# Analysis of Factors Affecting Data Science Job Salaries

Aditi Patil and Masha Volkova

STAT 4051 Applied Statistics I

University of Minnesota Twin-Cities

## Introduction:

In this report, we wish to analyze factors that impact the base salaries of Data Scientists. We chose this topic because as future data scientists we were interested in the factors which affect base salaries. As future data scientists we ask questions like does having a full-time job vs a part-time job affect our salary? Does being a full-time junior mid-level employee vs a part-time entry-level employee affect our salary? This report strives to explore the relationship between experience level, employment type, and work yea, in order to provide an answer to these questions.

The Analysis of Variance (ANOVA) model is used to compare nested models (one nested inside the other) with the null model being the simpler model and the alternative being the more complex one. Specifically, it can be used to test whether there is a difference in the group means in the levels of the explanatory variables. The underlying distribution of the ANOVA is the F-distribution. In an ANOVA fixed effect model, the levels of the data are selected intentionally by the experimenter (fixed numbers). In a random effect model, the random effect's "treatment" levels are selected randomly from the population "treatments" and so have a variance term associated with it. A mixed-effect model contains both fixed effects and random effects[1].

In our dataset, we chose to explore how experience level, employment type, and work year affect the salaries of Data Scientists. These variables can be treated as categorical variables, so we decided to analyze the data with a continuous response variable salary. Therefore, an ANOVA model can be used to perform an analysis of variance of the levels of the independent categorical variables. Furthermore, the data only contains three levels (2020, 2021, 2022) without revealing why the author of the dataset chose these years. Therefore, both fixed-effects and mixed-effects ANOVA models were explored. First, we considered work year as a fixed effect. Then we considered work year as a random effect while leaving the other two variables as fixed effects. Using the final models in each category we can determine which variables (and their interactions) affect the mean salary indirectly by determining if there is a difference between group means of the levels.

## Methods and Materials:

Due to the restriction of time and resources, we found the Data Science Job Salaries dataset on Kaggle[2]. The dataset contains variables such as work year, experience level, employment type, job title, remote ratio, company size, employee residence,

[1] Oehlert (3.5,3.6, 11.1,12,4),2018

[2] Bhatia, 2022

company location, and salary in USD. For our analysis, we considered work year, experience level, and employment type as the explanatory variables. The response variable was the total gross salary amount paid in US dollars.

## Preprocessing

For data preprocessing, we dropped variables such as `X`, `job_title`, `salary`, `salary_currency`, `employee_residence`, `remote_ratio`, `company_location`, and `company_size`. This left us with three categorical explanatory variables: `work_year`, `experience_level`, and `employment_type`, and the response variable: `salary_in_usd`.

We dropped the columns `salary` and `salary_currency` in favor of the `salary_usd` column where all the salaries were converted to the same currency. We treated the `work_year` column as a categorical variable with three levels as it only had data from 3 years (2020, 2021, 2022). The `experience_level` column was treated as a categorical variable with four levels: EN (Entry-level / Junior), MI (Mid-level / Intermediate), SE (Senior-level / Expert), and EX (Executive-level / Director). The `employment_type` column was treated as a categorical variable with 4 levels: FT (Full-time), PT (Part-time), CT (Contract), and FL (Freelance). The response variable `salary_in_usd` was a numeric, continuous variable.

## Exploratory Analysis

We created boxplots for salaries, employment type, experience level, and work year to assess if there were any outliers in the data that could impact our results. We also created three two-way interaction plots to check for the presence of two-way interactions between our predictor variables[3]. This was just a visual exploration and the significance of these interactions was tested in the ANOVA analysis discussed in the next sections.

## Model Selection via Analysis of ANOVA Assumptions

To select the model which violates the least amount of assumptions, we considered three models with transformations on the response variable `salary_in_usd`:

T1: salary_in_usd~ work_year + experience_level + employment_type
T2: log(salary_in_usd)~ work_year + experience_level + employment_type
T3: sqrt(salary_in_usd)~ work_year + experience_level + employment_type

Using these three models, the ANOVA assumptions of normality and constant variance were assessed. To accomplish this, we generated a residual plot, a QQ plot, and a

[3] Torsten Hothorn (2014)

box-cox normality plot for the linear models. These transformations were considered because assumptions of constant variance and normality were violated for the regular model (T1)[4]. We applied both the log transformation and square root transformation on the response variable, both of which improved the results of the residual and QQ plots. In the end, we decided to apply the square root transformation to the response variable. The discussion on this is in the results section below.

Fixed Effect Models:

We considered two fixed effects models: one with main effects and the other with interactions. Based on the principle of hierarchy, we considered more than one interaction model to determine a final interaction model. We compared the main effects model to the interaction effects model to determine a final fixed effects model using a p-value with a significance level of 0.05. Finally, we assessed the final fixed-effect model's ANOVA assumptions for violations.

Mixed Effect Models:

Since `work_year` consists of only three levels 2020, 2021, and 2022, we can also consider it as a random effect as those levels represent a sample from the population of all the years to this date. Therefore, we analyzed the mixed effect model with `work_year` as the random effect, and the experience level and employment were fixed effects. As before we considered two mixed effects models: mixed effects without interactions and mixed effects with interactions. We checked if the assumption of work year normality was met in both models. We then compared Model 1 (main effects) and Model 2 (with interaction) to determine our final fixed effect model.

For the final mixed-effect model, using variances for the random effect and residuals using the REML approach, we calculated the ICC. This allowed us to find the percentage of variation attributed to the random component for each model.

