Data Science Salaries Project

Aditi Patil and Masha Volkova

2023-04-17

Aditi Patil and Masha Volkova

```
ds.salaries <- read.csv(file = 'ds_salaries.csv')
str(ds.salaries)
```

```
## 'data.frame':    607 obs. of  12 variables:
##  $ X                 : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ work_year         : int  2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
##  $ experience_level  : chr  "MI" "SE" "SE" "MI" ...
##  $ employment_type   : chr  "FT" "FT" "FT" "FT" ...
##  $ job_title         : chr  "Data Scientist" "Machine Learning Scientist" "Big Data Engineer" "Produc
##  $ salary            : int  70000 260000 85000 20000 150000 72000 190000 11000000 135000 125000 ...
##  $ salary_currency   : chr  "EUR" "USD" "GBP" "USD" ...
##  $ salary_in_usd     : int  79833 260000 109024 20000 150000 72000 190000 35735 135000 125000 ...
##  $ employee_residence: chr  "DE" "JP" "GB" "HN" ...
##  $ remote_ratio      : int  0 0 50 0 50 100 100 50 100 50 ...
##  $ company_location  : chr  "DE" "JP" "GB" "HN" ...
##  $ company_size      : chr  "L" "S" "M" "S" ...
```

Preprocessing and Exploratory Analysis

Since `salaries` is in different currencies we can drop it for in favor of `salary_in_usd`. Furthermore we can also drop variables like `X`, `company_location`, `employee_residence`, and `salary_currency`

```
cols = c("work_year","experience_level","employment_type","job_title", "remote_ratio", "company_size","
ds.salaries = subset(ds.salaries,select=cols)

# Treat the char variables as factors
ds.salaries$work_year = as.factor(ds.salaries$work_year)
ds.salaries$experience_level = as.factor(ds.salaries$experience_level)
ds.salaries$employment_type = as.factor(ds.salaries$employment_type)
ds.salaries$job_title = as.factor(ds.salaries$job_title)
ds.salaries$remote_ratio = as.factor(ds.salaries$remote_ratio)
ds.salaries$company_size = as.factor(ds.salaries$company_size)

# Treat employee_residence and company_location as continous variables
ds.salaries$employee_residence = unclass(as.factor(ds.salaries$employee_residence))
ds.salaries$company_location = unclass(as.factor(ds.salaries$company_location))

# View the final data
str(ds.salaries)
```

```
## 'data.frame':    607 obs. of  9 variables:
##  $ work_year         : Factor w/ 3 levels "2020","2021",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ experience_level  : Factor w/ 4 levels "EN","EX","MI",..: 3 4 4 3 4 1 4 3 3 4 ...
##  $ employment_type   : Factor w/ 4 levels "CT","FL","FT",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ job_title         : Factor w/ 50 levels "3D Computer Vision Researcher",..: 23 41 8 48 38 13 35 23
##  $ remote_ratio      : Factor w/ 3 levels "0","50","100": 1 1 2 1 2 3 3 2 3 2 ...
##  $ company_size      : Factor w/ 3 levels "L","M","S": 1 3 2 3 1 1 3 1 1 3 ...
##  $ employee_residence: int  15 33 21 24 56 56 56 26 56 42 ...
##   ..- attr(*, "levels")= chr [1:57] "AE" "AR" "AT" "AU" ...
##  $ company_location  : int  13 30 19 21 49 49 49 23 49 39 ...
##   ..- attr(*, "levels")= chr [1:50] "AE" "AS" "AT" "AU" ...
##  $ salary_in_usd     : int  79833 260000 109024 20000 150000 72000 190000 35735 135000 125000 ...
```

