STAT515 Final Project Report

Shravan Chintha

G01064991

Introduction:

The goal of this project is to select, explore, describe various aspects of a dataset using different combinations of graphics and visualize to observe insights. To achieve this goal, a dataset that satisfies this criterion and large enough to support a variety of visualizations is to be selected. The attributes from the dataset are selected such that information from the plots generated convey necessary insights.

Dataset:

The dataset that is chosen for this task is Universities offering Data science course in United States of America. The dataset gives information about all the colleges which offer Data Science course as part of their curriculum in US.

The main reason to choose this data set is, as most of the students in the class belong to Data Science department, it would be interesting see which all colleges in the US offer the course data science and know details about where they stand as compared to other colleges which offer data science courses with respect to other variables in the dataset. As the part of the task, correlated other variables in the dataset to see interesting patterns or variations that would analyze the dataset more comprehensively.

The dataset is downloaded from https://www.kaggle.com/datasets , which is one of the major sources in the web for data sets. The initial dataset that is downloaded looks as below:

	SCHOOL	STATE	CITY	NOC	PROGRAM	TYPE	DEPARTM	DELIVERY	DURATION	PREREQ	LINK	LOC_LAT	LOC_LON	WORLD_R	COUNTRY	TEACHING	INTERNAT	RESEARCH	CITATION	INCOME	TOTA
2	Albright C	Pennsylva	Reading	1	M.S. in Bu	M	Erivan K. H	Online	30 credits	Not Availa	http://ww	40.3602	-75.9101	NA	NA	NA	NA	NA	NA	NA	NA
3	American	Colorado	Aurora	1	Master of	M	Not Availa	Online	36 credits	Not Availa	http://ww	39.6766	-104.831	NA	NA	NA	NA	NA	NA	NA	NA
4	American	District of	Washingt	2	Online ME	M	Kogod Sch	Online	15 months	Not Availa	https://or	38.9378	-77.0901	401-500	United Sta	42.2	28.9	16.5	41.1	35.9	-
5	American	District of	Washingt	2	Master of	M	Kogod Sch	Online	33 credit h	Not Availa	http://ww	38.9378	-77.0901	401-500	United Sta	42.2	28.9	16.5	41.1	35.9	-
6	Arizona St	Arizona	Tempe	1	Master of	M	W.P. Care	Online or	9 month (	Not Availa	https://pr	33.4219	-111.94	146	United Sta	33.8	28.6	35.9	83.6	31.4	
7	Arizona St	Arizona	Tempe	1	Master of	M	W.P. Care	Online or	9 month (	Not Availa	https://pr	33.4219	-111.94	161	United Sta	43	24.1	44.1	66.9	-	
8	Arizona St	Arizona	Tempe	1	Master of	M	W.P. Care	Online or	9 month (	Not Availa	https://pr	33.4219	-111.94	148	United Sta	38.4	27.4	45.2	79.9	31.8	
9	Arizona St	Arizona	Tempe	1	Master of	M	W.P. Care	Online or	9 month (	Not Availa	https://pr	33.4219	-111.94	127	United Sta	38.2	26.1	39	80.3	28.7	
10	Arizona St	Arizona	Tempe	1	Master of	M	W.P. Care	Online or	9 month (	Not Availa	https://pr	33.4219	-111.94	182	United Sta	35.7	29.5	37.5	73.1	32.6	
11	Arizona St	Arizona	Tempe	1	Master of	M	W.P. Care	Online or	9 month (	Not Availa	https://pr	33.4219	-111.94	189	United Sta	32.4	31.9	38.1	84.6	32	
12	Aspen Un	Colorado	Denver	1	Master of	M	Not Availa	Online	36 credits	Not Availa	http://ww	39.7037	-104.94	NA	NA	NA	NA	NA	NA	NA	NA
13	Auburn Ur	Alabama	Auburn Ur	1	Online Ma	M	Raymond	Online	36 credit h	Not Availa	http://har	32.6002	-85.4924	501-600	United Sta	30.2	27.9	19.4	21	37.2	-
14	Auburn Ur	Alabama	Auburn Ur	1	Online Ma	M	Raymond	Online	36 credit h	Not Availa	http://har	32.6002	-85.4924	350-400	United Sta	33.7	22.5	18.7	10.3	47.3	-
15	Aurora Un	Illinois	Aurora	1	Master of	M	Dunham S	Campus	30 credit h	Not Availa	http://ww	41.7545	-88.3495	NA	NA	NA	NA	NA	NA	NA	NA
16	Austin Pe	Tennesse	Clarksville	2	Professior	M	Departme	Online	2 years	Not Availa	http://ww	36.5333	-87.3541	NA	NA	NA	NA	NA	NA	NA	NA
17	Austin Pe	Tennesse	Clarksville	2	Professior	M	Departme	Online	2 years	Not Availa	https://w	36.5333	-87.3541	NA	NA	NA	NA	NA	NA	NA	NA
18	Babson Co	Massachu	Wellesley	1	MBA with	M	F.W. Olin	Campus	41-55 cred	Not Availa	http://ww	42.2965	-71.2695	NA	NA	NA	NA	NA	NA	NA	NA
19	Baker Coll	Michigan	Allen Park	1	MBA in Bu	M	Baker Coll	Online	53 hours	Not Availa	http://ww	42.2745	-83.2032	NA	NA	NA	NA	NA	NA	NA	NA
20	Bay Path U	Massachu	Longmeac	1	MS in App	M	Graduate	Campus o	36 credits	math, stat	http://ww	42.0554	-72.584	NA	NA	NA	NA	NA	NA	NA	NA

As seen from above, the dataset contains lot of missing values (NAs). In order to produce visualizations as desired, the dataset needs to be cleaned and preprocessed.

Data Pre-processing:

The missing values in the data set are replaced with mean of the other values in the column using mean data imputation method. All the NAs are thus replaced with means of their respective columns.

The next step in data cleaning is to find duplicate or repeated values which convey the same meaning but repeated to cause deviations in data in any of the columns present in the dataset. Found “DELIVERY” column in the data set as this kind.

DELIVERY column in the dataset contained values which have same meanings but are written in different format.

For example:

The delivery column has the below values,

Online or Campus

Campus or Online

On Campus or Online

All the three describes about the same thing i.e., Course offered in Online or Campus mode. But, the values are different. Hence changed all such values in the data set as below:

Online or Campus

Campus or Online → ***Campus or Online***

On Campus or Online

Similarly, the rest values which belong to the above type are modified as required to get better results.

After the data is pre-processed and ready for analysis, looks as below:

