DS 3000 Group 34 Phase 2: Data Acquisition and Preparation

Teammates:

Isabella Hernandez, Han-Mac Kim, Julia Ouritskaya, Henry Renninger

Data Source Information:

1. Data Source Name: Motor Vehicle Collisions – Crashes
    - Hyperlink to Raw Data: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95
2. Data Source Name: Motor Vehicle Collisions – Person
    - Hyperlink to Raw Data: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Person/f55k-p6yu

Related Work:

Paper 1: An analysis of the New York City traffic volume, vehicle collisions, and safety under COVID-19

Hyperlink: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9364745/

- Examines traffic volume, vehicle collisions, and safety during and after the COVID-19 pandemic in New York City.
- Identifies a positive correlation between increased traffic volume and higher collision rates.
- Highlights the complexities of the relationship between traffic volume and safety, emphasizing the need for a nuanced approach.
- Calculates the social cost of collision during the pandemic and discusses the implications for policy making.

Paper 2: Mapping Motor Vehicle Collisions in New York City

Hyperlink: https://toddwschneider.com/posts/nyc-motor-vehicle-collisions-map/

- Utilizes New York Police Department data to analyze motor vehicle collisions in NYC since July 2012.
- Provides an interactive heatmap showcasing the number of collisions in each area, and can be customized to reflect injuries and fatalities.
- Addresses challenges faced by cyclists, including vehicles parked in bike lanes.
- Explores injury trends, road congestion, vehicle types, and contributing factors in traffic collisions, offering insights into traffic safety in the city.

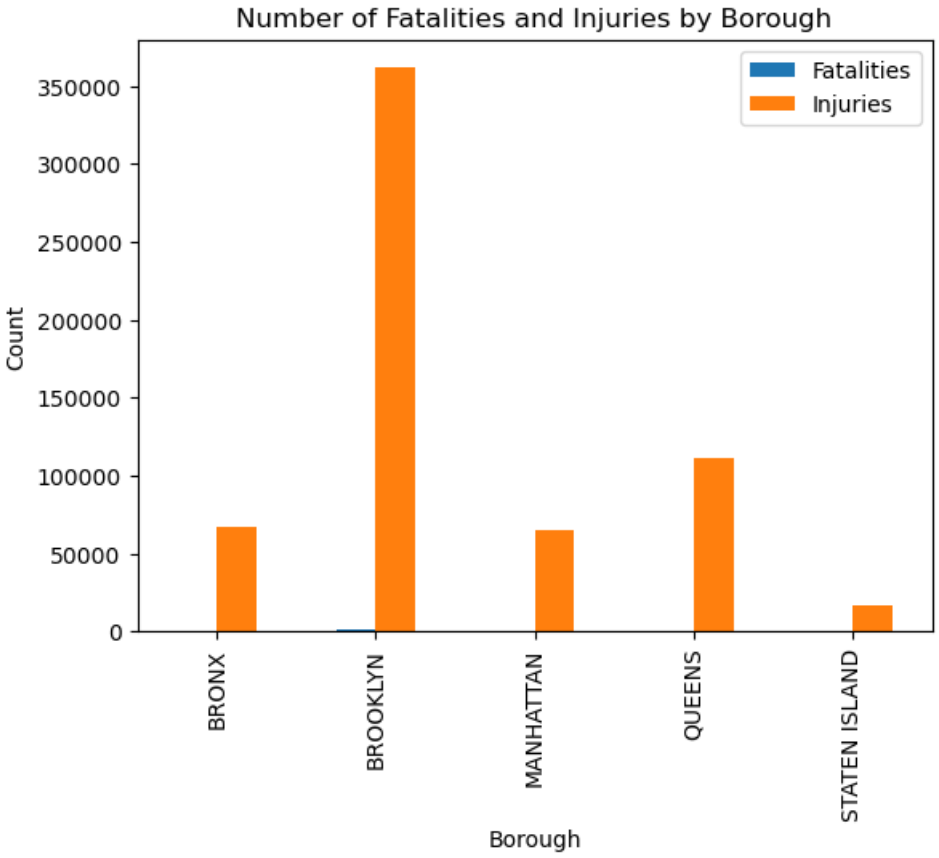
Additional Resources and Libraries:

- **Pandas:** For data manipulation and analysis.
- **NumPy:** For numerical operations and computations.
- **Matplotlib** and **Seaborn:** For data visualization and creating plots.
- **Scipy:** For scientific and statistical computing.
- **Statsmodels:** For statistical modeling and hypothesis testing.
- **Sklearn:** For statistical analysis.

