**End of Course Assignment**

Kelly Glueck Tayong

N2181528

School of Business, Singapore University of Social Science

ANL201: Data Visualisation for Business

Monday, 6 March 2023

***Tutorial Group***

*TG02*

**Section A**

**Question 1**

**Table with Summary Statistics: Identifying and Justifying Data Type**

The data type assigned for the variables in this dataset is shown in the table below and the summary statistics were computed within an excel spreadsheet and can be validated using Tableau > Summary Tab and Data Analysis > Descriptive Statistics function in excel.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Data Type** | **Summary Statistics** | | **Number of Observations** |
| Name | Nominal | Mean | - | 231 |
| Median | - |
| Mode | NA |
| Maximum | 0 |
| Minimum | 0 |
| Location | Nominal | Mean | - | 231 |
| Median | - |
| Mode | New York |
| Maximum | 0 |
| Minimum | 0 |
| State | Nominal | Mean | - | 231 |
| Median | - |
| Mode | CA |
| Maximum | 0 |
| Minimum | 0 |
| Rank | Ordinal | Mean | 113,9892 | 231 |
| Median | 111 |
| Mode | 220 |
| Maximum | 220 |
| Minimum | 1 |
| Tuition and Fees | Ratio | Mean | $33,427 | 228 |
| Median | $31.744 |
| Mode | $26,334 |
| Maximum | $100,000 |
| Minimum | -$51,265 |
| Undergrad Enrolment | Ratio | Mean | 15,008.88 | 226 |
| Median | 12,928.50 |
| Mode | 6,883.00 |
| Maximum | 54,513.00 |
| Minimum | 1,001.00 |

Refer to Part A of the Appendix section for the summary statistics workings. Very little can be derived from nominal variables. Only the mode of “Location” and “State” can be computed.

Finally, these functions are repeated for other columns or variables and the final table generated is shown in Figure 1. These values can also be derived using the Data Analysis > Descriptive Statistics function in Excel. The results are shown in Figure 1.

Table

Description automatically generated

Figure 1. Excel Summary Statistics

**Data Type Justification**

***Name*** – the values in the “name” column were all the names of the universities. The data type assigned is nominal because it contains the labels of the university, in this case, their names. Furthermore, no ranks can be made, and no mathematical operations can be performed. These values are strings as identified by Tableau and strings are nominal variables.

Graphical user interface, application

Description automatically generated

Figure 2. Tableau "Name" Type

***Location*** – like “name”, the values in the “location” column label the locations where the universities are located. In Tableau, the type is assigned as a string and a geographic role of City.

Graphical user interface, application

Description automatically generated

Figure 3. Tableau "Location" Type

***State*** – like “location”, the values in the “state” column are also labels of states in which the universities are located. In Tableau, the type is assigned as a string and a geographic role of State/Province.

Graphical user interface, application

Description automatically generated

Figure 4. Tableau "State" Type

***Rank*** – ranks are ordinal variables because they are ranked in order. The difference between the values is usually unknown and there is no true zero point. Within this column, the ranks of the universities are sorted from 1 – 220. Only some statistical analysis can be performed such as median, mode and non-parametric tests.

Graphical user interface, application

Description automatically generated

Figure 5. Tableau "Rank" Type

***Tuition and fees*** – tuition and fees represent the costs of tuition and fees incurred to enrol into the university. These values are mainly numbers that are represented in the United States dollar currency. Hence, the data type assigned to them is Ratio. This is because the values in the column are continuous values that have a true zero point and have known differences that can be compared. For ratio variables, statistical analyses such as standard deviation, correlation, and regression can be performed.

Graphical user interface, application

Description automatically generated

Figure 6. Tableau "Tuition and Fees" Type

***Undergrad Enrollment*** – undergrad enrolment represents the number of undergraduates that are enrolled on universities. In this case, the data type should be Ratio. The undergrad enrolment has defined zero points and known differences can be compared, however, the variables should be discrete.

Graphical user interface, application

Description automatically generated

Figure 7. Tableau "Undergrad Enrollment" Type

**Question 2**

**Data Quality Issues and Treatment Methods**

***Inappropriate/Inconsistent Data Format for Rank and Undergrad Enrollment (Must be Discrete)***

Based on the given dataset, the data formatting for the Rank and Undergrad Enrollment values is not suitable because they are both set as continuous variables. As shown in Figure 8, some of the values within the variable “Rank” are displayed in decimals and all values “Undergrad Enroll” columns are displayed in two decimal places.

Table, Excel

Description automatically generated

Figure 8. Decimal Point in Rank and Undergrad Enrollment

Both variables are supposed to be discrete variables because there should be no numbers in between. Rankings should be represented in the whole number because it is ordinal data. Undergrad Enrollment should also be represented in the whole number because there is no half-body or half-a-person when students enrol on a university.

***Treatment Method (Number Format Standardization)***

The data quality issue can be treated by formatting and standardizing the number format for the values under both variables. By decreasing the decimal place, the data quality issue is resolved.

Graphical user interface, application

Description automatically generatedTable, Excel

Description automatically generated

Figure 9. Number Format Standardization

This can also be done in Tableau by changing the data type from Number (decimal) to Number (whole) for both variables.

Table

Description automatically generatedTable

Description automatically generated

Figure 10. Tableau Number Format Change

***Missing Rank Orders and Repeated Rankings***

Based on the given dataset, it is observed that some ranks are repeated, and the ranking orders are incomplete. There are missing ranks such as 6,7,9 in the dataset shown in Figure 11.

Graphical user interface, application, table, Excel

Description automatically generated

Figure 11. Missing and Unordered Ranks

For the dataset to be appropriate, the ranks should be ordered completely from 1 to 231 without repeated or missing orders.

***Treatment Method (Sorting and Autofill Function)***

The Rank column can be sorted naturally from the smallest to the largest using the Sort and Filter function as shown in Figure 12.

Graphical user interface, application

Description automatically generated

Figure 12. Sorting and Autofill Function

Thereon, the autofill function can be used to replace the repeated ranks with a natural order from 1 to 231. This can be done by highlighting the first three values and dragging the autofill function down to the last value or record. The fixed data is shown in Figure 13.

Graphical user interface, application, table, Excel

Description automatically generatedGraphical user interface, application, table, Excel

Description automatically generated

Figure 13. Autofill and Fixed Rank Numbers

***Extreme Observations and Outliers***

In the given dataset, there are some extreme observations or outliers that can be observed in the “Tuition and fee” column as shown by the Boxplot in Figure 14.

Chart

Description automatically generated

Figure 14. Outliers in Tuition and Fees

While in the “Undergraduate Enrollment Column”, there are two outliers.

A picture containing chart

Description automatically generated

Figure 15. Outliers in Undergrad Enrollment

***Treatment Method (Replacing Values with Column’s Mode)***

Replacing extreme values and outliers with mean values is not recommended because it is highly susceptible to outliers. Hence, a better option is to replace the extreme values and outliers with the mode. To do this, we can simply replace the extreme data value outlined by the boxplot with the mode of their respective columns. Which is $26,334 for Tuition and fees and 6883 for Undergrad Enrollment.

Graphical user interface, application, table, Excel

Description automatically generatedTable

Description automatically generated

Figure 16. Replacing Outliers with Mode

***Missing Observations***

There are three missing observations in the “Tuition and Fee” column and five missing observations in the “Undergrad Enrollment” column as shown in Figure 17. These missing values may give an undesirable result during the process of data visualization.

Graphical user interface, application

Description automatically generated

Application, table, Excel

Description automatically generated

Figure 17. Missing Observations in Tuition and Fees and Undergrad Enrollment

***Treatment Method (Filling Up Missing Values with Mean)***

Since the outliers are replaced, the mean of each column should change. These records are independent of each other; hence interpolation cannot be performed. Furthermore, the records with empty values should be kept making the analysis more meaningful. Hence, replacing the empty cells with the mean is the most appropriate treatment.

The new mean is $33,686 for Tuition and Fees and 14,612 for Undergrad Enrollment after replacing the outliers.

Table

Description automatically generated

Application, table, Excel

Description automatically generated

Figure 18. Filling Up Empty Observations with Mean

***Misspelt Location and State Names***

Twelve state codes couldn’t be identified and twelve locations (cities) that are misspelt. This data issue can be observed in Tableau as shown in Figure 19.

Graphical user interface, application

Description automatically generated

Figure 19. Misspelt Locations (City)

***Treatment Method***

This issue can be simply fixed by editing the location and assigning the correct state codes as shown in Figure 20. The cities can be verified in a search engine and referred from the university names. The longitude and latitude of some cities need to be inputted because some city names are not available in Tableau. The state codes can be fixed by setting the Country/Region to USA instead of the default setting None.

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

Figure 20. Fixing and Matching the Location

**Question 3**

**Developing and Interpreting Data Visualisation**

Underlying trends and patterns can be dug out using data visualisation. We can analyse how rank orders affect the amount of tuition and fees needed to be paid and how the university rankings affect the undergraduate enrolment numbers. Computation for the average tuition and fee can also help universities to understand students’ ability to pay for tuition and fees, thereon, universities can choose to adjust their tuition and fees according to the student’s demography. Understanding the trend of how tuition and fee affect the number of undergraduate enrolments can also be important.

Furthermore, we can make use of spatial maps to plot the locations and states where undergraduate enrolments are more than others to study the demography of university enrolments. With this, a dashboard can be developed to oversee the percentage of total enrolments in each location and let users make decisions on how to get more undergraduate enrolments in certain areas. Moreover, universities can make use of the dashboard to understand and compare the tuition and fees incurred by students studying in their university and use it to attract more upcoming undergraduates.

***Scatter Plot with Trend Line – Understanding the “Tuition and Fees” and “Undergrad Enrollment” Patterns/Trend by University Ranking***

Chart, scatter chart

Description automatically generated

Figure 21. Scatterplot with Trend Line

The scatter plot seen in Figure 21 is plotted to compare the relationship between the Tuition and Fees and Undergraduate Enrollments by the University Rank that is sorted in ascending order. The trend line feature is plotted to better understand the underlying patterns.

The trend line in both scatter plots is negatively correlated, meaning that the higher the rank is, the lower the Tuition and Fee and Undergraduate Enrollment numbers. Both have a negative relationship with each other, but the magnitude is different. For Tuition and Fees, since it has a greater slope and gradient value, the magnitude at which the Tuition and Fee decrease is higher when the rank increases by one. This change is lesser for Undergraduate Enrollment because the slope is steeper with a lesser gradient value.

***Line Chart with Trend Line – Understanding the Relationship between Tuition and Fees with Undergrad Enrollment***

Chart, bar chart

Description automatically generated

Figure 22. Line Chart with Trend Line

The line chart seen in Figure 22 is plotted to compare the relationship between Undergraduate Enrollment by Tuition and Fees. The trend line feature is added to study the pattern and correlations between the two variables.

The trend line in this line chart is sloping downwards, meaning that the relationship of both values is negatively correlated. This means that when the Tuition and Fees increase, there will be lesser Undergraduate Enrollments in the university.

This trend is useful to explain how Tuition and Fees affect the willingness of undergraduates to enroll and continue their further studies and can be used by universities to further analyse and adjust their Tuition and Fees independently to attract more potential undergraduates.

***Bar Chart – Using Parameters and Calculated Field to Group Universities with “Tuition and Fees” and “Undergrad Enrollment” into Two Sections (Above and Below Average) and Compare with Reference Line***

Chart, bar chart

Description automatically generated

Figure 23. Bar Chart with Reference Line

The Bar Chart seen in Figure 23 is plotted to compare the relationship between the Tuition and Fees and Undergraduate Enrollments by the University Rank that is sorted in ascending order. These two charts are plotted using the Parameter and Calculated Field function to study and compare the Tuition and Fees and Undergraduate Enrollment values with their average values. The Reference Line is used to segment the universities into two distinguished parts, one that is above average and the other that is below average.

The pattern that can be deduced from the Tuition and Fees by University Rank chart is that higher-ranked universities are charging more Tuition and Fees compared to other universities and the average. While lower-ranked universities charge less. Yet, there are still lower-ranked universities that charge higher and higher rank universities that charge lower than others.

The pattern in the Undergraduate Enrollment by University Rank chart doesn’t seem to pop out because Enrollments seem to be unaffected by the University Ranking. Even though the University’s Rank increases, the number of Enrollments for high-ranked universities is still below average and others with lower ranks turn out to have numbers that are above average.

This chart can be useful to study if the university can attain the targeted Undergraduate Enrollment numbers given their Tuition and Fees. Universities can use the chart to highlight their names and studies if the Tuition and Fees and the Undergraduate Enrollment numbers are below or above the average numbers. Moreover, Universities can study and adjust the Reference Line to represent the average Tuition and Fee students can afford to pay and observe if they are charging the right price by comparing their values to other universities.

***Spatial Heatmap – Understand the Demography of University Enrollment based on Location, State and Tuition and Fee***

Map

Description automatically generated

Figure 24. Spatial Heatmap of %Total Undergrad Enrollment

The spatial heatmap is used to understand the Percent of Total Undergraduate Enrollments based on the State level. This can be used to see the population distribution of students in each state and how each state performed in attracting undergraduate students.

The colour gradation is used to represent states with the most percentage of total Undergraduate Enrollments. In the spatial heatmap shown in Figure 24, it is shown that California (CA) states have the most Undergraduate Enrollment numbers compared to other states while Idaho (ID) states have the least Undergraduate Enrollment numbers compared to other states.

The spatial heatmap only shows the Maximum and Minimum values, however, universities can use the highlight and interact with the available tooltips to study the demographics and performance of each state independently.

**Question 4**

**Business Performance Dashboard**

**Chart

Description automatically generated**

Figure 25. Business Performance Dashboard

***Keep it Simple***

The dashboard and chart are simple because features are not overlapped in a single chart and the different variables used in plotting each chart are only limited to three. Furthermore, the x-axis label/header in the bar chart is removed to provide more space and make the dashboard less cluttered than it was. This header can be removed because it is already mentioned in the chart title. Very few legends are displayed in the dashboard to avoid over-cluttering.

***Don’t Display Everything***

All data required for the dashboard to display are displayed appropriately. Although there are many data points in each chart, the data points meaningfully contribute to the analysis that has been generated. With some data points missing, the trend and patterns will change.

***Keep to a Single Page***

The dashboard is displayed on a single page without the need to navigate to other pages.

***Avoid Fancy Formatting***

Other than the spatial heatmap, other graphs only use two main colours to make comparisons easier. No additional formatting was made to make the dashboard fancy.

***Use Layout and Placement***

The spatial heatmap is prioritised in the dashboard so that users can have an overview of undergraduate enrollments around the different states in the United States of America. Thereon, the other charts are used to provide supplementary information such as trends, relationships, and demographics of other variables so users can view them in more detail from the previous chart.

***Format Numbers Effectively***

Monetary values are formatted are displayed in denominations of thousands (K) and the currency used is labelled in the axis headers.

***Use Titles and Labels Effectively***

Titles are described appropriately to represent the graph. Column headers effectively represent the values in the chart and the legends provide a supplementary explanation of the details in the chart.

**Actions for Advanced Dashboard Navigation**

Three separate and different actions are available for advanced dashboard navigation so users can Highlight data points; Filter information; and direct users to an external HTTP link.

Graphical user interface, text, application

Description automatically generated

Figure 26. Dashboard Actions

***Highlight Action in Bar Chart***

Graphical user interface, text, application, email

Description automatically generatedMap

Description automatically generated with medium confidence

Figure 27. Highlight Action Setting and Result

This action lets users hover around the Bar Chart and highlight the data points in the Bar Chart. Once the users hover over the data points in the Bar Chart, it will only focus on those data points and show the corresponding values in the Scatter Plot and the Spatial Heatmap.

***Filter Action in Spatial Heatmap***

Graphical user interface, application, table

Description automatically generatedA picture containing graphical user interface

Description automatically generated

Figure 28. Filter Action Setting and Result

This action allows users to filter out the data that are related to each state. When a user selects a state in a map, the corresponding information is filtered and displayed in the Scatter Plot and Bar Chart as well.

***Navigating to External HTTP Link in Spatial Heatmap***

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedChart

Description automatically generated

Figure 29. External HTTP Setting and Result

This action allows users to investigate in more detail and compare the undergraduate enrollment numbers in Google Search Engine. By selecting the state and triggering the URL in the menu bar, users are directed to Google with the search prompt/query of “Student Enrollment in <State> state.”

**Appendix**

A) Excel Computation for Summary Statistics

1. For mean values, the =AVERAGE function is used.

Table

Description automatically generated with medium confidence

Figure . MEAN function

2. For median, the =MEDIAN function is used.

Graphical user interface, application, table

Description automatically generated with medium confidence

Figure . MEDIAN function

3. For mode, the =MODE function is used.

Graphical user interface

Description automatically generated with medium confidence

Figure . MODE function

For Modes of Location and State

The modes of Location and State can be found using the =COUNTIF function. To do this, duplicates should be removed from each column.

Graphical user interface, application

Description automatically generated

Figure . COUNTIF function

Thereon, the =COUNTIF function is used to fund how frequently the unique values appear. The most frequent Location and State are the mode for the column.

4. For maximum, the =MAX function is used.

Graphical user interface, application

Description automatically generated

Figure . MAX function

5. For minimum, the =MIN function is used.

Graphical user interface, application, table

Description automatically generated

Figure . MIN function

6. For the number of observations, the =COUNTIF function is used.

Graphical user interface, application

Description automatically generated

Table

Description automatically generatedTable

Description automatically generated

Figure .COUNTIF function

**--- END OF ECA PAPER ---**