Franco Bueno Mattera
Final Term Project
Class: MET CS544

Project Description Summary:

The main objective of this project is to analyze a dataset that contains information about different data analyst positions around the United States. There are three main questions that this project intends to answer:

1. As of July 2020, what was the general distribution of the variable salary for data analysts' jobs?
2. As of July 2020, what was the most common type of company that offers data analyst positions across the United States?
3. As of July 2020, what was the variability of salaries between the cities of New York, Chicago, Los Angeles, Charlotte (NC), San Francisco and Austin (TX)?

To answer question 3, three different approaches using four different sampling methods were used:

First, to demonstrate the applicability of the Central Limit Theorem, multiple samples were drawn from each of the variables of interest and their sampling means used to answer the proposed questions. In order for this theory to be applied, the samples drawn from the population need to be representative, meaning that every unit in the population needs to have the same chance of being selected(James Mclave and Terry Sincich 2013a).To comply with this, the population of this study was assumed to be the population of interest.

Second, inferential parametric methods were used to answer question 3, and third, non-parametric methods were performed to assess the same.

Finally, the results were compared to the whole dataset.

The sampling methods used were simple random sampling, systematic random sampling, systematic random sample with unequal probabilities and stratified random sampling.

Dataset Description:

The dataset contains information about different data analyst job positions across the United States. It was acquired from Kaggle Datasets (Original data obtained from Glassdoor) and it was created in July 2020. Initially, it contained a total of fifteen columns and 2253 rows. After selecting the columns and factors of interest and eliminating missing values (presented as -1), the cleaned dataset contains 531 rows and four columns. One of the columns of interest (Salary) was transformed from a character variable into two numeric variables (Maximum Salary and Minimum Salary).

Salary.Estimate	Location	Type.of.ownership
37K–66K (Glassdoor est.)	New York, NY	Nonprofit Organization
37K–66K (Glassdoor est.)	New York, NY	Nonprofit Organization
37K–66K (Glassdoor est.)	New York, NY	Company - Private
37K–66K (Glassdoor est.)	New York, NY	Subsidiary or Business Segment
37K–66K (Glassdoor est.)	New York, NY	Company - Private
37K–66K (Glassdoor est.)	New York, NY	Company - Private

Figure 1: Dataset with selected variables of interest after eliminating missing values

min.salary	max.salary	Location	Type.of.ownership
37	66	New York, NY	Nonprofit Organization
37	66	New York, NY	Nonprofit Organization
37	66	New York, NY	Company - Private
37	66	New York, NY	Subsidiary or Business Segment
37	66	New York, NY	Company - Private
37	66	New York, NY	Company - Private

Figure 2: Dataset with selected variables of interest after eliminating missing values and creating columns "min.salary" and "max.salary".

Data Analysis:

Question 1:

The general distribution of the variables Minimum Salary and maximum Salary are significantly right skewed, and they are not normally distributed (Figure 3). Table 1 shows the Shapiro-Wilk Normality test values confirming that the distributions are not normal, based on the null hypothesis that the populations are normally distributed (Long and Teetor 2019).

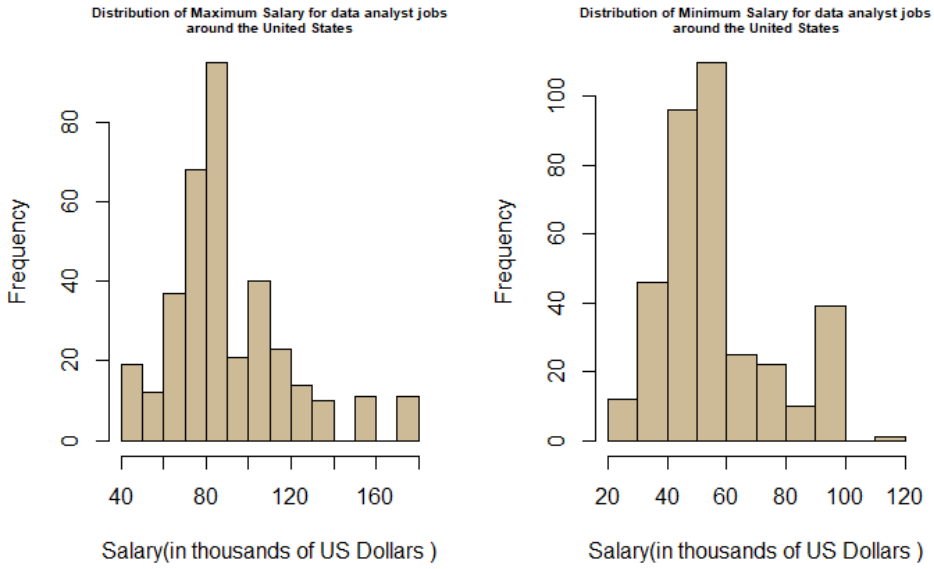

Figure 3: Distribution of minimum and maximum salaries around different cities in the United States.

Table 1: General summary of the variables Minimum Salary and Maximum Salary.

	Minimum Salary	Maximum Salary
Minimum	27.00	42.00
Quantile 25	42.00	75.00
Median	53.00	86.00
Mean	56.60	91.45
Quantile 75	65.00	104.00
Maximum	113.00	178.00

Standard Deviation	18.72	28.54
Shapiro-Wilk p-value	7.167e-14	2.75e-13

Question 2:

The analysis of the categorical variable "Type of Ownership" shows that for the data analyst jobs positions, private companies are the ones that have the greatest number of positions, while colleges and universities have the lowest.

Table 2: Percentage of data analyst positions in different types of companies.

College	Private	Public	Hospital	Non-Profit	Subsidiary
1.51%	59.7%	23.9%	1.88%	7.34%	5.65%

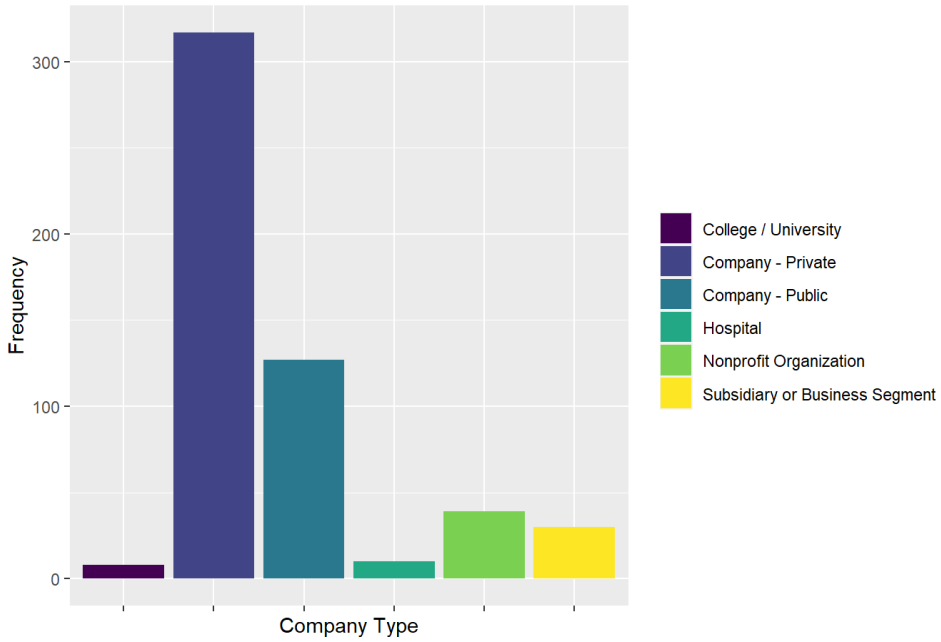

Figure 4: Bar plot Showing different frequencies of different data analyst job positions per type of company.

Question 3:

Looking at the whole dataset, for the variable 'Minimum Salary', it can be appreciated in Figure 5 that there are some evident differences in income between the cities. The highest income was for the city of San Francisco, followed by the city of Chicago. For the variable 'Maximum Salary' (Figure 5), the situation is the same but apart from San Francisco, it appears that there are no major differences in maximum salaries between the rest of the cities. More detailed pairwise comparisons are shown in Table 3.