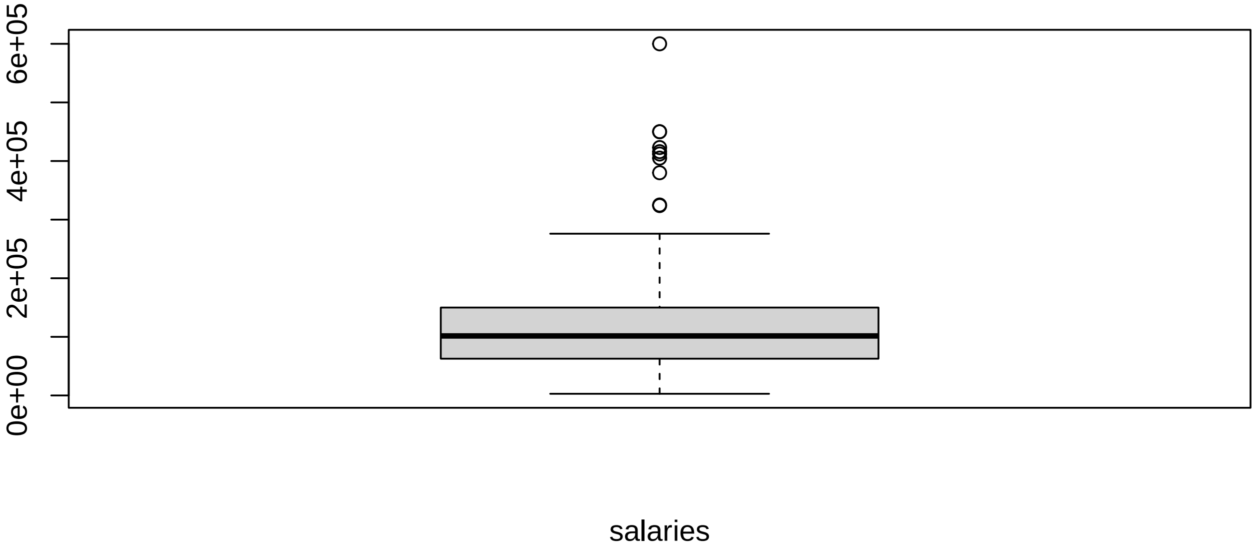
In this case we have factors `work_year`, `experience_level`, `employment_type`, `job_title`, `remote_ratio`, `work_year` and `company_size`.

`employee_residence` and `company_location` will be treated as continuous variables.

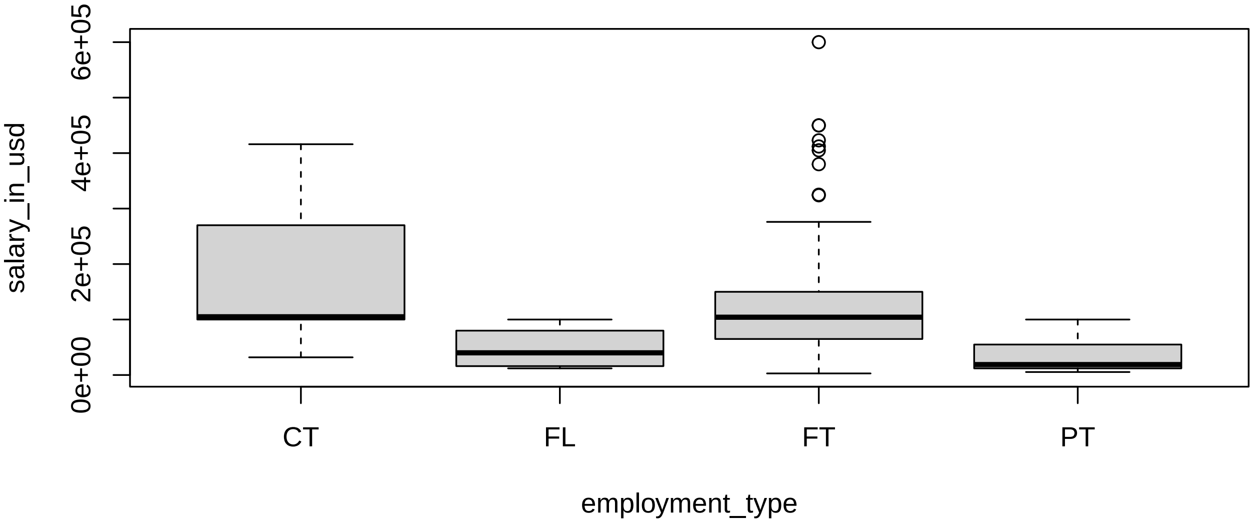
`job_title` has 50 levels

The response variable `salary_in_usd` is continuous.

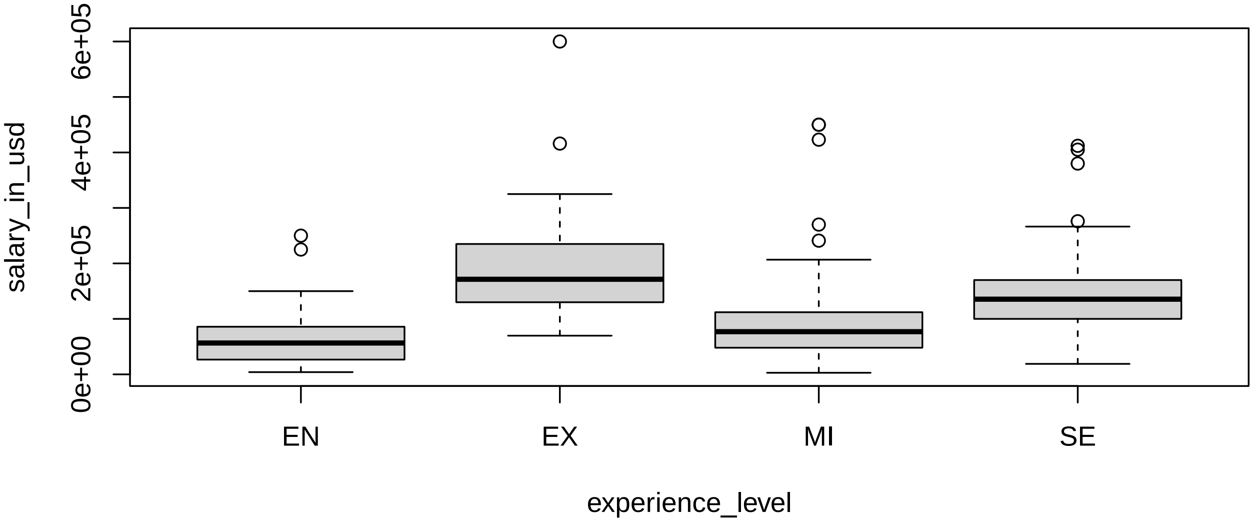
```
boxplot(ds.salaries$salary_in_usd, xlab="salaries")
```



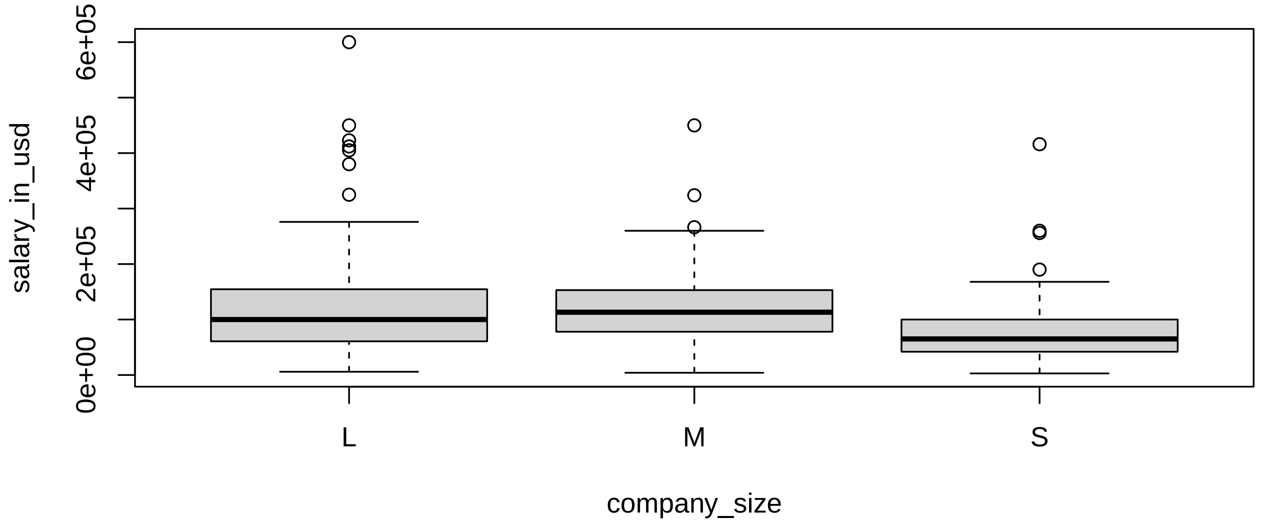
```
boxplot(salary_in_usd~employment_type,
        data=ds.salaries)
```



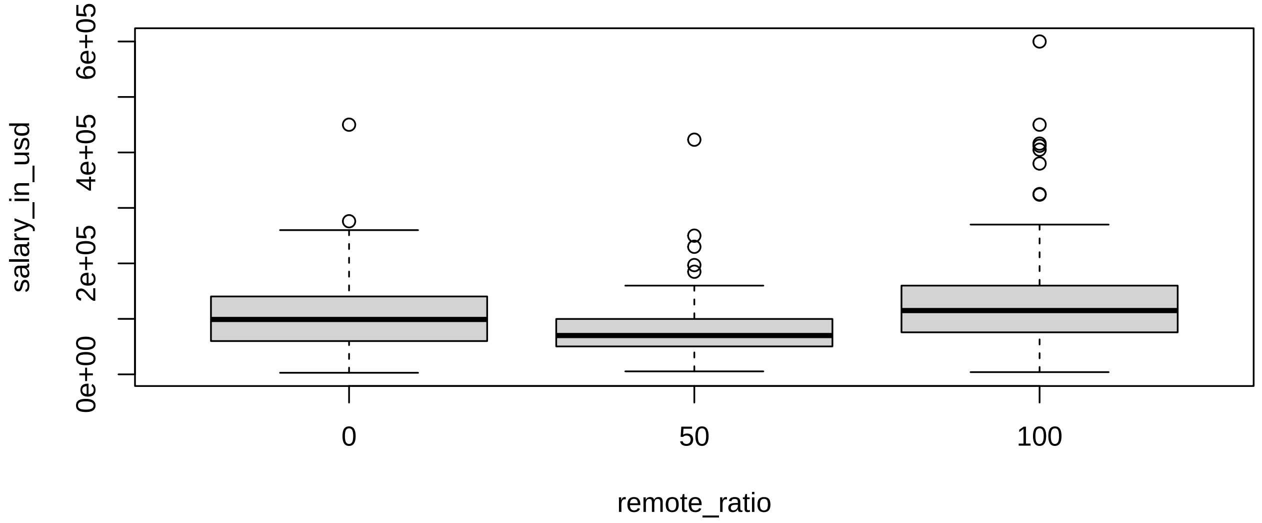
```
boxplot(salary_in_usd~experience_level,
        data=ds.salaries)
```



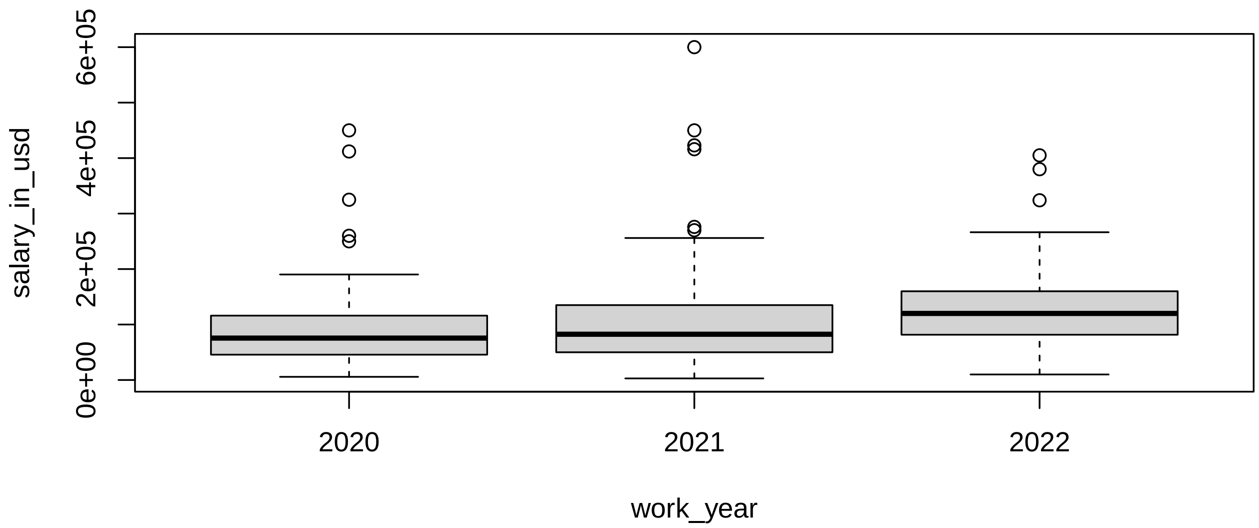
```
boxplot(salary_in_usd~company_size,
        data=ds.salaries)
```



```
boxplot(salary_in_usd~remote_ratio,
        data=ds.salaries)
```



```
boxplot(salary_in_usd~work_year,
        data=ds.salaries)
```



Possible techniques

4-way anova? Tukey HSD to perform multiple comparison tests?

Linear Regression for Model Selection

```
lm.salaries = lm(salary_in_usd~ work_year + experience_level + employment_type + job_title + employee_r

summary(lm.salaries)
```

```
##
## Call:
## lm(formula = salary_in_usd ~ work_year + experience_level + employment_type +
##     job_title + employee_residence + remote_ratio + company_location +
##     company_size, data = ds.salaries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -163133  -26170   -3270   22055  289748
##
## Coefficients:
##                                                    Estimate Std. Error t value
## (Intercept)                                          32151.9    62177.4   0.517
## work_year2021                                       -16632.1     7250.1  -2.294
## work_year2022                                        -4257.1     8074.4  -0.527
## experience_levelEX                                  116045.5    14781.2   7.851
## experience_levelMI                                   24148.2     7007.4   3.446
## experience_levelSE                                   53723.1     7345.0   7.314
## employment_typeFL                                  -102687.0    39897.8  -2.574
## employment_typeFT                                   -48827.3    28559.7  -1.710
## employment_typePT                                   -53797.1    33626.8  -1.600
## job_titleAI Scientist                                51734.3    56098.3   0.922
## job_titleAnalytics Engineer                          42370.3    59833.2   0.708
## job_titleApplied Data Scientist                     117663.8    58691.2   2.005
## job_titleApplied Machine Learning Scientist          86760.0    59787.4   1.451
## job_titleBI Data Analyst                             26572.7    57814.4   0.460
## job_titleBig Data Architect                          86030.2    73748.1   1.167
## job_titleBig Data Engineer                          22786.6    56874.2   0.401
## job_titleBusiness Data Analyst                      28778.8    58901.9   0.489
## job_titleCloud Data Engineer                        88203.6    64985.2   1.357
## job_titleComputer Vision Engineer                   25692.2    57136.8   0.450
## job_titleComputer Vision Software Engineer         104106.4    61913.2   1.681
## job_titleData Analyst                               22742.1    54034.1   0.421
## job_titleData Analytics Engineer                    10527.7    59507.8   0.177
## job_titleData Analytics Lead                       297673.7    74177.4   4.013
## job_titleData Analytics Manager                     29390.5    57112.6   0.515
## job_titleData Architect                             97907.3    55940.1   1.750
## job_titleData Engineer                              44313.3    53911.1   0.822
## job_titleData Engineering Manager                   42154.5    58604.7   0.719
## job_titleData Science Consultant                    51745.9    57609.3   0.898
## job_titleData Science Engineer                      48934.6    61470.8   0.796
## job_titleData Science Manager                       73156.2    55735.3   1.313
## job_titleData Scientist                             52922.1    53933.7   0.981
## job_titleData Specialist                            70048.7    73916.2   0.948
## job_titleDirector of Data Engineering               88145.9    64651.6   1.363
## job_titleDirector of Data Science                   82599.7    58409.4   1.414
## job_titleETL Developer                              28098.8    64660.1   0.435
## job_titleFinance Data Analyst                       25545.9    73886.5   0.346
## job_titleFinancial Data Analyst                    221491.0    64935.8   3.411
## job_titleHead of Data                               65043.5    58678.0   1.108
## job_titleHead of Data Science                       17441.2    60229.1   0.290
## job_titleHead of Machine Learning                  -41690.9    75296.8  -0.554
## job_titleLead Data Analyst                          23743.3    61367.5   0.387
## job_titleLead Data Engineer                         74223.9    57863.8   1.283
## job_titleLead Data Scientist                        66212.8    61455.0   1.077
## job_titleLead Machine Learning Engineer             41181.3    73999.7   0.557
## job_titleMachine Learning Developer                 78995.3    61390.4   1.287
## job_titleMachine Learning Engineer                  61926.8    54498.3   1.136
## job_titleMachine Learning Infrastructure Engineer   31960.2    61015.3   0.524
## job_titleMachine Learning Manager                   80720.2    74353.7   1.086
## job_titleMachine Learning Scientist                103414.5    56785.8   1.821
## job_titleMarketing Data Analyst                     41086.1    74117.6   0.554
## job_titleML Engineer                                81362.8    57119.7   1.424
## job_titleNLP Engineer                              -28553.8    74146.9  -0.385
## job_titlePrincipal Data Analyst                     80577.3    64605.8   1.247
## job_titlePrincipal Data Engineer                   214634.2    61432.8   3.494
## job_titlePrincipal