A close up of a logo

Description automatically generated

**Week 2 Assignment - Qlik Visualization**

*Traffic Accidents in Cambridge from January 2010 to June 2016*

**Huskies Viz Team**

Siva Ramakrishna Sai Chand Chivukula

Thi Thai Nigan Le

Yang Wu

Bodian Li

80926 – Communication & Data Visualization - Spring 2019

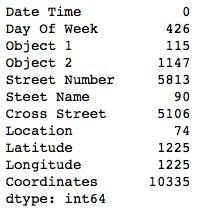
Instructor: Jack Bergersen (Prof.)

**Introduction**

Traffic accident is a concerning topic these days especially in the US. This paper would show the overview of crashes involving motor vehicles, bicycles and/or pedestrians reported in the City of Cambridge from January 2010 through June 2016. The dataset used in this paper is from Cambridge Police Department and visualized by Qlik Sense – one of the most common data visualization tools.

**Dataset overview**

From an overall perspective, the dataset has 12 variables with 10,335 records which describe the information about crashes in Cambridge. We figured out that there are missing or null values in the dataset which we can see below.



***Figure 1:*** The number of Null Values in Each Column

There are three types of information that can be exploited from the dataset including time (years, quarters, months, days of week, hours of day), reasons of crashes (object 1, object 2), location (street number, street name, cross street, address, latitude, longitude). Defining the kind of information combined with missing data can suggest what and how many insights that can be drawn from the dataset. Regarding of time, we can come up with the number of crashes by year, quarter, month, days of week, hours of day. This information can answer questions about comparison of number of accidents in Cambridge between years, months, and so forth; or they can be used to show the trend patterns with number of occurrences over a period of time. We can take advantage of these as they do not have any missing values. **Therefore, we thought it would be perfect to exploit the insights of data in terms of time.**

Related to accidents, the relationship is not clear between object 1 and object 2 as for example whether the object 1 hit by object 2 or vice versa. Besides, there are duplications in the types of object such as School Bus, Bus (Other), Bus seats more than 15 people. Therefore, it is hard to answer to question what vehicle or object causes the most crashes in Cambridge for a given amount of time. Moreover, there is nearly 10% null values in the object 2 which can cause the bias to our conclusion.

In term of location, “location” column shows the most detail information about location, but the inputs are not proper, the addresses are recorded manually without any format, therefore there are so many ways to define a name of a street in the column. For instance, the Garden Street is recorded as “GARDEN ST.”, “GARGEN ST” and “GARDEN STREET”. This leads to a strong bias in the outcomes. The “Street Name” is extracted from “Location” field, therefore it would cause the same problem as using “Location” column. Besides, we are not able to use “Cross Street” and “Street Number” as they have about 50% null values. “Latitude” and Longitude” are ideal to draw a map of the dataset. However, it is no avail to draw a distribution map of just Cambridge city because it does not show the differences within places. Overall, visualization of the location can have some bias and misunderstanding conclusions.

From the above conclusions we figured out that time would be best used to be deep into and questions of this assignments are built regarding data of “Date Time” and “Crash Number”.

**Visualize the data with Qlik Sense with research questions**

**Q: How many numbers of crashes are reported and documented in the given dataset?**

The data will be approached by the time order which ranges from year to days of week and hours of day. In total, there are 10,335 accidents happening in Cambridge from Jan 2010 to Jun 2016 calculated by counting the number of “Crash Number”. This single information can be shown simply by KPI chart in Qlik Sense with color. This number describe the information about traffic accident, therefore it should be emphasized to be a alarming number so we represented with red color . The tittle of the KPI can be underlined using “Text and image” chart in Qlik Sense which enable us to change font size and color of the text. The blue color can emphasize the red number i.e. total accidents.



***Figure 2:*** Total Number of Crashes in Cambridge from Jan 2010 – June 2016.

**Q: How many accidents or crashes occurred in terms of years?**

The total number would be separated by years i.e. 2010 to 2016. To compare the absolute values in a given period, bar chart is well suited as it has zero baseline. Red and blue is the best contrast combination to highlight highest values in 2011 compared to other years. Besides, steps such as removing gridlines, changing chart and axis titles, adding value labels are made to discard clutters. Subtitle is also added to note the features of the chart.