We then tested the significance of the random effects and fixed effects by generating a hypothesis test using the REML approach and confidence intervals for the random effect in the final mixed effect model and the Kenward-Roger F-test for the final fixed effect model and the final mixed effect model. [4]

Goal

These statistical methods will help us address our research goal of analyzing the factors that impact Data Science job salaries because we can identify the best fixed and random effect models that fit the data. We can also evaluate the impact that the random

[4] Lee DK. (2020)

effect of the work year had on salaries based on its variance contribution. In addition, we can analyze the significance of the various random and fixed effects using hypothesis testing with exactRLRT() and Kenward-Roger F-tests.

## Results:

### Exploratory Analysis:

The box plot for the entire data (Figure 1) reveals that there are a few higher outliers in the salaries. Whether these outliers can significantly affect the model or not can be investigated in further analysis. It can be observed that the median salary is around $100,000.

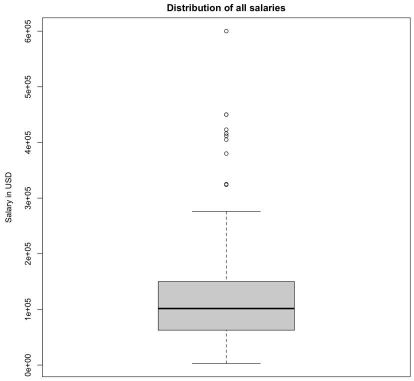

Figure 1: Box plot for Salaries in USD

We can further investigate the box plots for Employment Type, Experience Level, and Work Year and observe any visible differences in the levels of these variables.

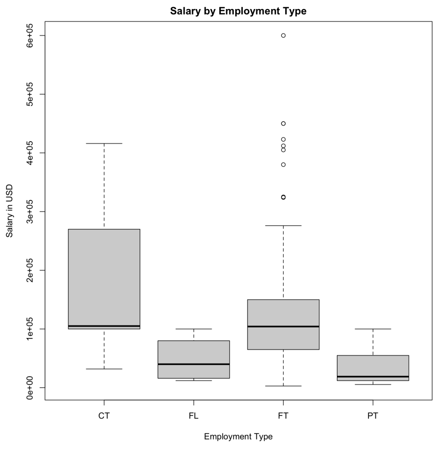

Figure 2: Box plot for Employment Type

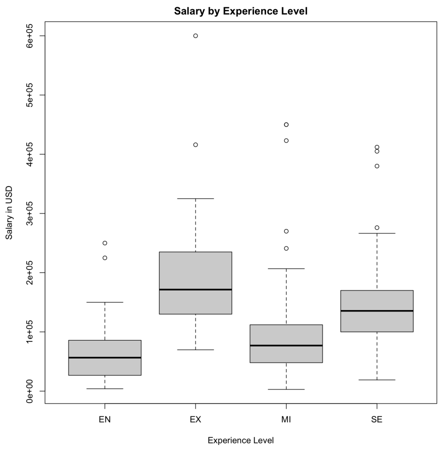

Figure 3: Box plot for Experience Level

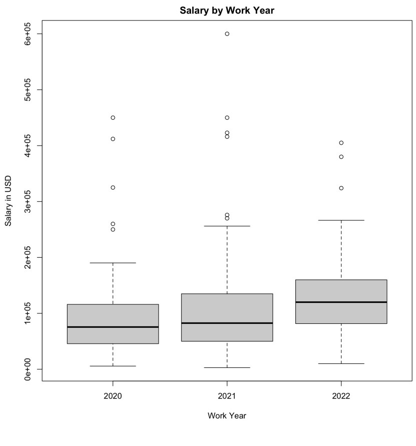

Figure 4: Box plot for Work Year

The box plots in Figure 2 show that the medians for levels CT and FT are higher than those for FL and PT. In Figure 3, the box plots reveal that the medians for EX and SE are much higher than those for EN and MI positions. In Figure 4, the box plots suggest that 2022 may have a higher median salary, however, this cannot be said for sure until we test for the difference in variance. Since this is a preliminary analysis, no conclusions can be made on whether there is a difference in the means for the different levels of Employment Type, Experience Level, and Work Year.

In addition, the interaction plots for the three variables can also be explored.

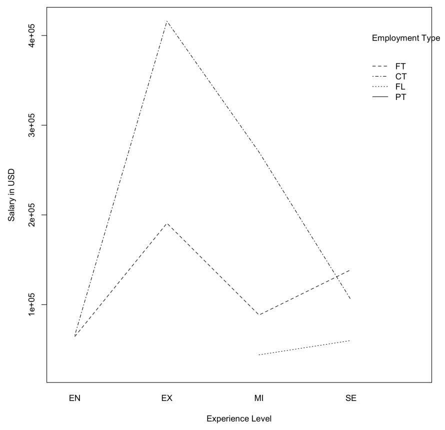

Figure 5: Interaction of Experience Level vs Employment Type

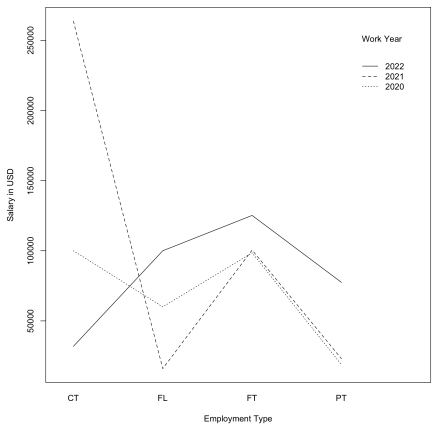

Figure 6: Interaction of Employment Type vs Work Year

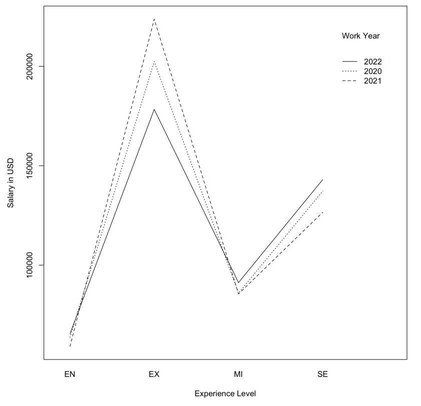

Figure 7: Interaction of Experience Level vs Work Year

In Figure 5, there seems to be an interaction between the Experience Level and Employment Type as the lines for most levels don't intersect. For example, the line for FL remains below those for FT and CT. The line for CT remains above the line FT for most levels of Experience Level.

On the contrary, Figures 6 and 7 don't reveal interactions with the variable Work Year as the lines intersect about once between each level on the x-axis.

Model Selection and Evaluation of ANOVA Assumptions**:**

Next, the assumptions of ANOVA were analyzed across the three models (regular, log transformation, and square root transformation). First, we explored the constant variance assumption for the model using residual plots. Then, we assessed the normality of each of the models using QQ plots. Finally, we looked at the Box-Cox normality plots to also assess which model is better for applying ANOVA.

Residual PlotsTorsten Hothorn, & Brian S. Everitt

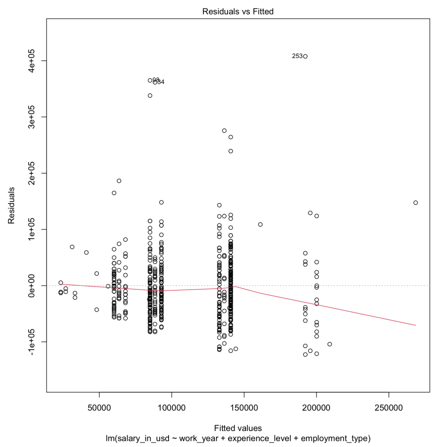

Figure 8: Residual plot for the model with no transformation

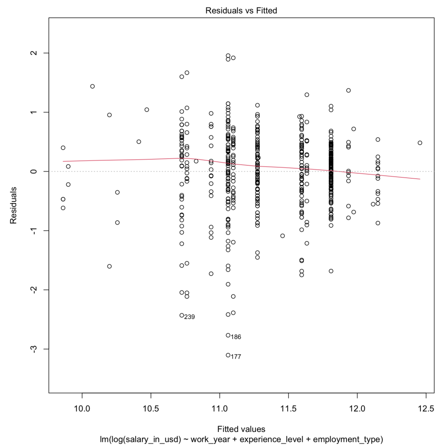

Figure 9: Residual plot for the model with log transformation

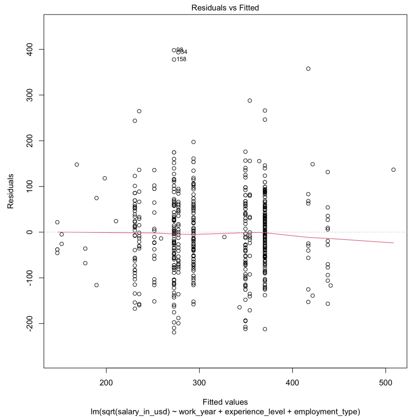

Figure 10: Residual plot for the model with square root transformation

In Figure 8, the residual plot's red line curves toward the bottom revealing a violation in the constant variance assumption. In Figure 9, the slope to the right decreases, but the line is above the 0 line. Therefore, while the log transformation improves the imbalance of the variance, it is not a perfect solution. In Figure 10, the red line gets closer to the 0 line and the slope to the right is much less than that of the one in Figure 8. Based on the residual plot alone, the square root transformation is a better model.

QQ Plots

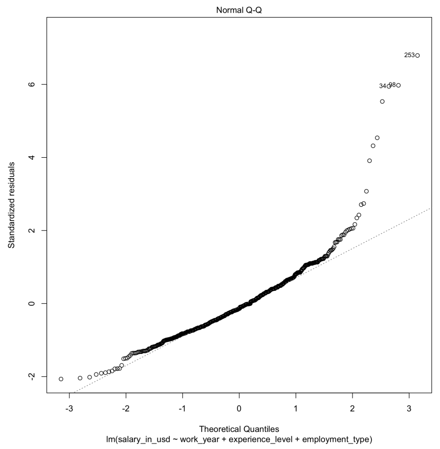

Figure 11: QQ plot for the model with no transformation

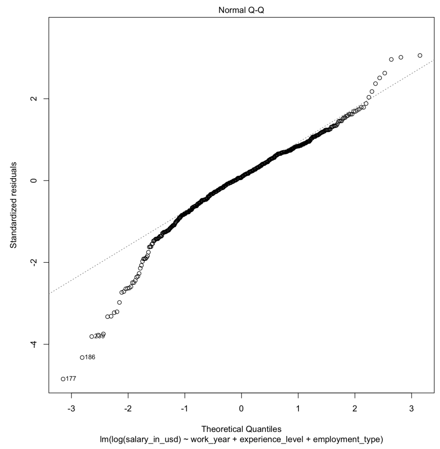

Figure 12: QQ plot for the model with log transformation

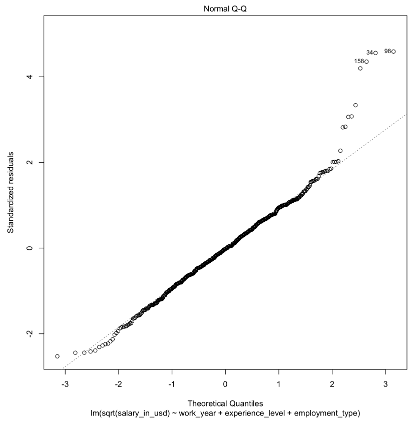

Figure 13: QQ plot for the model with square root transformation

In Figure 11, there are about 253 outliers marked in the QQ plot. These decrease to 186 for the log transformation (Figure 12). The outliers further decrease to 158 in the square root transformation (Figure 13) along with the range of the y-axis. Therefore, the square root transformation provides a model with better normality.

Box-Cox Normality Plots

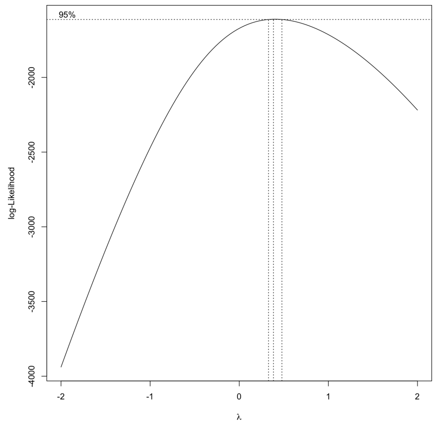

Figure 14: Box-Cox Normality plot for the model with no transformation

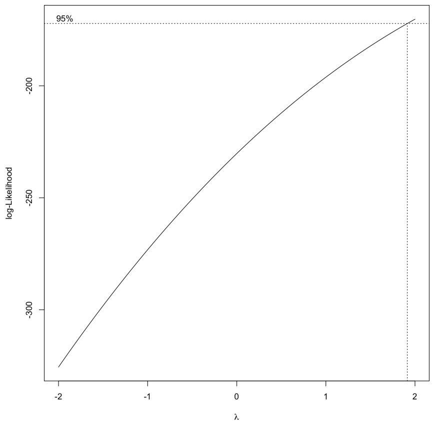

Figure 15: Box-Cox Normality plot for the model with log transformation

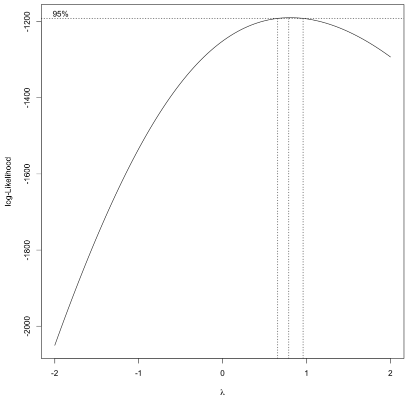

Figure 16: Box-Cox Normality plot for the model with square root transformation

The ideal confidence interval for lambda in a box-cox transformation should be 1. In Figure 14,  the confidence interval for lambda is below 1. The log transformation offsets it to a value much larger than 1 as can be seen in Figure 15. When we do a square root transformation this gets closer to 1 than the regular model (Figure 16 vs Figure 14).

With all the given points we can see that the square root transformation is appropriate as it has the least amount of outliers in the QQ plot, the residual plot has a much better equal variance, and the box-cox transformation's lambda value is closer to 1 compared to the other transformations (or no transformation).

*Henceforth*, when we talk about models in the future we will reference the square root transformation model.

Fixed-Effect Model

The fixed-effect model with main effects only resulted in all the variables being significant. However, for the fixed-effects model with interaction terms, we discovered that the highest-order interaction term was not significant. So we reconstructed the model with only the two-way interaction terms.

We then compared the two following models with LRT using the anova() function with the Chi-Squared test.

$H_0$: sqrt(salary_in_usd) ~work_year + experience_level + employment_type
$H_a$: sqrt(salary_in_usd) ~work_year+experience_level+employment_type + work_year:experience_level + work_year:employment_type + experience_level:employment_type

The p-value was 0.2546 > 0.05, so we failed to reject the null model in favor of the alternative model. In this case, the interaction effect doesn't lead to a better model, so we will consider the main-effects model as the final fixed-effects model.

We then evaluated the assumptions of the main effects (no interaction) fixed-effects model by computing a QQ plot to check for residual normality and a residual plot to check for constant variance. The plots don't reveal any ANOVA violations.

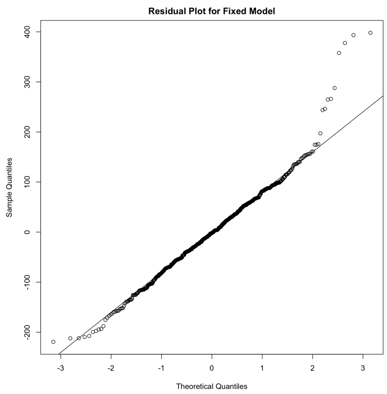

Figure 17: QQ plot of final fixed effect model error term

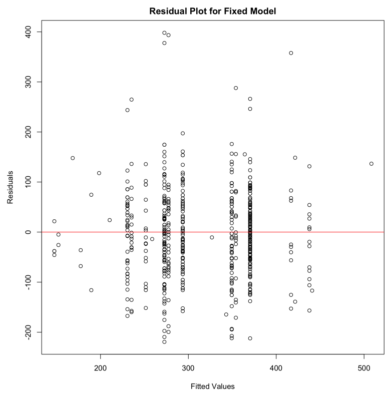

Figure 18: Residual plot of final fixed effect model

We analyzed the significance of fixed effects using exactRLRT() and Kenward-Roger F-tests and found that all three factors were statistically significant with p-values less than 0.05.

Mixed Effect Model

The first model was a mixed effect model containing just the main effects. The experience level and employment were fixed effects, and the work year was considered as the random effect. The second model we generated was a mixed-effect model with interaction. We checked if the assumption of work year normality was met in both models.

We then compared the two following models with LRT using the anova() function with the Chi-Squared test.

$H_0$: sqrt(salary_in_usd) ~experience_level+employment_type+ (1|work_year)

**$H_a$** : sqrt(salary_in_usd) ~experience_level*employment_type+ (1|work_year)

Since the p-value of 0.08407 is greater than the significance level of 0.05, we failed to reject the null model in favor of the alternative model. In this case, the interaction effect is not significant at the 0.05 significance level, so we will consider it as the final mixed effects model with the main effect only.

The next step was analyzing the final mixed-effect model. We found the variances for the random effect and residuals using the REML, and calculated the ICC as follows:

| | |
|---|---|
| Variance component of work_year: | 101.7 |
| Variance component of residuals: | 7578.9 |
| ICC (Intraclass Correlation): | (101.7) / (7578.9) = 0.01324115 |

Percentage of variation attributed to the random component: 0.01324 x 100 = 1.324%.

Since the percentage of variation attributed to work_year is 1.24%, this means that there is a lack of variability in the work years. Therefore, it is not appropriate to treat `work_year` as a random effect.

We tested the effect of the random effect using the following hypothesis:

**$H_0$**: Variance components are equal to 0
**$H_a$**: Variance components are greater than 0

Since the p-value of 0.0201 was less than the significance level of 0.05, the random effect is significant and the variance components are statistically greater than 0.

The confidence interval for the random effect and residual were calculated to be:

```
            2.5 %    97.5 %
.sig01    0.00000  27.84882
.sigma   81.95547  91.73973
```

Based on the confidence interval results, we are 95% confident that the true variance component of work_year is in the interval (0, 27.84882) and that the true variance component for residuals is in the interval (81.95547, 91.73973).

We analyzed the significance of fixed effects using exactRLRT() and Kenward-Roger F-tests and found that both the fixed effects (Employment Type and Experience Level) were statistically significant with p-values much less than 0.05.

Finally we checked the final mixed effects model for any ANOVA assumption violations.

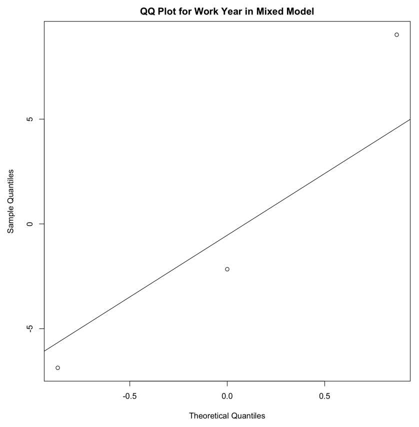

Figure 19: QQ plot for Work Year (random effect)

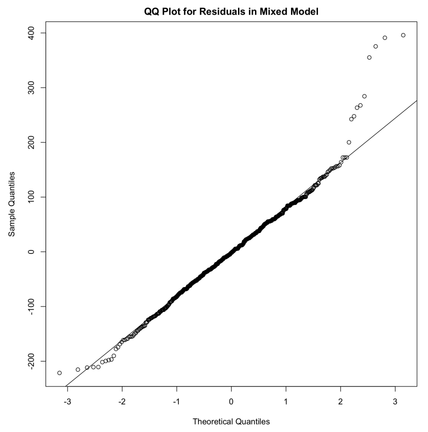

Figure 20: QQ plot of final fixed effect model residual

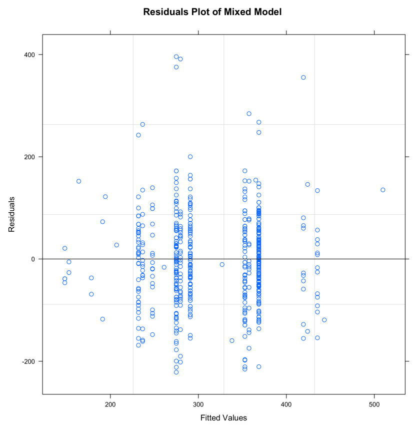

Figure 21: Residual plot of final fixed effect model

The figures above don't reveal any violations in the Mixed Effects ANOVA assumptions.

## Discussion/Conclusion:

As future data scientists in the industry, we wanted to investigate how factors such as year, experience level, or employment type would affect our future salaries.

One question we wanted to answer was whether factors such as year, experience level, and employment type had a statistically significant impact on the response variable which was salaries in USD. After applying the Kenward-Roger F-test on the final mixed effect model, we found that experience level and employment type had a statistically significant impact on salaries as their p-values were both <0.001. When we applied the exactRLRT() test to work year, we were able to determine that work year also had a statistically significant impact on salaries. Therefore, we were able to conclude that all three explanatory variables year, experience level, and employment type were significant factors that determine the salaries of data scientists.

Another question we posed was whether interactions between year, experience level, and employment had a significant impact on salaries. In order to answer this question, we conducted Likelihood Ratio Tests using the anova() function, to choose the best model and determine whether the interaction between variables was significant or not. When comparing fixed effect models we failed to reject the main effects model, so our final fixed effect model was without interaction. When comparing the mixed effect model we also failed to reject the main effects model, so our final mixed effect model was without interaction. Therefore, we were able to conclude that while all three variables of year, experience level, and employment type were significant, their interactions were not significant.

A limitation of our statistical method is that we did not assess if outliers in the data had a significant impact on our models. We would need to conduct further exploration, such as calculating Cook's distance, to determine if we need to make adjustments to the model and potentially exclude outliers that have a negative impact on our results. We also have outliers in all QQ plots which suggest a slight deviance from the Normal assumption. However, ANOVA is quite resistant to small deviances in the normality assumption. It is important to note that this analysis doesn't allow us to make conclusions about how the specific levels of the Employment Type, Experience Level, Work Year differ from each other. Another limitation of our analysis is that the percentage of variation attributed to the random component (work year) was very small at only 1.324%. Since the variance contribution of work year is very small when it's being considered a random effect, this means that we should treat it as a fixed effect instead.

This leads us to conclude that the final fixed-effects model (with no interactions) is the best model to analyze the data. Since all three main effects were significant, this leads us to make the following conclusions.

1. The final fixed-effects model shows that Experience Level did have a statistically significant effect on salaries as the p-value was less than 2e-16.
2. The final fixed-effects model shows that Employment Type did have a statistically significant effect on salaries as the p-value was 3.08e-4.
3. The final fixed-effects model shows that Work Year did have a statistically significant effect on salaries as the p-value was 1.23e-11.

## References:

[1] Oehlert, G. W. (2018). *A First Course in Design and Analysis of Experiments.* New York, NY: W. H. Freeman and Co

[2] Bhatia, R. (2022, June 15). *Data Science Job Salaries.* Kaggle. Retrieved May 4, 2023, from https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries

[3] Torsten Hothorn, & Brian S. Everitt. (2014). *A Handbook of Statistical Analyses Using R: Vol. 3rd edition*. Chapman and Hall/CRC.

[4] Lee DK. (2020)Data transformation: a focus on the interpretation. Korean J Anesthesiol. 2020 Dec; 73(6):503-508. doi: 10.4097/kja.20137. Epub 2020 Nov 20. PMID: 33271009; PMCID: PMC7714623.

## Appendix:

R code and Output:

```
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"
"Product Data Analyst" ...
##  $ 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","salary_in_usd")
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)

# View the final data
str(ds.salaries)

## 'data.frame':        607 obs. of  4 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 ...
##  $ salary_in_usd   : int  79833 260000 109024 20000 150000 72000 190000 35735 135000 125000 ...

boxplot(ds.salaries$salary_in_usd, ylab="Salary in USD", main="Distribution of all salaries")

boxplot(salary_in_usd~employment_type, data=ds.salaries, xlab="Employment Type", ylab="Salary in USD", main="Salary by Employment Type")

boxplot(salary_in_usd~experience_level, data=ds.salaries, xlab="Experience Level", ylab="Salary in USD", main="Salary by Experience Level")

boxplot(salary_in_usd~work_year,data=ds.salaries, xlab="Work Year", ylab="Salary in USD", main="Salary by Work Year")

library(lattice)
interaction.plot(ds.salaries$experience_level,ds.salaries$employment_type,ds.salaries$salary_in_usd, ylab ="Salary in USD", xlab="Experience Level", trace.label ="Employment Type")

interaction.plot(ds.salaries$employment_type,ds.salaries$work_year,ds.salaries$salary_in_usd, ylab ="Salary in USD", xlab="Employment Type", trace.label ="Work Year")

interaction.plot(ds.salaries$experience_level,ds.salaries$work_year,ds.salaries$salary_in_usd, ylab ="Salary in USD", xlab="Experience Level", trace.label ="Work Year")
```

## ANOVA Assumptions

```
lm.salaries = lm(salary_in_usd~ work_year + experience_level + employment_type, data = ds.salaries)
lm.log.salaries = lm(log(salary_in_usd)~ work_year + experience_level + employment_type, data = ds.salaries)
lm.sqrt.salaries = lm(sqrt(salary_in_usd)~ work_year + experience_level + employment_type, data = ds.salaries)

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

```
library(MASS)
boxcox(lm.salaries,lambda = seq(-2,2,length=100)) # before transformation
boxcox(lm.log.salaries,lambda = seq(-2,2,length=100)) # after SQRT transformation
boxcox(lm.sqrt.salaries,lambda = seq(-2,2,length=100)) # after SQRT transformation
```

## Fixed effect model

```
library(car)

## Loading required package: carData
```

```
model.1 = lm(sqrt(salary_in_usd) ~work_year+
experience_level+employment_type,data=ds.salaries)
fixed.1 = aov(model.1)
summary(fixed.1)
```

```
##                  Df  Sum Sq Mean Sq F value   Pr(>F)
## work_year         2  397445  198723  26.210 1.23e-11 ***
## experience_level  3 1538687  512896  67.647  < 2e-16 ***
## employment_type   3  144304   48101   6.344 0.000308 ***
## Residuals       598 4534033    7582
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model.2 = lm(sqrt(salary_in_usd)
~work_year*experience_level*employment_type,data=ds.salaries)
fixed.2 = aov(model.2)
summary(fixed.2)
```

```
##                                             Df  Sum Sq Mean Sq F value   Pr(>F)
## work_year                                    2  397445  198723  26.349 1.11e-11
## experience_level                             3 1538687  512896  68.006  < 2e-16
## employment_type                              3  144304   48101   6.378 0.000295
## work_year:experience_level                   6   20486    3414   0.453 0.843172
## work_year:employment_type                    6   74167   12361   1.639 0.133993
## experience_level:employment_type             3   42862   14287   1.894 0.129353
## work_year:experience_level:employment_type   1    7136    7136   0.946 0.331092
## Residuals                                  582 4389383    7542
##
## work_year                                  ***
## experience_level                           ***
## employment_type                            ***
## work_year:experience_level
## work_year:employment_type
## experience_level:employment_type
## work_year:experience_level:employment_type
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the highest level interaction term was not significant we can drop it and consider the two way interactions

```
model.3 = lm(sqrt(salary_in_usd) ~work_year+experience_level+employment_type +
                work_year:experience_level + work_year:employment_type +
                      experience_level:employment_type,data=ds.salaries)
fixed.3 = aov(model.3)
summary(fixed.3)
```

```
##                                   Df  Sum Sq Mean Sq F value   Pr(>F)    
## work_year                          2  397445  198723  26.352 1.11e-11 ***
## experience_level                   3 1538687  512896  68.012  < 2e-16 ***
## employment_type                    3  144304   48101   6.378 0.000294 ***
## work_year:experience_level         6   20486    3414   0.453 0.843143    
## work_year:employment_type          6   74167   12361   1.639 0.133949    
## experience_level:employment_type   3   42862   14287   1.895 0.129322    
## Residuals                        583 4396519    7541                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can compare models fixed.1 and fixed.3 where fixed.1 is the null model and fixed.3 is the alternative model

```
anova(model.1,model.3)
```

```
## Analysis of Variance Table
## 
## Model 1: sqrt(salary_in_usd) ~ work_year + experience_level + employment_type
## Model 2: sqrt(salary_in_usd) ~ work_year + experience_level + employment_type + 
##     work_year:experience_level + work_year:employment_type + 
##     experience_level:employment_type
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    598 4534033                           
## 2    583 4396519 15    137514 1.2157 0.2546
```

Since the p-value of 0.2546 is greater than significance level of 0.05, we fail to reject the null model model.1 in favor of the alternative model model.3. In this case the interaction effect is not significant, so we will consider model.1 as the final fixed effects model.

### Assumptions

```
plot(fixed.1$residuals ~ fixed.1$fitted.values, xlab="Fitted Values", ylab="Residuals", 
main="Residual Plot for Fixed Model")
abline(h=0,col="red")

residuals = resid(fixed.1)
qqnorm(residuals,main="Residual Plot for Fixed Model")
qqline(residuals(fixed.1))

par(mfrow=c(1,2))
```

## Random Effects

Since we picked the model without interactions, we will now consider if `work_year` can be treated as a random effect. The study only pulled data from 2020,2021,and 2022 so we can say that the levels represent a random sample from all the years.

```
library(lme4)

## Loading required package: Matrix

mixed.1 = lmer(sqrt(salary_in_usd) ~experience_level+employment_type+ (1|work_year), data
=ds.salaries)
summary(mixed.1)

## Linear mixed model fit by REML ['lmerMod']
## Formula: sqrt(salary_in_usd) ~ experience_level + employment_type + (1 |
##        work_year)
##        Data: ds.salaries
##
## REML criterion at convergence: 7088.8
##
## Scaled residuals:
##        Min      1Q  Median      3Q       Max
## -2.5435 -0.6216 -0.0153  0.6367  4.5478
##
## Random effects:
##  Groups       Name                Variance Std.Dev.
##  work_year (Intercept)  101.7   10.08
##  Residual                    7578.9   87.06
## Number of obs: 607, groups:  work_year, 3
##
## Fixed effects:
##                          Estimate Std. Error t value
## (Intercept)              329.14            40.03   8.223
## experience_levelEX   187.51           19.65   9.542
## experience_levelMI42.98    11.29   3.807
## experience_levelSE   120.94           11.27  10.733
## employment_typeFL   -186.86         58.72  -3.182
## employment_typeFT           -90.35  39.34  -2.297
## employment_typePT   -174.10        47.91  -3.634
##
## Correlation of Fixed Effects:
##        (Intr) exp_EX exp_MI exp_SE emp_FL emp_FT
## exprnc_lvEX -0.146
## exprnc_lvMI -0.140  0.414
## exprnc_lvSE -0.145  0.432  0.742
## emplymnt_FL -0.639  0.019 -0.085 -0.058
## emplymnt_FT -0.960  0.031 -0.063 -0.063  0.666
## emplymnt_PT -0.806  0.093  0.045  0.065  0.540  0.805
```

#Assumptions

```
#Random effect - work year normality

random.work.year.1 = ranef(mixed.1)$work_year[["(Intercept)"]]

qqnorm(random.work.year.1, main="Work Year")
qqline(random.work.year.1)
```

Variance component of work_year: 101.7 Variance component of residuals: 7578.9

Compute the ICC (Intraclass Correlation):

Determine the percentage of variation attributed to the random component. Percentage of variation attributed to random component= ICC x 100

0.01324115 x 100 = 1.324%

```
mixed.2 = lmer(sqrt(salary_in_usd) ~experience_level*employment_type+ (1|work_year), data
=ds.salaries)

## fixed-effect model matrix is rank deficient so dropping 4 columns / coefficients

summary(mixed.2)

