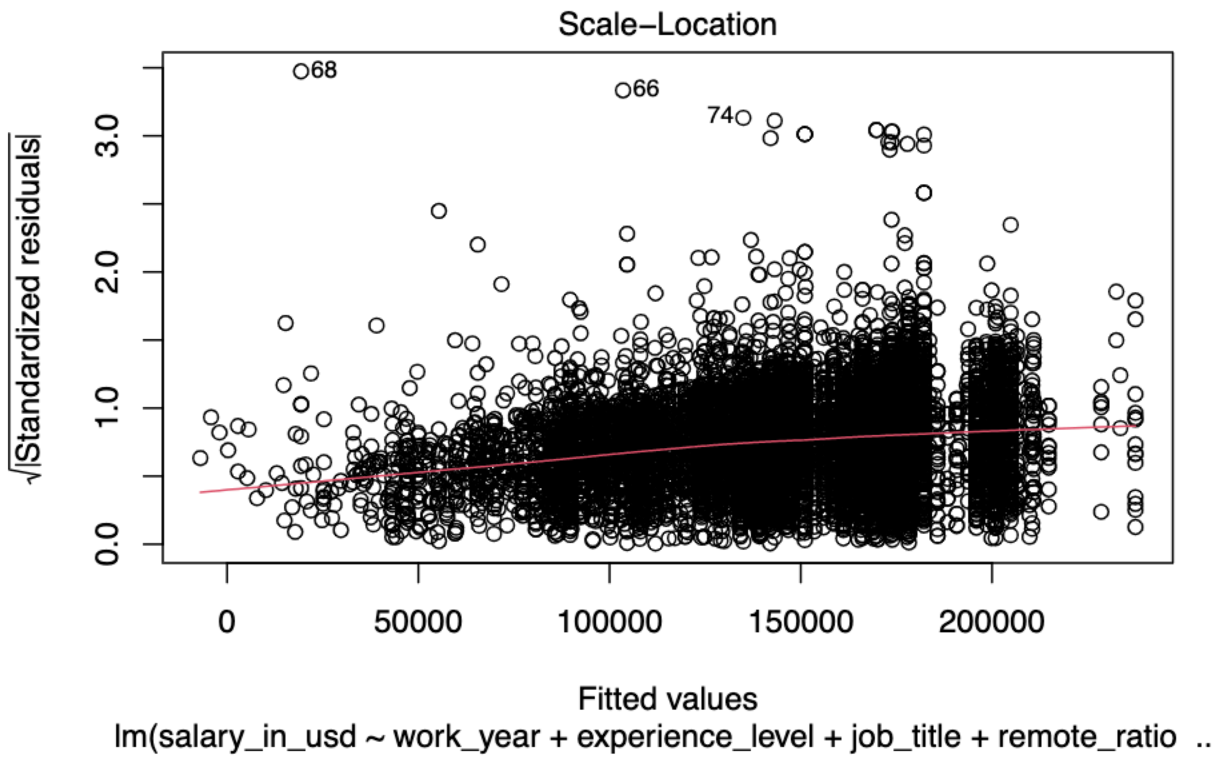
```
##
## Call:
## lm(formula = salary_in_usd ~ work_year + experience_level + job_title +
##     remote_ratio + company_location + company_size, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -145076  -39646   -8260   29673  754653
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    119213       2840  41.977  < 2e-16 ***
## work_year.L                      9976       1587   6.287 3.39e-10 ***
## work_year.Q                     -2536       1154  -2.198 0.027972 *
## experience_level.L              64702       2735  23.657  < 2e-16 ***
## experience_level.Q               5161       2140   2.411 0.015917 *
```

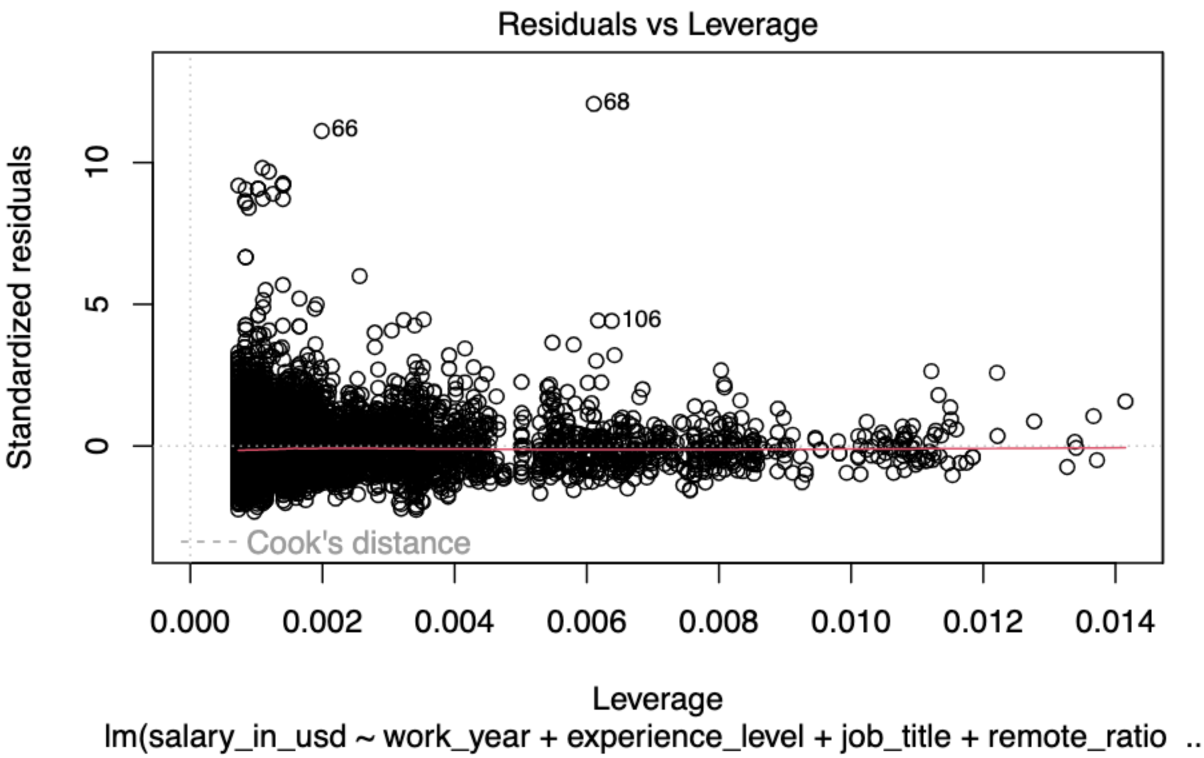
```
## experience_level.C              -1653       1368  -1.208 0.226892
## job_titleBI                      2791       3345   0.834 0.404048
## job_titleDE                     26921       2284  11.788  < 2e-16 ***
## job_titleDS                     34723       2207  15.735  < 2e-16 ***
## job_titleML                     61726       2495  24.738  < 2e-16 ***
## job_titleOther                  39054       2221  17.580  < 2e-16 ***
## remote_ratioHybrid             -15131       4743  -3.190 0.001427 **
## remote_ratioRemote              -5010       1441  -3.477 0.000510 ***
## company_locationFirst World    -42517       1941 -21.904  < 2e-16 ***
## company_locationDeveloping     -70359       4273 -16.468  < 2e-16 ***
## company_size.L                  14136       3971   3.560 0.000373 ***
## company_size.Q                  -6034       2640  -2.286 0.022285 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62730 on 9044 degrees of freedom
## Multiple R-squared:  0.2673, Adjusted R-squared:  0.266
## F-statistic: 206.2 on 16 and 9044 DF,  p-value: < 2.2e-16
```

```r
plot(model1)
```




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(salary_in_usd ~ work_year + experience_level + job_title + remote_ratio ...
