Analysing educational investment for the government

Bharti Pandit

1.Introduction

Institutions and administrators regularly have to make difficult choices about how best to invest resources to serve students. Grants from federal, state, and philanthropic agencies provide additional funds for the development, testing, and evaluation of innovative undergraduate education programs, which, if demonstrated to be effective, often are expected to be sustained from other sources when grant funding ends. Yet the changing landscape in postsecondary education, including increasing enrolment, expanding access, and decreasing state-level investment, is putting added pressure on postsecondary education budgets. But economic evaluation, or the systematic analysis of the relationship between costs and outcomes of a program or policy, is relatively uncommon in higher education. This type of evaluation can be an important tool for decision-makers considering questions of resource allocation.
In this Project, I am focusing on the problem of USA government who wants to invest in various Data Science universities across USA. need to analyse and cluster various institutions to decide the structure of their investment and where should they invest more.

2.Data acquisition and cleaning

2.1 Data sources and Data cleaning

The dataset having the names and various other features of Data Science universities in the USA was found, and obtained from Kaggle. There were few problems with the dataset.
At first, this dataset had many columns with large numbers of missing data. Also, some data was not very useful or was in a complex format which, was difficult to be dealt with it. So, I decided to remove all these columns, which were ‘Time Data’, ‘Year’, ‘F M Ratio’, ‘Income’, ‘Country’, ‘World Rank’, ‘International’, ‘Prereq’, ‘Student-Staff Ratio’, ‘International Students’, ‘Link’, ‘Duration’. After this, the features that were left was 'School' which has the university name, 'State', 'City', 'NOC' which tells about the level of jobs, 'Program', 'Type', 'Department', 'Delivery' which tells about the mode of teaching, 'Loc-Lat', 'Loc-Long' which gives the latitude and the longitude of the universities respectively, 'Teaching' which gives the teaching score, 'Research' which gives the research score, 'Citations' which gives the citation score, 'total score' which gives the overall score of the university, 'Num_Students' gives the total number of students.

The Second problem was the delivery column which had multiple names for a similar kind of delivery mode of the program. Some were typed in the capital, some separated by commas, or written in reverse order. To address those issues, I decided to merge the similar types of mode of teaching. This was done by replacing the duplicate names with another common name so that the column can be used further.

The Third Problem was the column ‘Total Score’ which had ‘-‘ in place of ‘nan’ which was not in the correct format to be converted to float datatype. Hence, they were first converted to ‘nan’. Then the datatypes of the columns, ‘Total Score’ and ‘Num_Student’ were converted to type float.

After fixing all the above problems, I checked for the outliers and the data distribution of all the columns and found some noticeable outliers only in the ‘Num_Students’ column. I used this information for the data imputation of the 5 features which contained missing values. The columns ‘Teaching’, ‘Research’, ‘Total Score’ was imputed with their mean values since the data distribution was nearly symmetrical, while the columns ‘Citations’, ‘Num_Students’ were imputed by their median values due to skewed

data and presence of outlier. There were 32 columns ‘delivery’ and ‘Type’ which were having data in the qualitative format, which was also converted to the quantitative format.

2.2 Feature Selection

In Feature Selection, I inspected the correlation of the independent variables and found that the columns ‘Teaching’ and ‘Research’ (score of 0.93), ‘Research’ and ‘Total Score’ (score of 0.74), ‘Teaching’ and ‘Total Score’ (score of 0.73) were highly correlated. So, in order to remove the redundant data, I decided to remove one of the two features that were correlated that is ‘Teaching’ and ‘Total Score’

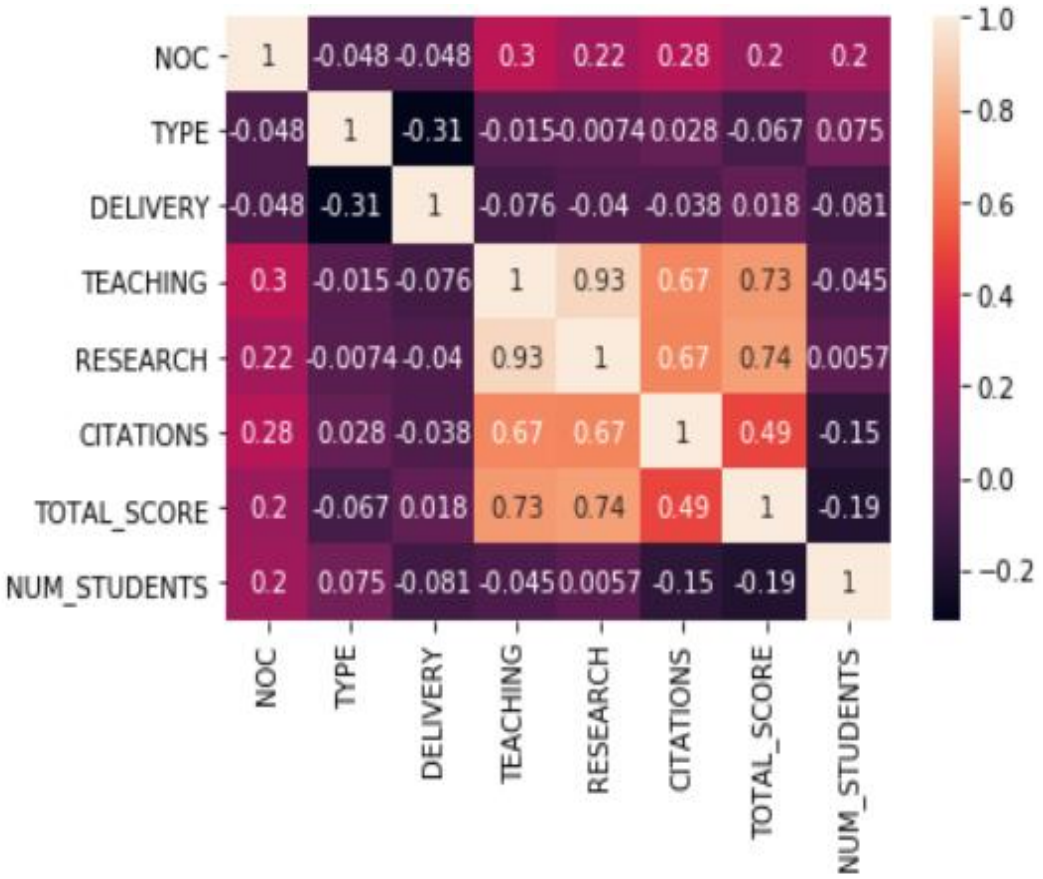


Fig1: Corelation Heat Map

3. Exploratory data analysis

The histogram of continuous variable and bar graphs of the categorical variable were plotted, and from these plots I observed that the universities mostly had a higher citations score, a variable research score and the number of students were in a lower range of 15000-20000. The Universities mostly preferred in delivering on campus program than the online mode. They mostly belong to category M and having NOC scores of 1 and 3. Most of the University were from the state New York and ‘Master of Science in Business Analytics’ program was offered in large number than any other program, mostly from Department of Statistics. We saw that there is a fairly linear relationship between research and citations, which is justifiable since a greater number of researches will give a greater number of citations.

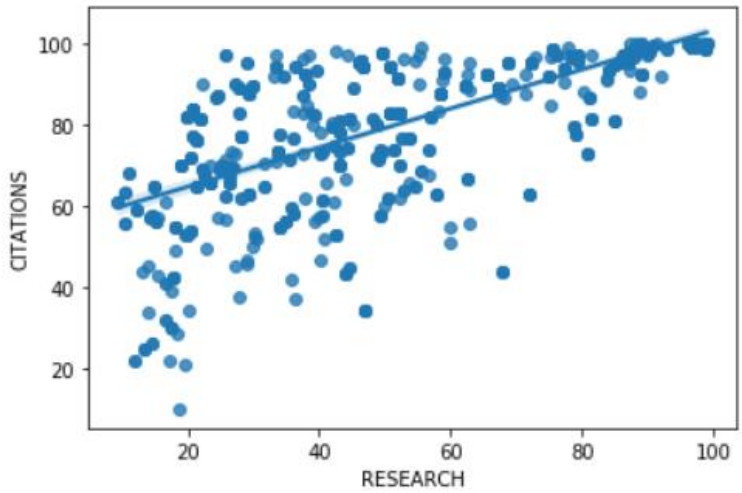


Fig2: Reg plot between Citations and Research

I plotted a box plot between the type of delivery and the research score. It is visible from the graph that research scores for on-campus and online modes were higher due to the outliers present. While the flexible programs with either online or on-campus mode along with the hybrid mode had the most distributed research score.

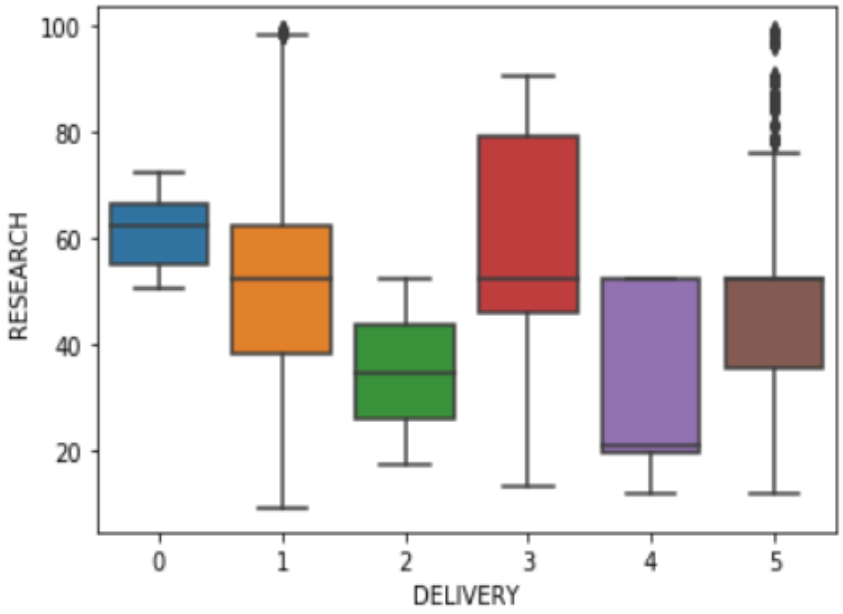


Fig3: Box plot between Research and Delivery

4. Predictive Modelling

There is no value that needs to be predicted while deciding the distribution of money and resources to various institutions despite we could decide the distribution by using some factors, for example, investing more in research in some institution than other might be a better decision. Hence, I decided to cluster the dataset and then look at the pattern of the different cluster, and decide what resources be distributed to which cluster.

First of all, I normalized all the dataset. Then, I applied k-means clustering to my dataset and clustered the whole dataset into 5 clusters. After which I predicted the labels of all the data that depended on the cluster they belonged to and merged the labels to my dataset.

5. Results

A new dataset containing all the information about the universities and the labels of the cluster to which they belonged was obtained. The five clusters consisted of the nearly same number of data points, except the 2nd cluster which had relatively a greater number of data points and the 4th cluster with relatively lower number of data points.

I also plotted a folium map using the latitude and longitude of the universities to get a better visualizing of the different clusters of university located in various cities which was found out by our model.