## Linear mixed model fit by REML ['lmerMod']
## Formula: sqrt(salary_in_usd) ~ experience_level * employment_type + (1 |
##     work_year)
##     Data: ds.salaries
##
## REML criterion at convergence: 7024.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5361 -0.6282 -0.0090  0.6270  4.5837
##
## Random effects:
##  Groups    Name           Variance Std.Dev.
##  work_year (Intercept)  112.7   10.62
##  Residual               7518.4   86.71
## Number of obs: 607, groups:  work_year, 3
##
## Fixed effects:
##                                    Estimate Std. Error t value
## (Intercept)                         243.500     61.660   3.949
## experience_levelEX                  409.243    106.416   3.846
## experience_levelMI                  283.878    106.416   2.668
## experience_levelSE                   88.300    106.416   0.830
## employment_typeFL                   -85.166    122.972  -0.693
## employment_typeFT                    -2.491     62.197  -0.040
## employment_typePT                   -85.679     69.624  -1.231
## experience_levelMI:employment_typeFL -251.121    158.540  -1.584
## experience_levelEX:employment_typeFT -229.262    108.306  -2.117
## experience_levelMI:employment_typeFT -243.750    107.077  -2.276
## experience_levelSE:employment_typeFT   30.803    107.111   0.288
## experience_levelMI:employment_typePT -248.986    122.038  -2.040
##
## Correlation of Fixed Effects:
##             (Intr) exp_EX exp_MI exp_SE emp_FL emp_FT emp_PT e_MI:_FL e_EX:_
## exprnc_lvEX -0.575
## exprnc_lvMI -0.575  0.336
## exprnc_lvSE -0.575  0.336  0.336
## emplymnt_FL  0.002 -0.004 -0.004 -0.579
## emplymnt_FT -0.982  0.572  0.572  0.572 -0.004
## emplymnt_PT -0.878  0.511  0.511  0.511 -0.003  0.871
## expr_MI:_FL -0.001  0.003 -0.443  0.448 -0.775  0.002  0.002
## expr_EX:_FT  0.565 -0.983 -0.330 -0.330  0.004 -0.576 -0.502 -0.002
## expr_MI:_FT  0.572 -0.334 -0.994 -0.334  0.004 -0.583 -0.508  0.441    0.337
## expr_SE:_FT  0.572 -0.335 -0.335 -0.994  0.575 -0.583 -0.508 -0.445    0.338
## expr_MI:_PT  0.501 -0.292 -0.871 -0.292  0.002 -0.498 -0.571  0.387    0.288
##             e_MI:_FT e_SE:_
## exprnc_lvEX
## exprnc_lvMI
## exprnc_lvSE
## emplymnt_FL
## emplymnt_FT
## emplymnt_PT
## expr_MI:_FL
## expr_EX:_FT
## expr_MI:_FT
## expr_SE:_FT  0.342
## expr_MI:_PT  0.866 0.291
## fit warnings:
## fixed-effect model matrix is rank deficient so dropping 4 columns / coefficients
```

#Assumptions

```
#Random effect - work year normality

random.work.year.2 = ranef(mixed.2)$work_year[["(Intercept)"]]

qqnorm(random.work.year.2, main="Work Year")
qqline(random.work.year.2)
```

Variance component of work_year: 112.7 Variance component of residuals: 7518.4

Compute the ICC (Intraclass Correlation):

Determine the percentage of variation attributed to the random component: Percentage of variation attributed to random component= ICC x 100

0.01476851 x 100 = 1.477%

```
anova(mixed.1,mixed.2)

## refitting model(s) with ML (instead of REML)

## Data: ds.salaries
## Models:
## mixed.1: sqrt(salary_in_usd) ~ experience_level + employment_type + (1 | work_year)
## mixed.2: sqrt(salary_in_usd) ~ experience_level * employment_type + (1 | work_year)
##         npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## mixed.1    9 7159.4 7199.1 -3570.7   7141.4
## mixed.2   14 7159.7 7221.5 -3565.9   7131.7 9.7039  5    0.08407 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value of 0.08407 is greater than significance level of 0.05, we fail to reject the null model mixed.1 in favor of the alternative model mixed.3. In this case the interaction effect is not significant at the 0.05 significance level, so we will consider mixed.1 as the final mixed effects model.

### Mixed Effects Assumptions

```
# Random Effect Normality
random = ranef(mixed.1)$work_year[["(Intercept)"]]
```

```
qqnorm(random,main="QQ Plot for Work Year in Mixed Model")
qqline(random)

# Residuals - Normality
residuals = resid(mixed.1)
qqnorm(residuals,main="QQ Plot for Residuals in Mixed Model")
qqline(residuals(mixed.1))

# Residuals - Constant Variance
plot(mixed.1,xlab ="Fitted Values",ylab="Residuals",main="Residuals Plot of Mixed Model")
```

## Testing the significance of the Random and Fixed effects

### *Random effect*

Ho: Variance components are equal to 0 Ha: Variance components are greater than 0

```
library(RLRsim)
exactRLRT(mixed.1)

##
##  simulated finite sample distribution of RLRT.
##
##  (p-value based on 10000 simulated values)
##
## data:
## RLRT = 3.0806, p-value = 0.0183
```

Since the p-value of 0.0201 is less than the significance level of 0.05, the random effect is significant/variance components statistically greater than O.

#Confidence intervals

```
confint(mixed.1)

## Computing profile confidence intervals ...

##                        2.5 %    97.5 %
## .sig01             0.00000  27.84882
## .sigma            81.95547  91.73973
## (Intercept)              250.45629 406.26135
## experience_levelEX  149.65916 226.57541
## experience_levelMI   21.29233  65.55629
## experience_levelSE   99.54519 144.53503
## employment_typeFL  -301.56769 -72.18910
## employment_typeFT  -166.64727 -12.89700
## employment_typePT  -267.39921 -80.21688
```

### *Fixed Effect*

```
##Final fixed effects model
library(car)
car::Anova(model.1, test="F")

## Anova Table (Type II tests)
##
## Response: sqrt(salary_in_usd)
##                    Sum Sq  Df F value       Pr(>F)
## work_year               53729   2  3.5432 0.0295294 *
## experience_level 1406379   3 61.8298 < 2.2e-16 ***
## employment_type   144304   3  6.3441 0.0003077 ***
## Residuals     4534033 598
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Final mixed effects model
library(car)
car::Anova(mixed.1, test="F")

## Analysis of Deviance Table (Type II Wald F tests with Kenward-Roger df)
##
## Response: sqrt(salary_in_usd)
##                    F Df Df.res       Pr(>F)
## experience_level 61.8978  3 572.67 < 2.2e-16 ***
## employment_type   6.3782  3 598.55 0.0002935 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```