4/18/2019

DATA VISUALISATION AND DASHBOARDING

COURSEWORK

HARSH KADIA  
MSc BUSINESS INTELLIGENCE AND ANALYTICS (FT)

UNIVERSITY OF WESTMINSTER

Table of Contents

[**RESEARCH QUESTION:** 2](#_Toc6445079)

[**ACQUISITION:** 2](#_Toc6445080)

[*Who created the data and why?* 2](#_Toc6445081)

[How has the data been compiled? 2](#_Toc6445082)

[What are the analysis implications? 2](#_Toc6445083)

[**PREPARATION:** 3](#_Toc6445084)

[**DISCOVERY:** 4](#_Toc6445085)

[**VISUALISATIONS:** 7](#_Toc6445086)

[What visual encodings did you consider? 7](#_Toc6445087)

[Which visual encoding did you choose and why? 8](#_Toc6445088)

[How have you used the software package tools and data visualization theory to create and refine your final data graphics? 8](#_Toc6445089)

[**NARRATIVE:** 8](#_Toc6445090)

[**APPENDIX:** 12](#_Toc6445091)

[**REFERENCES:** 21](#_Toc6445092)

## **RESEARCH QUESTION:**

*“What are the key factors for the number of accidents in the UK for the year 2014 which results in accidents based on severity (i.e. fatal, serious, slight)? The key factors (present in the dataset) may include rural/urban areas, accident severity across time, and the dimensions which contribute to accidents, accidents based on speed and the impact of light conditions on the time of day when most of the casualties occur based on the variables given.”*

## **ACQUISITION:**

### *Who created the data and why?*

The study focuses on road traffic accidents in the United Kingdom during the year 2014. The data has been created and published by the Department for Transport on data.gov.uk because of the demand was increasing up to date report information on road accidents for the stakeholders, public and researchers.

The link to the dataset is <https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data> . There are many datasets for Road Safety from which I chose Road Safety – Accidents 2014. For the supporting documents i.e. for the names of the measures in each attribute, I have used the XLS file named Lookup tables for variables which were added on 05 October 2015 under SUPPORTING DOCUMENTS.

### How has the data been compiled?

It is important to maintain the analytical and compilation quality for which reliable measuring techniques are the ideal key. The Department for Transport has gathered the data with respect to the factors which were contributing to the road accidents in the year 2014.

*How reliable can we expect the data to be?*

The Road Safety – Accidents 2014 dataset for used for my data visualization is from <https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data> which is licensed by <http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>.

### What are the analysis implications?

With data insufficiency, information of casualties and vehicles weren’t considered for the exact meaningful data for visualization of the accident severity. But now the data has been updated in the data.gov.uk and added another dataset especially for the variables having the names instead of numeric values relating to the unnamed values of the variables of all the datasets in Road Safety Data for the public, this will determine the severity of road traffic accidents in the United Kingdom more efficiently. In some cases the accidents might not have reported, so those accidents couldn’t be in the dataset.

## **PREPARATION:**

The most important factor in generating great visualizations is clean and great data. Data comes in many formats and quite often it already contains some formatting of presentation and layout that another software cannot understand. Therefore, I need to get the data as clean and raw as possible in order to better analyse the data and create interactive visualisations so that any other software can process the data.

I have checked 7 things before starting with my analysis.

1. Identify the table of your data.

Excel sheets sometimes contain multiple tables of data within the same sheets. This is not semantically correct, there must be one table of data for each sheet. We do not have to use multiple data tables to upload files/copy and paste data. Identifying the desired data table which will be used for visualization.

1. The table’s first row should be headers

The table’s first row should begin with the “headers” describing the columns. Remove or exclude unnecessary text from the introductory and other text.

1. Remove and merge multiple headers into one
2. Columns and detect data types

Use the dot ”.” for example 1200.25 as the decimal delimiter. Put the information in the column headers, not besides the number, if you need to indicate numbers in percentage or currency.

1. Clean the column values for any annotations

Numbers should be numbers only. The date should only be dates, etc. If you need to annotate a value, add it in a clear text in the field of visualization description. E.g. “Accidents value for UK 2014 does not include a Manchester”.

1. Annotations and blank cells

Sometimes for layout reasons, the data contains blank cells, they actually have the value of the cell of the parent category.

In other cases, having blank cells is legitimate, they denote an “unknown/unreported” value.

1. All rows must have same number of columns corresponding to the headers of the columns

Checking my data before using it, so there are no rows with more column headers than the actual column headers.

## **DISCOVERY:**

Exploratory data analysis is the process to get to know your data, so that you can generate and test your hypothesis. Visualization techniques are usually applied. I have done my Exploratory Data Analysis using R and using a package named DataExplorer which automatically generates the report for the Exploratory Data Analysis in the browser. Also, I have used packages dplyr, ggplot2.

First of all, after loading my dataset into R, I used glimpse function of dplyr package which is like a transposed version of print: columns runs down the page and data runs through it. This allows each column to be viewed in a data frame. It is a bit like a str applied to a data frame but it’s trying to show as much data as it can.

It gives output as number of observations as well as variables with the names and values as much as it can display. It is shown in the **Figure 1** in the Appendix.

Then I looked for the summary of my dataset with summary function which is used to produce result summaries of the different fitting functions on the model. The function invokes specific methods that depends on the first argument’s class. It shows the minimum value, maximum value, 1st quartile, 3rd quartile, median and mean of the each variable in the dataset. It is shown in the **Figure 2** in the Appendix.

After then, I used head() and tail() to see the first and last 6 values of each variables which is shown in the **Figure 3 (a)** and **(b)** in the Appendix.

Also, I have used boxplot to see the values in quartiles with mean and median. It is shown in **Figure 4** in the Appendix.

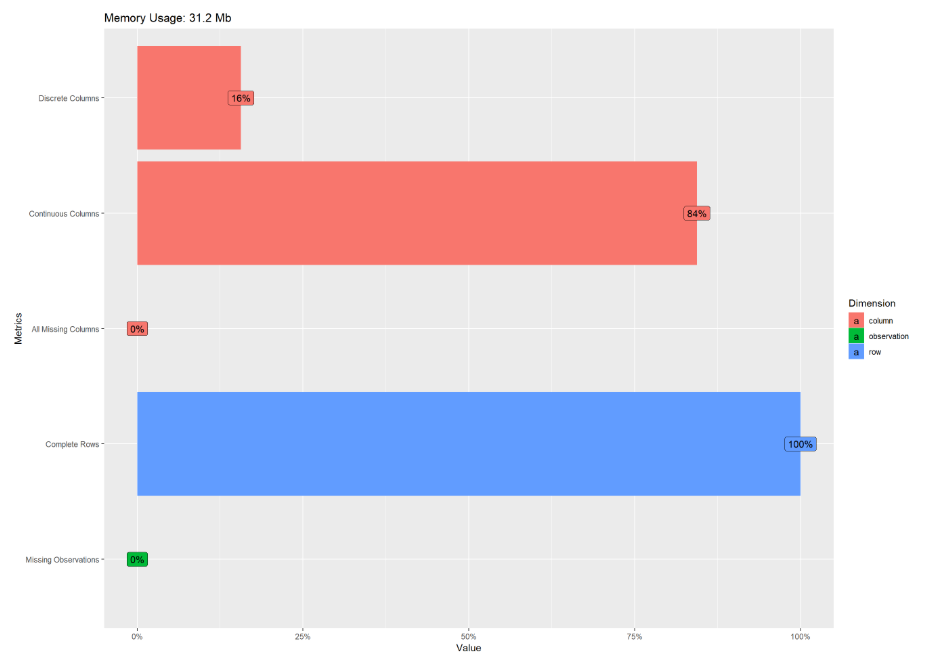
Moreover, I have used is.na() function to find out the missing values using its sum. But there were no missing values in my dataset.



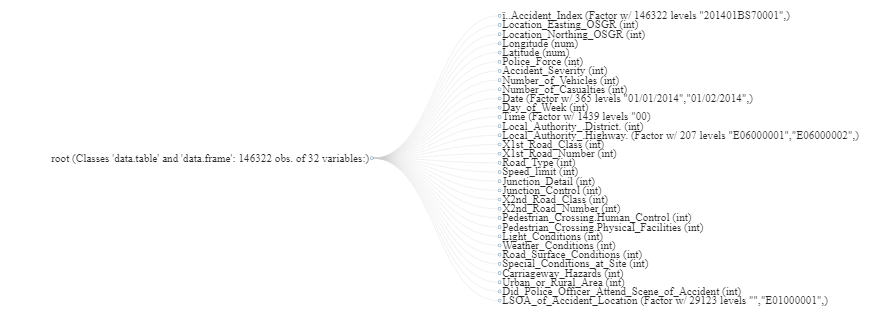
The most effective method for Exploratory Data Analysis was by using create\_report function of DataExplorer package, we can get Exploratory Data Analysis to get to know your data, so that you can generate and test your hypothesis. Visualization techniques are usually applied. The basic statistics of my dataset is as follows:



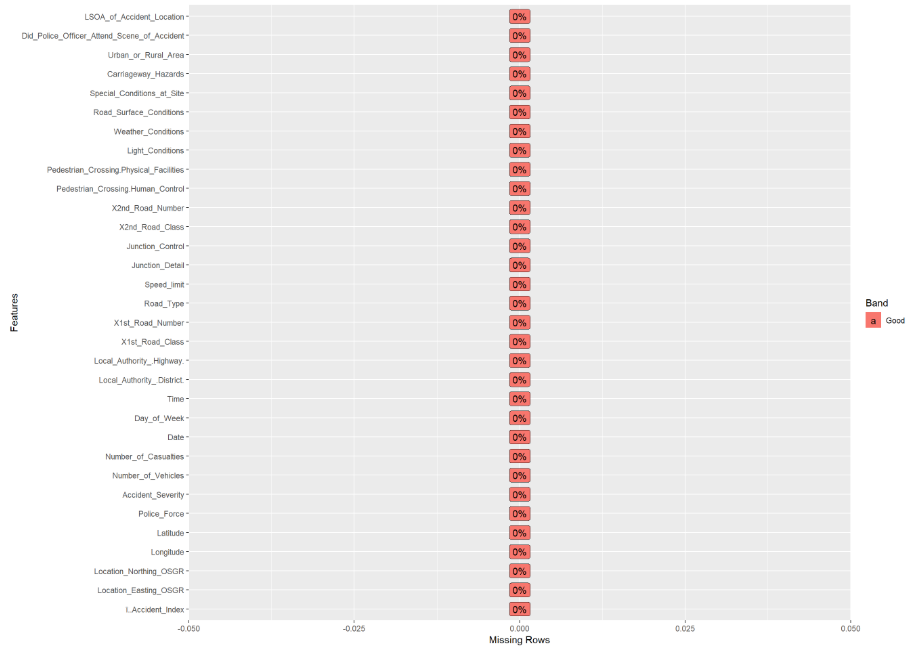
The percentage of discrete columns, continuous columns, missing columns, complete rows and missing observations are 16%, 84%, 0%, 100% and 0% respectively.



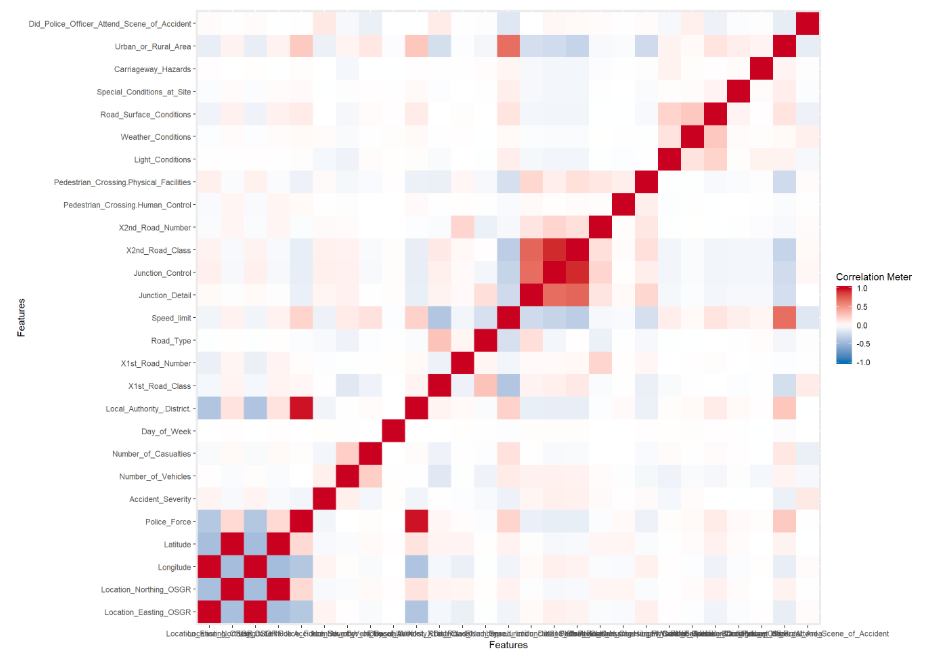
The data structure of my dataset is also shown. This is mostly used when you have integrated two or more dataset for your analysis.



The most important is finding out the missing vales in your dataset or else the analysis would not be accurate for analysis. There were no missing values in my dataset proven with the graphical manner.



The correlation analysis is used to study the strength of the relationship between two continuous numerically measured variables, it is useful when a researcher wants to determine if connections between variables are possible.



The link to the automated generated report from the create\_report function from DataExplorer package is [file:///C:/Users/harsh/OneDrive/Documents/report.html](file:///C:\Users\harsh\OneDrive\Documents\report.html).

## **VISUALISATIONS:**

### What visual encodings did you consider?

Visual encodings is a mapping of elements from the data to display. I have considered size, colors (brightness, saturation, and hue), orientation, shape, and color saturation. There is quite a bit of science about how our brains perceive and process a universal colour. Some colours may be lucky, some unlucky, some may be positive or negative, some may be associated with life events such as weddings, funerals, or new-borns.

Also, I have used points with line graph as I wanted to highlight the overall shape of the time series data, but also focus on individual values, the combination of points and lines works well. This is particularly useful when displaying multiple datasets across a time series, each as a separate line, and wanting to make it easy to compare individual values at specific times with each other factors.

### Which visual encoding did you choose and why?

The dataset I am using is a referring to quantitative values as such the best encoding to use would be the bar charts as they harness the principles of position and length. In addition to this, I have used line charts which also represents quantitative values through the principles of a slope and a connection. Color has been used to distinguish between factors of variables and also used to denote substantial values for my visualisation as it is effective for displaying quantitative nominal and ordinal variables which are all present in my data (Jock Mackinlay, 1986).

With the presence of geographic data, it was also suitable to use geospatial encoding as this harness the principle of the past experience which can give a spatial context to the user.

### How have you used the software package tools and data visualization theory to create and refine your final data graphics?

For the discovery phase, I used R which provided me with a flexible set of tools to analyse and manipulate my dataset. In this phase, I conducted Exploratory Data Analysis which generates hypothesis from the data instead of using data to prove hypothesis. The R scripts that I wrote generated an automated report that provided a look at how the data was distributed and split between categories.

Once the data provided me with a view of the types of patterns to explore, I switched to tableau to run my analysis. Tableau is a powerful BI tool that creates best practice graphics quickly. So, the analysis was conducted by splitting quantitative values (measure) by nominal values (dimension). In doing this, it is possible to see the way data is distributed in between nominal values.

## **NARRATIVE:**

I have created 8 sheets in total and 2 dashboards which gives answers to many questions as per the sheets such as:

1. *Which day of the week has the highest number of accidents segmenting them on the basis of urban and rural areas?*

Sheet 1 named Week explains about the number of accidents on weekly basis. I have added variables named Urban or Rural Area and Day of the Week in Columns whereas the Count of Accident Index in Rows. To visualise clearly I have added Day of the week to the Color inside Marks field.

The bar graph depicts that Sunday has the lowest number of accidents with the count of 9610 in the urban area. Whereas, Friday has the highest number of accidents with the count of 16007 in the urban area. Conversely, the lowest number of accidents with the count of 6330 in the rural area. Whilst, Friday has the highest number of accidents with the count of 7953 in the rural area. Overall, we can say that the highest number of accidents for both rural and urban areas occurs on Friday and the lowest number of accidents occurs on Monday.

**Figure 5** in the Appendix shows the sheet.

1. *Under what weather conditions do the most accidents occur?*

Sheet 2 named Weather explains about the count of accidents in weather conditions on the basis of accident severity based on 3 factors i.e. fatal, serious and slight. Fine no high winds (weather condition) has the maximum number of fatal, serious and slight accidents which has the count of 1409, 17,180 and 1,00,290 respectively.

Snowing no high winds (Weather Condition) has the least number of slight, serious and fatal accident with the count of 196, 24 and 3 accidents respectively.

**Figure 6** in the Appendix shows the sheet.

1. *How the date plays a role in determining the rise or drop in accidents?*

The sheet named TIME shows the number of accidents on a quarterly basis.

The line graph depicts that slight accidents are most volatile in nature and reach their peak in the second half of every quarter, however, records the lowest mark during the initial days of every quarter. Slight accidents reached its bottom to 170 on 4th day of the 3rd quarter while peaked to 3192 in the 4th quarter.

Serious accidents follow a trend similar to that of slight accidents but they are comparatively less sensitive. They have also shown a downward trend during the beginning of the quarter then followed an upward trend till climax which is then followed by a downward trend. Serious accidents peaked on the 17th day of the 3rd quarter and touched its bottom to 461 in 3rd quarter.

Fatal accidents remain stable throughout the time period.