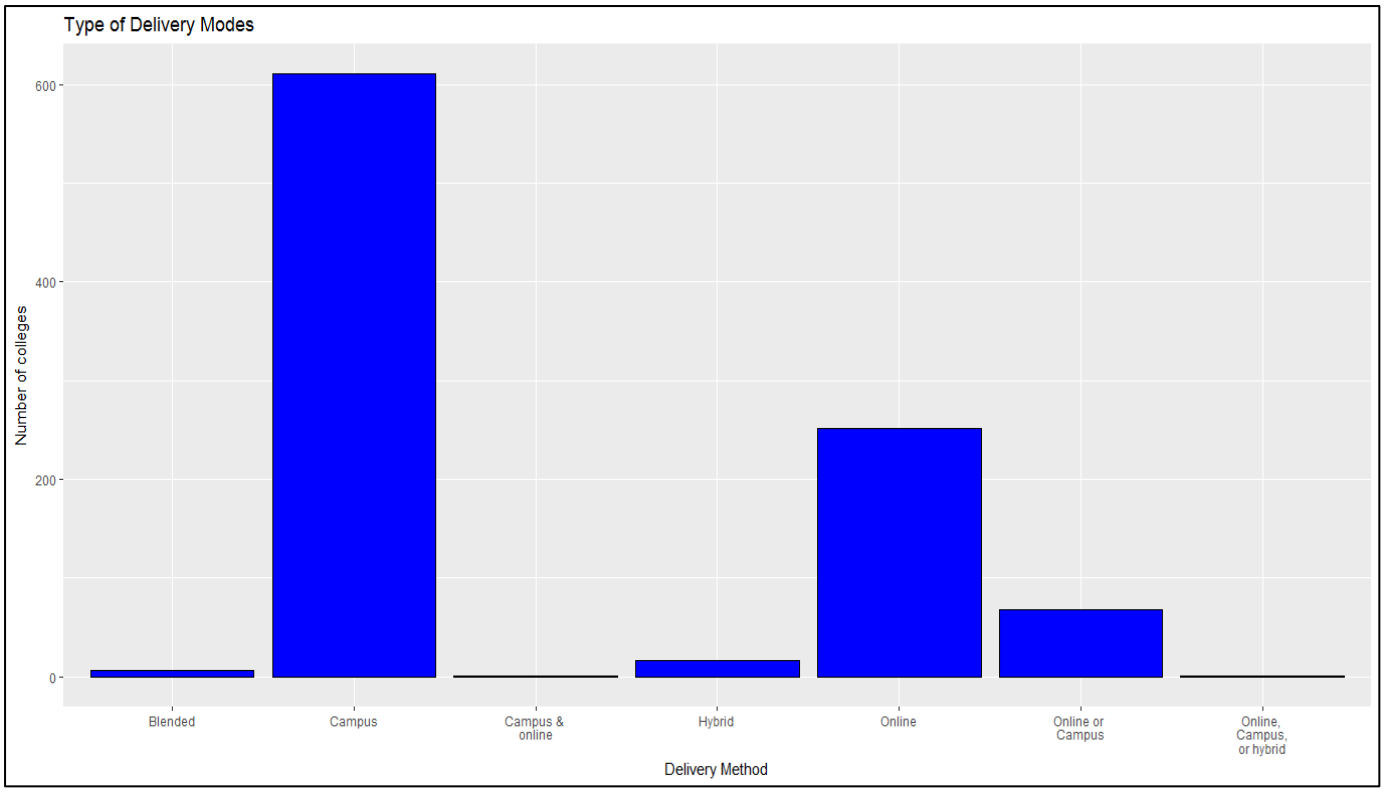
	SCHOOL	STATE	CITY	NOC	PROGRAM	TYPE	DEPARTMENT	DELIVERY	Credits
2	Albright College	Pennsylvania	Reading	1	M.S. in Business Intelligence	M	Erivan K. Haub School of Business	Online	
3	American Sentinel University	Colorado	Aurora	1	Master of Science Business Intelligence and A	M	Not Available	Online	
4	American University	D.C.	Washington	2	Online MBA with Analytics Concentration	M	Kogod School of Business	Online	
5	American University	D.C.	Washington	2	Master of Science in Analytics	M	Kogod School of Business	Online	
6	Arizona State University	Arizona	Tempe	1	Master of Science in Business Analytics	M	W.P. Carey School of Business	Online or	
7	Arizona State University	Arizona	Tempe	1	Master of Science in Business Analytics	M	W.P. Carey School of Business	Online or	
8	Arizona State University	Arizona	Tempe	1	Master of Science in Business Analytics	M	W.P. Carey School of Business	Online or	
9	Arizona State University	Arizona	Tempe	1	Master of Science in Business Analytics	M	W.P. Carey School of Business	Online or	
10	Arizona State University	Arizona	Tempe	1	Master of Science in Business Analytics	M	W.P. Carey School of Business	Online or	
11	Arizona State University	Arizona	Tempe	1	Master of Science in Business Analytics	M	W.P. Carey School of Business	Online or	
12	Aspen University	Colorado	Denver	1	Master of Science in Technology and Innovati	M	Not Available	Online	
13	Auburn University	Alabama	Auburn University	1	Online Master of Business Administration wi	M	Raymond J. Harbert College of Business	Online	
14	Auburn University	Alabama	Auburn University	1	Online Master of Business Administration wi	M	Raymond J. Harbert College of Business	Online	
15	Aurora University	Illinois	Aurora	1	Master of Science in Digital Marketing and Ar	M	Dunham School of Business	Campus	
16	Austin Peay State University	Tennessee	Clarksville	2	Professional Science Master's in Data Manage	M	Department of Computer Science and In	Online	
17	Austin Peay State University	Tennessee	Clarksville	2	Professional Science Master's Degree in Pred	M	Department of Mathematics and Statisti	Online	
18	Babson College	Massachusetts	Wellesley	1	MBA with Business Analytics Concentration	M	F.W. Olin Graduate School of Business	Campus	
19	Baker College	Michigan	Allen Park	1	MBA in Business Intelligence	M	Baker College	Online	
20	Bay Path University	Massachusetts	Longmeadow	1	MS in Applied Data Science	M	Graduate school	Online or	

Now, as seen from above, the data set is clean without any missing values and ready for analysis.

Visualizations:

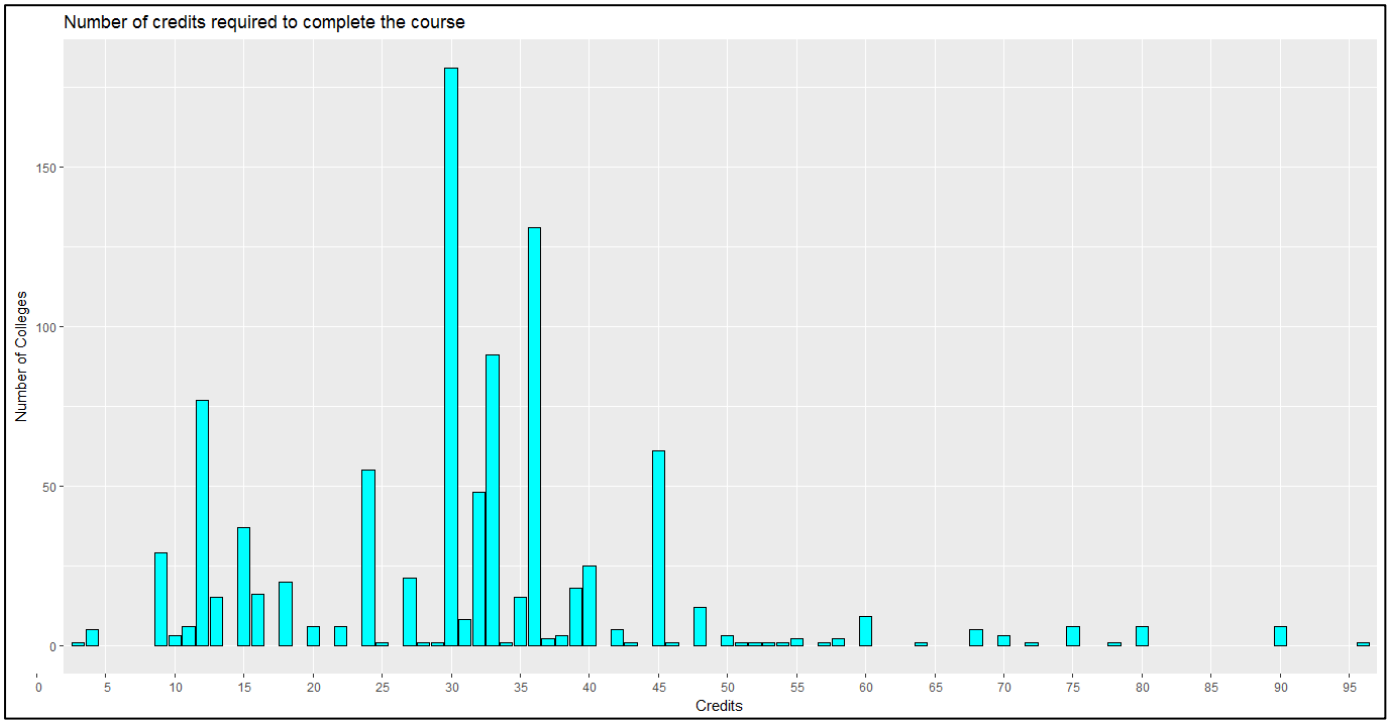
To start with, plotted different bar plots using **ggplot2** in R for different attributes in the data set to compare and gain insights from the plots.