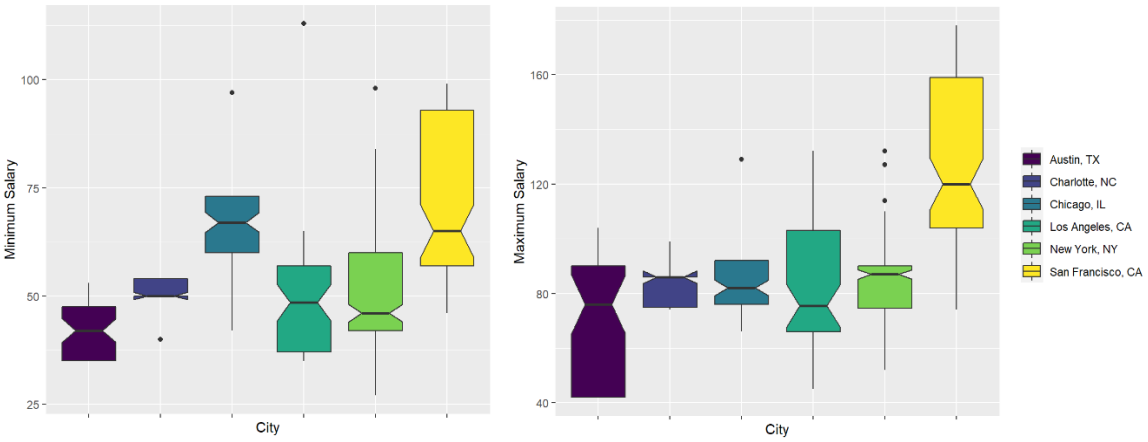

Figure 5. Boxplot showing Minimum and Maximum Salary (in thousands of US Dollars) for the six cities with the most data analyst positions. If the notches do not superpose, the medians are different (Chang 2019).

Table 3: Pairwise differences for the whole dataset based on the boxplot from Figure 5.

City Pair	Differences	
	Minimum Salary	Maximum Salary
Charlotte, NC-Austin, TX	TRUE	FALSE
Chicago, IL-Austin, TX	TRUE	FALSE
Los Angeles, CA-Austin, TX	FALSE	FALSE
New York, NY-Austin, TX	FALSE	FALSE
San Francisco, CA-Austin, TX	TRUE	TRUE
Chicago, IL-Charlotte, NC	TRUE	FALSE
Los Angeles, CA-Charlotte, NC	FALSE	FALSE
New York, NY-Charlotte, NC	FALSE	FALSE
San Francisco, CA-Charlotte, NC	TRUE	TRUE
Los Angeles, CA-Chicago, IL	TRUE	FALSE
New York, NY-Chicago, IL	TRUE	FALSE
San Francisco, CA-Chicago, IL	FALSE	TRUE
New York, NY-Los Angeles, CA	FALSE	FALSE
San Francisco, CA-Los Angeles, CA	TRUE	TRUE
San Francisco, CA-New York, NY	TRUE	TRUE

After taking one thousand samples with replacement of size one hundred for the variable salaries and comparing its sampling distribution of the sample means with the whole dataset, the differences were negligible (Table 4). This means that if given enough time and resources, that is the most accurate way of estimating the means. Also, the change in the distributions is clear when comparing the distributions of Minimum Salary per city and the sampling distribution of the sampling means of the same factors (Figure 6 and Figure 7). The situation is the same for the variable Maximum Salary.

Table 4: Population mean, sampling means and their respective differences between dataset means and sampling means(in thousands of dollars) per city.

	Minimum Salary			Maximum Salary		
	Population Mean	Sampling mean	Difference	Population Mean	Sampling mean	Difference
New York	52.6634	52.6355	0.0279	87.2624	87.2136	0.0488
Los Angeles	47.7679	47.723	0.0449	80.9107	80.8581	0.0526
San Francisco	72.4253	72.4201	0.0052	127.046	126.846	0.2
Austin	42.6078	42.5921	0.0157	72.4706	72.5148	-0.0442
Charlotte	50.8364	50.8471	-0.0107	84.7455	84.7793	-0.0338
Chicago	68.375	68.4073	-0.0323	87.4	87.3053	0.0947

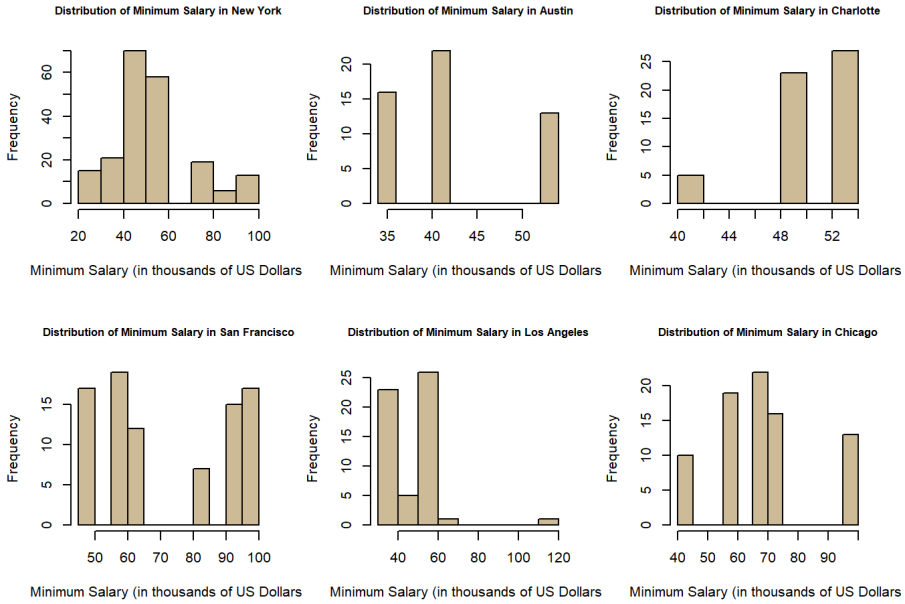

Figure 6: Distribution of Minimum Salary for the six most important cities considering the amount of data analyst job positions.

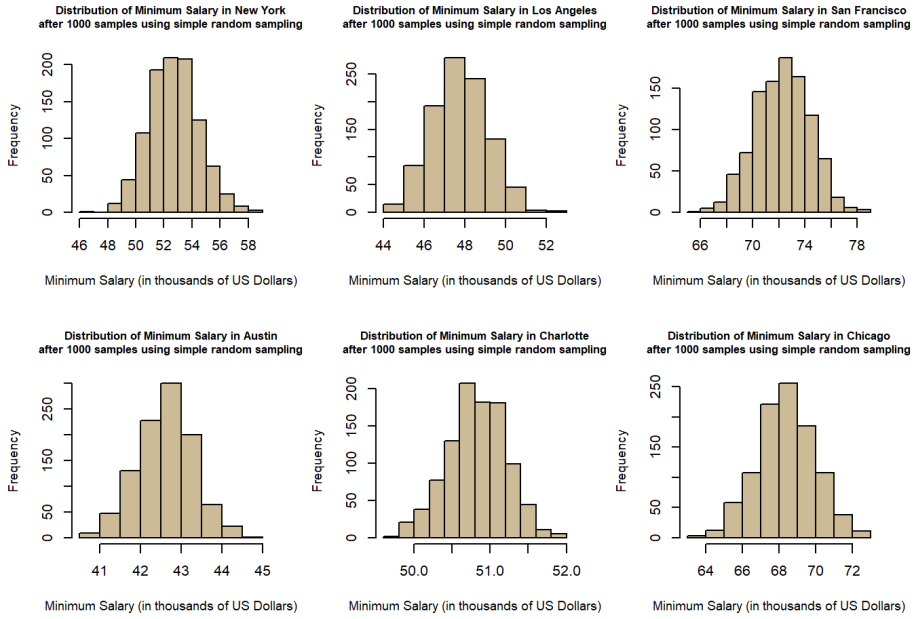

Figure 7: Distribution of the sampling distribution of the sampling mean of Minimum Salary for the six most important cities considering the amount of data analyst job positions.

Samples of size one hundred were drawn from the whole dataset. Four sampling methods were used: Simple Random Sampling, Systematic Random Sampling, Systematic Random Sampling with unequal Probabilities and Stratified Random Sampling. Parametric and non-parametric test were used to study the variability of the variables Minimum and Maximum Salary across the chosen cities of interest. Those results where compared with the variability of the whole dataset (Table 5).

Table 5: Success probabilities at estimating differences by using parametric and non-parametric post hoc tests (pairwise t-test with Bonferroni correction and multiple Wilcoxson rank-sum tests performed after ANOVA and Kruskal Wallis tests, respectively). Respective p-values for general parametric and non-parametric test are shown (for pairwise p-values, see section 2.2.1.3 from attached html code file)

	Parametric tests				Non-Parametric tests			
	Minimum Salary	p-value	Maximum Salary	p-value	Minimum Salary	p-value	Maximum Salary	p-value
Sampling Distribution Data	0.46	<2e-16	0.60	<2e-16	-	-	-	-
Simple Random Sampling	0.80	2.02e-05	1	5.36e-06	0.60	3.333e-05	0.867	0.001013
Systematic Random Sampling	0.733	1.26e-05	1	9e-07	0.60	4.173e-05	0.73	0.001206
Systematic w/unequal probs	0.60	0.000189	1	3.71e-08	0.60	5.539e-05	0.867	3.206e-05
Stratified Random Sampling	0.40	0.0277	0.67	0.95	0.40	0.02497	0.67%	0.8953
Simple Random Sampling n <30	0.40	0.148	0.73	0.0473	0.40	0.1713	0.67%	0.1742

As it can be appreciated in Table 5. Using the Sampling distribution of the sample means performed badly compared to the other sample methods. This is due in part that the sampling distribution of the sample mean has much less spread than the original data. This is because the standard deviation of the sampling distribution of the sampling mean is the standard deviation of the population divided by the squared root of the sample size(James Mclave and Terry Sincich 2013a). Having said that, the applicability of the Central Limit Theorem is appreciated when comparing the results of the test for a sample of size one hundred with a smaller sample of size of 28. The results for the smaller sample size performed poorly, in contrast, using a bigger sample sized yielded much better results. When using a big sample, the parametric tests performed even better than the non-parametric tests even though the normality assumptions (which are required when doing this parametric tests) are not met.

It is important to note that a pairwise z test would have been more appropriate in this situation for a large sample since the dataset is assumed to be the whole population, but such test is not available for R (t-test has the same formula than the z test, and its distribution is closely related to z, especially when the sample size is large). Still, the applicability of the central limit theorem is evident, and thanks to it, the fact that the sampling distribution of the differences in sample means follows a normal distribution centered around zero. In consequence, the pairwise t-test should be resistant to the departures of normality, just as is stated in the literature, as long as the sample is sufficiently large enough(James Mclave and Terry Sincich 2013b; Jerold H. Zar).

Conclusions and lessons learned:

In relation to question one, the distribution for minimum and maximum salary its consistent with the typical income distribution of salaries around the United States. Also, in this case, the median is a much better point estimate of salary for the whole dataset than the mean.

For data analyst positions as of July 2020, almost 60 percent belonged to private companies, followed by public companies with almost 24 percent. Most of the jobs offers should come from private companies, and (at least for the locations studied) it is possible that this rate is similar for the city of Boston which was not included in this study since data was not available.

While the Central Limit Theorem is applicable while using parametric tests (when the sample size is large), using its sampling distribution of sample means to estimate the variability of the population did not produce the expected results. The means of the sampling distributions are almost identical to the means of the whole dataset, but since its standard deviations are very small, any small deviation could be considered a significant difference.

Also, stratified sampling is not very good at estimating differences in the populations. This is since the sampling method follows a particular pattern which in consequence makes it easier to violate the assumption of randomness which is important for most tests. Stratified sampling should be done with care and with good domain knowledge since it is applicable when the population parameters are known to differ considerably (clear presence of subgroups in the population to be sampled). Systematic random sample yielded good results also, making it a good option when time and resources and scarce and when there is no pattern in the data(Andriy Blokin and Peter Westfall).

References

When Is It Better to Use Simple Random vs. Systematic Sampling? [Internet] [cited 2020 Nov 27,]. Available from: https://www.investopedia.com/ask/answers/071615/when-it-better-use-systematic-over-simple-random-sampling.asp.

Chang W. 2019. R graphics cookbook. Second edition, fourth release ed. Beijing, Boston, Farnham, Sebastopol, Tokyo: O'Reilly.

James Mclave, Terry Sincich. 2013a. Statistics. 413 p.

James Mclave, Terry Sincich. 2013b. Statistics. 285 p.

Jerold H. Zar. Biostatistical analysis. 127 p.

Data Analysts Jobs Dataset [Internet]. Available from: https://www.kaggle.com/andrewmvd/data-analyst-jobs .

Long JD, Teetor P. 2019. R cookbook. Second edition ed. Beijing: O'Reilly.

Pros and Cons of Stratified Random Sampling [Internet] [cited 2020 Nov 27,]. Available from: https://www.investopedia.com/ask/answers/041615/what-are-advantages-and-disadvantages-stratified-random-sampling.asp.

US Income Distribution [Internet]. Available from: https://www.census.gov/library/visualizations/2015/demo/distribution-of-household-income--2014.html.