ProjectProposal

July 26, 2023

1 IST 652 PROJECT PROPOSAL

The final project for IST652 involves locating an open data set or a group of data sets of interest, formulating an inquiry or set of inquiries that could be addressed with the data,processing the data set(s) in a Jupyter Notebook environment using Python, and conducting some analyses on the data to illuminate the inquiry. The project focuses on open data in order to ensure that your chain of transformations and analysis is reproducible.

This is the FIRST DELIVERABLE

1.1 Project Objective

Primary objectives for the project are ..

- Demonstrate your ability to write Python scripts to access and process data.
- Describe steps taken to prepare the data for analysis. For example how did you access and ingest the data, data wrangling, formatting, feature engineering and other steps.
- Develop a research questions you are hoping to answer from the data collected.
- Clearly articulate findings from analysis and summarizes impactful findings.
- Collaborate as a team.

1.2 Analysis Team

List team members below and their roles (note roles may be modified in the second deliverable)

1. Maruthamuthu K Sowmeya
2. Rujuta Prakash Lanke
3. Aditya Kulkarni

1.3 Phase 1: Ideation

The goal of this phase is to outline the specific goals and objectives of your project; include evidence of its feasibility by including citations of resources you will use to complete the code.

1.3.1 Step 1: Project Summary

Write a brief summary of your project ideas, In 250 - 500 words.

Identifying Factors Influencing Motor Vehicle Collisions in NYC

1. The project aims to analyze vehicle collision in New York City (NYC), identify important factors contributing to accidents. By leveraging data exploration, visualization, and statistical analysis, the study seeks to gain insights into traffic safety and improve decision-making for targeted interventions.
2. The project's findings will contribute to evidence-based decision-making for traffic safety initiatives in NYC. By identifying important factors contributing to collisions, stakeholders and policymakers can develop targeted interventions to reduce accident rates and enhance traffic safety in the city.
3. The objectives would be:
    1. Data Exploration and Preprocessing: Explore the dataset to understand the characteristics of the variables and handle any missing or inconsistent data.
    2. Data Visualization: Visualize the data through plots and graphs to uncover trends, patterns, and potential correlations between variables.
    3. Data Analysis and Hypothesis Testing: Conduct statistical analysis to test hypotheses and identify significant factors that contribute to motor vehicle collisions.
    4. Feature Importance: Utilize machine learning algorithms to determine the importance of each variable in predicting collisions.

1.3.2 Step 2: Datasets Research

Select a dataset or a combination of datasets for your project. Many data sets are available at sites such as the World Bank (http://data.worldbank.org), the U.S. Federal Government (http://www.data.gov), - other potential sites for data sets will be provided by the instructor but it is recommended that you search for open data sets too on your own. However, do not use datasets from Kaggle.com.

Note: The number of records (rows) present in your dataset (or total combination of datasets) must exceed 4,000 with at least 8 different categories (columns) of data.

Clearly describe from where your data was located.Why is this resource an authority. Provide a shortlist of datasets your team is considering for your final project. Provide references to the dataset as applicable. Include any other components necessary.

Data Location

1. The data for the project on "Identifying Factors Influencing Motor Vehicle Collisions in NYC" was located on the data.gov website, specifically from the following dataset: "Motor Vehicle Collisions - Vehicles. which consists of 25 vraiables and millions of rows." The dataset is publicly available and provided by the New York City Police Department (NYPD). It contains information about each vehicle involved in motor vehicle collisions reported by the NYPD since April 2016. The data is part of the larger "Motor Vehicle Collisions" database, which consists of comprehensive reports on all motor-vehicle collisions in NYC provided by police. https://catalog.data.gov/dataset/motor-vehicle-collisions-vehicles

Why is this resource an authority

1. The dataset "Motor Vehicle Collisions - Vehicles" is an official record from the NYPD, a highly respected law enforcement agency responsible for reporting and managing motor vehicle collisions in New York City. Being sourced directly from the NYPD, the dataset holds significant authority and reliability for conducting in-depth analyses related to traffic safety.

Provide a shortlist of datasets your team is considering for your final project.

1. Local Area Unemployment Statistics (LAUS) of California State

URL: https://catalog.data.gov/dataset/local-area-unemployment-statistics-laus

2. Crime Data

URL: https://catalog.data.gov/dataset/crime-data-from-2020-to-present

3. 2. U.S. State and Territorial Orders Closing and Reopening Restaurants

URL: (https://data.cdc.gov/api/views/azmd-939x/rows.csv?accessType=DOWNLOAD

Provide references to the dataset as applicable.

1. We will reference official documentation and metadata provided by data.gov to ensure data quality, understand data field definitions, and validate the reliability of the datasets. Additionally, relevant academic papers, government reports, and traffic safety initiatives like Vision Zero will be cited to support the analysis and interpretation of the findings. Proper citation and acknowledgment of the data sources will be maintained throughout the project to adhere to ethical practices and give due credit to the authoritative resources.

1.3.3 Step 2a: Objectives

What have you learned about your dataset(s) so far, and what are the questions you plan to answer with the data (a minimum of 5 questions is a good start).

1. Which vehicle types are involved in the highest number of collisions, and do certain types have a higher likelihood of being in severe accidents compared to others?
2. What are the most common contributing factors to collisions, and do these factors vary based on vehicle characteristics or driver demographics?
3. Is there any correlation between the age of vehicles and the severity of collisions they are involved in?
4. Are certain neighborhoods or intersections experiencing a higher concentration of collisions, and can these locations be identified as high-risk areas?
5. How does driver gender impact collision rates and severity? Are there notable differences between male and female drivers in terms of the types of collisions they are involved in?

1.3.4 References

https://catalog.data.gov/dataset/motor-vehicle-collisions-vehicles

[ ]: