Assignment 1

Geethu Girish, 20985805

Data:

This dataset is the result of a survey conducted by Hacker News to find the user salaries in various tech industries across the world in the year 2016. The data frame has 1655 objects of 19 variables which includes salary id, employer name, location details such as location name, state, country, latitude, and longitude, job details like title, category and rank, total years of experience, employer experience years, annual base pay, signing bonus, annual bonus, stock value bonus, comments and submission details. In this report, the average salaries of employees in "*Software*" category and all other categories collectively named "*Other*" for beginners (0 to 5 years experience), intermediate (5 to 10 years experience) and seniors (more than 10 hears experience) in United States of America is analysed.

To carry out the analysis, few variables from the dataset were chosen including *total_experience_years:* total years of work experience (numeric), *annual_base_pay:* bas pay of the employee for the year (US Dollars, numeric) and *job_title_category*: category of job(character) for location_country "US". Furthur looking into the dataset, there were missing values and "NA" which was removed from the subset of data. I checked for outliers using Interquartile range(IQR) method and omitted those *annual_base_pay* less than $1000 which is logically incorrect.

Planning and Analysis:

Null Hypothesis: Annual base pay of employees in the Software category is always greater that Other category for all job ranking.

For my analysis, I have added a new variable "*rank*" wrt the variable *total_experience_years* where employees with 0 to 5 years of experience is ranked "*beginner*", employees with 5 to 10 years of experience is ranked as "*intermediate*" and employees with more than 10 years of experience is ranked as "*senior".* I also created another variable "*category*" for the 2 categories of job ie; "*Software*" and merged all the other *job_title_category* into "*Other*".

The figure 1 illustrates the bar graph of annual base pay for employees in the two categories based on their ranks. From figure 1, it is clear that the salaries of employees in Software field is always greater than the other employees.

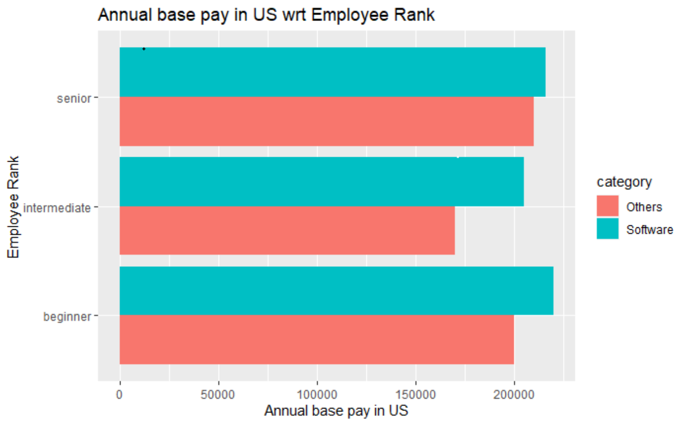
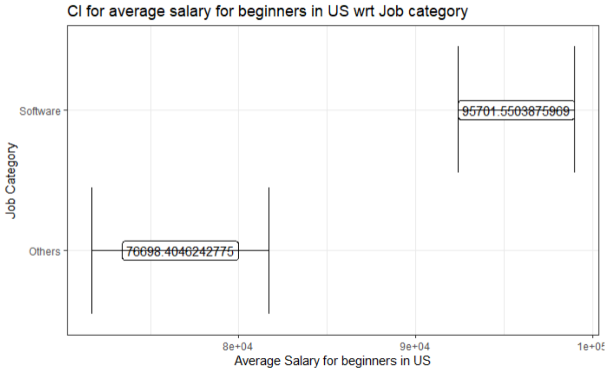
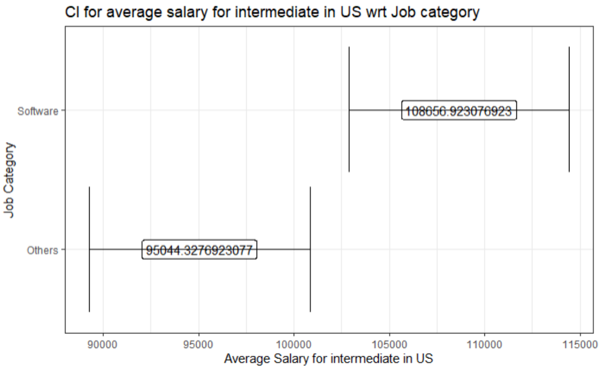
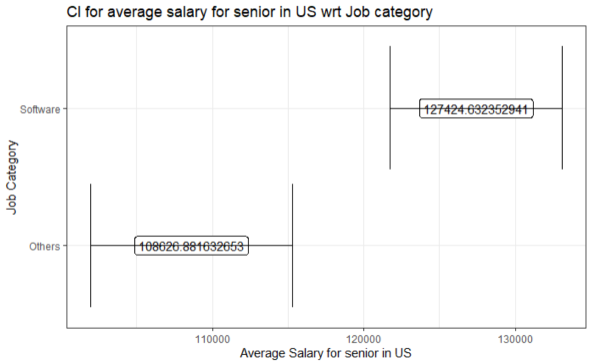


Figure 1

In this analysis we have 6 data subsets for each combination of category and rank. The normality of each data subset was checked using *Shapiro-Wilk test* and found that except the subset of intermediate-software employees (p = 0.2039) and senior-other category employees (p = 0.07072), all other data subsets were significantly non-normal at 5% level of significance *(p < 0.05)*. To further confirm the normality of each data subset, I also plotted *QQ plots and histograms*. However, all the data subsets in this analysis have more than 30 data points, hence by Central Limit Theorem the mean is normally distributed. The confidence intervals were calculated for the the subsets and it clearly shows the datasets are independent of each other.





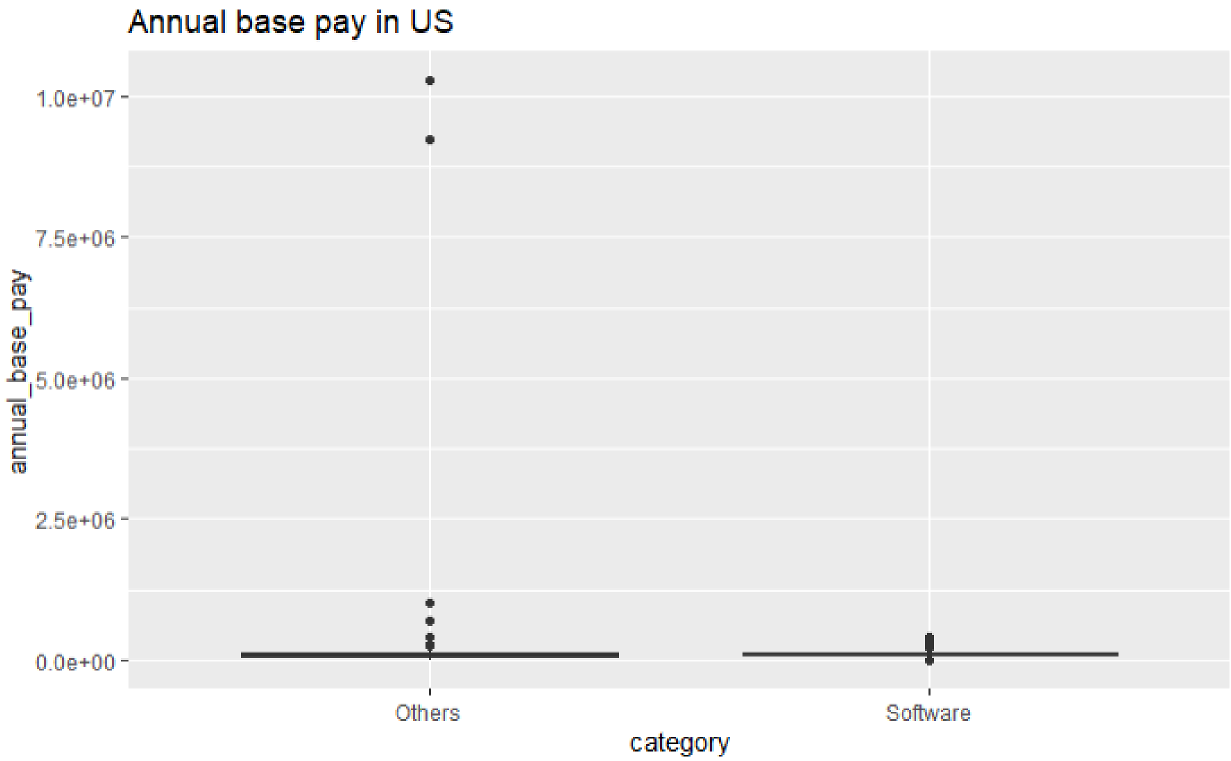


Conclusion:

Based on the results from the previous section, we can say that our null hypothesis is not rejected ie; the annual base pay of employees in the Software category is always greater that Other category for all job ranking.

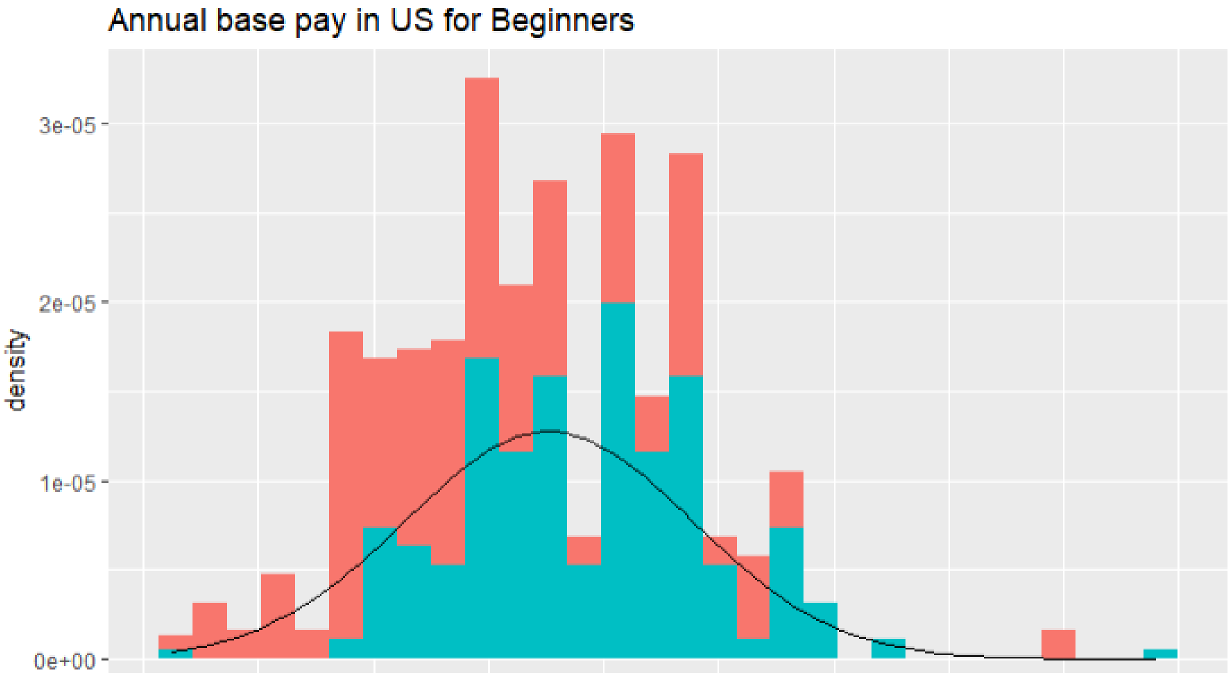
Appendix 1:

Figure below shows the outliers in the initial dataset.

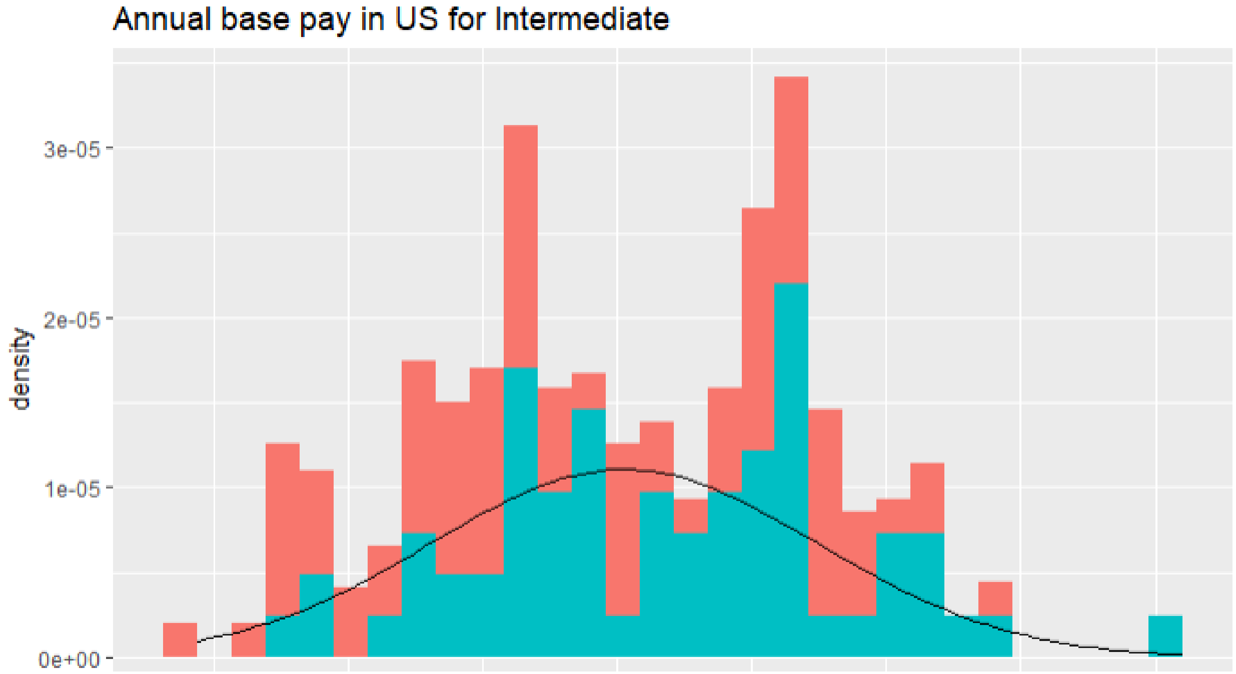


Appendix 2:

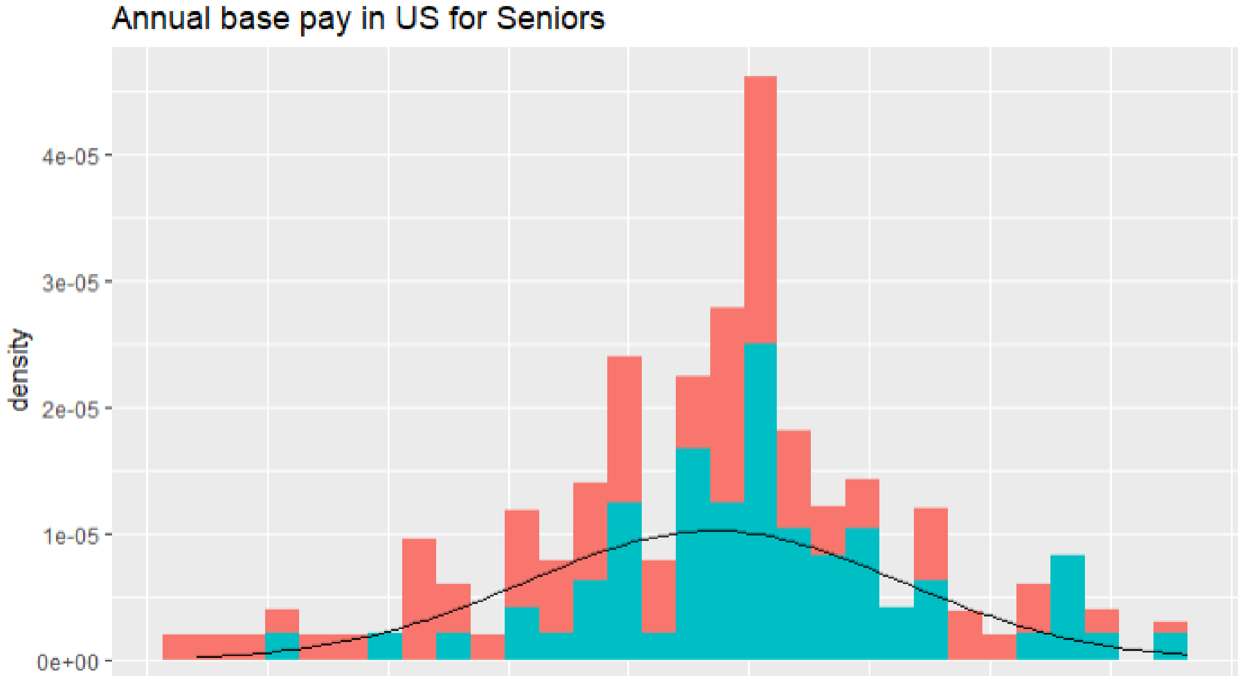
Histogram for beginners



Histogram for intermediate

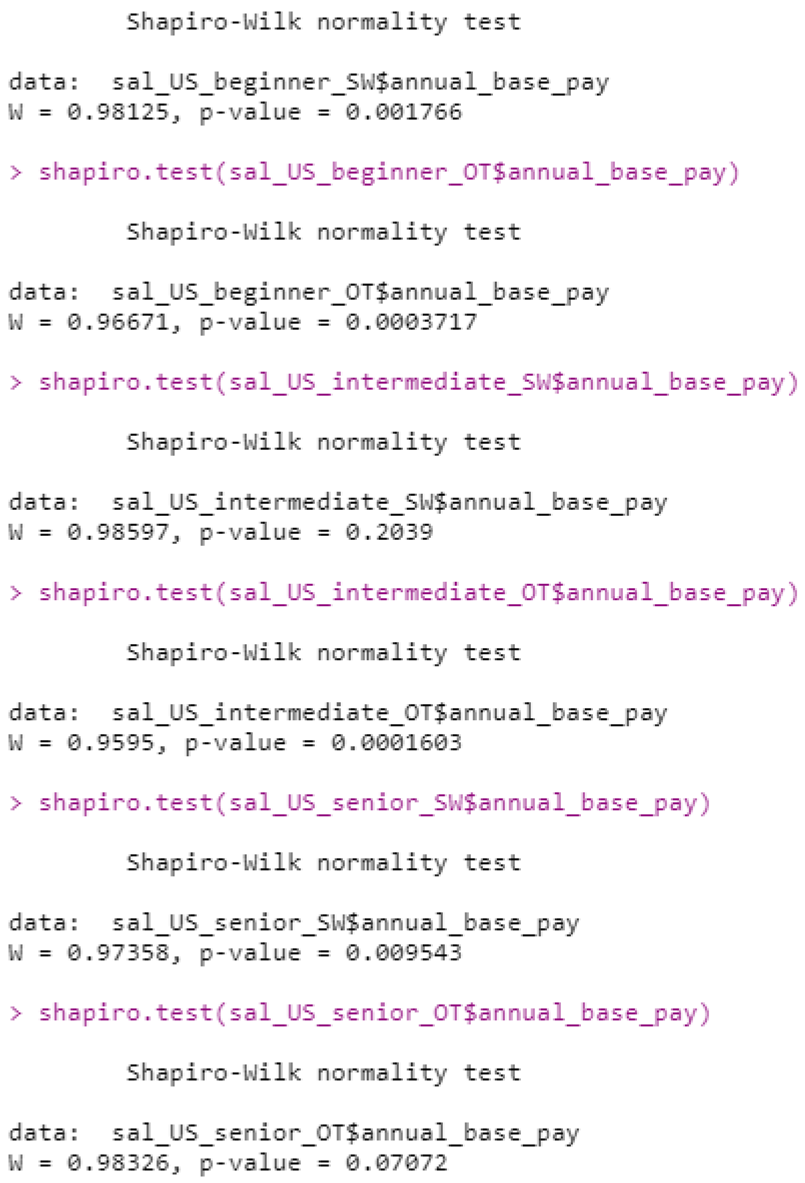


Histogram for seniors



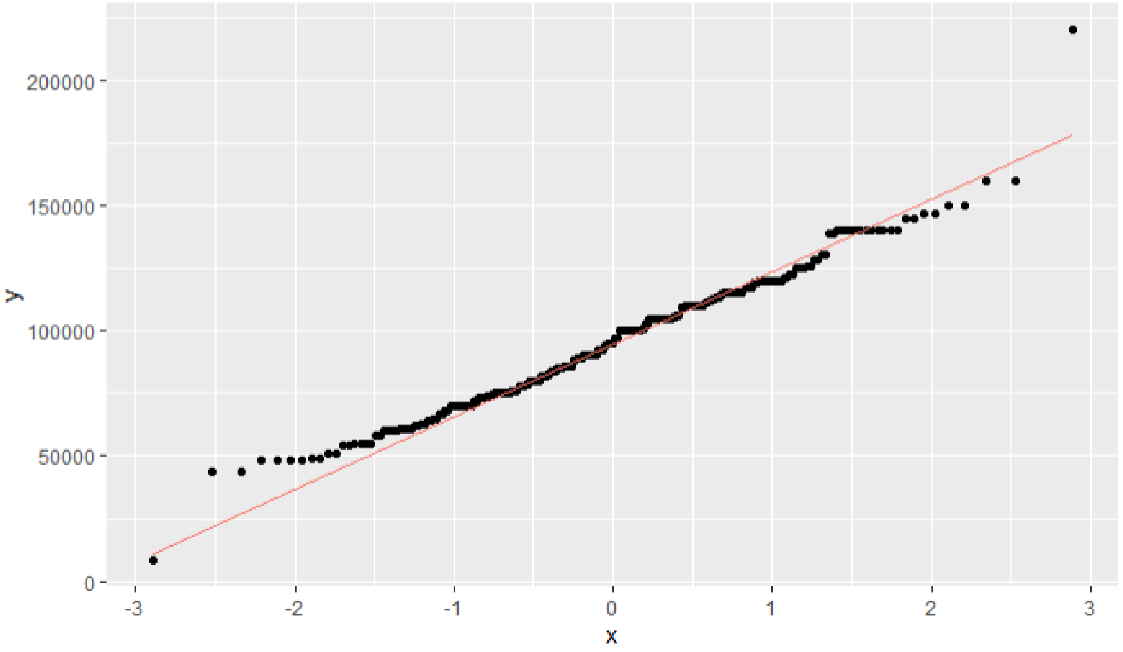
Appendix 3:

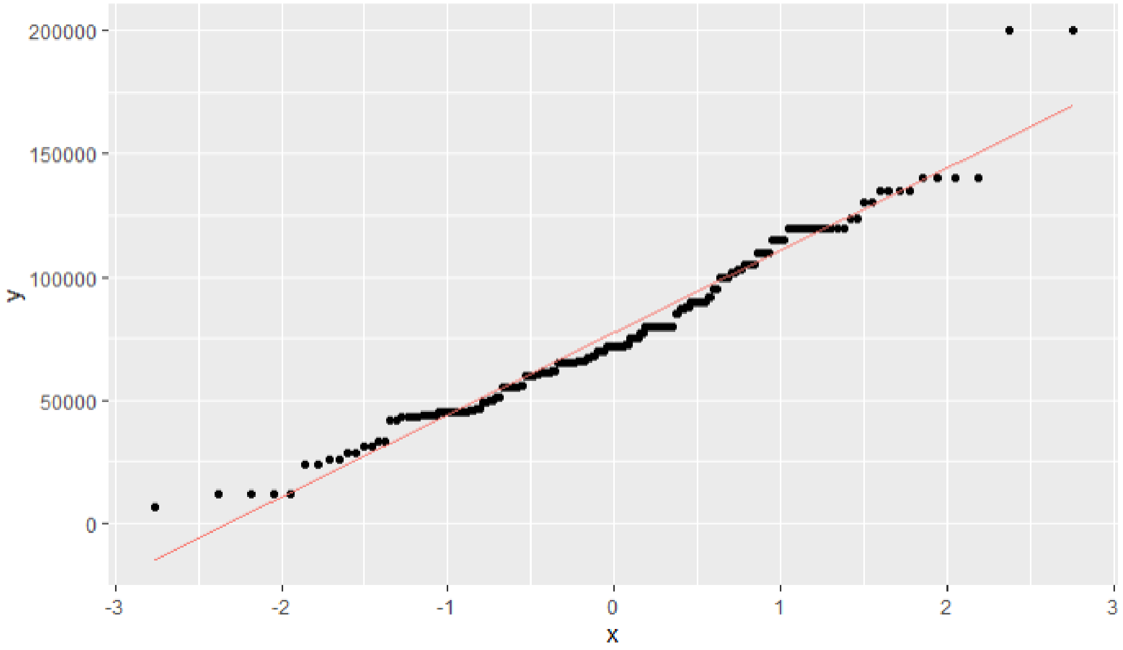
The result of Shapiro-Wilk test for normality

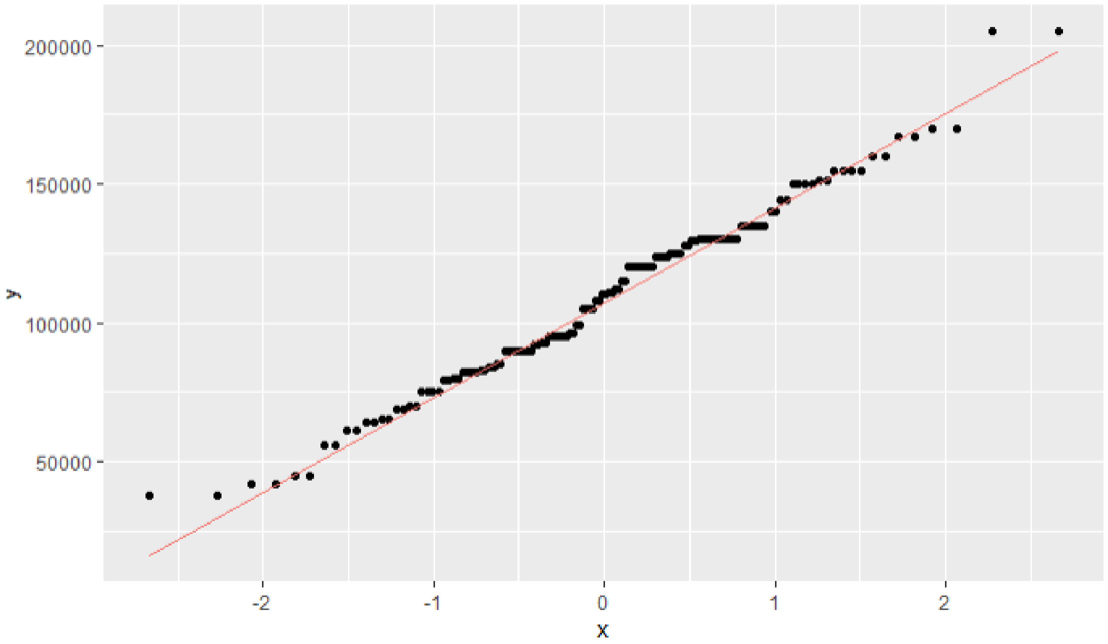


```
        Shapiro-Wilk normality test

data:  sal_US_beginner_SW$annual_base_pay
W = 0.98125, p-value = 0.001766

> shapiro.test(sal_US_beginner_OT$annual_base_pay)

        Shapiro-Wilk normality test

data:  sal_US_beginner_OT$annual_base_pay
W = 0.96671, p-value = 0.0003717

> shapiro.test(sal_US_intermediate_SW$annual_base_pay)

        Shapiro-Wilk normality test

data:  sal_US_intermediate_SW$annual_base_pay
W = 0.98597, p-value = 0.2039

> shapiro.test(sal_US_intermediate_OT$annual_base_pay)

        Shapiro-Wilk normality test

data:  sal_US_intermediate_OT$annual_base_pay
W = 0.9595, p-value = 0.0001603

> shapiro.test(sal_US_senior_SW$annual_base_pay)

        Shapiro-Wilk normality test

data:  sal_US_senior_SW$annual_base_pay
W = 0.97358, p-value = 0.009543

> shapiro.test(sal_US_senior_OT$annual_base_pay)

        Shapiro-Wilk normality test

data:  sal_US_senior_OT$annual_base_pay
W = 0.98326, p-value = 0.07072
```

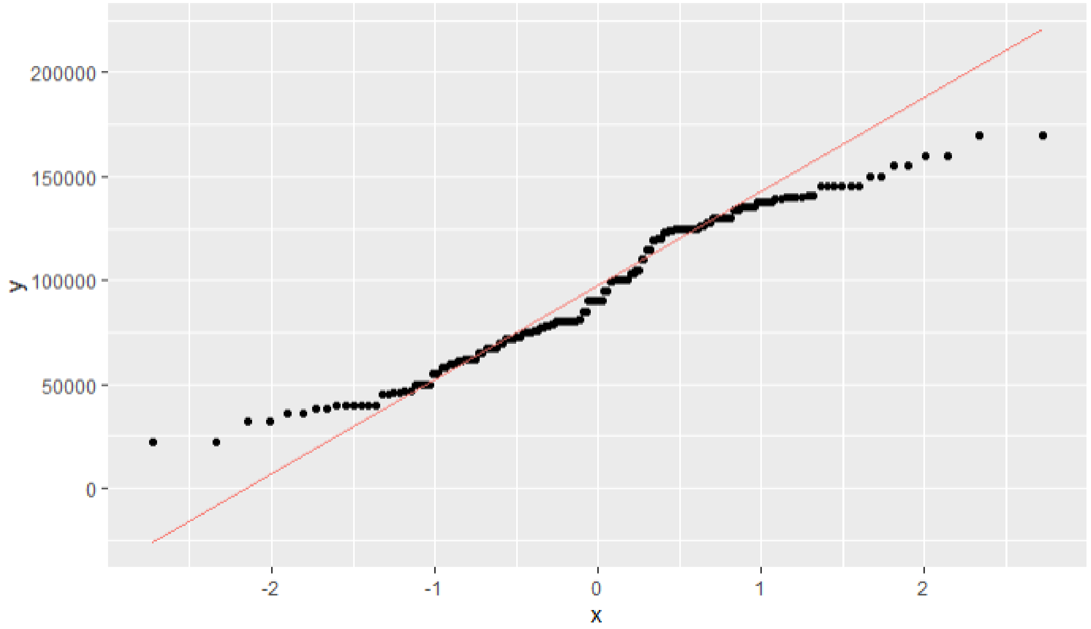
Q-Q plots for 6 data subsets


Annual base pay in US for Beginners in Software category

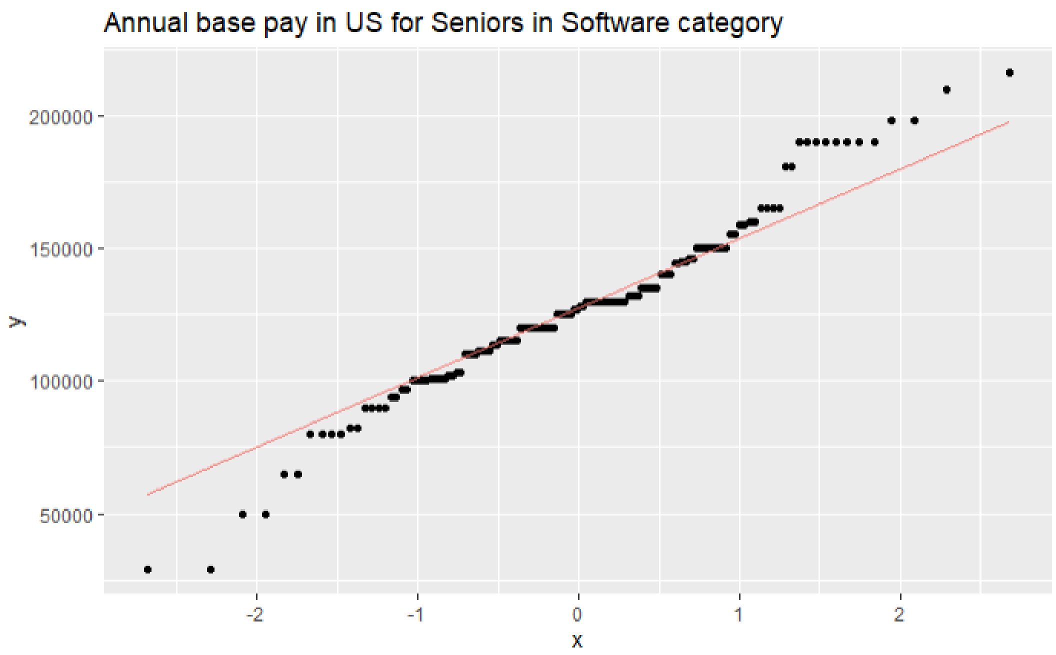

Annual base pay in US for Beginners in Other category

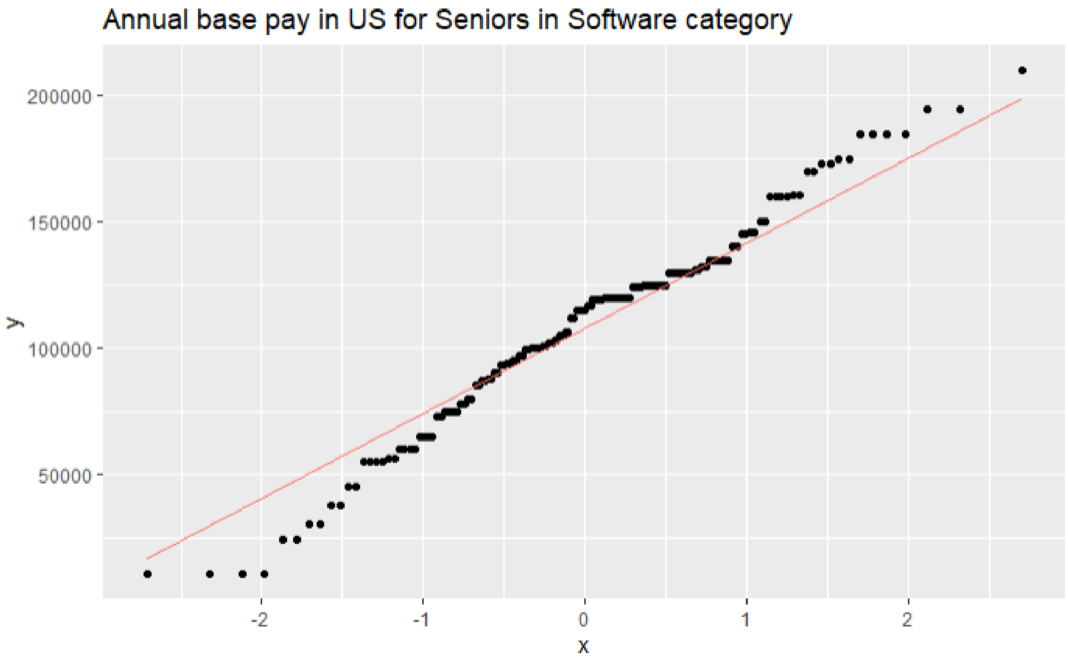


Annual base pay in US for Intermediates in Software category



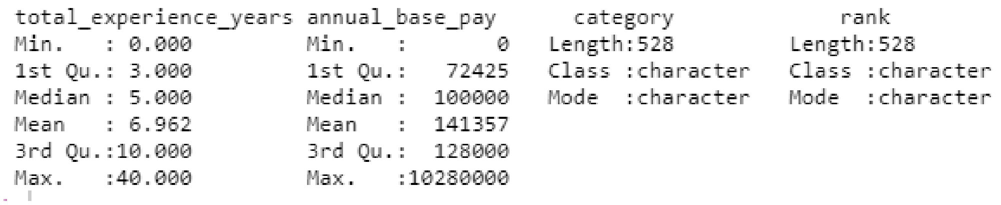
Annual base pay in US for Intermediate in Other category





Appendix 4:

Summary of the data



```
 total_experience_years annual_base_pay      category             rank          
 Min.   : 0.000         Min.   :       0   Length:528         Length:528        
 1st Qu.: 3.000         1st Qu.:   72425   Class :character   Class :character  
 Median : 5.000         Median :  100000   Mode  :character   Mode  :character  
 Mean   : 6.962         Mean   :  141357                                        
 3rd Qu.:10.000         3rd Qu.:  128000                                        
 Max.   :40.000         Max.   :10280000                                        
```