68
66
74
10
5
0
−4
−2
0
2
4

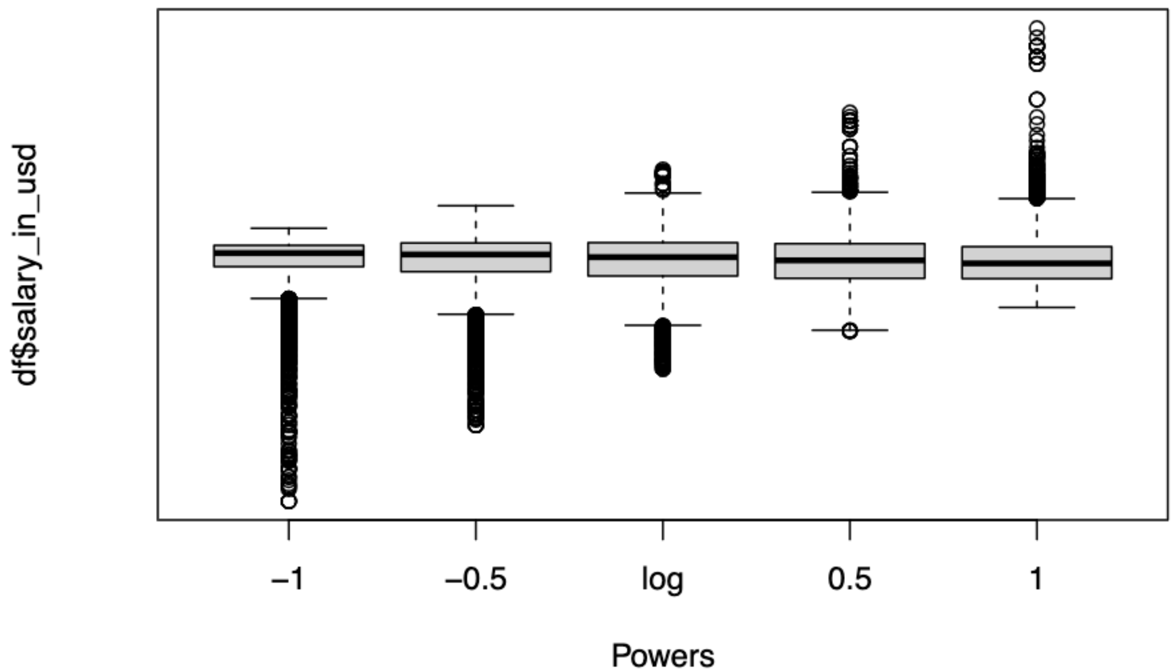

Scale−Location
68
66
74
√|Standardized residuals|
0.0
1.0
2.0
3.0
0
50000
100000
150000
200000
Fitted values
lm(salary_in_usd ~ work_year + experience_level + job_title + remote_ratio ...



