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" "Produ
##  $ 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","salary_in_usd","remote_ratio","c
ds.salaries = subset(ds.salaries,select=cols)

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

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

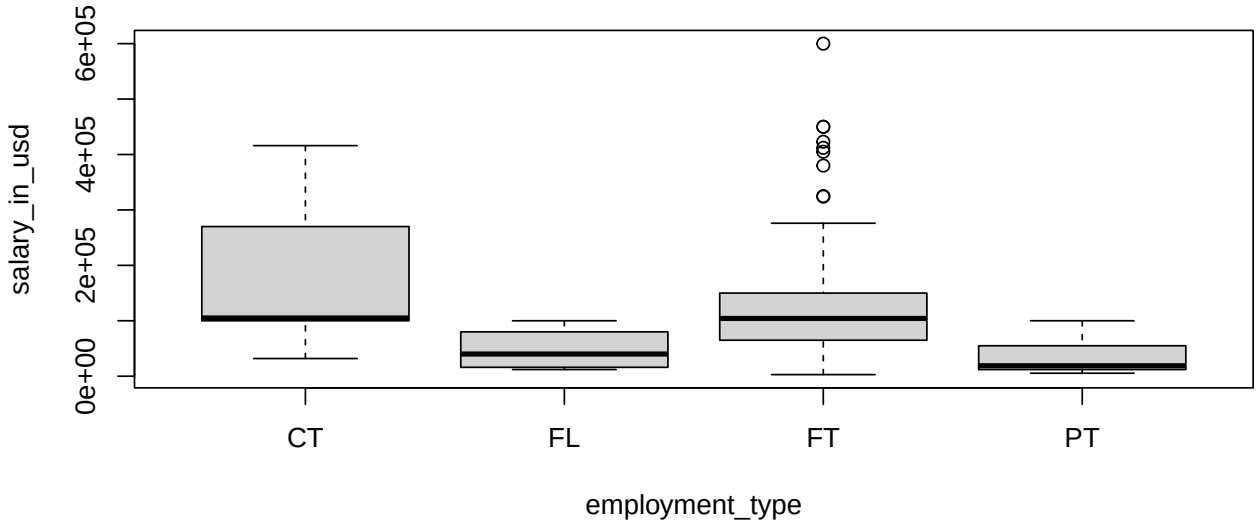
```
## 'data.frame':    607 obs. of  7 variables:
##  $ work_year       : int  2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
##  $ experience_level: Factor w/ 4 levels "EN","EX","MI",..: 3 4 4 3 4 1 4 3 3 4 ...
##  $ employment_type : Factor w/ 4 levels "CT","FL","FT",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ job_title       : Factor w/ 50 levels "3D Computer Vision Researcher",..: 23 41 8 48 38 13 35 23 
##  $ salary_in_usd   : int  79833 260000 109024 20000 150000 72000 190000 35735 135000 125000 ...
##  $ 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 ...
```

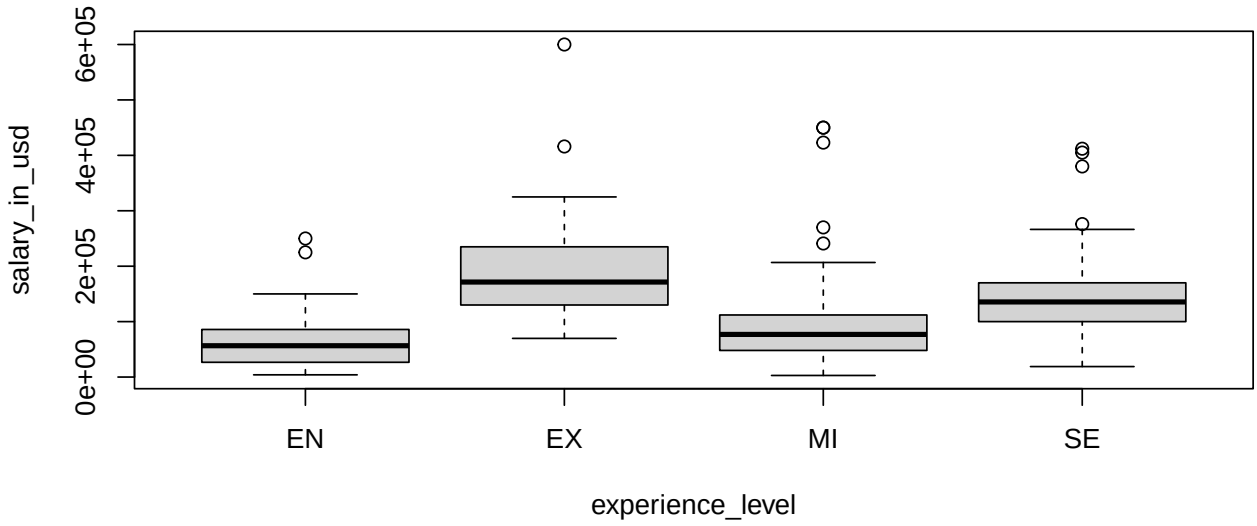
In this case we have factors `experience_level`, `employment_type`,and `company_size`, and `remote_ratio`. `job_title` and `work_year` are categorical variables with more than 3 levels

The response variable `salary_in_usd` is continuous.

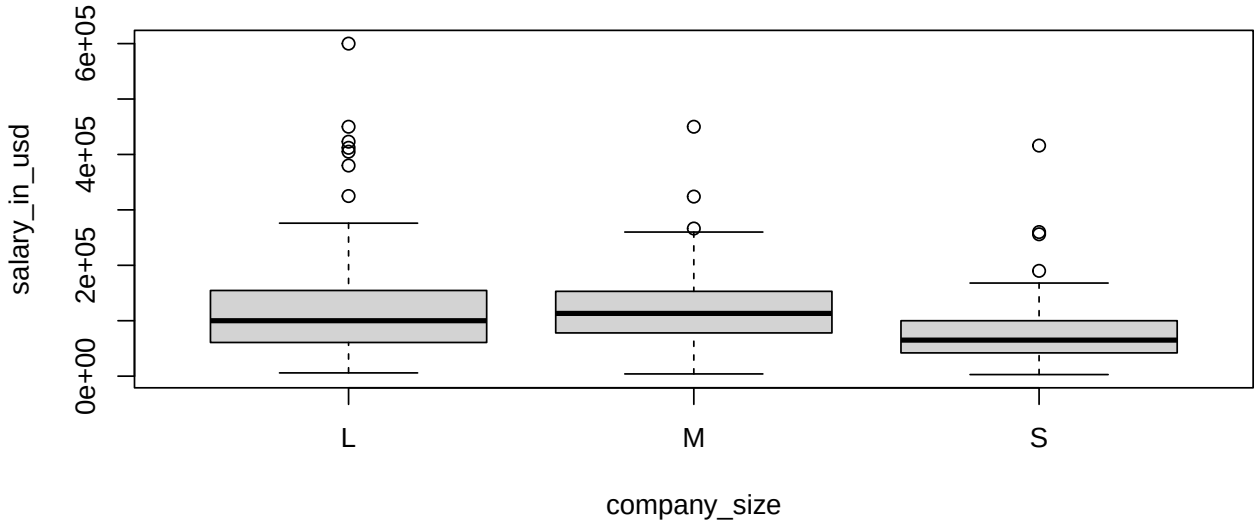
```
boxplot(salary_in_usd~employment_type,
        data=ds.salaries)
```



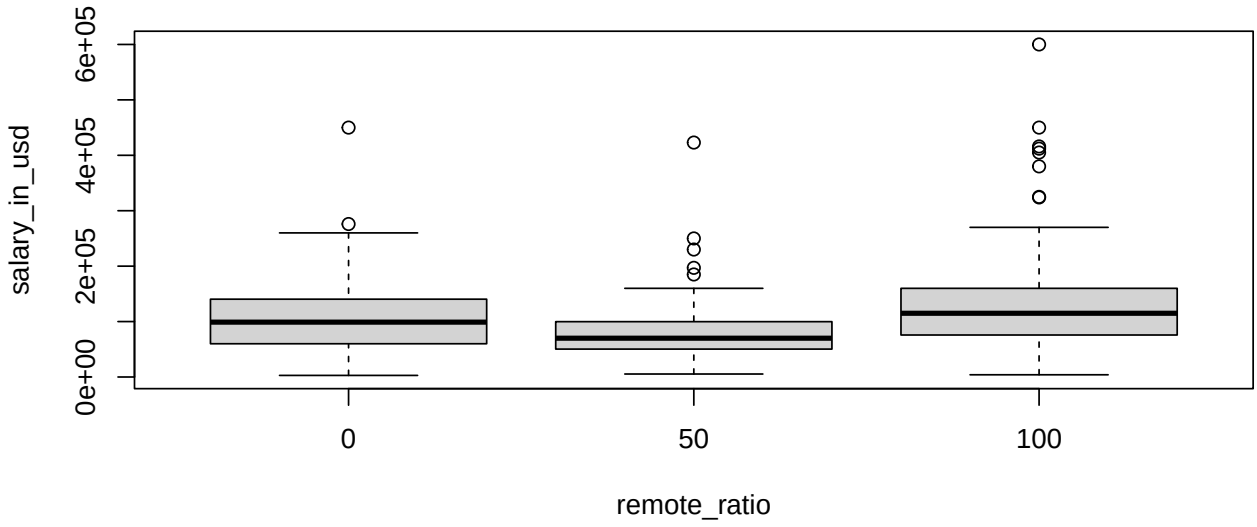

```
boxplot(salary_in_usd~experience_level,
        data=ds.salaries)
```




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




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



Possible techniques

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

ANOVA 4-way

```
library(car)
```

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

```
Anova(lm(salary_in_usd~company_size*experience_level*employment_type*remote_ratio,data=ds.salaries), ty
```

```
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
```

```
## Anova Table (Type II tests)
##
## Response: salary_in_usd
##                                                          Sum Sq  Df
```

```
## company_size                                                  8.0668e+10   2
## experience_level                                              5.3707e+11   3
## employment_type                                               4.2302e+10   3
## remote_ratio                                                  6.0331e+10   2
## company_size:experience_level                                 7.1301e+10   6
## company_size:employment_type                                  4.7433e+09   3
## experience_level:employment_type                              1.1817e+10   3
## company_size:remote_ratio                                     1.5981e+10   4
## experience_level:remote_ratio                                 6.8294e+10   6
## employment_type:remote_ratio                                  4.7220e+09   2
## company_size:experience_level:employment_type                               0
## company_size:experience_level:remote_ratio                    3.6788e+10  10
## company_size:employment_type:remote_ratio                     6.8740e+08   1
## experience_level:employment_type:remote_ratio                               0
## company_size:experience_level:employment_type:remote_ratio                  0
## Residuals                                                     1.8943e+12 557
##                                                               F value    Pr(>F)
## company_size                                                  11.8600 9.035e-06
## experience_level                                              52.6411 < 2.2e-16
## employment_type                                                4.1462 0.0063848
## remote_ratio                                                   8.8700 0.0001614
## company_size:experience_level                                  3.4943 0.0021161
## company_size:employment_type                                   0.4649 0.7068812
## experience_level:employment_type                               1.1582 0.3250644
## company_size:remote_ratio                                      1.1748 0.3208523
## experience_level:remote_ratio                                  3.3470 0.0030072
## employment_type:remote_ratio                                   0.6942 0.4998868
## company_size:experience_level:employment_type
## company_size:experience_level:remote_ratio                     1.0817 0.3740840
## company_size:employment_type:remote_ratio                      0.2021 0.6531843
## experience_level:employment_type:remote_ratio
## company_size:experience_level:employment_type:remote_ratio
## Residuals
##
## company_size                                                  ***
## experience_level                                              ***
## employment_type                                               **
## remote_ratio                                                  ***
## company_size:experience_level                                 **
## company_size:employment_type
## experience_level:employment_type
## company_size:remote_ratio
## experience_level:remote_ratio                                 **
## employment_type:remote_ratio
## company_size:experience_level:employment_type
## company_size:experience_level:remote_ratio
## company_size:employment_type:remote_ratio
## experience_level:employment_type:remote_ratio
## company_size:experience_level:employment_type:remote_ratio
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Princicple of hierarchy: remove highest interaction terms 4-way interaction and recompute

