Multivariate analysis of data science job salaries

Final Project - STAT 3690, Winter 2022

Andrea Unrau, student no. 7731916

Contents

Abstract

The data set contains several hundred Data Science-related job listings in the United States scraped from the website Glassdoor. The purpose of this research was to determine which of a particular subset of explanatory variables had an effect on the Lower, Upper, and Average salaries of the jobs. 1-way and 2-way MANOVA, multivariate linear regression, and testing for nested models was used to identify the best explanatory variables for a model. After selecting a model, Cook's distance and normality of residuals were tested to verify the appropriateness of the model. The results were that the Sector, Location and Size of the company advertising the jobs were good predictors of Upper and Lower salaries. This information could be used by any person seeking a job in this field in the United States who wants to maximize their salary.

Introduction

The dataset of Data Scientist Salaries was obtained from Kaggle and can be found at the URL https://www.kaggle.com/datasets/nikhilbhathi/data-scientist-salary-us-glassdoor?select=data_cleaned_2021.csv. It contains 742 Data Science-related job listings in the United States scraped from the website Glassdoor in December 2021. Examples of the job titles are: Data Scientist; Data Analyst; Data Engineer; Data Analytics Manager; and Machine Learning Engineer. It contains 41 variables including: Company name, Location, Size, Sector, Revenue, Age, and an estimate of the Lower, Upper and Average salary for each job posting. 16 of the variables are specific skill sets that employers could potentially mention in their job listing. Some of these are Python, AWS, SQL, SAS, Tableau, etc.

My aim was to answer the questions: Does the Location, Size, Sector, Revenue or Age of the company predict the Lower, Upper and Average salary of the job being advertised? Which of those variables are the best predictors of salary?

A similar analysis was performed by Junting Lai and his research also focused on job titles. Using exploratory analysis and Backward-stepwise Regression, he found that giant companies tend to only pay higher salaries for Data Scientists whereas small to medium sized businesses paid higher salaries for other job titles. He also found that salaries were certainly dependent on the location of the job and that job titles including the phrase "Machine Learning" increased the salary by over 10%.

An analysis was done by Nikhil Bhathi on the same dataset as this research and his analysis included looking at the effects of skills requested in the job descriptions. Some of his findings were that the largest companies had the highest average salaries and there was a decline in average salary as companies got smaller. Company revenue had mixed effects on the average salary, and the top 3 industries with the most data sciene jobs

were: Biotech & Pharmaceuticals (15.09%), Insurance Carriers (8.49%) and Computer Hardware & Software (7.95%).

Exploratory Data Analysis

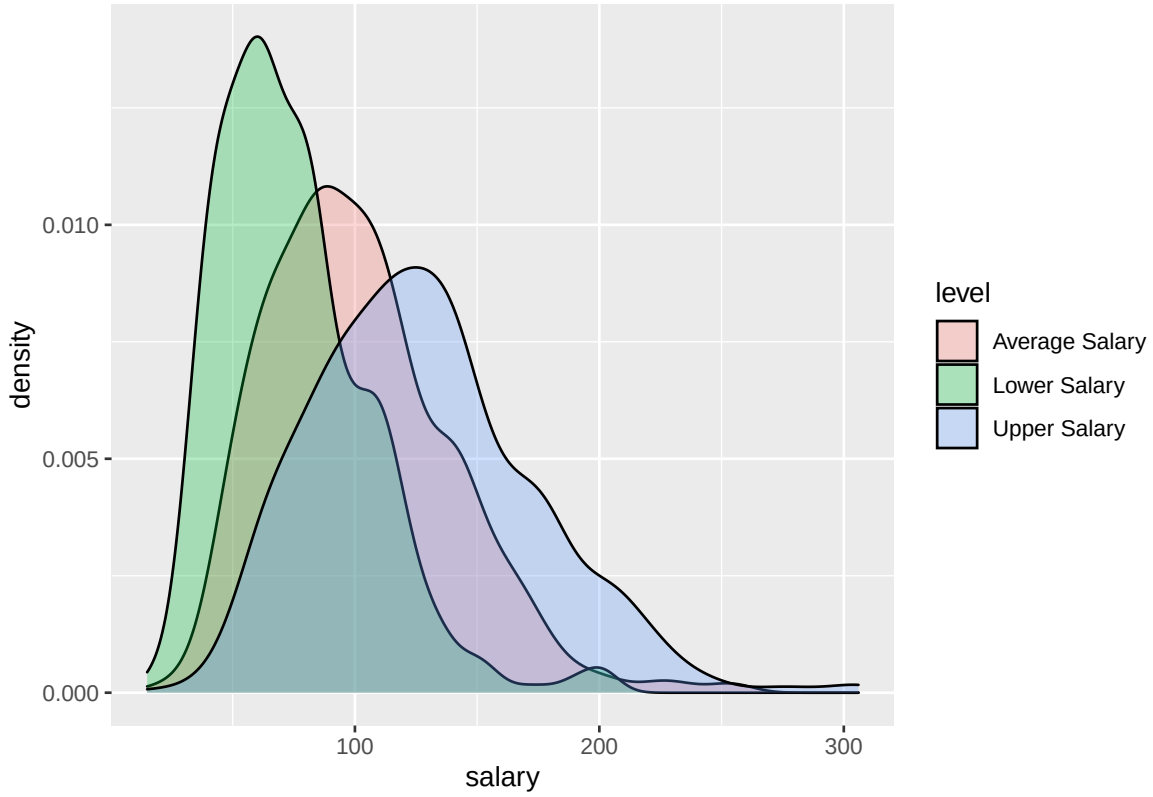
From the summary of the Lower, Upper and Average salaries it can be seen that the means are $74,800, $128,000 and $101,500 respectively. The mean of each salary is slightly higher than the medians, indicating skewness in the upper ranges of the salaries. The minimums are all around $15,000 which is very low for a salary so these should be investigated for correctness.

```
##   Lower.Salary     Upper.Salary   Avg.Salary.K.
##  Min.   : 15.00   Min.   : 16.0   Min.   : 15.5
##  1st Qu.: 52.00   1st Qu.: 96.0   1st Qu.: 73.5
##  Median : 69.50   Median :124.0   Median : 97.5
##  Mean   : 74.75   Mean   :128.2   Mean   :101.5
##  3rd Qu.: 91.00   3rd Qu.:155.0   3rd Qu.:122.5
##  Max.   :202.00   Max.   :306.0   Max.   :254.0
```

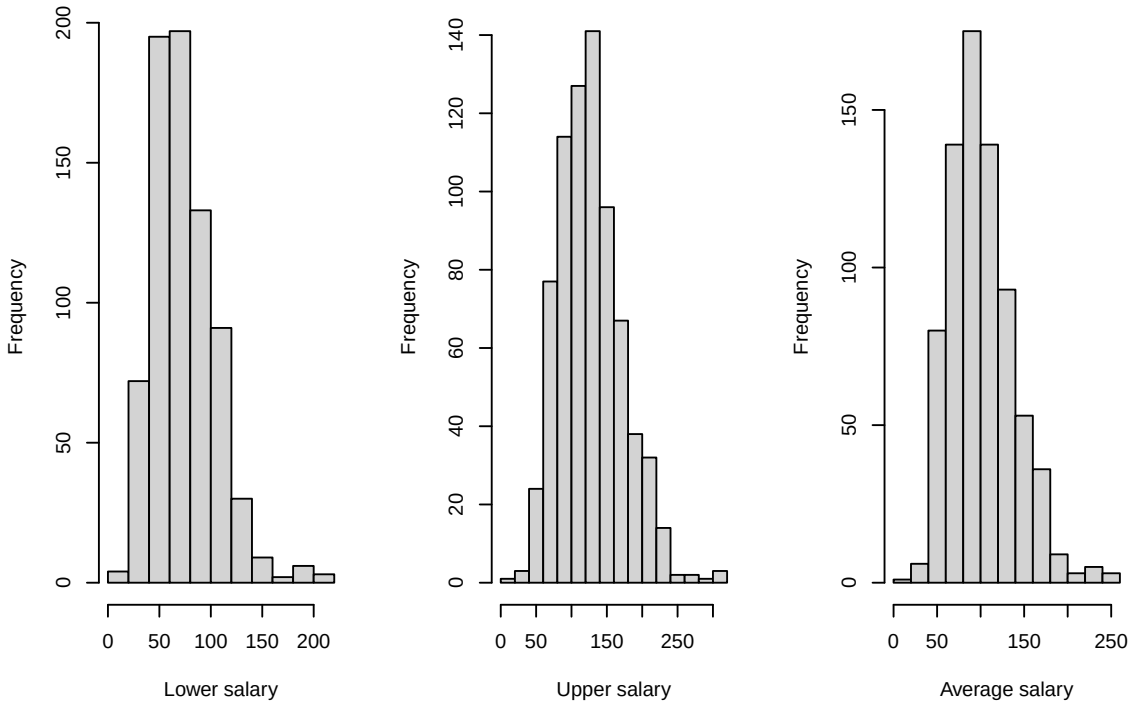
The variables Location, Size, Sector and Revenue are all factor-type variables, so only Lower salary, Upper salary, Average salary and Age could be considered for the correlation matrix. I was interested in the relationship between Age and each of the salary variables, and from the matrix I found that the correlations were all less than 0.035. This indicates an insignificant relationship.

```
##                Lower.Salary Upper.Salary Avg.Salary.K.         Age
## Lower.Salary    1.000000000   0.93999479    0.97867923 0.003010269
## Upper.Salary    0.939994790   1.00000000    0.99003173 0.034606622
## Avg.Salary.K.   0.978679229   0.99003173    1.00000000 0.022075740
## Age             0.003010269   0.03460662    0.02207574 1.000000000
```

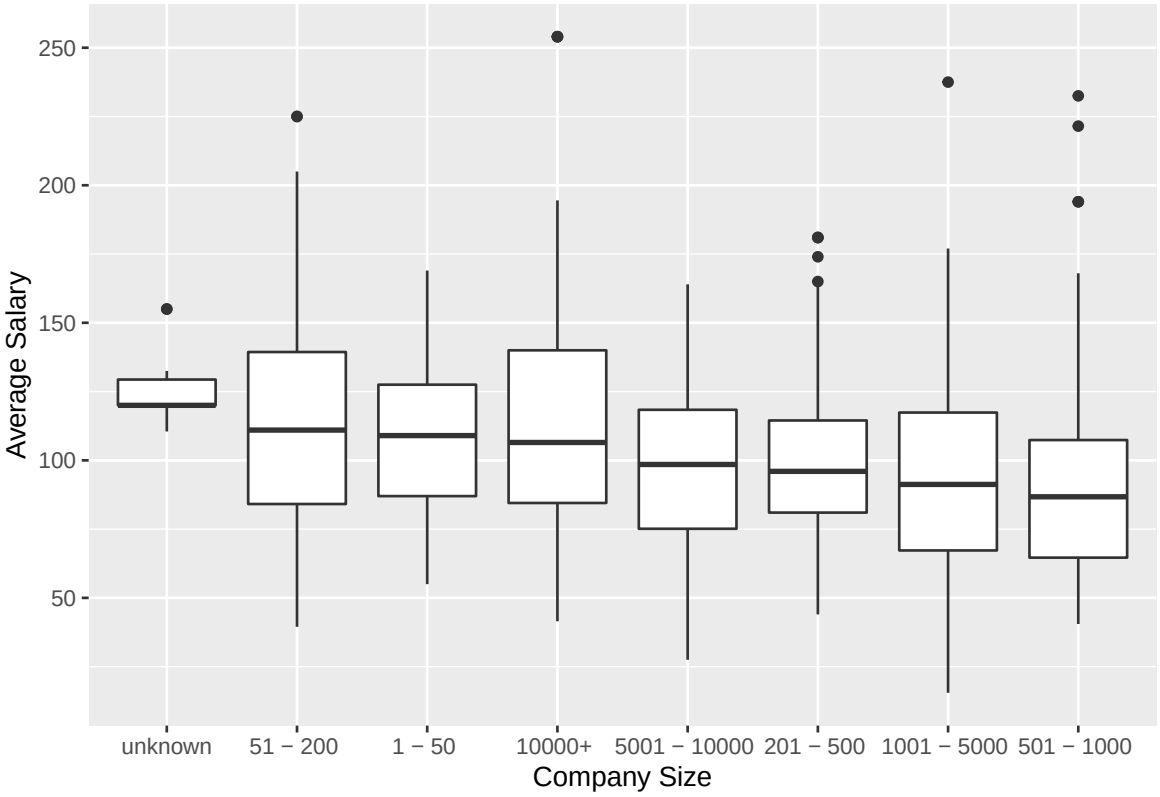
Comparing the density curves of the three salary estimates, it seems that there is more variation in the Upper salaries than in the Lower salaries, as the curve is not as high and is definitely wider. It also appears that $\approx 50\%$ of the Lower and Upper salaries overlap.



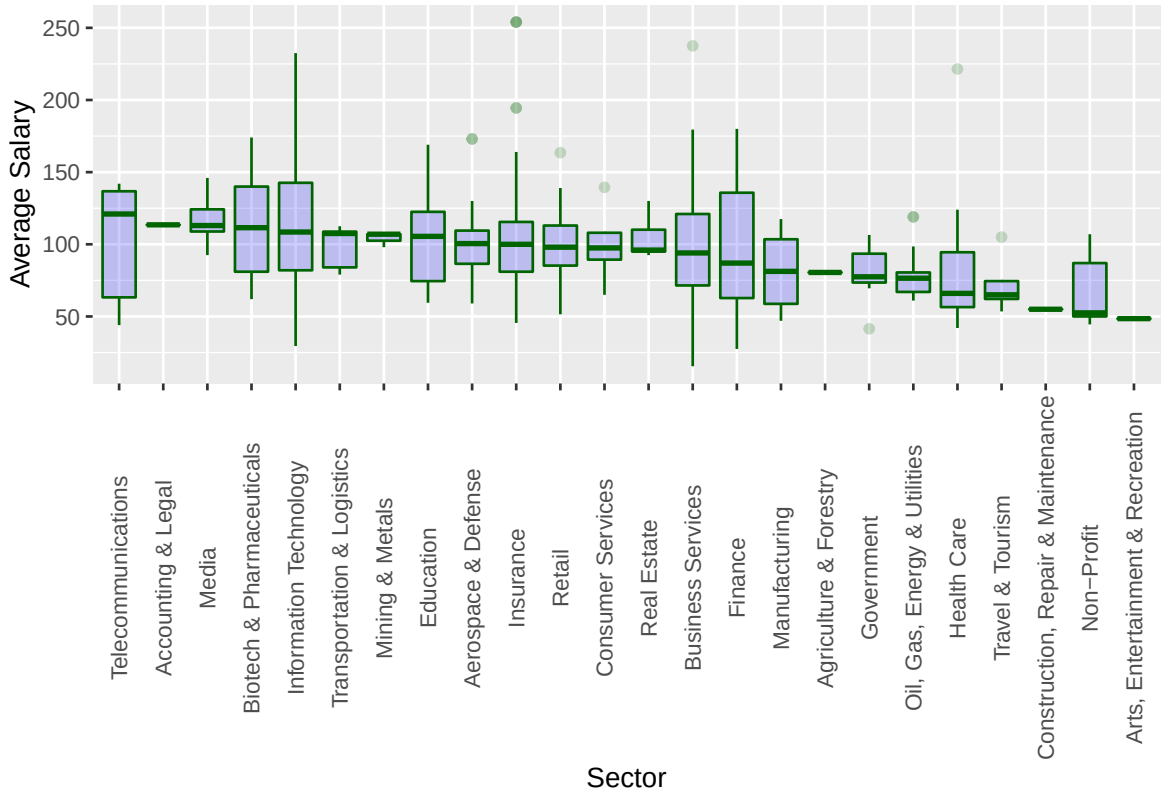
The Lower, Upper and Average salaries do not appear to be normally distributed as seen in these histograms. There is some right skewness occurring in all three.



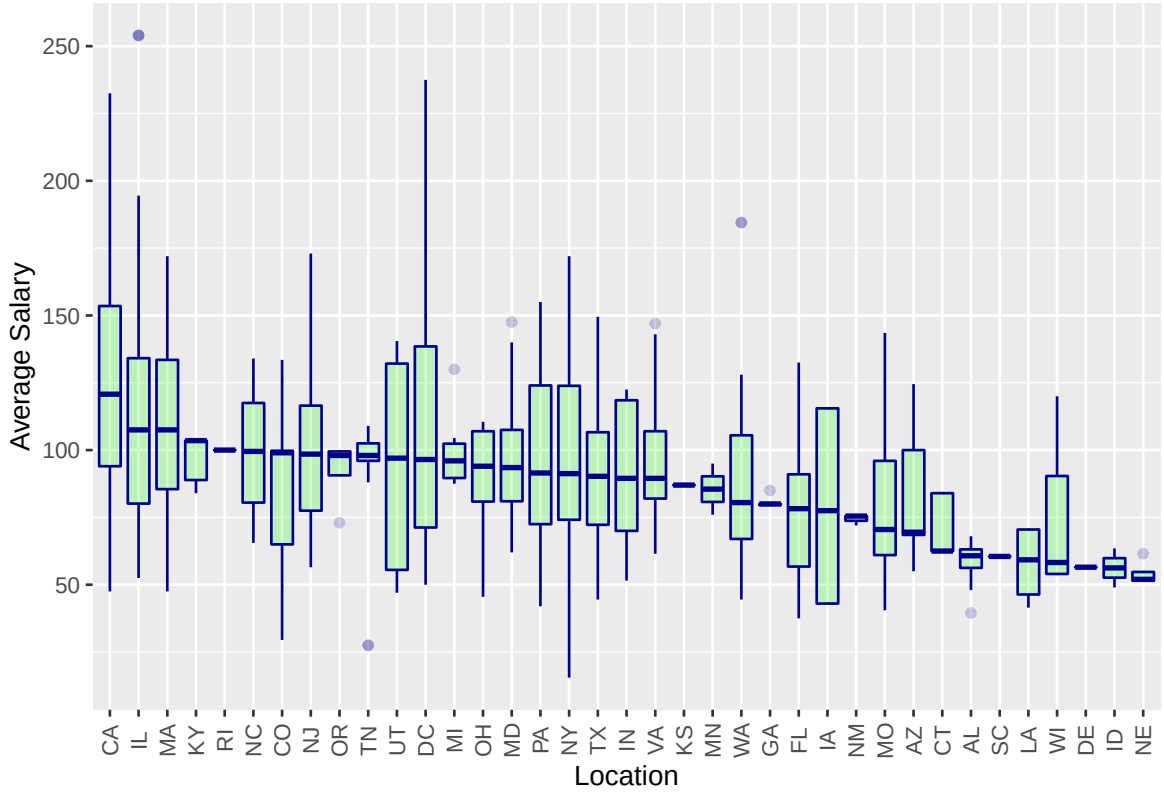
Investigating the Average salaries by company Size, this boxplot reveals something somewhat unexpected - the largest companies do not pay the highest salaries. Company sizes of 51-200 and 1-50 actually pay the highest average salary, and company sizes of 10,000+ and 5,001-10,000 follow closely behind. There is more variation though in the average salaries paid by companies of size 51-200 as seen by the length of the whiskers of its boxplot. The interquartile range of the company size 51-200 looks nearly the same as that of the company size 10,000+. The outliers are for manager-type job positions. The company size 501-1,000 had the lowest average salary.



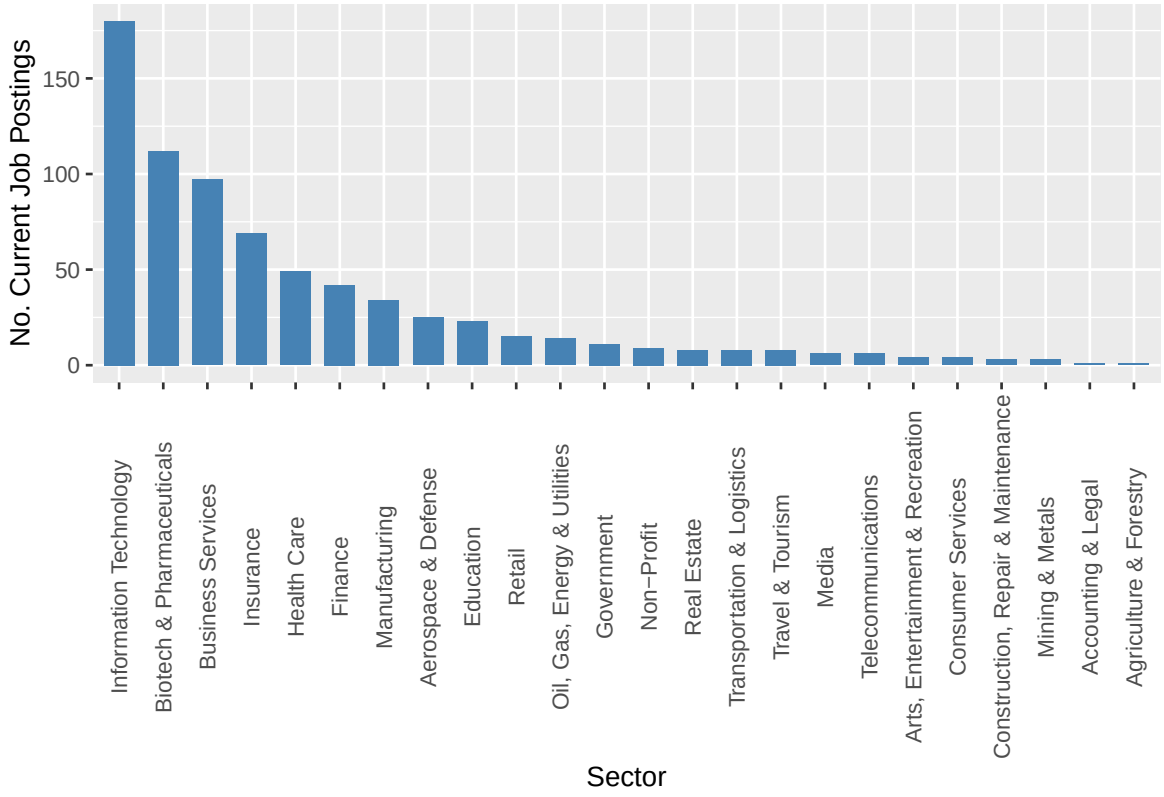
Looking at the Average salaries by sector I found that Telecommunications has the highest median average salary. This sector along with Finance have the widest interquartile ranges. Accounting & Legal, Media, and Pharmaceuticals have the next highest median average salaries. There is a lot of variation in the average salary in Information Technology indicated by the very long whiskers. The sector Arts, Entertainment & Recreation had the lowest median average salary.



The five states paying the highest average salaries are California, Illinois, Massachusetts, Kentucky, and Rhode Island. The states of Utah, Iowa and Washington, D.C all appear to have the widest interquartile ranges. Nebraska had the lowest average salary.

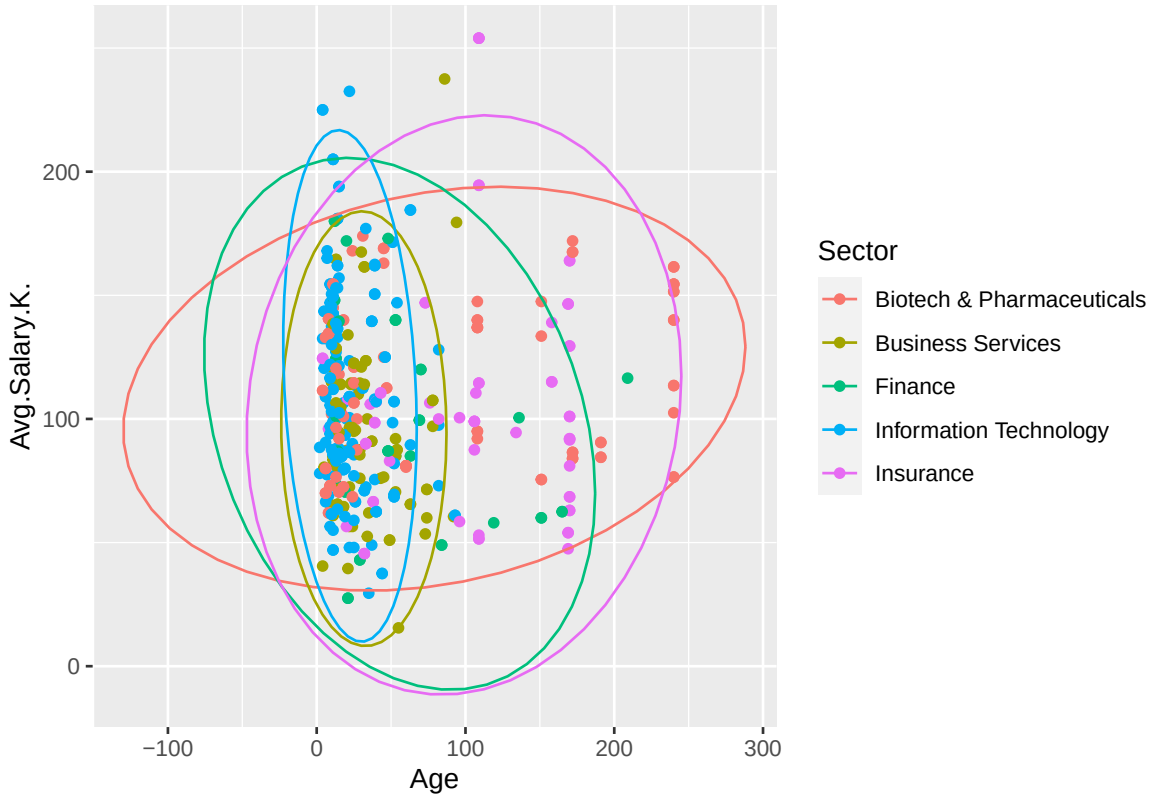


By far, the sector Information Technology had the most job postings, over 175, and Biotech & Pharmaceuticals, Business Services, and Insurance were the next three sectors with the most job postings, ranging from about 110 to 70 respectively. All remaining 20 sectors had less than 50 job postings each.



I looked at the relationship between the five sectors with the most job postings (which accounted for 71% of the job postings), Age, and Average salary. The older companies tend to be in the sectors Biotech & Pharmaceuticals, Insurance and Finance, and the younger companies are in Business Services and Information Technology. This was not a surprise.

The ellipses represent 95% confidence levels. Judging by the fact that the dots for the older companies are largely concentrated in the same salary range as the younger companies, it didn't seem that I could conclude that Age and Sector together have an effect on Average Salary.



Methods

To begin, I tested the means of the Lower, Upper and Average salaries to see if they were equal to the rounded means of the sample. This was to determine if the sample could give a good indication of the true mean salaries. To test μ, the assumption is $X_1, ..., X_n \overset{iid}{\sim} \text{MVN}_p(\mu, \Sigma)$ and $n > p$.
For a 1-way MANOVA test, it is assumed that the observations of each of the m samples are $X_{1m}, ..., X_{nm} \overset{iid}{\sim} \text{MVN}_p(\mu_m, \Sigma)$. I tested Lower and Upper salaries using 1-way MANOVA on these four factors individually: Size, Sector, Revenue, and Location. This

was to determine whether the levels of these factors had an effect on Lower and Upper salaries.
For a 2-way MANOVA test, the model is $X_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + E_{ijk}$ where $E_{ijk} \overset{iid}{\sim} MVN_p(\mathbf{0}, \mathbf{\Sigma}), i = 1, ..., m, j = 1, ..., b, k = 1, ..., n$. There is also an identifiability assumption:

$$\sum_i \tau_i = \sum_j \beta_j = \sum_i \gamma_{ij} = 0.$$

I tested all six pairs of the variables Size, Sector, Revenue, and Location to get an idea of whether each variable individually had an effect on Lower and Upper salaries, as well as whether the interaction of each pair of variables had an effect on Lower and Upper salaries.
Due to the assumption for 1-way MANOVA tests, the equality of covariance matrices was tested for Size with its eight levels and Revenue with its 13 levels. It is assumed that the observations of each of the m samples are $X_{1m}, ..., X_{nm} \overset{iid}{\sim} \text{MVN}_p(\mu_m, \Sigma_m)$.
Since Revenue, Sector, Size and Location are all factors with at least 8 factors, using multivariate linear regression for them was beyond my ability to interpret and analyze. Age is a continuous variable though, so multivariate linear regression of Age on Lower and Upper salaries was interpretable. It is assumed that $[Y_{i1}, ..., Y_{ip}, X_{i1}, ..., X_{iq}]^T \overset{iid}{\sim} [Y_1, ..., Y_p, X_1., , , .X_q]^T$ as well as $E_i \overset{iid}{\sim} (0_p, \Sigma)$.
I tested nested models to find the best explanatory variables using the AIC and BIC information criteria and a forward selection method.
To test the quality of the selected model, I investigated the leverage of the observations, calculated and analyzed Cook's Distance, and checked the normality of the residuals.
I performed Principle Component Analysis to see if the dimension of the selected model could be reduced. The assumptions for PCA are that there is a matrix of explanatory variables $X = [X_1, ..., X_n]^T_{n \times p}$ and that for each row of the matrix, $X_i \overset{iid}{\sim} (\mu, \Sigma)$.
Finally, I tested to see if the Lower and Upper salaries were MVN distributed.

Results

Testing on μ

The sample means for Lower, Upper and Average salaries were: \$74,750, \$128,210, and \$101,480 respectively. Since there was no previous data, I assumed σ^2 was unknown.

```
## [1] "The test statistic is 81583762780979248"
```

```
## [1] "The critical point is 7.87210828586338"
```

```
## [1] "The p-value is 0"
```

Report: Testing hypotheses $H_0 : \mu = [75, 130, 100]^T$ v.s. $H_1 : \mu \neq [75, 130, 100]^T$. We carried out the LRT and obtained a test statistic > 8,158,000. The p-value was 0 and the corresponding rejection region was $[7.872, \infty)$. At a significance level of 0.05, we rejected H_0 and concluded there was strong evidence the mean vector was not $[75, 130, 100]^T$.

1-way MANOVA

Here I investigated the response variables Lower and Upper salary, and the factor Size. Size has 8 levels: unknown, 1 - 50, 51 - 200, 201 - 500, 501 - 1,000, 1,001 - 5,000, 5,001 - 10,000, and 10,000+. I wished to test if the means of the Lower and Upper salaries were the same for each factor level of Size.

```
##            Df   Wilks approx F num Df den Df    Pr(>F)
## size        7 0.83757   9.7042     14   1466 < 2.2e-16 ***
## Residuals 734
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Report: Testing hypotheses H_0 : no company Size effect on Lower and Upper salaries v.s. H_1 : otherwise. We carried out the Wilk's lambda test and obtained 0.838 as the value of the test statistic. The p-value was ≈ 0, and the corresponding rejection region was $(-\infty, 0.9683]$. At a significance level of 0.05, we rejected H_0 and concluded there was strong evidence that there was an effect from company Size on Lower and Upper salaries.

Additionally using 1-way MANOVA, I tested the response variables Lower and Upper salary with the Sector factor. Sector has a total of 25 levels.

```
##            Df   Wilks approx F num Df den Df    Pr(>F)
## sector     24 0.74516    4.727     48   1432 < 2.2e-16 ***
## Residuals 717
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Report: Testing hypotheses H_0 : no Sector effect on Lower and Upper salaries v.s. H_1 : otherwise. We carried out the Wilk's lambda test and obtained 0.745 as the value of the test statistic. The p-value was ≈ 0, and the corresponding rejection region was $(-\infty, 0.9143]$. At a significance level of 0.05, we rejected H_0 and concluded there was strong evidence that there was an effect from Sector on Lower and Upper salaries.

In the same manner, I tested the effect of Revenue on Lower and Upper salary. Revenue has 13 levels.

```
##            Df  Wilks approx F num Df den Df    Pr(>F)
## revenue    12 0.8862   3.7774     24   1456 2.448e-09 ***
## Residuals 729
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Report: Testing hypotheses H_0 : no Revenue effect on Lower and Upper salaries v.s. H_1 : otherwise. We carried out the Wilk's lambda test and obtained 0.886 as the value of the test statistic. The p-value was ≈ 0, and the corresponding rejection region was $(-\infty, 0.9516]$. At a significance level of 0.05, we rejected H_0 and concluded there was strong evidence that there was an effect from Revenue on Lower and Upper salaries.

Likewise, I tested the effect of Location on Lower and Upper salary. Location has a 37 levels.

```
##            Df   Wilks approx F num Df den Df    Pr(>F)
## location   36 0.73574    3.243     72   1408 < 2.2e-16 ***
## Residuals 705
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Report: Testing hypotheses H_0 : no Location effect on Lower and Upper salaries v.s. H_1 : otherwise. We carried out the Wilk's lambda test and obtained 0.73574 as the value of the test statistic. The p-value was ≈ 0, and the corresponding rejection region was $(-\infty, 0.8792971]$. At a significance level of 0.05, we rejected H_0 and concluded there was strong evidence that there was an effect from Location on Lower and Upper salaries.

2-way MANOVA

I tested all six pairs of the variables Size, Sector, Revenue, and Location on Lower and Upper salaries.

Testing Lower and Upper salaries ~ Revenue and Sector:

```
##                  Df   Wilks approx F num Df den Df    Pr(>F)
## revenue          12 0.84784   4.4736     24   1248 6.454e-12 ***
## sector           24 0.70670   4.9282     48   1248 < 2.2e-16 ***
## revenue:sector   80 0.68554   1.6206    160   1248 6.728e-06 ***
## Residuals       625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing Lower and Upper salaries ~ Revenue and Size:

```
##                Df   Wilks approx F num Df den Df    Pr(>F)
## revenue        12 0.86900   4.2061     24   1388 6.056e-11 ***
## size            7 0.82711   9.8702     14   1388 < 2.2e-16 ***
## revenue:size   27 0.82757   2.5511     54   1388 1.027e-08 ***
## Residuals     695
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing Lower and Upper salaries ~ Sector and Size:

```
##               Df   Wilks approx F num Df den Df    Pr(>F)
## sector        24 0.69140   5.4629     48   1294 < 2.2e-16 ***
## size           7 0.86839   6.7570     14   1294 1.794e-13 ***
## sector:size   62 0.69275   2.1024    124   1294 2.392e-10 ***
## Residuals    648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing Lower and Upper salaries ~ Location and Size:

```
##                 Df   Wilks approx F num Df den Df    Pr(>F)
## location        36 0.65418   3.9462     72   1202 < 2.2e-16 ***
## size             7 0.78892  10.8059     14   1202 < 2.2e-16 ***
## location:size   96 0.57597   1.9886    192   1202 5.135e-12 ***
## Residuals      602
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing Lower and Upper salaries ~ Location and Sector:

```
##                   Df   Wilks approx F num Df den Df    Pr(>F)
## location          36 0.61792   4.3164     72   1142 < 2.2e-16 ***
## sector            24 0.66518   5.3796     48   1142 < 2.2e-16 ***
## location:sector  109 0.52233   2.0098    218   1142 2.566e-13 ***
## Residuals        572
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing Lower and Upper salaries ~ Location and Revenue:

```
##                    Df   Wilks approx F num Df den Df    Pr(>F)
## location           36 0.63943   4.0158     72   1154 < 2.2e-16 ***
## revenue            12 0.84359   4.2681     24   1154 4.330e-11 ***
## location:revenue 115 0.50661   2.0318    230   1154 3.165e-14 ***
## Residuals        578
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All six pairs of the variables Size, Sector, Revenue, and Location were tested using 2-way MANOVA. All six tests produced p-values of ≈ 0. From this I concluded that each variable individually has an effect on Lower and Upper salaries, as well as the interaction of each pair of variables has an effect on Lower and Upper salaries.

Testing for equality of covariance matrices

Testing equality of covariance matrices for lm(X ~ size):

```
## Chi-Sq (approx.)
##          160.483
```

```
## [1] 1.749616e-23
```

Report: Testing hypotheses H_0: the covariance matrix does not vary with the level of Size v.s. H_1: otherwise. We carried out the Box's M test and obtained 160.483 as the value of the test statistic. The corresponding p-value was ≈ 0. At a significance level of 0.05, we rejected H_0 and concluded there was strong evidence that the covariance matrix does vary with the level of Size.

Since I concluded that the covariance matrix does vary with the level of Size, I attempted to improve normality of the Lower and Upper salaries using Box Cox transformation and retested the equality of covariance matrices. The result was still a very small p-value leading me to conclude that the covariance matrix still varies with the level of Size.

```
## Chi-Sq (approx.)
##         73.51529
```

```
## [1] 9.49798e-08
```

Testing equality of covariance matrices for lm(X ~ revenue):

```
## Chi-Sq (approx.)
##         268.5238
```

```
## [1] 2.36152e-37
```

Report: Testing hypotheses H_0: the covariance matrix does not vary with the level of Revenue v.s. H_1: otherwise. We carried out the Box's M test and obtained 268.5238 as the value of the test statistic. The corresponding p-value was ≈ 0. At a significance level of 0.05, we rejected H_0 and concluded there was strong evidence that the covariance matrix does vary with the level of Revenue.

Here again I attempted to improve normality of the Lower and Upper using Box Cox transformation salaries and retested the equality of covariance matrices. It still produced a very small p-value leading me to conclude that the covariance matrix still varies with the level of Revenue.

```
## Chi-Sq (approx.)
##          152.301
```

```
## [1] 2.97383e-16
```

Multivariate linear regression

Checking the linear relationship between the Y-vector containing Lower and Upper salary, and the X-vector containing Age:

```
## Response df$Lower.Salary :
##
## Call:
## lm(formula = `df$Lower.Salary` ~ df$Age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.768 -22.748  -5.328  16.284 127.139
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 74.67249    1.51681  49.230   <2e-16 ***
## df$Age       0.00173    0.02113   0.082    0.935
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 30.97 on 740 degrees of freedom
## Multiple R-squared:  9.062e-06,  Adjusted R-squared:  -0.001342
## F-statistic: 0.006706 on 1 and 740 DF,  p-value: 0.9348
##
##
## Response df$Upper.Salary :
##
## Call:
## lm(formula = `df$Upper.Salary` ~ df$Age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -112.43  -32.10   -3.64   26.21  176.00
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 126.83571    2.21066  57.375   <2e-16 ***
## df$Age        0.02901    0.03079   0.942    0.347
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45.13 on 740 degrees of freedom
## Multiple R-squared:  0.001198,   Adjusted R-squared:  -0.0001521
## F-statistic: 0.8873 on 1 and 740 DF,  p-value: 0.3465
```

Report: Testing hypotheses $H_0 : \beta = 0$ v.s. $H_1 : \beta \neq 0$. Through multivariate linear regression we obtained a p-value of 0.93 for the model Lower salary ~ Age, and a p-value of 0.35 for the model Upper salary ~ Age. At a significance level of 0.05, we failed to reject H_0 and concluded there was not enough evidence that $\beta \neq 0$.

Therefore, I concluded that Age is insignificant and should not be in the model.

Testing for nested models

Since Age was deemed an unnecessary fit for the model, I aimed to find which of the variables Revenue, Sector, Size and Location were the best.

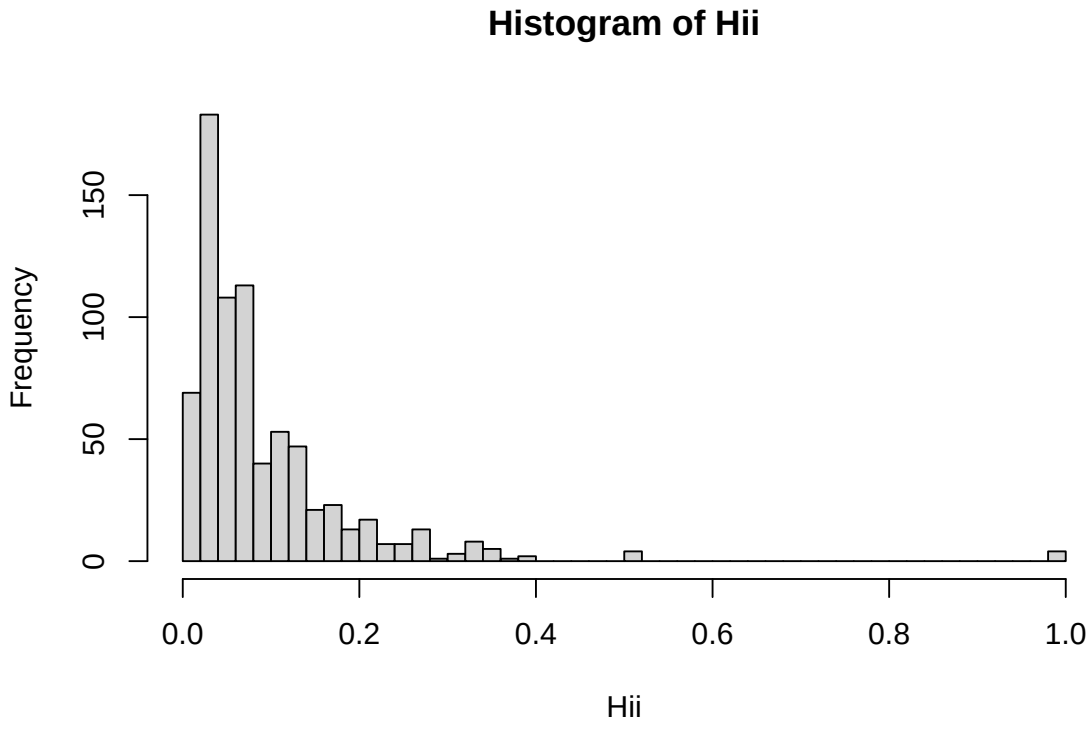
I used Excel to track and compare the AIC and BIC criteria as I tested the different models. I calculated the information criteria for the empty model first and obtained the values 13370 and 13393 respectively. I calculated the criteria for the model with just Age to verify my conclusion that Age did not belong in the model, and found that it barely lowered the criteria values. This confirmed my conclusion. The two potential single variables that lowered the criteria the most were Sector and Size. Those two variables

together in a model lowered the criteria even further. For three-variable models, Sector, Location and Size together lowered the criteria further. Since the model with all four variables didn't significantly lower the criteria, I settled on the model containing the three variables Sector, Location and Size.

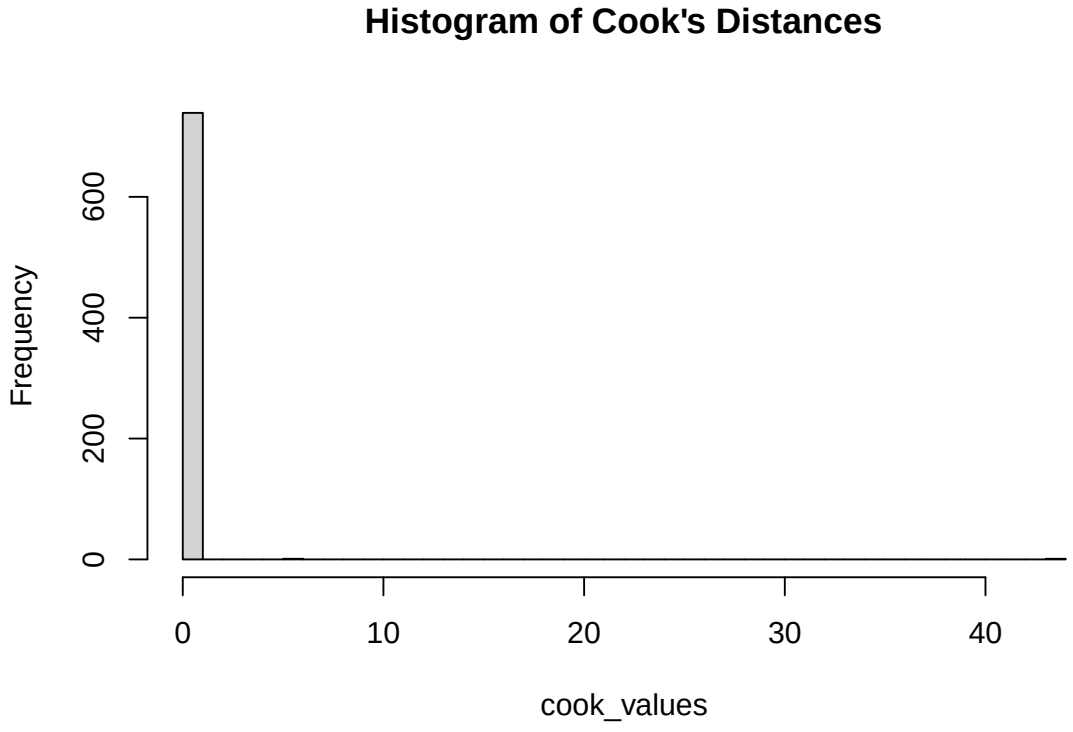
Testing the quality of the model fit

Since settling on the model Lower salary, Upper salary ~ Sector + Location + Size, I checked the quality of the model fit using three methods.

Influence of observations To see the influence the observations had on the predicted values, I first calculated and made a histogram of the leverage of the observations. I found that there were some potential outliers past 0.4.



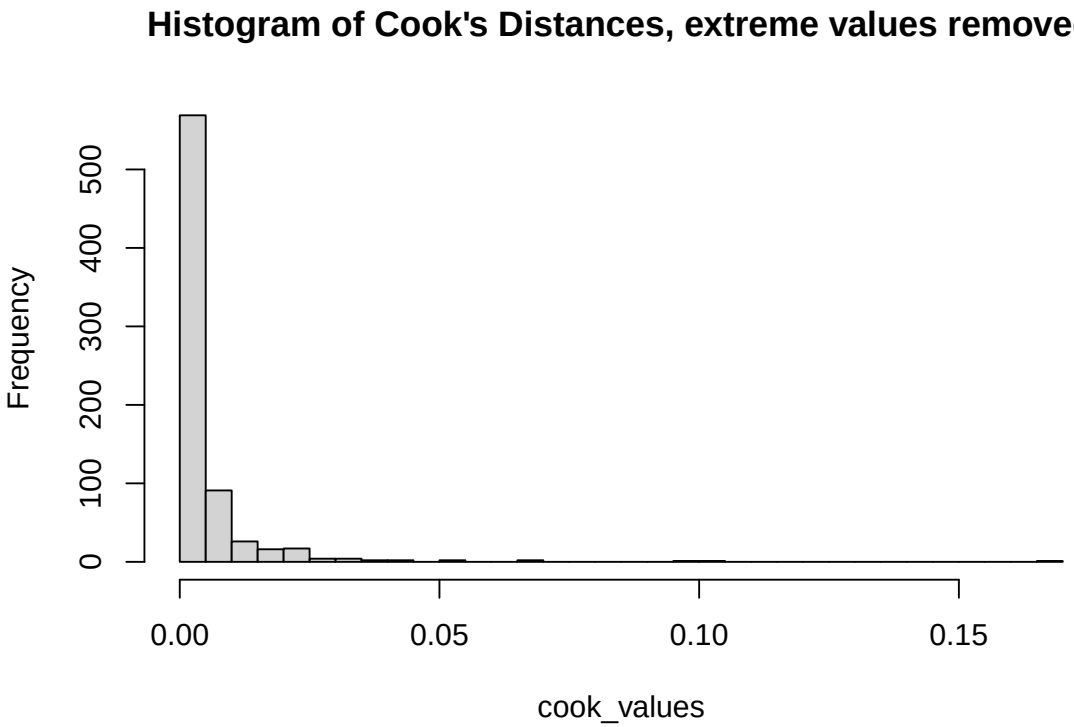
Cook's Distance I also calculated the Cook's distance to pinpoint which observations were the potential outliers. I found that at least one of the Cook's distances was infinite, so after removing 4 of the most extreme observations and making a new histogram, I got a better picture of them. Investigating the job listings of the 4 removed, I was not able to determine the cause of their extreme Cook's distances.




```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000000 0.0006518 0.0018635       Inf 0.0045879       Inf

##     99.5%
## 0.2240633

## 116 245 291 301
## 116 245 291 301
```


Histogram of Cook's Distances, extreme values removed
Frequency
0
100
200
300
400
500
0.00
0.05
0.10
0.15
cook_values

Normality of residuals Since one of the assumptions to this point has been that the residuals are normal, I investigated this. qqPlots showed that the residuals for Upper salary seemed to be normal - there were a few points outside the confidence bands in the positive tail but they were minimal. For the Lower salary however, there were a few more outside the confidence bands but still an acceptable proportion.

Principle Component Analysis

Since the selected explanatory variables Sector, Location and Size have different units of measurement, PCA needed to be performed on the sample correlation matrix rather than the sample covariance matrix. I found the cumulative contributions from the PCs to be 0.3707, 0.6947, 1.0000. Using 95% contribution as the threshold for the stopping rule, I found that no dimension reduction could occur and that all three PCs should remain in the model.

```
##             [,1]      [,2]       [,3]
## [1,] -0.6529170 0.1004106  0.7507443
## [2,] -0.5783095 0.5740025 -0.5797234
## [3,]  0.4891395 0.8126739  0.3167076

## [1] 0.3706717 0.6947424 1.0000000
```

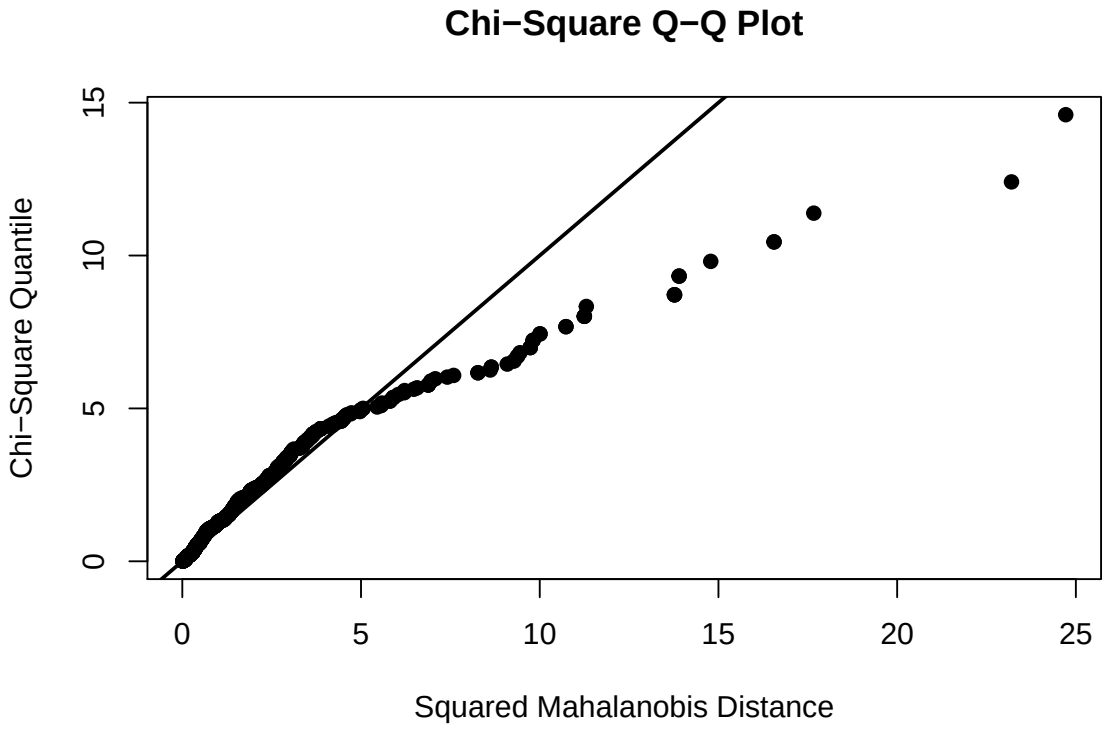
Testing salary data for MVN distribution

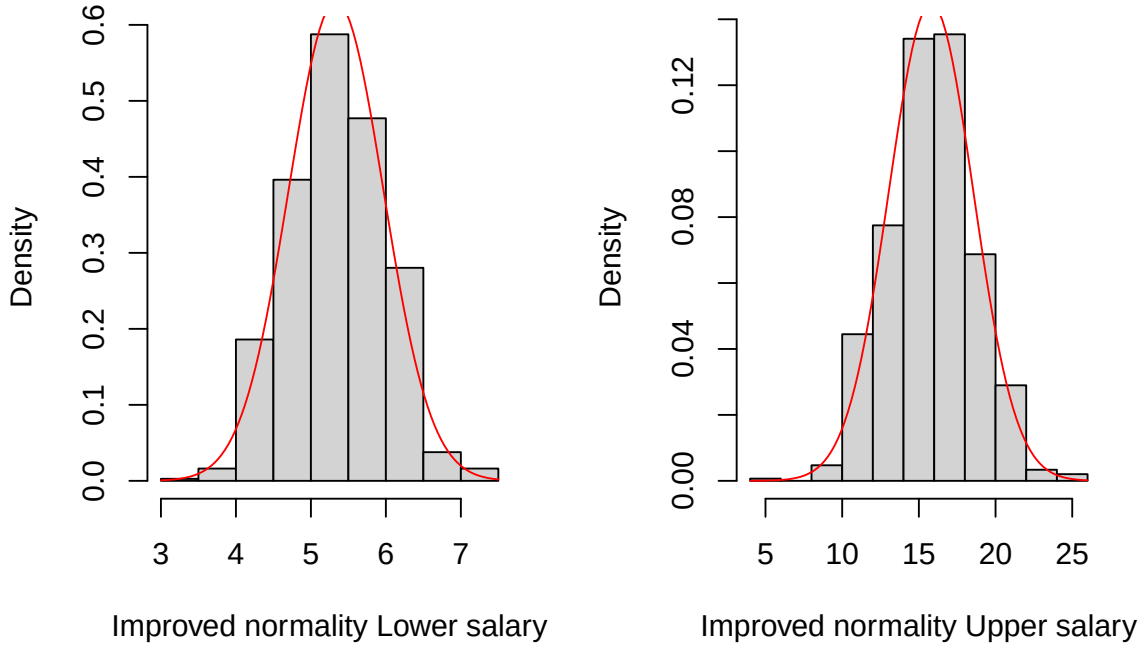
As several methods assumed that the data was MVN, I performed a test to find out.

```
##              Test        Statistic                p value Result
## 1 Mardia Skewness 481.274776481863 7.51059390241336e-103     NO
## 2 Mardia Kurtosis 36.0644736821848                      0     NO
## 3             MVN             <NA>                   <NA>     NO
```

```
##           Test     Variable Statistic   p value Normality
## 1 Shapiro-Wilk lower salary    0.9331  <0.001      NO
## 2 Shapiro-Wilk upper salary    0.9754  <0.001      NO
```

