Data Engineering Project
Milestone 1
Deadline 22-11-2022

- Load dataset
- Explore the dataset and ask at least 5 questions to give you a better understanding of the data provided to you.
- Visualize the answer to these 5 questions.
- Cleaning the data
- Observe missing data and comment on why you believe it is missing(MCAR,MAR or MNAR)
- Observe duplicate data
- Observe outliers
- After observing outliers, missing data and duplicates, handle any unclean data.
- With every change you are making to the data you need to comment on why you used this technique and how has it affected the data(by both showing the change in the data i.e change in number of rows/columns, change in distribution, etc. and commenting on it).
- Data transformation and feature engineering
- Add a new column named 'Week number' and discretize the data into weeks according to the dates. Tip: Change the datatype of the date feature to datetime type instead of object.
- Encode any categorical feature(s) and comment on why you used this technique and how the data has changed.
- Identify feature(s) which need normalization and show your reasoning. Then choose a technique to normalize the feature(s) and comment on why you chose this technique.
- Add at least two more columns which adds more info to the dataset by evaluating specific feature(s). I.E(Column indicating whether the accident was on a weekend or not).
- For any imputation with arbitrary values or encoding done, you have to store what the value imputed or encoded represents in a new csv file. I.e if you impute a missing value with -1 or 100 you must have a csv file illustrating what -1 and 100 means. Or for instance, if you encode cities with 1,2,3,4,etc what each number represents must be shown in the new csv file.
- Load the new dataset into a csv file.
- **Extremely Important note** - Your code should be as generic as possible and not hard-coded and be able to work with various datasets. Any hard-coded solutions will be severely penalized.
- Bonus: Load the dataset as a parquet file instead of a csv file(Parquet file is a compressed file format).