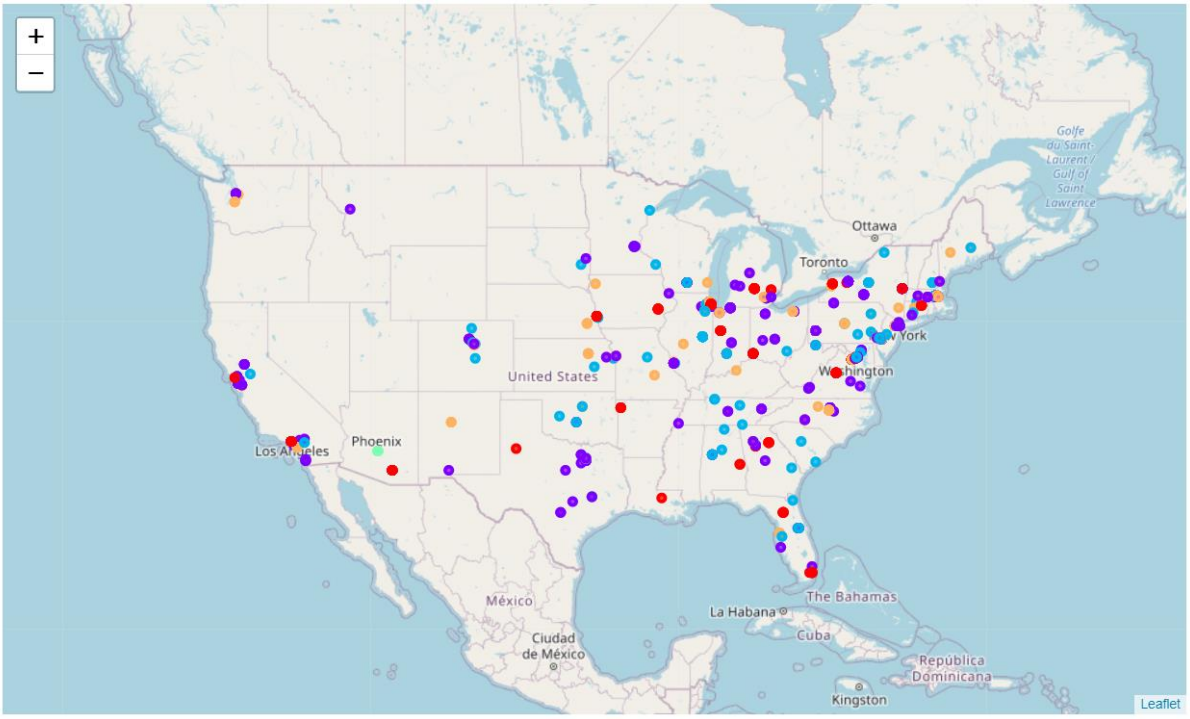


Fig4: Folium map of the clusters of the universities across USA

6. Discussion

The Predictive Model which I have made gave five clusters of various universities. All those five clusters where analysed separately.

The universities in the 1st cluster showed a higher number of on-campus programs but lacked in online, hybrid and other types of program delivery modes. It contained almost all types of NOC except NOC type 6 and 9. As this 1st cluster gave peaks at the research score of 15-20, and 40-45 and being nearly symmetrical in nature, it lacked in the research area. Most of the universities have a lower research score in the 1st cluster. But this cluster showed a good score of citations with a peak at 60-70, besides being slightly skewed to left. The number of students in the universities of this cluster goes up to 50000 with a maximum of them having students between 25000-35000.

Hence, for the universities in 1st cluster, more emphasis can be put on improving other means of delivering the programs for example, investment can be mode on developing online websites and improving the digitalization of the classrooms. More number of NOC of types 6 and 9 can be introduced. For these universities more investments in research can be made as compared to others, more funding can to given to research fellows, a greater number of research programs may be introduced or more apparatus for the labs may be provided. As more people will enrol in various programs and get more funding for their research, they will do their work with more interest which will result in earning more citations for the universities.

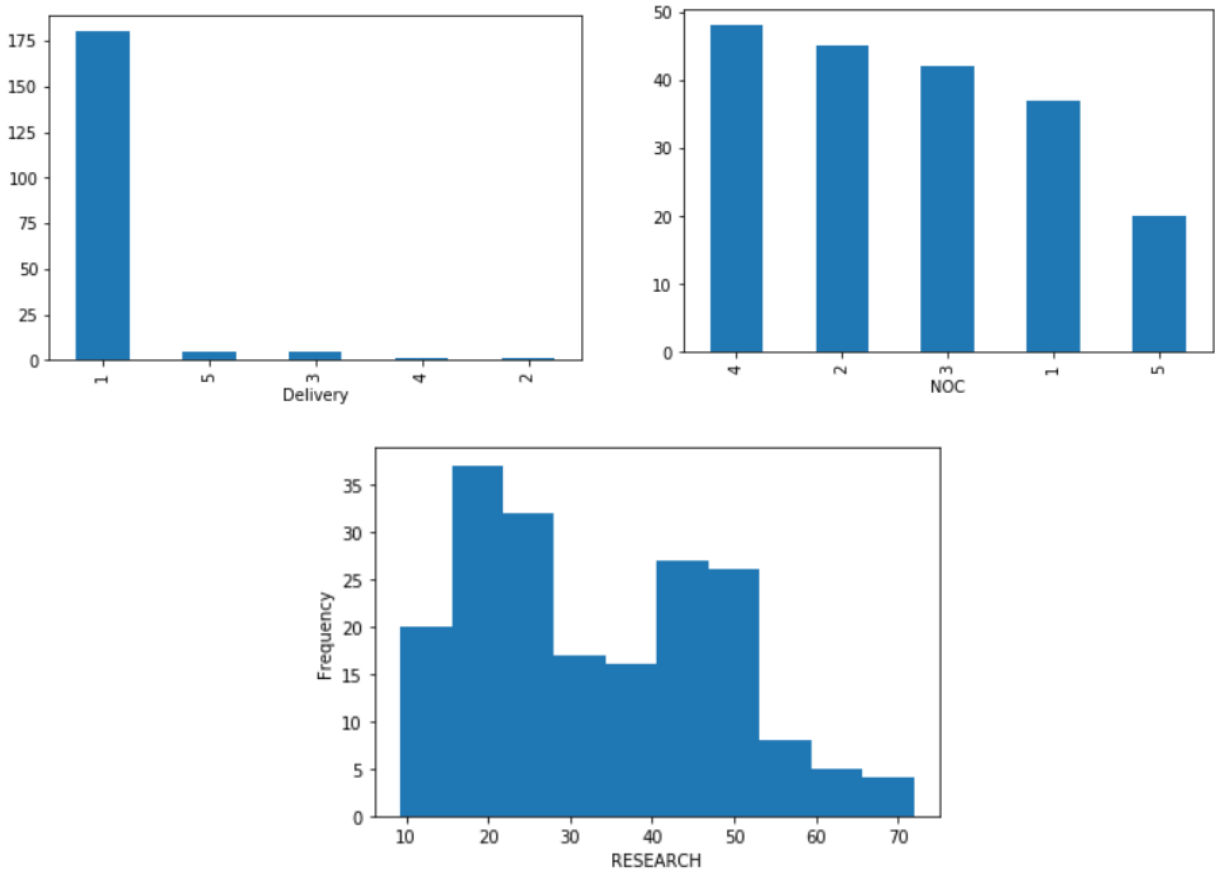


Fig5: frequency plots of various factors for cluster 1

The universities in the 2nd cluster also delivered their programs mostly on campus and lacked in other modes. It lacked in NOC of types 6,5 and 9, but also having largest of type 2 NOC. The research score in these universities varied up to 100, with most of them having a score between 45-55, which shows the research here is slightly in better condition. But the condition of citations in these universities is not as good, with most of them having a score between 80-85. These universities have a moderate number of students.

Hence, the universities in 2nd cluster may get investments for improving modes of education other than on-campus for example for online education and a hybrid mode where flexibility can be there between online and on-campus mode. They may be more focused more to increase the numbers of type 6,5 and 9 NOC. They have fairly good conditions for research so they may be given lesser funding in this area. But the citations are in poor condition which could be due to having fewer collaborations with research publishing websites or having less access to online platforms, so investment has to be made in this area for more improvement.

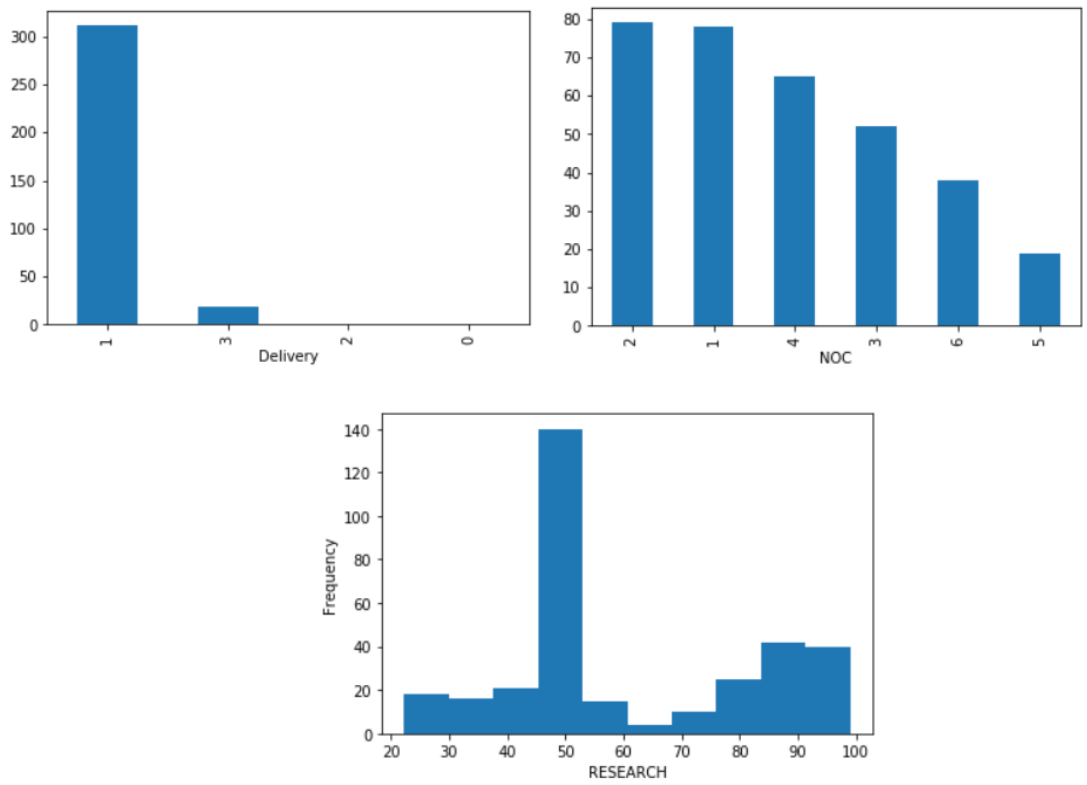
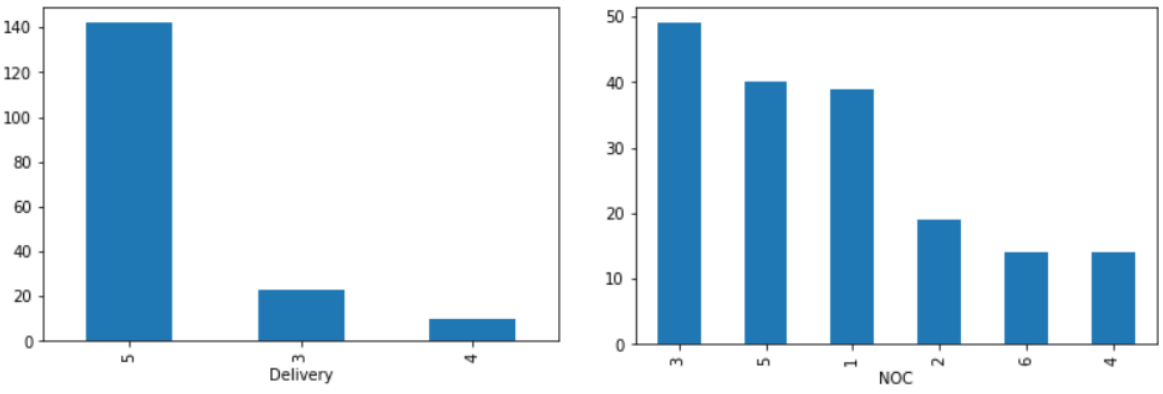


Fig6: Frequency plots of various factors for cluster

The 3rd cluster universities have the largest number of online programs and have some hybrid programs but lack the on-campus programs. They lag in the types 2,6,4 and 9 NOC. They also have a fairly good score for research, with most of the universities having a score between 50-60 and a fewer low score universities, but some having a score of nearly 90. Talking about the citations they mostly have a good citation score but few universities have a less citation score.

Universities in this cluster might get more investment to improve their on-campus programs, as they might be lagging in infrastructure and also it is clear from NOC that they have fewer types of NOC's so a greater number of faculties and other staffs should be recruited. Also, increasing some research labs might suffice.



