MSDS 660 Week 6 Project Assignment

Jeremy Beard

2022-08-03

Introduction The purpose of this week's assignment is to perform two-way analysis of variance (AOV) modeling. AOV is where the means of variables within a dataset are tested to see if they are significantly different or not. Different combinations of variables and the differences between them are also tested. For this project, we will be working with an engineer salary dataset. This dataset was provided to us as part of the class assignment. It is a clean dataset with no null values! This assignment is important as it shows data scientists which variables and specifically combinations of variables are significant. This can help improve model performance and prediction.

```
# Load the libraries - I probably loaded some unnecessary libraries but I'm keepin em, I've got enough
library(ggplot2)
library(devtools)
```

```
## Loading required package: usethis
```

```
library(data.table)
library(ggpubr)
library('magrittr')
library('dplyr')
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Load in the data first
dt <- read.csv("C:\\Users\\jerem\\OneDrive\\Documents\\School\\_REGIS\\2022-05_Summer\\MSDS660\\Week5\\
# Load 'data set to data.table
dt <- as.data.table(dt)

# Check structure of dt with different metadata probing commands
head(dt)
```

```
##    X Salary     Profession        Region
## 1: 1 126411 Data Scientist San Francisco
## 2: 2 108402 Data Scientist San Francisco
## 3: 3  99399 Data Scientist San Francisco
```

```
## 4: 4  91381 Data Scientist San Francisco
## 5: 5 105023 Data Scientist San Francisco
## 6: 6 108944 Data Scientist San Francisco
```

```
nrow(dt)
```

```
## [1] 180
```

```
ncol(dt)
```

```
## [1] 4
```

```
summary(dt)
```

```
##        X              Salary        Profession           Region
##  Min.   :  1.00   Min.   : 57646   Length:180         Length:180
##  1st Qu.: 45.75   1st Qu.: 80409   Class :character   Class :character
##  Median : 90.50   Median : 92284   Mode  :character   Mode  :character
##  Mean   : 90.50   Mean   : 94199
##  3rd Qu.:135.25   3rd Qu.:105932
##  Max.   :180.00   Max.   :140179
```

```
str(dt)
```

```
## Classes 'data.table' and 'data.frame':   180 obs. of  4 variables:
##  $ X         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Salary    : int  126411 108402 99399 91381 105023 108944 123952 108217 103722 140179 ...
##  $ Profession: chr  "Data Scientist" "Data Scientist" "Data Scientist" "Data Scientist" ...
##  $ Region    : chr  "San Francisco" "San Francisco" "San Francisco" "San Francisco" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

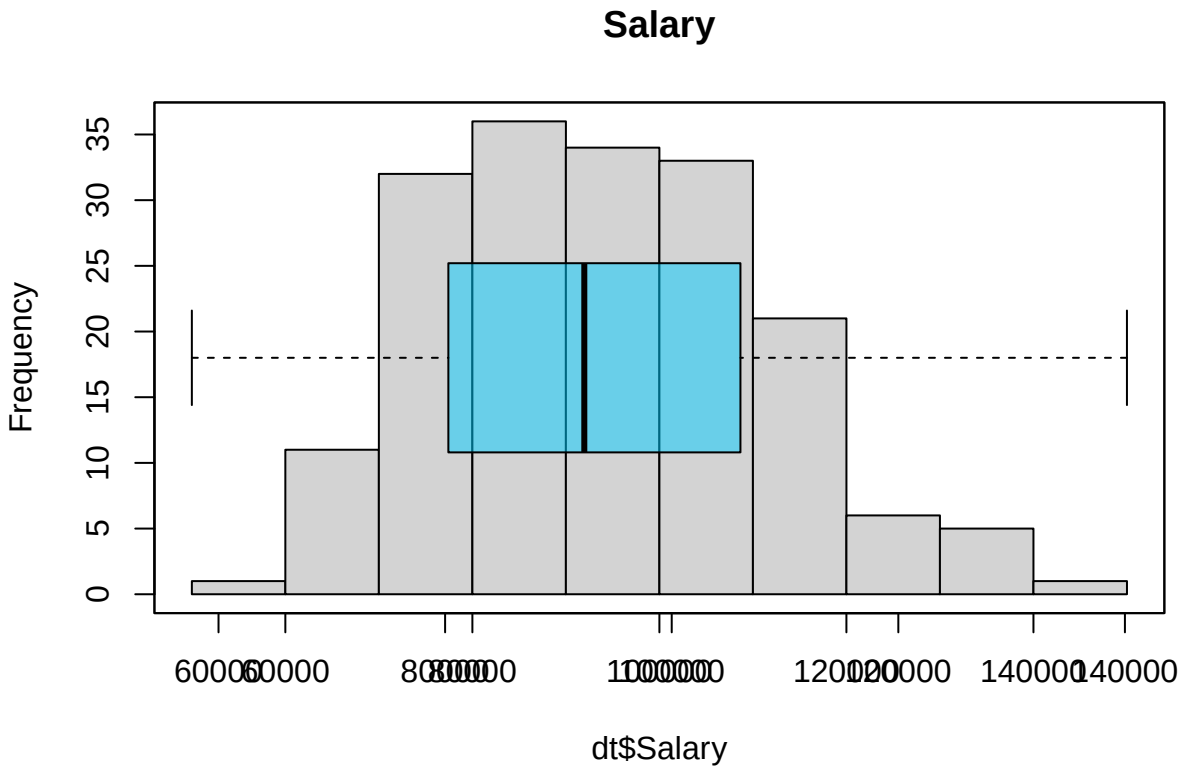
```
# I have visually looked at the data and there are no null values!
```

Methods For this week's assignment, we will be first creating many plots to show the shape of the data and how Salary is related to the different other parameters.

After creating plots, we will create an AOV model and will optimize it using analysis of p-values. After optimizing the model, we will perform a TukeyHSD analysis and see which combinations of variables have significant differences to other specific combinations of variables.

Finally, we will perform a Shapiro test of the residuals to see if the residuals are normally distributed.

```
# Plot histogram of Salary
hist(dt$Salary, main="Salary")
par(new = TRUE)
boxplot(dt$Salary, horizontal = TRUE, col = rgb(0, 0.8, 1, alpha = 0.5))
box()
```

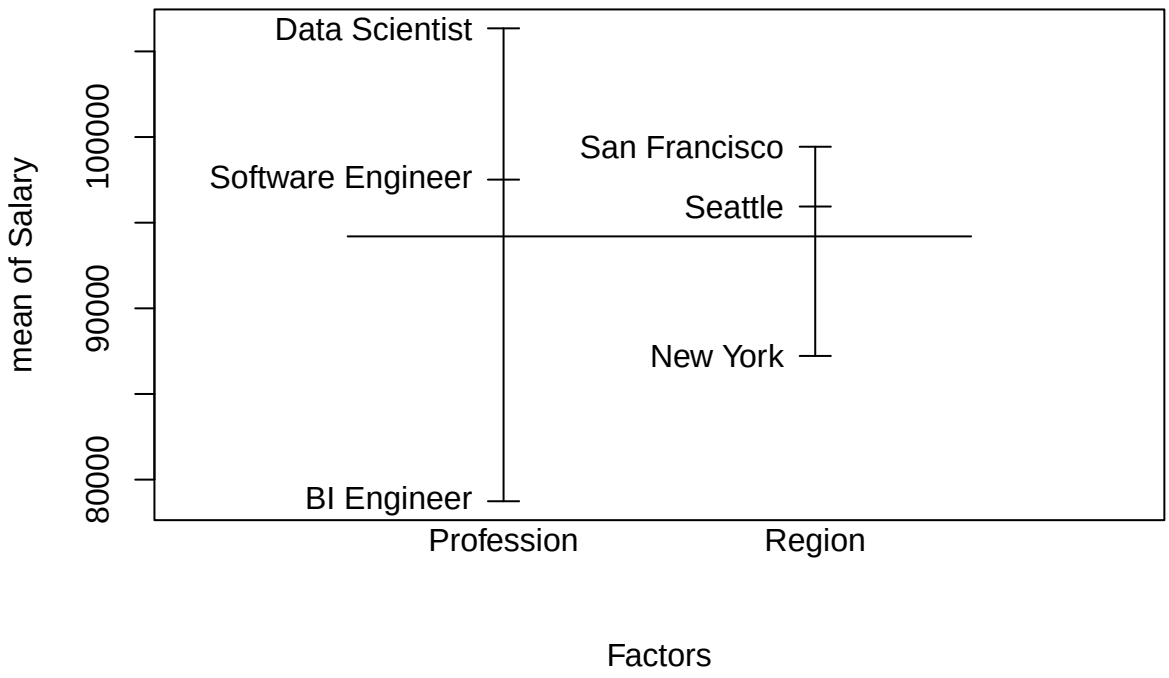


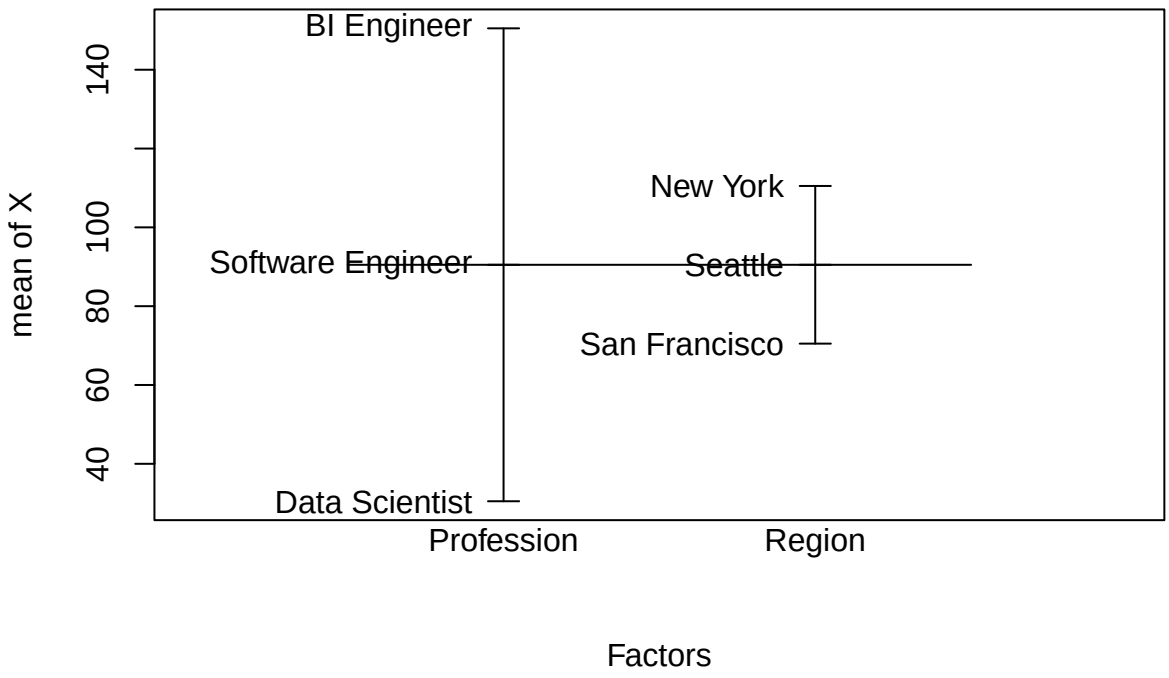

```
# Convert Profession and Region to factors
cols <- c("Profession", "Region")
dt %<>% mutate_each_(funs(factor(.)),cols)
```

```
## Warning: `mutate_each_()` was deprecated in dplyr 0.7.0.
## Please use `across()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

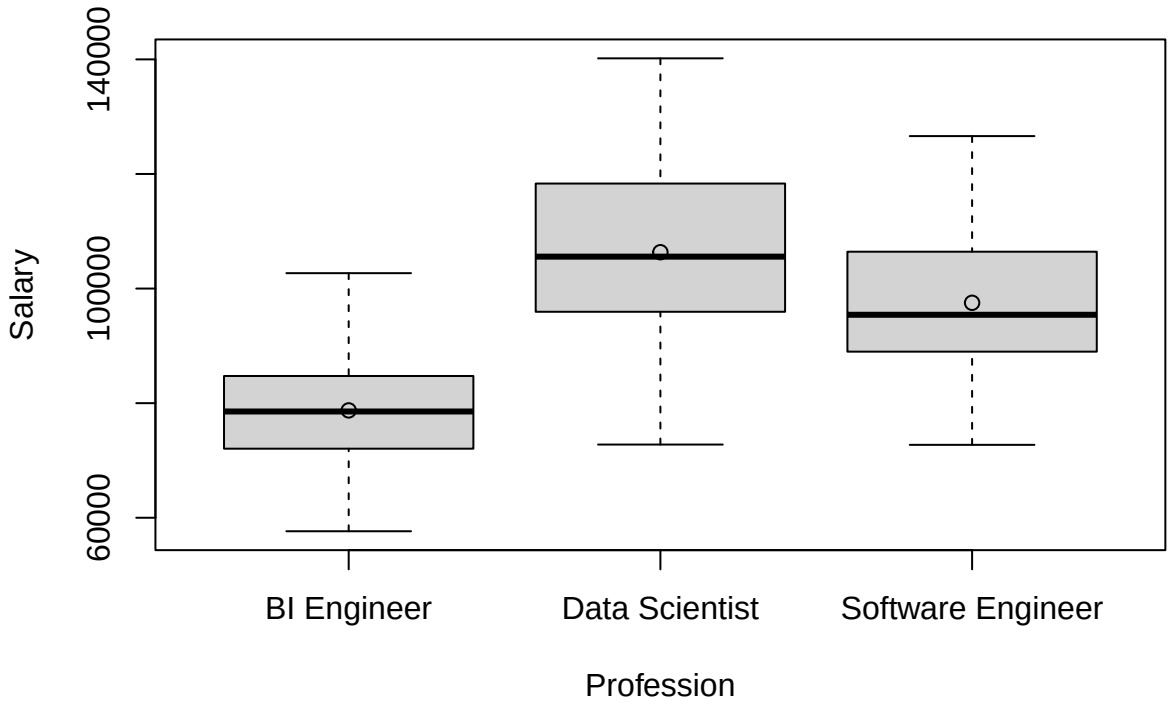
```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

```
# Plot Salary vs the 2 other factors
plot.design(Salary ~ ., data = dt)
```

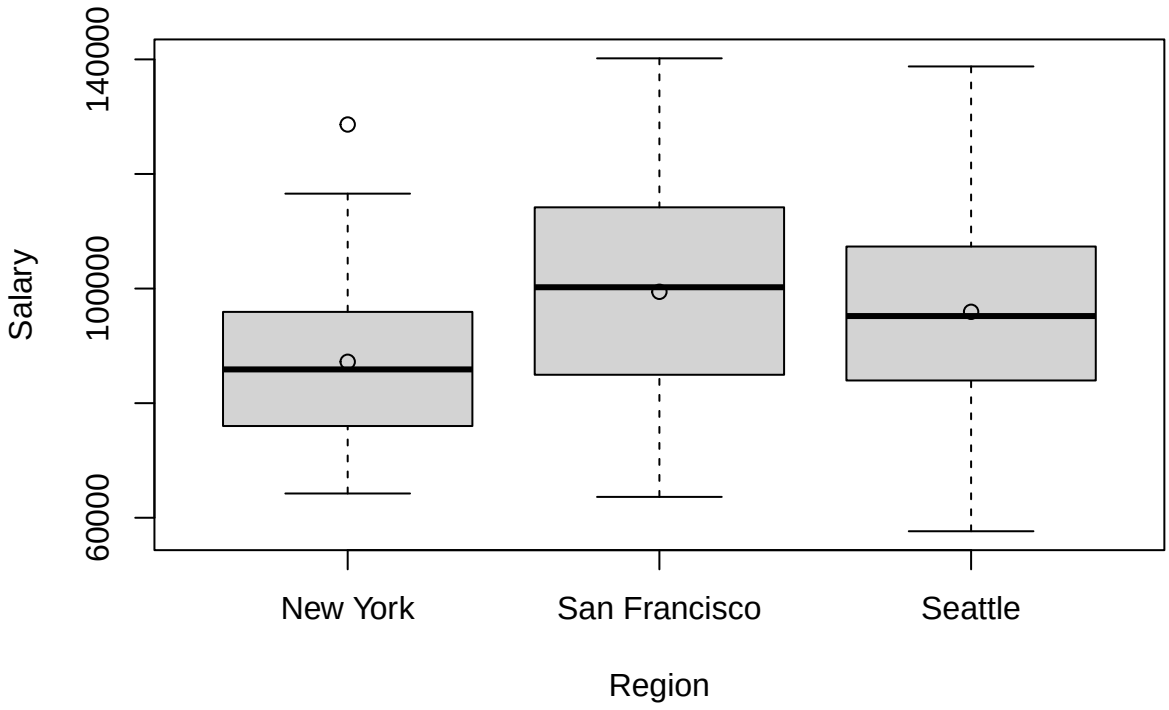

Data Scientist
Software Engineer
BI Engineer
San Francisco
Seattle
New York
mean of Salary
80000
90000
100000
Profession
Region
Factors



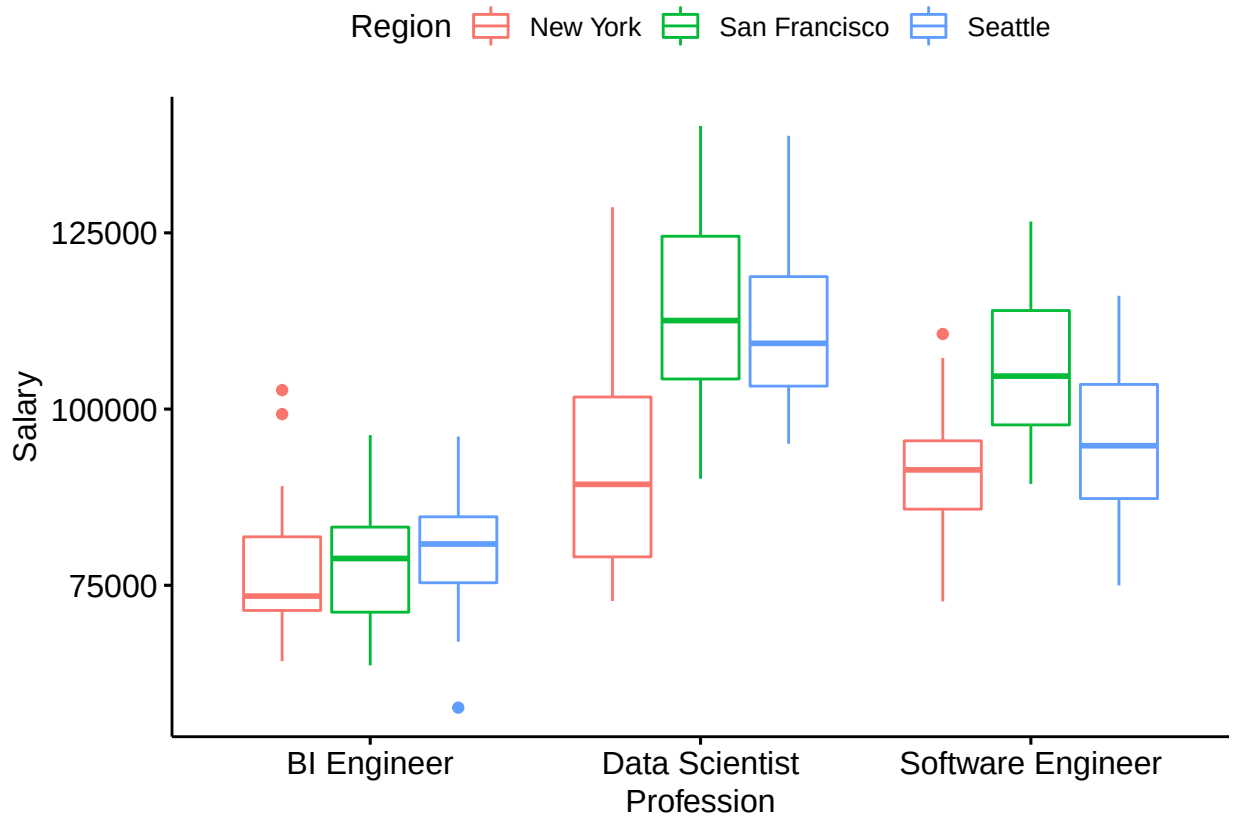

```
# Plot Individual Boxplots with means
boxplot(Salary ~ Profession, data = dt)
points(dt[, mean(Salary), by=Profession])
```



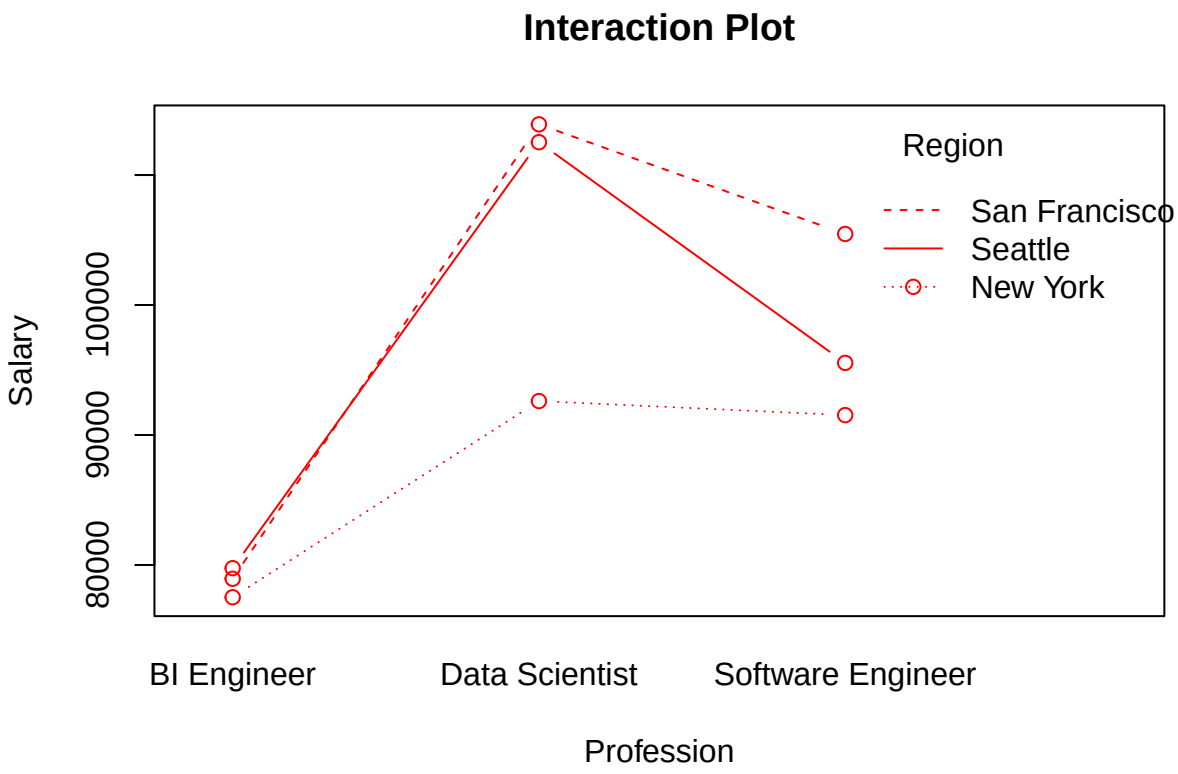

```
boxplot(Salary ~ Region, data = dt)
points(dt[, mean(Salary), by=Region])
```




```
# Plot 3-dimensional boxplot
ggboxplot(dt, x = "Profession", y = "Salary", color = "Region")
```




```
# Create interaction plot looking at Region and Profession
interaction.plot(x.factor = dt$Profession,
                 trace.factor = dt$Region,
                 response = dt$Salary,
                 fun = mean,
                 type = "b",  # shows each point
                 main = "Interaction Plot",
                 legend = TRUE,
                 trace.label = "Region",
                 xlab = "Profession",
                 ylab="Salary",
                 pch=c(1),
                 col = c("Red"))
```



Results From the analysis above and below, it was found that all variables had significant differences. Both Region, Profession, and Profession:Region had p values under 0.05 which indicated the previous statement. From that, only one iteration of the AOV model was performed. Lucky break!

After optimizing the model, a TukeyHSD analysis was performed on the AOV model and it was found that only Seattle-San Francisco have insignificant differences, from any of the Professions or Regions. All other combinations of Professions and combinations of Regions had significant differences.

The residuals of the model showed a great fit to the data! Wow, and only on one iteration of the AOV parameters. We got lucky with this dataset.

Finally, in the Shapiro test of the residuals, it was found that the p-value was only 0.032. This showed that the residuals could not be considered normal.

```
# report results

# Build ANOVA model - the * is giving interactions. Show anova fit summary
model <- aov(Salary ~ Profession * Region, data = dt)
summary(model)
```

```
##                    Df    Sum Sq   Mean Sq F value   Pr(>F)
## Profession          2 2.386e+10 1.193e+10  86.098  < 2e-16 ***
## Region              2 4.750e+09 2.375e+09  17.143 1.64e-07 ***
## Profession:Region   4 3.037e+09 7.593e+08   5.481 0.000355 ***
## Residuals         171 2.369e+10 1.385e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Everything is significant!


model2 <- aov(Salary ~ Profession + Region + Profession:Region, data = dt)
summary(model2)
```

```
##                     Df    Sum Sq   Mean Sq F value   Pr(>F)
## Profession           2 2.386e+10 1.193e+10  86.098  < 2e-16 ***
## Region               2 4.750e+09 2.375e+09  17.143 1.64e-07 ***
## Profession:Region    4 3.037e+09 7.593e+08   5.481 0.000355 ***
## Residuals          171 2.369e+10 1.385e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Based on the model people like hot dogs and ice cream the same.  There is a Profession Salary depends
# Profession and Region together interact and affect people Salary

# Perform TukeyHSD to check if which interactions have a significant difference
TukeyHSD(model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Salary ~ Profession * Region, data = dt)
##
## $Profession
##                                      diff       lwr       upr     p adj
## Data Scientist-BI Engineer       27608.02  22527.33 32688.707 0.0000000
## Software Engineer-BI Engineer    18776.57  13695.88 23857.257 0.0000000
## Software Engineer-Data Scientist -8831.45 -13912.14 -3750.759 0.0001807
##
## $Region
##                               diff       lwr       upr     p adj
## San Francisco-New York 12214.900  7134.209 17295.591 0.0000002
## Seattle-New York        8723.683  3642.993 13804.374 0.0002197
## Seattle-San Francisco  -3491.217 -8571.907  1589.474 0.2380471
##
## $`Profession:Region`
##                                                                    diff
## Data Scientist:New York-BI Engineer:New York                    15092.65
## Software Engineer:New York-BI Engineer:New York                 14010.80
## BI Engineer:San Francisco-BI Engineer:New York                   1421.35
## Data Scientist:San Francisco-BI Engineer:New York               36380.45
## Software Engineer:San Francisco-BI Engineer:New York            27946.35
## BI Engineer:Seattle-BI Engineer:New York                         2236.10
## Data Scientist:Seattle-BI Engineer:New York                     35008.40
## Software Engineer:Seattle-BI Engineer:New York                  18030.00
## Software Engineer:New York-Data Scientist:New York              -1081.85
## BI Engineer:San Francisco-Data Scientist:New York             -13671.30
## Data Scientist:San Francisco-Data Scientist:New York           21287.80
## Software Engineer:San Francisco-Data Scientist:New York        12853.70
## BI Engineer:Seattle-Data Scientist:New York                   -12856.55
## Data Scientist:Seattle-Data Scientist:New York                 19915.75
## Software Engineer:Seattle-Data Scientist:New York               2937.35
## BI Engineer:San Francisco-Software Engineer:New York          -12589.45
```

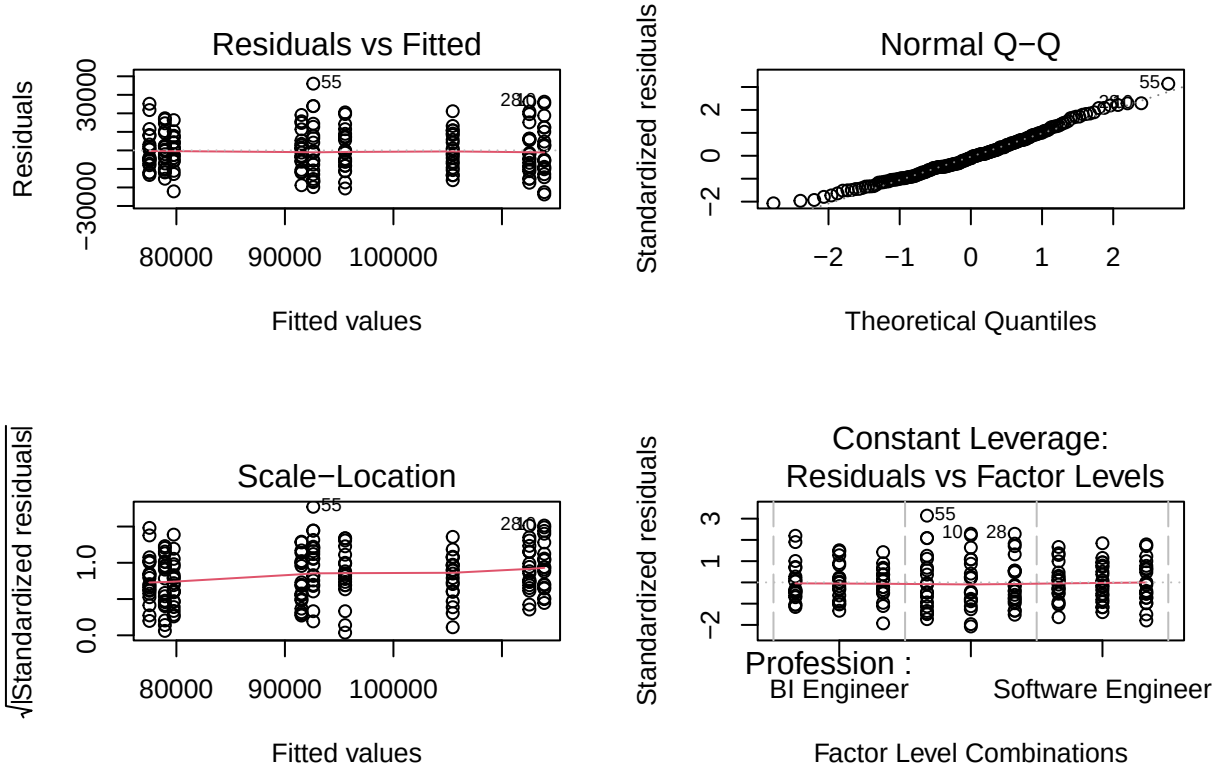
```
## Data Scientist:San Francisco-Software Engineer:New York       22369.65
## Software Engineer:San Francisco-Software Engineer:New York    13935.55
## BI Engineer:Seattle-Software Engineer:New York               -11774.70
## Data Scientist:Seattle-Software Engineer:New York             20997.60
## Software Engineer:Seattle-Software Engineer:New York           4019.20
## Data Scientist:San Francisco-BI Engineer:San Francisco        34959.10
## Software Engineer:San Francisco-BI Engineer:San Francisco     26525.00
## BI Engineer:Seattle-BI Engineer:San Francisco                   814.75
## Data Scientist:Seattle-BI Engineer:San Francisco              33587.05
## Software Engineer:Seattle-BI Engineer:San Francisco           16608.65
## Software Engineer:San Francisco-Data Scientist:San Francisco  -8434.10
## BI Engineer:Seattle-Data Scientist:San Francisco            -34144.35
## Data Scientist:Seattle-Data Scientist:San Francisco          -1372.05
## Software Engineer:Seattle-Data Scientist:San Francisco      -18350.45
## BI Engineer:Seattle-Software Engineer:San Francisco         -25710.25
## Data Scientist:Seattle-Software Engineer:San Francisco        7062.05
## Software Engineer:Seattle-Software Engineer:San Francisco    -9916.35
## Data Scientist:Seattle-BI Engineer:Seattle                   32772.30
## Software Engineer:Seattle-BI Engineer:Seattle                15793.90
## Software Engineer:Seattle-Data Scientist:Seattle            -16978.40
##                                                                   lwr
## Data Scientist:New York-BI Engineer:New York                 3398.181
## Software Engineer:New York-BI Engineer:New York              2316.331
## BI Engineer:San Francisco-BI Engineer:New York             -10273.119
## Data Scientist:San Francisco-BI Engineer:New York           24685.981
## Software Engineer:San Francisco-BI Engineer:New York        16251.881
## BI Engineer:Seattle-BI Engineer:New York                    -9458.369
## Data Scientist:Seattle-BI Engineer:New York                 23313.931
## Software Engineer:Seattle-BI Engineer:New York               6335.531
## Software Engineer:New York-Data Scientist:New York         -12776.319
## BI Engineer:San Francisco-Data Scientist:New York          -25365.769
## Data Scientist:San Francisco-Data Scientist:New York         9593.331
## Software Engineer:San Francisco-Data Scientist:New York      1159.231
## BI Engineer:Seattle-Data Scientist:New York                -24551.019
## Data Scientist:Seattle-Data Scientist:New York               8221.281
## Software Engineer:Seattle-Data Scientist:New York           -8757.119
## BI Engineer:San Francisco-Software Engineer:New York       -24283.919
## Data Scientist:San Francisco-Software Engineer:New York     10675.181
## Software Engineer:San Francisco-Software Engineer:New York   2241.081
## BI Engineer:Seattle-Software Engineer:New York             -23469.169
## Data Scientist:Seattle-Software Engineer:New York            9303.131
## Software Engineer:Seattle-Software Engineer:New York        -7675.269
## Data Scientist:San Francisco-BI Engineer:San Francisco      23264.631
## Software Engineer:San Francisco-BI Engineer:San Francisco   14830.531
## BI Engineer:Seattle-BI Engineer:San Francisco              -10879.719
## Data Scientist:Seattle-BI Engineer:San Francisco            21892.581
## Software Engineer:Seattle-BI Engineer:San Francisco          4914.181
## Software Engineer:San Francisco-Data Scientist:San Francisco -20128.569
## BI Engineer:Seattle-Data Scientist:San Francisco           -45838.819
## Data Scientist:Seattle-Data Scientist:San Francisco        -13066.519
## Software Engineer:Seattle-Data Scientist:San Francisco     -30044.919
## BI Engineer:Seattle-Software Engineer:San Francisco        -37404.719
## Data Scientist:Seattle-Software Engineer:San Francisco      -4632.419
## Software Engineer:Seattle-Software Engineer:San Francisco  -21610.819
```

```
## Data Scientist:Seattle-BI Engineer:Seattle                    21077.831
## Software Engineer:Seattle-BI Engineer:Seattle                  4099.431
## Software Engineer:Seattle-Data Scientist:Seattle             -28672.869
##                                                                      upr
## Data Scientist:New York-BI Engineer:New York                  26787.11898
## Software Engineer:New York-BI Engineer:New York               25705.26898
## BI Engineer:San Francisco-BI Engineer:New York                13115.81898
## Data Scientist:San Francisco-BI Engineer:New York             48074.91898
## Software Engineer:San Francisco-BI Engineer:New York          39640.81898
## BI Engineer:Seattle-BI Engineer:New York                      13930.56898
## Data Scientist:Seattle-BI Engineer:New York                   46702.86898
## Software Engineer:Seattle-BI Engineer:New York                29724.46898
## Software Engineer:New York-Data Scientist:New York            10612.61898
## BI Engineer:San Francisco-Data Scientist:New York             -1976.83102
## Data Scientist:San Francisco-Data Scientist:New York          32982.26898
## Software Engineer:San Francisco-Data Scientist:New York       24548.16898
## BI Engineer:Seattle-Data Scientist:New York                   -1162.08102
## Data Scientist:Seattle-Data Scientist:New York                31610.21898
## Software Engineer:Seattle-Data Scientist:New York             14631.81898
## BI Engineer:San Francisco-Software Engineer:New York           -894.98102
## Data Scientist:San Francisco-Software Engineer:New York       34064.11898
## Software Engineer:San Francisco-Software Engineer:New York    25630.01898
## BI Engineer:Seattle-Software Engineer:New York                  -80.23102
## Data Scientist:Seattle-Software Engineer:New York             32692.06898
## Software Engineer:Seattle-Software Engineer:New York          15713.66898
## Data Scientist:San Francisco-BI Engineer:San Francisco        46653.56898
## Software Engineer:San Francisco-BI Engineer:San Francisco     38219.46898
## BI Engineer:Seattle-BI Engineer:San Francisco                 12509.21898
## Data Scientist:Seattle-BI Engineer:San Francisco              45281.51898
## Software Engineer:Seattle-BI Engineer:San Francisco           28303.11898
## Software Engineer:San Francisco-Data Scientist:San Francisco   3260.36898
## BI Engineer:Seattle-Data Scientist:San Francisco             -22449.88102
## Data Scientist:Seattle-Data Scientist:San Francisco           10322.41898
## Software Engineer:Seattle-Data Scientist:San Francisco        -6655.98102
## BI Engineer:Seattle-Software Engineer:San Francisco          -14015.78102
## Data Scientist:Seattle-Software Engineer:San Francisco        18756.51898
## Software Engineer:Seattle-Software Engineer:San Francisco      1778.11898
## Data Scientist:Seattle-BI Engineer:Seattle                    44466.76898
## Software Engineer:Seattle-BI Engineer:Seattle                 27488.36898
## Software Engineer:Seattle-Data Scientist:Seattle              -5283.93102
##                                                                    p adj
## Data Scientist:New York-BI Engineer:New York                  0.0024207
## Software Engineer:New York-BI Engineer:New York               0.0069368
## BI Engineer:San Francisco-BI Engineer:New York                0.9999868
## Data Scientist:San Francisco-BI Engineer:New York             0.0000000
## Software Engineer:San Francisco-BI Engineer:New York          0.0000000
## BI Engineer:Seattle-BI Engineer:New York                      0.9995865
## Data Scientist:Seattle-BI Engineer:New York                   0.0000000
## Software Engineer:Seattle-BI Engineer:New York                0.0000975
## Software Engineer:New York-Data Scientist:New York            0.9999984
## BI Engineer:San Francisco-Data Scientist:New York             0.0094978
## Data Scientist:San Francisco-Data Scientist:New York          0.0000017
## Software Engineer:San Francisco-Data Scientist:New York       0.0195719
## BI Engineer:Seattle-Data Scientist:New York                   0.0195243
```

```
## Data Scientist:Seattle-Data Scientist:New York                0.0000098
## Software Engineer:Seattle-Data Scientist:New York             0.9970431
## BI Engineer:San Francisco-Software Engineer:New York          0.0244634
## Data Scientist:San Francisco-Software Engineer:New York       0.0000004
## Software Engineer:San Francisco-Software Engineer:New York    0.0074423
## BI Engineer:Seattle-Software Engineer:New York                0.0470207
## Data Scientist:Seattle-Software Engineer:New York             0.0000024
## Software Engineer:Seattle-Software Engineer:New York          0.9764101
## Data Scientist:San Francisco-BI Engineer:San Francisco        0.0000000
## Software Engineer:San Francisco-BI Engineer:San Francisco     0.0000000
## BI Engineer:Seattle-BI Engineer:San Francisco                 0.9999998
## Data Scientist:Seattle-BI Engineer:San Francisco              0.0000000
## Software Engineer:Seattle-BI Engineer:San Francisco           0.0004900
## Software Engineer:San Francisco-Data Scientist:San Francisco 0.3687205
## BI Engineer:Seattle-Data Scientist:San Francisco              0.0000000
## Data Scientist:Seattle-Data Scientist:San Francisco           0.9999900
## Software Engineer:Seattle-Data Scientist:San Francisco        0.0000667
## BI Engineer:Seattle-Software Engineer:San Francisco           0.0000000
## Data Scientist:Seattle-Software Engineer:San Francisco        0.6165068
## Software Engineer:Seattle-Software Engineer:San Francisco     0.1687988
## Data Scientist:Seattle-BI Engineer:Seattle                    0.0000000
## Software Engineer:Seattle-BI Engineer:Seattle                 0.0011759
## Software Engineer:Seattle-Data Scientist:Seattle              0.0003253
```

```r
#We can see Seattle-San Francisco is not significant.
#We can also see this when we look at the interaction plot.
#The lines of Seattle and San Francisco were very similar.

# Plot the residuals of the fit
par(mfrow = c(2,2))
plot(model)
```

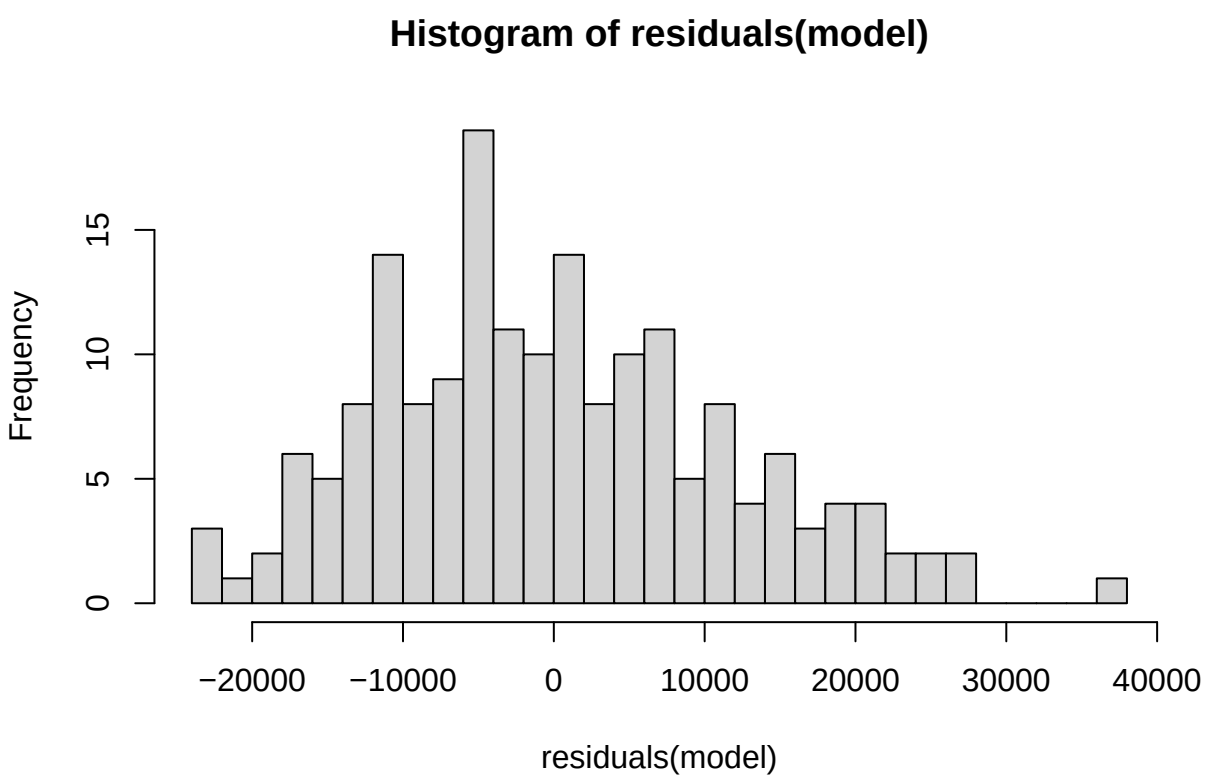



```
par(mfrow = c(1,1))
```

```
# Perform Shapiro test to see if residuals are normally distributed.
shapiro.test(residuals(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model)
## W = 0.98346, p-value = 0.03161
```

```
hist(residuals(model), breaks=40)
```



Conclusion From the results, it can be said in a general sense that both Region and Profession have significant differences and should be included in the AOV model. They both have p values under 0.05 when a summary of the AOV model was given. The model showed us that the salaries between engineers of Seattle and San Francisco were very similar compared to comparisons of those variables with New York. The plot of Salary vs. the other design factors showed us that data scientists get paid far more than software engineers or BI engineers. Additionally, People in San Francisco get paid more than those in Seattle or those in New York. However, as BI engineers, those who live in Seattle generally get paid more than those in New York or San Francisco. The interaction plot showed us that BI engineers all get paid a similar amount regardless of region. The plot of the model showed a good fit to the data!

Some things to consider in the future are more parameters to compare such as relationship status, race, and more. We can also explore more of the combinations from the TukeyHSD analysis and try to define why the p values are the way that they are, why they are significant differences or not.

Thank you!

Jeremy Beard

References

1. What P-Value Tells Us. (2022, May 18). Investopedia. Retrieved August 3, 2022, from https://www.investopedia.com/terms/p/p-value.asp