

2021

# Final Project Report

US Road Crash Data Visualization

Umakanth Sai Balguri

(U01403700)

Data Visualization | Prof. Dr. [Johannes Reichgelt](https://usflearn.instructure.com/courses/1509474/users/3869216)

Table of Contents

[Final Project Report 0](#_Toc70953987)

[1. Introduction 2](#_Toc70953988)

[2. Methodology 3](#_Toc70953989)

[2.1. Background of NTHSA and its Reliable Data collection Mechanism: 3](#_Toc70953990)

[2.2. Data Cleaning and processing for Analysis in Tableau 4](#_Toc70953991)

[3. Analysis 4](#_Toc70953992)

[3.1. US Road Crash - Public View 5](#_Toc70953993)

[3.1.1. Details of Visualizations 5](#_Toc70953994)

[3.1.2. Insights 6](#_Toc70953995)

[3.2. US Road Crash – Transport Department View 7](#_Toc70953996)

[3.2.1. Details of Visualizations 7](#_Toc70953997)

[3.2.2. Insights 11](#_Toc70953998)

[3.3. Vehicle Crash History Report 12](#_Toc70953999)

[3.3.1. Details of Visualizations 12](#_Toc70954000)

[3.3.2. Insights 13](#_Toc70954001)

[5. References 14](#_Toc70954002)

# Introduction

Road traffic crash data are useful tools to support the development, implementation, and assessment of highway safety programs that tend to reduce road traffic crashes. Road traffic crash data gathering aims at gaining a better understanding of road traffic operational problems, locating hazardous road sections, identifying risk factors, developing accurate diagnosis and remedial measures, and evaluating the effectiveness of road safety programs across various geographical areas. Furthermore, the data can be visualised and can be used by many agencies and businesses such as: insurance companies to identify previous vehicle history, insurer history; road safety researchers to access traffic crash reliable database; decision makers to develop long-term, state-wide strategic plans for traffic and highway safety; and highway safety administrators to help educate the public. Given the practical importance of vehicle crash data, this project aims to find the following useful insights:

* From a Road User Standpoint:
  + How many Crashes have been recorded in each county in United States?
  + How many of these crashes have been Fatal?
  + What are the chances of survival of a Pedestrian/Bike-pedestrian survival in a crash when compared to a motor vehicle Occupant?
  + Have the deaths in US due to Road accidents increases or decrease overtime. Which states have been consistently outperforming and which states have high fatality rate that the road user should be aware of.
* From the Transport Authorities Standpoint:
  + How many cases/deaths have been recorded in each county. What portion of these have been due to driving under the influence of Alcohol
  + Which brand/make have higher cases of crash? Had there been an improvement over time due to the improvement of safety standards?
  + What is the top-10 and bottom-10 states/counties with highest road crash incidents?
  + What are the major weather conditions leading to road crashes?
  + What are the peak hours for crashes? And is it consistent through the week? Is there a different pattern during the weekends?
  + Which age group is highly involved in road accidents?
* From a Used Car buyer Standpoint:
  + How many crashes was the vehicle involved in.
  + When did each of the crash take place?
  + Were the accidents major? Did it lead to fatality?
  + Which location did the accident occur? Was it on a highway or on the local streets? Which county was it and the state ?

# Methodology

The data required for this analysis should be complete and free from errors as otherwise would lead to biased inferences. Hence it was sourced from National Highway Traffic Safety Administration (NTHSA) website which provided CSV format of US road crash investigation data which can be locally downloaded. It also provides flexibility to call using api to source the data. However, standalone download files have been used instead due to the large size of the data and its static nature.

# Background of NTHSA and its Reliable Data collection Mechanism:

NHTSA uses data from many sources, including the Fatality Analysis Reporting System (FARS) which began operation in 1975. Providing data about fatal crashes involving all types of vehicles, the FARS is used to identify highway safety problem areas, provide a basis for regulatory and consumer information initiatives, and form the basis for cost and benefit analyses of highway safety initiatives.

FARS is a census of fatal motor vehicle crashes with a set of data files documenting all qualifying fatalities that occurred within the 50 States, the District of Columbia, and Puerto Rico since 1975. To qualify as a FARS case, the crash had to involve a motor vehicle traveling on a trafficway customarily open to the public, and must have resulted in the death of a motorist or a non-motorist within 30 days of the crash.

This multi-year analytical user’s manual provides documentation on the historical coding practices of the Fatality Analysis Reporting System from 1975 to 2019. In other words, this manual presents the evolution of FARS coding from inception through present. The manual includes the data elements that are contained in FARS and other useful information that will enable the users to become familiar with the data system.

The Manual: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813023

The DataSource Download Link: <https://www.nhtsa.gov/node/97996/251>

# Data Cleaning and processing for Analysis in Tableau

* There was no specific data cleaning activity that was required. However certain fields were required to be split to bring all the yearly data to single format.
* The data source was loaded to tableau
* Accident, Vehicle and Person data tables were joined using Union feature in tableau to have 2016-2019 data in single file.
* Union tables of Accident Vehicle and Person were joined using ST\_CASE (the unique identifier for a crash record).
* The data types which were incorrectly loaded i.e Date, geographical location due to duplicates were corrected to have final data source ready for analysis.