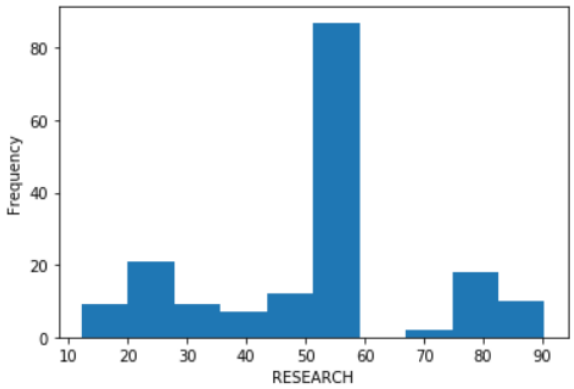


Fig7: Frequency plots of various factors for cluster 3

The universities in the 4th cluster are fairly good in number with on-campus programs but they lack in other modes of teaching. It has only two types of NOC's available that are type 9 and type 2. The research here is fairly good with most of the universities comes within a score of 50-80 and some under 50. These also have a good score for citations with most of them under a score of 85-95.

The universities under cluster 4 needs to target in improving their online programmes along with other flexible programs. More technology needs to be added to these institutions. These universities are in most need to improve their NOC, hence they need better faculty and staff members so that it would also increase the quality of education there. Some focus on research can be put for these institutions to compensate for those with a lower score in the research area.

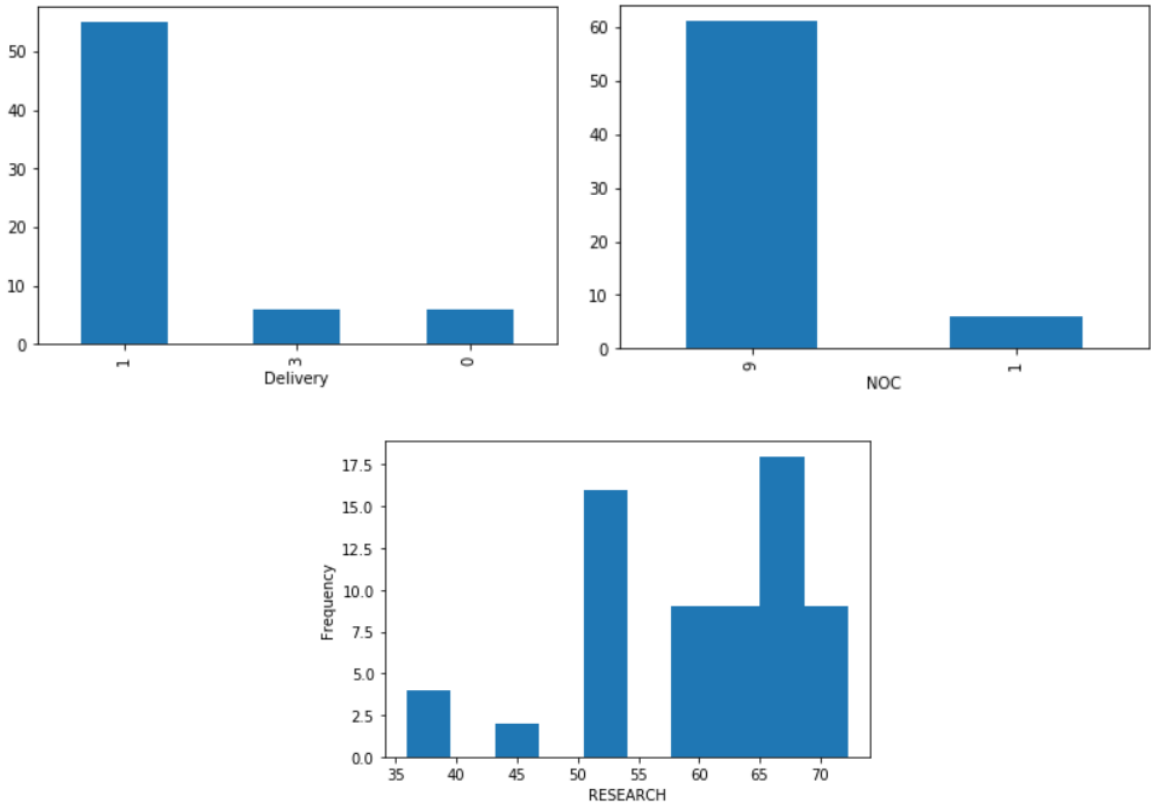


Fig8: Frequency plots of various factors for cluster

In the 5th cluster, the universities have a fairly good number of balance between on-campus and online programs and a few hybrids. They have less of NOC of type 9. They lag in research with most of the universities having a score between 45-50 and some also lesser. Although they seem to have a fair citation score.

For the universities of these clusters, some investment in more hybrid programs may be done. They are doing well with their faculties and staffs so lesser investment in this area might suffice. They lag in research and more investment in research would help them to elevate. Introducing some research programs with more funding in research is needed.

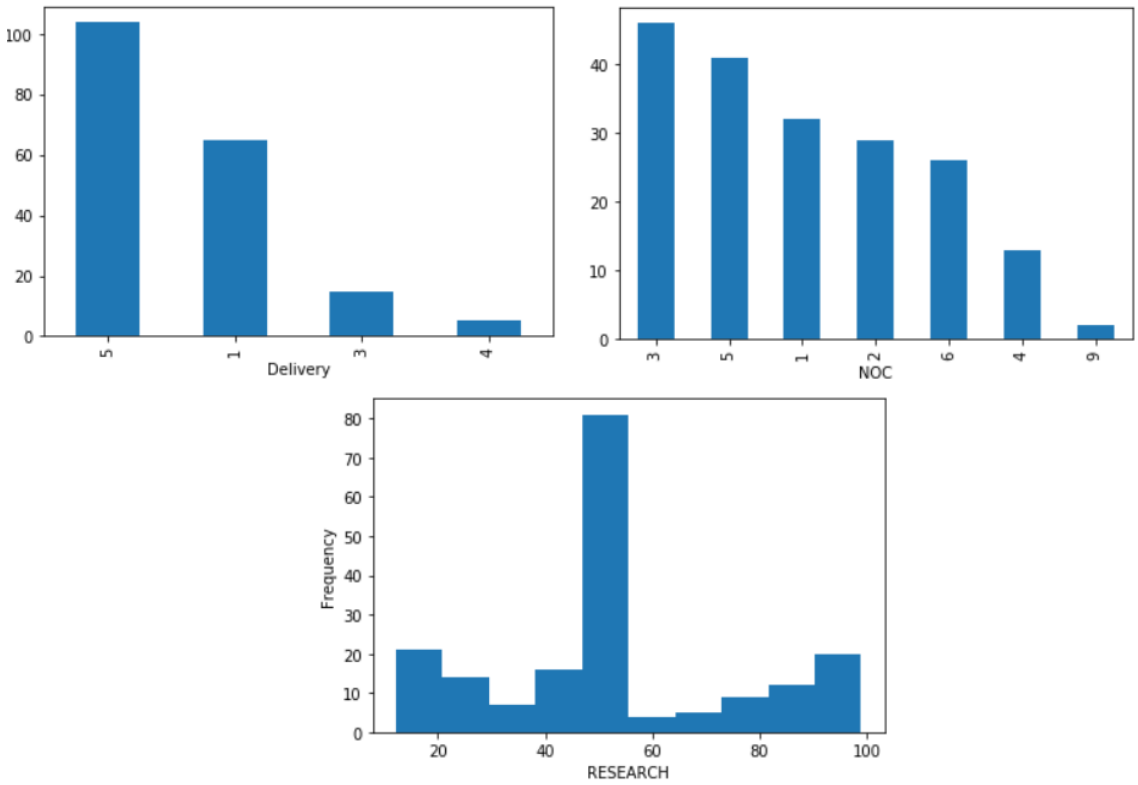


Fig9: Frequency plots of various factors for cluster 5

7. Conclusion

In this study, I analysed various Data science universities all over the USA based on many of their features and tried to structure a way in which the government should make investments in these universities. Features such as NOC, research score, delivery mode, citations and number of students were identified and used to group them. A clustering model was also built to group similar type of universities. These models can be very helpful for the government, while making any educational investment as they will be able to distribute the investment in a structured manner and make a more efficient and fruitful investment.