**Figure 7(a) s**hows the first three quartilesand **Figure (b)** shows 2nd, 3rd and 4th quartilein the Appendix shows the sheet.

1. *What is the combination of values that contributes to most accidents?*

The sheet named Combination is the combination of dimension attributes of light conditions, road surface conditions and speed limit which contributes to most of the accidents. Here two factors Pedestrian crossing – Human control and Pedestrian crossing – Physical facilities are considered vs. Special conditions at site. The most number of accidents occurred by having No physical facilities within 50 meters which has an accident count index of 115,092.

**Figure 8** in the Appendix shows the sheet.

1. *What speed the number of casualties is the highest and at what speed the number of casualties is the least.*

The sheet named SPEED shows the relationship between number of causalities and speed of vehicle.

It can be easily construed from the line graph that vehicles with speed limit of 30 recorded 102,908 which the highest number of causalities however speed limit of 30 recorded 214, the least number of causalities.

Albeit causalities in fatal accidents were the lowest, but they reached its peak at the speed of 70. However serious accidents and slight accidents reached climax at speed of 30 and touched nadir at the speed of 20.

**Figure 9** in the Appendix shows the sheet.

1. *Using the latitude and longitude dimensions, highlight the regions of the geography that has the least and highest severity of accidents.*

Basically, it is seen that the average number of accidents occurred in the region of London having fatal, serious and slight, all types of accident severity. While, Shetland has the lowest average number of accidents for all types of accident severity (fatal, serious and slight).

**Figure 10** in the Appendix shows the sheet.

1. *What is the number of vehicles and number of casualities are directly proportional to each other?*

The sheet named Vehicles is a mixed graph showing relationship between number of vehicles and number of causalities on each day of the week.

It can be easily observed that number of casualties are in direct proportion to number of vehicles because causalities count follow the exact same trend as that of vehicle count. Both number of causalities and number of vehicles marked their lowest record on Sunday with values 22544 and 28365 respectively. However both the variables showed an upward trend till Tuesday after which a marginal downfall occurred till Thursday followed by a sudden rise on Friday to climax at 31613 causalities and 44298 vehicles, after which both collapsed on Saturday.

**Figure 11** in the Appendix shows the sheet.

1. *What is the impact of light conditions on the time of the day when most casualties occur?*

It is delineated that during daylight around 5 PM in the evening where the most number of casualties occurred having the sum of 11,446. Also, during darkness – lights lit the most number of accidents occurred at 6 PM in the evening with the sum of 4,620. Moreover, during darkness – lights unlit, the most number of accidents occurred at 5 PM in the evening with the sum of 105. During darkness – no lightening the most number of accidents were occurred at 5 PM in the evening with the sum of 1,257. Additionally, during darkness – lighting unknown, the most number of accidents were occurred at 6 PM in the evening with the sum of 277. Overall, the maximum number of casualties occurred in between 3 PM to 6 PM during daylight.

**Figure 12** in the Appendix shows the sheet.

My size of the Dashboard is Desktop Browser (1000px \* 800px) in visualisation.

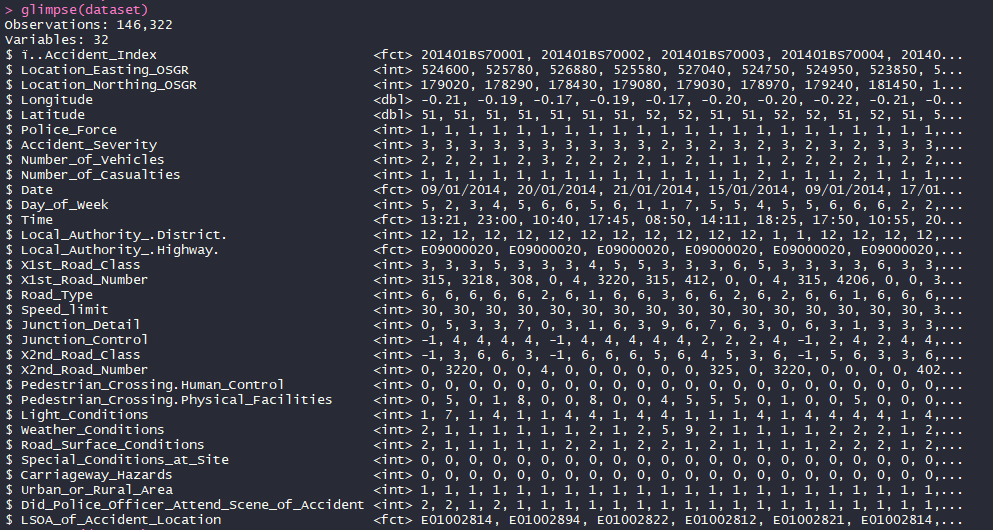
Dashboard 1 explains about the Statistics of a number of accidents and casualties at a day and time of the day level. Briefly, it delineates that most number of accidents and casualties occurs on Friday in between 3 PM to 6 PM during daylight in rural as well as urban areas with the most number of vehicles. **Figure 13** in the Appendix shows the dashboard.

On the other hand, Dashboard 2 describes the major three factors involved in most accident causes along with speeds and geographic locations of all the accidents. Vehicles with a speed limit of 30 recorded the highest number of causalities however speed limit of 40 recorded 214, the least number of causalities.

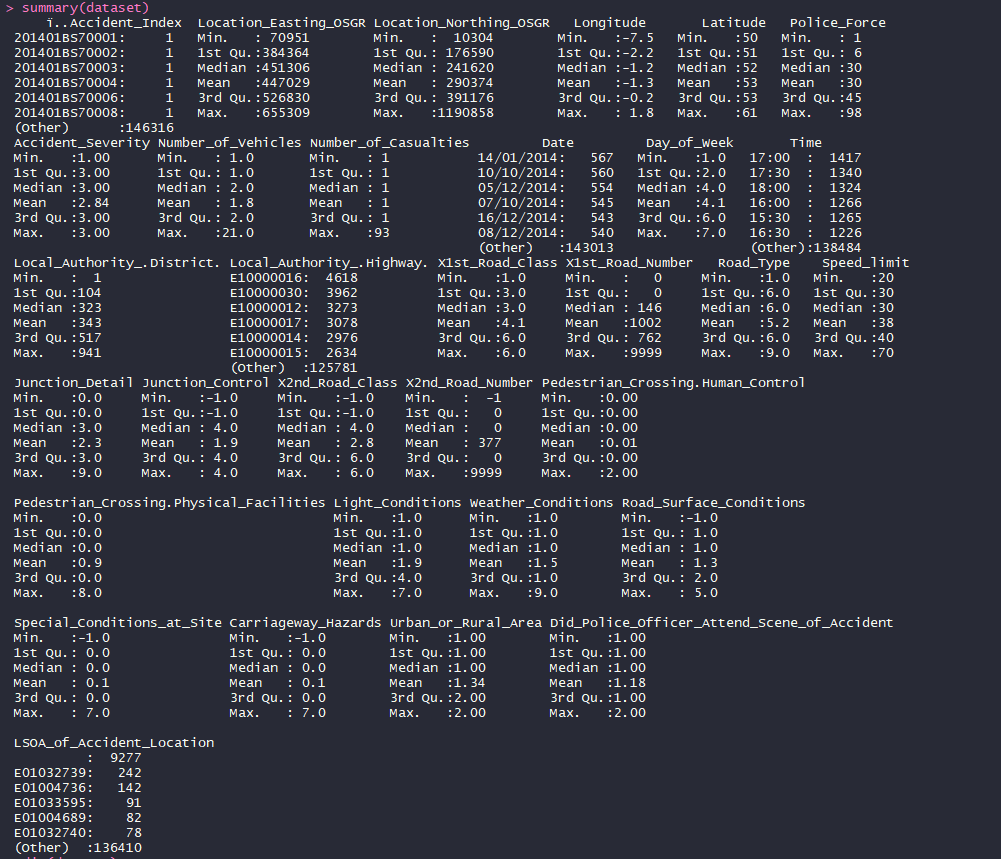
Albeit causalities in fatal accidents were the lowest, but they reached their peak at the speed of 70. However serious accidents and slight accidents reached climax at speed of 30 and touched nadir at the speed of 20 in the London region with accident severity of fatal, slight and serious. The major cause that it relates is that due to no physical crossing facilities for pedestrians within 50 meters of the road which resulted in the most number of accidents for a count of 115,092. **Figure 14** in the Appendix shows the sheet.

### **APPENDIX:**

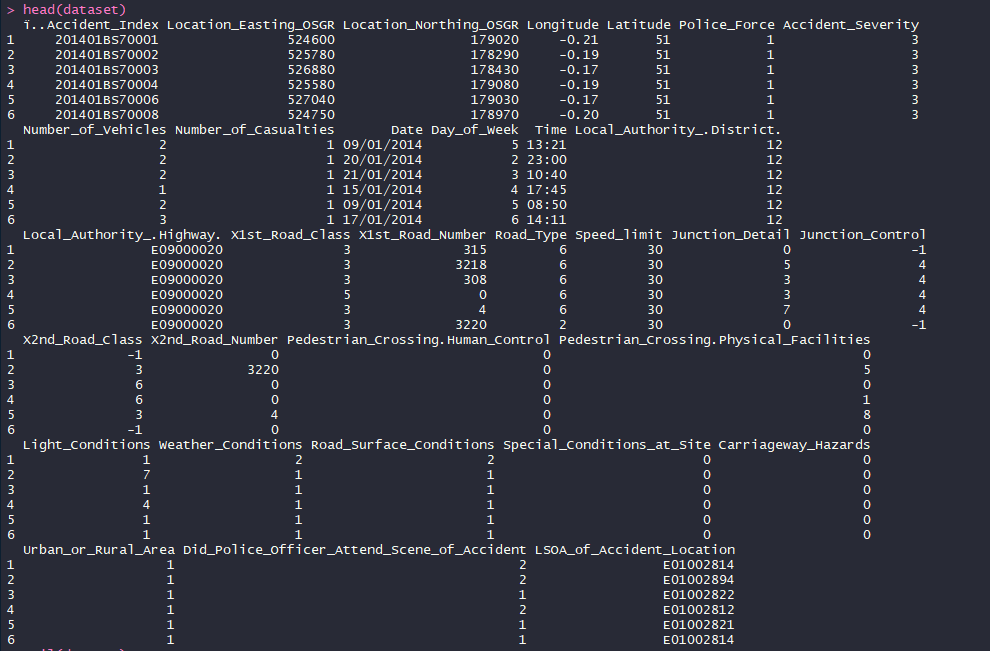
**Figure 1:**



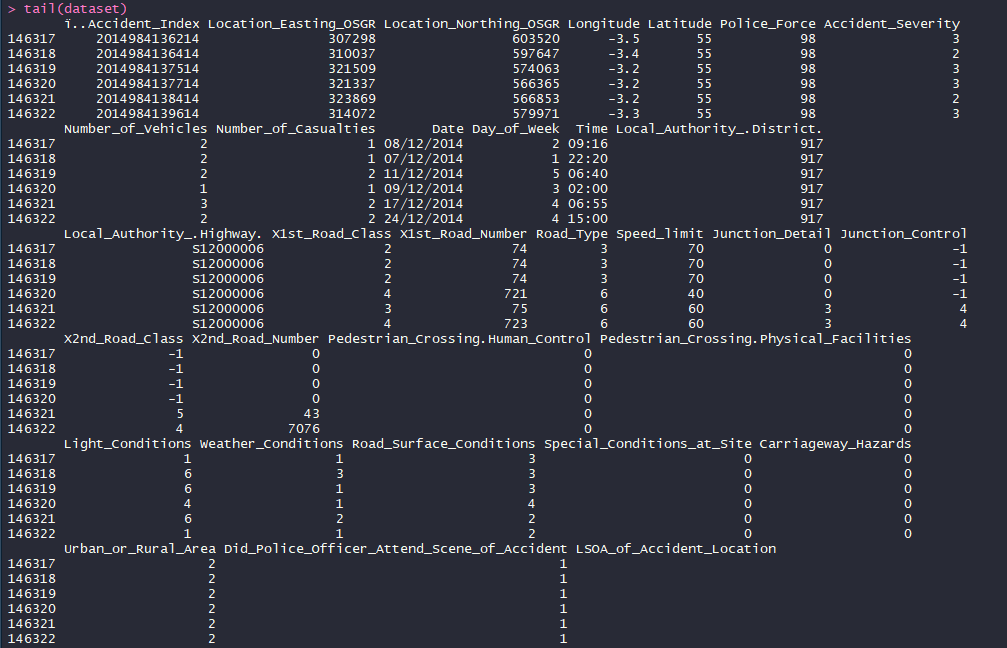
**Figure 2:**



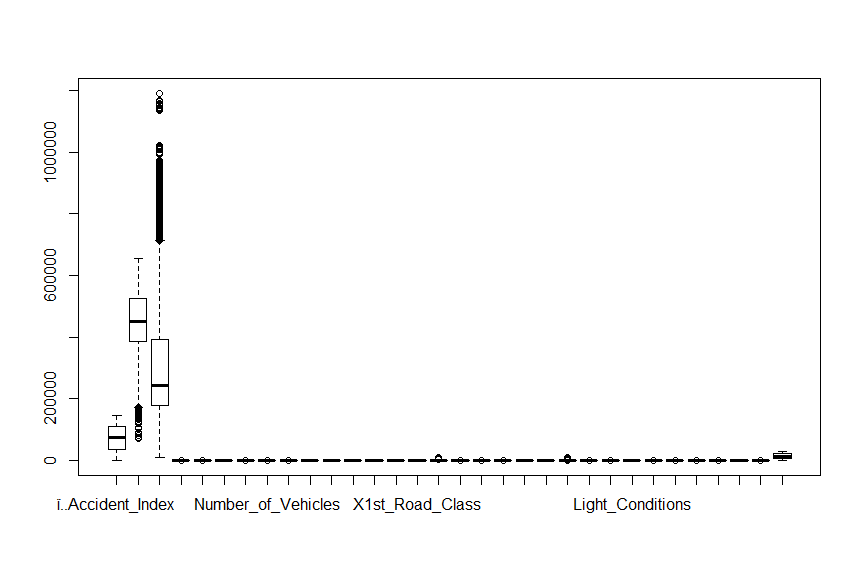
**Figure 3 (a):**



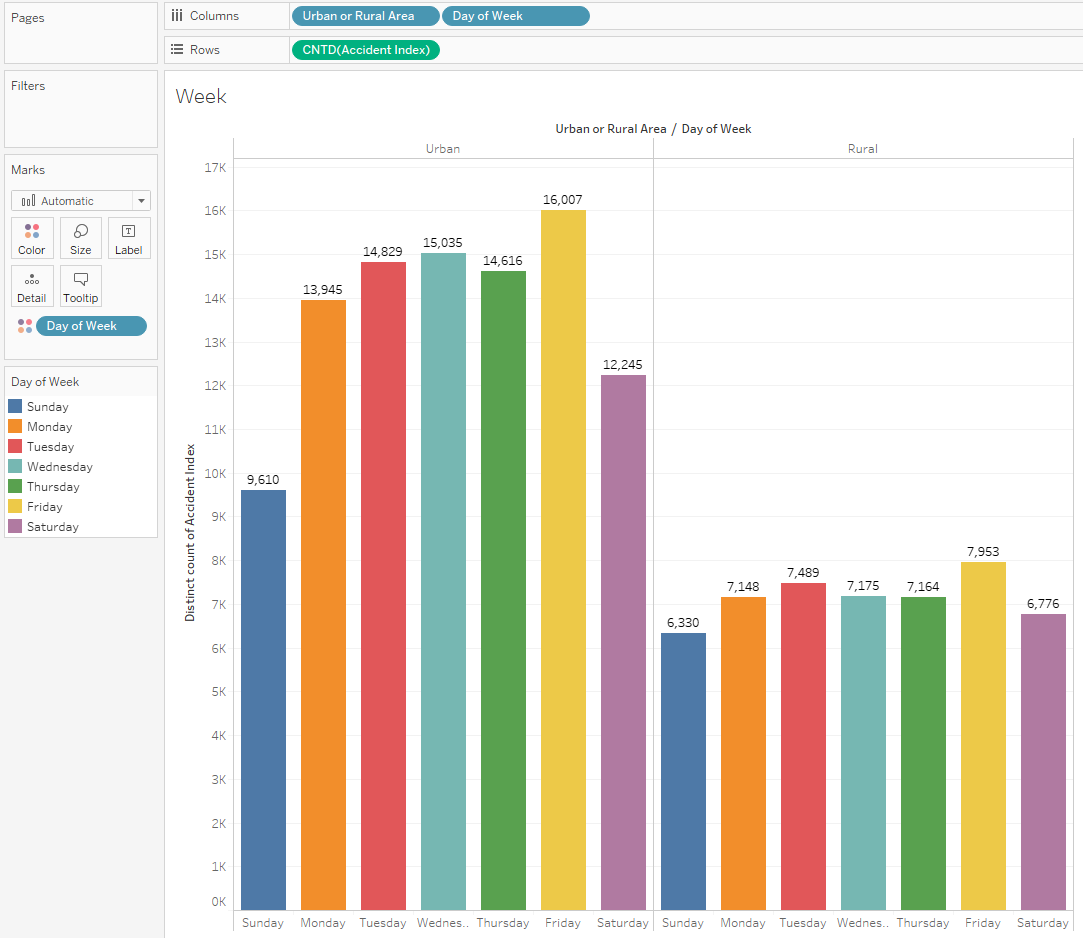
**Figure 3 (b):**



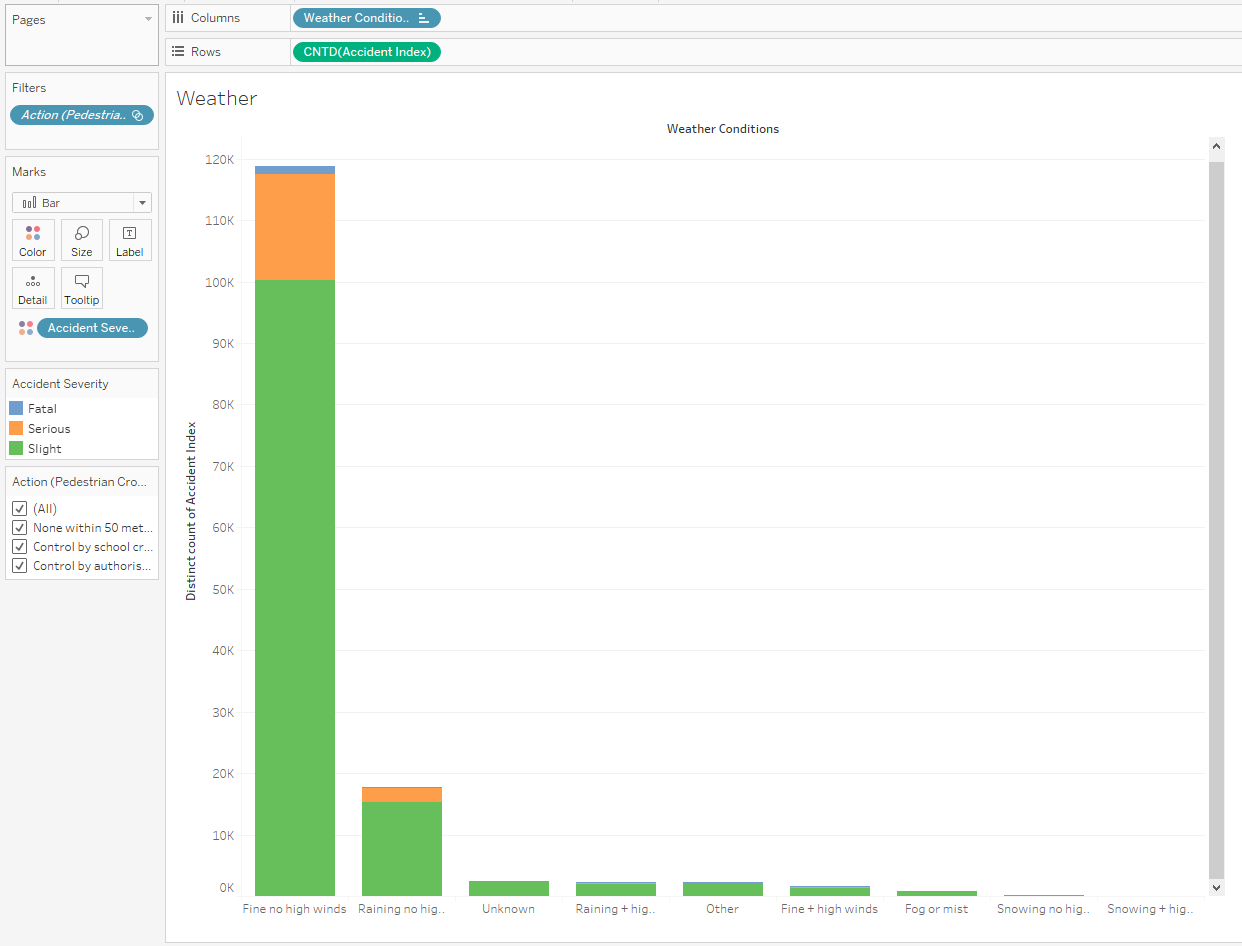
**Figure 4:**



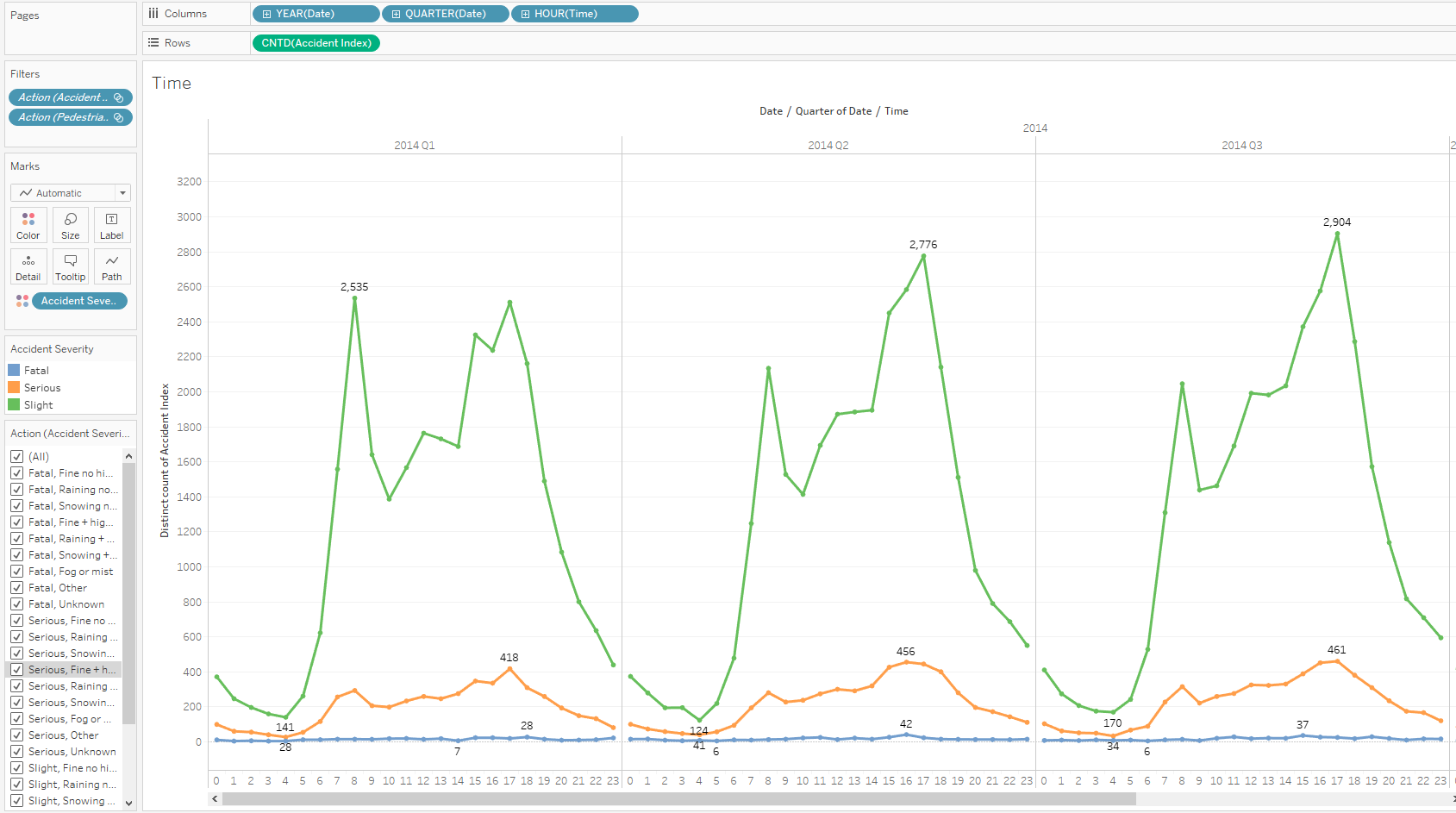
**Figure 5:**



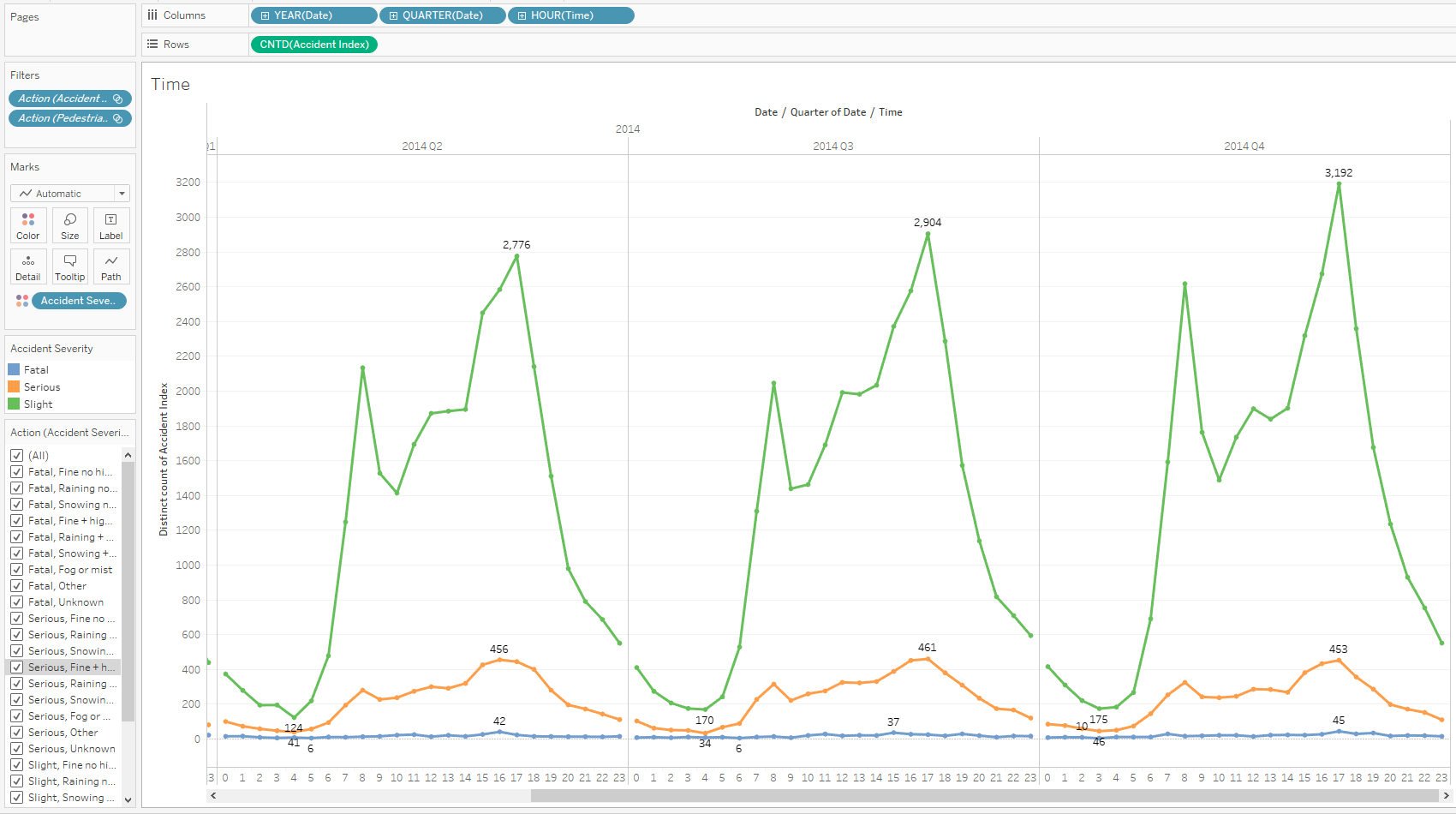
**Figure 6:**



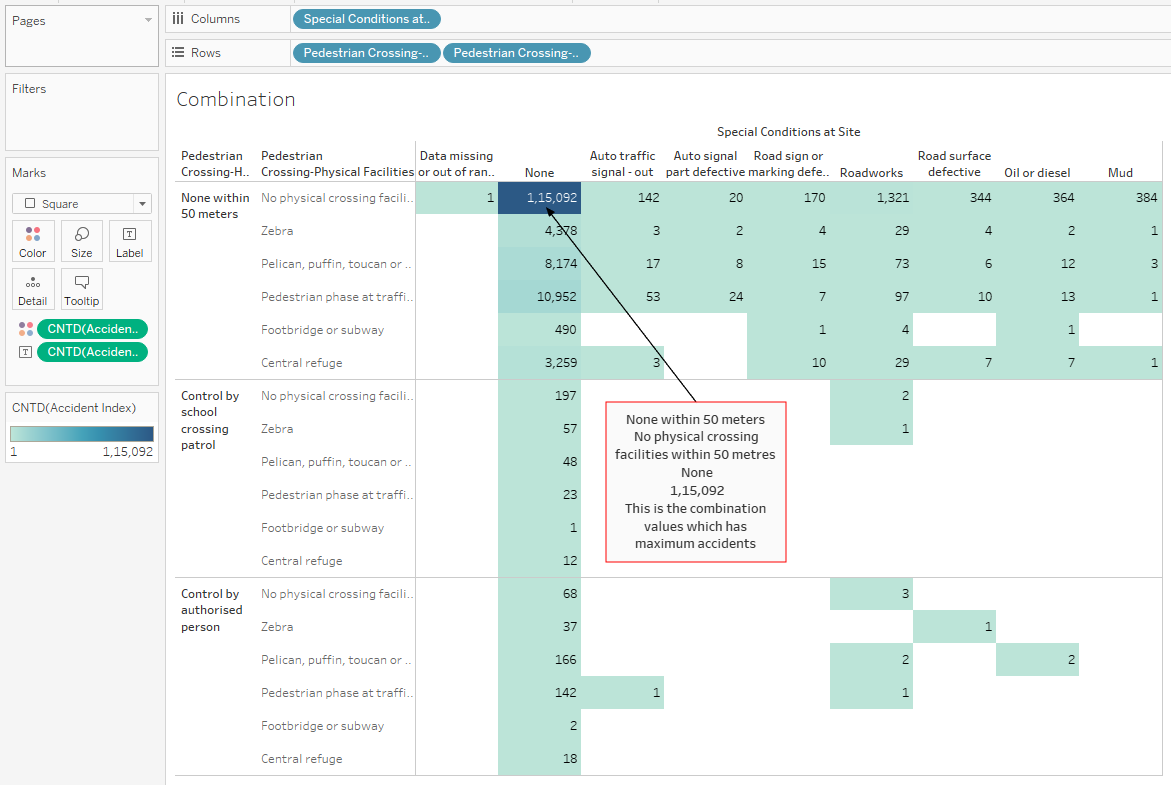
**Figure 7 (a):**



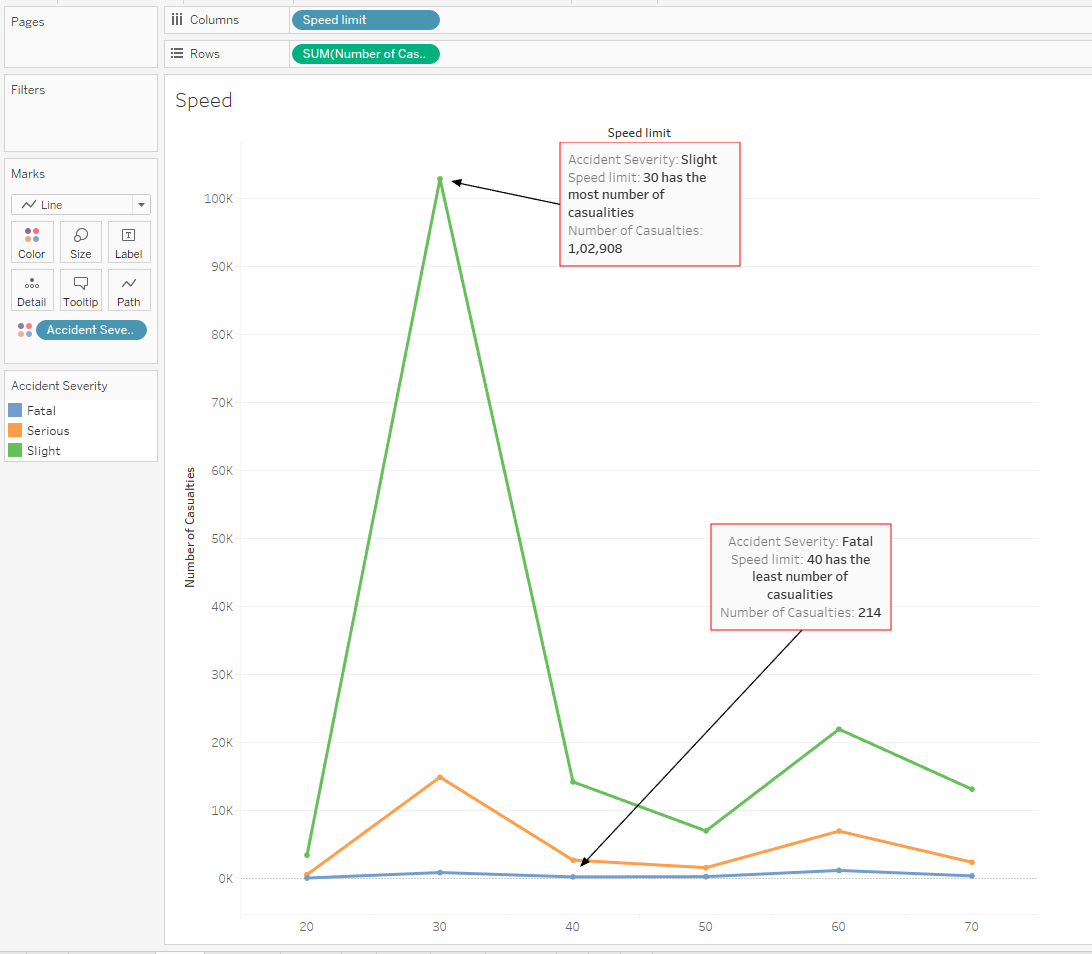
**Figure 7 (b):**



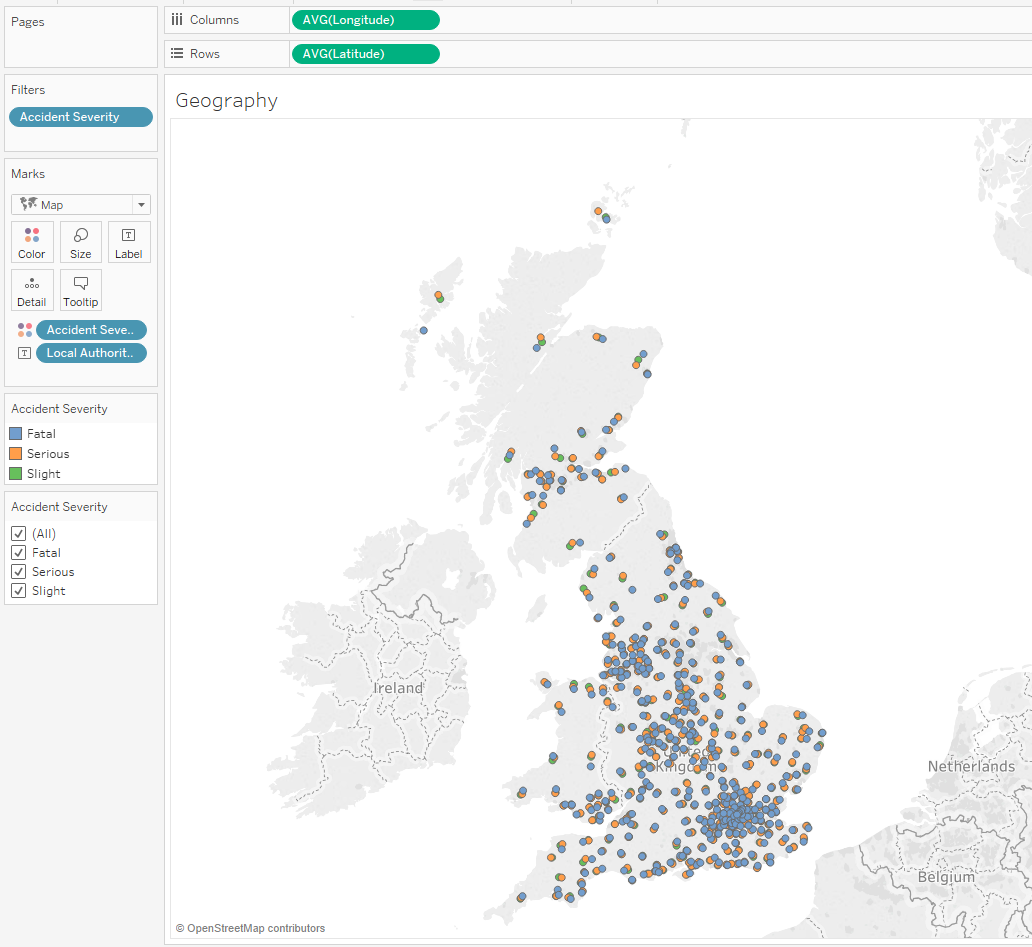
**Figure 8:**



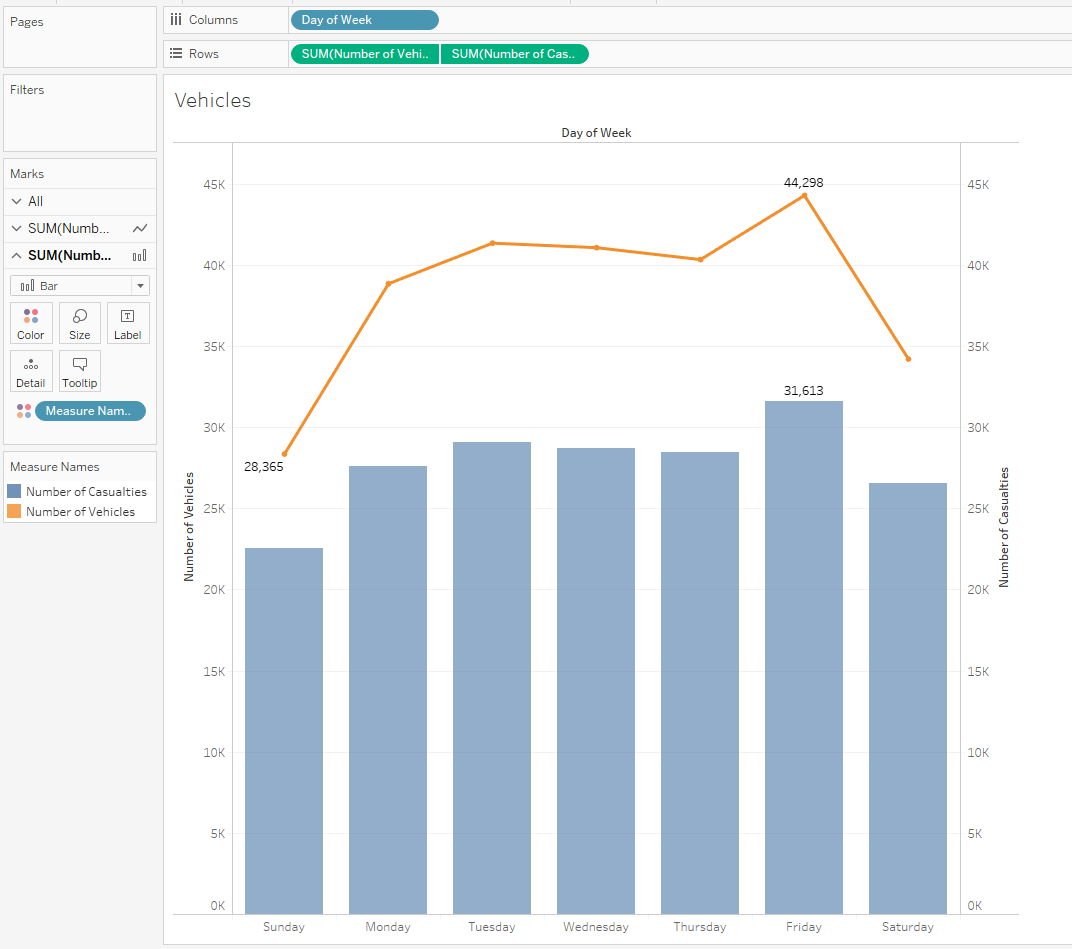
**Figure 9:**



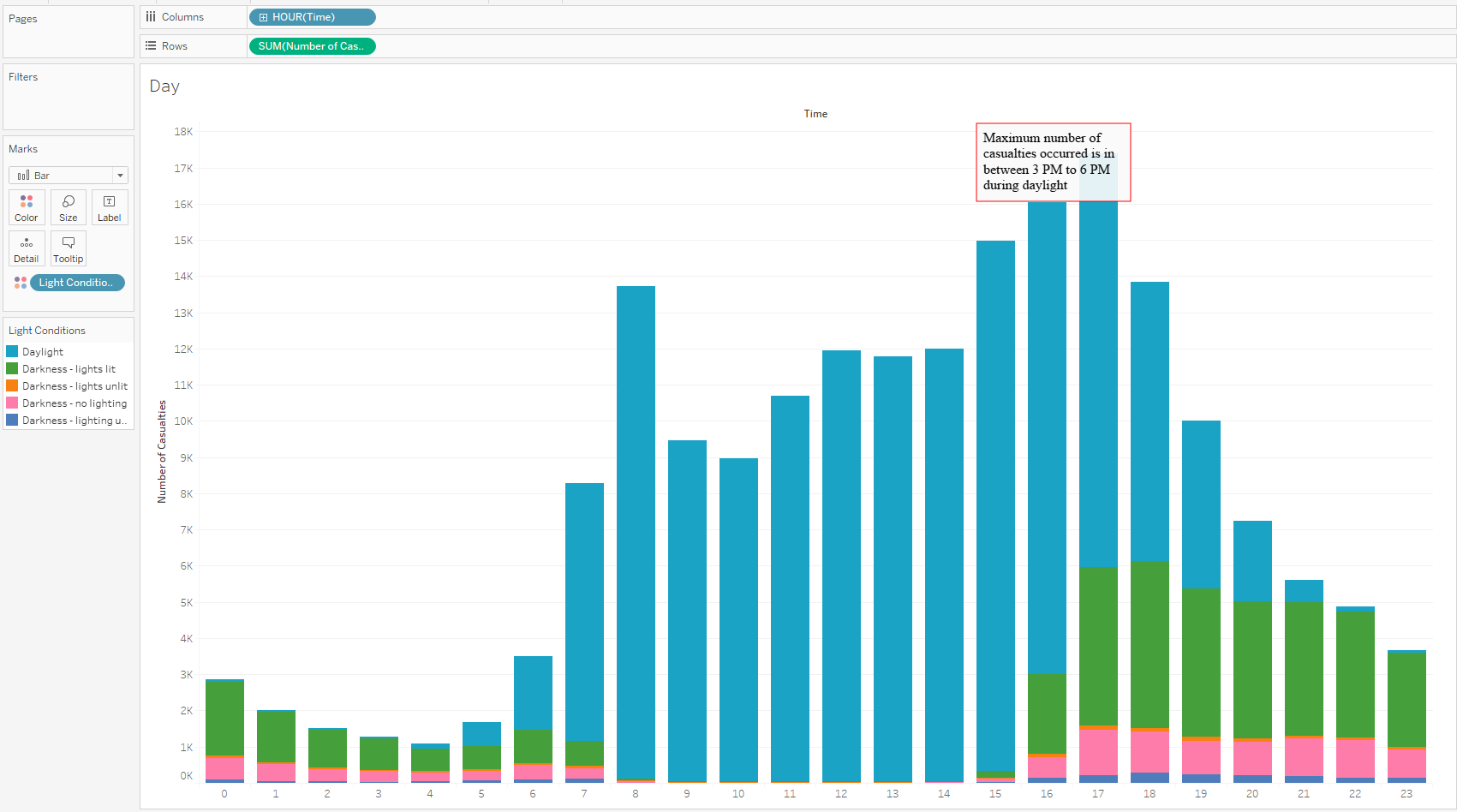
**Figure 10:**



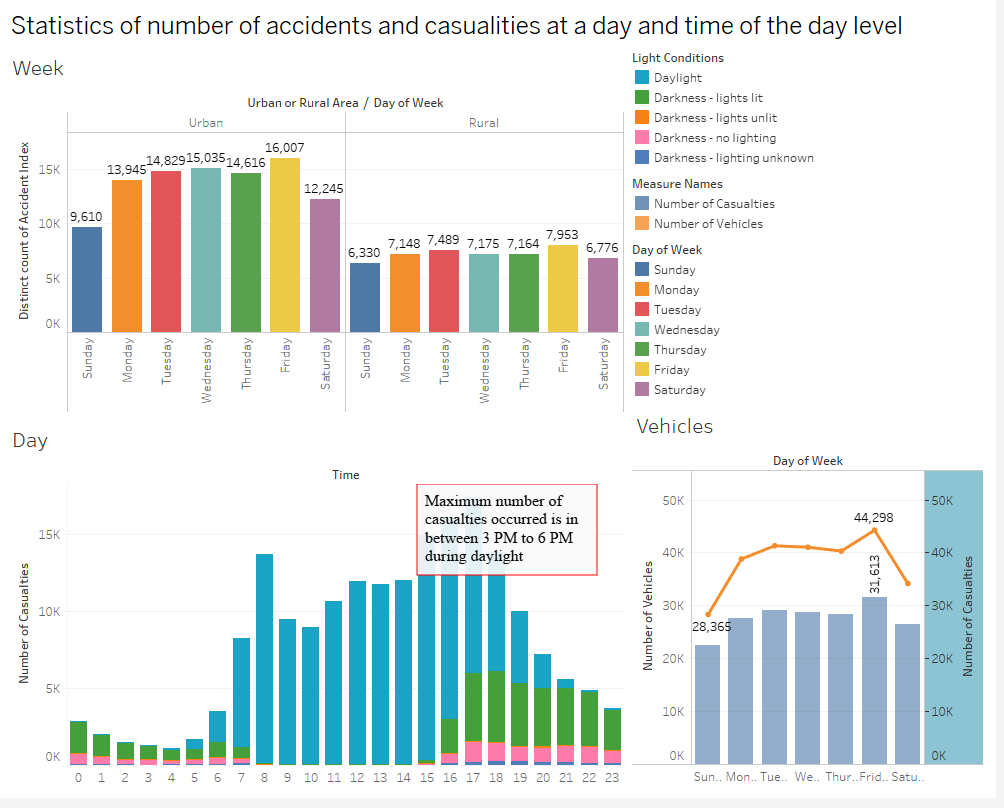
**Figure 11:**



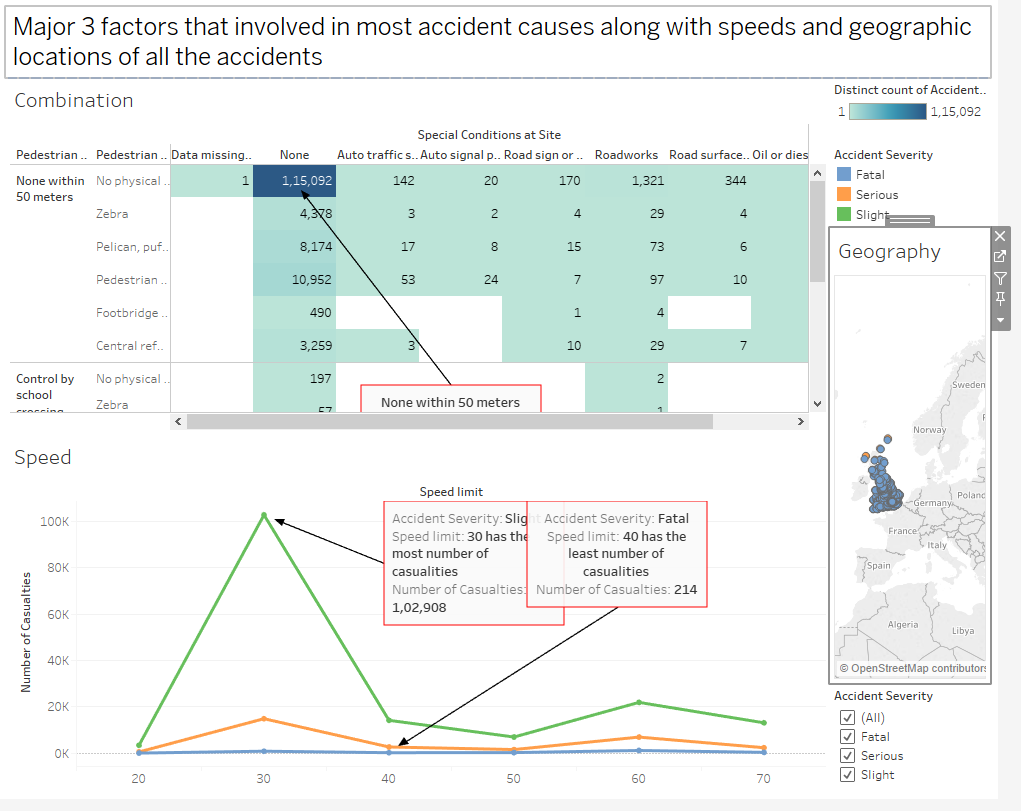
**Figure 12:**



**Figure 13:**



**Figure 14:**



**Exploratory Data Analysis R code:**

dataset <- read.csv("H:/MASTERS/Data Visualization and Dashboarding/Assessment/New folder/DfTRoadSafety\_Accidents\_2014.csv", header=TRUE)

install.packages("DataExplorer")

library(DataExplorer)

install.packages("tidyverse")

library(tidyverse)

library(funModeling)

library(Hmisc)

glimpse(dataset)

summary(dataset)

dim(dataset)

set.seed(12)

str(dataset)

head(dataset)

tail(dataset)

create\_report(dataset)

boxplot(dataset)

is.na(dataset)

sum(is.na(dataset))

### **REFERENCES:**

Road Safety Data - data.gov.uk. 2019. Road Safety Data - data.gov.uk. [ONLINE] Available at: <https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data>. [Accessed 07 April 2019].

European Environment Agency. 2019. Prepare data for analysis and visualisations — European Environment Agency . [ONLINE] Available at: <https://www.eea.europa.eu/data-and-maps/daviz/learn-more/prepare-data>. [Accessed 15 April 2019].

research.tableau.com. 2019. Eviza: A Natural Language Interface for Visual Analysis. [ONLINE] Available at: <https://research.tableau.com/sites/default/files/uist4832-setlurA_0.pdf>. [Accessed 14 April 2019].

O’Reilly | Safari. 2019. 4. Choose Appropriate Visual Encodings - Designing Data Visualizations [Book]. [ONLINE] Available at: <https://learning.oreilly.com/library/view/designing-data-visualizations/9781449314774/ch04.html>. [Accessed 03 April 2019].

Visual Best Practices - Tableau. 2019. Visual Best Practices - Tableau. [ONLINE] Available at: <https://onlinehelp.tableau.com/current/pro/desktop/en-us/visual_best_practices.htm>. [Accessed 15 April 2019].

David Baldwin. 2019. The 10 Commandments of Visual Analytics in Tableau. [ONLINE] Available at: <https://www.teknionusa.com/blog/the-10-commandments-of-visual-analytics-in-tableau>. [Accessed 12 April 2019].