# Analysis

# US Road Crash - Public View

# Details of Visualizations

For Public View, 3 individual visualizations have been created.

* The first visualization, PV-1 (﻿US Road Crash Cases by County - Map View) plots total number of cases per county in US that took place between 2016-2019. As the map is zoomed in the labels are displayed for each county.

Diagram, engineering drawing

Description automatically generated

* The second visualization PV-2 (﻿Number of individuals involved and fatality per Month and Year) is a bar graph indicating total individuals involved in an accident and the number of fatalities. There is a filter for the type of road user to filter out the data.

Graphical user interface, chart

Description automatically generated

* The second visualization PV-3 ﻿Animated View of Total Fatalities per Month and Year) is a side-by-side bubble chart indicating the count of total fatalities per each type of road user. There is a drill down view from which user can obtain data pertaining to each month of the year. Also, this data is animated on state name, meaning, the data keeps changing for each state as the animation is played.

Graphical user interface, chart, application, scatter chart

Description automatically generated

* The dashboard created has all the filters and can be used to filter data on all three visualization being displayed. There is a small notes field at the bottom right corner which explains the details and usage of this dashboard

Graphical user interface, application, table, Excel

Description automatically generated

# Insights

* + How many Crashes have been recorded in each county in United States?

The Map view shows this information per county. Eg: County Cumberland in North Carolina recorded total cases of 173

* + How many of the crashes have been Fatal?

The Visualization PV-2 gives information of total crashes vs fatalities. Eg: For Driver of a Motor Vehicle, in 2016, total cases were above 200K and fatalities were around 39K

* + What are the chances of survival of a Pedestrian/Bike-pedestrian survival in a crash when compared to a motor vehicle Occupant?

Both Pedestrian and Bicyclist have very few chances of survival when compared to other road users involving in a crash. 98% of the Bicyclist involved in a crash die whereas 75% of the pedestrians involved in the crash die.

* + Have the deaths in US due to Road accidents increases or decrease overtime. Which states have been consistently outperforming and which states have high fatality rate that the road user should be aware of.

From 2016 to 2019, the deaths of driver and passenger of a motor vehicle have reduced by around 10%. However, for the other road users, the count remains almost same.

# US Road Crash – Transport Department View

# Details of Visualizations

For Transport Department View, 7 individual visualizations have been created.

* The first visualization, TDV-1 (﻿﻿Total Cases by County Animated View) plots cases each county, death and alcohol involve in US that took place between 2016-2019. There is an animated view for county and state, meaning the county wise data keeps displaying one after the other as the animation is played.

Map

Description automatically generated

* The second visualization, TDV-2 (﻿﻿﻿Total Cases by Make and Model) is a bar graph indicating total number of cases per make and model (which is a drill down view) that took place between 2016-2019.

Chart, histogram

Description automatically generated

* The third visualization, TDV-3 (﻿﻿﻿﻿Most Affecting weather conditions) is a circle marks where size shows % of Total St Case. The marks are labelled by weather. The data is filtered on Year, which ranges from 2016 to 2019. The view is filtered on Weather, which excludes Clear, Not Reported, Other and Reported as Unknown.

Chart, bubble chart

Description automatically generated

* The fourth visualization, TDV-4 (﻿﻿﻿﻿﻿States and Countys by highest number of Crashes-Top 10) is a table of Count of Case and Fatals broken down by STATENAME. The view is filtered on STATENAME (copy) with highest no of cases, which keeps top 10 of 51 members.

Graphical user interface, text, application, email

Description automatically generated

* The fifth visualization, TDV-5 (﻿﻿﻿﻿﻿States and County by least number of Crashes-Bottom 10) is a table of Count of Case and Fatals broken down by STATENAME. The view is filtered on STATENAME (copy) with lowest no of cases, which keeps bottom 10 of 51 members.

Graphical user interface, text, application, email

Description automatically generated

* The sixth visualization, TDV-6 ﻿(Time of day with Peak Case count on each weekday) is a table showing top three hours each weekday with highest recorded cases.

Graphical user interface, application

Description automatically generated

* The seventh visualization, TDV-7 (﻿Count of Crashes by Age Group) is a table showing count if crashes per each age group of the driver.

Graphical user interface, application

Description automatically generated

* The dashboard created has all the filters and can be used to filter data on all seven visualization being displayed. This has been created to give user single screen view of all the important data they would require on day-to-day basis.

Graphical user interface, website

Description automatically generated

# Insights

* + How many cases/deaths have been recorded in each county. How many of these have been due to driving under the influence of Alcohol

The Map view shows this information per county. Eg: County TUSCARAWAS in Ohio recorded total cases of 31 which resulted in 34 deaths and 18 persons were under influence of Alcohol

* + Which brand/make have higher cases of crash? Had there been an improvement over time due to the improvement of safety standards?

Ford has the highest number of cases of crash followed by Chevrolet and Toyota. And Ford 150 has highest number of cases recorder (around 31K). Yes Crashes have decreases by slightly from 2016 to 2019. However, none of the brand has shifted their position in the graph.

* + What is the top-10 and bottom-10 states/counties with highest road crash incidents?

Top 10 counties include California (4705), Texas(3165), Florida(2098), Ohio etc and Bottom 10 states include District of Colombia, Vermont, Rhode Island etc. The drill down gives counties with lowest cases in these states.

* + What are the major weather conditions leading to road crashes?

Cloudy followed by Rain and snow are the major weather conditions leading to road crashes.

* + What are the peak hours for crashes? And is it consistent through the week? Is there a different pattern during the weekends?

09:00 PM to 09:59 PM on Saturday recorded the highest number of road crashes (1362). On weekdays, most of the crashes occur during 6 to 7 PM.

* + Which age group is highly involved in road accidents?

Age Group 20-29 is highly involved in accidents, they account for 114K accidents of the total 136K accidents recorded in 2016-2019

# Vehicle Crash History Report

# Details of Visualizations

* The Vehicle Crash History Report Visualization provides a map view of all the cases where the vehicle (with the specified VIN no) has been involved in an accident. The label on the map gives important information such as record locator, date of the event, crash location, road, county, state, number of deaths occurred, and the vehicle information.

A picture containing graphical user interface

Description automatically generated

# Insights

* + How many crashes was the vehicle involved in.

The Map View shows now of crash incidents for a selected vehicle VIN. For Eg: VIN:1B6MM36C7PS2 has 4 crashes recorded on various roads in Kansas

* + When did each of the crash take place?

A date of crash has been added to the label indicating day, month and year. Eg: VIN:1B6MM36C7PS2 has 4 crashes recorded on

Oct 23, 2019

Sep 10, 2018

Aug 28, 2017

Oct 27, 2016

* + Did it lead to fatality?

No of deaths occurred in each incident has been added to the label

Eg: VIN:1B6MM36C7PS2 has 4 crashes recorded on and the deaths were:

Oct 23, 2019:1

Sep 10, 2018:1

Aug 28, 2017:1

Oct 27, 2016:0

* + Which location did the accident occur? Was it on a highway or on the local streets? Which county was it and the state ?

The road, county, state has been added to the label.

Eg: VIN:1B6MM36C7PS2 has 4 crashes recorded at:

Oct 23, 2019:CR-3 in CHEYENNE county, Kansas

Sep 10, 2018: QUINDARO RD in WYADATTE county, Kansas

Aug 28, 2017: I-70 in GEARY county, Kansas

Oct 27, 2016: 140RD in SMITH County, Kansas

1. **Conclusion**

With the clear data visualization of Crash data obtained from NTHSA, the following conclusions can be made:

* Although Florida is geographically smaller, less densely populated, the number of road crashes are quite high.
* Los Angeles county in California alone records 2823 cases in 4 years which is half of the state’s recoded and equal to third highest in the list.
* Almost all the Bicyclist involved in Crash are dead. Hence both motor vehicle drivers and bicyclists should be extra vigilant to each other.
* The cases involving bicyclists are recorded more in densely populated cities. So, drivers should be extra careful while driving through the city streets.
* Ford has recoded highest number of crashes as it is most sold brand. Although Toyota is the second most sold, Chevrolet recorded the second highest number of crashes. Authorities need to investigate into this to improve safety standards of this vehicle.
* Nearly half of the accidents involved alcohol. Authorities should bring in some strict measures to avoid dunked driving.
* Cloudy, Rain and Snow are the weather conditions leading to crashes. Road users should be extra careful when using the road under these conditions.
* Saturdays, 09 to 10 PM and 6 to 07PM on other weekdays, there are high volumes of crashes. Hence there should be temporary speed checks and patrolling during these times.
* Age group 20-29 are accounted for more than 90% of the road crashes. Hence, they should be put to continuous checks by introducing driving habit logging application.
* As a free alternative to CARFAX, the Vehicle Crash History Report Visualization can be used to obtain information about an used car previous crash history to make a sound decision on purchasing it.

# References

* <https://www.statista.com/statistics/264362/leading-car-brands-in-the-us-based-on-vehicle-sales/>
* <https://www.edgarsnyder.com/car-accident/who-was-injured/teen/teen-driving-statistics.html#:~:text=5%2C864%20teen%20drivers%20ages%2015,day%20from%20motor%20vehicle%20injuries>.
* <https://www.nhtsa.gov/ratings?cmpid=TSGSNF0317&gclid=Cj0KCQjwvr6EBhDOARIsAPpqUPG5NzcIGd7pg3JbCV69CH7Sak_48LY9yEWSaA7kvxQpsQRe5sa9J0UaAk5NEALw_wcB&gclsrc=aw.ds>
* <https://www.nhtsa.gov/crash-data-systems/crash-report-sampling-system>
* <https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm>