# Motor Vehicle Collision - Analysis

DASC 5300 - Foundations of Computing

**Team 40**

Abhijit Challapalli - 1002059486

Samarth Mahendra - 1001974557

**Overall Status:**

Initially we were having a discussion whether to use Pandas or Python as both of us were new to Pandas. We decided to use Pandas as it is a library specifically for data manipulation and analysis. After spending a week of learning about Pandas and understanding the problem statement we started on the project. Then we worked on the code independently to bring out all the different ideas we could think of. For preprocessing and cleaning the data we understood and made use of the helper slides which were provided. The approach for the analysis of three queries were described in depth in the below sections of the document.

**Code Developed:**
https://colab.research.google.com/drive/1PZu3i9IYqpDar6veygCkDf-ioasm9GV_?userstoinvite=enaredla02%40gmail.com&actionButton=1#scrollTo=0uBFMek4jAiT

**File description:**

We extracted the data from **Motor_Vehicle_Collisions_-_Vehicles.csv** and later we filtered the data for the required years(2019,2020,2021) for analysis and created another csv file named **Analysis_file.csv.**

**Division of labor:**
As both of us were not too familiar with the pandas library we spent the initial week learning the concepts of pandas. Later, we shared the concepts mutually. Pre-processing and Query 1 is done by Samarth Mahendra & Query 2/Query3/analysis was done by Abhijit Challapalli.

**Time spent:** The total time spent on the project is 2-3 weeks.

**Problems encountered:**

- We were not able to get the count of accidents per year by grouping the CRASH_DATE and VEHICLE_MAKE columns using multi-level indexing. So, instead we created individual DataFrames for VEHICLE_MAKE's and later applied the group by on the CRASH_DATE.

- Found it hard to sort the months in a chronological order for plotting after we applied a group by to the column (CRASH_DATE) as the output of months was not in an order. We used sort_values() function with key attribute to overcome it.

- In the third query we found it hard to clean the vehicle types and replace it with a particular string. We used regular expressions to cleanse the column and populated it with the required string.

- Encountered plotting challenges in the visualization part in the pie-chart/Line Graph. Post referring to google/youtube we got familiar with the approach and presented it.

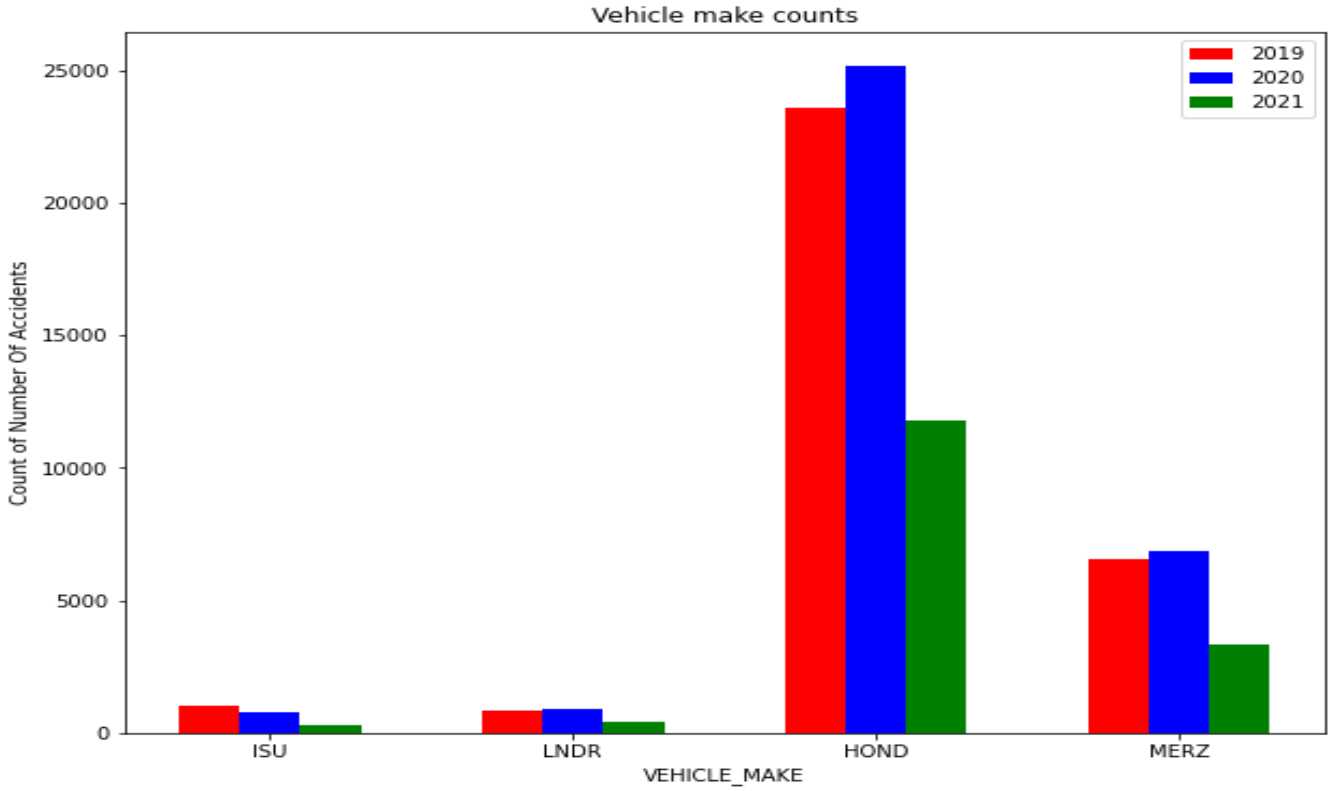
## Pre- processing:

1. Read the Motor_Vehicle_Collisions_-_Vehicles.csv file for analysis into the DataFrame **data.**
2. Converted the Crash_date column to datetime and extracted the years(2019,2020,2021) accordingly using the query method and pulled the data to Df **my_data** and into **Analysis_file.csv** file.
3. Checked for null values with respective to column vehicle_make and dropped the respective rows.
4. Extracted/Cleansed the vehicle make names using clean_names function(regular expressions inside it).
5. Took a sample extract (based on our birth date) from the dataset and analyzed the queries.

## Approach for Query 1:

After extraction of data for required years the steps we followed are as follows

1. Dropped the null values from the VEHICLE_MAKE column.
2. Cleansed the Vehicle names using clean_names function(regular expressions inside it).
3. Created individual DataFrames for each of the Vehicle makes respectively that are used for analysis as given in the parameter file.
4. Converted the CRASH_DATE to datetime.
5. Updated the crash_date with year , applied groupby and count on it.



## Analysis On Query 1:

Car maker ISUZU had exited the US Market in 2009 thereby causing a huge drop in the number of vehicles sold so there is a drop in the number of accidents it is involved in. But Honda has been a top player in the US by selling close to 3000K vehicles ie. only 2%(60509) of the total vehicles were part of an accident. Analyzing this, ISUZU had a lower count of accidents in comparison with Honda.

References:


1. https://www.autonews.com/article/20180130/CCHISTORY/180139984/isuzu-decides-to-exit-u-s

