

# MSDS 660 Week 4 Homework Assignment

Jeremy Beard

2022-07-27

## Introduction

In this week's assignment, we will be exploring a dataset found on Kaggle. This dataset explores the salaries of data scientists from different locations, experience levels, and more. This assignment will focus on one-way ANOVA analyses and also performing post hoc analyses. The post hoc analysis that I chose to perform during this assignment was the Tukey HSD pairwise comparison. My null hypothesis when performing the regression model with categorical data was that there is no significant difference between the means of the different pairwise comparisons of categorical levels of experience. My alternate hypothesis when performing the model and analysis was that there would be a significant difference between the means of the different Tukey HSD pairwise comparisons of levels of experience. This type of analysis is important because it helps tell the data analyst or data scientist whether the results they are seeing are significant or not when fitting models to data. This helps provide a degree of confidence in the conclusions reached at the end of the analysis.

In the code below, we simply load the csv file, probe the data for its size, shape, and metadata. Finally, we count the amount of null values in the dataset which in this case is zero!

```
# code may include summaries, head/tail, na, imputation, etc.

##### MSDS660 Homework Assignment - Week 4 - ANOVA #####
##### Jeremy Beard

# Load the required libraries
# import and load "ds_salaries.csv" data
# It is a data science salary dataset from :
https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries
dt <- read.csv("C:\\Users\\jerem\\OneDrive\\Documents\\School\\_REGIS\\2022-
05_Summer\\MSDS660\\Week4\\ds_salaries.csv", sep = ",")

# Convert data to data table or data frame or whatever
dt <- as.data.frame(dt)

head(dt)
```

##	X	work_year	experience_level	employment_type	job_title
## 1	0	2020	MI	FT	Data Scientist
## 2	1	2020	SE	FT	Machine Learning Scientist
## 3	2	2020	SE	FT	Big Data Engineer

```
## 4 3      2020      MI      FT      Product Data Analyst
## 5 4      2020      SE      FT      Machine Learning Engineer
## 6 5      2020      EN      FT      Data Analyst
## salary salary_currency salary_in_usd employee_residence remote_ratio
## 1 70000      EUR      79833      DE      0
## 2 260000     USD      260000     JP      0
## 3 85000      GBP      109024     GB      50
## 4 20000      USD      20000     HN      0
## 5 150000     USD      150000     US      50
## 6 72000      USD      72000     US      100
## company_location company_size
## 1      DE      L
## 2      JP      S
## 3      GB      M
## 4      HN      S
## 5      US      L
## 6      US      L
```

```
nrow(dt)
```

```
## [1] 607
```

```
ncol(dt)
```

```
## [1] 12
```

```
str(dt)
```

```
## '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" ...
```

```
summary(dt)
```

```
##      X      work_year  experience_level  employment_type
## Min.   : 0.0    Min.   :2020    Length:607    Length:607
## 1st Qu.:151.5  1st Qu.:2021    Class :character  Class :character
## Median :303.0  Median :2022    Mode  :character  Mode  :character
```

```

## Mean :303.0 Mean :2021
## 3rd Qu.:454.5 3rd Qu.:2022
## Max. :606.0 Max. :2022
## job_title salary salary_currency salary_in_usd
## Length:607 Min. : 4000 Length:607 Min. : 2859
## Class :character 1st Qu.: 70000 Class :character 1st Qu.: 62726
## Mode :character Median : 115000 Mode :character Median :101570
## Mean : 324000 Mean :112298
## 3rd Qu.: 165000 3rd Qu.:150000
## Max. :304000000 Max. :600000
## employee_residence remote_ratio company_location company_size
## Length:607 Min. : 0.00 Length:607 Length:607
## Class :character 1st Qu.: 50.00 Class :character Class :character
## Mode :character Median :100.00 Mode :character Mode :character
## Mean : 70.92
## 3rd Qu.:100.00
## Max. :100.00

which(is.na(dt$work_year))

## integer(0)

which(is.na(dt$experience_level)) # EN entry-level, MI mid-level, SE senior,
EX executive

## integer(0)

which(is.na(dt$employment_type)) # PT part-time FT full-time CT contract FL
freelance

## integer(0)

which(is.na(dt$job_title))

## integer(0)

which(is.na(dt$salary))

## integer(0)

which(is.na(dt$salary_currency))

## integer(0)

which(is.na(dt$salary_in_usd))

## integer(0)

which(is.na(dt$employee_residence))

## integer(0)

which(is.na(dt$remote_ratio))

```

```
## integer(0)
which(is.na(dt$company_location))
## integer(0)
which(is.na(dt$company_size)) # S: <50, M: 50<x<250, L: 250+
## integer(0)
# it looks like the data is clean already! Thank goodness
```

## Methods

The specific models and tests we are using in this week's assignment are boxplots of the dependent variable (salary\_in\_usd) vs. the categorical variable (experience\_level), fitting a regression model to the two variables, fitting the model to the two variables using ANOVA, and finally performing a post-hoc analysis using the Tukey HSD analysis. The significance of the results will be discussed afterward.

```
# run tests

# Plot the dependent variable vs the categorical variables (should be a
boxplot)
# in this case, the dependent variable is salary_in_usd and i will choose the
categorical
# variable of experience_level
par(mfrow = c(1,1))
#specify logical order for box plots, per here: https://r-graph-gallery.com/9-ordered-boxplot.html
dt$experience_level <- factor(dt$experience_level , levels=c("EN", "MI",
"SE", "EX"))
boxplot(salary_in_usd ~ experience_level, data = dt)
```

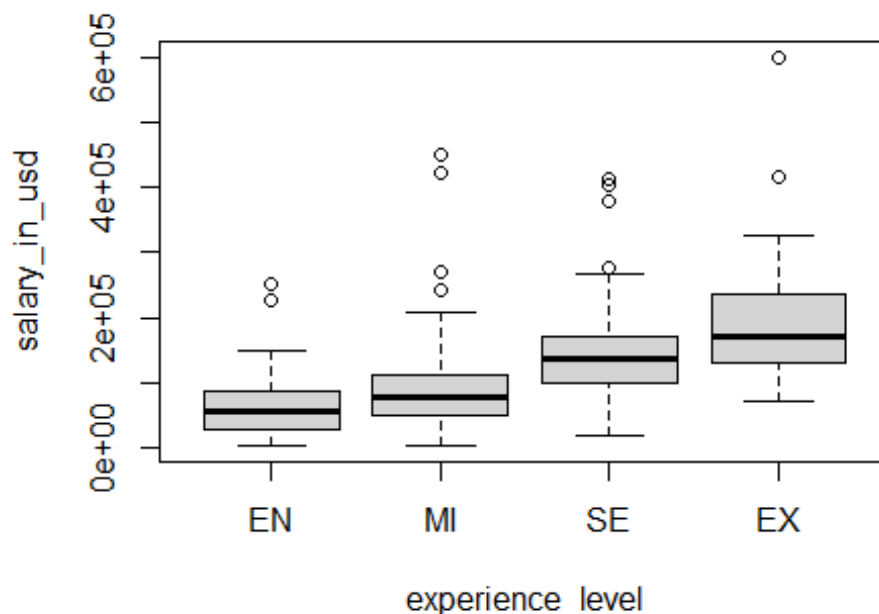


Figure 1: dependent variable (salary\_in\_usd) vs. categorical variable (experience\_level)

```
# Fit the dependent variable to the categorical variables using ANOVA
# First I will just fit a regression model to the two variables
fit <- lm(salary_in_usd ~ experience_level, data = dt)
summary(fit)
```

```
##
## Call:
## lm(formula = salary_in_usd ~ experience_level, data = dt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -129651  -39592   -7996    27930   400608
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      61643      6596   9.346  < 2e-16 ***
## experience_levelMI  26353      7841   3.361 0.000826 ***
## experience_levelSE  76974      7561  10.180  < 2e-16 ***
## experience_levelEX 137749     13811   9.974  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61870 on 603 degrees of freedom
## Multiple R-squared:  0.2434, Adjusted R-squared:  0.2397
## F-statistic: 64.68 on 3 and 603 DF,  p-value: < 2.2e-16
```

```

anova(fit)

## Analysis of Variance Table
##
## Response: salary_in_usd
##              Df      Sum Sq   Mean Sq F value    Pr(>F)
## experience_level  3 7.4277e+11 2.4759e+11  64.675 < 2.2e-16 ***
## Residuals       603 2.3084e+12 3.8282e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

par(mfrow=c(2,2))
plot(fit)

```

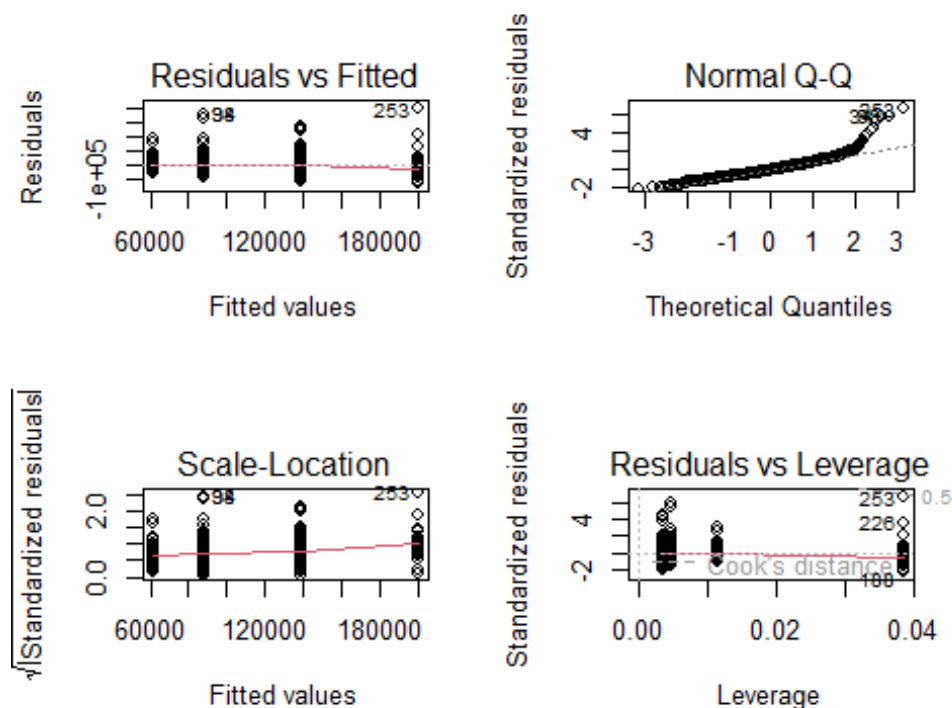


Figure 2: plot of the model fitted to the data

```

# Now I will perform a fit using ANOVA
afit <- aov(salary_in_usd ~ experience_level, data = dt)

```

## Results

In this section of the assignment, we will show the summary of the ANOVA fit, view the coefficients and plot the data, and perform the Tukey HSD post-hoc analysis discussed earlier in the assignment. The results of the post-hoc analysis were interesting. It showed that all of the pairwise comparisons created during the Tukey HSD analysis had a significant difference. This was strange as I wouldn't expect all of the comparisons to have

such low p-values. I even had to add digits to the Tukey HSD results so they wouldn't be displayed/interpreted as zero.

```
# report results
```

```
# View the ANOVA summary
```

```
summary(afit)
```

```
##              Df      Sum Sq   Mean Sq F value Pr(>F)
## experience_level  3 7.428e+11 2.476e+11   64.68 <2e-16 ***
## Residuals      603 2.308e+12 3.828e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# View the coefficients of the ANOVA fit
```

```
coefficients(afit)
```

```
##      (Intercept) experience_levelMI experience_levelSE
## experience_levelEX
##      61643.32      26352.74      76973.97
137748.72
```

```
# Change the plot window to a 2x2
```

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

```
# Plot the residuals
```

```
plot(afit)
```

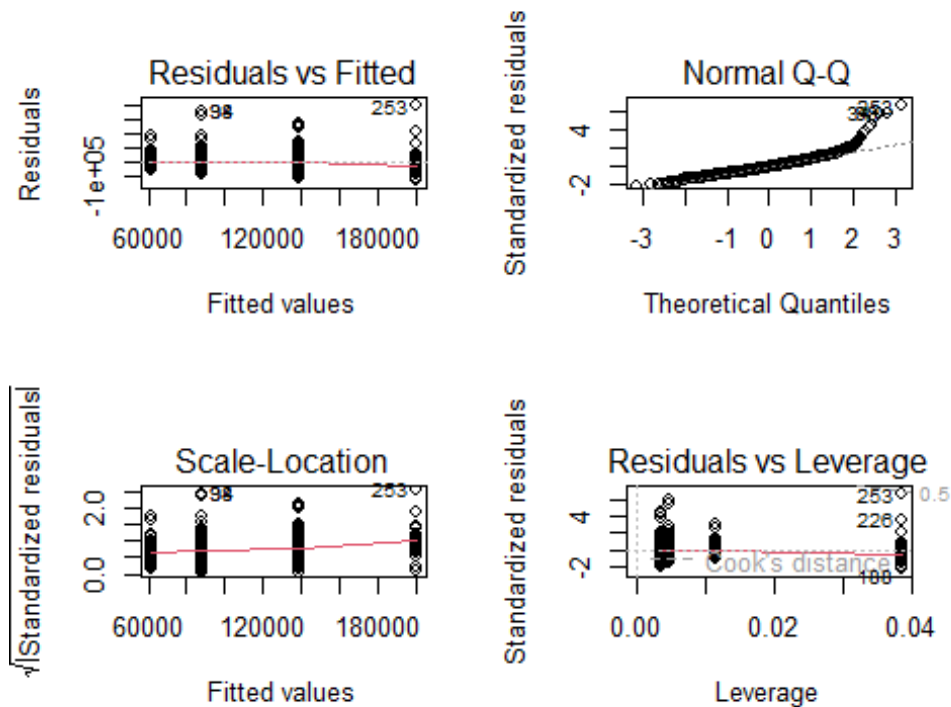


Figure 3: Plot of the ANOVA fit, quite similar to Figure 2

```
# Perform the post hoc analysis that you were assigned.

# I'm choosing to perform a TukeyHSD pairwise comparison
tfit <- TukeyHSD(afit, conf.level = 0.95) # TukeyHSD pairwise comparison
str(tfit)

## List of 1
## $ experience_level: num [1:6, 1:4] 26353 76974 137749 50621 111396 ...
## ..- attr(*, "dimnames")=List of 2
## .. ..$ : chr [1:6] "MI-EN" "SE-EN" "EX-EN" "SE-MI" ...
## .. ..$ : chr [1:4] "diff" "lwr" "upr" "p adj"
## - attr(*, "class")= chr [1:2] "TukeyHSD" "multicomp"
## - attr(*, "orig.call")= language aov(formula = salary_in_usd ~
experience_level, data = dt)
## - attr(*, "conf.level")= num 0.95
## - attr(*, "ordered")= logi FALSE

print(tfit,digits=15)

## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = salary_in_usd ~ experience_level, data = dt)
##
## $experience_level
##              diff              lwr              upr              p
adj
## MI-EN 26352.7381562097 6153.73307224441 46551.7432401750
0.004573630183886
## SE-EN 76973.9746753246 57494.31226682142 96453.6370838278
0.000000000427262
## EX-EN 137748.7202797205 102169.02009851398 173328.4204609270
0.000000000427263
## SE-MI 50621.2365191149 36129.09685623105 65113.3761819988
0.000000000427262
## EX-MI 111395.9821235108 78282.84380284080 144509.1204441809
0.000000000427262
## EX-SE 60774.7456043959 28095.43455579783 93454.0566529940
0.000012421852371

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



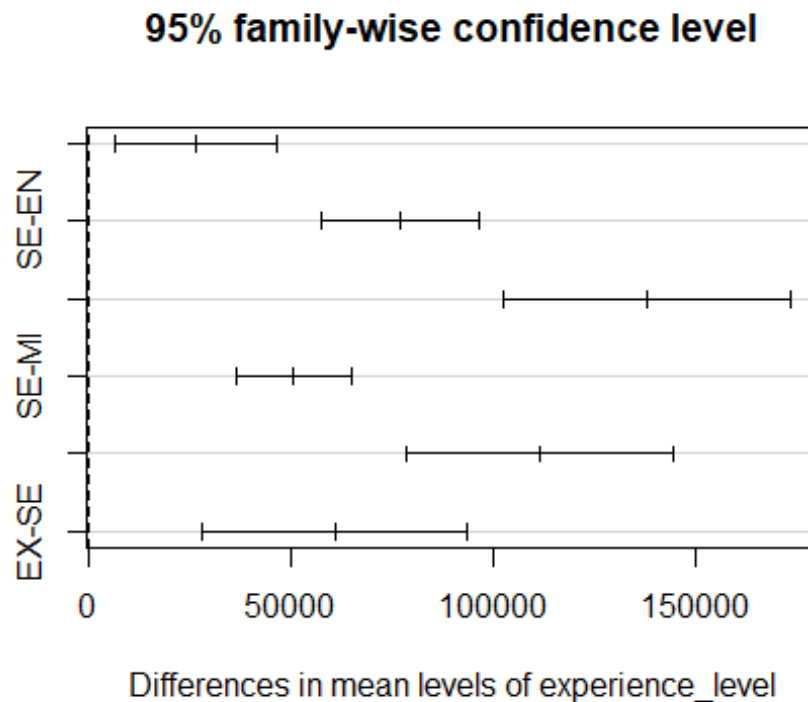


Figure 4: Tukey HSD pairwise comparisons, plotted

```
# Which post hoc analysis did you perform and which variables(s) have means
that are significantly different?
# I performed a TukeyHSD pairwise comparison post hoc analysis. This analysis
showed that
# it seems all p adj values are under p=0.05. this is interesting, this
implies that
# all differences between means are significant. I tried using difference
confidence levels
# of 0.90, 0.95, 0.97, and 0.99 and received the same p values using all of
these.
# What could be causing these universally low p values?
```

```
# source: https://stats.stackexchange.com/questions/253588/interpreting-tukeyhsd-output-in-r
```

### Conclusion

In conclusion, this week's assignment told us that there were significant differences in the salaries earned by each level of experience (Entry, Mid-Level, Senior, Executive). The null hypothesis could be rejected and the alternate hypothesis was accepted. However, I believe further analysis is warranted as the p-values of the Tukey HSD analysis were all exceptionally low. This seems a bit strange and should be checked and possibly corrected if

an error was made. Future analyses should definitely be made, iterating using insights gained from the strange p-values received in the first round of analyses. Thank you!