FUNDAMENTALS OF DATA SCIENCE

MA7419 Coursework 2

Team DABIb3

Lakshmi Mahitha Kodali, lmk22

Pranav S Asalekar, psa8

Srijeet Mohan Nair, smn24

Sahithi Devarapalli, sd592

Identifying and Evaluating Data Sources (Slide 2)

(Lakshmi Mahitha Kodali, lmk22)

Identifying Data Sources:

For a data analysis project the preliminary focus is the goal-oriented objective. Once we have a definitive objective the next step would be identifying its relevant data sources.

Many different sorts of data sources can be utilized in the analysis. Some of them are:

- Primary data sources - Gathered straight from the source, including census, tests, or direct observations.
- Secondary data sources - Gathered and made accessible for reuse by someone else, such as official statistics, scholarly works, or media stories.
- Tertiary data sources - Databases or digital portals that compile and consolidate information from both primary and secondary data sources.

The authenticity, trustworthiness, and appropriateness of the data should all be taken into account when determining the sources of the data. Additionally, you want to think about the particular kind of data (qualitative or quantitative) you require and the type that it is presented in.

A few methods for retrieving data sources are as follows:

- Exploring internet repositories and databases
- Trying to find research articles
- Using tools for data visualisation:
- Querying authorities

By considering all these factors in our mind, the data source used in this analysis is Stat19 data. Stat19 Dataset have all the road traffic accidents which were reported to police in span of 30 day of accidents with personal injury. Department for Transport (DfT) defined the fields and variables which are collected and SCRAS & ACPO committee has agreed.

Evaluating Data Sources :

In the qualitative research, evaluating data sources is crucial because it enables you to confirm the accuracy, dependability, and applicability of the data being employed.

Following elements are taken into account while assessing data sources:

- Source authority – How and by whom were the data gathered? Do you trust and respect the source?
  The Stat19 data source is completely valid and can be trusted as it's a government website. UKVI updates the dataset. For users to make a change or modify the data, one must create an account in www.data.gov.uk to proceed.

- Bias – Has there been any proof that the data's collection or analysis was biased ?

As the dataset entirely belongs to the government there would be no sort of biasness based on religion, color, caste, sex etc.

- Date – At what point was information gathered? Is the information still valid and up to date?
  Every September UKVI updates the data in-order to keep it up to date. So, whatever be the date and time of Stats19 data collection, the information would be accurate and valid.
- Quality – Is the data accurate and of good quality?
  As explained above the data would be 100% accurate and of good quality as it is entirely handled by UKVI office with no biasness and information updated every year.

Also, to analyse such huge data, we cleanse the data usually where we extract proper and even more accurate information from the bulk data by removing incorrect data, incorrectly formatted data, insufficient data, and duplicate data.

Ethical, Regulatory and Privacy Issues potentially encountered during a data project (Slide 3)

(Pranav S Asalekar, psa8)

Issues encountered while selecting data and evaluation:

Now, the Stat19 Dataset is published by UK government which is open for everyone. Stat19 annually releases anonymised records except sensitive fields. The full dataset is available is available on DfT or UK Data Archive which is available for certified researchers under an end user licence.

The open data is more transparency and with this there is less corruption as this data is accomplished by UK government anti-corruption policies. The one of disadvantage is missing data and data out of range. Other issue faced is data interpretation done incorrectly. Example here is analysing the gender which had more casualty in both year and it might be mistreated by viewers that the particular gender is worse at driving vehicle, which is not entirely correct.

We are mainly focusing on year 2019 and 2020 to analyse the pre-covid and during covid road casualties. We have used library stat 19 for Casualty and Accident Dataset. And used Vehicle dataset from UK government website in CSV format. Vehicle dataset for year 2020 was not correctly loaded from Library Stat19 as there were 2 files on the name of Vehicle and Vehicle-e-Scooter. Even selecting Vehicle dataset for 2020 in various, system was fetching by default Accident’s data.

