ProjectProposal

November 3, 2024

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)

Berber Bakermans: Data exploration. Marjan Abedini: Data exploration.

2 Phase 1: Ideation

2.1 Primary Goal

We want to take a closer look at the Citibike usage data for July 2024. Our goal is to better understand riding behavior, peak usage times, popular routes, and any differences between member and casual riders.

2.2 Objectives

- Analyze trends in ride duration and how often rides happen.
- Identify the most frequently used stations.
- Look at the distribution of bike types (e.g., electric vs. standard bikes).
- Compare the behavior of members and casual riders.

- Create visualizations to give useful insights for better transportation planning and resource management.

2.3 Feasibility and Supporting Resources

Below are some tools and resources that will help us with the analysis:

- **Pandas Library for Data Manipulation**: Pandas will be used to load, clean, and organize the data, making it easier to work with.
- **Matplotlib and Seaborn for Visualization**: These libraries will help create visuals like line charts for trends, bar graphs for popular stations, and heatmaps to show usage at different times of the day.
- **Citibike NYC Data**:
  - *Source*: NYC Department of Transportation. (2024). *Citibike Trip Data - July 2024.*
  - *Retrieved from*: NYC Open Data
  - *Details*: This dataset includes trip details such as timestamps, station information, and user types. It will help us understand bike-sharing patterns in New York City.

2.3.1 Step 1: Project Summary

3 Project Overview: Analyzing Citibike Usage Data for July 2024

3.1 Primary Objective

The main goal of this project is to analyze Citibike usage data for July 2024. We aim to discover trends, understand user behavior, and provide data-driven insights that can inform management strategies for transportation in New York City. Our focus will be on understanding key aspects of user patterns such as ride frequency, trip durations, popular routes, and the differences between members and casual users.

3.2 Project Plan

3.2.1 Data Preparation

We will start by loading and preprocessing the Citibike data. This includes cleaning and formatting the data to ensure it is ready for analysis. We will make decisions that best fit the dataset to maximize its usefulness.

3.2.2 Exploratory Data Analysis (EDA)

We will conduct an exploratory data analysis to dive deeper into the dataset and uncover actionable insights. This phase will also help us identify any new questions that may need further investigation. Our specific areas of focus will include:

- **Identifying Peak Usage Times**: This will help with resource planning and ensure Citibike stations are properly equipped during busy times.
- **Mapping Popular Stations and Routes**: This can highlight which areas need more attention or infrastructure improvements.

- **Comparing Member and Casual User Behavior**: Understanding differences between these groups can support targeted marketing efforts and service improvements. For example, we want to find out if casual riders prefer certain routes or times compared to members.
- **Electric vs. Standard Bike Usage**: Analyzing which type of bike is preferred will help transportation planners make more informed decisions about future investments.

3.3 Tools and Resources

For analysis and visualization, we will use: - **Pandas**: To manipulate and clean the data. - **Matplotlib and Seaborn**: To create charts and graphs that visualize our findings.

3.4 Expected Outcomes

By completing this project, stakeholders such as city planners, transportation authorities, and Citibike itself will gain a better understanding of user needs. They will have clearer insights into peak usage times, high-traffic routes, and user demographics. This can drive strategic decisions that improve user satisfaction and operational efficiency.

4 Dataset Research

4.1 Primary Dataset: Citibike Trip Data (July 2024)

- **Description**: This dataset contains records of Citibike trips in New York City. It has 3,217,063 rows and 13 columns for the month of July 2024. The dataset includes attributes like ride ID, bike type, start and end times, station names and IDs, geographic coordinates (latitude and longitude), and user type (member or casual).
- **Authority**: NYC Open Data is a reliable source managed by the city government. It provides access to accurate and up-to-date data, ensuring the credibility of the information for our analysis.

4.2 Supplementary Datasets (if necessary):

4.2.1 Weather Data (July 2024)

- **Source**: NOAA National Centers for Environmental Information (https://www.ncdc.noaa.gov)
- **Description**: Daily or hourly weather conditions, including temperature, precipitation, and humidity. This data can help us see how weather affects rider behavior.
- **Authority Justification**: The NOAA is a government organization that provides climate and weather data for research and public use.

4.2.2 Demographic Data by NYC Boroughs

- **Source**: U.S. Census Bureau (https://www.census.gov/data.html)
- **Description**: Demographic statistics on population density, age distribution, and income levels for different areas in NYC. This information can help us analyze Citi Bike usage in relation to population demographics.
- **Authority Justification**: The U.S. Census Bureau is the main source of demographic data, ensuring reliable and accurate information.

5 Initial Insights from the Citibike Dataset

- The dataset includes detailed trip information with columns such as ride ID, bike type, start and end timestamps, station names and IDs, latitude and longitude for both start and end points, and user type (member or casual).
- It contains **3,217,063 rows** and **13 columns**.
- There are **6 columns with multiple NA values**.
- There are **9 columns with data type 'object'** and **4 columns with data type 'float64'**.

5.1 Questions We Plan to Answer

1. **Which stations are the most popular for starting and ending rides?**
   - Determine the busiest stations.
2. **What is the average duration of rides, and how does it vary between different bike types (e.g., electric vs. standard bikes)?**
   - Analyze whether certain bike types are preferred for longer or shorter rides.
3. **How does user behavior differ between members and casual users in terms of trip duration, start times, and station preferences?**
   - Provide insights into membership trends.
4. **What are the peak usage hours and days for Citibike in July 2024?**
   - Identify the most common times for bike usage to understand demand patterns.
5. **How do weather conditions (e.g., temperature, precipitation) affect Citibike usage?**
   - Use the supplementary weather data to correlate environmental factors with ride patterns and predict demand under different weather scenarios.
6. **What geographic trends are observable in ride routes and bike type distribution across different NYC areas?**
   - Map out usage patterns by area to identify regions with higher electric bike demand or unique route characteristics.

With these questions, we aim to achieve actionable insights that can help the rideshare in New York City. These questions are a starting point, and we are open to switching or adding more questions as we discover new insights.

6 References

1. NYC Department of Transportation. (2024). *Citibike Trip Data - July 2024.* pd.read_csv('202404-citibike-tripdata.csv')
2. NOAA National Centers for Environmental Information. (n.d.). *NOAA Climate and Weather Data.*https://www.ncdc.noaa.gov
3. U.S. Census Bureau. (n.d.). *Demographic Data.* https://www.census.gov/data.html
4. Pandas (import pandas as pd)
5. Matplotlib Documentation (import matplotlib as mpl)
6. Seaborn Documentation (import seaborn as sb)

[ ]: