Data Mining Homework 1

Option2: Data Analytics

Analysis of Accidents of USA dataset.

Dataset : Link
Tablue analysis : Link
Colab : Link

Name: Vaibhavi Hiteshkumar Savani
SJSU ID: 017456972

Professor: Kaikai Liu

Analysed NHTSA traffic dataset to gain insight about the traffic crashes and the causes behind it.

Dataset Link : https://console.cloud.google.com/marketplace/product/nhtsa-data/nhtsa-traffic-fatalities?project=engaged-rope-400320

The goal is to determine which factors have the most significant impact on crashes in the USA. To gain insights, we will analyze the data using the following questions:

1. Which State/City/County have the most accident. List top 5 State and City with highest accident rate.
2. Have accidents increased or decreased over time? Analyzing the top 5 states at highest risk?
3. Analysis for drunk drivers per state. Does some state have more drunk drivers accident then other?
4. When do accidents happen the most? Are they more likely to occur during certain months or on specific days of the week?
5. What kind of atmosphere/weather impacts the most?
6. Does accident happens more in Rural Area or Urban area? Which area has more drunk drivers
7. Does the gender of the driver affect the likelihood of an accident?

Visualizations

Question 1 : Which State/City have the most accident. List top 5 State,City and County with highest accident rate.

We aim to analyze accident data to identify the areas with the highest accident rates. This analysis will enable the department to prioritize these locations, ensuring the availability of adequate emergency treatment resources.

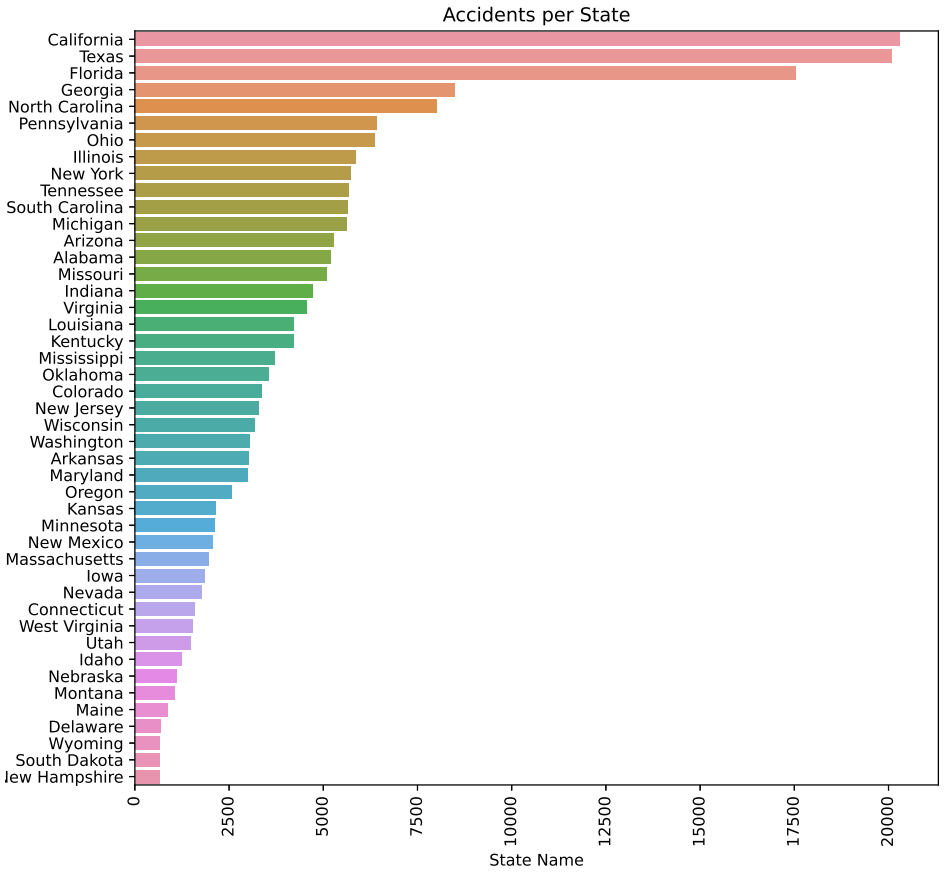


Figure 1: Accidents per State.

