# Project 4 Data Summaries

Ben Goldstone

4/28/2023

## 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.

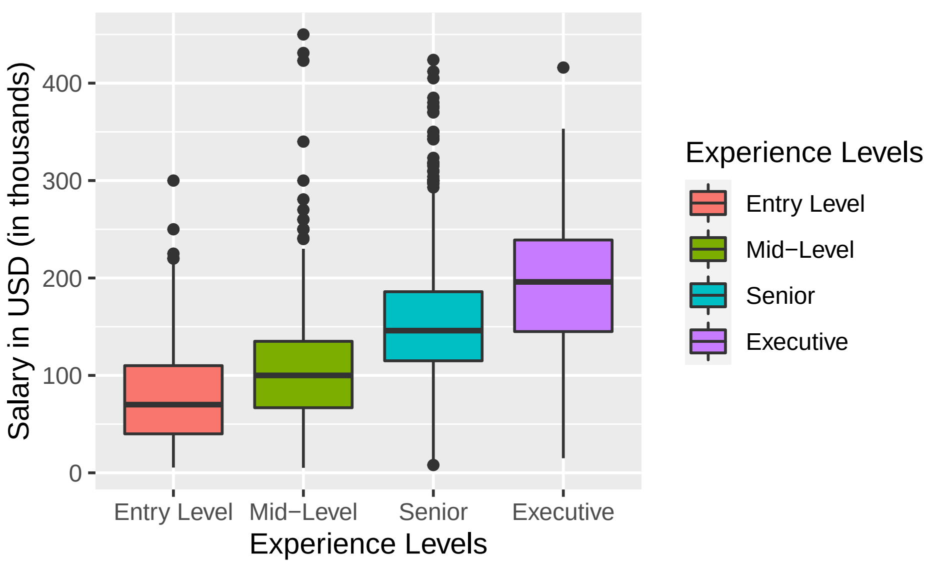
## Data Source

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
```

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
```

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.

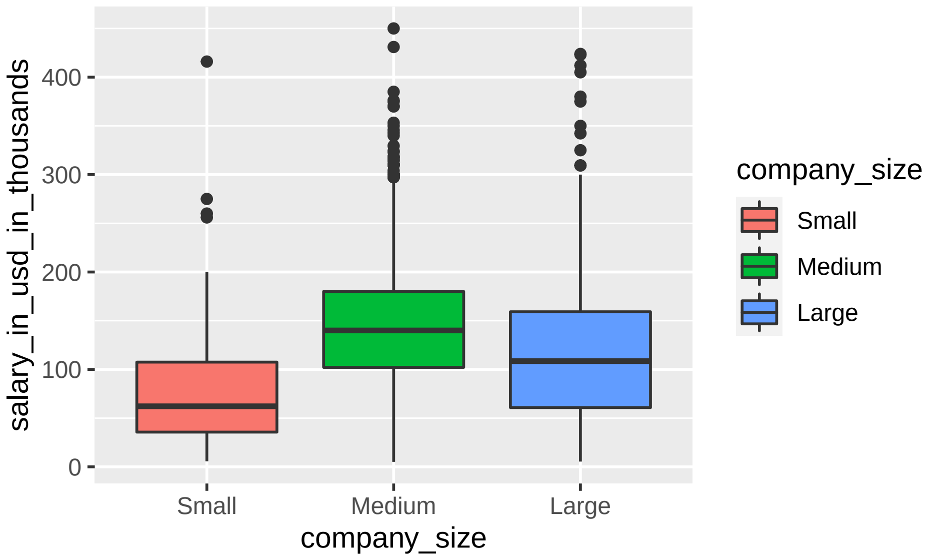
```
TukeyHSD(experience_model)
```

```
##   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
```

Given the substantial number of outliers in this dataset, we cannot get a consistent adjusted p-value across the different experience levels.

## Company Size and Salary

```
datascience_salaries$company_size <- factor(datascience_salaries$company_size, levels=c("S","M","L"), la
gf_boxplot(salary_in_usd_in_thousands~company_size,fill=~company_size,data=na.omit(datascience_salaries)
```




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

```
##   company_size   min        Q1  median       Q3     max      mean       sd    n
## 1        Small 5.679  35.66800  62.146 107.4818 416.000  78.22668 61.95514  148
## 2       Medium 5.132 102.10000 140.000 180.0000 450.000 143.13055 58.99281 3153
## 3        Large 5.409  60.83075 108.500 159.1750 423.834 118.30098 75.83239  454
##   missing
## 1       0
## 2       0
## 3       0
```

Data Scientists at a medium company seem to generally earn more than those at a small or large company.The Standard-Deviation rule seems to fit and be able to run an ANOVA Test.

```
size_model=aov(salary_in_usd_in_thousands~company_size,data=datascience_salaries)
summary(size_model)
```

```
##                Df   Sum Sq Mean Sq F value Pr(>F)    
## company_size    2   787259  393629   104.5 <2e-16 ***
## Residuals    3752 14138690    3768                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the P-value is very small, we can conclude that salaries are different across different sizes of companies.

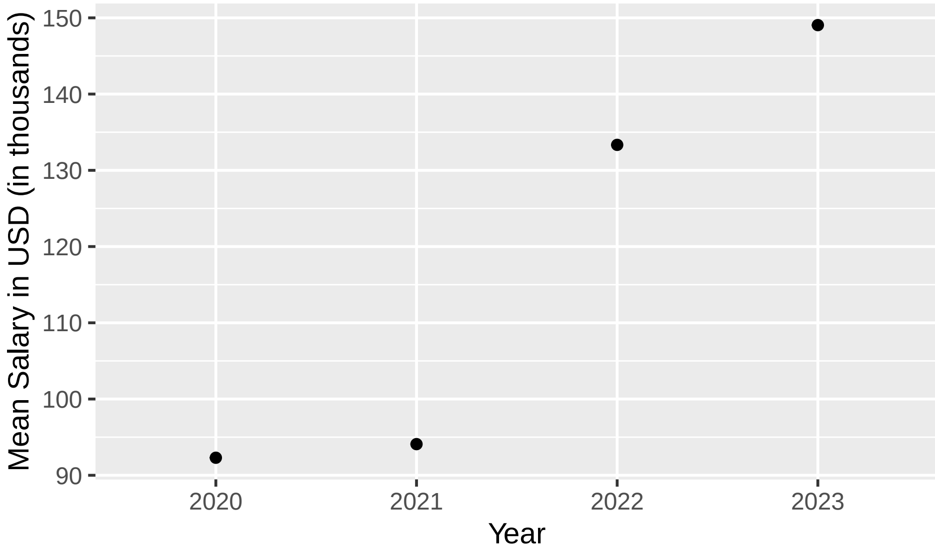
```
TukeyHSD(size_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = salary_in_usd_in_thousands ~ company_size, data = datascience_salaries)
## 
## $company_size
##                    diff       lwr       upr p adj
## Medium-Small   64.90387  52.79850  77.00923     0
## Large-Small    40.07430  26.45084  53.69776     0
## Large-Medium  -24.82957 -32.05446 -17.60468     0
```

Given the dataset, it seems the biggest difference in salary is between Medium and Small companies

## Year and Salary

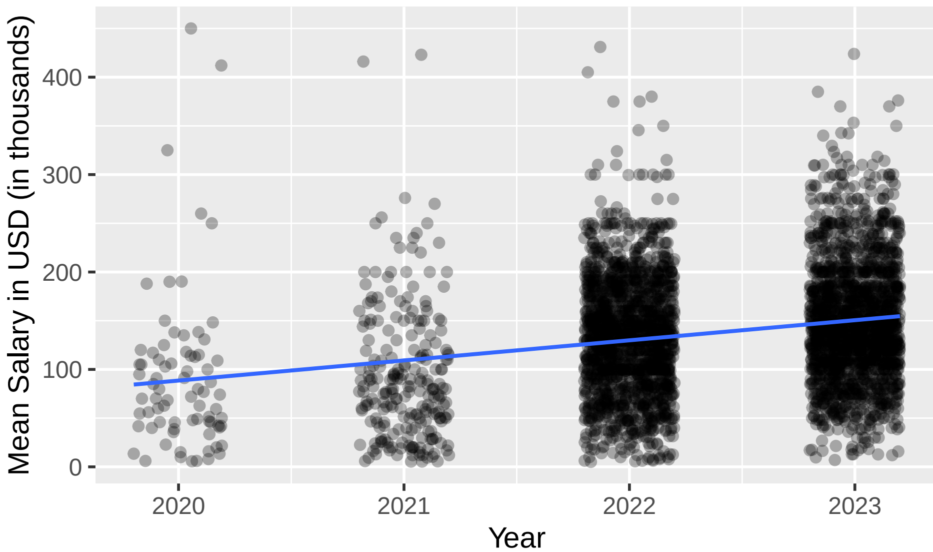
```
# gets means and salaries
means <- favstats(salary_in_usd_in_thousands~work_year, data=datascience_salaries)$mean
years <- favstats(salary_in_usd_in_thousands~work_year, data=datascience_salaries)$work_year

#puts years into categorical variables
datascience_salaries$years_cat = as.character(datascience_salaries$work_year)

gf_point(means~years, data=data.frame(years,means))%>%
  gf_lm() + xlab("Year") + ylab("Mean Salary in USD (in thousands)")
```



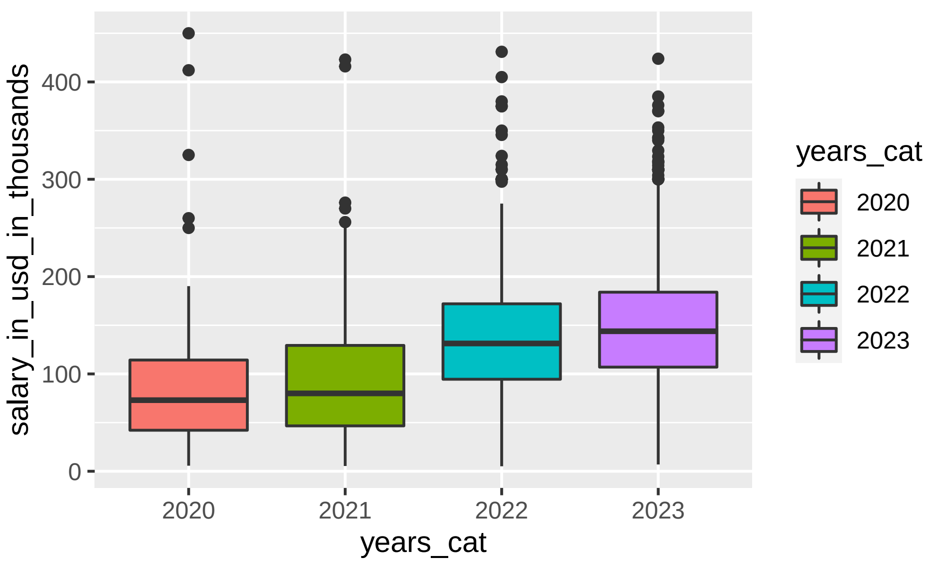

```
favstats(salary_in_usd_in_thousands~work_year,data=datascience_salaries)
```

```
##   work_year   min        Q1  median       Q3     max      mean       sd    n
## 1      2020 5.707  42.14775  73.065 114.2852 450.000  92.30263 82.37005   76
## 2      2021 5.409  46.65000  80.000 129.3053 423.000  94.08721 68.60047  230
## 3      2022 5.132  94.54500 131.300 172.0500 430.967 133.33862 58.94716 1664
## 4      2023 7.000 107.00000 143.860 184.0000 423.834 149.04554 61.30771 1785
##   missing
## 1       0
## 2       0
## 3       0
## 4       0
```

```
gf_point(salary_in_usd_in_thousands~jitter(work_year), data=datascience_salaries, alpha=0.3)%>%
  gf_lm() + xlab("Year") + ylab("Mean Salary in USD (in thousands)")
```




```
gf_boxplot(salary_in_usd_in_thousands~years_cat, data=datascience_salaries, fill=~years_cat)
```




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

```
##   work_year   min        Q1  median       Q3     max      mean       sd    n
## 1      2020 5.707  42.14775  73.065 114.2852 450.000  92.30263 82.37005   76
## 2      2021 5.409  46.65000  80.000 129.3053 423.000  94.08721 68.60047  230
## 3      2022 5.132  94.54500 131.300 172.0500 430.967 133.33862 58.94716 1664
## 4      2023 7.000 107.00000 143.860 184.0000 423.834 149.04554 61.30771 1785
##   missing
## 1       0
## 2       0
## 3       0
## 4       0
```

Given the data here, it looks like salaries increase each year by around 5-10 thousand dollars.

```
cor(salary_in_usd_in_thousands~work_year, data=datascience_salaries)
```

```
## [1] 0.22829
```

The Correlation between average salary and year is not very strong.

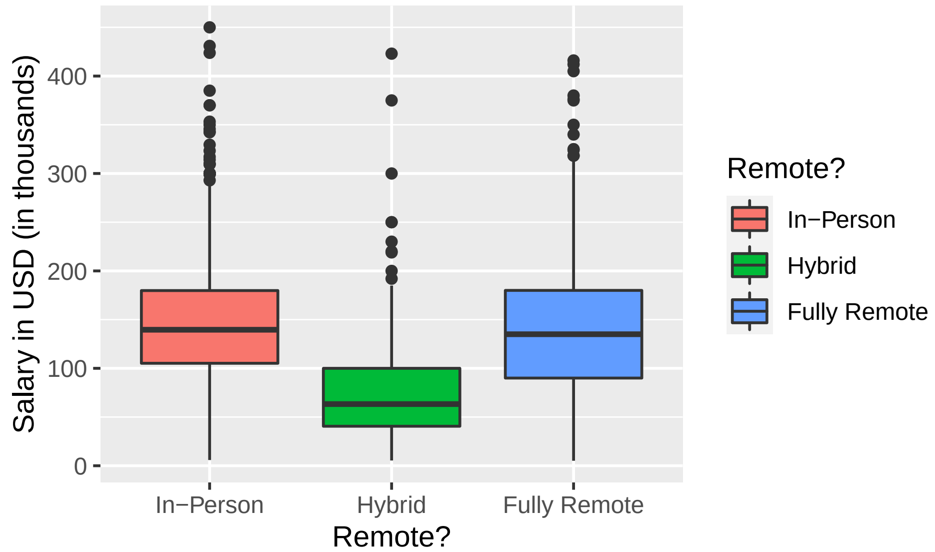
```
yearModel=lm(salary_in_usd_in_thousands~work_year, data=datascience_salaries)
summary(yearModel)
```

```
##
## Call:
## lm(formula = salary_in_usd_in_thousands ~ work_year, data = datascience_salaries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -143.61  -42.61   -3.82   35.29  361.85
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -41965.376   2930.987  -14.32   <2e-16 ***
## work_year       20.819      1.449   14.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.4 on 3753 degrees of freedom
## Multiple R-squared:  0.05212,    Adjusted R-squared:  0.05186
## F-statistic: 206.3 on 1 and 3753 DF,  p-value: < 2.2e-16
```

A linear Model only accounts for 5% of the data and is not a good fit. I also think if we had more data, there would be more of a trend to see. In addition the pandemic could've brought down salaries.

## Remote Ratio and Salary

```
# makes remote ratio categorical
datascience_salaries$remote_ratio_cat <- cut(datascience_salaries$remote_ratio,breaks=c(-100,0,50,100),l
gf_boxplot(salary_in_usd_in_thousands~remote_ratio_cat, fill=~remote_ratio_cat, data=datascience_salari
```




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

```
##   remote_ratio_cat   min     Q1  median     Q3 max      mean       sd    n
## 1        In-Person 5.882 105.20 139.600 179.82 450 144.31620 59.79997 1923
```

```
## 2           Hybrid 5.409  40.57  63.312 100.00 423  78.40069 61.32613  189
## 3     Fully Remote 5.132  90.00 135.000 180.00 416 136.48145 63.37351 1643
##   missing
## 1       0
## 2       0
## 3       0
```

People who work in-person seem to earn more than those who are hybrid. The Standard-Deviation rule seems to fit and be able to run an ANOVA Test.

```
remote_model=aov(salary_in_usd_in_thousands~remote_ratio_cat, data=datascience_salaries)
summary(remote_model)
```

```
##                    Df   Sum Sq Mean Sq F value Pr(>F)    
## remote_ratio_cat    2   751155  375578   99.41 <2e-16 ***
## Residuals        3752 14174793    3778                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There seems to be a difference between the average salary of those who are remote, hybrid, and in-person.

```
TukeyHSD(remote_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = salary_in_usd_in_thousands ~ remote_ratio_cat, data = datascience_salaries)
## 
## $remote_ratio_cat
##                               diff       lwr        upr     p adj
## Hybrid-In-Person        -65.915514 -76.90123 -54.929802 0.0000000
## Fully Remote-In-Person   -7.834749 -12.67630  -2.993203 0.0004419
## Fully Remote-Hybrid      58.080765  47.01160  69.149931 0.0000000
```

People who are In-Person seem to earn on average $66,000 more than those who are Hybrid.