

# RECOMMENDATIONS FOR AUTO INSURANCE

**Professor: Antonio Lora**  
**MIS 515 Business Analytics Project**

# Table of Contents

<b>Executive Summary</b>	3
Objective	3
<b>Description</b>	3
Background Information	3
Products and Services	5
<b>Project Detail</b>	5
Methodology	5
Dataset Description	5
Big Data Architecture and pipeline	6
Data Visualization	Error! Bookmark not defined.
<b>Recommendations &amp; Future Scope</b>	17
<b>Cost Analysis</b>	18
<b>References</b>	20
<b>Appendix</b>	21

## Executive Summary

---

Recoflexi is a data analytics firm that believes in making value with data. The company helps its customers develop recommendations based on insights drawn from the data. The company has a dedicated team working towards understanding and mapping business policies, requirements of its clients and designing a big data solution for them. The product developed by the company using crash data is targeted towards auto insurance companies that would benefit from grasping the extent to which external and driver-related factors are critical to causing fatalities in vehicle accidents which will help clients analyse the necessary attributes that need to be factored in when determining insurance premiums.

It is expected that the product developed will allow the client

- to interact with the data allowing them to make simple queries on the data
- to see business intelligence reports that give key performance indicators
- to develop recommendations through Spatio-temporal analysis provided through interactive dashboards

## Objective

The overall objective of the solution is to map out fatalities in road accidents against the significant variables and gain good insights on how fatalities vary by their geographical locations based on hour of the day, day of the week, month, highway, Tollway, the vehicle makes/model/type. The goal is to thoroughly analyze & understand the factors affecting fatalities that will provide a sound groundwork for Recoflex's working prototype which will consist of interactive dashboards allowing customers to perform ad hoc analysis on their data.

## Keys to Success

- Provide reliable visualizations and insights to clients particularly new entrants in the market giving them a competitive edge
- Allow easy interaction with the interactive dashboards allowing quick queries, visualizations and drill through the data based on temporal or spatial elements
- Cost-efficient big data architecture solution taking into consideration infrastructure (hardware) available to the client
- When clients are already established players in the market, value addition would be to better their existing business intelligence capabilities to readapt quickly to the changing demands of the market

## Description

---

### Background Information

With ever-increasing Auto Insurance companies always competing to establish a monopoly in the market, it is absolutely necessary for the niche companies to better their existing BI capabilities by catering to the fluctuating demands of the market and for the new entrants to be able to penetrate into less explored segments in order to gain unique insights of the market, setting their BI unit apart from the older firms .

The core idea of our product is to utilize crash data, a publicly available dataset in order to identify trends and patterns in fatalities, analyze hotspots for the crashes and identify temporal and spatial features that play a critical role in these crashes. The end deliverable includes interactive dashboards and business intelligence reports that would facilitate general trends and visualizations specific to a region and time frame and that would help determine if a particular customer falls into which of the three categories and what custom product might suit him/her the best.

Auto-Insurance companies incur losses known as unwriting losses occur on the insurance contracts in which the company had to pay claims. When the claims are more than the premiums paid by the insurers underwriting loss occurs. The company loses money because of mispricing insurance through underestimation of the risks. Thus appropriately identifying the risks associated with each insurer is vital for the insurance company to not lose money. Chart 1 shown below is the losses Incurred for Auto-Insurance companies from 2014-2018.



*Figure 1: the losses Incurred for Auto-Insurance companies from 2014-2018*

Auto insurance companies need to recognize and customize their products according to the specific segments of customers. A quick look at the loss trends in the auto insurance sector reveals that there has been a steady increase in loss. There have been around 30% increase[3] in losses between the years 2014 to 2018. From the graphs shown in Section 3, it is clear that the percentage increase in losses was highest in the years 2014-15 and 2015-16(around 9% and 12%) and has come down to 3% by 2017-18 showing that auto insurance companies are proactively applying measures to tackle the situation. The availability of large volumes of data like mileage data, driving behavior data and a myriad of other data available with the increased number of sensor/IoT devices on board the new generation vehicles has now paved way for ‘Usage-based Insurance’(UBI)[4]. Insurance Companies are looking at ways to utilize this data in order to customize their products and develop flexible policies for their consumers and cover more segments of the market.

The traditional auto insurance relies on actuarial studies of historical data relating to credit-based insurance scores, personal characteristics of the driver, vehicle use, etc., and has fixed costs associated with the policy. However, UBI makes use of data obtained from telematics devices that record miles drive, rapid acceleration, GPS data and so on to determine if a customer is a low, medium or high-risk customer based on the analytics of the data obtained. The concept of UBI must evolve and must not be limited to the data of the drivers or vehicles alone. As the number of machine to machine interactions increases and the concept of Intelligent Vehicles, a reality there are huge volumes of untapped potential lying in all the big data that millions of IoT devices generate every minute.

However, the concept of UBI is evolving and it has been facing legal challenges to limit the data the Insurance Companies collect and declare tracking procedures. Thus, amidst all the legal tangles and peer competition, drawing the most useful insights with available data along with other publicly available data is very critical to the success of their business.

## Products and Services

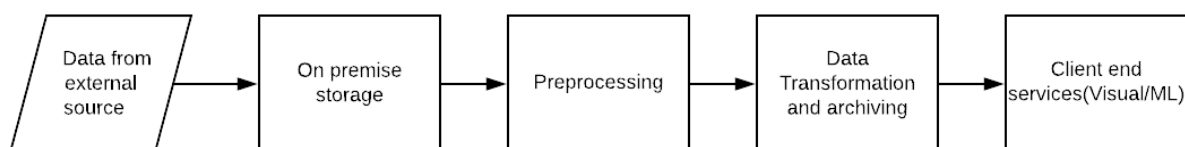
One of the primary services offered will include an interactive dashboard developed in Power BI which will help insurance companies interact with the data provided. The dashboard will also help develop a deeper understanding of how factors like time, weather, geographical location, demographic factors, etc affect crashes. It will allow insurance companies to perform geospatial analysis on the data by selecting a time frame and geographical region they want to look at. The analytical reports can be generated by the region or time and can be used to analyse the data specific to a county or city within a state of their interest. Besides the visualizations, a detailed performance report that includes the average, median, mean number of crashes aggregated over year, month, day or hour can also be made available through big data platforms like Apache Spark. The second line of products involves a machine learning model that helps with the identification of customers into low, medium and high-risk customers based on well-performing clustering algorithms like DB Scan allowing further data analytics. This module is in the developmental phase. The developed product will use publicly available datasets, thus helping companies save huge amounts of money otherwise required to acquire data from other sources. The product is user-friendly requiring little to no programming knowledge. The aims to give insurance companies a better edge over their fellow competitors and also help make insurance policies more affordable to low-risk customers while taxing high-risk customers for their risky behavior.

## Project Detail

### Methodology

The general methodology consisted of the following steps:

- ❖ the collection and archiving of the FARS dataset from the official website on inhouse premises
- ❖ the data pre-processing involved removing null values, inserting vehicle names, inserting names and zip codes corresponding to various states and counties to form sxxx where s corresponds to the state number and xxx corresponds to the county code. In the case of state California and county 1, the code would be 6001. The state name was also changed from the number code to the name string. Snapshots of the data loaded into PowerBI
- ❖ data transformation included various joins between datasets based on the state case and the year of occurrences, grouped aggregations of data to get a net number of fatalities and an average number of fatalities by county, city, by the year, month, day of the week and hour.
- ❖ data archival involved storing the data as tables and querying later using Databricks services.



*Fig 2: Shows the general framework of the product*

### Dataset Description

Two datasets related to vehicle crashes were analyzed. Dataset 1 is sourced from a dataset called the “Fatality Analysis Reporting System (FARS)” and is part of the U.S. Department of Transportation (USDOT)/Bureau of Transportation Statistics' (BTS') National Transportation Atlas Database (NTAD). It consists of fatalities recorded in 50 states, either of the motorist or a non-motorist occurred within 30 days of the crash. The data set also contains information regarding crash characteristics and the environmental conditions at the time of the crash. Each record

represents one crash and its respective crash characteristics. The dataset has 34,011 records and 55 attributes. Dataset 2 is sourced from the Traffic Safety Data and Analysis on the Iowa DOT website. The Data contains persons involved in the aforementioned accidents. The data on these persons involved in the fatalities is used to identify if either the individual's characteristics like age, race, gender have a role to play in determining the crash frequency or whether vehicle characteristics including the brand of the car, its year of make and so on play a role or not. The description of the important variables has been provided in the appendix.

## **Analytical Questions**

A few questions that are critical in the development of our product have been mentioned below:

- Which are the hotspots for accidents of each state/city/county?
- What are the key performance indicators or metrics that give useful insights into the crash analysis data?
- Which customers are likely to be categorized as low, medium and high-risk customers and what would be the criterion for the same?
- What are the general trends with respect to the category of the accident recorded, frequency of accidents over the day/week/year and so on?

## **Big Data Architecture and pipeline**

Azure DataFactory can be used to build a self-sufficient ecosystem having ingestion, transformation and storage options all accommodated and managed efficiently. An empty canvas can be used to drag and drop elements. The 'Create Pipeline' Option helps specify input and output connections routing the data efficiently from source to sink. In our use case scenario, we have an input module that reads the input from the HDFS or any other local storage or database and it is pooled into the Azure Databricks Module. There is a facility to fetch the authentication token from our Client's Databricks Account and can be inserted in the azure data bricks module settings. A spark notebook may thus be created in Databricks and efficiently accessed from here. Just as the source was specified the sink may also be specified to be SQL Database or Hive etc which in turn can be connected to the PowerBI for visualization. Once the pipeline is built the data is efficiently routed from source to sink. The architecture solution we propose is very flexible and can be easily readapted depending on the infrastructural requirements of the client. Azure Data factory allows the use of a large variety of sources and sinks as shown in the figure below.

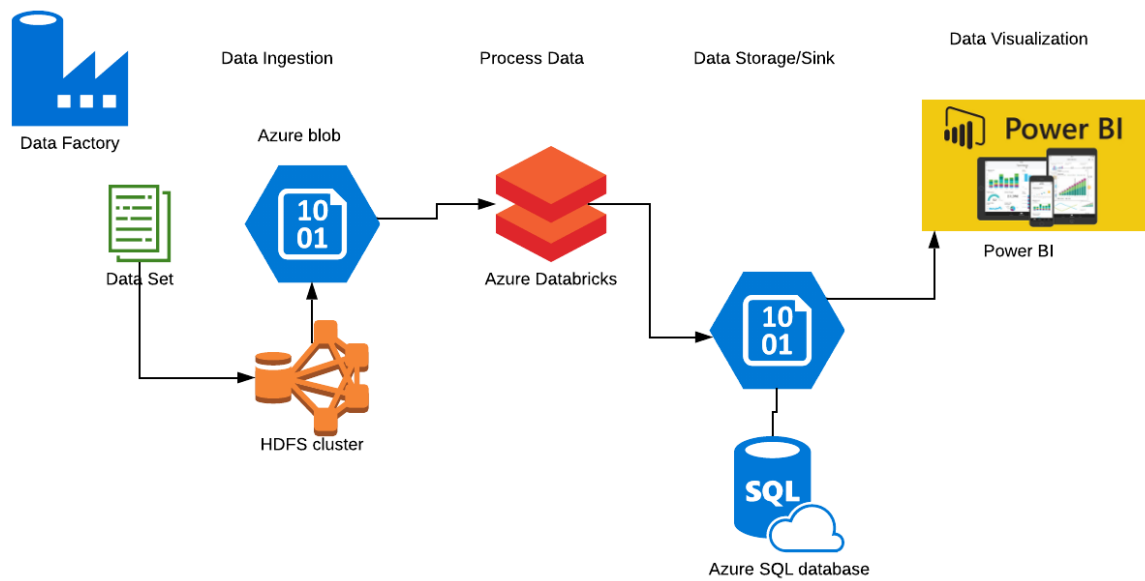


Fig 3a: Big Data Architecture Solution to ingest, process and visualize data



















 FTP	 File System	 Google AdWords	 Azure Data Lake Storage Gen2	 Azure Database for MariaDB	 Azure Database for MySQL
 Google BigQuery	 Google Cloud Storage (S3 API)	 Greenplum	 Azure Database for PostgreSQL	 Azure File Storage	 Azure SQL Database
 HBase	 HDFS	 HTTP	 Azure SQL Database Managed Instance	 Azure Synapse Analytics (formerly SQL DW)	 Azure Table Storage

Fig3b: Sample Data sources/Sinks available for connection in Azure data Factory

## Description of Components of the Big Data Architecture

## Data Ingestion:

Data can be pulled in from an external API using Azure Databox Blob Storage or alternatively downloaded into a cluster/system setup on-premises using simple shell scripts and uploaded into an azure blob. The data factory has options to include the path of the blob( through its URL) in the pipeline. An illustration of the setup process has been shown in the figures below.

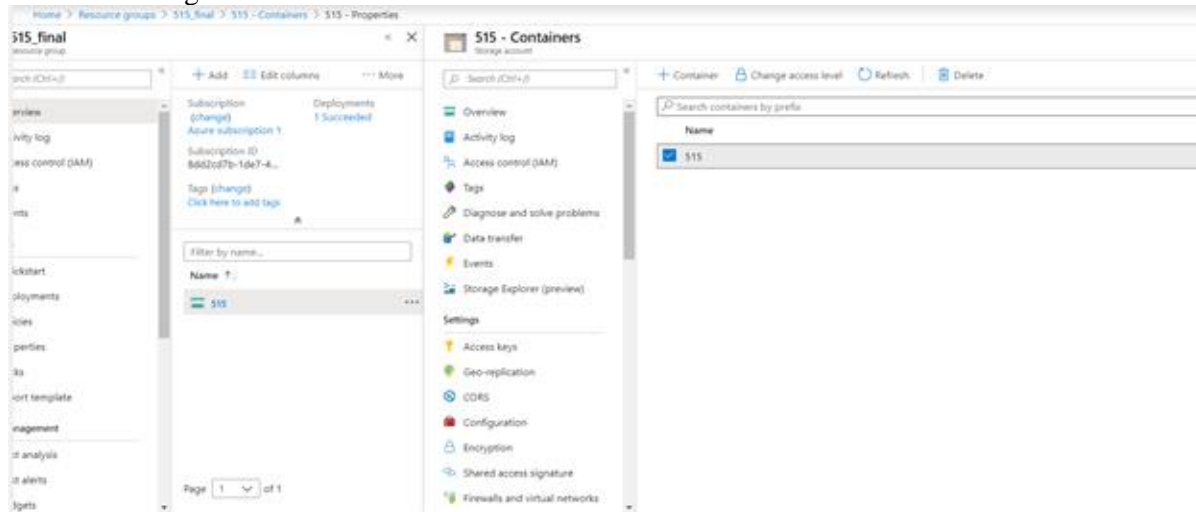


Fig 4a: Shows creation of a blob storage

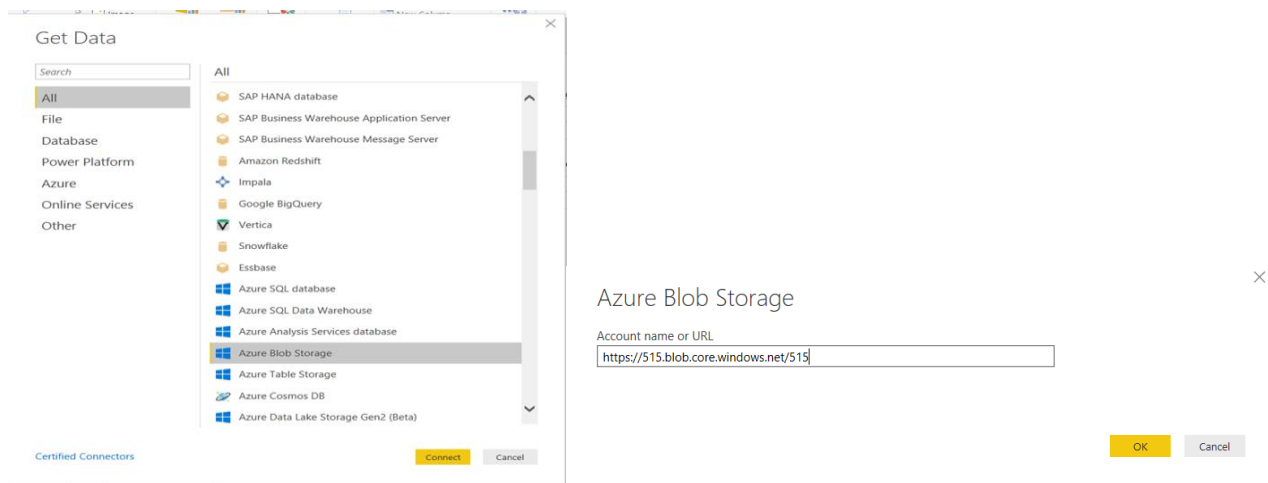


Fig 4b: Shows how the blob storage is connected to PowerBi

## Data Transformation:

The data transformation is carried out using Apache Spark on the Databricks platform. Transformations include joins between two datasets like the 'accidents.csv' and 'persons.csv' files in our case and also include aggregation done over temporal and spatial elements to find total, average or median number of fatal. It can be used for reporting these



parameters in a systematic manner. These transformed datasets can be pushed into a SQL database and can be queried or connected with front end BI and Visualization tools. An illustration of the notebook has been shown below. The data transformations involved deriving the sum, average, median of fatals. The aggregations were performed over the years, month, day and so on. A sample of the aggregation has been shown below. This data can be either stored as tables to be queried later or may remain in the notebooks allowing anyone with access to the workspace to be able to do quick visualizations of the data or generate a performance report. The figures below show an illustration of how the notebook can be used to effectively see the data aggregated over the month, year and city and subsequently filter out values corresponding to the city code 1370.

```

1 # File location and type
2 file_location = "/FileStore/tables/accidentsstate6.csv"
3 file_type = "csv"
4
5 # CSV options
6 infer_schema = "true"
7 first_row_is_header = "true"
8 delimiter = ","
9
10 # The applied options are for CSV files. For other file types, these will be ignored.
11 df = spark.read.format(file_type) \
12     .option("inferSchema", infer_schema) \
13     .option("header", first_row_is_header) \
14     .option("sep", delimiter) \
15     .load(file_location)
16
17 display(df)

```

▶ (3) Spark Jobs

▶ df: pyspark.sql.dataframe.DataFrame = [\_c0: integer, STATE: integer ... 51 more fields]

Displaying 50 out of 53 columns. [Display all columns](#) (may affect performance).

_c0	STATE	ST_CASE	VE_TOTAL	VE_FORMS	PVH_INVL	PEDS	PERNOTMVT	PERMVT	PERSONS	COUNTY	CITY	DAY	MONTH	YEAR	DAY_WEEK
0	6	60001	1	1	0	0	0	2	2	79	0	1	1	2016	6
1	6	60002	1	1	0	1	1	1	1	71	0	2	1	2016	7
2	6	60003	2	2	0	0	0	3	3	37	1980	3	1	2016	1
3	6	60004	1	1	0	0	0	1	1	37	1980	4	1	2016	2
4	6	60005	1	1	0	0	0	2	2	67	0	3	1	2016	1

Fig 5a: Shows sample code in the DataBricks Notebook that reads the data

```

1 from pyspark.sql.types import *
2 # from iwzudfs import *
3 from pyspark.sql import functions as F
4
5 import datetime as dt
6 from datetime import datetime, timedelta
7 import os
8 df1=df.groupBy('YEAR','MONTH','CITY').agg(F.sum('FATALS').alias('fatals'))
9 display(df1)
10 display(df1.where(F.col('city')==1370))

```

▶ (5) Spark Jobs

▶ df1: pyspark.sql.dataframe.DataFrame = [YEAR: integer, MONTH: integer ... 2 more fields]

YEAR	MONTH	CITY	fatals
2016	1	1370	2
2017	9	1370	8
2018	1	1370	5
2017	10	1370	5
2017	7	1370	2
2017	2	1370	4
2016	5	1370	2
2016	9	1370	1
2017	4	1370	2

Fig 5b: Shows sample code in DataBricks that performs aggregation of fatals

## Data Consumption or Storage and Visualization:

The data can be stored in an output blob or alternatively put into a SQL database and accessed using a Business Intelligence tool like Power BI to draw useful insights. The visualizations have been provided towards the end of the document.

## Machine Learning Component

Azure ML Classic was used to build an intuition of what may be the most critical factors affecting crashes or accident fatality. Filter based feature selection module in Azure Classic ML Service provides multiple feature selection algorithms to choose from, including correlation methods such as Pearson's or Kendall's correlation, mutual information scores, and chi-squared values. Fisher method was used to get an intuition on which factors or variables play an important or critical role in determining the dependent variable fatality. Fisher score indicates how much of the information resides in a variable. When run for the accident data alone it showed that none of the variables had much information. However, on a scale of relative comparison, none of the intuitively expected variables showed any strong correlation with the dependent variable 'FATALS' also indicating that individually none of these factors could be having a strong impact but in combination with certain other variables were capable of strongly influencing the dependent variable. This became the motivation to find a solution to the problem through visualizations.

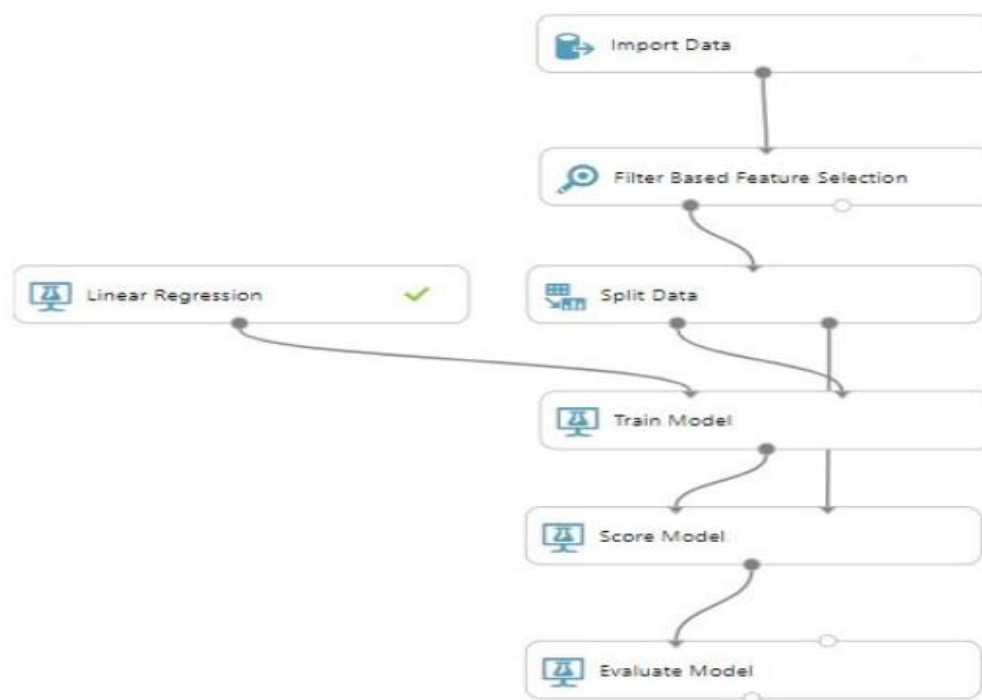


Fig 6a: Shows machine learning module built using Azure ML Classic

FATALS	PERSONS	PERMVIT	VE_FORMS	VE_TOTAL	PEDS	DRUNK_DR	PERNOTM
1	0.081624	0.080347	0.016118	0.014106	0.010656	0.009374	0.008384

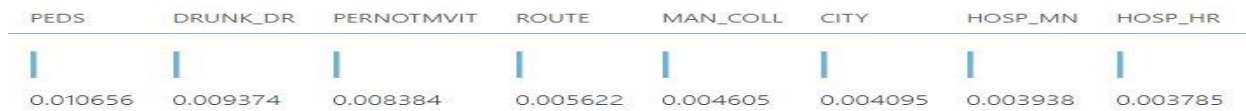


Fig 6b: Shows relative importance of variables obtained from Filter based feature selector in Azure ML

## Data Visualizations

Power BI and Tableau were used for visualizations since both of them have their inherent advantages and disadvantages. A sample dashboard has been shown below. The employee can select those that are of interest to him and the dashboard very clearly shows him what are the exact number of fatalities on that particular road and simultaneously shows whether these roads fall under the category of Interstate, highway, county road, etc based on Route ID. After getting an overall picture of the states with the highest fatalities we have limited the scope of further investigation to California in order to help us focus better on specific factors affecting crash fatality at a county or city level. The first part of the visualization description shows the overall picture of all the crashes throughout the states from 2016-2018. We have tried to show the insights based on spatial, temporal and demographic factors for all the states. Some of the insights drawn from the accidents dataset:

- When inspecting the fatalities by the state, the highest number of fatalities were for the states Texas, California, Florida, Georgia, North Carolina.
- Hotspots in California terms of counties are Los Angeles, Riverside, San Bernardino, San Diego, Orange, Sacramento, Kern, Fresno, San Joaquin in the decreasing order.
- When analysing the total number of fatalities that have occurred in the three-year window 2016-2018, Los Angeles seems to have as many as over 2000 fatalities with nearly 62% of them being on the Urban routes and 37.03% on rural routes.
- The age group of 18 to 27 was found particularly susceptible to accidents
- In terms of 'TWAY\_ID', the highest fatalities are on the I5, US101, SR99, I15, I10, I8, SR20, SR4, I405, SR74. Yearly patterns don't show much variation among these routes.
- Heat maps were developed to see the variation in fatalities with time in each of the counties. The time variation can be set to, day of the week or hour to see weekly, hourly trends. Los Angeles, Riverside, San Bernardino, San Diego, Orange, Sacramento, Kern, Fresno, and San Joaquin are all counties that have hotspots. When we examine the fatalities grouped by the day of the week we see Los Angeles constantly appearing a hot spot.
- As an example, the heatmap corresponding to odd days of the week has been shown in the figure in the latter half of the report. To understand these trends very clearly in terms of numbers one has to visit the static dashboards that have been developed. The heat map reports also bear another graph containing the number of fatalities by county. Selecting the county of interest allows one to see an animated heatmap for that county over the days of the week. Additionally, a slicer has been provided to see the numbers statically on the side corresponding to each day.
- Temporal analysis shows Saturdays and Sundays are most prone to accidents with Friday not very far behind. Probable causes could be that people travel to other places unfamiliar to them or even be hasty to reach home soon over the weekend.

- The make of the vehicle also seems to have a role to play. The models of vehicles seem to be having the highest fatalities in the three-year window included Chevrolet, Toyota, Ford and Toyota
- Vehicle models from 2015 onwards seem more susceptible to crashes. The one made 2008 to 2014 on a relative scale have been involved in fewer crashes.
- Demographic features like age and sex were other factors that showed an interesting impact. Those that were in the age group 18 to 27 were found to be most vulnerable to accidents.
- It may be seen that the fatal by harmful environment factor 12 is very high. Effect of weather factor 1 and the manner of collision 0 dominate. Weather 1 corresponds to clear conditions indicating that it may not be the dominant reason for collisions causing the user to probe further into the factors. The interactive dashboard allows a person to select a particular factor value and it will show the joint effect of other variables playing a role giving a better intuition of how these factors are interacting with each other.
- On route I5 the fatalities have increased from 2016 to 2018 while on route I10 it decreased from 2016 to 2017 and increased in 2018. I55 and SR99 show similar patterns with an increase from 2016 to 2017 and then a decrease in 2018. US 101 has almost been consistent over the three years.
- Insights on the spatial distribution of fatalities in the three windows helped determine the routes that have maximum fatalities and based on that static dashboards were developed to show any temporal trends that may exist on these routes using tableau. The idea behind this is that there is a constant watch out for these routes if the trends change or not.

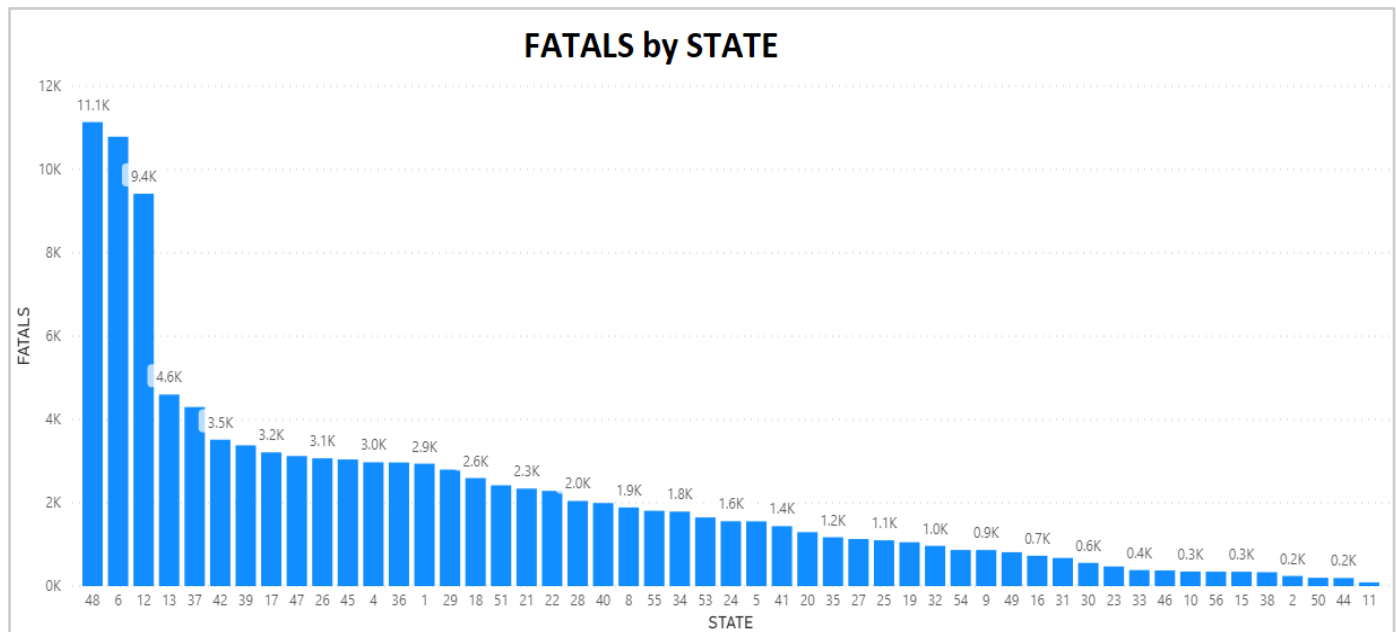
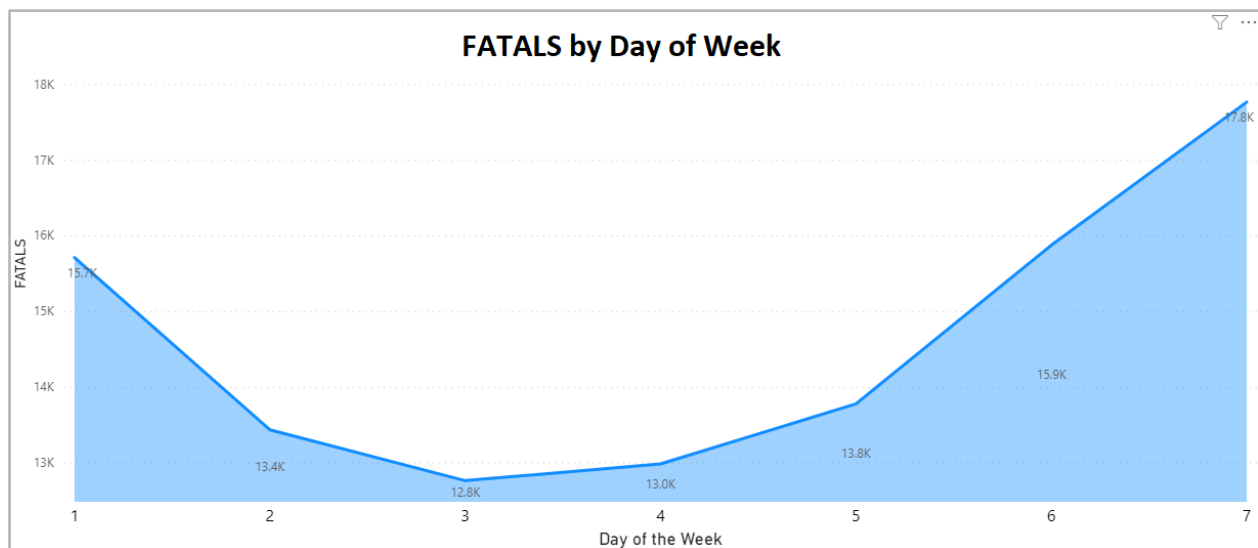
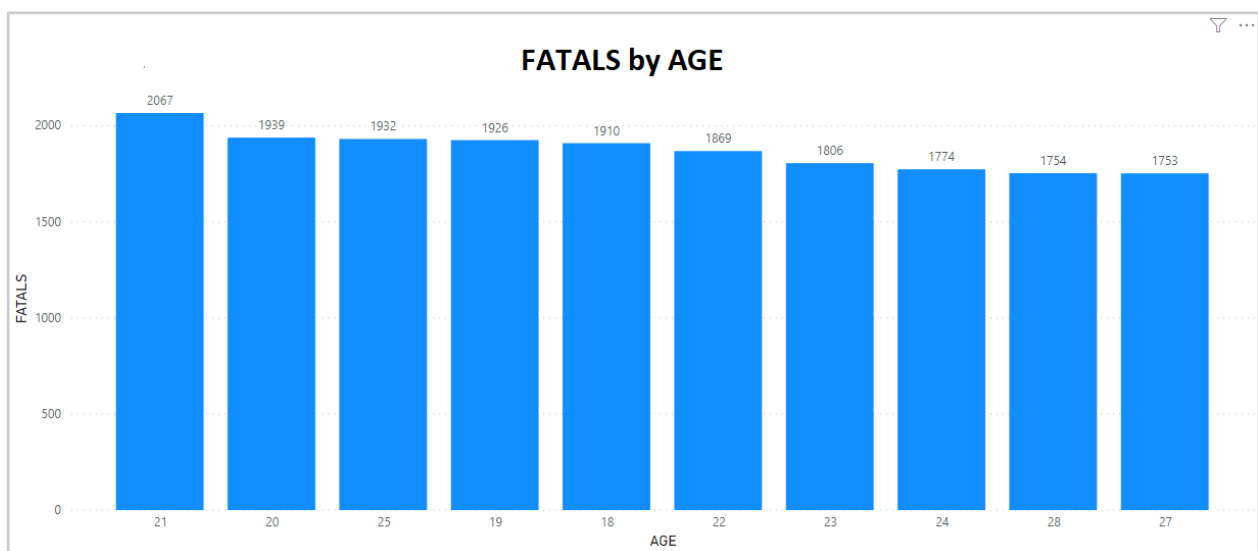


Fig 7: Shows top 5 most affected States: 48 - Texas, 6 - California, 12 - Florida, 13 - Georgia, 37 - North Carolina



*Fig 8: Shows most fatalities occur during Saturday(7) and then Sunday (1)*



*Fig 9: Shows fatalities plotted against age group*

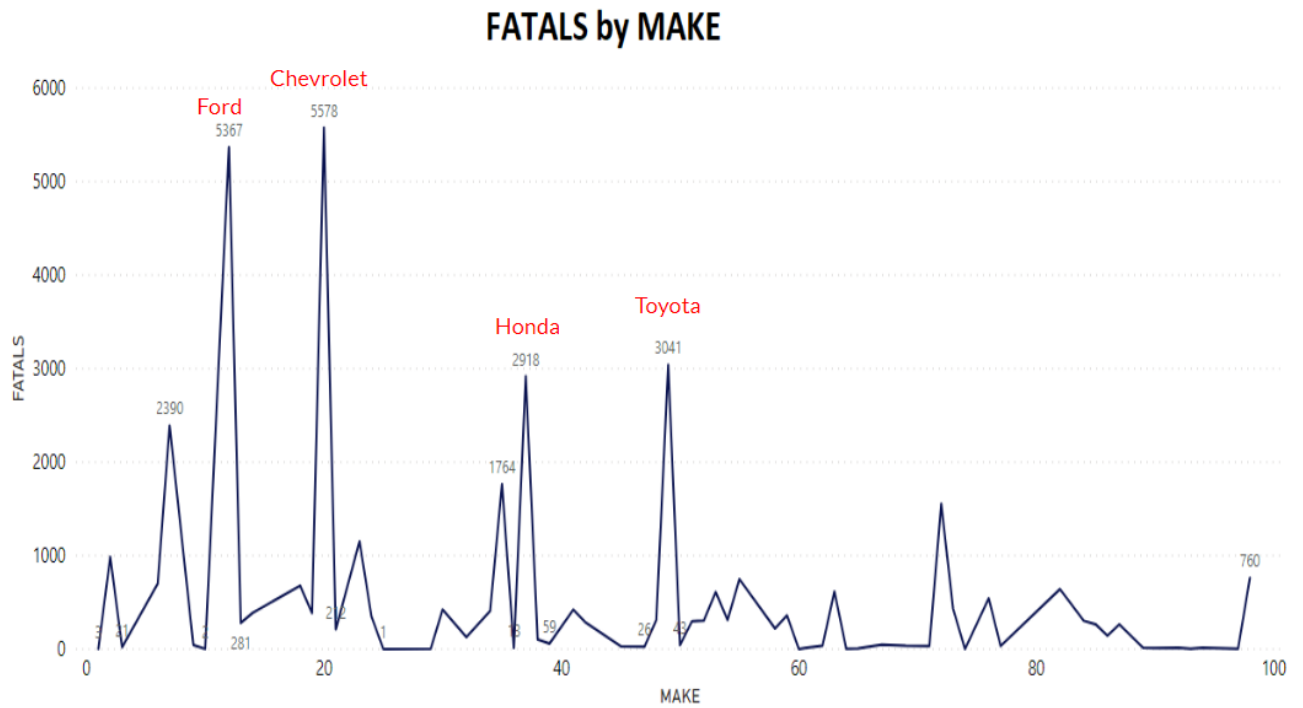


Fig 10: Shows fatalities occurred by Vehicle's Make. Most affected Vehicle - Top effected sequentially are Chevrolet, Ford, Toyota, and Honda

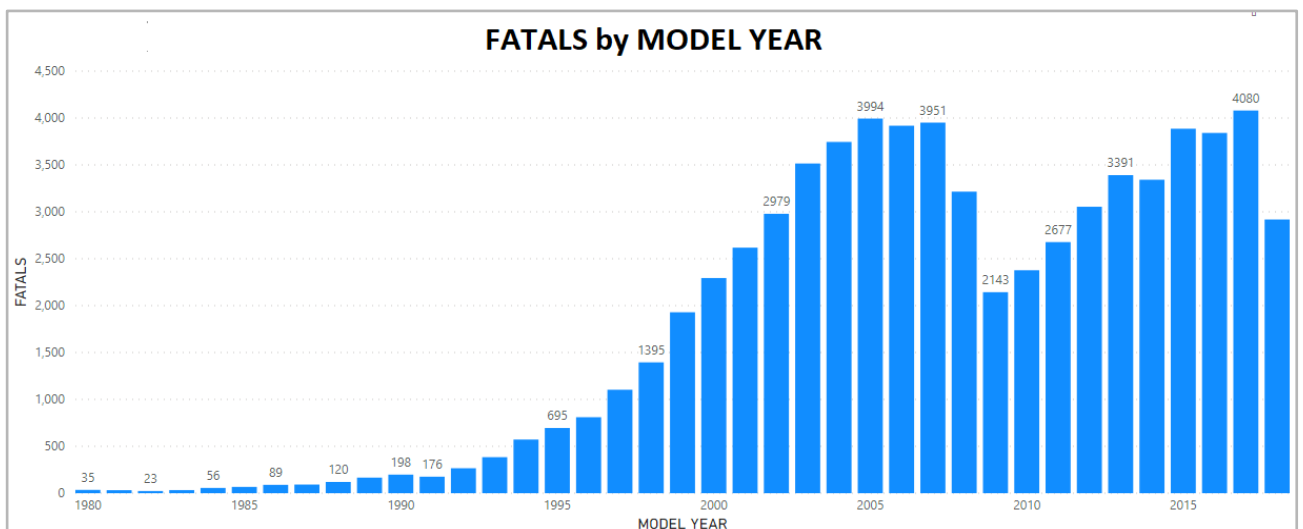


Fig 11: Shows fatalities by Vehicle Model Year

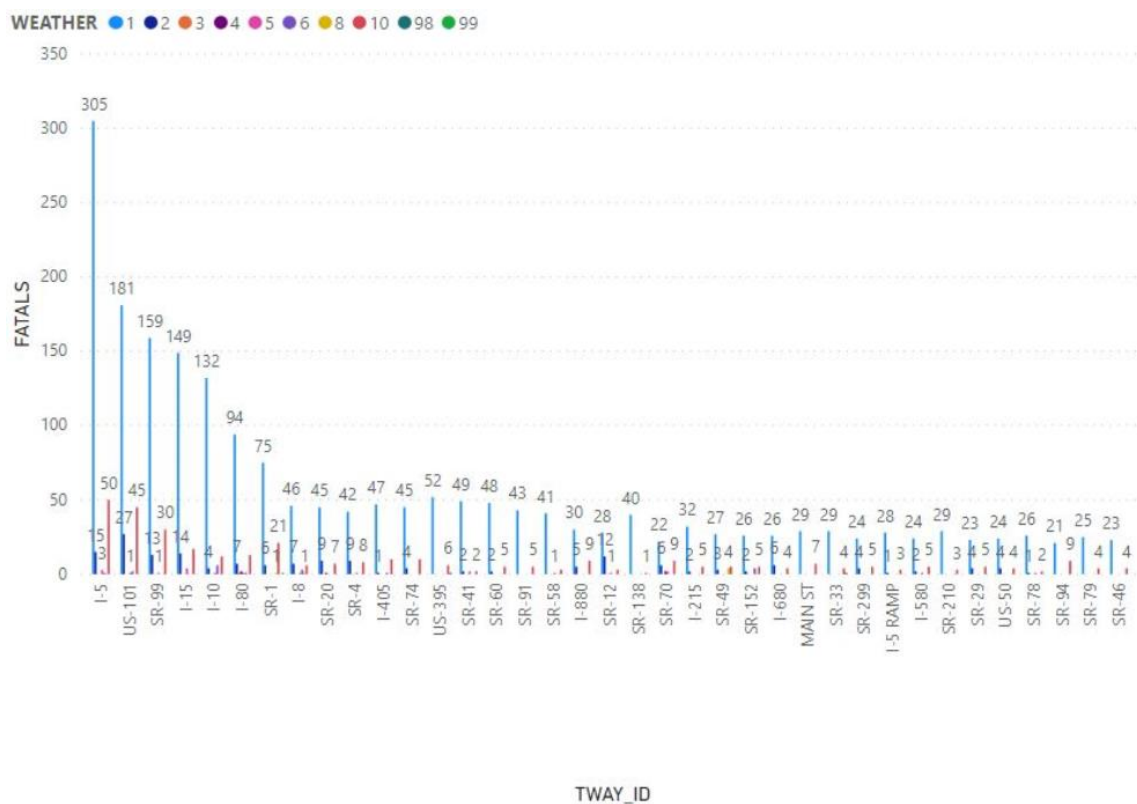


Fig 12: Shows the number of fatalities on different Tollways

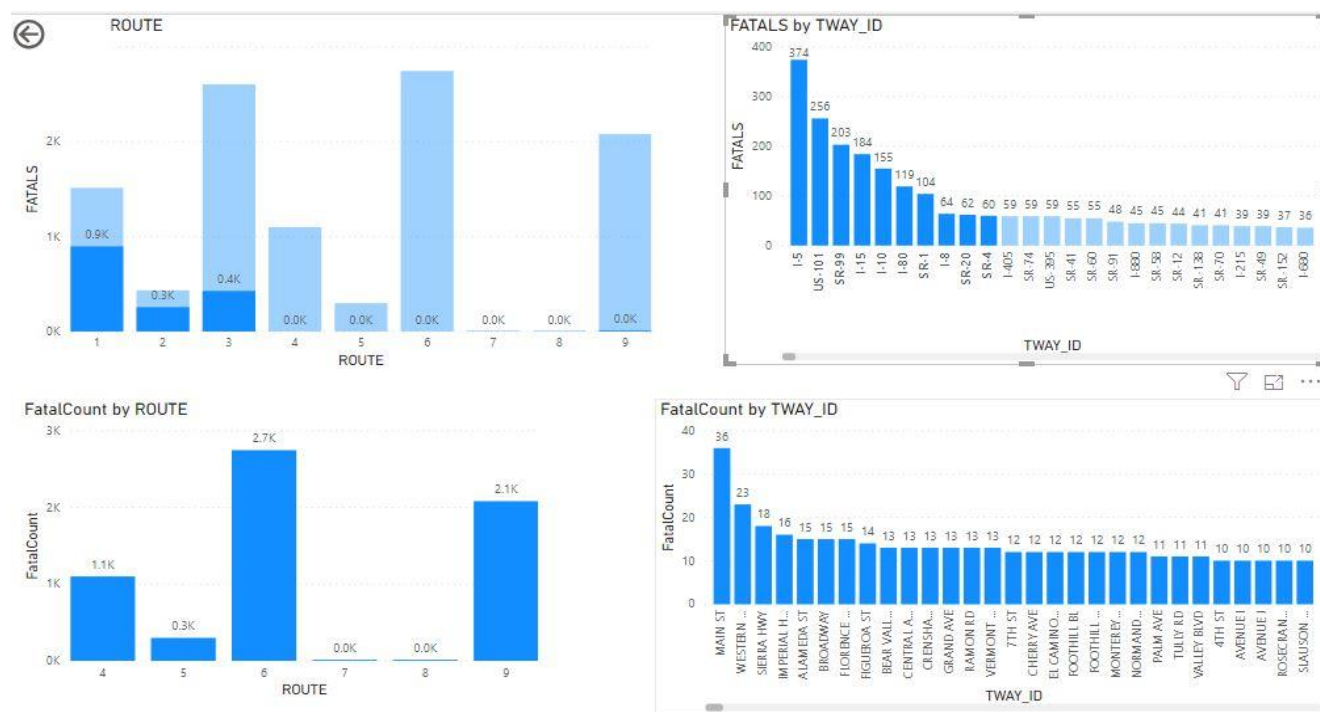


Fig 13a : Fatalis by the TWAY\_ID (Insights of Fatalities across state California from year 2016-2018)