***Figure 3:*** Number of Crashes by Years.

This bar chart can answer the question which year has the highest number of accidents in Cambridge over the years. There is an increase in the number of crashes in 2011, then it drops in 2012. The number of crashes tends to go up between 2013 and 2015. In 2016, data only shows for six months, therefore it is not enough to conclude the trend of data in 2016. Overall, Cambridge witnessed the highest number of crashes in 2011 but in general the number of crashes ranges from approximately 1500 to 1700 (crashes per year) in a given period.

**Q: Which days of the week have a greater number of accidents?**

Line graph is the best to show the trend of the data by the days of week. The only purpose is figuring the pattern of the data, it is not necessary to add colors to the chart as it clearly differentiates the data. The same clutter steps are applied, and the subtitle is also added as above.

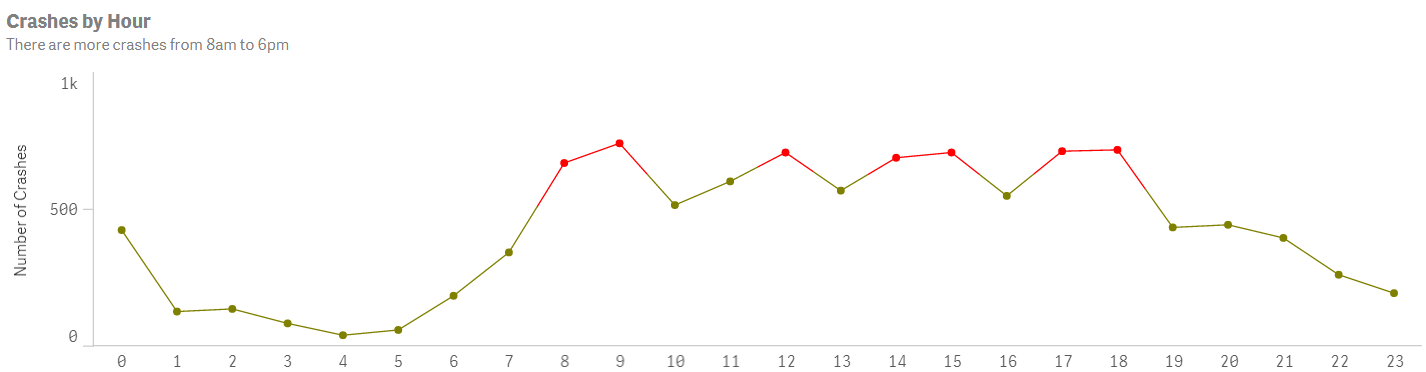


***Figure 4:*** Number of Crashes by Days of Week.

It can be seen from the graph that the number of accidents on weekends is significantly smaller than that of any day in the weekdays. The reason might be that people rarely go out on weekends than weekdays. We can also conclude that Friday have the highest number of accidents as people travel a lot to enjoy on this day.

**Q: Which hours have the greatest number of accidents in a day?**

Regarding hours of day, the pattern of crash can be demonstrated hourly by the line graph. There are 24 values in x-axis representing 24 hours in a day which can easily cause cognitive load to reader if we use any other graphs. Hence, line graph would be the best visualization for the number of crashes per hour. Gridlines is also removed, and similar formats are applied as above.



***Figure 5:*** Number of Crashes by Hours of Day.

The color is adjusted using expression in Qlik Sense. It would be turned to red when the count number of crashes is more than 670, otherwise the line is green. This makes it easier to see which time the high amount of crashes and they have are 8-9am, 12pm, 2-3pm, 5-6pm. It makes sense because they are the time frames when people go to and off from work or school, therefore there would be more crashes happening at these times.

**Conclusion.**

Let us conclude this by using dashboard which includes created charts that could give information about number of crashes in Cambridge between Jan 2010 and Jun 2016 annually, days of week and hours of day. Audience can figure out what are the major points over a given amount of time.



***Figure 6:*** Dashboard of Crashes in Cambridge from Jan 2010 to June 2016.