# **MIS 6380 - Data Visualization**

# **Project Report** **Group 1**

# **FACTORS CONTRIBUTING TO ROAD ACCIDENTS**

# **(UNITED STATES: 2015)**

# **Instructor: Dr. Judd Bradbury**

**Group Members:**

**Alok Sanjeev Sheth**

**Himanshu Yadav**

**Piyush Mittal**

**Pranusha Yanamandala**

**Prapti Singh**

**Introduction**

Every year the lives of more than 1.25 million people are cut short as a result of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability because of their injury. Road traffic injuries cause considerable economic losses to individuals, their families, and to nations.

Accidents are responsible for a huge loss of human lives and the number of deaths, and deaths relative to the total US population, declined over most of the previous two decades, the trend reversed in 2015.The Nation lost 35,092 people in crashes on U.S. roadways during 2015, an increase from 32,744 in 2014. The 7.2-percent increase is the largest percentage increase in nearly 50 years. The estimated number of people injured on the Nation’s roads increased in 2015, rising from 2.34 to 2.44 million injured people. Fatalities increased from 2014 to 2015 in almost all segments of the population—passenger vehicle occupants, passengers of large trucks, pedestrians, pedal cyclists, motorcyclists, alcohol-impaired driving fatalities, male/female, daytime/nighttime.

We have taken account of the Road accidents in The United States in the year 2015. We analyzed the effect of various factors on accidents and found out some interesting insights which are explained below.

This analysis will provide a deep dive into the most important causes and how the combination of these causes can increase the fatality of a road accident. The purpose of this analysis is to create awareness about some uncommon factors leading to road accidents. Moreover, these causes can be used to plan and implement certain safety measures in the accident-prone zones which can play an important role in saving human life and property.

**Data Sources**

The Fatality Analysis Reporting System (FARS) provides information on all motor vehicle traffic crashes in the U.S. in which one or more involved people die of their injuries within 30 days of the crash. FARS data is available for every year since FARS was established in 1975. Users can query the FARS database directly using download files via FTP at [**ftp://ftp.nhtsa.dot.gov/fars/2015/National/**](ftp://ftp.nhtsa.dot.gov/fars/2015/National/)

* Main Data Source

**File Name:** Accidents.csv

**File Description:** The Accident data file includes crash data. It contains the data elements ST\_CASE and STATE, which are described at the beginning of the Data Description section. The Accident data file also contains the data elements on the following pages. ST\_CASE is the unique case identifier for each record.

* Secondary Data Source

**File Name:** Vehicles.csv

**File Description:** The Vehicle data file includes in-transport motor vehicle data as well as driver and pre-crash data. ST\_CASE and VEH\_NO are the unique identifiers for each record. ST\_CASE is used to merge the Vehicle data file with the Accident data file.

**File Name:** Persons.csv

**File Description:** The Person data file includes motorist and non-motorist data. It contains the data elements ST\_CASE, STATE, VEH\_NO, and PER\_NO. ST\_CASE, VEH\_NO, and PER\_NO are the unique identifiers for each record.

ST\_CASE is used to merge the Person data file with the Accident data file for a set of all motorists and non-motorists.

ST\_CASE and VEH\_NO should be used to merge the Person data file with the Vehicle data files for a set of all motor vehicle occupants.

ST\_CASE and PER\_NO is used to merge the Person data file with a non-motorist person-level dataset

**File Name:** Factors.csv

**File Description:** The Factors data file includes encoded data for the factors responsible for the accident. It contains the data elements

ST\_CASE, STATE, VEH\_NO, and MFACTOR. ST\_CASE, VEH\_NO are unique identifiers for each record.

ST\_CASE VEH\_NO is used to merge the Factors data into Vehicle data.

* Supporting Data Source –

**Car Sales Data from 1990 to 2015**

**Link** - <http://carsalesbase.com/us-car-sales-data/>

**Description** – We have used this source to find the number of cars sold across 1990 to 2015 for a respective car manufacturer.

**File Name** – Vehicle Make.xlsx

**File Description –** File has a column of MAKE which tells the unique id of each MAKE, column MAKER tells the name of a respective make of the vehicle, column VEHICLE SALES tell the number of cars sold for each make from the year 1990 to 2015. This file is linked to Vehicle.csv using MAKE as the primary key.