Here, we can observe that the majority of accidents have occurred in California, Texas, Florida, Georgia, and North Carolina.

Now Lets see which city and county has highest number of accidents in USA?

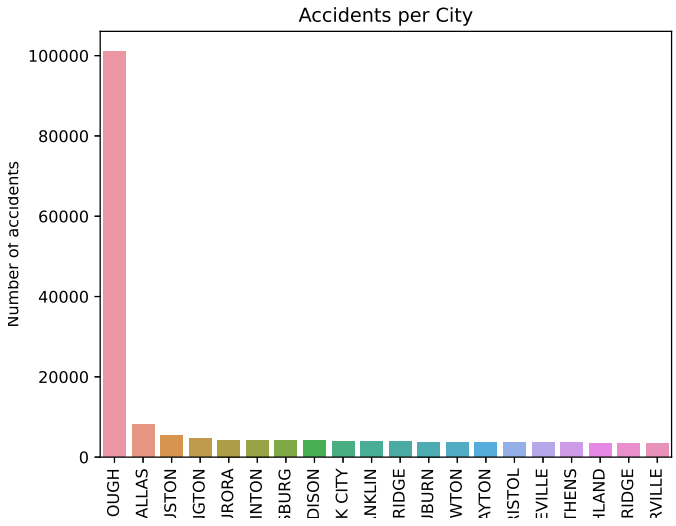


Figure 2: Accidents per city, [?].

As we can see most frequent accidents happened in city Denalo Borough, Dallas, Houston, Arlington and Aurora

Question 2 : Have accidents increased or decreased over time?

Lets analyse if the accident trends have increased of decreased over the time.

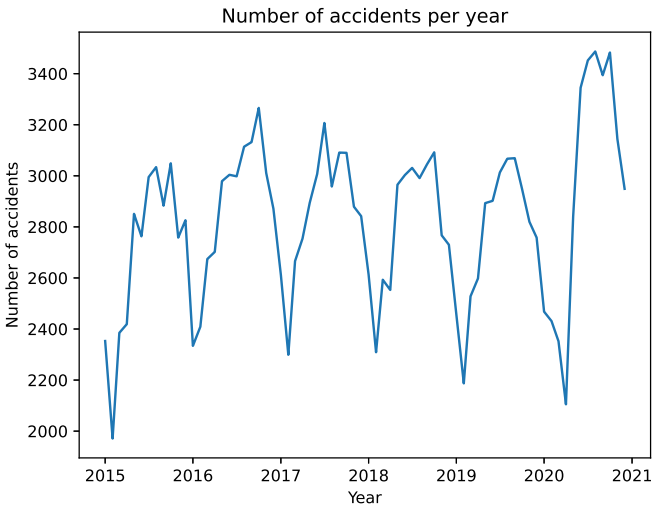


Figure 3: Accidents trends.

We observe a decline in the number of accidents over time, except for the year 2021, which experienced the highest accident rate. It is imperative for the government to conduct an analysis to ascertain whether any policy changes implemented before 2021 have contributed to this upward trend.

Question 3 : Analysis for drunk drivers per state. Does some state have more drunk drivers accident then other?

Accidents that includes drunk drivers vs accidents that doesn't include drunk drivers in every states.

Here are some interesting observations: In some states, even though the total number of accidents is lower compared to other states, the number of accidents involving drunk drivers seems to be higher.

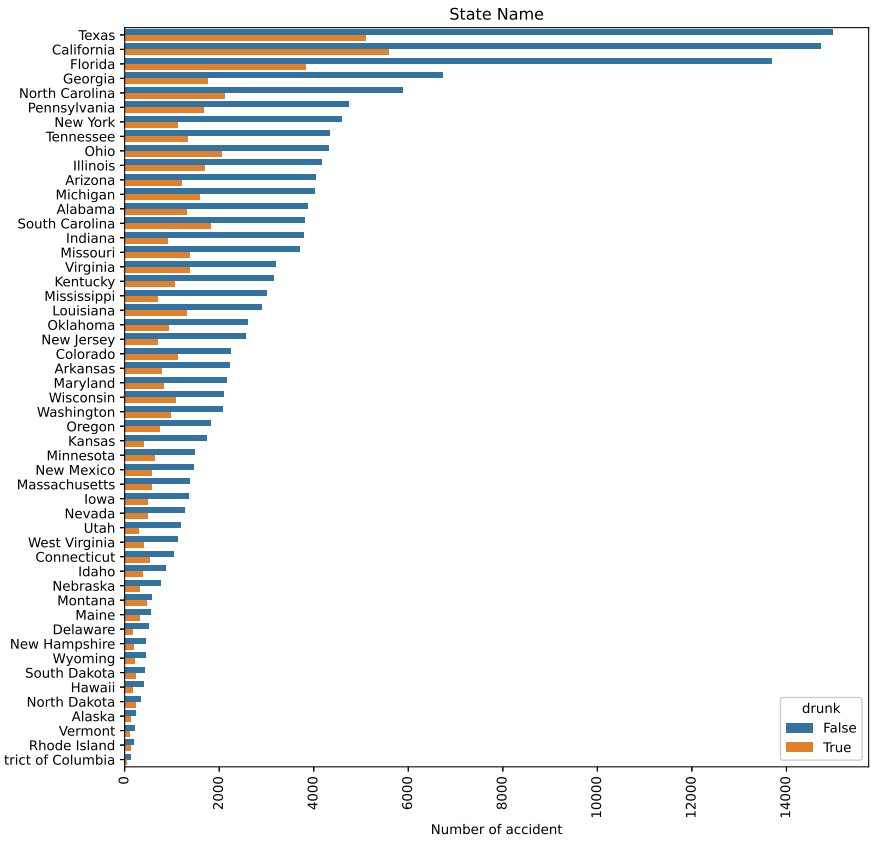


Figure 4: Drunk vs Non drunk drivers per state.

For example, the total number of accidents in Mississippi (3,700) is higher than in Washington (3,059), but the number of accidents involving drunk drivers is higher in Washington (986) than in Mississippi (695)

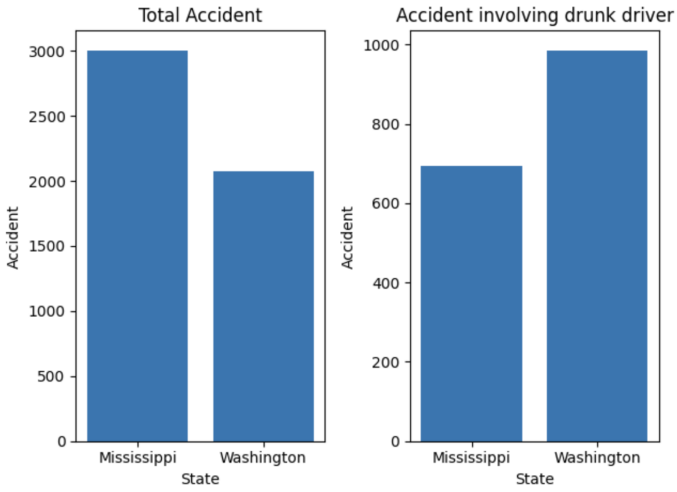
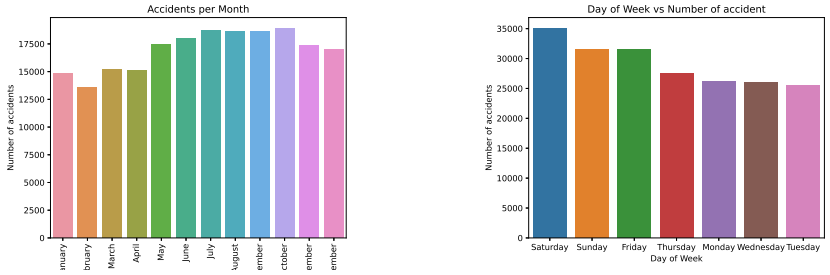


Figure 5: Mississippi vs Washington Drunk driver analysis.

Quetion 4:When do accidents happen the most? Do they tend to occur more frequently during certain months or on specific days of the week?

- As we can observe, period from July to October experiences the highest number of accidents. Summer and early fall often bring outdoor activities and events, which can result in more people traveling to and from these events, leading to higher traffic volumes
- The peak accident hours appear to be between 6 pm and 9 pm coincide with the time when most offices are typically closed and traffic is at



(a) Accidents per Month. (b) Accidents per day.