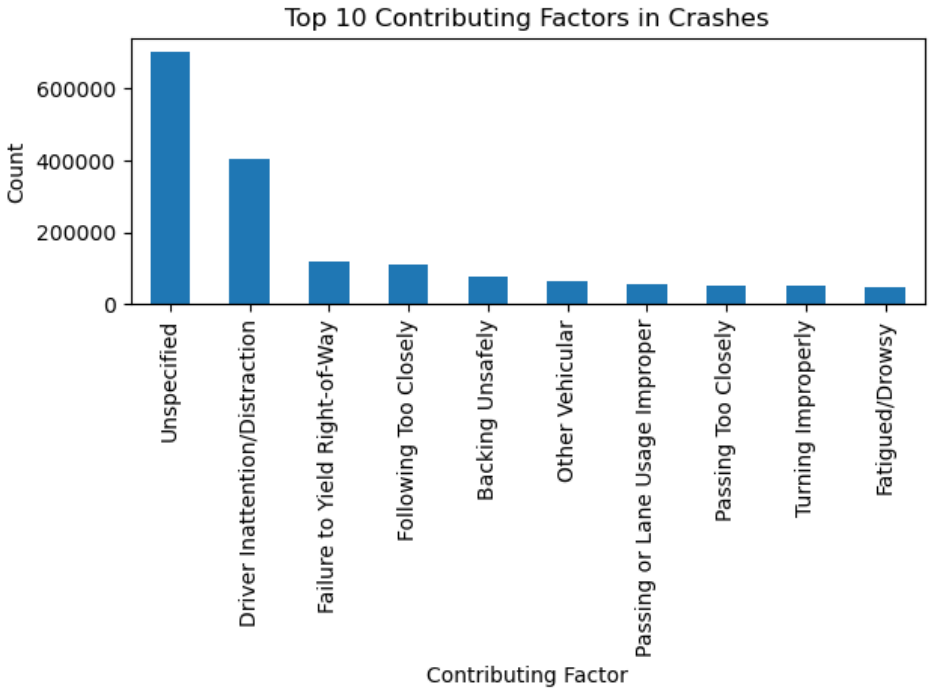
Preliminary Exploratory Data Analysis (EDA):

1. **Column Selection:**
    - For the Crashes Dataset, selected relevant columns including:
        - COLLISION_ID, CRASH DATE, CRASH TIME, BOROUGH, LATITUDE, LONGITUDE, CONTRIBUTING FACTOR VEHICLE 1, NUMBER OF PERSONS INJURED, NUMBER OF PERSONS KILLED, NUMBER OF CYCLIST INJURED, NUMBER OF CYCLIST KILLED, NUMBER OF PEDESTRIANS INJURED, and NUMBER OF PEDESTRIANS KILLED
    - For the Persons Dataset, selected relevant columns including:
        - UNIQUE_ID, COLLISION_ID, CRASH_DATE, CRASH_TIME, PERSON_TYPE, PERSON_AGE, and CONTRIBUTING_FACTOR_1
2. **Data Type Conversion:**
    - Converted 'CRASH DATE' and 'CRASH TIME' columns in both datasets to datetime data types using pandas.to_datetime().
3. **Handling Missing Values:**
    - Checked for missing values in both datasets.
    - Decided not to drop rows with missing values and instead imputed them as follows:
        - For the Crashes Dataset:
            1. Imputed missing values in 'BOROUGH' with the mode.
            2. Imputed missing values in 'LATITUDE' and 'LONGITUDE' with the median.
            3. Imputed missing values in 'CONTRIBUTING FACTOR VEHICLE 1' with 'Unspecified'.
            4. Imputed missing values in 'NUMBER OF PERSONS INJURED' and 'NUMBER OF PERSONS KILLED' with zeros.
        - For the Persons Dataset:
            1. Imputed missing values in 'PERSON_AGE' with the median.
            2. Imputed missing values in 'CONTRIBUTING_FACTOR_1' with 'Unspecified'.
4. **Removing Duplicates:**
    - Remove duplicated rows in both datasets.
5. **Data Visualization:**
    - **Visualization 1: Number of Fatalities and Injuries by Borough**

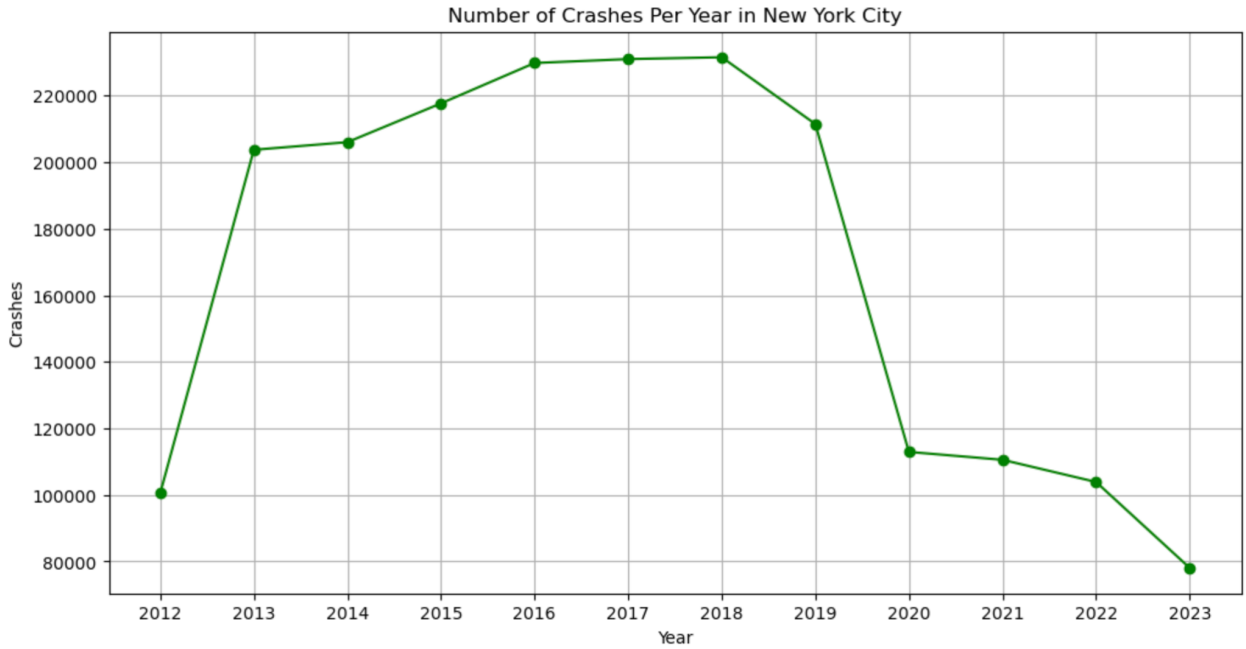
- Created a visualization showing the number of fatalities and injuries by borough. Found that Brooklyn has the highest number of injuries and fatalities.



- **Visualization 2: Most Common Contributing Factor**
  - Calculated and visualized the top 10 contributing factors in crashes. Found that the most common contributing factor is 'Unspecified.'



During the data preparation and visualization process, we observed patterns in the data, such as the most common contributing factors and the distribution of fatalities and injuries by borough. We also noticed a trend when evaluating crashes per year that there was a steep drop during the pandemic, making it notable to look into causations of certain tendencies in the data. To evaluate whether these patterns are statistically significant, we need to perform further statistical tests and analysis. This could include hypothesis testing, regression analysis, or other statistical methods.



This example is a good visual of the influence the pandemic may have had on vehicular collisions. It was also a case where we used data conversion, changing ‘CRASH DATE’ to the datetime type to make it easier to group by year, highlighted below:

```
plt.figure(figsize=(12, 6))
crashes_df['CRASH DATE'] = pd.to_datetime(crashes_df['CRASH DATE'])
crashes_df['Year'] = crashes_df['CRASH DATE'].dt.year
crashes_per_year_df = crashes_df.groupby('Year').size()
crashes_per_year_df.plot(kind='line', marker='o', linestyle='-', color='green')
plt.xticks(range(2012, 2024))
plt.title('Number of Crashes Per Year in New York City')
plt.xlabel('Year')
plt.ylabel('Crashes')
plt.grid(True)
plt.show()
```

Works Cited:

[1] T. Schneider, “Mapping motor vehicle collisions in New York City,” toddwschneider.com, https://toddwschneider.com/posts/nyc-motor-vehicle-collisions-map/ (accessed Oct. 29, 2023).

[2] P. Cappellari and B. S. Weber, “An analysis of the New York City Traffic Volume, vehicle collisions, and safety under covid-19,” Journal of safety research, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9364745/ (accessed Oct. 29, 2023).

[3] Motor vehicle collisions - person | NYC open data, https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Person/f55k-p6yu (accessed Oct. 29, 2023).

[4] Motor vehicle collisions - crashes | NYC open data, https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95 (accessed Oct. 29, 2023).