A close up of a logo

Description automatically generated

**Huskies Viz**

Siva Ramakrishna Sai Chand Chivukula, Thi Thai Nigan Le, Yang Wu, Bodian Li

Tableau Visualization

80926 – Communication & Data Visualization

Spring 2019

Jack Bergersen

**Introduction**

What is meant by Data Visualization? Data Visualization is nothing but the representation of the information in the form of a chart, diagram, picture, etc. Data visualization is crucial in the present world because as we have massive data, which is generated and generating, it is essential to understand the data for getting the insights. It is also necessary because the way the human brain interprets the complex data which can be easily represented with the minimum amount of cognitive load on the human brain. There are so many data visualization tools outside the world but why most people are fascinated about the Tableau! It is the fastest growing tool in the Business Intelligence industry because it helps in simplifying the raw data very quickly in an understandable format; we can create the visualizations very fast. The most important thing is that it is available with more than 51 database connections and recent versions are supported with almost all the databases in the world, which makes the Tableau most popular. The user interface is very intuitive compared to the remaining competitors in the sector. We cannot believe this, but it is right Tableau is recognized as a leader in Business Intelligence tools for seven consecutive years, which is impossible for any others. I know it is boring! Let us now dive into some of the basics of the visualizations in the Tableau; for doing these data is the crucial step. We will be using the “Cambridge Police Department crash dataset” for all visualizations.

**Analysis & Visualizations**

Data Analysis is nothing but the collection of 5 steps and those are:

1. **Define your questions:** we will be defining each problem in the next page of the document with an explanation.
2. **Set clear measurement priorities:** our priorities are Number of Records, Days of the Week and Date Time
3. **Collect Data:** Data is nothing but the Cambridge Police Crash Dataset
4. **Analyze the Data:** Explained each research questions in a detailed way under the questions
5. **Interpret the Results by Visualization:** Explained each visualization in a broadway under each research question.

Before proceeding further regularly doing the visualizations will make bore, so we had come up with an idea to explain each and everything by a research question which will cause us to do in a better way.

Data cleaning is an essential step for any Analytics professional, and it is analytically proved that 75% of the time is spent on doing the data cleaning. As we are talking about the visualizations, we will be representing the data cleaning in a broad way for all the visualizations.

A screenshot of a cell phone

Description automatically generated

***Figure 1:*** The General Summarization of the dataset

The dataset illustrates the crush number in Cambridge from January 2010 through June 2016. There are 11 columns and 10335 entries, which contain the float data and object data. Regarding the Date time, we found there is no missing value, and each data is identical. Furthermore, regarding the longitude and latitude, which contain many 0 amounts, indicating that the location was not found through their geocoding efforts.

A screen shot of a social media post

Description automatically generated

***Figure 2:*** The null number of each column

As we found before, there is no missing value existed in date time. Regarding the latitude and longitude, we can see there is the same number of missing values. Regarding the cross street and street number, the missing value is exceptionally high, which mean that the missing information in the street number usually doesn't have a specific name of the cross street. Most importantly, we focused much on "the number of records" and "day of the week," in the later part, we found that there are 426 missing values.

**Q: Are the average traffic conditions of each day on weekends are more than the average on each day on weekends?**

First, the necessary step is to clarify the context of the question. This question can be asked by customers who want to understand how the number of accidents changes within days of the Week. With the need to visualize the range and the trend of data by time (by day of Week), the line graph is the best option. The lines are applied to continuous data.

The line graphs are drawn showing a single series of data. It is crucial to eliminate the clutter which includes removing the border and gridlines, labeling the x, y-axis, the lines, and the charts.

**A close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generated**

***Figure 3:*** Number of records during weekdays and weekends

**Conclusion:** When we drew the visualizations for weekdays and weekends, we observed that a more significant number of accidents happened on Thursday and Friday when compared to remaining days in a week the reason might be people travel a lot during the end of the Week, for example, long weekends. When we drew for weekends, we observed that Saturday is having a higher number of accidents compared to the Sunday and it is evident because people travel on Saturday to enjoy their weekends. As a result, we had seen more accidents on that day. When we compare the below tabular results:

|  |  |  |  |
| --- | --- | --- | --- |
| **Weekdays** | **No of Records** | **In Percentage** | **In % for Week** |
| Monday | 1430 | 18% | 14% |
| Tuesday | 1533 | 20% | 15% |
| Wednesday | 1540 | 20% | 16% |
| Thursday | 1619 | 21% | 16% |
| Friday | 1616 | 21% | 16% |
| **Total** | **7738** | **100%** | **78%** |

***Figure 4:*** Tabular data for Weekdays in term of % in weekdays and Week

|  |  |  |  |
| --- | --- | --- | --- |
| **Weekend** | **No of Records** | **In Percentage** | **In % for Week** |
| Sunday | 1194 | 55% | 12% |
| Saturday | 976 | 45% | 10% |
| **Total** | 2170 | 100% | 22% |

***Figure 5:*** Tabular data for Weekdays in term of % in weekdays and Week

From the above tabular results, we can conclude that on average, it is very less on weekends when compared with the average for any two days in the Week. Our perception is incorrect with the data that usually a more significant number of accidents happens at the weekend.

**Q2:** **On what time does the higher number of accidents occur?**

The dataset shows the crashed vehicles in Cambridge from January 2010 through June 2016. In this case, we will use a stacked bar chart, line chart, and histogram to handle the data. A stacked bar chart is a graph that is used to break down and compare parts of a whole. Each bar in the table represents a whole, and segments in the bar represent different parts or categories of that whole. The stacked bar chart allows us to compare totals across category and see the subcomponent pieces within a given group

Let us focus on the stacked bar chart. As we can see, we separate the data into seven totals according to the days of a week. Then we divide each bar into hours. In this stacked bar chart, we can find that the crashed car numbers on the weekend are less than those on weekdays. By analyzing the crashed numbers of the hours in each weekday, we find that the most significant amount of accidents appears at 9 am and 5 pm. During these periods, most people are on their way to work or home, which will lead to massive traffic jam, increasing the risk of accidents. Also, all the bars show the minimum traffic accidents appearing from 1 am to 6 am. As we all know, most people are sleeping at that period. To some degree, this stacked bar chart can reflect human's behavior in every day. The police department should strengthen the road management of people's commuting time.

A screenshot of a computer

Description automatically generated

***Figure 6:*** Bar chart representation each hour in term of 0 to 23

**Conclusion:** From the above graph we can conclude that the higher number of accidents were happening during the hours of 8,9 and 17,18 and if we observe closely these are mainly occurring in the weekdays. Cambridge city is famous for the tech and pharmaceutical companies many people travel to their workplaces during these times. This would be the primary reason behind this.

**Q3: What are the trends in the crashes?**

In the third question, it is focused on showing the trend of the data by years. The context might be similar to that in issue one, but it is separated by different ranges of time (years instead of a day of weeks). Therefore, the line graph is also an ideal option for this question. However, there are slightly different advantages of using a bar chart, which make it a more efficient way to apply in this case. The bar chart has a zero baseline which does not apply to the line graph. The line graph focus on the relative position in space, therefore it is unnecessary to have zero baselines. While the height of each bar illustrating the actual values ranging from zero. These bars are put side-by-side, which enable us to compare the costs of each year. The bar chart is more comfortable for eyes to read.

The cluttering elimination process is also made to improve visualization in term of legends or labels. The width of the bar should be more comprehensive than the space between bars. In this graph, it is crucial to draw attention by coloring the bar. The hot red in 2011 demonstrates the highest number of crashes, and the shade will fade correlated to the number of accidents. In 2016, the green color illustrated the smallest amount of crashes compared to other years.

**A screenshot of a cell phone

Description automatically generated**

***Figure 7:*** Bar graphs with the trend over the years

**Conclusion:** Above trend line shows a positive sign that crashes are decreasing year by year**,** we represented the color legend in such a way the red color represents the highest number of accidents happened in a year while the green one represents the smaller number of crashes. We can conclude that countermeasures are taken by the respective departments for reducing the number of accidents.