Type of delivery methods offered by different colleges:



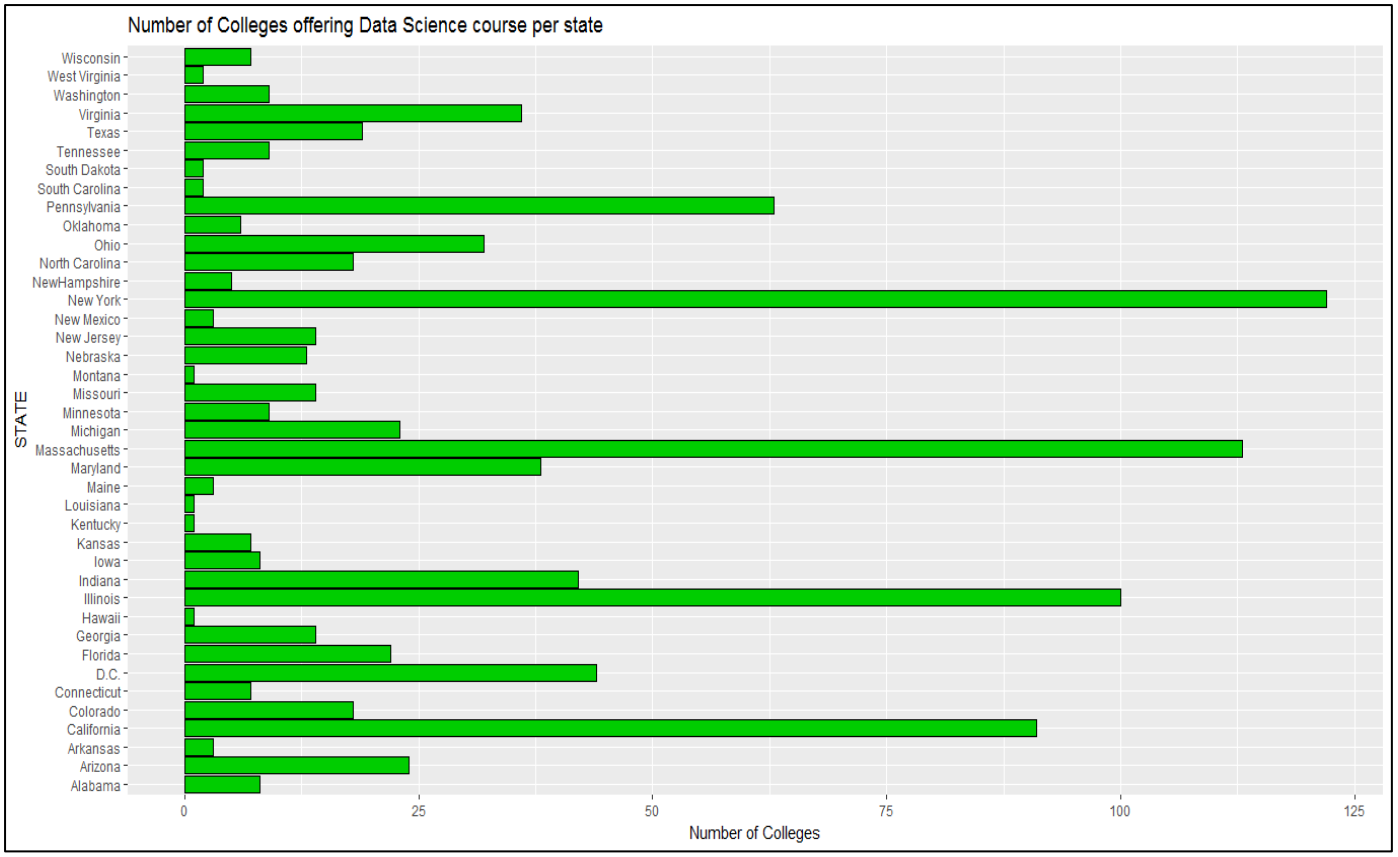
The above plot shows the type of delivery methods offered in different colleges. Clearly the method type “Campus” which means the course that is offered in On Campus mode is high with a value around 610, among all the other modes of delivery. Next stands Online method of delivery with value around 250 colleges. It means that the number of colleges that offer data science course in US offer them mostly in On Campus mode.

Number of Credits required to complete the course:



The above plot describes about the number of Credits that are required to complete the course in number of colleges in US. Value for colleges that offer course in 30 credits is high. It is around 180 for 30 Credits system and around 130 for 36 Credits system. More number of colleges offer the course requiring 30 Credits and next is 36 credits.

Number of colleges offering Data Science per state:



The above plot shows the number of colleges that offer data science course in each state. The number of colleges that offer Data Science courses are more in New York with a value around 122 and the next state is Massachusetts with a value around 111. The least states being Louisiana, Kentucky and Hawaii with once each college in each state.

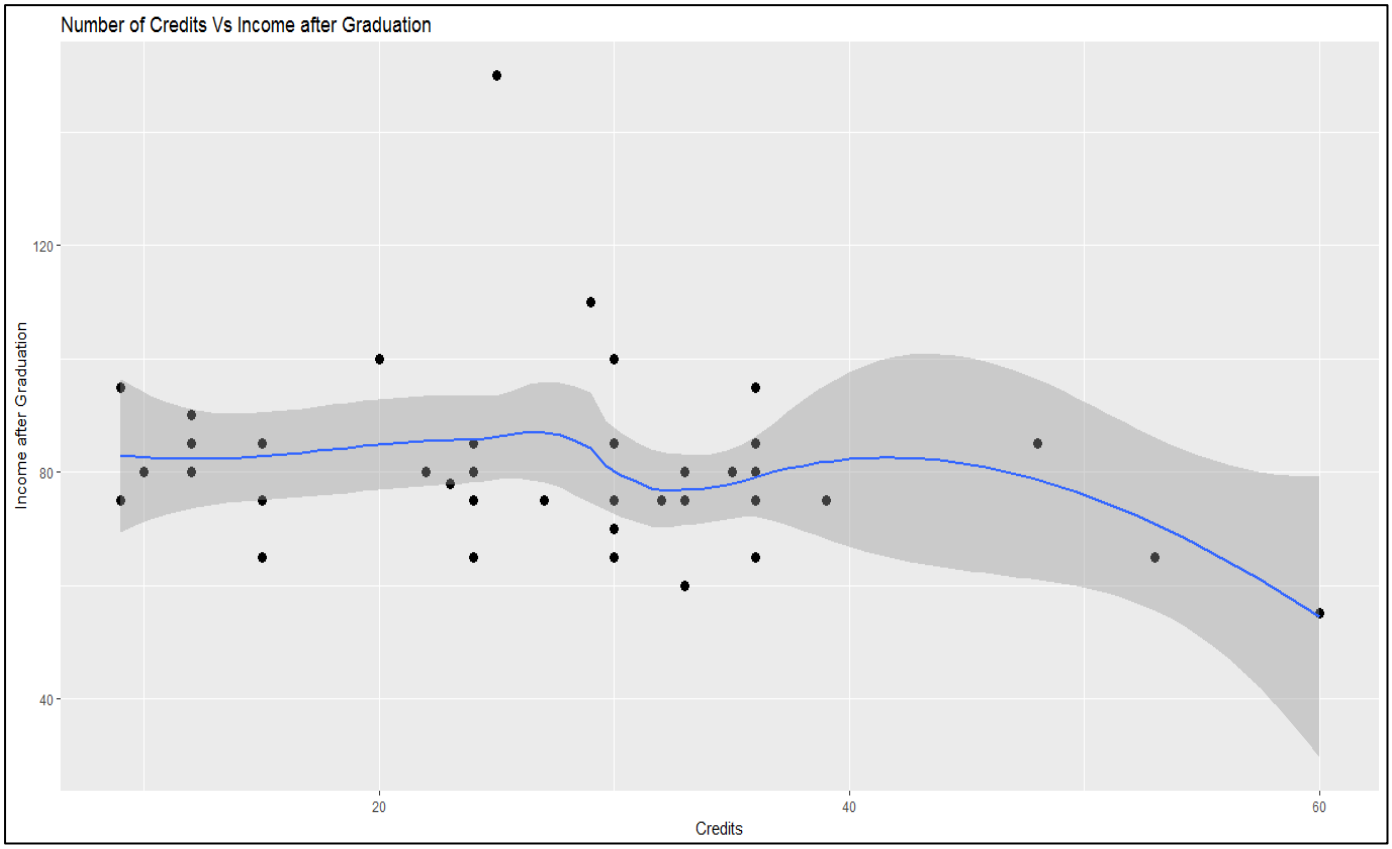
Universities in each state and their stats comparison per state:

To visualize data per each state, dataset needs to be modified. To get comparisons between different factors in the dataset as per state, the data set is modified as below:

Order	State	Credits	TEACHING	INTERNATIONAL	RESEARCH	INCOME	TOTAL_SCORE	STUDENT_STAFF_RATIO
1	Massachusetts	25	55.9	44.56	52.23	150	67.6	10.93
2	Minnesota	48	55.9	44.56	52.23	85	67.6	11.56
3	New Hampshire	39	55.9	44.56	52.23	75	67.6	11.56
4	New Jersey	36	55.7	46.6	59.3	95	66.4	12.5
5	Vermont	30	54.2	56.4	64.2	75	59.4	14.2
6	DoDEA	23	22	12	23	78	23	12
7	North Dakota	24	56.6	45.5	70.25	80	81	14.6
8	Washington	30	55.9	44.56	52.23	75	67.6	11.56
9	Pennsylvania	9	45	60.6	35.9	95	46.9	11.92
10	Maine	15	55.9	44.56	52.23	85	67.6	11.56
11	Kansas	15	22.9	41.7	17.5	75	67.6	13.6
12	Ohio	33	51	35.4	38.6	80	55	12.92
13	Wyoming	33	65.2	47.2	53.2	60	62.7	13.2
14	Colorado	36	55.9	44.56	52.23	85	67.6	18
15	Montana	32	55.9	44.56	52.23	75	67.6	11.56
16	Wisconsin	36	55.9	44.56	52.23	85	67.6	12
17	Texas	32	55.9	44.56	52.23	75	67.6	12.05
18	Indiana	10	55.9	44.56	52.23	80	67.6	15
19	South Dakota	30	55.9	44.56	52.23	70	67.6	11.56
20	Virginia	22	55.9	27.9	43	80	55	14.12

The dataset is modified such that all the variables are put across each state as per their average values in colleges present in each state. Colleges are filtered as per each state and values in all the columns are averaged and imputed against each state in the data set. Now the dataset is ready for visualizing values for each state.

Number of Credits Vs Income after Graduation:

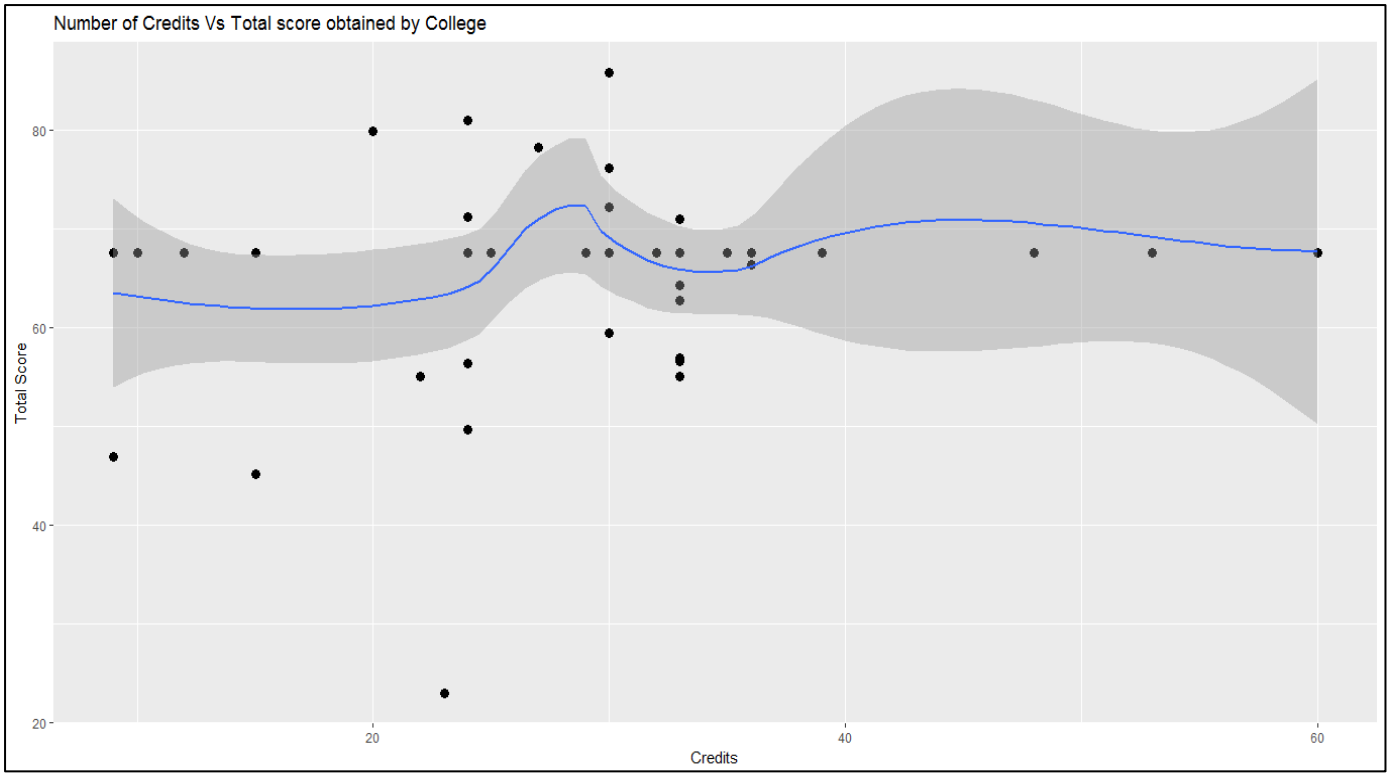


The above plot i.e., scattered plot with regression line shows the comparison between the average number of credits that are needed to complete a course in each state and the average income obtained after they graduate in each state. This comparison is made to see how the number of credits in a course varies with the income they obtain after the students graduate and start their jobs. More number of credits value lies between 22 to 38 and the income value lies between 65K to 95K dollars.

There are also few states for which the number of credits required are 60 and income is less than 60K.

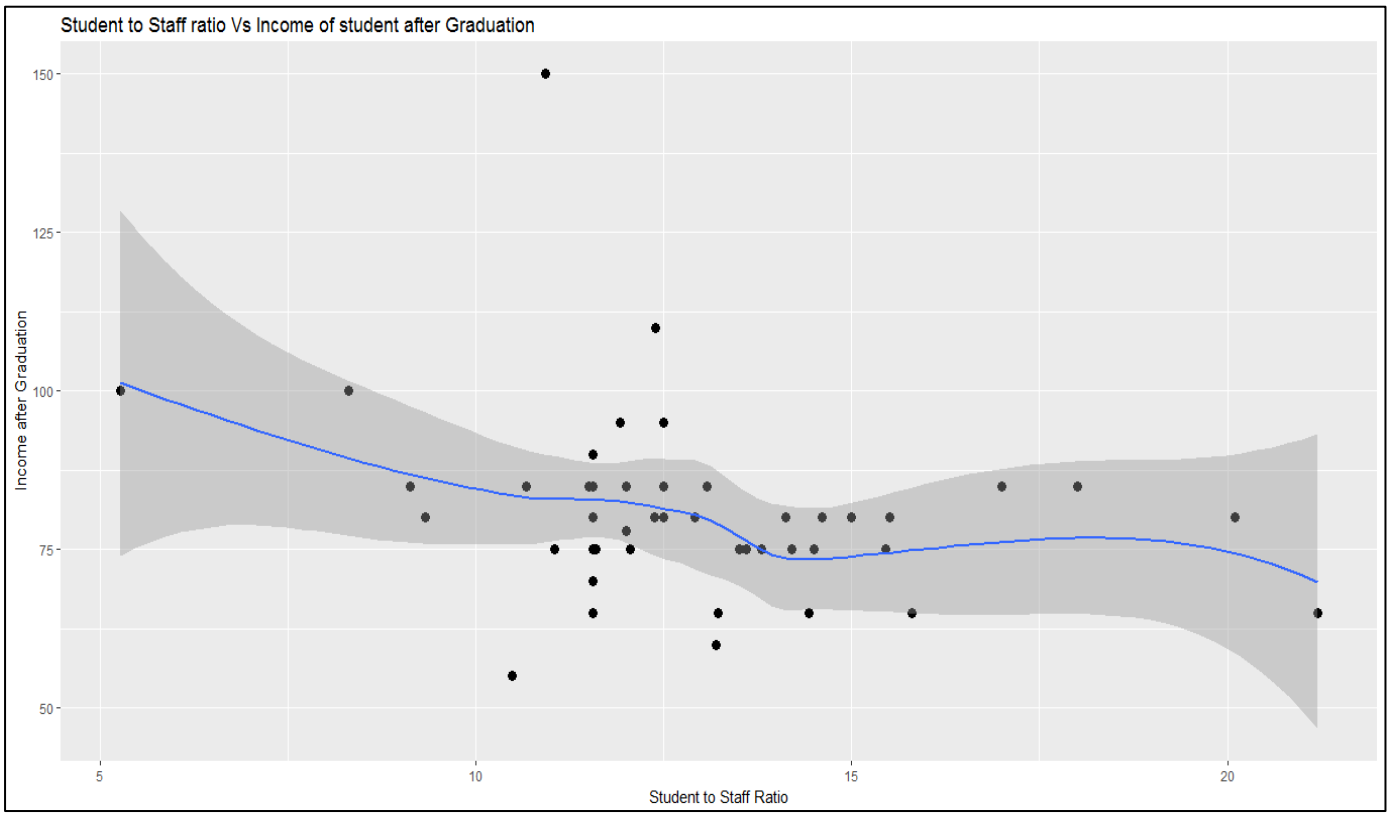
Number of Credits Vs Total Score obtained:

Total score is the value obtained by each college in state based on their performance in other factors like Research, Teaching, Citations and Income. Comparison is made between total score obtained by each state and the average number of credits required to complete the course to see how the number of credits required could impact the total score they get by performing.



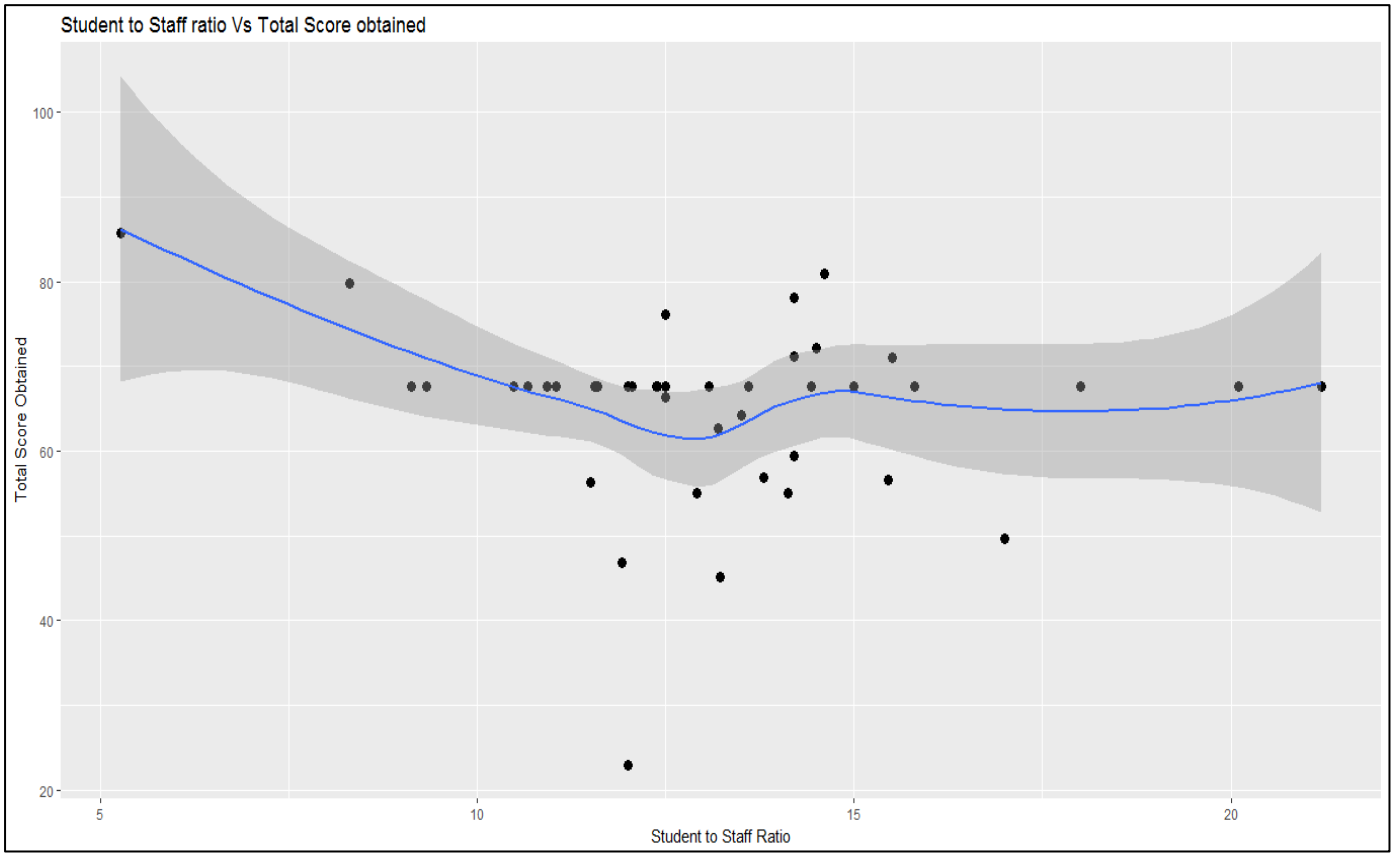
In the above plot, higher number of values for Total Score are between 60 and 80. And the credits are between 22 and 38. As in most colleges, the number of credits required are around 30, values for them are plotted around 30.

Student to Staff Ratio Vs Income of student after graduation:



The above plot shows comparison between the Student to Staff ratio and Income the students obtain after they graduate. This comparison is made to see how the Income of the students after they graduate vary with the student to staff ratio levels maintained in the colleges in each state. Most student to staff ratio ranges between 11 and 14.

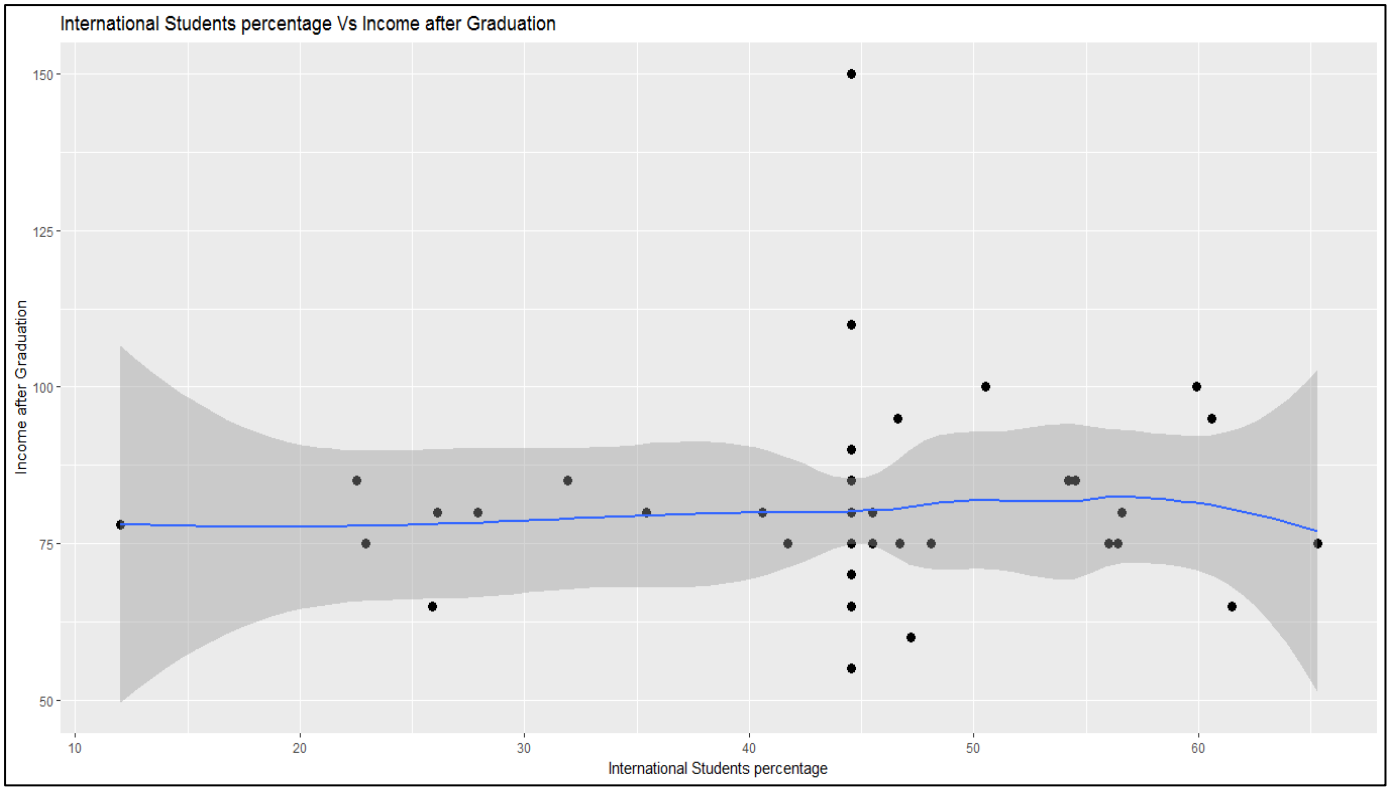
Student to Staff Ratio Vs Total Score obtained:



The above plot shows comparison between Student to Staff ratio and Total Score obtained by colleges in each state. This comparison is made to see how the Total score obtained by each college varies with the Student to Staff ratio they maintain in colleges in each state. Most of the values for Student to Staff ratio lies between 11 to 14 with values for Total score ranging between 65 to 75.

International Students Percentage Vs Income after graduation:

The comparison is made between the International students percentage and Income the students obtain after graduation to see how the in the Income of the student after graduation varies with International students percentage.



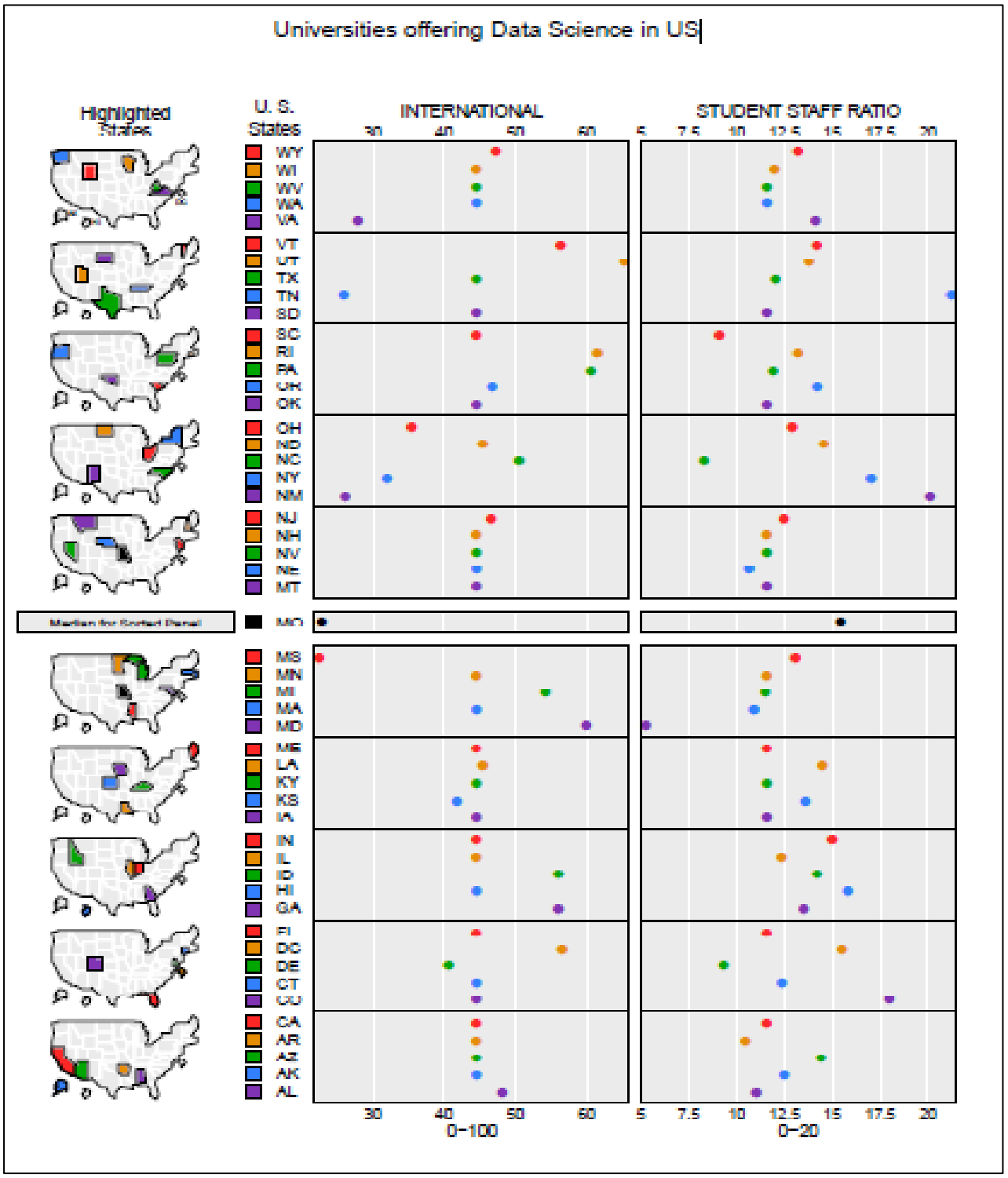
The above plot shows the variation of Income of students after they graduate with International students percentage in their colleges. The line structure for a value 44.56 in international students percentage is because of the mean imputation. Since the missing values are high for International students percentage variable, the imputed mean value in place of missing values is been displayed.

Micromaps:

The dataset is been converted to visualize the data per each state. The better representation of this data would be by plotting micro maps.

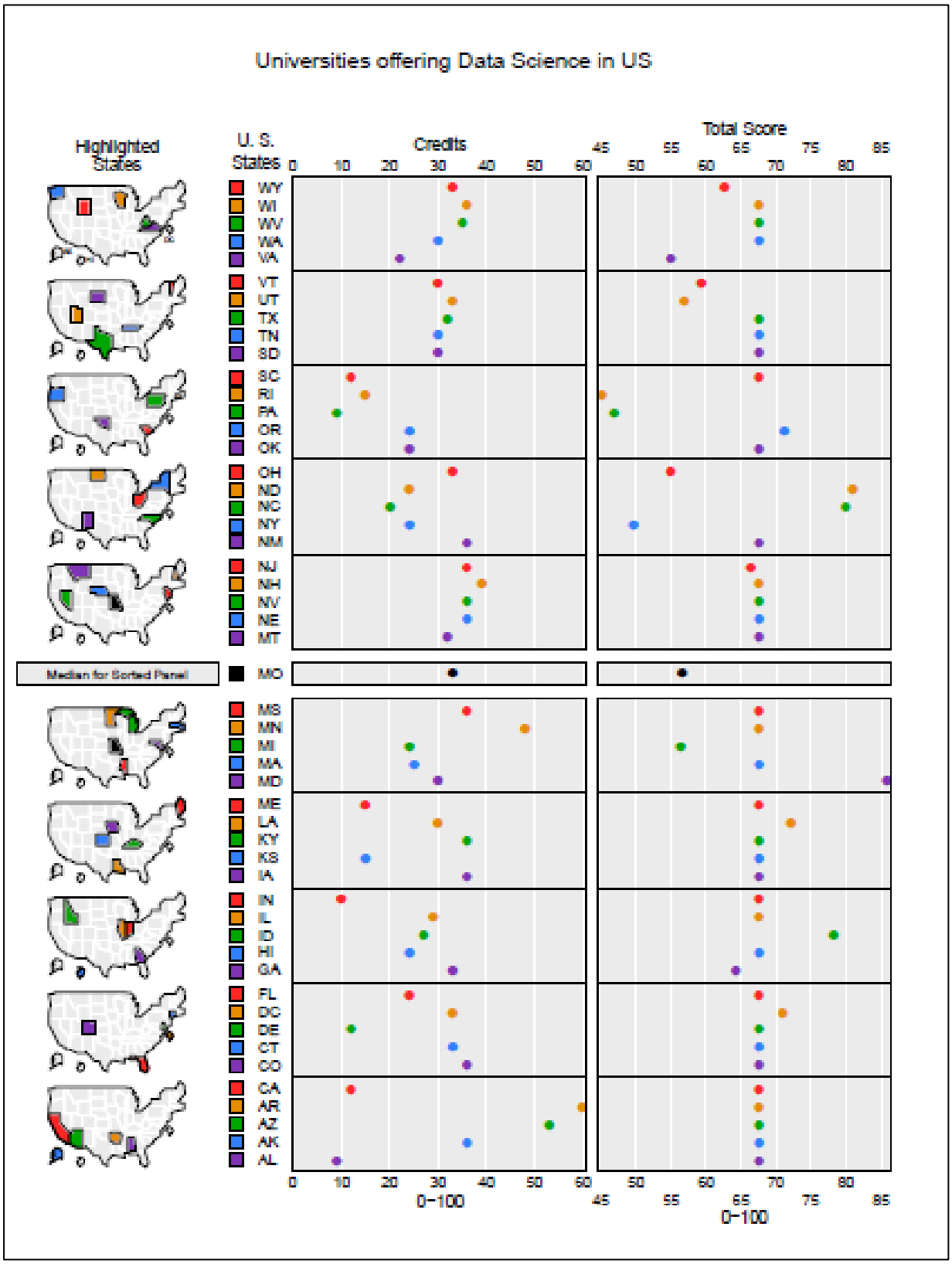
Hence micro maps are plotted for different variable in the modified dataset. Comparisons are made between International Students Percentage and Student to Staff ratio, Number of credits per state and Income obtained by students after they graduate, Number of credits per state and Total score obtained, Total score obtained by colleges in each state and Income of students after they graduate using micro maps in R.

