ST 511 Project

Pavan Sai Nallagoni

Summer 2022

Introduction:

The dataset **Data Science Job Salaries** (2020 – 2022) by Ruchi Bhatia (owner) is extracted from Kaggle website. The dataset contains twelve columns namely work year, experience level, employment type, job title, salary currency, salary in USD, employee residence, remote ratio, company location, company size, salary, and id. In this report, the researcher wants to analyze this data particularly on the Company size (S, M, L) and Salary in USD to test whether the salary is same across the company sizes. Company size (S, M, L) is defined by the average number of people that worked for the company during the year: S less than 50 employees (small) M 50 to 250 employees (medium) L more than 250 employees (large).

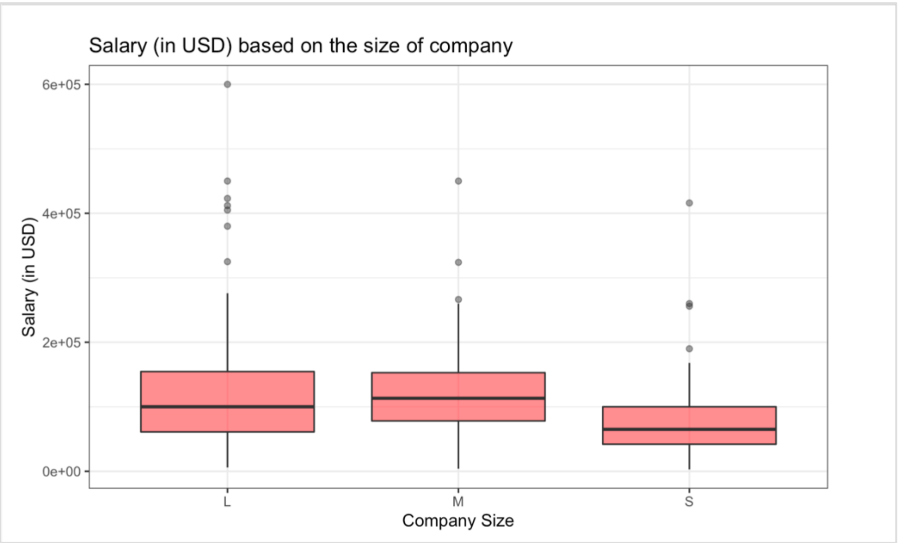
Scientific questions of interest:

In this study, the researcher wanted to know which, if any, of the company sizes resulted in higher salaries in USD such that a question was raised:

Q: Are the population means of salaries (USD) of each company size the same or different?

Population of Interest	:	Salaries in USD
Variable of Interest	:	Company Sizes (S, M, L)
Parameter	:	Mean

Exploratory plot:



The above box plot clearly indicates that the population mean salaries of all the different company sizes are not the same since there's a shift in the centers.

Statistical Method:

In this study design, the researcher wants to perform one-way analysis of variance (one-way ANOVA).

One-way analysis of variance (one-way ANOVA):

Assumptions:

The selected data is assumed to have the following assumptions:

- Independence within groups
- Independence between groups
- Normality of populations (or large enough sample size that central limit theorem approximation is good)
- Equal variances in all populations

The researcher used one-way analysis of variance (ANOVA test) since it involves a quantitative comparison of variation between the groups and variation within the groups such that it is possible to know whether the population mean salaries of all the different company sizes are same or not.

Hypothesis:

$Ho: \mu(S) = \mu(M) = \mu(L)$

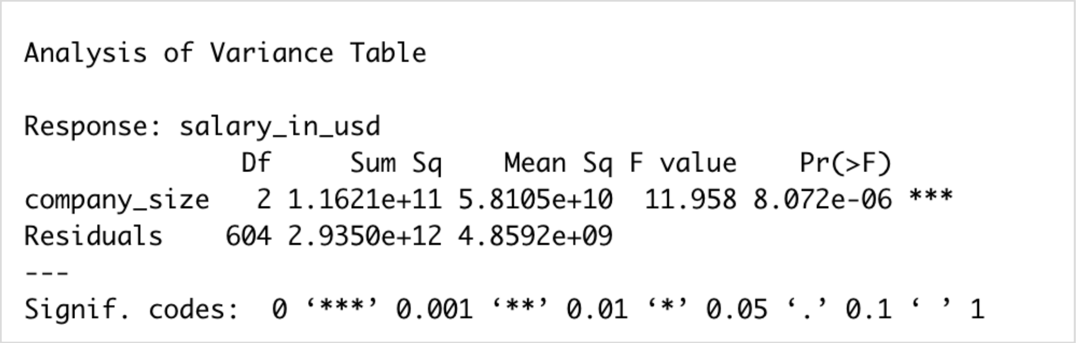
Null Hypothesis: The population mean salaries of all the different company sizes are same.

$H_A: \mu(S) \neq \mu(M) \mid \mu(S) \neq \mu(L) \mid \mu(M) \neq \mu(L)$

Alternative Hypothesis: At least two population mean salaries of all the different company sizes are different.

Results:

The following table provides the results from ANOVA test.

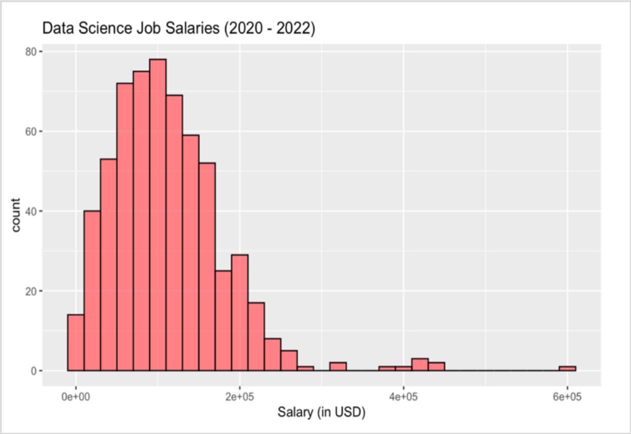


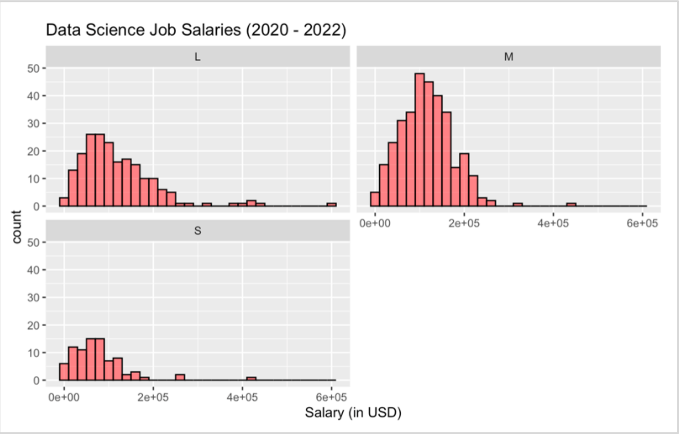
```
Analysis of Variance Table

Response: salary_in_usd
              Df     Sum Sq    Mean Sq F value    Pr(>F)
company_size   2 1.1621e+11 5.8105e+10  11.958 8.072e-06 ***
Residuals    604 2.9350e+12 4.8592e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the ANOVA table, the value of the test statistic is $F - statistic \ = \ 11.958$ and the $p - value \ = \ 8.072e - 06$. Therefore, the data is statistically significant for the company sizes since the $F - statistic$ is a greater value and $p - value$ is a very small value.

Assesment:





From the above histogram, it is clear that the population means of salaries (in USD) for different company sizes is normally distributed. Hence, the assumption of normality of populations is validated to perform the one-way ANOVA analysis. In the adjacent histogram, the variances seem to be the same with an equal spread. Finally, the people in the groups are not connected, and the groups are made up of different people. Therefore, the selected data satisfies all the assumptions of One-way analysis of variance (one-way ANOVA)

Conclusion:

In conclusion, the researcher rejects the null hypothesis $Ho: \mu(S) \ = \ \mu(M) \ = \ \mu(L)$ since the obtained $p - value \ = \ 8.072e - 06$ is much lesser than the significance level $\alpha \ = \ 0.05$, in favor of alternative hypothesis.

So basically $p - value$ is close to $\boldsymbol{0}$ and provides strong evidence that to reject the null hypothesis. Hence, the researcher concludes that at least two population mean salaries of the different company sizes are different.

In the future work, the researcher wants to know whether the population means of salaries (USD) of the smaller company size is different from that of the rest of the company sizes. To do this, the researcher suggests performing Kruskal-Wallis Test to test whether the centers of all the groups are the same or not $Ho: median(S) \ = \ median(M) \ = \ median(L)$.

---------------- END ----------------