```
Anova(lm(salary_in_usd~ company_size+experience_level+employment_type+remote_ratio +
           company_size:experience_level + company_size:employment_type +
           company_size:remote_ratio + experience_level:employment_type +
           experience_level:remote_ratio + employment_type:remote_ratio +
           company_size:experience_level:employment_type  +
           company_size:experience_level:remote_ratio  +
           company_size:employment_type:remote_ratio +
           experience_level:employment_type:remote_ratio ,data=ds.salaries), type=2)
```

```
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
```

```
## Anova Table (Type II tests)
##
## Response: salary_in_usd
##                                                Sum Sq  Df F value    Pr(>F)
## company_size                                8.0668e+10   2 11.8600 9.035e-06
## experience_level                            5.3707e+11   3 52.6411 < 2.2e-16
## employment_type                             4.2302e+10   3  4.1462 0.0063848
## remote_ratio                                6.0331e+10   2  8.8700 0.0001614
## company_size:experience_level               7.1301e+10   6  3.4943 0.0021161
## company_size:employment_type                4.7433e+09   3  0.4649 0.7068812
## company_size:remote_ratio                   1.5981e+10   4  1.1748 0.3208523
## experience_level:employment_type            1.1817e+10   3  1.1582 0.3250644
## experience_level:remote_ratio               6.8294e+10   6  3.3470 0.0030072
## employment_type:remote_ratio                4.7220e+09   2  0.6942 0.4998868
## company_size:experience_level:employment_type             0
## company_size:experience_level:remote_ratio  3.6788e+10  10  1.0817 0.3740840
## company_size:employment_type:remote_ratio   6.8740e+08   1  0.2021 0.6531843
## experience_level:employment_type:remote_ratio             0
## Residuals                                   1.8943e+12 557
##
## company_size                                ***
## experience_level                            ***
## employment_type                             **
## remote_ratio                                ***
## company_size:experience_level               **
## company_size:employment_type
## company_size:remote_ratio
## experience_level:employment_type
## experience_level:remote_ratio               **
## employment_type:remote_ratio
## company_size:experience_level:employment_type
## company_size:experience_level:remote_ratio
## company_size:employment_type:remote_ratio
## experience_level:employment_type:remote_ratio
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since none of the 3way interactions are significant, we can drop all three-way interactions

```
Anova(lm(salary_in_usd~company_size + experience_level + employment_type+
           remote_ratio +company_size:experience_level +
           company_size:employment_type + company_size:remote_ratio +
```

```
          experience_level:employment_type + experience_level:remote_ratio +
          employment_type:remote_ratio,data=ds.salaries), type=2)
```

```
## Note: model has aliased coefficients
##       sums of squares computed by model comparison
```

```
## Anova Table (Type II tests)
##
## Response: salary_in_usd
##                                     Sum Sq  Df F value    Pr(>F)
## company_size                     8.1393e+10   2 11.9845 7.977e-06 ***
## experience_level                 5.3844e+11   3 52.8540 < 2.2e-16 ***
## employment_type                  4.8798e+10   3  4.7901 0.0026400 **
## remote_ratio                     6.0331e+10   2  8.8832 0.0001589 ***
## company_size:experience_level    7.1692e+10   6  3.5187 0.0019907 **
## company_size:employment_type     5.5163e+09   3  0.5415 0.6540544
## company_size:remote_ratio        1.6235e+10   4  1.1953 0.3117832
## experience_level:employment_type 1.3087e+10   3  1.2847 0.2787645
## experience_level:remote_ratio    6.8617e+10   6  3.3678 0.0028551 **
## employment_type:remote_ratio     3.4195e+09   2  0.5035 0.6046806
## Residuals                        1.9322e+12 569
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```