Issues encountered while cleaning the data:

The columns from vehicle table contained only numeric values and to make it more sense, selected required columns and modified numeric codes to sensible values. For this we have taken help from Stat19 dataset guide.

Now Taking up the cleaning and maintaining code systematics, we removed non-required columns from the datasets. We dropped the column ‘Age of Casualty’ only contained ‘NA’ & ‘Data Missing’ records; dropped column ‘datetime’ which was combination of another present columns ‘Date’ & ‘Time’.

The data which had values such as ‘NA’, ‘Data missing or Out of Range’, ‘Not Known’, ‘-1’ were removed from the dataset. This data was not of any use for project. This made data with actual useful values and the size of dataset reduced. We tried to replace the all the not required values with value ‘Unknown’ but the some of the values still got projected as NA by the system.

To work on our questionaries, we used Inner join using common column ‘Accident reference’ and created two datasets for year 2019 and 2020, which brought up some duplicate records in data. we removed rows which were duplicate with help of distinct().

To make use of the data for questionaries, we built data frames df2019 and df2020. Most of the column names were sensible to use, though we have renamed it for personal dataset use. This does not make a big difference though but will be helpful to use it further if required.

Finding precise dataset which was related to accidents as most of the data was available on stat19 and the data which we had found but was not used due to not being free to use.

Adding Visualisation to the Data for analysis

(Srijeet Mohan Nair, smn24)

What is Data Visualisation?

Data Visualisation is the procedure of representing visual plots for the data in-order to more effectively comprehend, evaluate, and convey the information contained.

Use of Data Visualisation

Techniques for visualising the data come in a wide variety, and the properties of the data set and our objectives determine which one should be used. Here, the data set considered is the Birmingham – Wolverhampton Stats19 data for the year 2019 and 2020 which conveys the information about the casualties reported during these respective years in these two major cities.

2019 VS 2020 comparative analysis charts:

1. **Pie chart (Slide 4)**

The pie chart presented compares the statistics of casualties recorded in the cities like Birmingham, Manchester, Leicester, Sheffield, Leeds, and Wolverhampton during the year 2019 and the year 2020.

Clearly, the maximum number of casualties were recorded in Birmingham for both 2019 and 2020 with 2623 and 1801 casualties respectively. Moreover, when the respective 2019 and 2020 Birmingham records are combined with Wolverhampton records, the reported number of casualties covers half the total number of casualties with 3112 out of 6920 in 2019 and 2177 out of 4943 in 2020.

2. **Mosaic plot (Slide 6)**

To estimate the count of the seriously affected driver or rider, pedestrian casualties reported only during the weekends of the year 2019 and 2020 and considering the casualties to be a driver or rider, pedestrian, mosaic plots are illustrated. A mosaic plot is a unique variety of stacked bar chart. The rectangular splits that can be seen in the plot is with respect to casualty class.

For fatal casualties, in 2019, two road accidents were recorded, a pedestrian and a driver or rider, and in 2020 only 1 fatal casualty was reported and that of a driver or rider. In addition to this, in 2019, around 57 driver or rider were articulated as seriously affected casualty which is more than the double of serious pedestrian casualties, 22, whereas in 2020, 11 pedestrians were admitted as serious casualty and 44 driver or rider were addressed as serious casualty, that is exactly four times the pedestrians’ count.

A total of 490 driver or rider casualties and 52 pedestrian casualties were observed in 2019 out of which 432 driver or rider and 29 pedestrians had a minor effect. Considering the same point of comparison, in 2020, there were 326 accidents involving vehicles and 31 accidents involving pedestrians, of which 281 accidents involving vehicles and 20 accidents involving pedestrians had a minimal impact.

3. **Line graph on bar chart (Slide 5)**

For analysing the number of driver or rider, pedestrian casualties reported for each age band or group during the weekends of the year 2019 and 2020, a line graph on bar chat is created for each.

The age groups are specified in the x-axis and the number of driver or rider, pedestrian casualties in the y axis. The age group that resulted in maximum casualties were from 26 to 35, and age group 0-5 recorded minimum, for both the comparison years 2019 and 2020.