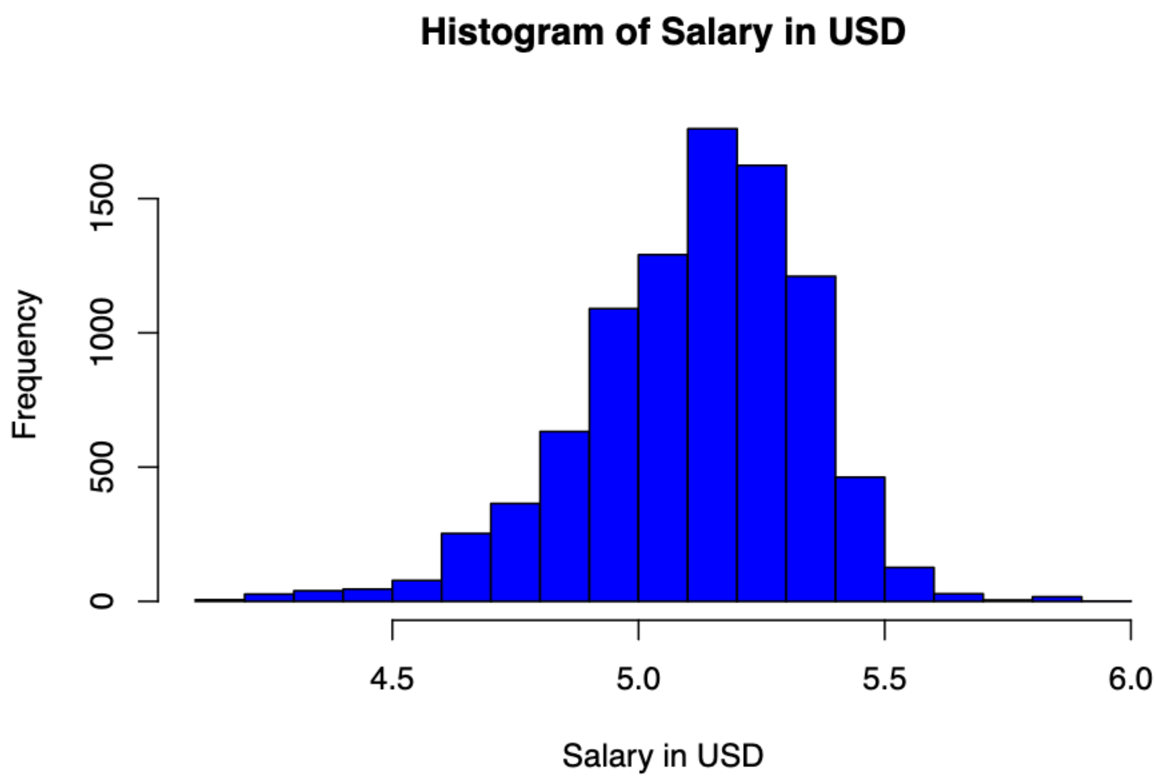

```
#see funnel pattern in residual, and qqplot does not follow the normal line
```

We can see see funnel pattern in residual, and qqplot does not follow the normal line. From the Q-Q residual plot that the data is mostly linear until it gets about 2 standard deviations above the mean. (Consider subsetting this data?). All VIF values are pretty low, suggesting low multicollinearity.

```
library(car)
symbox(df$salary_in_usd)
```




```r
hist(log10(df$salary_in_usd),xlab="Salary in USD", col = "blue", main = "Histogram of Salary in USD")
```

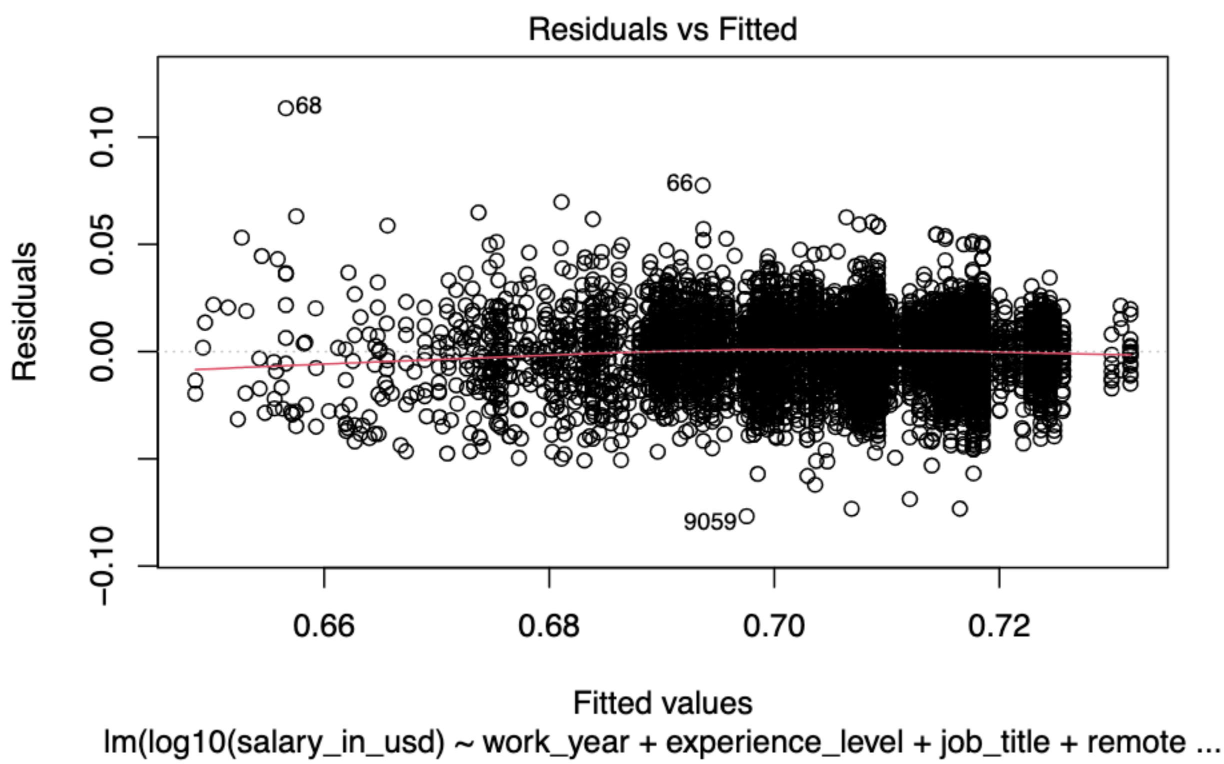


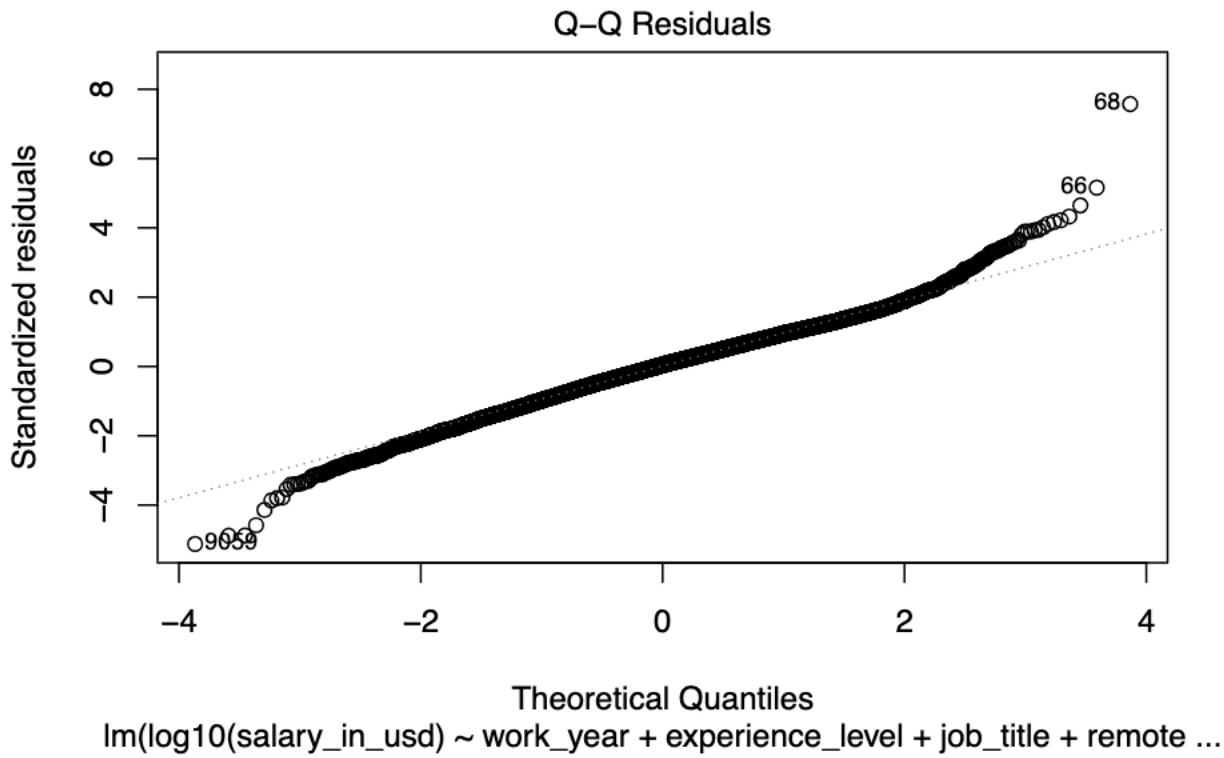

```r
df$salary_in_usd = log10(df$salary_in_usd)

#MLR model
model2 <- lm(log10(salary_in_usd) ~ work_year + experience_level + job_title + remote_ratio + company_lo
summary(model2)
```

```
##
## Call:
## lm(formula = log10(salary_in_usd) ~ work_year + experience_level +
##     job_title + remote_ratio + company_location + company_size,
##     data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.076771 -0.009387  0.000690  0.009883  0.113430
##
## Coefficients:
##                                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                   0.6997693  0.0006803 1028.678  < 2e-16 ***
## work_year.L                   0.0025341  0.0003801    6.667 2.76e-11 ***
## work_year.Q                  -0.0008098  0.0002763   -2.931  0.00339 **
## experience_level.L            0.0191278  0.0006551   29.197  < 2e-16 ***
## experience_level.Q           -0.0008953  0.0005127   -1.746  0.08082 .
## experience_level.C           -0.0005430  0.0003276   -1.657  0.09747 .
## job_titleBI                   0.0016121  0.0008012    2.012  0.04425 *
```

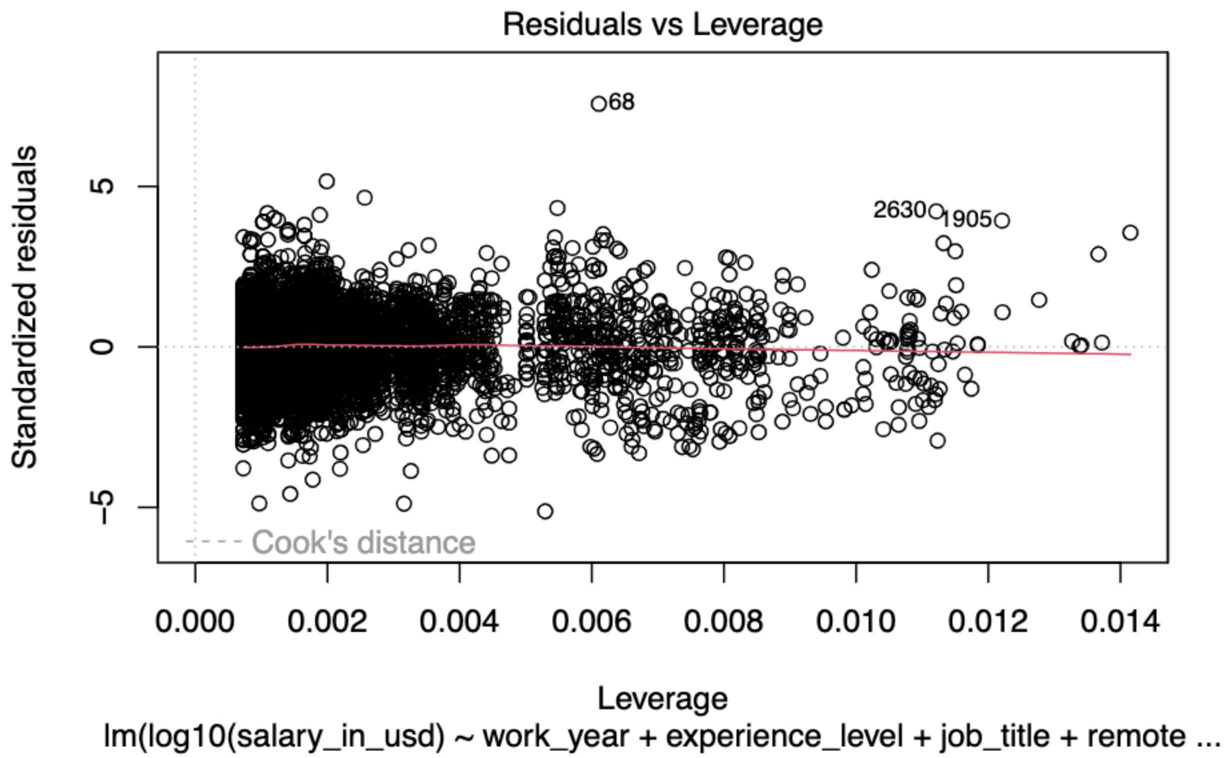
```
## job_titleDE                    0.0078178  0.0005470   14.291  < 2e-16 ***
## job_titleDS                    0.0097674  0.0005286   18.478  < 2e-16 ***
## job_titleML                    0.0157984  0.0005977   26.433  < 2e-16 ***
## job_titleOther                 0.0098223  0.0005321   18.459  < 2e-16 ***
## remote_ratioHybrid            -0.0065648  0.0011361   -5.778 7.80e-09 ***
## remote_ratioRemote            -0.0008650  0.0003451   -2.506  0.01222 *
## company_locationFirst World -0.0147211  0.0004650  -31.661  < 2e-16 ***
## company_locationDeveloping  -0.0338327  0.0010234  -33.058  < 2e-16 ***
## company_size.L                 0.0040289  0.0009513    4.235 2.31e-05 ***
## company_size.Q                -0.0028812  0.0006323   -4.557 5.27e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01503 on 9044 degrees of freedom
## Multiple R-squared:  0.4076, Adjusted R-squared:  0.4066
## F-statistic:   389 on 16 and 9044 DF,  p-value: < 2.2e-16
```

```r
plot(model2)
```




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(log10(salary_in_usd) ~ work_year + experience_level + job_title + remote ...
68
66
−4
−2
0
2
4
6
8


Scale−Location
68
6690059
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
0.66
0.68
0.70
0.72
Fitted values
lm(log10(salary_in_usd) ~ work_year + experience_level + job_title + remote ...




```r
#see funnel pattern in residual, and qqplot does not follow the normal line
```

```r
library(knitr)
model <- lm(log10(salary_in_usd)~.+experience_level*company_location,data = df)
ncvTest(model)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 486.1625, Df = 1, p = < 2.22e-16
```

```r
vif(model,type = "predictor")
```

```
## GVIFs computed for predictors
```

```
##                      GVIF Df GVIF^(1/(2*Df))   Interacts With
## work_year        1.255330  2        1.058497               --
## experience_level 1.303991 11        1.012138 company_location
## job_title        1.099184  5        1.009502               --
## remote_ratio     1.358309  2        1.079567               --
## company_location 1.303991 11        1.012138 experience_level
## company_size     1.387064  2        1.085236               --
##                                                                     Other Predictors
## work_year        experience_level, job_title, remote_ratio, company_location, company_size
## experience_level                         work_year, job_title, remote_ratio, company_size
```

```
## job_title        work_year, experience_level, remote_ratio, company_location, company_size
## remote_ratio        work_year, experience_level, job_title, company_location, company_size
## company_location                              work_year, job_title, remote_ratio, company_size
## company_size        work_year, experience_level, job_title, remote_ratio, company_location
```