**Population of US states for 2015**

**Link-** <https://www.infoplease.com/us/population/us-population-state-1790-2015>

**Description** – We have used this data to find the population of each state of US for the year 2015

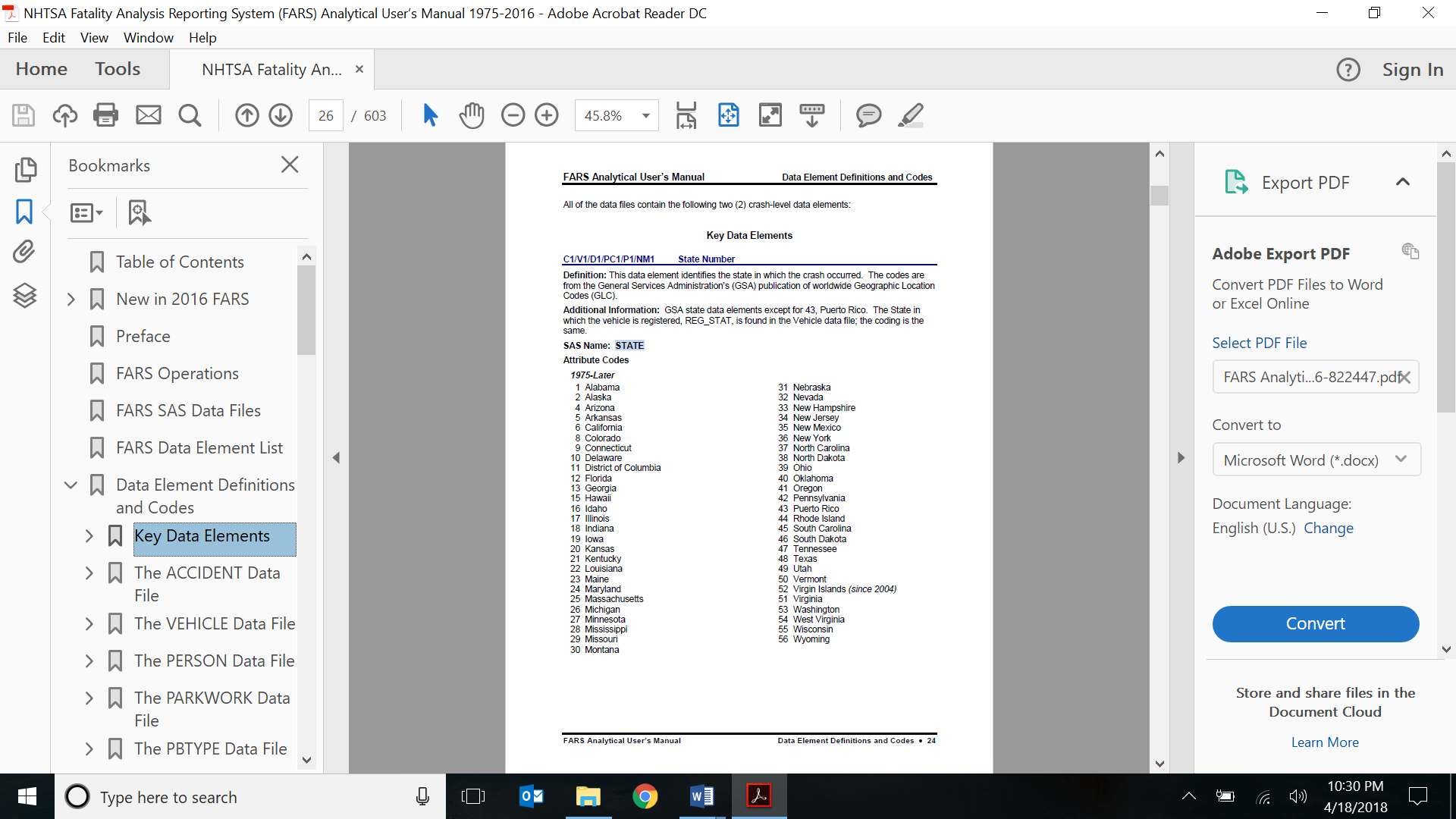
**File Name** – Population.xlsx

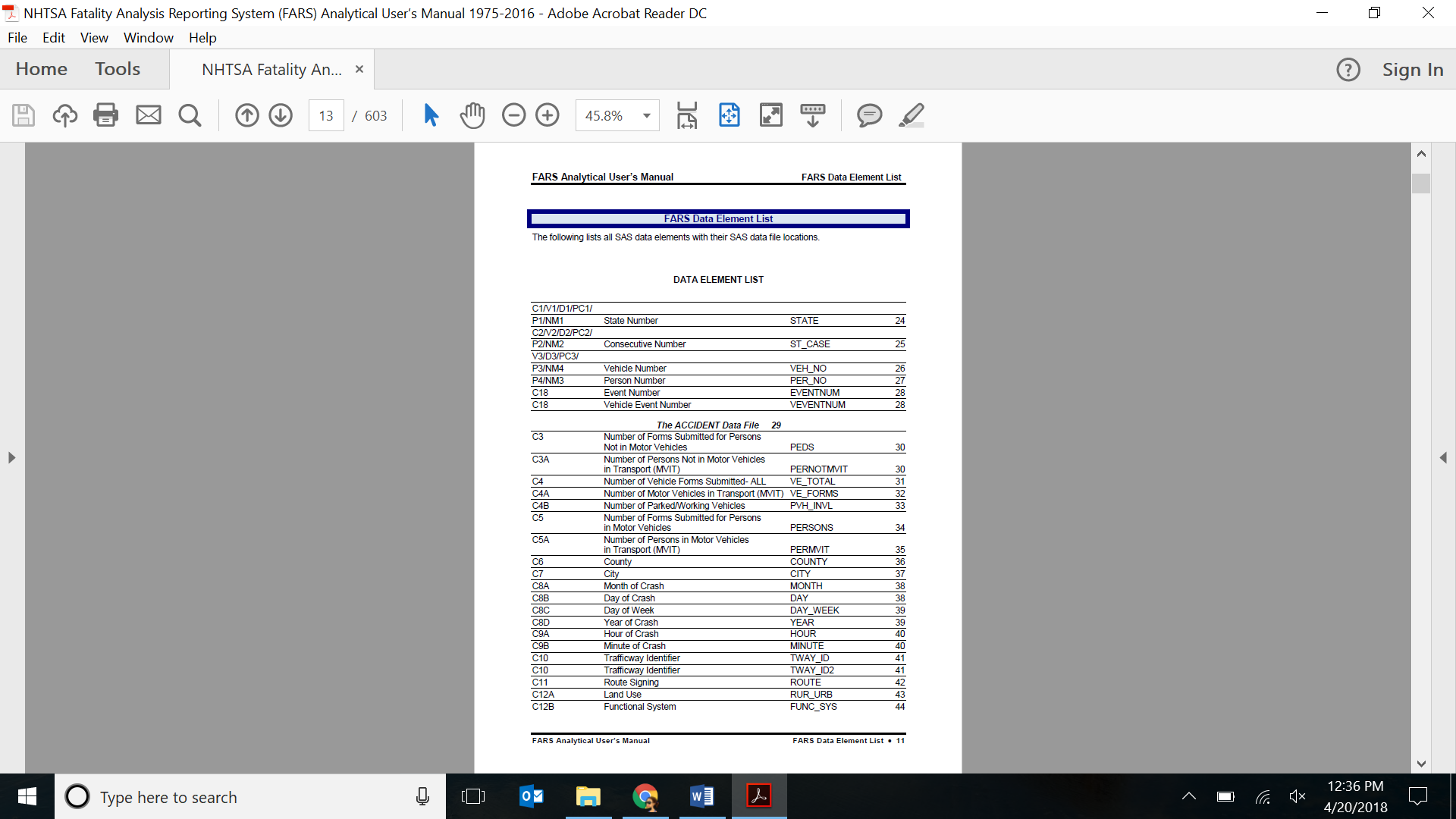
**File Description** – File has a column for STATE which represent the id of the state, Column Geography which represents the name of the state respective to STATE, column Population which represents the population of the respective state, column Accident Count which represent the number of accident cases for the respective state for the year 2015. Accident\_Count is a derived column from Accident.csv file. This file is linked with Accident.csv with the unique id of STATE.

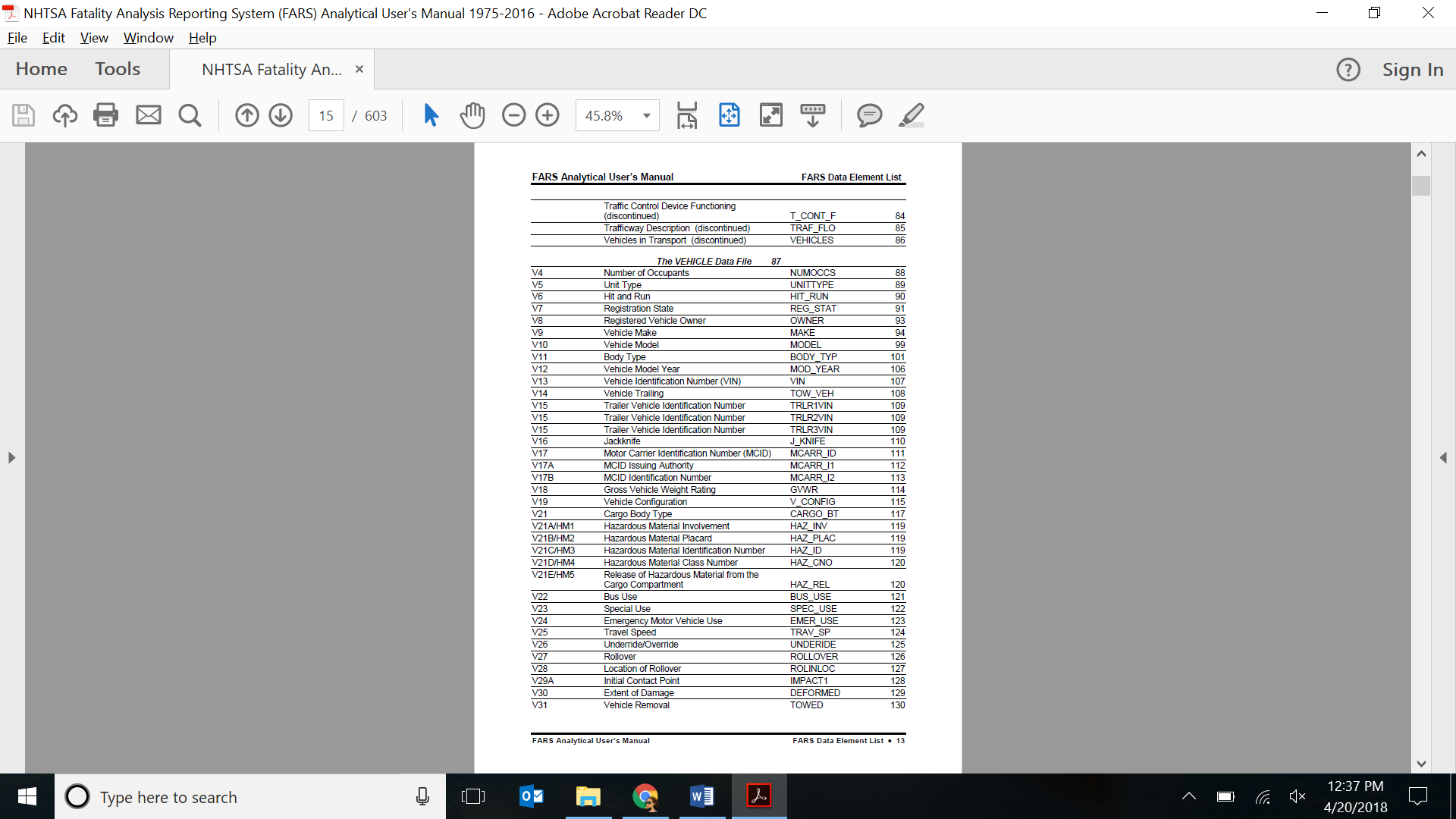
**Description of Data**

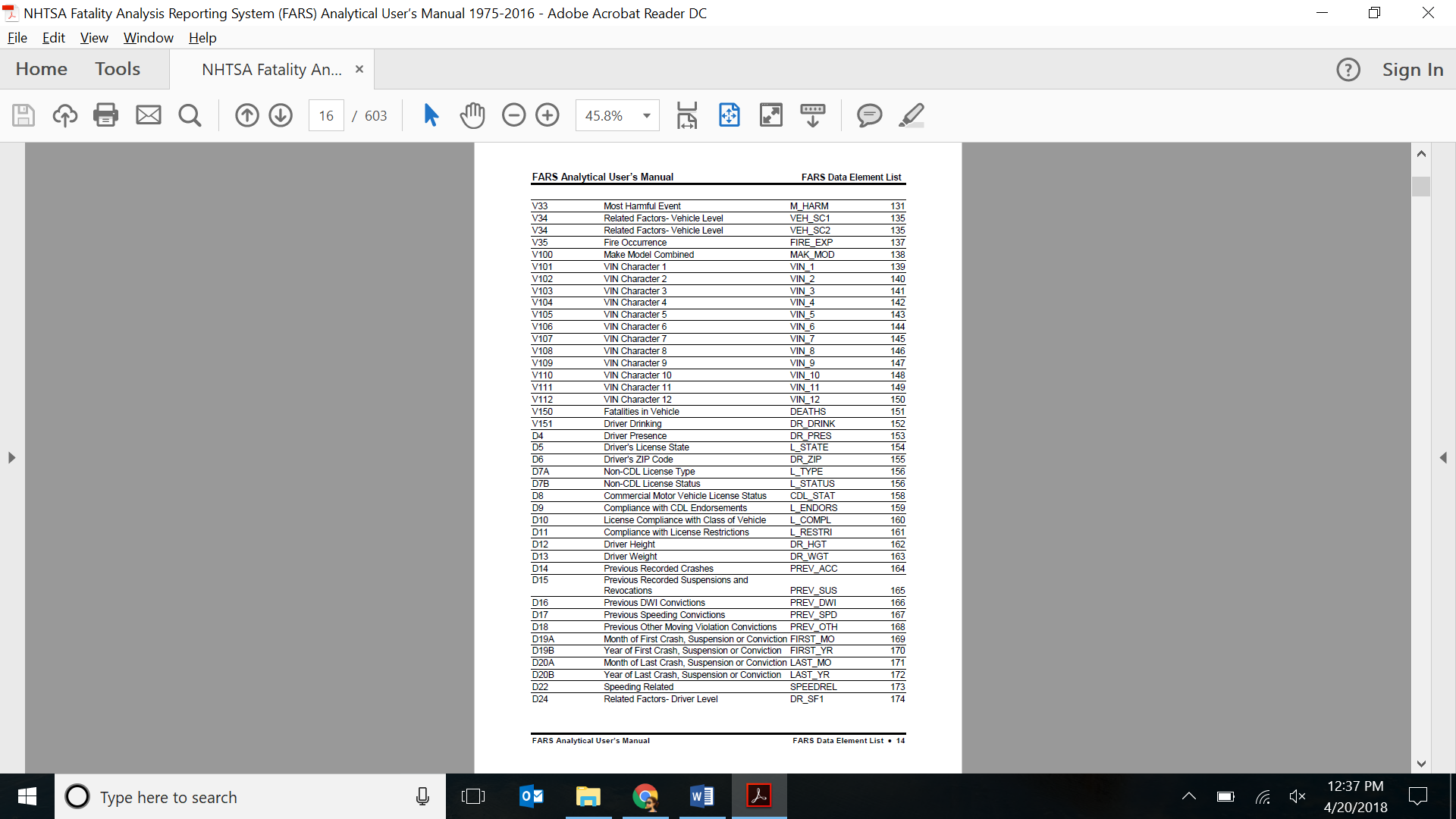
Below are the main data elements in the primary and secondary data source. These descriptions have been taken from Fatality Analysis Reporting System (FARS) Analytical User’s Manual 1975-2016.

