Project 4 Final Report

Ben Goldstone

5/3/2023

Data Set

```
library(readr)
datascience_salaries <- read_csv("~/CSVs/ds_salaries.csv")
```

```
## Rows: 3755 Columns: 11
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): experience_level, employment_type, job_title, salary_currency, empl...
## dbl (4): work_year, salary, salary_in_usd, remote_ratio
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(datascience_salaries)
```

```
## # A tibble: 6 x 11
##   work_year experience_level employment_type job_title    salary salary_currency
##       <dbl> <chr>            <chr>           <chr>         <dbl> <chr>
## 1      2023 SE               FT              Principal D~  80000 EUR
## 2      2023 MI               CT              ML Engineer   30000 USD
## 3      2023 MI               CT              ML Engineer   25500 USD
## 4      2023 SE               FT              Data Scient~ 175000 USD
## 5      2023 SE               FT              Data Scient~ 120000 USD
## 6      2023 SE               FT              Applied Sci~ 222200 USD
## # ... with 5 more variables: salary_in_usd <dbl>, employee_residence <chr>,
## #   remote_ratio <dbl>, company_location <chr>, company_size <chr>
```

Columns

Data Science Job Salaries Dataset contains 11 columns, each are:

work_year: The year the salary was paid.

experience_level: The experience level in the job during the year. (SE:Senior, EN:Entry level, EX:Executive level, MI:Mid/Intermediate level)

employment_type: The type of employment for the role.

job_title: The role worked in during the year.

salary: The total gross salary amount paid.

salary_currency: The currency of the salary.

salary_in_usd: The salary in USD.

employee_residence: Employee's primary country of residence in during the work year as an ISO 3166 country code.

remote_ratio: The overall amount of work done remotely.

company_location: The country of the employer's main office or contracting branch.

company_size: The median number of people that worked for the company during the year.

Data Source

This data was sourced from AI-jobs.net. AI-Jobs.net seems to connect directly to employers for posting information on jobs. This dataset is from Kaggle datasets.

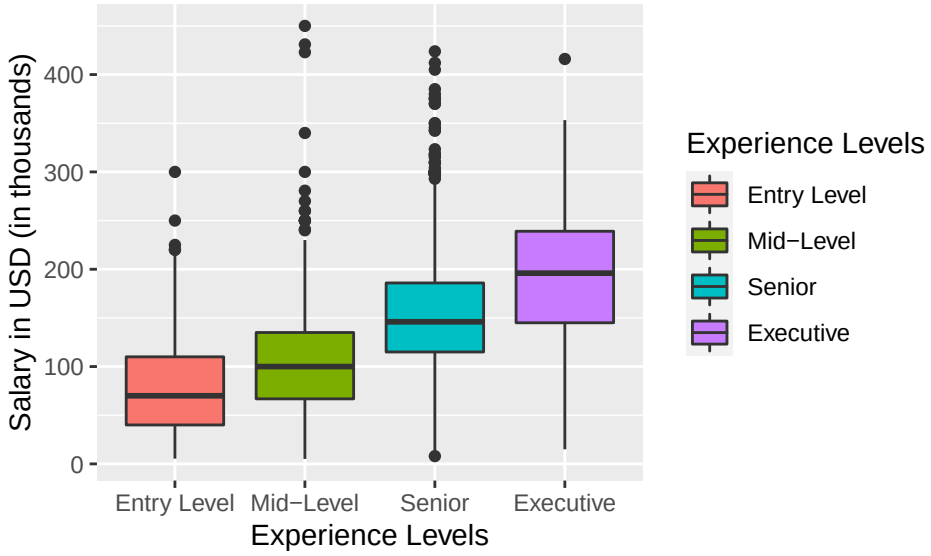
The data does not seem to have a direct source for it's data besides specifying that they "Provide a direct link between candidates and employers."

Experience Level and Salary

```
# convert to thousands
datascience_salaries$salary_in_usd_in_thousands <- datascience_salaries$salary_in_usd/1000

# order and relabel experience_level
datascience_salaries$experience_level_ordered <- factor(datascience_salaries$experience_level, levels=c

# boxplot with axis and legend labels
gf_boxplot(salary_in_usd_in_thousands~experience_level_ordered, fill=~experience_level_ordered, data=na
```




```
favstats(salary_in_usd_in_thousands~experience_level_ordered, data=datascience_salaries)
```

```
##   experience_level_ordered    min      Q1 median       Q3     max      mean
## 1              Entry Level  5.409  40.000     70 110.0093 300.000  78.54628
## 2                Mid-Level  5.132  66.837    100 135.0000 450.000 104.52594
## 3                   Senior  8.000 115.000    146 185.9000 423.834 153.05107
## 4                Executive 15.000 145.000    196 239.0000 416.000 194.93093
```

```
##         sd    n missing
## 1 52.22542  320       0
## 2 54.38769  805       0
## 3 56.89626 2516       0
## 4 70.66193  114       0
```

Hypotheses

H0: mu_entry = mu_mid = mu_senior = mu_executive

Ha: Not all means are the same

General Trends

In General as your experience level increases, your salary increases. Given the substantial number of outliers in this dataset, we should not run an ANOVA test, but for the sake of practice, we will run one anyway also the Standard-Deviation rule fits an ANOVA test.

```
experience_model=aov(salary_in_usd~experience_level, data=datascience_salaries)
summary(experience_model)
```

```
##                    Df    Sum Sq   Mean Sq F value Pr(>F)
## experience_level    3 2.972e+12 9.906e+11   310.8 <2e-16 ***
## Residuals        3751 1.195e+13 3.187e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analysis

Since the P-Value is very small, we can conclude that salaries are greater when you have more experience. In addition, we can run a Tukey Comparisons Test to see how different the salaries are based on the position.

Tukey Comparisons Test

```
TukeyHSD(experience_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = salary_in_usd ~ experience_level, data = datascience_salaries)
##
## $experience_level
##            diff        lwr       upr p adj
## EX-EN 116384.65  100558.84 132210.45     0
## MI-EN  25979.65   16391.13  35568.18     0
## SE-EN  74504.79   65893.44  83116.14     0
## MI-EX -90404.99 -104924.63 -75885.35     0
## SE-EX -41879.86  -55773.58 -27986.14     0
## SE-MI  48525.13   42649.84  54400.43     0
```

Tukey Comparisons Test Analysis

Given the substantial number of outliers in this dataset, we cannot get a consistent adjusted p-value across the different experience levels. All groups are very significantly different given the p-value is so small leading us to reject the null Hypothesis that all salaries are the same across different experience levels.

Final Conclusions

The dataset is very unbalanced in terms of the number of collected entries per category, so it is difficult to tell how balanced the ANOVA test is and with more data and less outliers, we can probably see a different difference between each category.