# DABI_44_PORJECT_REPORT

20/12/2021

## Identifying and evaluating data sources

- “First of all, What exactly is data? Many people associate the term with a variety of meanings. (Take, for instance,”information” as it relates to your phone plan!) “Units of information observed, collected, or created in the field of study” is a reasonable description for our needs”. After deciding what to do in the project, the next step is to identify what dataset is needed for the project. The correct dataset should be selected to derive the needed result. An irrelevant dataset can lead to an improper answer and also waste time and effort. Research says most people choose irrelevant datasets which leads them to get wrong answers.
- The next step is obtaining the data needed for the project. The dataset needed for the project can be obtained from various sources like library databases, publicly available, government sources, and identifying data websites. Using these options, the datasets can be obtained for our project or research work. The dataset to be used for the project is the “road accident by causality “ dataset from stats19. The dataset can be taken from the government website or the inbuild package stats19.
- In our project, the dataset used is taken from the stats19 package. The dataset needed for the project is taken for four years namely 2017,2018,2019 and 2020. “Another thing to remember about data is that there is no such thing as a perfect source. As mentioned in Evaluating Sources, you’ll have to make educated judgments (inferences) about whether the data is suitable for your needs.”
- As said before, the evaluation part is where you should check whether the dataset is relevant. For example, if we take a dataset from a research paper we have to make sure that we read the entire article and then use the data in our project. Because by seeing the data it may look like it may be the relevant data. But when you read the description of it, you can know the entire features of the data. Then critical thinking is one of the skills which should be used here to verify the data. Critical means asking different questions like, In which the road type where the event happened? For this question, we have to check with the data whether it has relevant information for the questions to be answered. Likewise, we have to ask various questions which are to be answered in this project or article. And also you have to validate the data by checking for duplicate columns, empty cells, or NA in the dataset because this helps you proceed further and also helps you to know if you find errors or improper output while doing the project. Because if we don’t do this, we will face a lot of problems while doing our project and also get us the improper result. The last step in the data evaluation is the dissemination and reporting part
- AUTHOR NAME : Tarun Shekdar
- UNIVERSITY ID : ts386

## Ethical, regulatory, and privacy issues potentially encountered during a data project of this type

- Data collection has been carried out by the government of the United Kingdom for the past many decades. For the past 20 years, the collected data has been automated and digitalized rather than storing the data in paper format. It helps to reduce the manpower behind the data collection and has a great consistency between the records collected during an event. Data related to road safety data can be downloaded easily on the data.gov.uk website and it is made publicly downloadable regardless of what country the user resides. Any sort of data is information and can be termed as digital wealth living in a data-driven society. It can be used for abusive purposes.

- The United Kingdom has published a privacy note on stats19 on the government website [1]. The Department for Transport (DfT) is responsible for collecting the event's information and occasionally gathers personal or sensitive in regards to an accident event. Most of the personal information is not available in publicly accessible data. But every accident event is given a unique accident reference number. With the help of this unique accident reference number, the people involved in that particular event can access the whole data by providing some of their personal information such as their name, phone number, email address, full address, and the date and the time of the accident, etc.

- If someone could be able to taper the road accident's sensitive information of an event, the data can be sold to some private organizations for a price. Further, the organization can push personalized adware to that person who is involved in the accident. This adware can affect the person's psychological state and cause more mental stress to their life [2]. During my time with the stats19 data set and with the project work, I think the government should handle the data with more care and try to pass a strict law for the abusers tapering a person's sensitive information.

- Data can be used for many good purposes such as studying, analyzing, and researching. The analysis made from the stats19 annually could help the government of the United Kingdom to improve the people's standard of life. By detecting the root cause of the accidents can help the government to provide necessary restrictions or even improve the infrastructure of the road to prevent accidents occurring in an area. On the contrary, publicly available can be as dangerous as using sensitive information to an abusive activity and make the person's life more miserable. So, the government should constantly monitor the publicly available data to ensure it is free from any sensitive information and also should reinforce the regulations surrounding the people's details. Currently, the stats19 data set are validated annually by the government. Regulatory measures on privacy should be carried out in a consistent interval of time to ensure the ethics have been properly maintained during data collection and data analyses.

- AUTHOR NAME : Roshan ramesh

- UNIVERSITY ID : rrr13

## Designing the data visualizations used

- The data visualizations are designed based on the dataset for which we are doing data analysis. In this first part of the project, the visualizations are designed to understand the data clearly. Without understanding the dataset, there is no use in doing data analysis and also data analysis provides us a path to get a lot of questions that are to be found using the dataset.
- The first data looked at in this was the severity because that's a reason for doing this project, so the severities in the dataset are filtered and stored separately. The severities are ranked into 1,2 and 3 based on the severity rate after being met with an accident. The first visualization we used here is a pie chart each year individually. The reason for using the pie chart is because the dataset is categorized into three and the go-to plot would be a pie chart. The pie chart is created for different years based on their severity and they are visualized in percentage format to understand it more easily. And then a bar plot is also made to show the values of the categories to get a deeper understanding of the data. The reason for choosing the bar graph instead of the table is to make it more attractive to the data. The bar plots were then made advanced by joining multiple years' data for severity and made to make much clearer.
- The next visualization used was to answer the question " what are the days of week accidents take place?".For this question, the bar graph is used to show the count of accidents that happened on each day of a week in each year. This gave us the answer for the question which is the weekends ( Friday & Saturday) and in these those two Saturdays has the highest number of accident counts in each year from 2017 to 2020.
- The next visualization used was to answer the question "what are the road surface condition that faces more accident" which had three categories of roads. A pie chart is created for this question and the data are converted into percentage format. And based on the visualization, road type "1" has the highest number of accidents followed by road type 2.
- The next visualization used is done for finding the future rates of accidents. A regression plot is used for this visualization and the data for the regression plot is done by using the correlation created by the data of several accidents that happened in each year. And it showed us that the accident counts are in decreasing very rate which is a very good sign
- AUTHOR NAME : Sanjaykumar Thiyagarajan
- UNIVERSITY ID : skt18

### Possible future work on the topic using this, or other data.

- As we discussed, based on our trend analysis the accidents rate is in a decreasing rate from 2017 to 2020 which is a good sign. So the government and people are taking safety measures to reduce the accidents it's proven. But the decreasing rate is low and it should be increased from year by year. To achieve that people should follow the current rules and also be careful while driving. The government must also reduce the speed limit not only in the accident-prone zone. The speed should be controlled on all roads in the United Kingdom.
- Based on the research we made about the accidents happening, there are a lot of factors that cause accidents. For example, light conditions are a factor in causing an accident. From the dataset used in each year the more accidents happened when the drivers had more visibility. So, we must concentrate on all the factors.
- The data analysis we did here would be useful for people to know the days where accidents tend to happen more which informs them that the people should be careful on those days and the junction detail data analysis where we found the junction detail which has more number of accident. The speed limit should be controlled on the accident peak days and also the junction places where accidents happen in a large number followed providing a lot of safety signals on those places where the accidents happened.
- The next on this data analysis we did use the vehicle dataset where it showed us that the four-wheeler vehicle faces the most number of accidents in particular cars have more accident counts. To prevent this, the number of cars traveling should be reduced or the people should drive more safely. Driving the car in a safer cannot be done by the government, it is every individual's responsibility. But controlling the count of cars can be done by the government.
- The rules are used in a city called "Delhi, India" where the government has introduced new rules to control vehicle pollution. The rule is that it has allotted days for cars for traveling based on their number plate, that is they have separated the car numbers into two parts, one is even number plate and another one in the odd number plate. Each category of cars is allowed to travel only two days a week which resulted in a reduction in air pollution lot. This method can be applied here to reduce the number of accidents because more of several cars traveling leads to more number of accidents and lot of traffic problems. As far as now United kingdom stands at 5th position in less number of accidents but if we take still more efforts by creating new ideas will lead United kingdom to 1st place in future which will be done in our future project by finding out the best ways to reduce accidents.
- AUTHOR NAME : Rahul venukumar
- UNIVERSITY ID : rv85

## Reference

- [1]“Personal information and data protection - GOV.UK.” https://www.gov.uk/guidance/personal-information-and-data-protection (accessed Dec. 20, 2021).
- [2]“Ads blamed for childhood stress | Media | The Guardian.” https://www.theguardian.com/media/2006/dec/12/lifeandhealth.advertising (accessed Dec. 20, 2021).