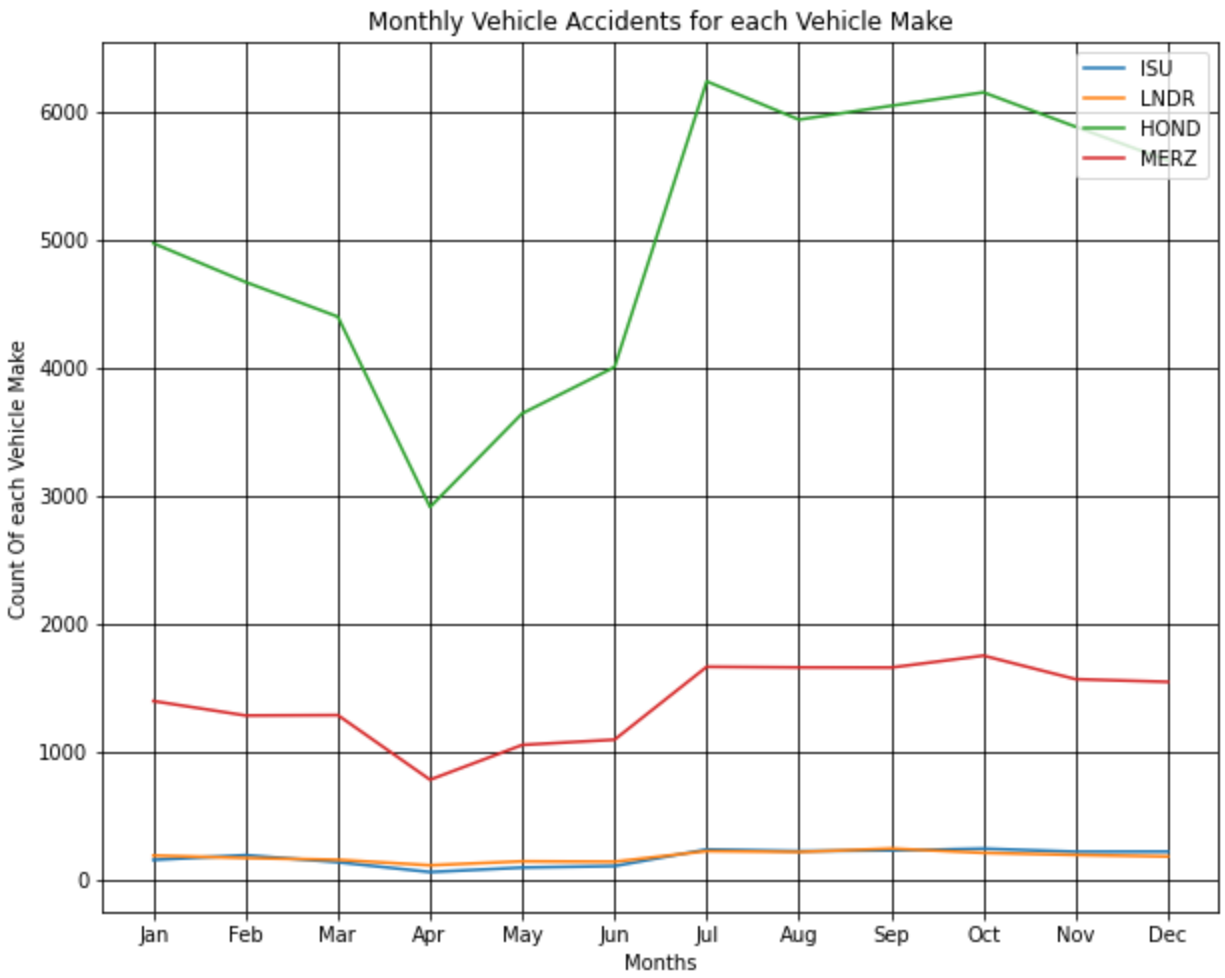
2. https://www.goodcarbadcar.net/honda-us-sales-figures/

3. https://www.hotcars.com/surprising-problems-with-honda-cars/


## Approach for Query 2:

Used the Analysis_file.csv file/df1 which is used in query1 as it is the same required here.

1. Created respective DataFrames for vehicle makes that are used for analysis as given in the parameter file.
2. Converted the crash_date column to datetime and extracted month from it.
3. Updated the crash_date column with month, applied groupby and count function on it.
4. Segregated the count of accidents for each month in a chronological order using lambda functions.




## Analysis for Query 2:

More than half of Americans travel over the fourth of July week by road so there is a drastic increase in accidents in the summer season(July). The count of vehicle crashes continues to be in the higher range till December due to the winter season. The probability of car crash in winter months is also greater compared to other months due to snowy pavements and roads.

References:


1. https://ops.fhwa.dot.gov/weather/weather_events/snow_ice.htm

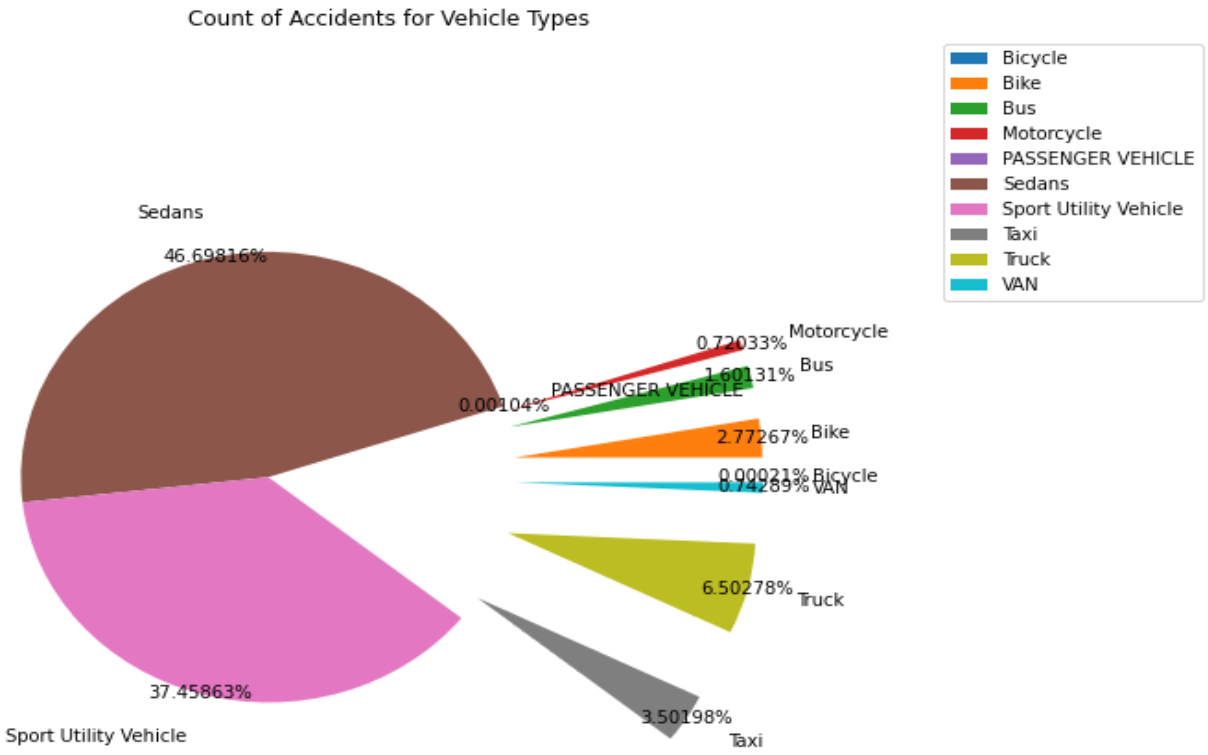
2. https://sobolaw.com/car-accident/winter-driving-accidents-in-new-york/

3. https://www.travelagentcentral.com/running-your-business/stats-nearly-half-americans-to-travel-over-july-4th

## Approach for Query 3:

Read the Analysis_file.csv to df3 for analysis.

1. Dropped the rows which have null values in the column Vehicle_type.
2. Created a function(Cleaning) using regular expressions which is used to clean the name of the Vehicle types.
3. Grouped the column vehicle_type and applied count on the same and pushed the result into the DataFrame vt_count.




## Analysis for Query 3:

Based on the data available Sedans were involved in the highest number of car crashes followed by SUVs and Trucks. Generally SUV's are comparatively safer than Sedans as they are bulkier doing minimum damage to the driver. But in case of rollover collisions sedans can prove safer as they has lower center of gravity. As per the data sedans are more prone to crashes compared to other vehicle types.