NYC MOTOR VEHICLE COLLISION ANALYTICS AND PREDICTIVE MODELING

Databricks notebooks

1. Initial data cleaning and EDA- https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/2223093725023385/3402554983214854/563989775462406/latest.html

2. Data Wrangling- https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/2223093725023385/1978338885242709/563989775462406/latest.html

3. data-cleaning-for-spatial-clustering - https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/2202773089231587/3270097802115930/6397218320977099/latest.html

4. features and target variable for predictive modeling - https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/2202773089231587/288185156849501/6397218320977099/latest.html

5. train-test-split-clustering-join-data - https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/2202773089231587/3538449047751016/6397218320977099/latest.html

6. Model Training - https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/2223093725023385/3086432350858478/563989775462406/latest.html

NOTE*

In the stages of preprocessing data, we had to displace our data from one user to the other, without getting rid of all the preprocessing as the joining and discretization were expensive operations. So, we utilized df.display() to get all the data and download it, transferred to another user and uploaded to their DBFS to continue with the subsequent steps in the preprocessing. Thus, at many points of our code, there have been calls to datasets other than the ones mentioned in our presentation.