Since this test revealed that the Lower and Upper salaries were not MVN, I retested the Lower and Upper salaries upon which I had improved the normality using Box Cox transformation. From the test I found that the improved-normality variables still could not be considered MVN. The univariate distribution of Upper salary could be considered normal, but the distribution of Lower salary did not pass for normal.






```
##              Test        Statistic             p value Result
## 1 Mardia Skewness 17.0431119339882 0.00189602120785399     NO
## 2 Mardia Kurtosis  12.217019621009                   0     NO
## 3             MVN             <NA>                <NA>     NO
```

```
##           Test                          Variable Statistic   p value Normality
## 1 Shapiro-Wilk Improved normality Lower salary    0.9954    0.0272    NO
## 2 Shapiro-Wilk Improved normality Upper salary    0.9972    0.2368    YES
```

Discussion/Conclusion

Summary of results My aim was to answer the questions: Does the Location, Size, Sector, Revenue or Age of the company predict the Lower, Upper and Average salary of the job being advertised? Which of those variables are the best predictors of salary?

In testing if the sample means of the Lower, Upper and Average salaries were representative of the true means as $75,000, $130,000, and $100,000, I found they were not. I believe this was due to the huge variation in the salaries. To perform the remaining methods, I had to eliminate Average salaries from my analysis as it had a linear relationship with Lower and Upper salaries and certain R functions could not handle it.

Through 1-way and 2-way MANOVA tests, I determined that each variable individually as well as each pair from Size, Sector, Revenue, and Location all had effects on Lower and Upper salaries. In Multivariate linear regression testing of the explanatory variable Age, I found insufficient evidence that it's β_i coefficient was something other than zero. I confirmed through the testing of nested models that Age as well as Revenue indeed did not belong in the model, but that Sector, Location and Size did. To test if this model selection was a good fit, I checked the Cook's distance of the residuals as well as the normality of the residuals. qqPlots showed that the residuals for Upper Salary seemed to be normal but the residuals for Lower salary were questionable. The results of performing Principle Component Analysis on this model with two response variables and three explanatory variables was as I expected - there was no dimension reduction that could occur.

Upon testing the assumption of equal covariance matrices for the m samples for 1-way MANOVA, I found that the covariance matrices varied with the levels of the factors. As such, the results of the 1-way MANOVA tests may be unreliable. I attempted to improve the normality of Lower and Upper salaries using the Box Cox method. Using this transformed data I ran tests for MVN distribution, but only the univariate distribution of Upper salary could be considered normal, not Lower salary.

Limitations and concerns Given that some of these assumptions were not met, further methods could be attempted at improving the normality of the Lower and Upper salary variables. Additionally, only ~4% of the salaries in the data set were employer provided, the rest were estimated by Glassdoor. The salary range estimated by Glassdoor for any given job posting was very wide - much wider than the employer provided ranges. This lead me to question the accurateness of Glassdoor's estimates. Future studies could find datasets with a larger proportion of employer provided salary ranges. Further, a more robust investigation into determining the best explanatory variables for salary could include the 16 different skill sets that employers may look for. With all these additional variables, the dimension reduction methods of Principle Component Analysis or Canonical Correlation Analysis may actually reveal if dimension reduction could occur. In this case Principle Component Regression would be an applicable method. Factor Analysis may also be more appropriate on this considerably larger dataset.

Appendix

Following is all the code for this report.

```
library(tidyverse)
library("MVN")


# Exploratory Data Analysis
data = read.csv("data_cleaned_2021.csv")

# remove unwanted columns
df = data %>%
  select(-Job.Description, -Salary.Estimate, -Competitors, -Hourly,
         -Employer.provided, -company_txt)

# create a data frame of just the salary variables
salary_df = as.matrix(df[,c("Lower.Salary", "Upper.Salary", "Avg.Salary.K.")])

# summary statistics of the salary variables
summary(salary_df)


# Correlation
data.sub = as.matrix(df[,c("Lower.Salary", "Upper.Salary", "Avg.Salary.K.", "Age")])
cor(data.sub)


### Comparing Lower, Upper and Average salaries

# setup to create an overlapped density plot of the 3 salary variables
low = data.frame(salary = df$Lower.Salary)
up = data.frame(salary = df$Upper.Salary)
avg = data.frame(salary = df$Avg.Salary.K.)

low$level = "Lower Salary"
up$level = "Upper Salary"
avg$level = "Average Salary"

all.salaries = rbind(low, up, avg)

ggplot(all.salaries, aes(salary, fill = level)) +
  geom_density(alpha = 0.3)
```

```r
# Histograms of Lower, Upper, Average salaries
par(mfrow=c(1,3))
hist(df$Lower.Salary, xlab = "Lower salary", main = "")
hist(df$Upper.Salary, xlab = "Upper salary", main = "")
hist(df$Avg.Salary.K., xlab = "Average salary", main = "")


### Salaries by Company Size

# Average salary grouped by company size
size = as_factor(df$Size)
df = cbind(df, size)

# arrange the plots in decreasing order by the median Average salary
ggplot(df, aes(x=reorder(size, -Avg.Salary.K., median), Avg.Salary.K.)) +
  geom_boxplot() +
  xlab("Company Size") +
  ylab("Average Salary")


### Salaries by Sector

# Change the Sector variable to a factor
df$Sector = as.factor(df$Sector)

# Remove rows where Sector = -1
df.Sector = df[!df$Sector == "-1",]

# Average salary grouped by sector
ggplot(df.Sector, aes(x=reorder(Sector, -Avg.Salary.K., median), Avg.Salary.K.)) +
  geom_boxplot(color="dark green", fill="blue", alpha=0.2) +
  xlab("Sector") +
  ylab("Average Salary") +
  theme(axis.text.x  = element_text(angle=90, vjust=0.5))


### Salaries by Location

# Average salary grouped by location
ggplot(df, aes(x=reorder(Job.Location, -Avg.Salary.K., median), Avg.Salary.K.)) +
  geom_boxplot(color="dark blue", fill="green", alpha=0.2) +
```

```
  xlab("Location") +
  ylab("Average Salary") +
  theme(axis.text.x  = element_text(angle=90, vjust=0.5))


# Sectors with the most jobs postings
ggplot(df.Sector, aes(x=reorder(Sector, Sector, function(x)-length(x)))) +
  geom_bar(stat="count", width=0.7, fill="steelblue") +
  theme(axis.text.x  = element_text(angle=90, vjust=0.5)) +
  xlab("Sector") +
  ylab("No. Current Job Postings")


### Salaries by Company Age

# Remove rows where Age = -1
df.Age = df[!df$Age == "-1",]

# Add extra level "Other" to the Sector factor
levels(df.Age$Sector) = c(levels(df.Age$Sector), "Other")

# Change Sector to "Other" for all those that are not a sector with
# the number of job postings in the top 5
for(i in 1:length(df.Age$Sector)) {
  if(!(df.Age$Sector[i] == "Information Technology" ||
       df.Age$Sector[i] == "Biotech & Pharmaceuticals" ||
       df.Age$Sector[i] == "Business Services" ||
       df.Age$Sector[i] == "Insurance" ||
       df.Age$Sector[i] == "Finance")) {

      df.Age$Sector[i] = "Other"
  }
}

# Create a df excluding all Sectors that are Other.
# Approx. 29% of the job postings are being excluded
df.Age_Sec = df.Age[!df.Age$Sector == "Other",]

# Scatterplot showing the salary range for the top 5 company sectors
# with the most job postings.
# Assuming a MVN distribution
ggplot(df.Age_Sec, aes(x=Age, y=Avg.Salary.K., color=Sector)) +
```

```
  geom_point() +
  theme(legend.position="right") +
  stat_ellipse(type = "norm")


# testing on mu for lower, upper and average salary
mu_hat = colMeans(salary_df)

mu_0 = c(75, 130, 100)
n = nrow(salary_df)
p = ncol(salary_df)

# sigma unknown
test.stat1 = drop(n*t(mu_hat - mu_0) %*% solve(cov(salary_df),
                          tol = 5e-19) %*% (mu_hat - mu_0))
paste("The test statistic is", test.stat1)

cri.point = (n-1)*p * qf(0.95, p, n-p)/(n-p)
paste("The critical point is", cri.point)

p.val = 1-pf((n-p)*test.stat1/((n-1)*p), p, n-p)
paste("The p-value is", p.val)


## 1-way MANOVA using factor Size

# salary vectors - lower and upper only
X = cbind(df$Lower.Salary, df$Upper.Salary)
colnames(X) = c("lower salary", "upper salary")

# using size vector previously made
summary(manova(X ~ size), test = "Wilks")

# calculate the critical point for the rejection region
n = nrow(X)
p = ncol(X)
m = 8
cri.point = exp(qchisq(0.95, p*(m-1))/((p+m)/2-n+1))


## 1-way MANOVA using factor Sector
```

```r
# convert the sector variable to a factor
sector = as.factor(df$Sector)
summary(manova(X ~ sector), test = "Wilks")

# calculate the critical point for the rejection region
n = nrow(X)
p = ncol(X)
m = 25
cri.point = exp(qchisq(0.95, p*(m-1))/((p+m)/2-n+1))


## 1-way MANOVA using factor Revenue

# convert the revenue variable to a factor
revenue = as.factor(df$Revenue)
summary(manova(X ~ revenue), test = "Wilks")

# calculate the critical point for the rejection region
n = nrow(X)
p = ncol(X)
m = 13
cri.point = exp(qchisq(0.95, p*(m-1))/((p+m)/2-n+1))


## 1-way MANOVA using factor Location

# convert the sector variable to a factor
location = as.factor(df$Job.Location)
summary(manova(X ~ location), test = "Wilks")

# calculate the critical point for the rejection region
n = nrow(X)
p = ncol(X)
m = 37
cri.point = exp(qchisq(0.95, p*(m-1))/((p+m)/2-n+1))


## 2-way MANOVA using factors Revenue & Sector
summary(manova(X ~ revenue*sector), test = "Wilks")


## 2-way MANOVA using factors Revenue & Size
```

```r
summary(manova(X ~ revenue*size), test = "Wilks")


## 2-way MANOVA using factors Sector & Size
summary(manova(X ~ sector*size), test = "Wilks")


## 2-way MANOVA using factors Location & Size
summary(manova(X ~ location*size), test = "Wilks")


## 2-way MANOVA using factors Location & Sector
summary(manova(X ~ location*sector), test = "Wilks")


## 2-way MANOVA using factors Location & Revenue
summary(manova(X ~ location*revenue), test = "Wilks")


## Testing for equality of covariance matrices

result1 = heplots::boxM(lm(X ~ size))
result1$statistic
result1$p.value


# Attempting to improve normality through Box Cox transformation
newFitLower = lm(df$Lower.Salary ~ size)
newFitUpper = lm(df$Upper.Salary ~ size)

lambdaLower = EnvStats::boxcox(newFitLower, optimize=T,
                               lambda = c(-10,10))$lambda
df$Lower.Salary.new = (df$Lower.Salary^lambdaLower-1)/lambdaLower

lambdaUpper = EnvStats::boxcox(newFitUpper, optimize=T,
                               lambda = c(-10,10))$lambda
df$Upper.Salary.new = (df$Upper.Salary^lambdaUpper-1)/lambdaUpper

fitNew = lm(cbind(df$Lower.Salary.new, df$Upper.Salary.new) ~ size)
result1_new = heplots::boxM((fitNew))
result1_new$statistic
result1_new$p.value
```

```r
## Testing for equality of covariance matrices

result3 = heplots::boxM(lm(X ~ revenue))
result3$statistic
result3$p.value


# Attempting to improve normality through Box Cox transformation
newFitLower = lm(df$Lower.Salary ~ revenue)
newFitUpper = lm(df$Upper.Salary ~ revenue)

lambdaLower = EnvStats::boxcox(newFitLower, optimize=T,
                               lambda = c(-10,10))$lambda
df$Lower.Salary.new = (df$Lower.Salary^lambdaLower-1)/lambdaLower

lambdaUpper = EnvStats::boxcox(newFitUpper, optimize=T,
                               lambda = c(-10,10))$lambda
df$Upper.Salary.new = (df$Upper.Salary^lambdaUpper-1)/lambdaUpper

fitNew = lm(cbind(df$Lower.Salary.new, df$Upper.Salary.new) ~ revenue)
result1_new = heplots::boxM((fitNew))
result1_new$statistic
result1_new$p.value


## Multivariate linear regression using variable Age

fit.a = lm(cbind(df$Lower.Salary, df$Upper.Salary) ~ df$Age)
summary(fit.a)


## Testing for nested models
logLik.mlm <- function(object, ...) {
  resids <- residuals(object)
  Sigma_ML <- crossprod(resids)/nrow(resids)
  ans <- sum(mvtnorm::dmvnorm(resids, sigma = Sigma_ML,log = TRUE))
  df <- prod(dim(coef(object))) + choose(ncol(Sigma_ML) + 1, 2)
  attr(ans, "df") <- df
  class(ans) <- "logLik"
  return(ans)
}
```

```r
# test the empty model
fit.empty = lm(X ~ 1)
logLik(fit.empty)
AIC(fit.empty)
BIC(fit.empty)

# single-variable models,
# confirming that Age is not a good fit for the model
AIC(lm(X ~ df$Age))
BIC(lm(X ~ df$Age))

AIC(lm(X ~ sector))
BIC(lm(X ~ sector))

AIC(lm(X ~ location))
BIC(lm(X ~ location))

AIC(lm(X ~ size))
BIC(lm(X ~ size))

AIC(lm(X ~ revenue))
BIC(lm(X ~ revenue))

# 2-variable models
AIC(lm(X ~ revenue + sector))
BIC(lm(X ~ revenue + sector))

AIC(lm(X ~ revenue + location))
BIC(lm(X ~ revenue + location))

AIC(lm(X ~ revenue + size))
BIC(lm(X ~ revenue + size))

AIC(lm(X ~ sector + location))
BIC(lm(X ~ sector + location))

AIC(lm(X ~ sector + size))
BIC(lm(X ~ sector + size))

AIC(lm(X ~ location + size))
BIC(lm(X ~ location + size))
```

```r
# 3-variable models
AIC(lm(X ~ revenue + sector + location))
BIC(lm(X ~ revenue + sector + location))

AIC(lm(X ~ revenue + sector + size))
BIC(lm(X ~ revenue + sector + size))

AIC(lm(X ~ revenue + location + size))
BIC(lm(X ~ revenue + location + size))

AIC(lm(X ~ sector + location + size))
BIC(lm(X ~ sector + location + size))

# all variables model
AIC(lm(X ~ revenue + sector + location + size))
BIC(lm(X ~ revenue + sector + location + size))


# Influence of observations
fit = lm(X ~ sector + location + size)

# Leverage
resids = residuals(fit)
n = length(resids)
Xc = model.matrix(fit)
H = Xc %*% solve(crossprod(Xc)) %*% t(Xc)
Hii = diag(H)
hist(Hii, 50)


# Cook's Distance
SigmaHatLS = crossprod(resids)/(n - ncol(Xc))
cook_values = Hii/((1 - Hii)^2 * ncol(Xc)) * diag(resids %*%
                            solve(SigmaHatLS) %*% t(resids))
hist(cook_values, 50, main = "Histogram of Cook's Distances")
(summary(cook_values))

# there is at least one Cook's Distance that is infinite.
# Find the 99.5th quantile and remove the values that are greater than it
quantile(cook_values, 0.995)
which(cook_values > 0.22)
```

```r
# remove the four observations above 0.05
cook_values = cook_values[-c(116, 245, 291, 301)]

# make a new histogram with the removed observations
hist(cook_values, 50, main = "Histogram of Cook's Distances, extreme values removed")

#investigate the rows with the extreme Cook's Distances
removed = data[c(116, 245, 291, 301),-4]


# Normality of residuals
name = c("Lower salary", "Upper salary")
op <- par(mfrow = c(1,2), oma = c(5,4,0,0), mar = c(1,1,2,2))

for (i in 1:ncol(resids)){
  car::qqPlot(resids[,i], main = name[i], id = F)
}

title(xlab = "Normal quantiles",
  ylab = "Sample quantiles",
  outer = TRUE, line = 3)
  par(op)


# PCA based on sample correlation matrix
X1 = as.matrix(cbind(sector, location, size))

pca1 = eigen(cor(X1), symmetric = T)
# loadings
pca1$vectors
(cumsum(pca1$values)/sum(pca1$values))
# PC scores
Z1 = scale(X1, center = T, scale = F) %*% pca1$vectors


# testing salary data for MVN distribution
Y = cbind(df$Lower.Salary, df$Upper.Salary)
colnames(Y) = c("lower salary", "upper salary")

# testing the original lower and upper salaries
result = mvn(Y, mvnTest = "mardia", univariateTest = "SW")
result$multivariateNormality
```

```
result$univariateNormality


# new Y variable with the Box Cox transformed Lower and Upper salary variables
Y_new = as.matrix(cbind(df$Lower.Salary.new, df$Upper.Salary.new))
colnames(Y_new) = c("Improved normality Lower salary",
                    "Improved normality Upper salary")

# testing the Box Cox transformed lower and upper salaries
result = mvn(Y_new, mvnTest = "mardia", univariateTest = "SW",
             univariatePlot = "histogram", multivariatePlot = "qq")
result$multivariateNormality
result$univariateNormality
```

Bibliography

Bhathi, Nikhil. "Data scientist salary." *Kaggle*, https://www.kaggle.com/datasets/nikhilbhathi/data-scientist-salary-us-glassdoor?select=data_cleaned_2021.csv.

Lai, Junting. "Predict Data Science Salaries with Data Science." *Towards Data Science*, 12 Oct 2020, https://towardsdatascience.com/the-us-data-science-job-market-in-2020-463520a9d5a.