Figure 6: Overall caption for the figure with two subfigures.

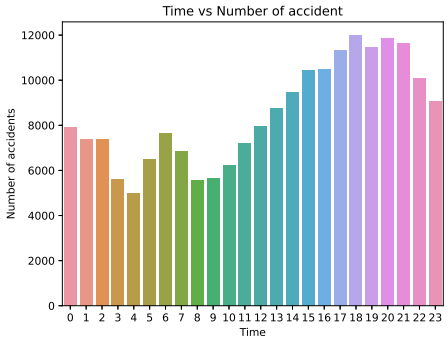


Figure 7: Accidents per hour.

its highest. This correlation suggest that increased traffic during these hours could be the reason for higher number of accidents.

- It appears that the majority of accidents occur on weekends.

Question 5: Does accident happens more in Rural Area or Urban area? Which area has more drunk drivers?

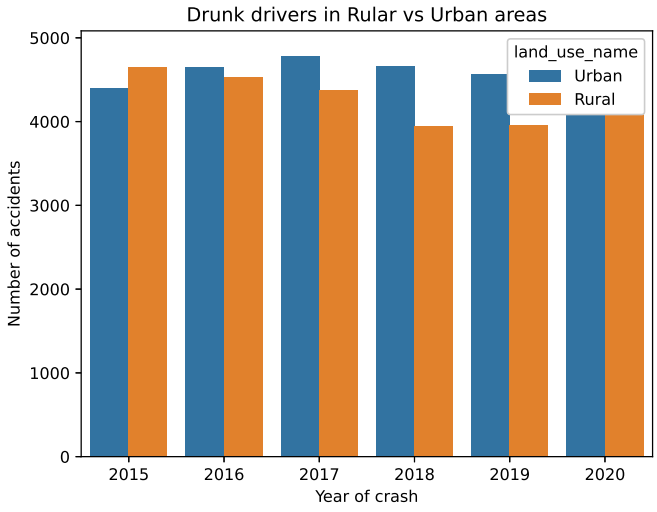


Figure 8: Drunk Driver per land.

Here, we can observe that after 2015, urban areas have more drunk drivers compared to rural areas. Additionally, the chart above shows that the difference between drunk and non-drunk drivers has increased over time.

Question 6: Does the gender of the driver affect the likelihood of an accident?

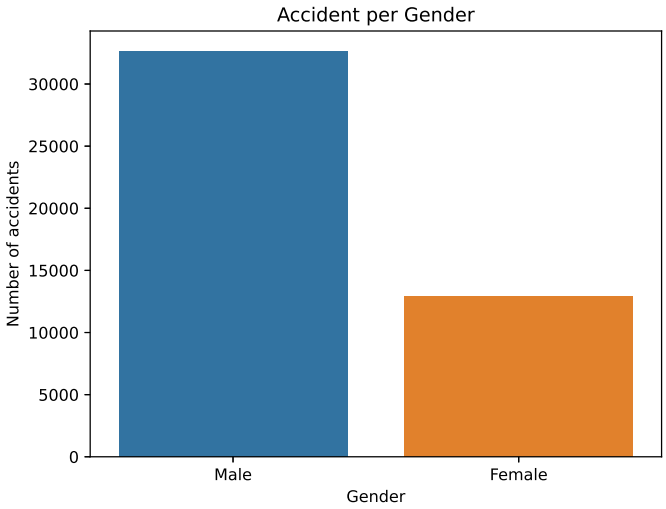


Figure 9: Drunk Driver per gender.

The number of male drivers has more accidents than the number of female drivers. However, we cannot conclude which gender is more accident-prone, as it is possible that the overall number of male drivers is higher than the number of female drivers. Nevertheless, according to the data from 2015 to 2022, the number of accidents involving female drivers is lower than that of male drivers.

Question 8: Utilize Machine Learning to identify the contributing factors that result in more than one fatality.

We have a dataset of accidents that led to fatalities, and we are interested in understanding the factors that contribute to multiple fatalities.

We will use day of week, hour of crash, manner of collision name, light condition name, land use name, state name, atmospheric conditions name columns to train our machine learning algorithm.

Given that we have a very large data set with approximately 90 different features, we will choose the significant features based on the above analysis, which may be contributing factors to the number of fatalities that occurred in the USA.

```
scores = []
cv = ShuffleSplit(n_splits=2, test_size=0.1, random_state=0)
for algo_name, config in algos.items():
    gs =  GridSearchCV(config['model'], config['params'], cv=cv, return_train_score=False)
    gs.fit(X,y)
    scores.append({
        'model': algo_name,
        'best_score': gs.best_score_,
        'best_params': gs.best_params_
    })

return pd.DataFrame(scores,columns=['model','best_score','best_params'])
```

Figure 10: Find best algorithm using GridSearchCV.

Then we will identify which machine learning algorithm and what parameters works best for our dataset using GridSearchCV. We will use that algorithm for training and we will identify the important features according to our model.

We have selected the Random Forest Classifier. According to our machine learning model, the most important factors contributing to multiple fatalities are as follows:

- Clear weather: This is likely because during clear weather conditions, there tends to be more traffic compared to when it is raining or foggy.
- Saturdays: As it is the weekend, more people are likely to be out and about, which can increase the risk of accidents.
- Drunk drivers: This makes sense because intoxicated drivers can exhibit poor driving behaviors, leading to dangerous accidents and multiple fatalities.
- Front to front collisions: According to our model, Front to front collisions often result in multiple fatalities.
- Sundays: Again, because it's the weekend, there may be increased risk.

Our model also considers cloudy weather and the month of July as additional factors.

	Feature	Importance
125	atmospheric_conditions_1_name_Clear	0.023804
18	day_of_week_6	0.021642
0	drunk_driver	0.021621
45	manner_of_collision_name_Front-to-Front	0.019309
19	day_of_week_7	0.018799
10	month_of_crash_10	0.018711
14	day_of_week_2	0.018664
13	day_of_week_1	0.018452
126	atmospheric_conditions_1_name_Cloudy	0.018127
7	month_of_crash_7	0.018086

Figure 11: Top 10 feature contributing to multiple fatalities according to our ML model

Union tables from accident 2015 to accident 2021

As NHTSA dataset has 6 different table for data of 2015 to 2021, One table was created by merging these 6 tables.

Save Dataframe to the bigQuery and use them in tablue for visualisation

Dataset uploaded to the bigQuery from colab.

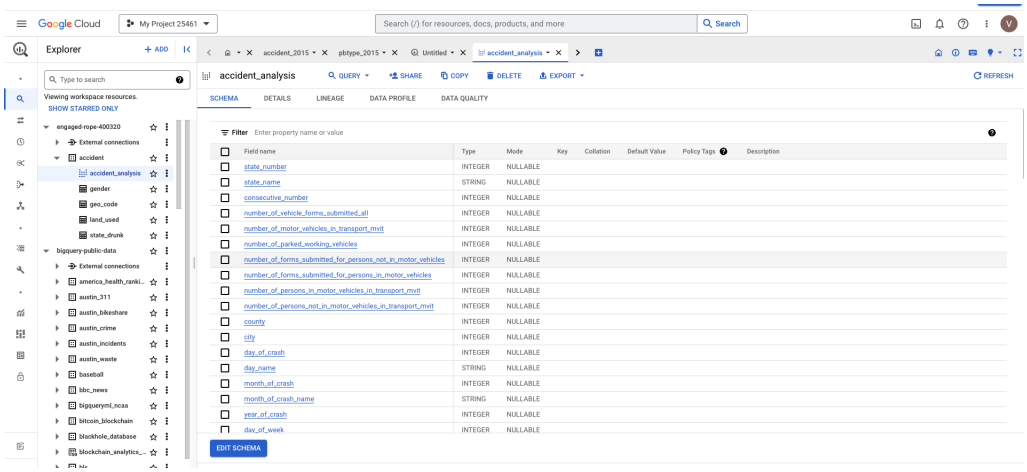


Figure 12: Drunk Driver per land.

```
[ ] ########## write the DataFrame to a BigQuery table
    table_name = 'land_used'
    pandas_gbq.to_gbq(land_use, f'{project_id}.{dataset_id}.{table_name}', project_id=project_id, if_exists='replace')

    100%|██████████| 1/1 [00:00<00:00, 1642.89it/s]
```

Figure 13: Save DF to bigQuery.

Question 7: Does the gender of the driver affect the likelihood of an accident?

Use the saved DF to plot a visualisation in tablue

Tablue Dashboard embedded in web and hosted on github server

Link: https://vaibhavisavani1910.github.io/

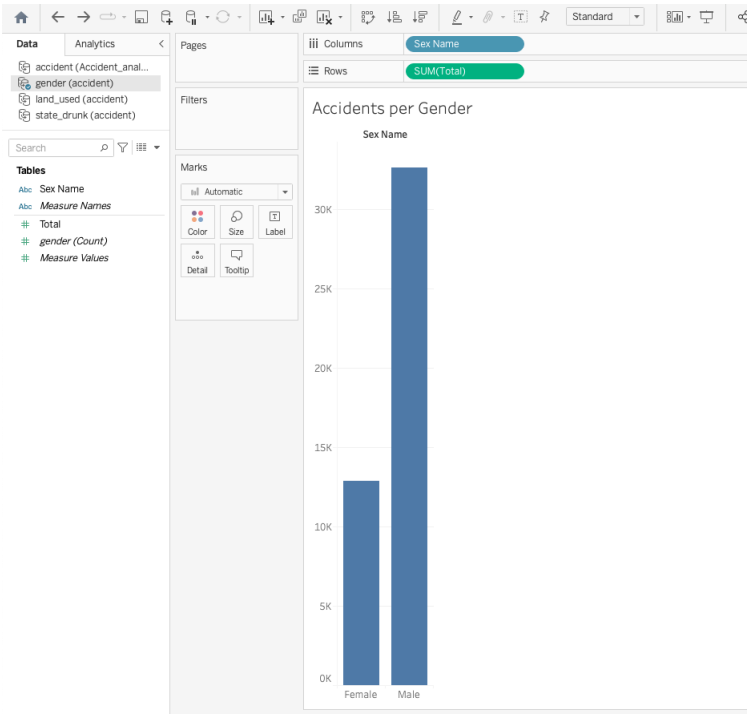


Figure 14: Visualisation via saved DF.

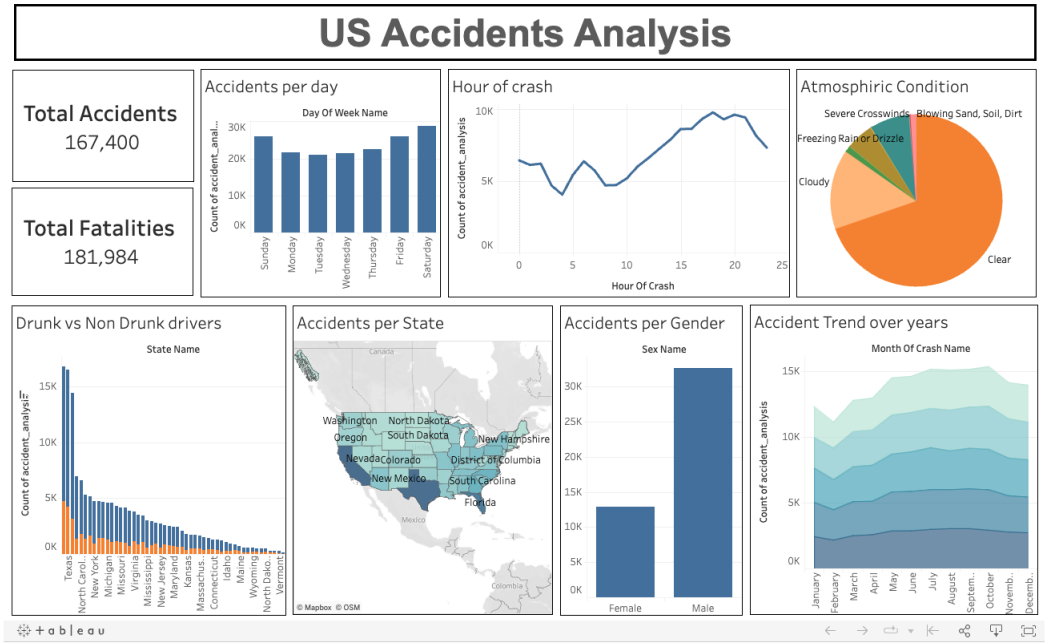


Figure 15: Tablue Dashboard