Data Scientist                  133372.5    57268.2   2.329
## job_titleProduct Data Analyst                      -19522.6    65034.4  -0.300
## job_titleResearch Scientist                         83879.9    55390.4   1.514
## job_titleStaff Data Scientist                      -32699.8    78651.6  -0.416
## employee_residence                                   1028.7      285.1   3.608
## remote_ratio50                                     -12403.7     7722.7  -1.606
## remote_ratio100                                      -442.1     5444.4  -0.081
## company_location                                      354.5      319.8   1.109
## company_sizeM                                       -6078.9     5837.6  -1.041
## company_sizeS                                      -19540.3     7331.9  -2.665
##                                                   Pr(>|t|)
## (Intercept)                                       0.605297
## work_year2021                                     0.022168 *
## work_year2022                                     0.598246
## experience_levelEX                                2.22e-14 ***
## experience_levelMI                                0.000613 ***
## experience_levelSE                                9.34e-13 ***
## employment_typeFL                                 0.010324 *
## employment_typeFT                                 0.087900 .
## employment_typePT                                 0.110219
## job_titleAI Scientist                             0.356830
## job_titleAnalytics Engineer                       0.479162
## job_titleApplied Data Scientist                   0.045481 *
## job_titleApplied Machine Learning Scientist       0.147318
## job_titleBI Data Analyst                          0.645973
## job_titleBig Data Architect                       0.243908
## job_titleBig Data Engineer                        0.688837
## job_titleBusiness Data Analyst                    0.625330
## job_titleCloud Data Engineer                      0.175254
## job_titleComputer Vision Engineer                 0.653134
## job_titleComputer Vision Software Engineer        0.093243 .
## job_titleData Analyst                             0.674007
## job_titleData Analytics Engineer                  0.859643
## job_titleData Analytics Lead                      6.84e-05 ***
## job_titleData Analytics Manager                   0.607038
## job_titleData Architect                           0.080645 .
## job_titleData Engineer                            0.411456
## job_titleData Engineering Manager                 0.472264
## job_titleData Science Consultant                  0.369466
## job_titleData Science Engineer                    0.426344
## job_titleData Science Manager                     0.189884
## job_titleData Scientist                           0.326910
## job_titleData Specialist                          0.343716
## job_titleDirector of Data Engineering             0.173322
## job_titleDirector of Data Science                 0.157891
## job_titleETL Developer                            0.664054
## job_titleFinance Data Analyst                     0.729668
## job_titleFinancial Data Analyst                   0.000696 ***
## job_titleHead of Data                             0.268145
## job_titleHead of Data Science                     0.772247
## job_titleHead of Machine Learning                 0.580020
## job_titleLead Data Analyst                        0.698979
## job_titleLead Data Engineer                       0.200132
## job_titleLead Data Scientist                      0.281772
## job_titleLead Machine Learning Engineer           0.578094
## job_titleMachine Learning Developer               0.198723
## job_titleMachine Learning Engineer                0.256329
## job_titleMachine Learning Infrastructure Engineer 0.600626
## job_titleMachine Learning Manager                 0.278127
## job_titleMachine Learning Scientist               0.069137 .
## job_titleMarketing Data Analyst                   0.579576
## job_titleML Engineer                              0.154898
## job_titleNLP Engineer                             0.700316
## job_titlePrincipal Data Analyst                   0.212857
## job_titlePrincipal Data Engineer                  0.000515 ***
## job_titlePrincipal Data Scientist                 0.020230 *
## job_titleProduct Data Analyst                     0.764148
## job_titleResearch Scientist                       0.130521
## job_titleStaff Data Scientist                     0.677754
## employee_residence                                       0.000337 ***
## remote_ratio50                                           0.108825
## remote_ratio100                                          0.935306
## company_location                                         0.268062
## company_sizeM                                            0.298185
## company_sizeS                                            0.007926 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50280 on 543 degrees of freedom
## Multiple R-squared:  0.5501, Adjusted R-squared:  0.4979
## F-statistic: 10.54 on 63 and 543 DF,  p-value: < 2.2e-16
```

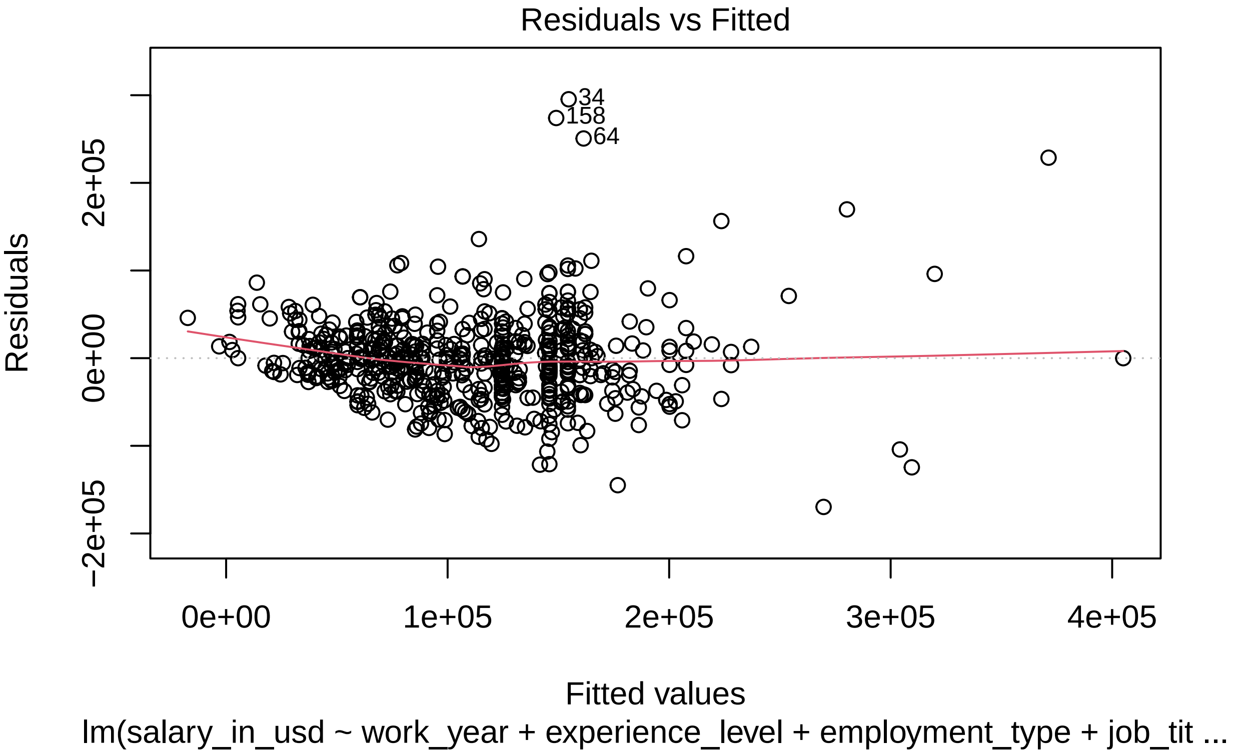
We can see that the variables `work_year`, `experience_level`, `employment_type`, `job_title`, `company_size`, and `employee_residence` are significant. So we will keep them in the model

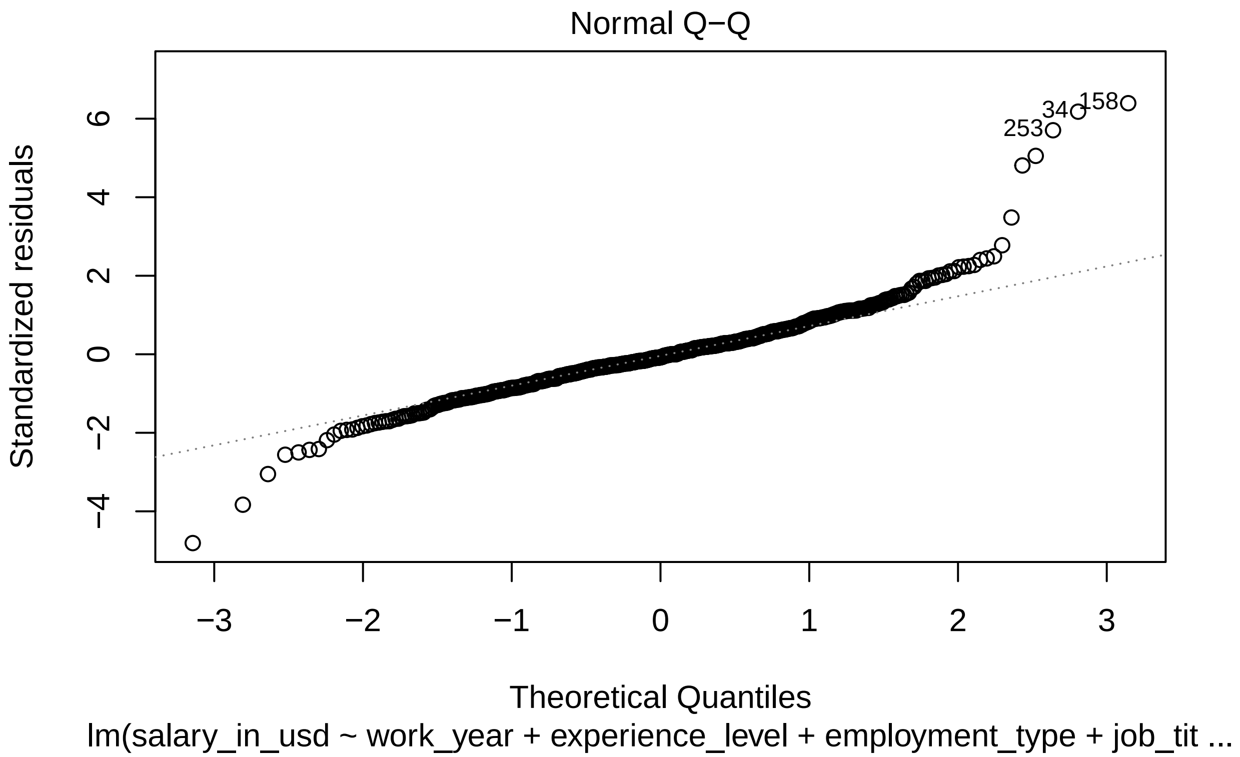
ANOVA Assumptions

```
lm.salaries = lm(salary_in_usd~ work_year + experience_level + employment_type + job_title + employee_re

plot(lm.salaries,which=c(1,2))
```

```
## Warning: not plotting observations with leverage one:
##   91, 166, 256, 385, 456, 524
```





ANOVA 6-way

The following code takes a very long time to run. **Is there a more efficient way to narrow down variables?**

```
library(car)
```

```
## Loading required package: carData
```

```
# Anova(lm(salary_in_usd~work_year*experience_level*employment_type*job_title*employee_residence* compa
```