Pattern Study And Analysis Of Motor Transport Statistics:1969-71

Abhishek Bapna, Ananay Gupta, Md Waquar Wasif, Vatsayayn Binay
Birla Institute of Technology and Science, Pilani-Hyderabad

***Abstract* : The purpose of this paper is to manifest the different techniques which can be used for the mining of the data sets given to us. It will describe how we modified our data and obstacles we encountered while working on the data and the extent it was brought into change. Visualization techniques of data were also utilized and this paper will touch upon what type of plots were used and what was sought to be represented with these graphics. Then finally two of the data analysis techniques-classification and clustering were implemented to come to the conclusion.**

***Keywords* : Cleaning, Reduction, Aggregation, Normalization, Attributes, Data classification, Clustering, Matplotlib**

I. Introduction

This research project aims to analyze and showcase varied transportation statistics of the public domain in different parts of the country and the globe. To do so, we have made use of data from a government website (https://data.gov.in /) which contains values varying across different months and geographical boundaries in the years 1969 to 1971. However, owing to the numerous amount of CSV files present in the original data set we shortlisted a small set of CSV files and as it was not possible to begin mining data from the CSV files as they were, our first objective was to transform and reduce our data sets to make them usable.

The major stages involved in Data Pre-processing are- data cleaning, data reduction and data transformation. Cleaning involved removing or modifying records which had Outliers,non-numerical,N/Avalues.

Reduction and normalization were used to was used to reduce and scale the data accordingly. The techniques that had been used during the course of the project will be looked in detail in the next part of this report.

Progressively, we worked on the visualization of the data. This was done to draw comparisons between different time periods and locations from which the data was collected and to provide a more approachable way to understand the data. The entire project was coded in python using Jupyter Notebook.

The purpose of this task is to give an in-depth analysis of the state of general transportation and locomotion trends and practices during 1969-71. The data set comprises of multitude of transportation services across the country.

II. Nature of the Data

The 19 data sets were initially comprised of numerous features -Accident Reports(state-wise and country-wise), State wise breakdowns, State wise operation costs, Seating capacity and profitability of certain bus transports, Investment undertakings, Foreign vehicle production and utilization, nature and cause of accidents and collisions.

When working with data, we found that there were N/A values, specifically for the accidents and breakdown attributes leading to an unsupervised data collection. Since we wanted to keep the integrity of the data set we opted to remove certain data instances.We also decided to merge and create certain attributes based on these features for our own convenience

We had data pertaining to each of the 12 months for 1969 and 1971 meticulously divided into separate attributes.Owing to the complexity of the the data structure, we merged certain data structures into singlular CSV files to ease our work as the data present in the above mentioned sheets were quite comparable and we believe it would be easier to work with in the future stages of this project. There were certain data points which were in different scales of measurement which was noticeable when we tried to visualize the data The issue was resolved by simply multiplying the values in that column by a suitable figure.

The features present in the data sets were mainly numerical and continuous.

III. Pre-Processing of the Data

The data was read and stored in data frames with the help of Pandas. Data cleaning was performed on all the sheets. Data reduction and data transformation was performed on selected sheets wherever necessary.

A. Data Reduction

We have excluded all those attributes and fields of data which has very high occurrence of NA.And also those attributes which has almost negligible significance on the overall data. It was applied in almost all the sheets. We set some threshold value for the number or NA's and if any attribute was having more than threshold NA's, it was discarded because it might result in errors.

Moreover there were some redundant attributes. For eg. - in sheet 11.23, we were given no. Of buses of each service, seating capacity of each bus and total seating capacity of the fleet. The third attribute was redundant as it could be found just by multiplying first and the second attribute.

B. Data Cleaning

Data cleaning in our case included the removal of noisy data i.e. replacing all non-numerical values with NA and removal of redundant attributes.

We also filled the NA's in a particular position with the best and most suitable measure of central tendency i.e. mean, median.This was done by using the fillna() function of python. Usually NA was replaced by the suitable central tendency corresponding to the column. But in some cases, it is being replaced by the central tendency corresponding to the row as it made more sense. Data Cleaning is performed in every sheet at the very beginning.

C. Data Transformation

In many datasets, few attributes had scales which were different from one another and it was wrong to use those values for plotting. To overcome this situation, the attributes were normalized. the normalization technique was performed by either multiplying or dividing the data with proper values to give the desired result.

Normalization was performed on sheets 10.8 and 10.9 as in one sheet, it was no. of break downs per 10,000 kms and in another sheet, it was no. of breakdowns per lakh kms. So we normalized it to bring it to a common scale.

In sheet 11.23, data values were too large for proper visualization. So we normalized them for good comparison between attributes.

Aggregation technique was used in few sets of data. Whenever two or more sets of data with same amount and type of fields were encountered and the attributes of both type of data were related, we aggregated the data to draw some conclusion and give some beneficial inference.

We performed aggregation in 8.1, 8.3, 11.5, 10.8-10.9, 11.29 sheets. In sheet 8.1, we clubbed 6 different attributes into 2 attributes. In sheet 8.3, we aggregated 9 different attributes into 2 attributes. In these two sheets, attributes were basically different types of vehichles, we divided them into two attributes- Four wheeler and two wheeler vehichles.

IV. Data Analysis

After the pre-processing was done, we continued with data analysis part. We applied two of the major data analysis techniques:-

A. Data classification

This technique was performed in sheets 7.15 and 8.4 .

(1). In 7.15, we observed that that a major proportion of accidents were happening due to the fault of drivers and other things like mechanical defects, bad weather, defective road surface were not playing a significant role. And the proportion of accidents because of the fault of passengers, pedestrians, cyclists were also very less. So we classified the states into two categories based on the proportion of accidents caused because of the fault of drivers.