International Students Percentage and Student to Staff ratio:



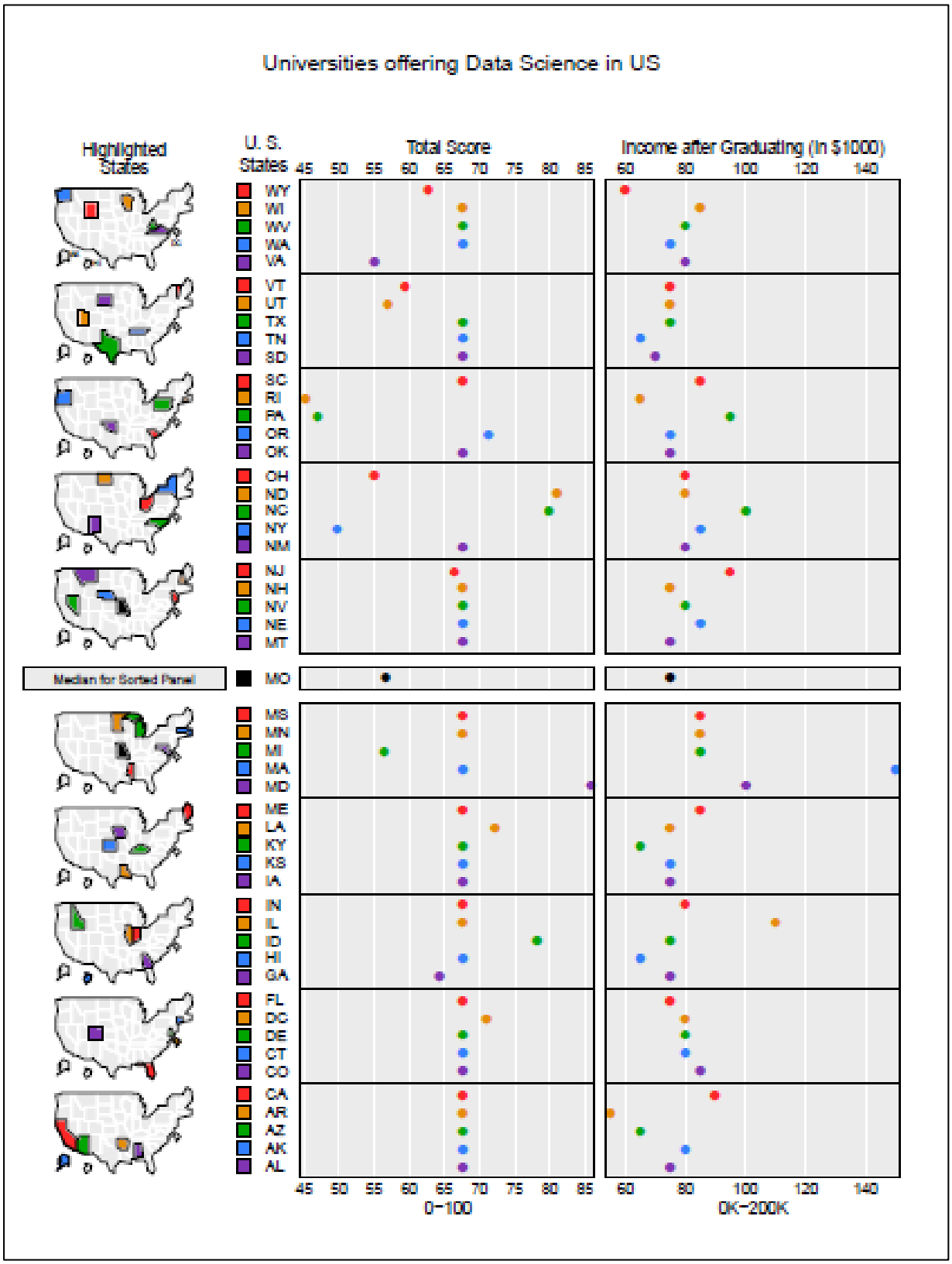
The above micro map describes the international students percentage in each state and Student to staff ratio in each state in the US. International Students ratio is highest for **UTAH** with **65.3%** of international students. Student to staff ratio is highest in **TENNESSEE** with **21.2** students per staff.

Number of Credits and Total Score obtained by college in each state:



The above micro map shows the number of credits needed to complete the course in each state and Total Score obtained by colleges in each state. Number of credits required to complete the course is highest in **ARKANSAS** state with **60** Credits. Total Score obtained is highest in **MARYLAND** with **85.8** value.

Total Score obtained by colleges in each state and Income obtained after students graduate:

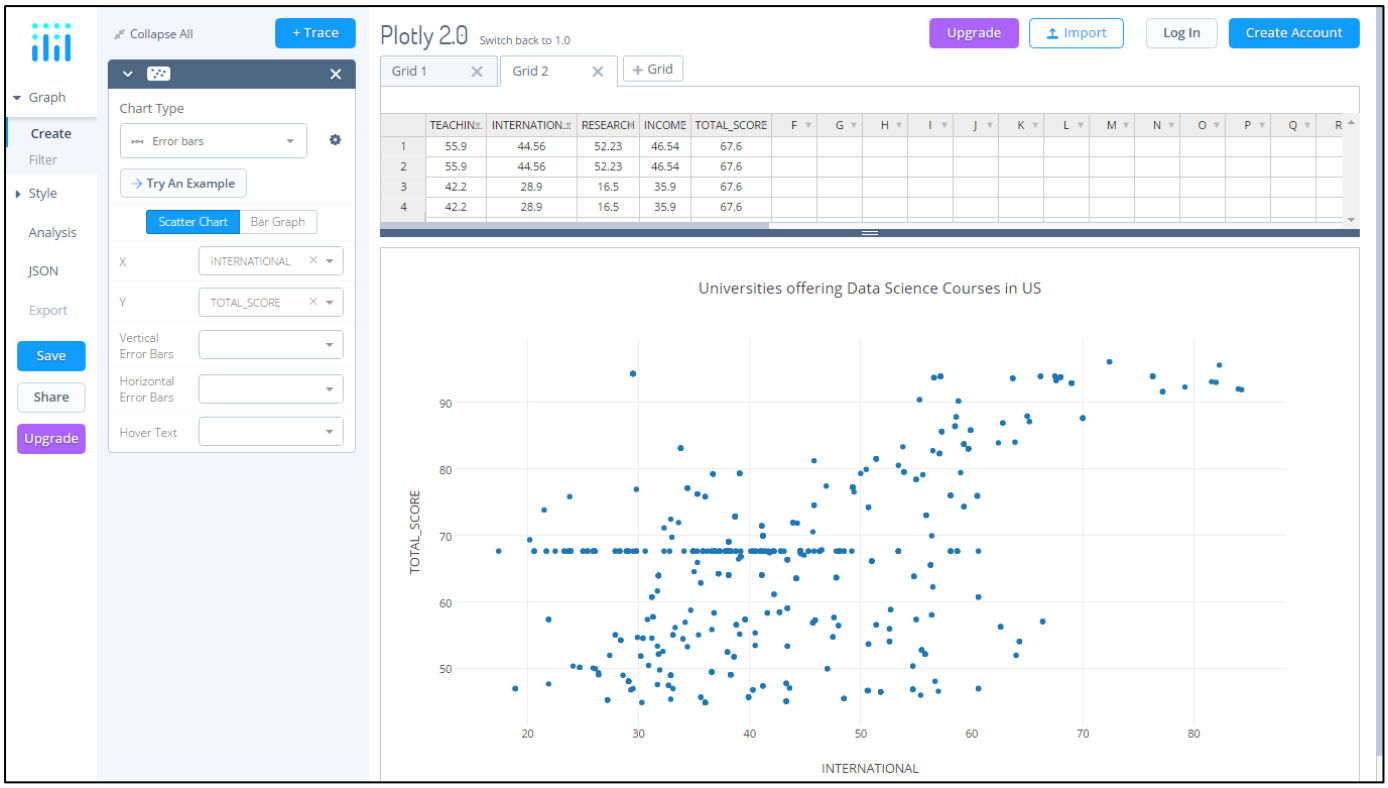


The above micro map shows the Total score obtained by colleges in each state and Income obtained by students after they graduate in each state. Total Score obtained is highest in **MARYLAND** with **85.8** value. Income obtained by students after they graduate is highest in **MASSACHUSETTS** with **150K** per year.

Plotly:

To add variety in visualizations, plotted graphs using Plotly in web. Plotly is an online data analytics and visualization tool which provides online graphing, statistics and analytics for individuals and collaboration, as well as scientific graphing libraries for languages like R, Python, MATLAB etc.

Total Score Vs International Students percentage:



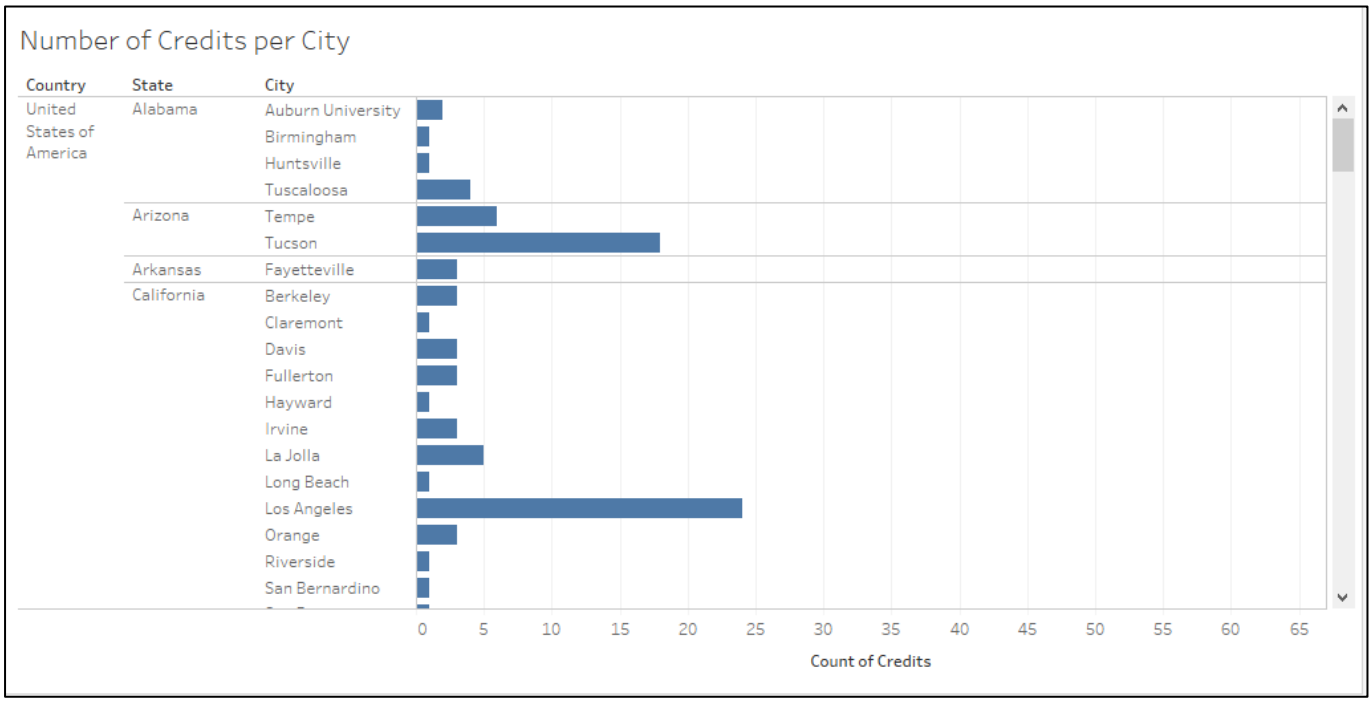
The above plot is plotted using Plotly, which compares Total score obtained in each college with International students ratio in each college. The whole data set is used to run the plot in Plotly. This provides us with the variation in Total score as the change in International students ratio in each college.

Tableau:

Tableau is a data visualization tool which generates visualizations with easy mechanisms as drag drop method by importing datasets. Used this tool to show the number of credits in each college with their city and state in US and Area graph showing Total score obtained in each state.

Below are the plots:

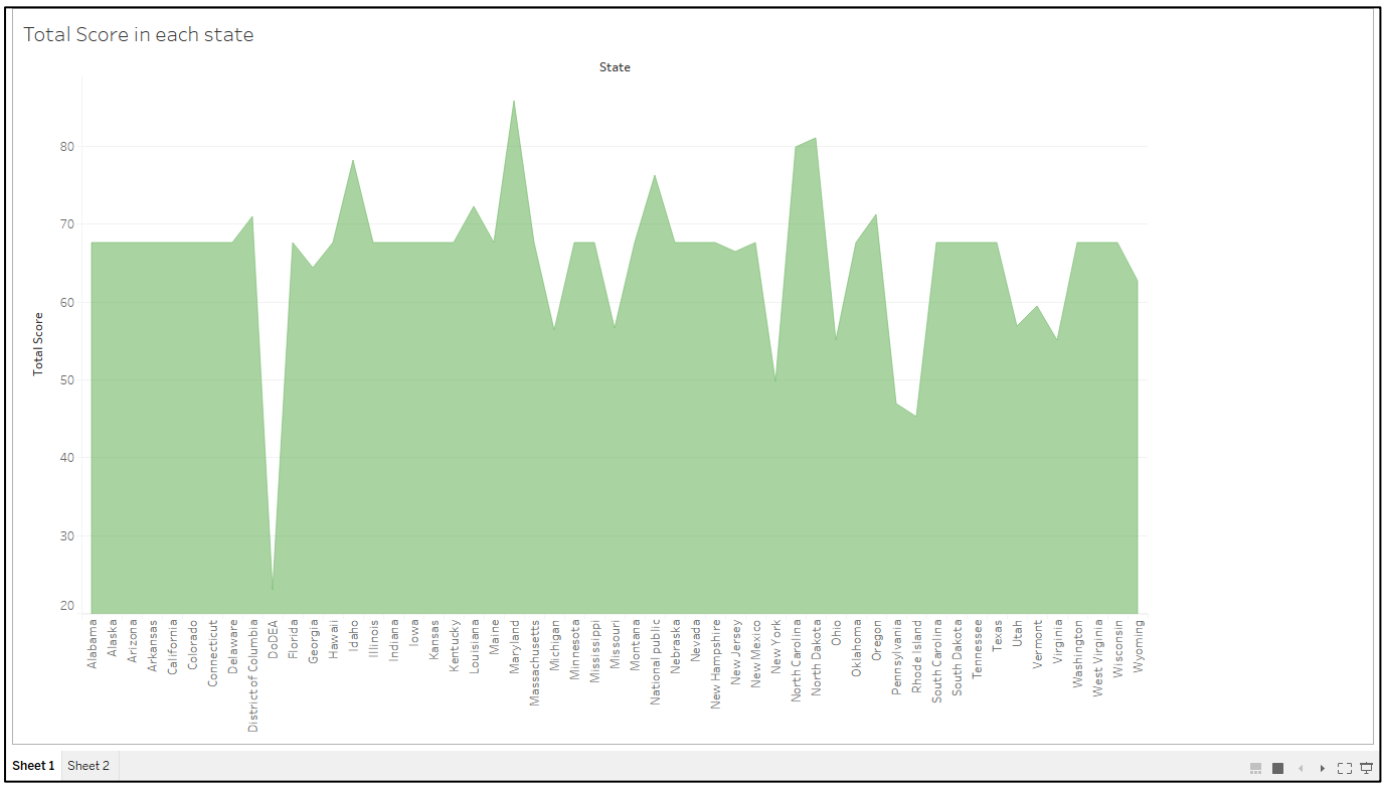
Number of Credits per City:



(Only a part of the plot is shown above as the whole plot is very big to fit in single page)

The above plot shows the credits required to complete the course per each city in the list.

Area graph (Total score in each state):



The above plot shows the area graph which describes the total score obtained by colleges in each state.

Conclusion:

- ✓ The dataset selected contains sufficient number of variables and columns to show variety of visualizations
- ✓ The dataset is preprocessed by cleaning the data set with R methods (Mean Imputation method to replace missing values) and removed duplicate values in columns.
- ✓ Modified dataset by adding values based on mean calculations with respect to the variables in the data set to produce variations in plots as desired
- ✓ Visualizations are made using a variety of methods using Bar plots, Scattered plots with regression line, Micromaps, Area graphs etc.
- ✓ Conclusions are drawn based on the comparisons made in each plot from the dataset.