**Case Study:**

**Dataset:**

Data Set folder has 6 csv files. Please use the data dictionary (attached in the mail) to understand the dataset and then develop your approach to perform below analytics.

**Analytics**:

• Application should perform below analysis and store the results for each analysis.

• Analytics 1: Find the number of crashes (accidents) in which number of persons killed are male?

• Analysis 2: How many two wheelers are booked for crashes?

• Analysis 3: Which state has highest number of accidents in which females are involved?

• Analysis 4: Which are the Top 5th to 15th VEH\_MAKE\_IDs that contribute to a largest number of injuries including death

• Analysis 5: For all the body styles involved in crashes, mention the top ethnic user group of each unique body style

• Analysis 6: Among the crashed cars, what are the Top 5 Zip Codes with highest number crashes with alcohols as the contributing factor to a crash (Use Driver Zip Code)

• Analysis 7: Count of Distinct Crash IDs where No Damaged Property was observed and Damage Level (VEH\_DMAG\_SCL~) is above 4 and car avails Insurance

• Analysis 8: Determine the Top 5 Vehicle Makes where drivers are charged with speeding related offences, has licensed Drivers, uses top 10 used vehicle colours and has car licensed with the Top 25 states with highest number of offences (to be deduced from the data)

**Expected Output:**

• Develop an application which is modular & follows software engineering best practices (e.g. Classes, docstrings, functions, config driven, command line executable through spark-submit)

• Code should be properly organized in folders as a project.

• Input data sources and output should be config driven

• Code should be strictly developed using Dataframe APIs (Do not use Spark SQL)

• Share the entire project as zip or link to project in Github repo.