Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

# Assessment Submission Form

| | |
|---|---|
| **Student Name** | Kul Gaurav |
| **Student ID Number** | 19307204 |
| **Course Title** | MSc. Computer Science (Data Science) |
| **Module Title** | DATA VISUALISATION |
| **Assessment Title** | A3 Novel Visualization |
| **Date Submitted** | 19-04-2020 |
| **Word Count** | 1170 |

**I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at:** http://www.tcd.ie/calendar

**I have also completed the Online Tutorial on avoiding plagiarism 'Ready, Steady, Write', located at**

http://tcd-ie.libguides.com/plagiarism/ready-steady-write

**I declare that the assignment being submitted represents my own work and has not been taken from the work of others save where appropriately referenced in the body of the assignment.**

Kul Gaurav

Date: 19-04-2020

# 1 Introduction

The visualization provides the exploratory analysis of college statistics for the USA. Various questions related to the data can easily be explored using the dashboard. We visualize the different features which a student may think of while choosing a college in the USA. This visualization highlights the insights for total enrolments, early-career pays, type of school, and cost for the study using multiple coordinated views. We look at the relationships in the data and connect to questions about the data.

# 2 Dataset

The dataset is openly available as a part of Tidy Tuesday, a weekly data project on GitHub [1]. The data is of CSV format and is also maintained on Kaggle [2] with the usability factor of 10.0 and is broadly classified into five different CSVs. Following four of the CSV files contain data that are used in the process of visualization:

1. Tuition_cost.csv

| Variable | Type | Description |
|---|---|---|
| state | Categorical | Name of the U.S. state |
| type | Categorical | Private school or public school |
| In_state_total | Quantitative, Continuous | Total cost for in-state residents in USD (sum of room & board + in state tuition) |
| Out_of_state_total | Quantitative, Continuous | Total cost for in-state residents in USD (sum of room & board + out of state tuition) |

2. Salary_potential.csv

| Variable | Type | Description |
|---|---|---|
| early_career_pay | Quantitative, Continuous | Estimated early career pay in USD |
| make_world_better<br>_percent | Quantitative, Continuous | Percent of alumni who think they are making the world a better place |
| stem_percent | Quantitative, Continuous | Percent of the student body in STEM |

2. Diversity_school.csv

| Variable | Type | Description |
|---|---|---|
| total_enrollment | Quantitative, Continuous | Total enrollment of students |

3. Historical_tuition.csv

| Variable | Type | Description |
|---|---|---|
| year | Quantitative, Discrete | Academic year |
| tuition_cost | Quantitative, Continuous | Tuition cost in USD |

The final data is prepared by taking unions on the state name for various files.

# 3 Task

The visualization aims at identifying the correlation between various factors which compare the education system for different states of the U.S. We analyze the following major tasks:

- The relation between science, technology, engineering, and mathematics (STEM) students and average early career pay.
- Distribution of total enrollments in the country compared to STEM enrollments in the state.
- The trend of total education cost based on the type of school and the total number of students in the state.

- Correlate the number of alumni believing that he or she is making the world a better place with the early career pay.

The visualization utilizes the commonality of the state for which we explore the above questions or compare if more than one region is selected.

## 4 Visual Encoding

For the effectiveness of the visualization following visual encodings are used:

- Position: The states in both the maps are placed as per the relative geolocation.
- Color: Various colors are utilized for identifying the commodities in exploration
- Brightness: The colors use the brightness level to determine the level of quantitative values.
- Mark: Two significant marks are used, circle and square, in different maps.
- Size: The size of the marks represents the difference in continuous quantitative values.
- Motion: As we just have the data of cost change for study for several years, we show the change by animation. Once the data is available for other features too over time, it can easily be extended for them.

## 5 Approach

Tableau is used for the visualization process. All the coordinated views are custom charts and use different default charts available in the tool. The following procedures are performed for various aspects:

- Dual-axis map for total enrollments and STEM students' distribution in each state of the U.S. The color's brightness and size of the mark encodes the visualization. Since the dataset is not including geolocation coordinates, another CSV file is created using the data from [6].
- Hex map for average pay distribution. This eliminates the visual perception by different sizes of the states but keeps the relative position. Hex map files are available on the blog of Tableau Zen Master Joshua Milligan [5]. The SHP file is added to the data source with union of states names. Similar to the first map, the brightness of the color gives the idea of the strength of early career pay distribution in the state and size of the squares represent the number of alumni of states who think they are making the world a better place by their work.
- Donut chart for the type of school for the selected states. This is created by creating two pie charts of different sizes in dual-axis mode and changing the background color for smaller pie chart.
- Butterfly chart [4] to display the cost of study for in-state and out-of-state students. The view gives a quick glance at both the cost for a given state at the same time. Color encodes the two costs and provides a glance of change with brightness. To create the views, one of the bar chart's scale is reversed, and the zero-axis calculated field is added, which holds the names of the states.
- Tuition cost change is again dual-axis synchronized chart to display the cost for a given year. Tableau Pages is utilized to provide the animation functionality.

Multiple calculated fields are generated in the measures and dimensions to achieve the visualization.

The filter of state name and type of school is kept universal for all the views. Custom tooltips are attached to the visualization to show on hover on the maps.

## 6 Conclusion

The visualization uses various data sources and can help explore multiple curious questions based on the statistics of the college data in the U.S. The dataset selected for the visualization is complex and prepared for the final use by Tableau Prep Builder and Python Scripts. No visualization of this dataset is found to the best of my knowledge except for some fundamental exploratory data analysis (EDA) using Python and R plots, hence making the visualization novel.

# 7 Source code and Links

The Tableau workbook files, presentation video, thumbnail image, hosted dashboard are available at: https://github.com/kulgaurav/College-Statistics-US

Presentation video: https://kulgaurav.github.io/College-Statistics-US/playVideo.html

Dashboard: https://kulgaurav.github.io/College-Statistics-US/

Thumbnail image: https://github.com/kulgaurav/College-Statistics-US/blob/master/Thumbnail.png

Tableau Files: https://github.com/kulgaurav/College-Statistics-US/tree/master/Tableau

Data used: https://github.com/kulgaurav/College-Statistics-US/tree/master/Data

Original Data: https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-03-10

Visualization on Tableau Public:
https://public.tableau.com/profile/kul4423#!/vizhome/Viz_15867291815930/Dashboard1?publish=yes

# References

[1] "College tuition, diversity, and pay" [Online]. Available: https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-03-10/readme.md.

[2] " TidyTuesday week 11" [Online] Available: https://www.kaggle.com/jessemostipak/college-tuition-diversity-and-pay

[3] " Data Visualization in Tableau" [Online] Available: https://www.udacity.com/course/data-visualization-in-tableau--ud1006

[4] " Tableau 2018: Hands-On Tableau Training For Data Science! " [Online] Available: https://www.superdatascience.com/courses/tableau-2018-hands-on-tableau-training-for-data-science

[5] Hex Map Spatial File [Online]. Available: https://vizpainter.com/hex-map-spatial-file/

[6] States in United States [Online]. Available: https://www.latlong.net/category/states-236-14.html