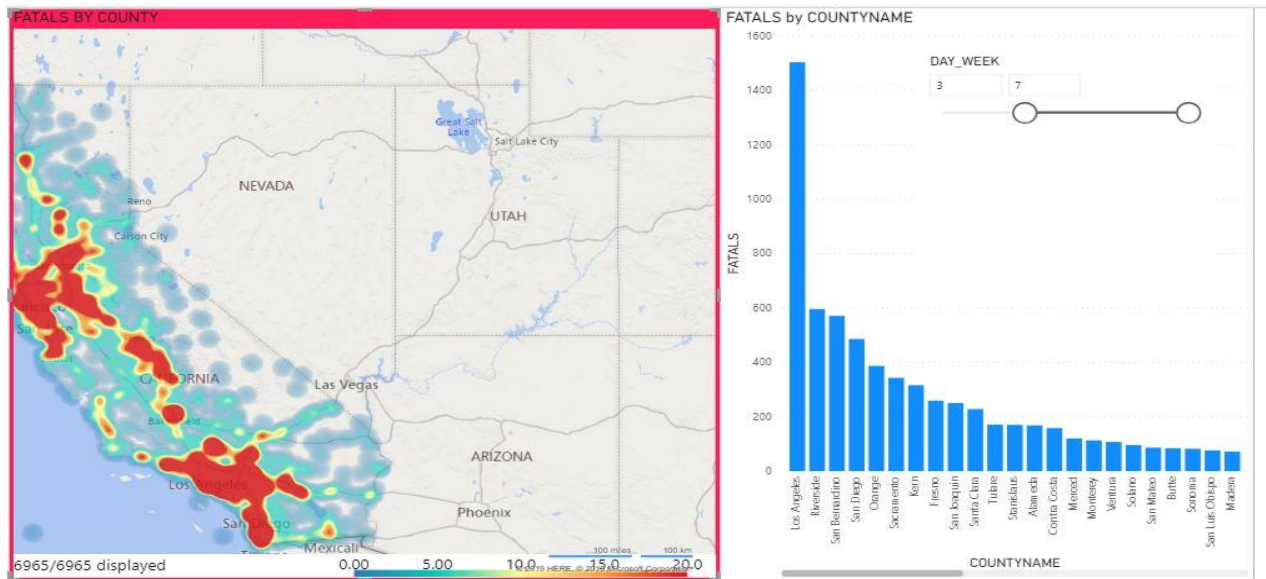


Fig 14a: County fatality heatmap with slicer showing the day of week

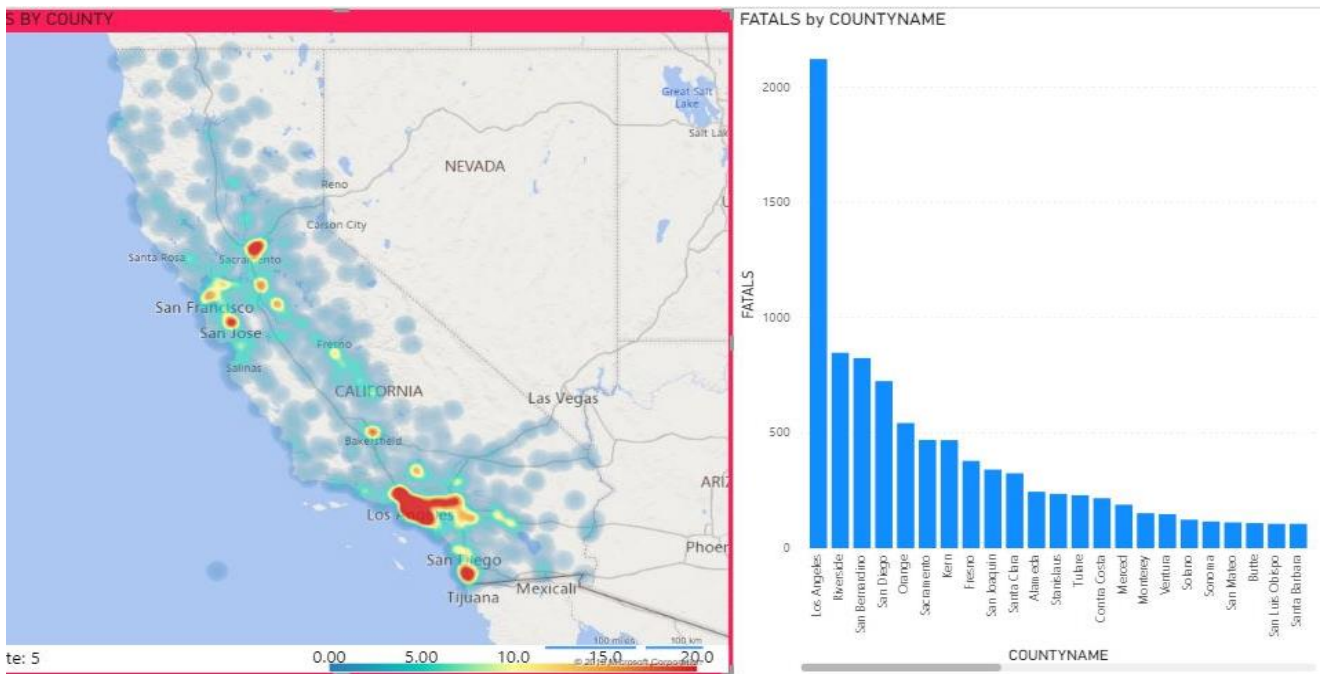


Fig 14b: County fatality heatmap for odd day 3 of the week

b)



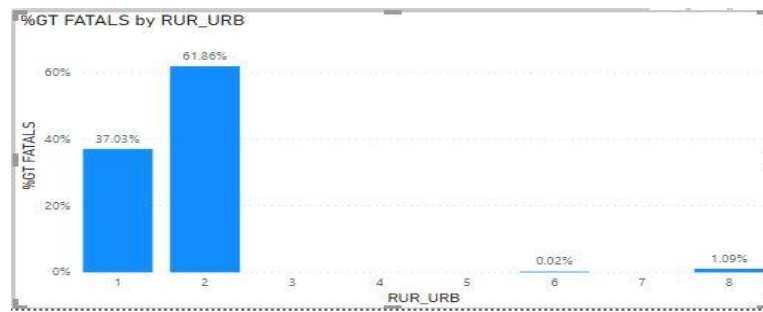


Fig 15: Percentage fatalities on the rural and urban routes

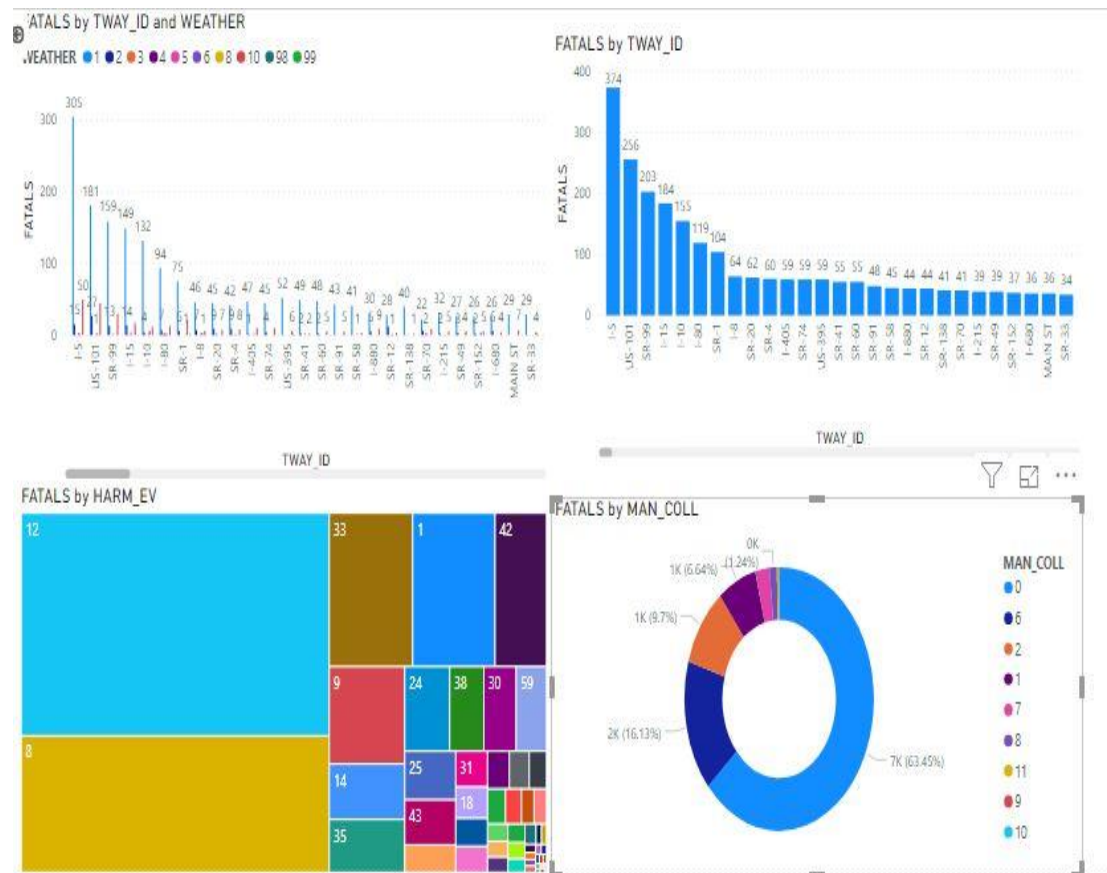


Fig 16: Effect of factors like harmful environment, weather and manner of collision


## Recommendations & Future Scope

From the spatiotemporal visualization and analysis, we find that certain days of the week are more susceptible to crashes and certain hotspots have been found along certain routes along the Interstate roads and State highways namely I-5, US-101, SR-99, I-10, I-8. As far as the demographic aspects are concerned, the age group of 18 to 27 was found particularly susceptible to accidents and males tend to be involved in a higher percentage of crashes than women. When considering the make of the vehicle there are particularly two models that were involved in more number of crashes which were Chevrolet & Ford. The most fatalities by model-year 2000- 2015. Insurance companies can use a combination of a few or all these factors when performing customer segmentation and designing flexible insurance solutions. The future scope involves using Machine learning algorithms, for example unsupervised algorithms like DB Scan to create clusters related to the crashes, identifying all the spots that come under the same

cluster to see if it is possible to identify a recurring pattern in the crashes in that area or if these clusters are based on demographic factors it may be used to know if the people in these clusters have some factor in common that could be related to the occurrence of a crash. The FARS data by itself has a limited scope in providing insights and may have to be analyzed along with other relevant data sources to draw very specific insights and recommendations.

## Cost Analysis

The figures below estimate prices for standard blob storage that does not take into consideration any additional operations required by the company.

 **Storage Accounts**

REGION:  
East US


TYPE:  
Block Blob Storage

PERFORMANCE TIER:  
Standard

STORAGE ACCOUNT TYPE:  
General Purpose V2

REDUNDANCY:  
LRS

ACCESS TIER:  
Hot

 **Storage Accounts**

REGION:  
Central US

TYPE:  
Block Blob Storage

PERFORMANCE TIER:  
Standard

STORAGE ACCOUNT TYPE:  
Blob Storage

REDUNDANCY:  
LRS

ACCESS TIER:  
Hot

### Capacity









#### Billing Option

Save up to 38% on pay-as-you-go prices with 1-year or 3-year Azure Storage Reserved Capacity. [Learn more about Azure Storage Reserved Capacity pricing.](#)

☐ Pay as you go  
☒ 1 year reserved  
☐ 3 year reserved

1 PB Reserved capacity	×	1 Number of units	=	\$15,049.17 Per month
100 TB Reserved capacity	×	1 Number of units	=	\$1,545.00 Per month

The following cost estimates for a single unit of processing capability per hour billed on per-second usage. The cost does not include the pricing for any other Azure resources such as computing instances.

Data Engineering Light		Data Engineering	Data Analytics
<i>Run jobs on Azure Databricks Automated Clusters</i>		<i>Run jobs on Azure Databricks Automated Clusters, with optimized runtime for better performance</i>	<i>Collaborate on projects, notebooks, and experiments with Azure Databricks Interactive Clusters</i>
Standard SKU 	\$0.07 / DBU 	\$0.15 / DBU 	\$0.40 / DBU 
Premium SKU 	\$0.22 / DBU 	\$0.30 / DBU 	\$0.55 / DBU 

## References

---

- Fatal Motor Vehicle Accidents. (2019, September 5). Retrieved from <https://data-usdot.opendata.arcgis.com/datasets/fatal-motor-vehicle-accidents?geometry=-232.151,-4.687,127.849,75.554>
- Facts Statistics: Auto insurance. (2019). Retrieved from <https://www.iii.org/fact-statistic/facts-statistics-auto-insurance>
- Crash Data Location. (2019, May 14). Retrieved from [https://data.iowadot.gov/datasets/cbd84abf01894f4a8404d6990ad2eb2e\\_0?geometry=-110.091,38.962,-43.338,44.686](https://data.iowadot.gov/datasets/cbd84abf01894f4a8404d6990ad2eb2e_0?geometry=-110.091,38.962,-43.338,44.686)
- USAGE-BASED INSURANCE AND TELEMATICS. (2019, May 17). Retrieved from [https://www.naic.org/cipr\\_topics/topic\\_usage\\_based\\_insurance.htm](https://www.naic.org/cipr_topics/topic_usage_based_insurance.htm)

## Appendix

---

**Table1: Description of Important variables in the dataset**

Relevant Variables	Description
FATALS	Number of fatally injured persons in the crash.
PERSONS	Number of person-level forms
COUNTY	Location of the event in regard to the county
CITY	Location of the event in regard to the city
DAY	Day of the month when the crash occurred
MONTH	Month when the crash occurred
YEAR	Year when the crash occurred
DAY_WEEK	Day of the week when the crash occurred
HOURL	Hour when the crash occurred
MINUTE	Minute when the crash occurred
NHS	If the crash occurred on the national highway
DRUNK_DR	Number of drinking drivers involved in the crash
LGT_COND	Type/level of light existed at the time of the crash
WORK_ZONE	Whether the crash happened in a work zone
REL_ROAD	Location of the crash in relation to the road
TYP_INT	Type of intersection
MAN_COLL	Orientation of the vehicle/ manner of collision
HARM_EV	First injury created at the crash site
SP_JUR	If the crash is under the special jurisdiction
MILEPT	Mile point near where the crash happened
State	State name
ST_CASE	Two characters for state code followed by four characters for case number
VE_FORMS	Number of motor vehicles in transport
VEH_NO	Assigned number of motor vehicle

STR_VEH	Number of motor vehicle striking non-motorists
RUR_URB	Land area where the crash occurred
MAKE	Vehicle make
BODY_TYP	Body type
MOD_YEAR	Vehicle model year
TOW_VEH	Vehicle trailing
FIRE_EXP	Fire occurrence
PER_TYP	Person type
INJ_SEV	Injury severity
SEAT_POS	Seating position
AIR_BAG	Airbag deployed
EJECTION	Ejection
EJ_PATH	Ejection path

Sample Dataset 1

OBJECTID	STATE	ST_CASE	VE_TOTAL	VE_FORMS	PVH_INVL	PEDS	PERNOTMVIT	PERMVIT	PERSONS
1	1	10001	1	1	0	0	0	1	1
2	1	10002	1	1	0	0	0	1	1
3	1	10003	3	3	0	0	0	3	3
4	1	10004	1	1	0	0	0	1	1
5	1	10005	1	1	0	0	0	2	2
6	1	10006	2	2	0	0	0	4	4
7	1	10007	2	2	0	0	0	2	2
8	1	10008	1	1	0	0	0	1	1
9	1	10009	1	1	0	0	0	1	1
10	1	10010	1	1	0	1	1	1	1

Sample Dataset 2

▼ OBJECTID	▼ Crash Key	▼ Case Number - Iowa DOT	▼ Law Enforcement Case Number	▼ Date of Crash	▼ Month of Crash	▼ Day of Week
1	20181032538	20181032538	2018003847	2/9/2018, 4:00 PM	February	Saturday
2	20181032539	20181032539	20183894	2/11/2018, 4:00 PM	February	Monday
3	20181032540	20181032540	390218-0398	2/4/2018, 4:00 PM	February	Monday
4	20181032541	20181032541	A18-012	2/10/2018, 4:00 PM	February	Sunday
5	20181032542	20181032542	B18-0518	2/8/2018, 4:00 PM	February	Friday
6	20181029459	20181029459	18-000330	1/23/2018, 4:00 PM	January	Wednesday
7	20181029460	20181029460	18001786	1/25/2018, 4:00 PM	January	Friday
8	20181029461	20181029461	201800303	1/19/2018, 4:00 PM	January	Saturday
9	20181029462	20181029462	20180422	1/25/2018, 4:00 PM	January	Friday
10	20181033815	20181033815	201800261	1/4/2018, 4:00 PM	January	Friday