Annual Salary Report

Hitesh Kumar Pounraj

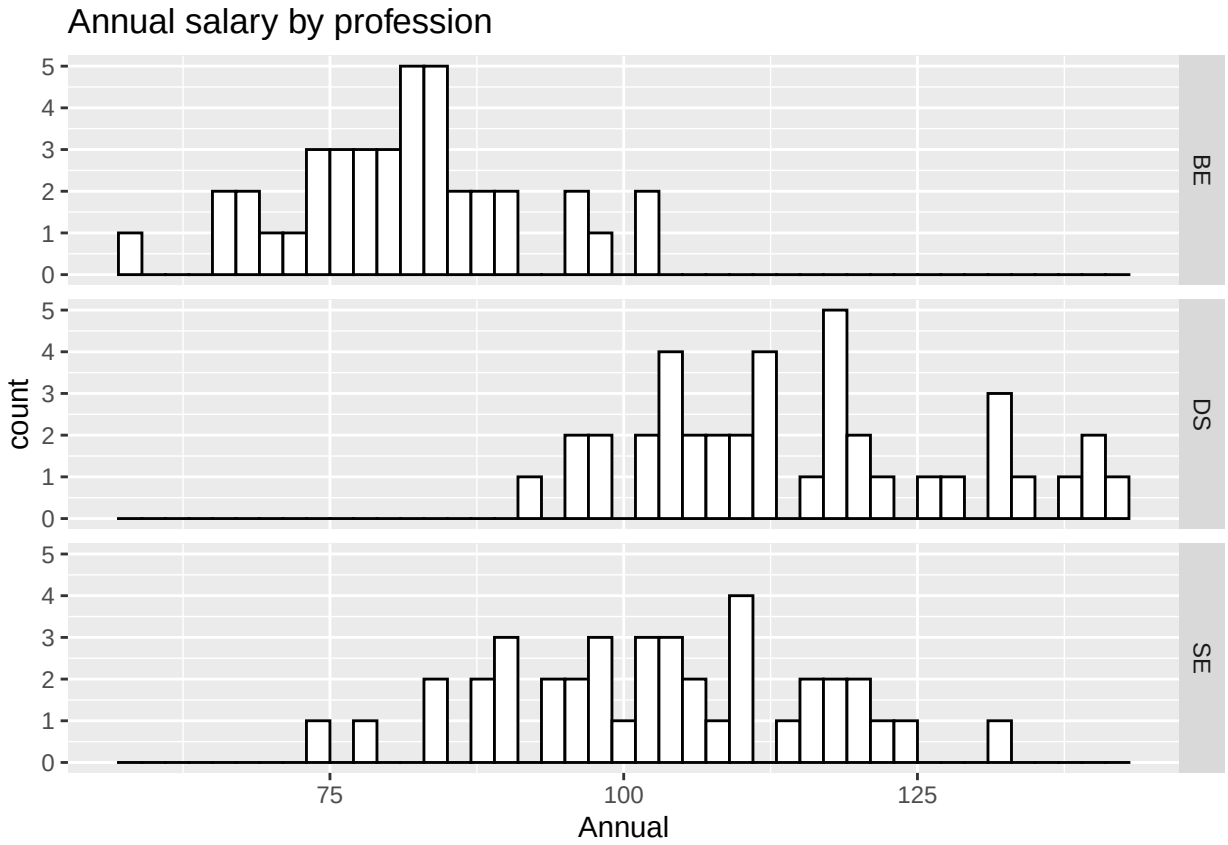
Introduction

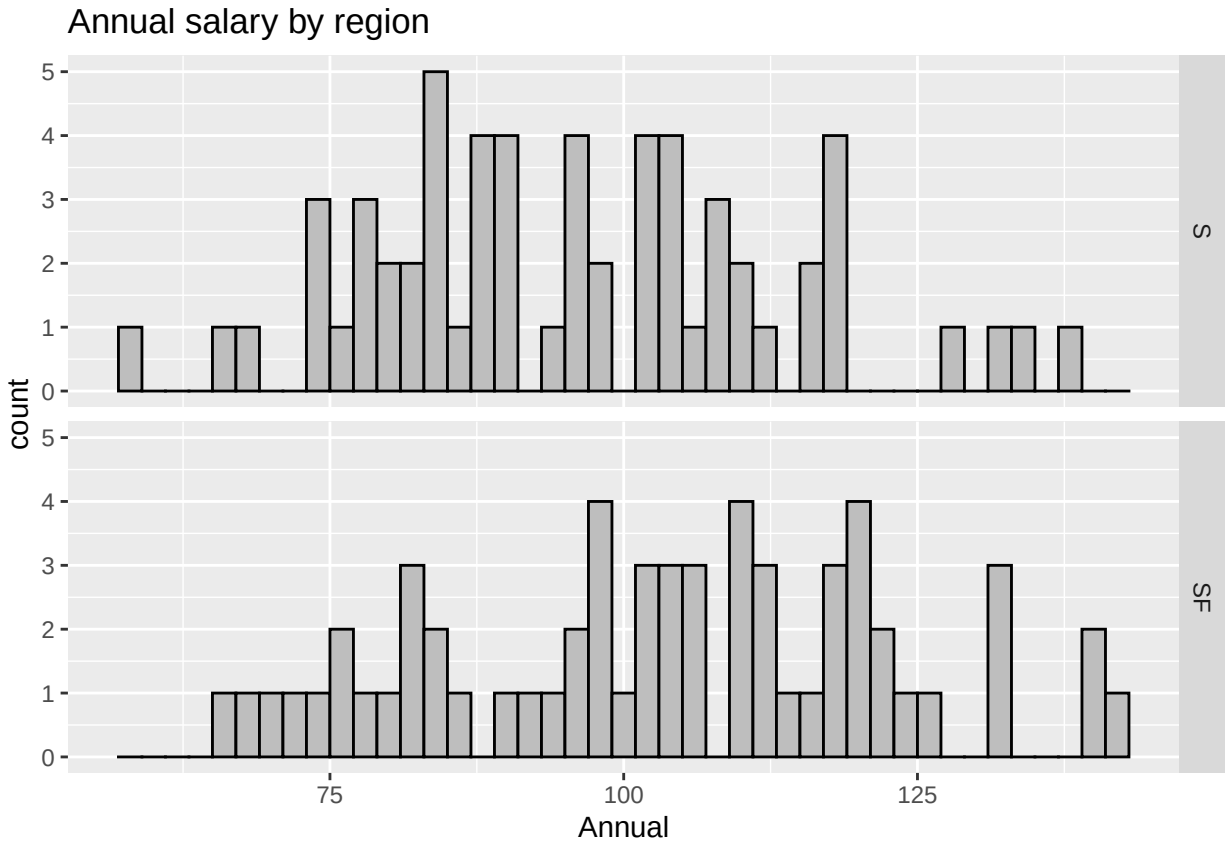
```
## Is there statistical evidence to suggest there is an interaction effect between Profession and city
```

Summary

(Histogram)

```
##      Annual       Prof    Region
##  Min.   : 57.65   BE:40   S :60
##  1st Qu.: 84.24   DS:40   SF:60
##  Median :100.75   SE:40
##  Mean   : 99.71
##  3rd Qu.:112.55
##  Max.   :142.31

## nT: 120 , a = 3 , b = 2
```

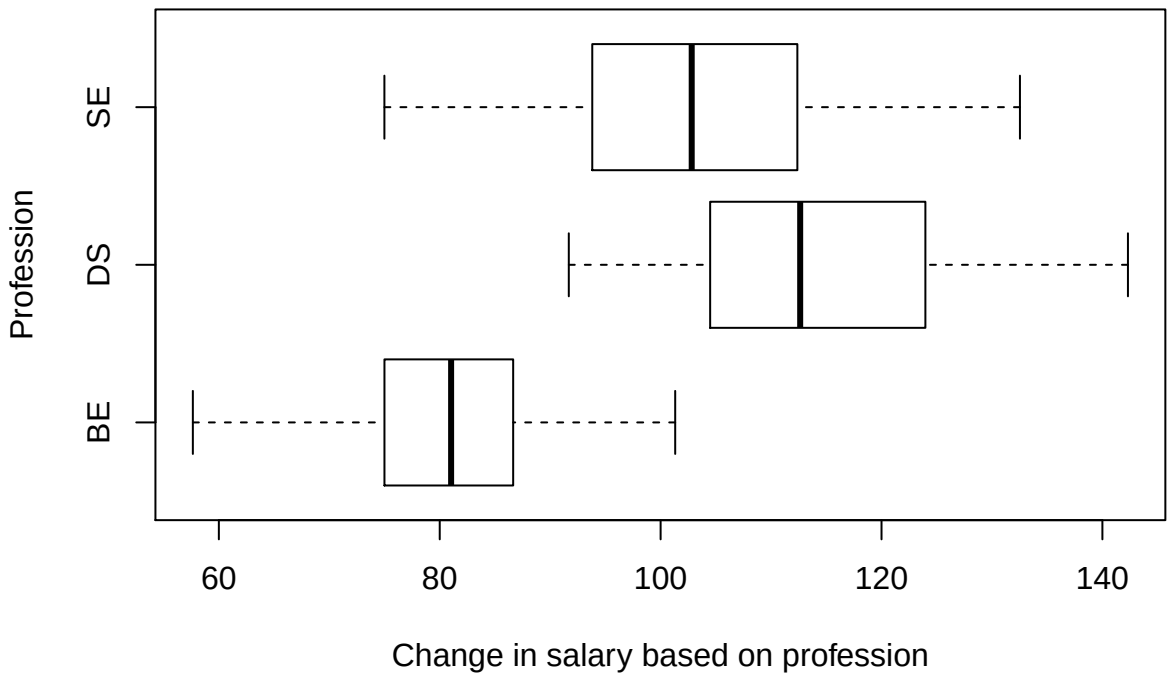

Annual salary by profession
count
Annual
BE
DS
SE
0
1
2
3
4
5
75
100
125




```
## The annual salary in thousands of dollars for 'Data Scientist', 'Software Engineer', 'Bioinformatics
```

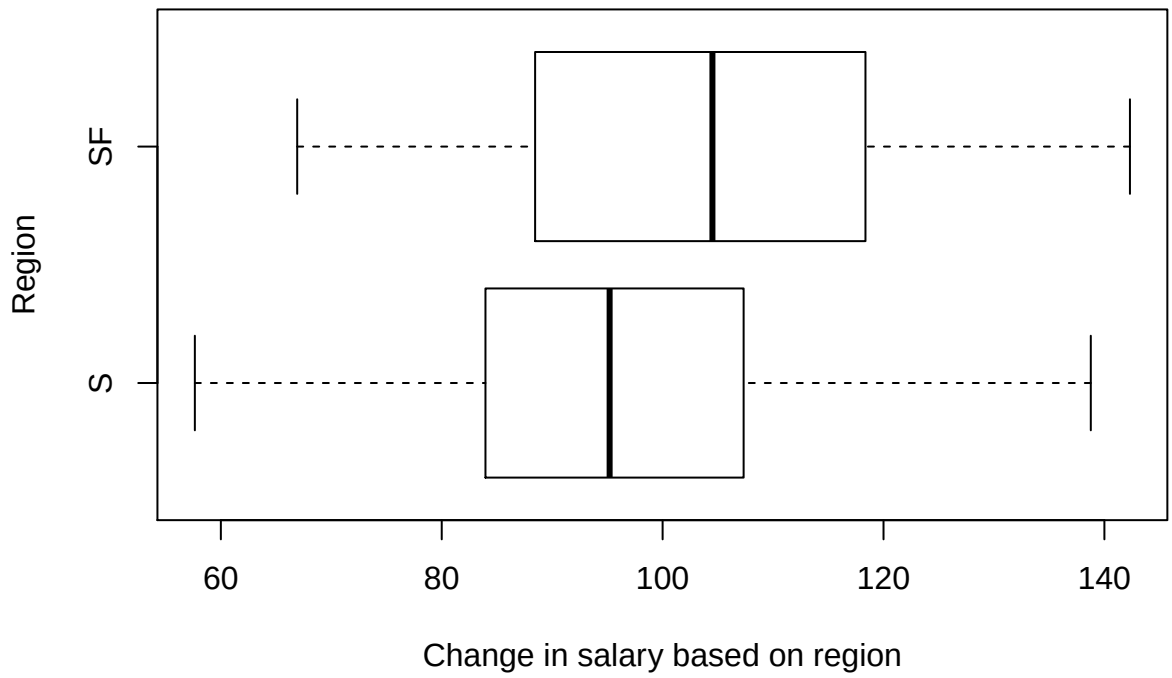

```
## By the plot, it seems to exist an interaction between Profession and Region
```

(BoxPlot)




```
## The average salary is comparetively the highest in the Data Scientist profession according to the his
```

```
##   Prof    Annual
## 1   BE  9.662515
## 2   DS 13.668190
## 3   SE 13.240313
```




```
## The average weight loss is comparetively the highest in San Francisco according to the histogram

##   Region   Annual
## 1      S 17.41791
## 2     SF 19.29842
```

(Means)

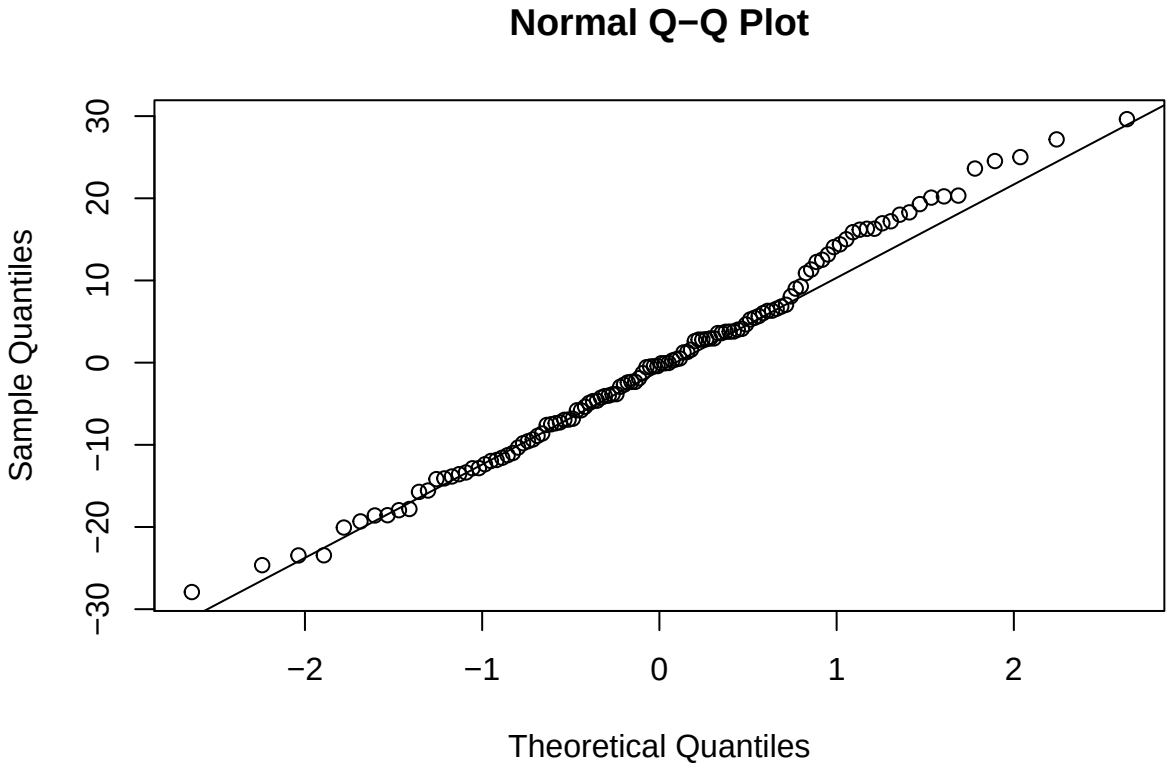
```
## Yij :

##          BE       DS        SE
## S  79.75485 112.5272  95.54875
## SF 82.41914 117.7688 110.26412
```

##Diagnostic

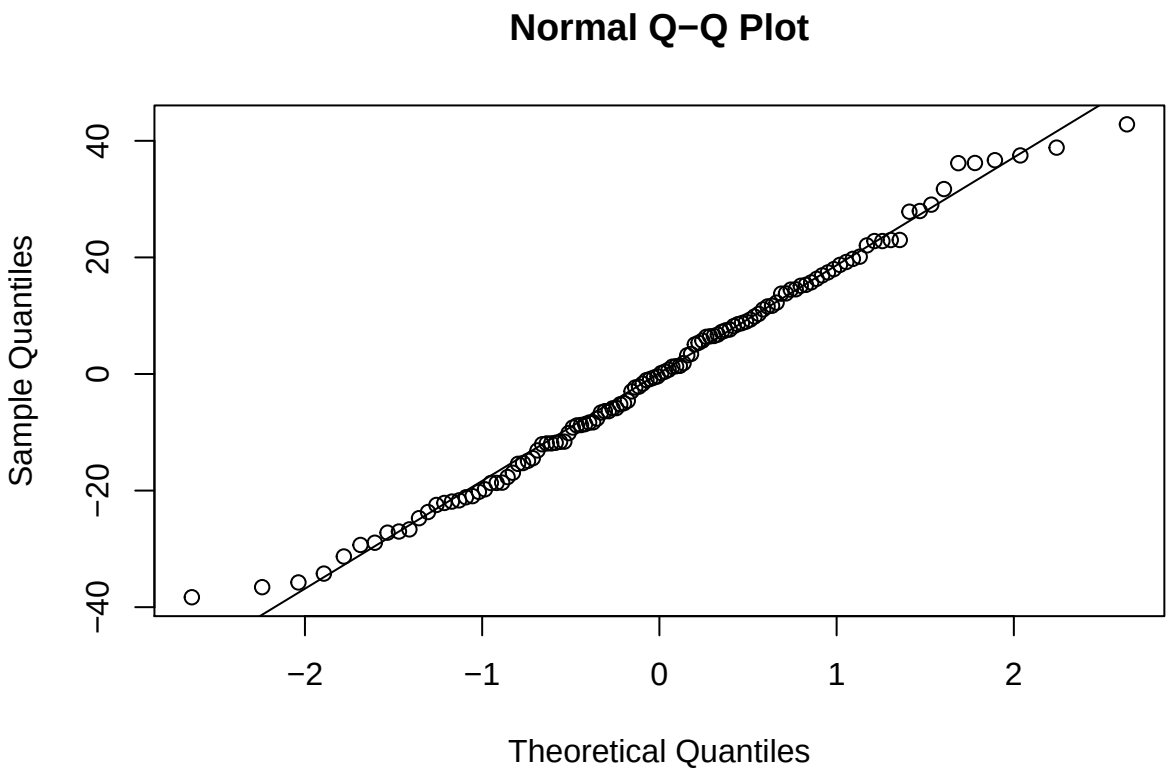
```
## We would like to find out if any form of this data will meet the assumption that all test statitics a
##     1: All subjects are randomly sampled
##     2: All levels of Factor A are independent
##     3: All levels of Factor B are independent
##     4: eijk ~ N(0 , sd = sigma-e)
##  Test for normal distrinution and check for outliers or any representation of skewed data and constan
```

(Assess Normality)

```
##
## Attaching package: 'EnvStats'

## The following objects are masked from 'package:stats':
##
##     predict, predict.lm

## The following object is masked from 'package:base':
##
##     print.default
```

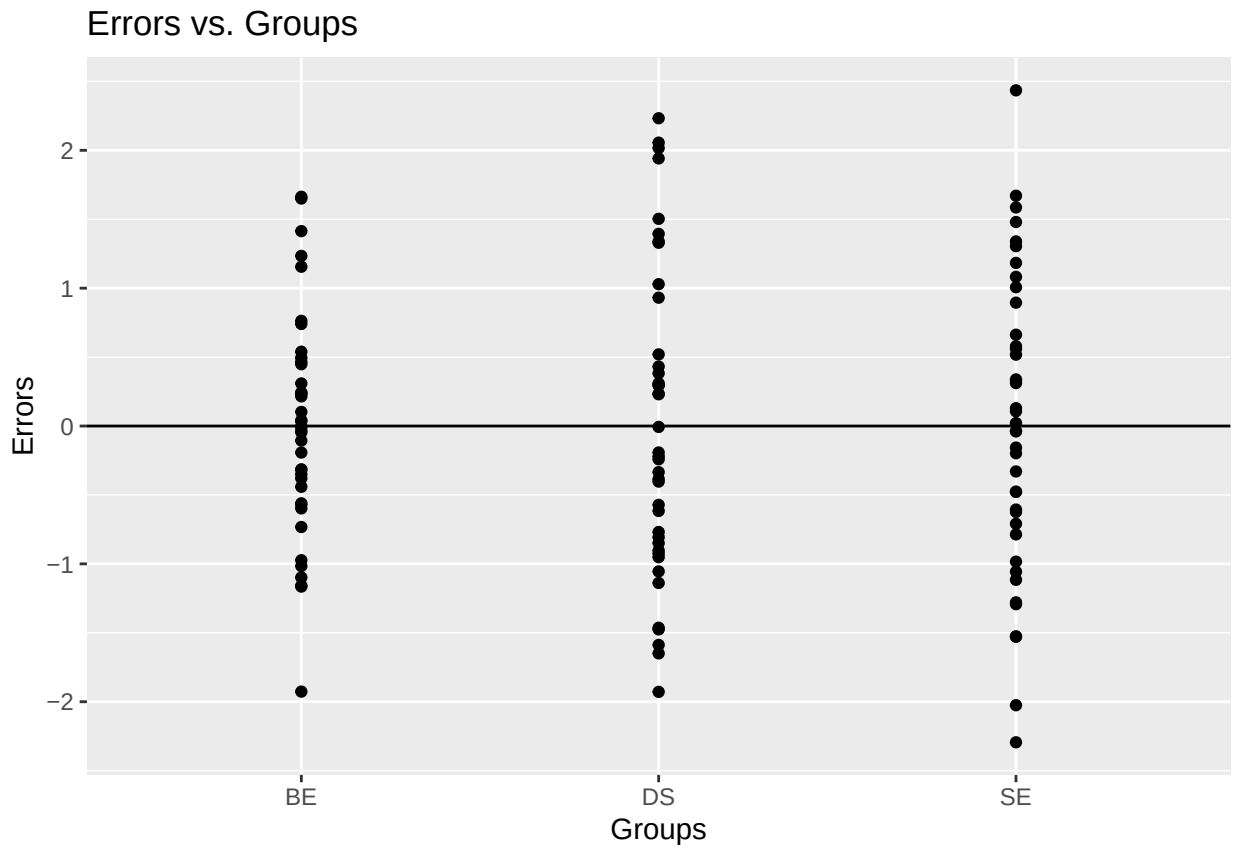


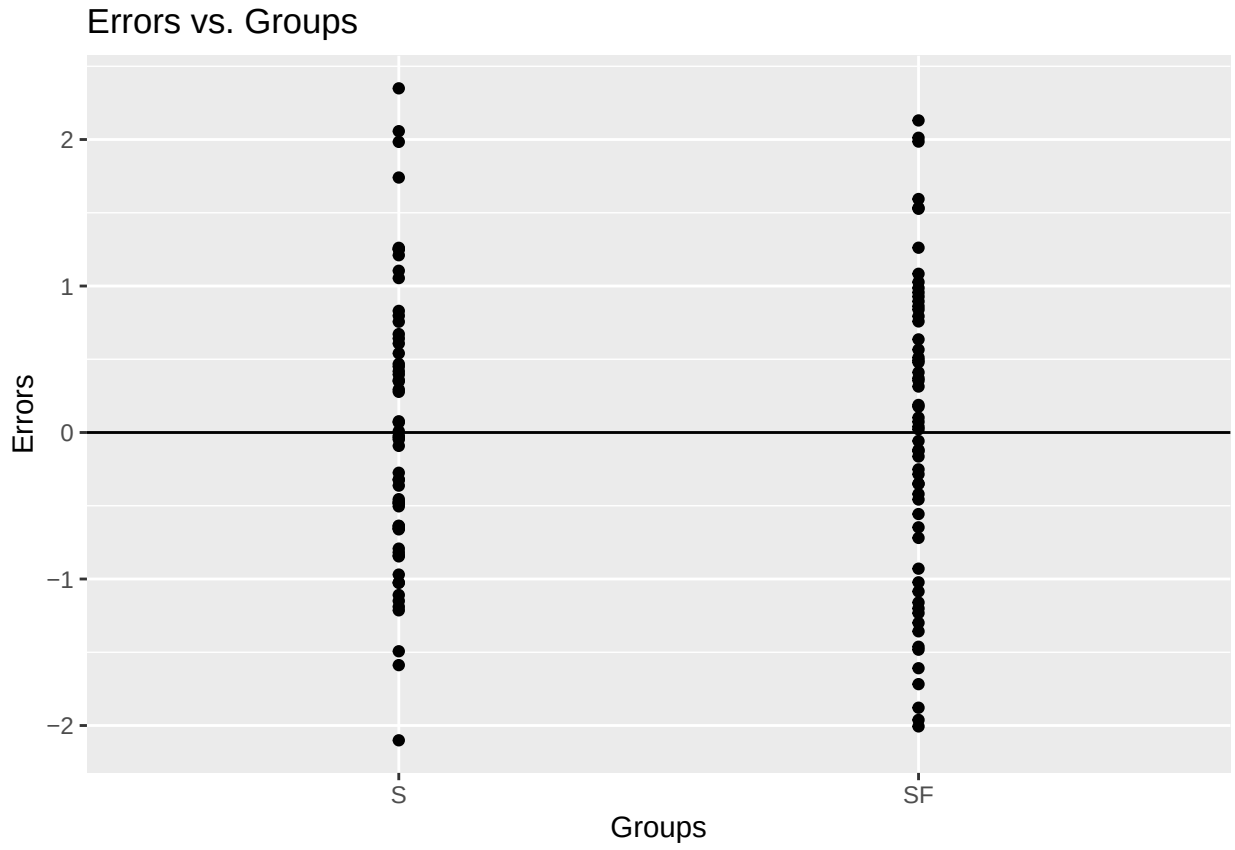

```
##
##  Shapiro-Wilk normality test
##
## data:  prof.e.i
## W = 0.99027, p-value = 0.5585

## SW p-val = 0.5585
##  The qq line and the plots seem to represent an approximate normal distribution, as y is not equal to
```




```
##
##  Shapiro-Wilk normality test
##
## data:  reg.e.i
## W = 0.98987, p-value = 0.5231

## SW p-val = 0.5231
##  The qq line and the plots seem to represent an approximate normal distribution, as y is not equal t
```

(homoscedasticity)


Errors vs. Groups
Errors
2
1
0
−1
−2
BE
DS
SE
Groups




```
## The errors for the sampled population seem to have roughly the same variance.
```

Analysis

(Interaction Effects)

```
## Assuming level of significance(alpha) as 0.05, test for interaction.

## (AB) : interaction model 15252.93

## (A+B): no-interaction model 16058.34

## test-statistic:  3.009798  p-value:  0.05323577
##  As the p-value is greater than the significance level, we fail to reject H0 and conclude that the m
##  As the no-interaction model is a better fit, we proceed testing for Factor A&B effects
```

(Factor Effects)

```
## test-statistic:  12.32177  p-value:  0.0006384655
##  As the p-value is less than alpha(0.05), we reject H0 and conclude that factor A effects exist
##  R2{A+B|B}:  0.5972622
##  the propotion of reduction in error when adding factor A to B is 59.7%.
```

```
## test-statistic:  86.0143  p-value:  1.233952e-23
##  As the p-value is less than alpha(0.05), we reject H0 and conclude that factor B effects exist
##  R2{A+B|A}:  0.09602243
##  the propotion of reduction in error when adding factor B to A is 9.6%.
```

(No-Interaction Two Factor ANOVA)

```
## Yijk =  µ.. + <U+03B3>i + dj + eijk
```

(Confidence Intervals)

```
## Factor A (Profession) and Factor B (Region), pairwise comparisions.

## µ11 - µ12, difference in annual salary for Bioinformatics Engineer in Seattle and San Francisco
##  -10.03494 4.706358 are the bounds
##  We are 95% confident that there exists no significant difference between a Bioinformatics Engineer i

## µ21 - µ22, difference in annual salary for Data Scientist in Seattle and San Francisco
##  -12.61233 2.128968 are the bounds
##  We are 95% confident that there exists no significant difference between a Data Scientist in Seattle

## µ31 - µ32, difference in annual salary for Software Engineer in Seattle and San Francisco
##  -22.08602 -7.344723 are the bounds
##  We are 95% confident that there exists a significant difference between a Software Engineer from Sea

## µ.1 - µ.2, difference in annual salary in Seattle and San Francisco
##  -15.18994 0.1090442 are the bounds
##  We are 95% confident that there exists no significant difference between on average in profession in

## µ21 - (µ11+µ31)/2, difference in annual salary between a Data Scietist and the average of the enginee
##  -37.13 -11.4362

## µ22 - (µ12+µ32)/2, difference in annual salary between a Data Scietist and the average of the enginee
##  -44.44423 -18.75044
```

Interpretation

```
##
##
## Alpha(0.05), is the probability of rejecting the claim that there is an interaction effect between pr
##
## With our data set and question of interest, we completed a Two Factor Anova Hypothesis test. First we
##
## Through our confidence intervals, we have concluded that software engineers from San Francisco earn a
```

Conclusion

```
## We can conclude that the best model for this report is the No-Interaction Two Factor ANOVA. We conclu
```

R Appendix