Furthermore, a steep inclination in the number of driver or rider, pedestrian casualties is observed from the adolescent age to the adults or the youth age and then from there a sharp declination is observed through the middle age to the old age. This trend is observed in 2019 as well as 2020.

One important difference that can be spotted easily is that in the year 2020, no age group recorded 100 or more casualties, whereas in 2019, one group reported approximately 150 casualties during the weekends and one group very close to 100 casualties.

4. Bar plot graph (Slide 7)

To analyse the direction of vehicle for number of casualties happened in year 2019 and 2020, a bar graph is generated for respective years.

In year 2019, we had the most of casualty for North and South direction of travel. The number of maximum casualties is 384. The least number of casualties was for direction East and west with count 300.

In year 2020, we the most casualty for happened where the direction of vehicle travel was East and West with the count of 144, while North and South had 100 records of casualty.

Here, the number of casualties happened in the direction of North and South decreased by 284 in year 2020 as well for East and West direction we can find the decrease by 256 records.

Covid might be the biggest reason for the decrease in the casualty happened overall.

Basic Conclusion (Slide 8)

An overall significance can be made from all the data visualisations carried out above that the casualties reported due to the road accidents in 2020 is less when compared to 2019 which make sense considering the fact that due to the rise of COVID in the year 2020 many restrictions, rules and regulations were imposed by the government.

Possible Future Work (Slide 9)

(Sahithi Devarapalli, sd592)

We estimated the casualty severity for the future forecast using the Stats19 package. The model used to generate the prediction is Multinomial Logistic Regression, which is partitioned into train and test data, with an accuracy average of 87% for train and test data. There is not that much variation between the two data sets, so we can believe that overfitting is not present in the data.

The programme has plenty of room to grow outside the network. Download, read, and format STATS19 data. The greatest potential lies in providing STATS19 data analysis tools to facilitate traffic research security. Many academic studies have been conducted based on this data; some of them are mentioned below to show the huge potential for future research.

Using research on the effectiveness of road safety regulations such as speed limits as an example, this study of 20 mph zones suggests that areas with 20 mph speed limits are safer. This raises the question of whether the same results can be obtained again using reproducible techniques. Does the conclusion we reached apply to the 20-mph zone, and what does this mean in areas where speed limits have not changed recently? I still don't know the answers to these questions. However, the study highlights some challenges researchers face if they wish to continue such research in the future.

1)Unravel the potential confounding effects of recent policy changes on crash rates after the changes are implemented and whether these effects persist over time and replicate the study in all geographic regions.

2) Investigate the effectiveness of other safety interventions using the same experimental design.

3) Identify the main risk factors for road accidents and study how these risks factors differed between regions with recent speed limit changes and no changes.

The estimated exposure rate is used to normalise the fall rate assessment (dangerous). The package's creator is an example of this type: according to research, the number of bicycle fatalities in Birmingham. This begs the question: are there any geographical comparisons? Inequality somewhere? What is the reason for the relatively high or low crash rate of a particular type? For example, what factors contribute to the difference in collision frequency between light vehicles and large vehicles? The answers to these types of questions are the most effective means of reduction, ultimately leading to improved road safety results for all road users. Furthermore, quantifying the specific effects of traffic calming measures can aid in future city planning decisions. The use of data in this exercise should ideally be extensive.

Covers enough streets and includes multiple streets' traffic and operating conditions; also, the data should be comprehensive (i.e., covering different types of roads) and cover a sufficiently long period of time to reflect the impact of policy changes implemented during this time.

The stats19 package, which makes open-access data on road accidents more easily available to researchers around the world, has the potential to support road safety research. Simplifying the data download and study clean-up stages may help facilitate repeatable on-site analyses. It also has the potential to enhance our understanding of road safety issues. This is especially important given that non-peer-reviewed or unpublished data are frequently used to draw conclusions in published studies.