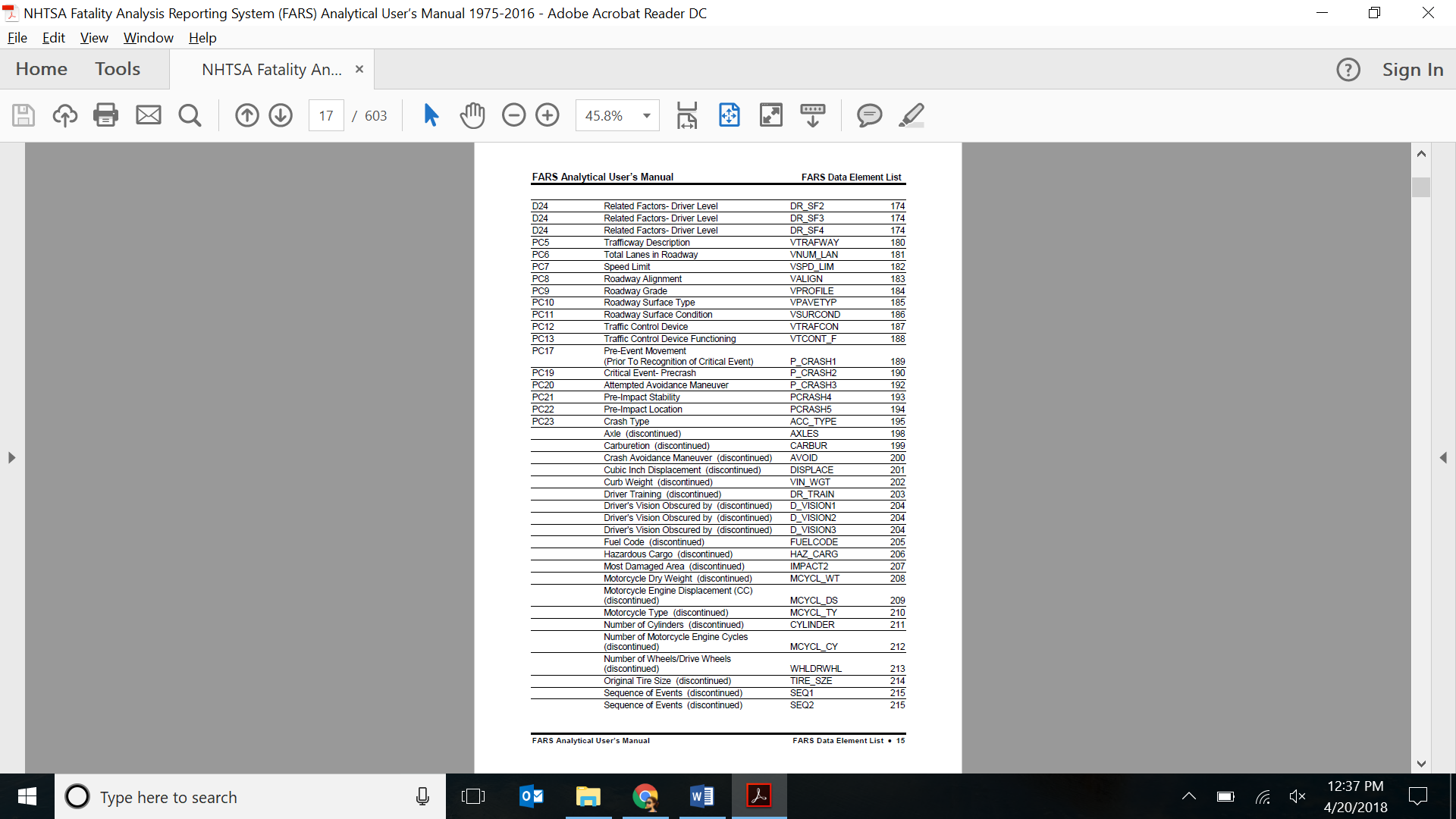
Note: These descriptions include all the elements present in the dataset but may or may not have been utilized as per their need. Many of them were not helpful for our analysis and have been excluded from the visualization.

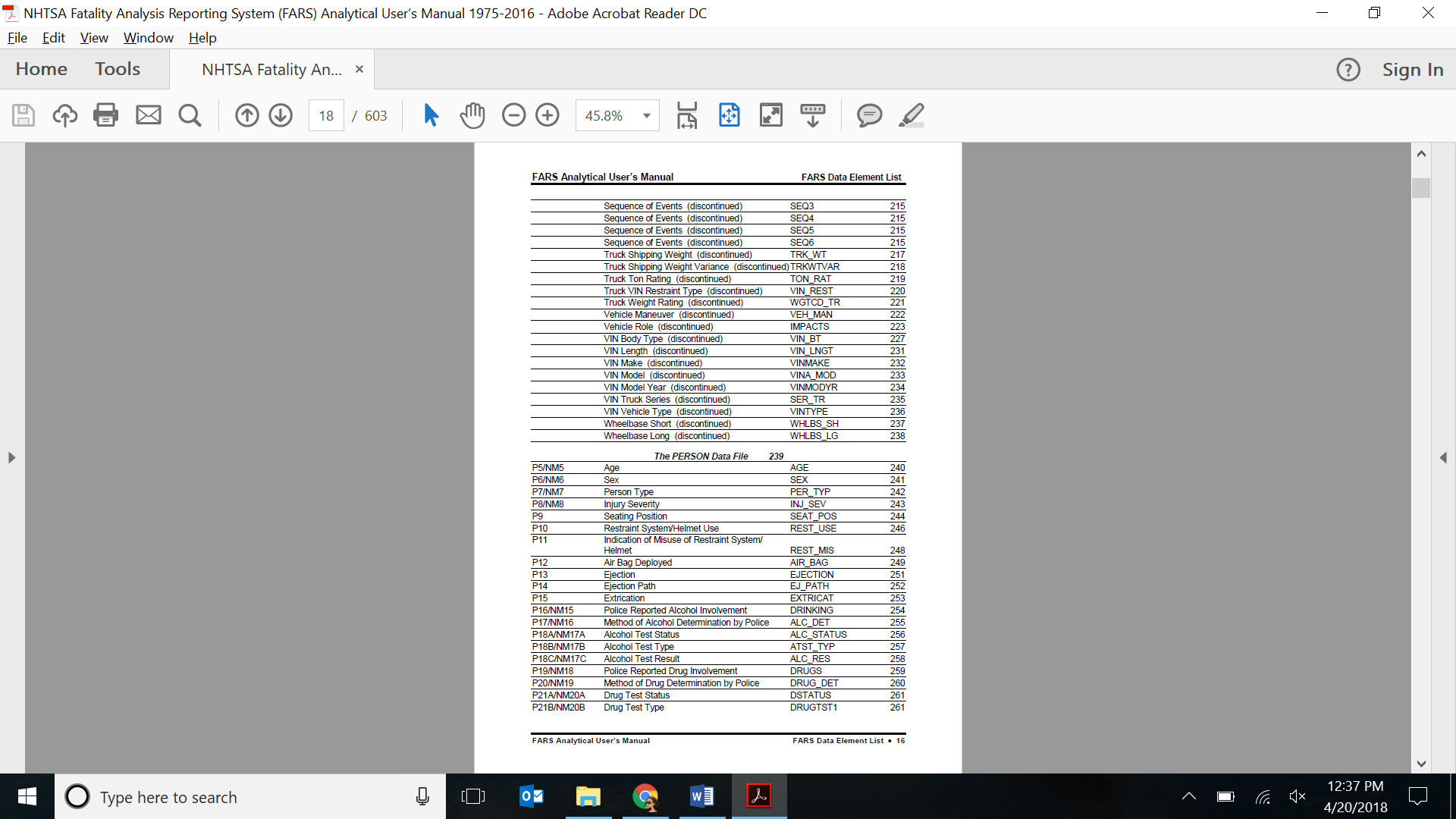


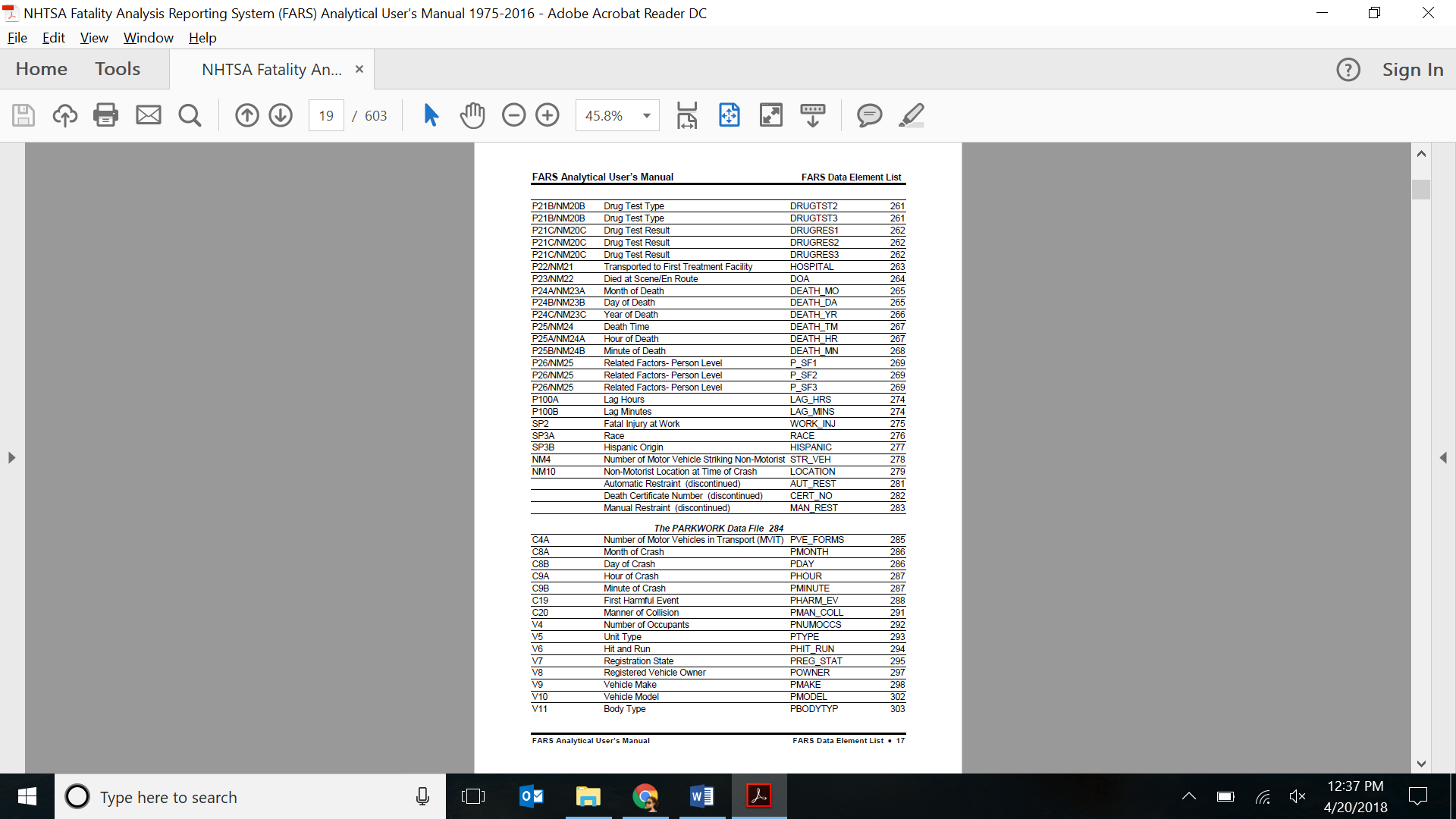












**Insights**

1. Wyoming is the most dangerous state to drive in.
2. Cloudy, Rain, Snow and Fog are the prominent bad weather conditions across the top three dangerous states to drive.
3. Most of the accidents occurred in Principal Arterial followed by Collector & Local segment of roads.
4. Drink & Drive accidents are less on weekdays and spiked during weekends around 1 am - 2 am, whereas nondrinking accidents spiked during evening hours of 4 pm - 5 pm during weekdays.
5. At peak hours of Drink & Drive accidents during 1 to 2 AM, Rural Roads should be monitored more for collector segment & Urban Roads for another segment of roads.
6. People in the 2nd cluster have highest survival rate than the most and those in the 5th cluster has highest death rate.
7. People seating on the Front Seat are at high risk of not surviving the accident than the ones seating in other seats.
8. The ratio of number car accidents to the number of vehicles sold by a maker is very high for Suzuki and non-human factor for accidents is Headlights.

**Story**

**Rhetoric Argument** – We have used Logical Rhetoric in our visualization for finding the most dangerous state to drive in.

The Story here tells us how the Accidents are a major concern in The United States. This story provides us the insights about the various factors which can be responsible for the increasing No. of Accidents. Starting with the identification of accident-prone areas, the story moves forward with finding out some important factors which are responsible for these accidents. This story uses **Logical Rhetoric** for the **Factors Contributing to Road Accidents**. This story is based on the **3-act play** model where we have a beginning, middle and an end. This story model has created a tension between the audience and resolves it by stating facts and guiding them towards our insights.

**Place**: United States of America

**Time**: 2015

**Characters**: Weather, Type of Roads, Human and Non-human factors

**Setting the Stage**

We set up the stage by introducing the intentness of the subject i.e. Accidents. This is visually represented on the Map of USA. The gravity of the issue is corroborated by pointing out the most and the least dangerous States to drive in. This is done by using a ratio of Accidents to Population as a variable.

**Introducing Dramatic Conflict**

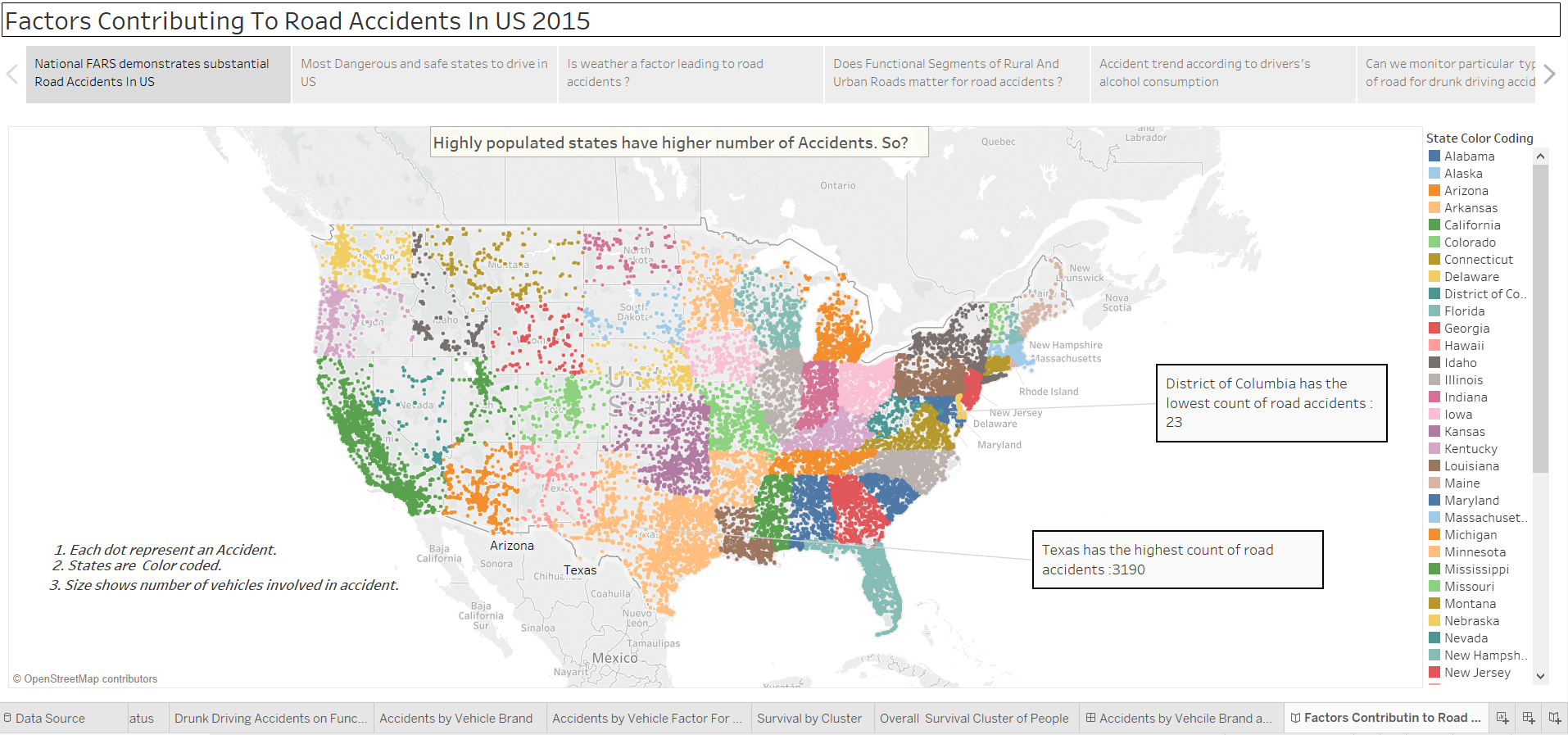
We used the logical Rhetoric by showing that it is logically correct that higher the population, higher will be the number of accidents. But we have proven it false by narrating that the number of accidents to population ratio of Wyoming is the highest which makes Wyoming as the most dangerous state to drive in. We have then narrated the factors like weather condition, drinking status of the driver, type of the road, a segment of the road, clustering of people in survived and died clusters and finally the factors related to the manufacturer of the vehicle.

* The next step to find out the potential cause of accidents in the top states. We have used multi-maps to drill down and find the top 4 weather conditions during the accident in the top 3 states.
* On finding out the type of roads which have maximum No. of Accidents, we found out that the Principal Arterial roads i.e. connecting collector roads to freeways or expressways, have more accidents.
* On finding how Alcohol can influence the probability of an Accident, we observed that there is a spike in No. of Accidents on weekend after midnight. On further analysis, we found that both Urban and Rural roads are unsafe at this hour but surprisingly for different road segments.
* To find the level of concern we thought about checking how fatal the accidents were. We used clustering to find out how can affect the survival of an individual in an accident.
* To know about the human involvement, we focused on factors such as drinking, restraint used, seating position, sex. We used a box plot to visualize the data and find out people with high risk of death in a car accident.
* We end the story by analyzing Car manufacturer and its components. This shows how a combination of all above-mentioned factors is responsible for the road accidents.

**Resolution**: In the end, we have narrated the resolution by stating the steps that need to be taken to reduce the number of road accidents in the USA.

**Visualizations**

1. National FARS demonstrates substantial Road Accidents In US:



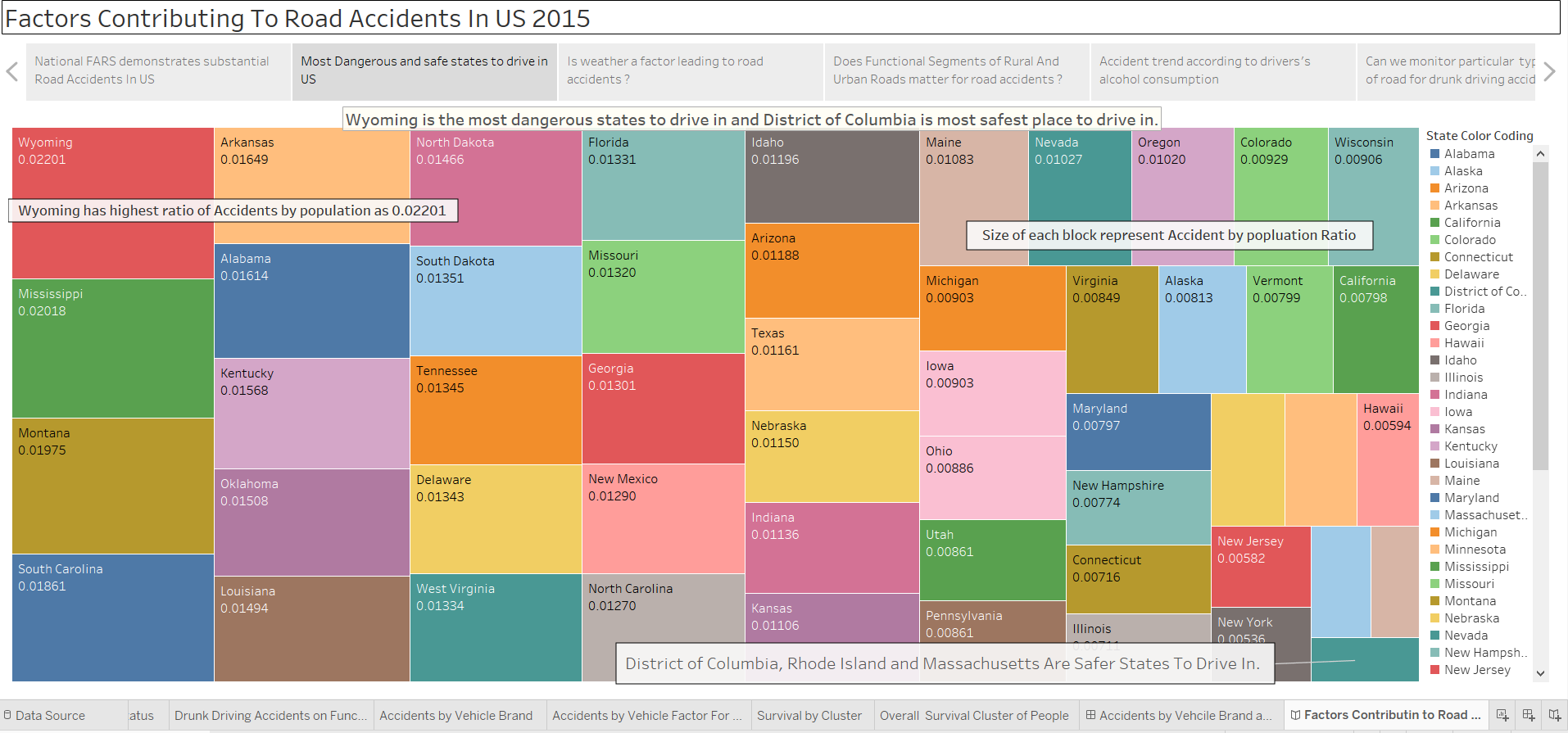
**Visual Encoding:**

Here we use latitudes and longitudes i.e. maps to plot the data points. The states are color coded i.e. each state is represented by different color. The density of points represents the density of accidents taking place in the state.

**Observation:** Highly populated states like Texas, California have a higher number of accidents.

**Insight:** Highly Populated are most dangerous states to drive in.

2. Most Dangerous and safe states to drive in the US:



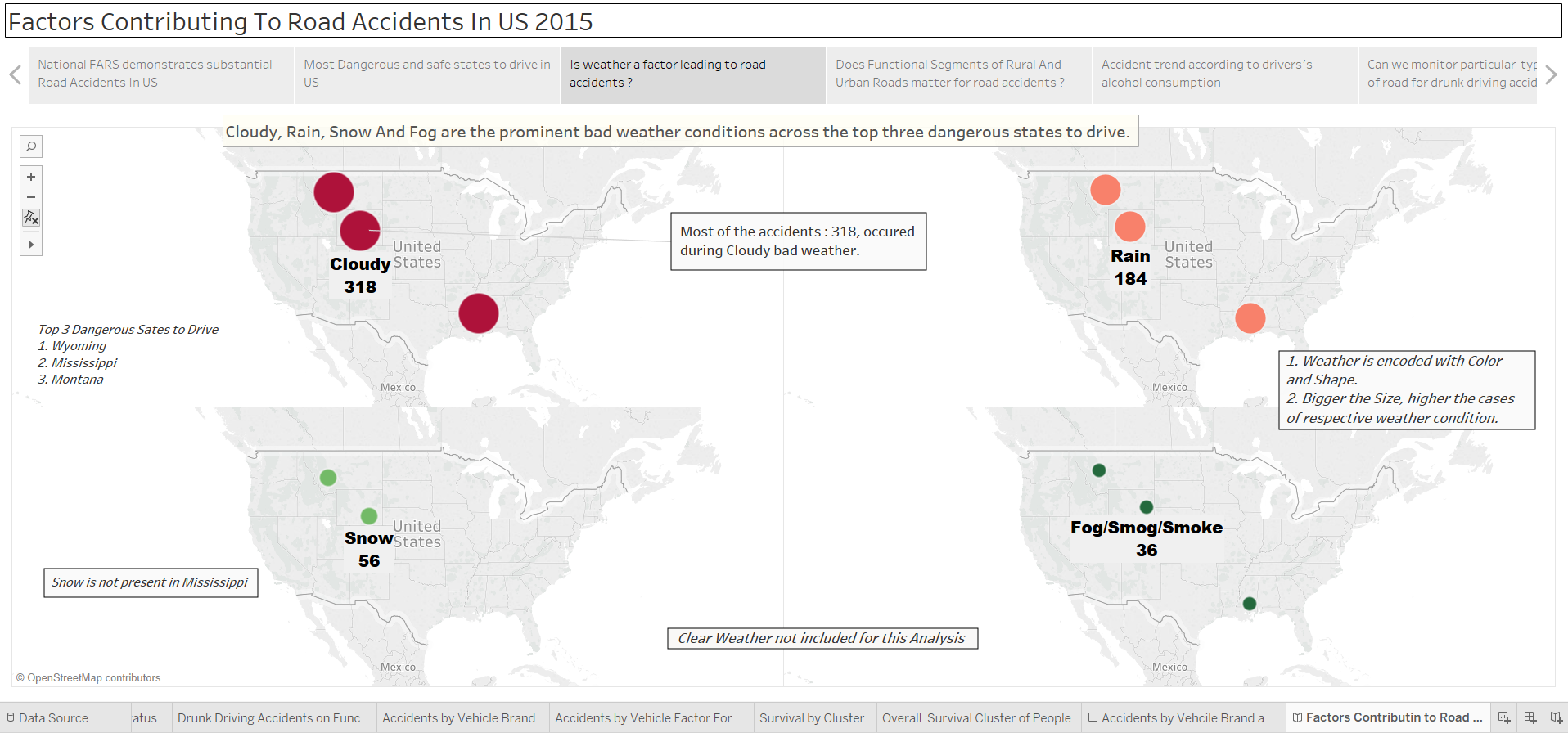
**Visual Encoding:**

For this visualization, we have used treemaps. We have used a calculated field (ratio) to estimate the most dangerous states to drive in. It is clearly visible on the map that as the ratio is decreasing (dangerous to safest), the size of the states is also decreasing.

**Observation:** Wyoming has the highest ratio of the number of accidents to the population.

**Insight:** Wyoming is the most dangerous state to drive in.

3. Is weather a factor leading to road accidents?



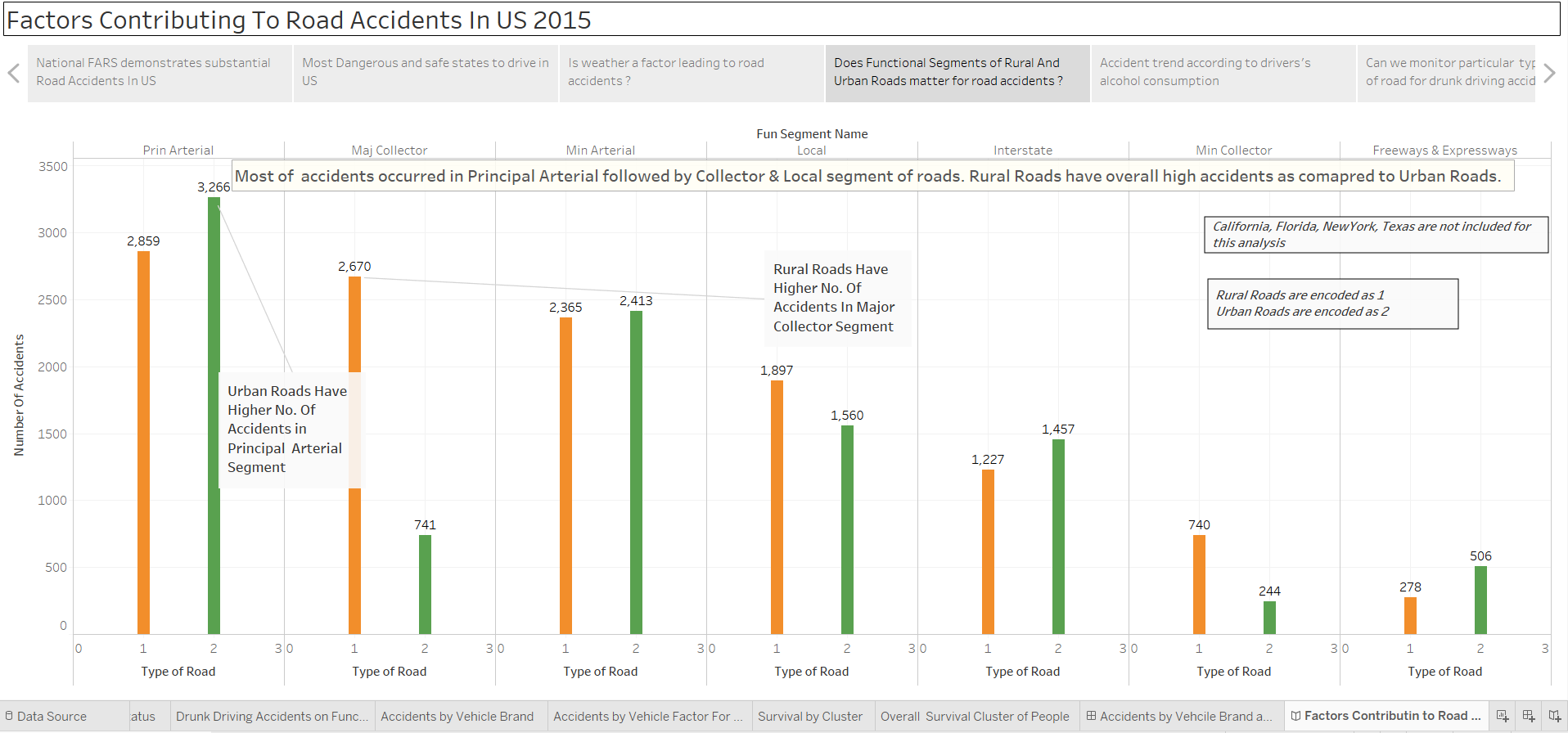
**Visual Encoding**

We have utilized multimaps to show the top 4 weather conditions affecting accidents in the most dangerous states which we found in the previous visualization. Color is representing different types of weather. The size of the circle represents the number of accidents taking place due to a weather.

**Note:** We have removed Clear weather for our analysis as most of the accidents occur during clear weather. Including clear weather will create a bias in our result.

**Insight:** Cloudy, Rain, Snow and Fog are the prominent bad weather conditions across the top three dangerous states to drive.

4. Does Functional Segments of Rural and Urban Roads matter for road accidents?



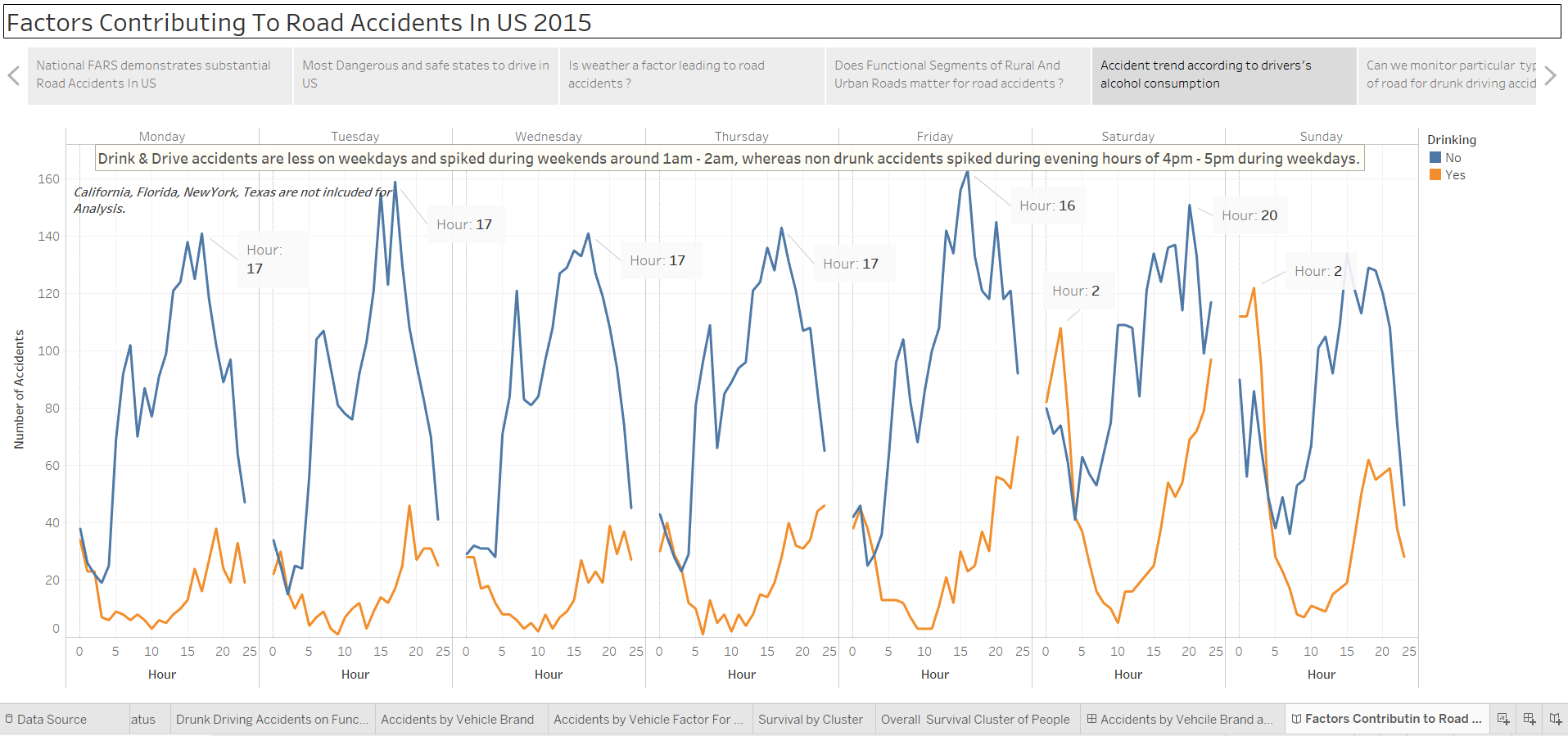
**Visual Encoding**

In this visualization, we have grouped the number of accidents as per functional segments of the roads using bar graphs. The green bar represents accidents which took place in urban roads and orange bar represents accidents which took place in rural roads. Sum of accidents is mentioned on top of the bars.

**Note:** We have removed California, Texas, Florida, New York from here on for our visualizations, as they will create a bias in our analysis

**Insight:** Most of the accidents occurred in Principal Arterial followed by Collector & Local segment of roads.

5. Accident trend according to drivers’ alcohol consumption

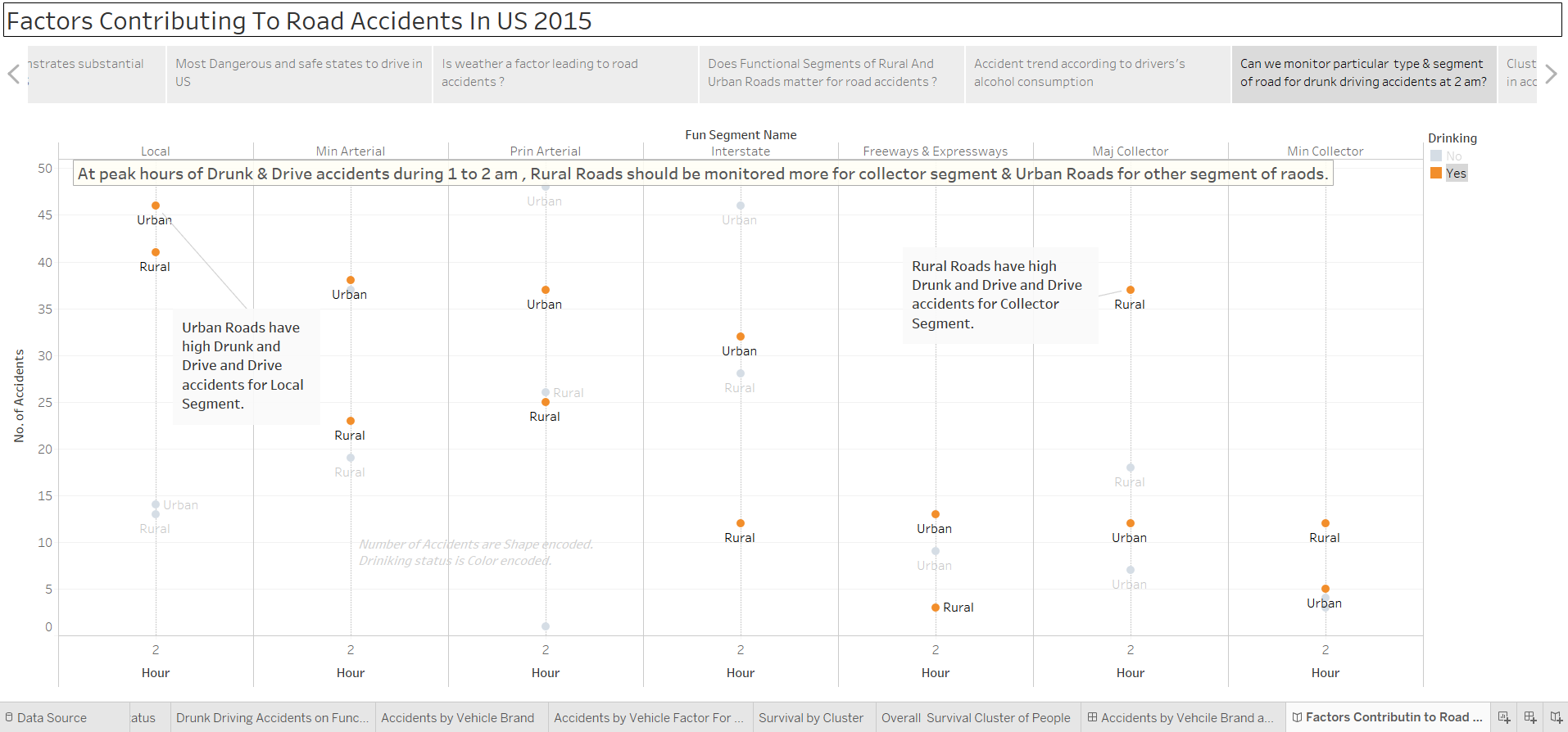


**Visual Encoding**

Trend lines are used to depict accidents taking place due to consumption and non-consumption of alcoholic drinks. The accidents are grouped by time and day of the week. The trend line in blue shows accidents taking place and orange trend lines show accidents taking place due to drinking and driving. Peak time (24-hr format) of accidents taking every week are annotated.

**Insight:** Drink & Drive accidents are less on weekdays and spiked during weekends around 1 am - 2 am, whereas nondrinking accidents spiked during evening hours of 4 pm - 5 pm during weekdays.

6. Can we monitor type & segment of road for drunk driving accidents at 2 am?

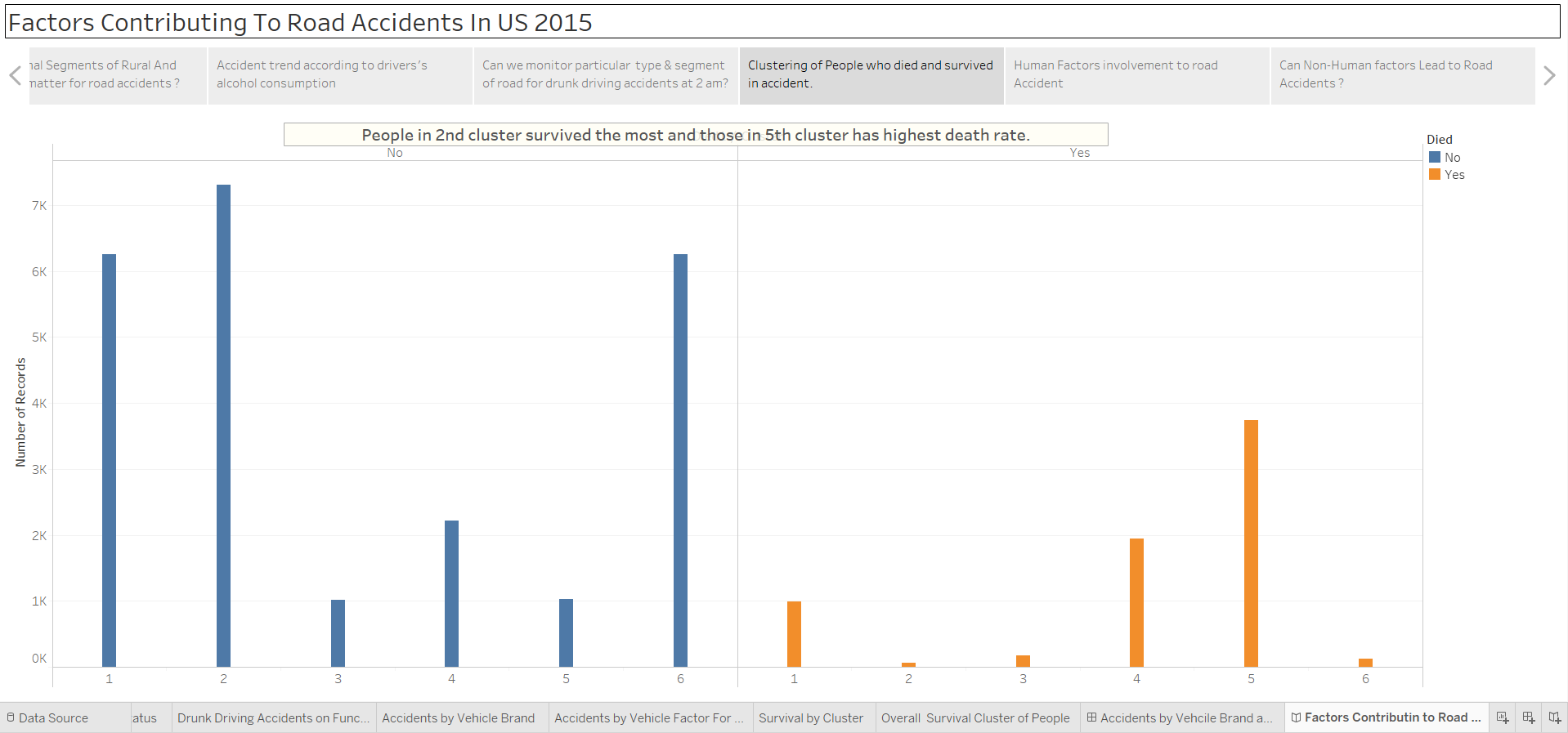


**Visual Encoding:**

Here we have applied filter for a drink and drive accidents according to road segments and rural/urban roads at the peak time of accidents (2 am). Orange dot represents Drink & drive accident cases & blue dot represents accidents not involving alcohol.

**Insight:** At peak hours of Drink & Drive accidents during 1 to 2 AM, Rural Roads should be monitored more for collector segment & Urban Roads for another segment of roads

7. Clustering of People who died and survived the accident

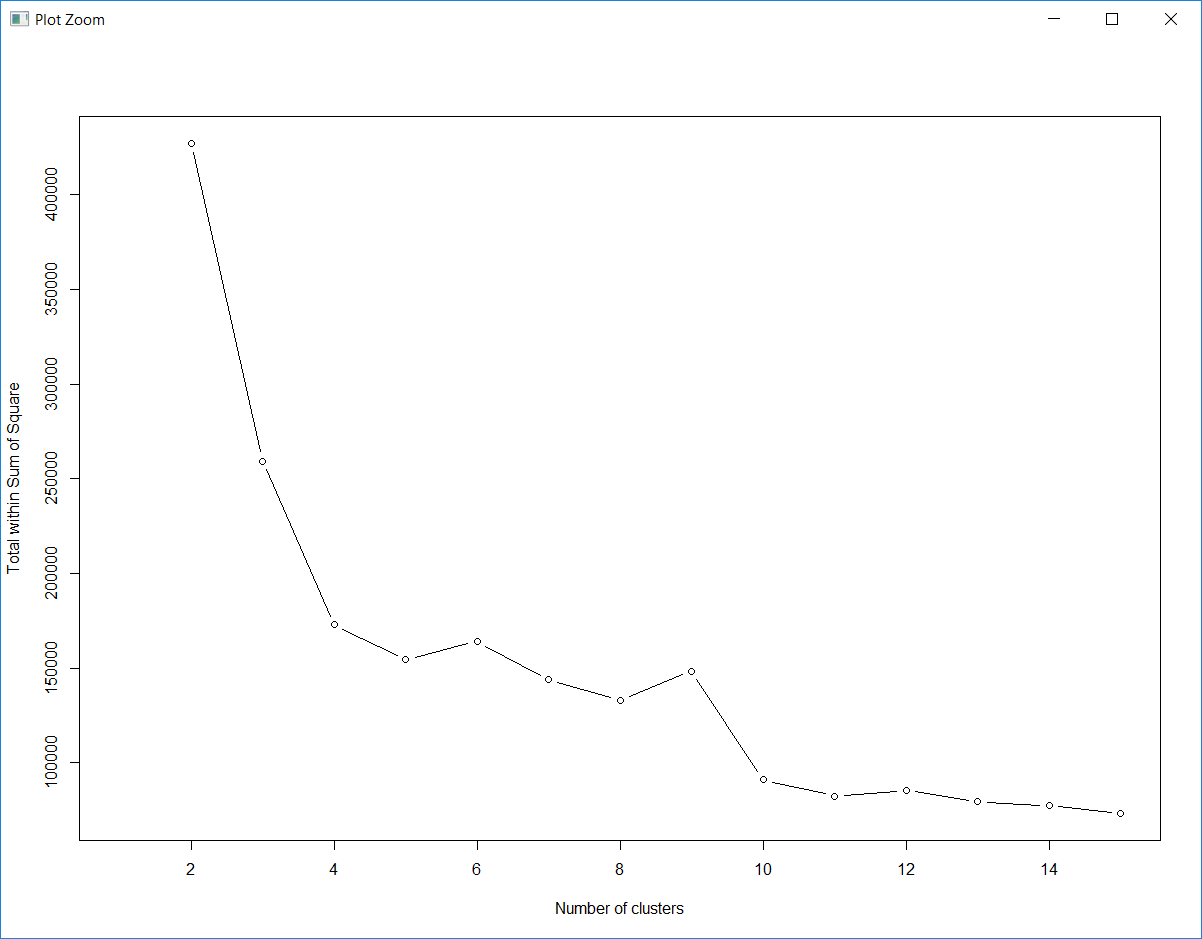


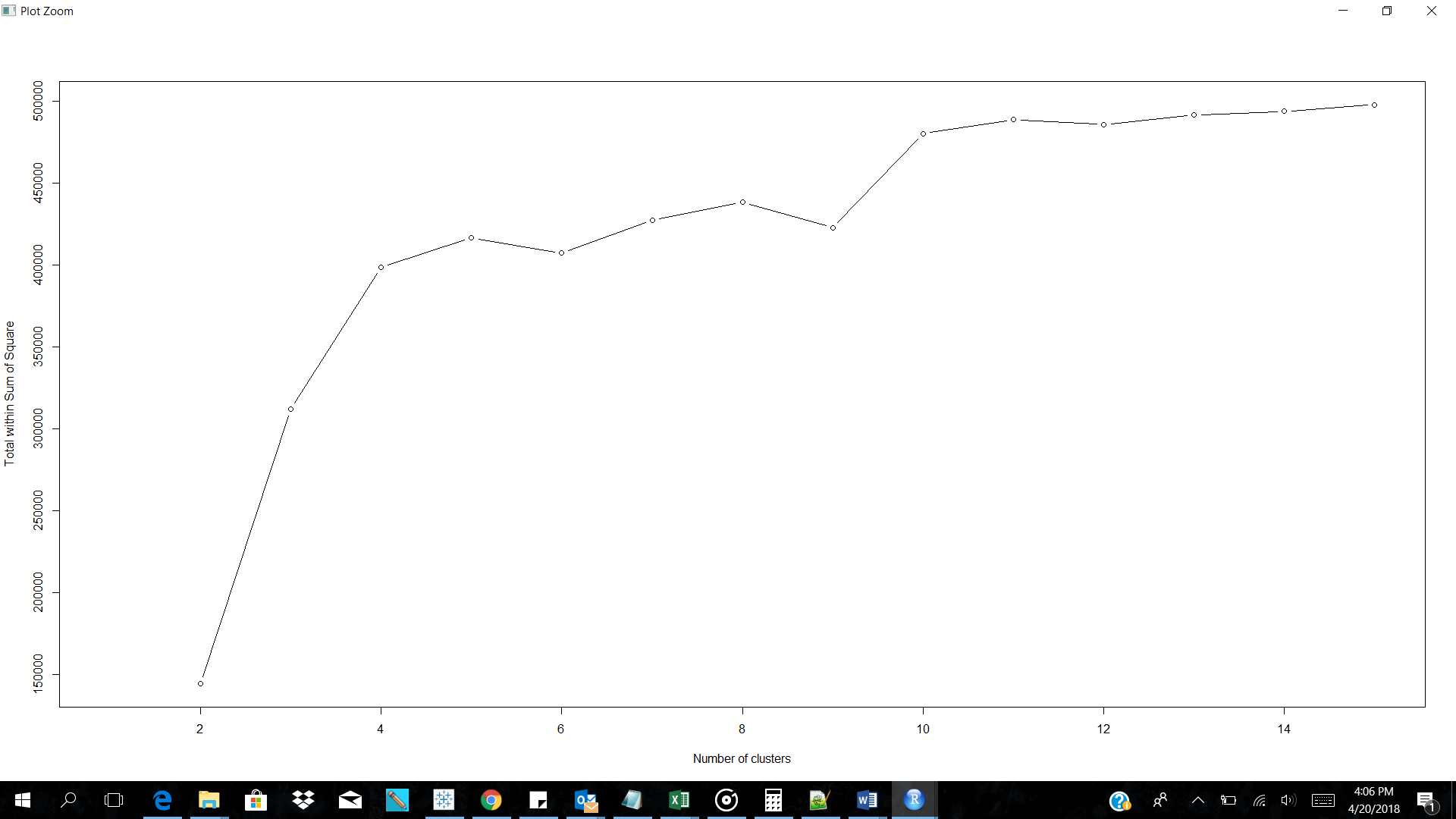
**Visual Encoding**

Bar graphs are used in this visualization. Blue coded data shows us survival cluster and orange-coded data shows us death clusters. Here we have accidents clustered in 6 clusters using bar graphs. The count of 6 clusters we got from clustering analysis using R. Above is the plot of the total within sum square vs a number of clusters. The plot tells us that curve stagnates at a number of clusters equal to 6. So we use 6 as an optimum number of clusters.

**R Clustering result:**

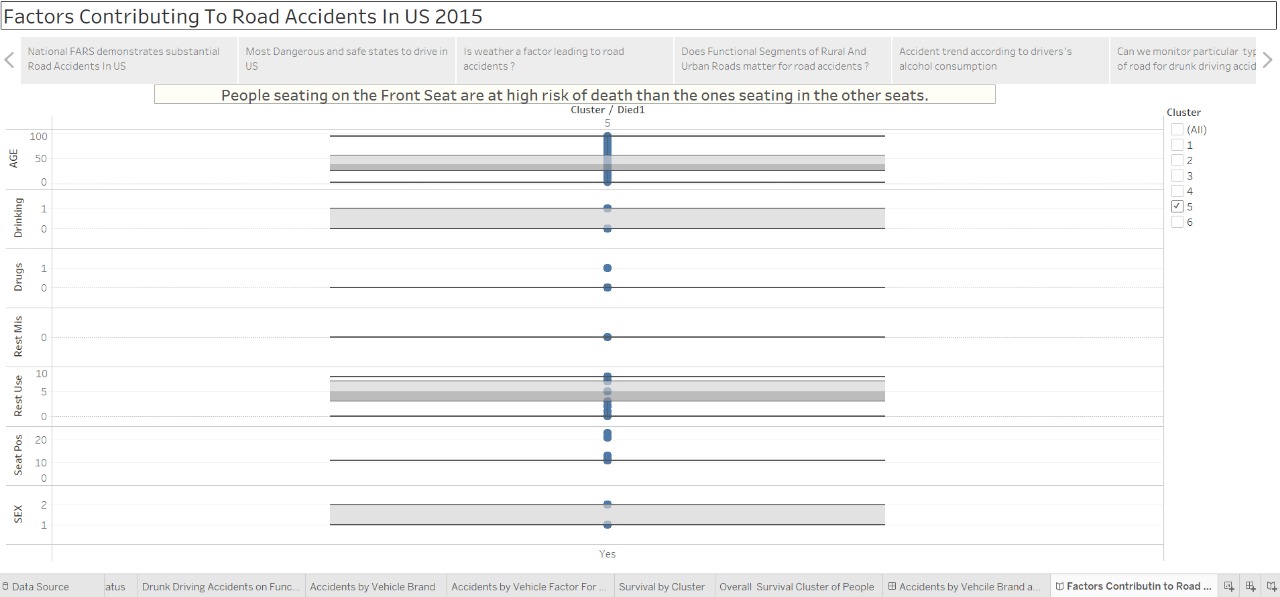
**Total Within Sum of Squares plot vs number of clusters –**





**Insight:** People in 2nd cluster survived the most and those in the 5th cluster has highest death rate.

8. Human Factors involvement in Road Accident

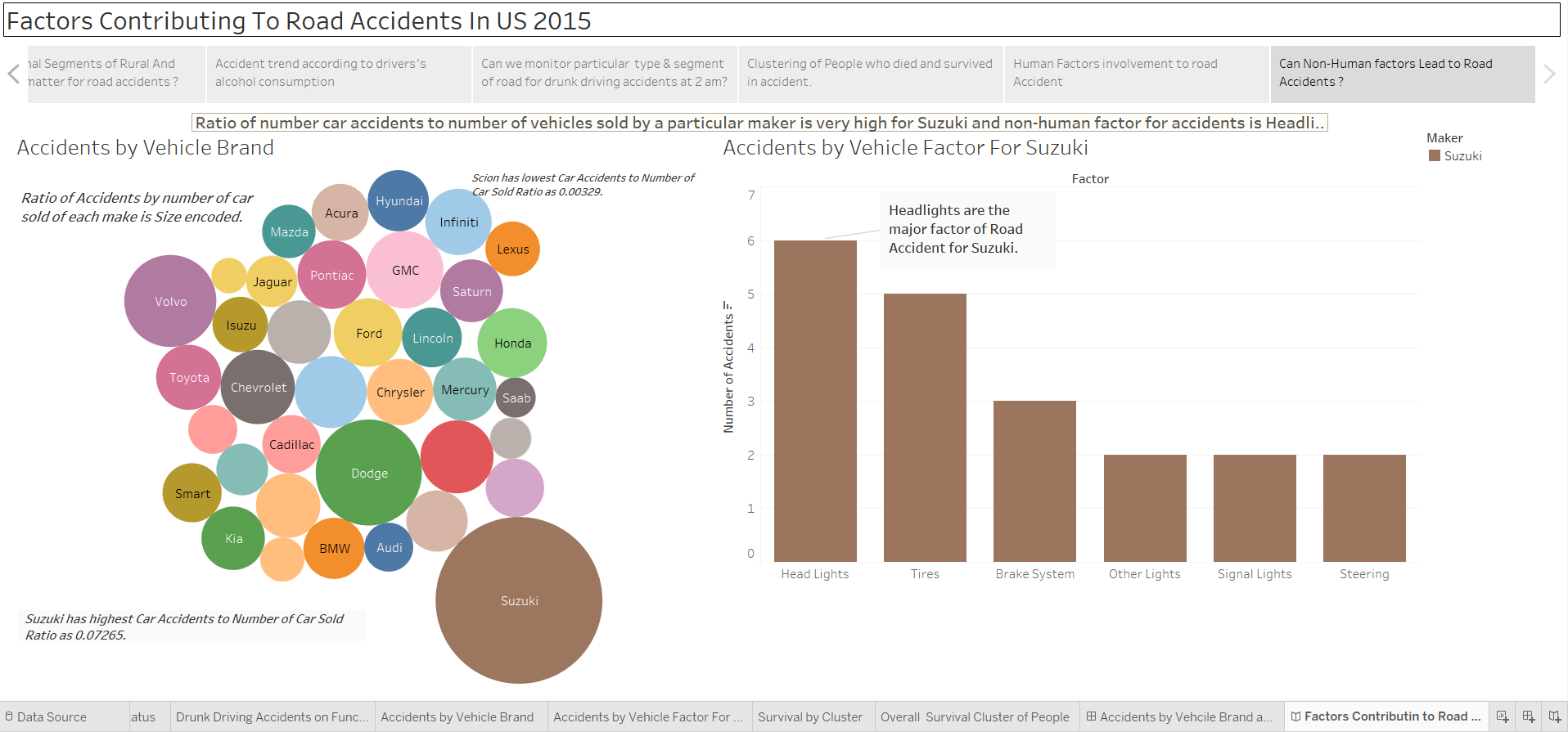


**Visual Encoding**

Box plots are used for this visualization. Yes at x-axis represents whether the person survived or not. Y-axis has various attributes of the victim i.e. gender, seating position, restraint used, drugs use, drinking and driving and age.

**Insight:** People seating on the Front Seat are at high risk of death than the ones seating in the other seats

9. Can Non-Human Factors Lead to Road Accidents?



**Visual Encoding**

This dashboard has two visualizations. The bubble-chart depicts accidents by vehicle brand. The size of the bubble represents the calculated field(ratio) of a number of accidents by brand. Color represents different car manufacturer.

The second visualization has a bar graph showing factors affecting road accidents. This visualization is for Suzuki. We have used a filter for Suzuki which is color-coded with brown color.

**Insight:** Ratio of number car accidents to a number of vehicles sold by a particular maker is very high for Suzuki and non-human factor for accidents is Headlights.