(2). In 8.4, data of country-wise people killed and injured was given. We classified the various countries into two categories based on how many people are killed in accidents among the injured ones in the years 1969-1971. We used the threshold of 5%. This reflects the medical services of the country and also their spontaneity. Countries where more than 5% among the total injured ones get ultimately killed, they need to improve their medical services so as to save the lives of people.

B. Clustering

We applied clustering in sheets 7.1, 7.8, 8.4, 11.28. In all of them we first used elbow method to find the suitable number of clusters. Then we generated all the possible clusters with that number of centroids and we took the one with minimum SSE.

(1). In 7.1 we clustered different months of year into different groups on the basis of number of accidents. We found that February belonged to category with very low number of accidents and whereas, January, March, April, May, June and December had higher number of accidents and the rest had moderate figures.

(2). In 7.8 we clustered different states into different groups on the basis of number of accidents. We found that Maharashtra and Tamil Nadu belonged to states with higher number of accidents and Andhra Pradesh, Assam and Rajasthan belonged to category with comparatively less number of accidents.

(3). In 8.4 we clustered different countries on the basis of the number of road accidents and number of deaths for three consecutive years, which will give us an idea about the kind of hospital facilities and traffic control in that country. The one with more deaths indicate poor hospital facilities and the one with more accidents indicate poor traffic control. We plotted these two on same graph as they are interlinked. There was also not much change in the pattern observed on those three years.

(4). In 11.28 we clustered different bus services on the basis of number of breakdowns and number of accidents. It helped us show a correlation between breakdowns and accidents, and the services which needed to review their bus's condition.

V. Visualization of Data

The visualization techniques were used to draw comparisons between transportation figures cross different months ,years and in location in a handful. The data was visualized using Matplotlib, with which we implemented a number of different graphs.

Since we are using a lot of varied data sets with different attributes and instances covering a wide range of topics it would be quite evident that we cannot utilize just a single type of visual aid. In our project, we have used Bar Graphs (eg.7.8,7.9), Faceted Bar Graphs(eg. 7.1,20), Pie Charts (eg. 11.5) Scatter Plot(eg. 8.1,8.4). While plotting certain graphs outliers were present in our data which were taken into account which will be quite evident by the fact that none of the graphs have the same data pre-processing code and algorithm. Many important conclusions were drawn from these graphs which are mentioned in the next section.

VI. Conclusions

Using the previously mentioned visualization techniques for data visualization and data analysis, we were able to make certain observations. They are all listed below.

A. Comparison of motor vehicle accidents through the years

1) *Comparing all states at Once* : A bar graph was plotted between no. of accidents across the year and the different states and UT's taken into account.

Accidents across several states were observed how it varied in 1969 and 1971. Over the 2 years Maharashtra and Tamil Nadu had the most no. of accidents towering over all the other states. A reasonable promotion would be that the officials were much more strict in recording what constitutes as an accident in the 2 states. Across the common states the number of accidents increased by a reasonable amount with a drastic increase in Jammu & Kashmir which almost tripled. Kerala remained reasonably unaffected and surprisingly the number of recorded accidents reduced in Delhi.

2) *Comparing the Data Across Month by Month* : We could not spot any major trends across the months which had reasonably stable numbers except for a few peaks and troughs in between which could be easily accounted as the 'tourist season' and 'off season' of the location.

3) *Comparing different countries:* From 1969-1971, in all the countries, number of accidents have increased by small percentage(around 5%). But in Canada, the number got almost 3 times and in Italy, number of accidents increased by 60%. This is an alarming rise in accidents in these two countries.

We observed that many people were dying out of road accidents in Poland and Ceylon as compared to other countries. Japan had the best medical facilities at that time as least number of people were dying because of road accidents there(only 1%).

B. Comparison of accidents defined by different class of vehicles : In Maharashtra and Delhi , we observed that maximum accidents were bus accidents. Surprisingly in many states , rickshaws were the reason behind a large number of accidents. Maybe in 1970, rickshaws was the most common means of public transport.

C. Cause and Effect of Accidents

By plotting a faceted bar graphs with states on the x axis and the number of defects in the y axis. The overloading of the vehicles came out as the leading cause of the accidents shortly followed by mechanical defects. Insufficient illumination came out to be the least reason for accidents.

Moreover we observed that in more than 50% cases, it is the fault of driver which result in accidents rather than fault of pedestrians, passengers, cyclists or some weather conditions.

D. Variation in production and utilization of vehichles in different countries.

This was identified with the help of a scatter plot. We were able to get a spread across the country and production of 4 wheelers and 2 wheelers.

Japan had the most produced 2 wheelers in both years and USA emerged as the clear winners for producing 4 wheelers.

But if we talk about the usage, maximum 2 wheelers were running in Japan in both the years. Maximum 4 wheelers were running in USA in 1969. But in 1970, the number decreased sharply and in Japan, number increased as a result of which in 1970, maximum 4 wheelers were running on the roads of Japan.

E. Public transportation operation costs and services

This was identified with the help of a horizontal bar graph and a pie chart.

The costs were least during 65-66 and reached their maximum during the years 70-71. Cost of materials consumed was the main reason of these costs.

Among all the bus services the district plain service has the maximum seating capacity which is quite evident as it contributes to the maximum expenditure and negligible expenditure in goods and transport.

Conclusion

Preprocessing and data analysis based on various techniques to make it more useful and appreciable. The visualization helped us in understanding what the data is actually about and gave us a rough idea of how the trends are going to be shaped.

Acknowledgement

We would like to avail this opportunity to thank Prof. Manik Gupta for giving us this chance to experience this project and her very informative lectures which were crucial for us in this project. We would also like to thank the teaching assistants whose help has been the most helpful and crucial.

References

[1]data.gov.in

[2]Lecture tools and resources