INFO 3300/5100 - Project 2 Milestone 2

Group Member Information

Name	NetID	Year
Jeana Hoffmann	jmh472	UGrad
Simon Tian	jt886	Grad
Colin Wu	cww72	UGrad

Introduction

In this project, we combined multiple geographic data regarding vehicle collisions occurring in New York City (NYC), sourced from NYC Open Data. We have found three high-quality datasets, including Motor Vehicle Collisions - Crashes dataset focuses on the crash details, the Motor Vehicle Collisions - Vehicles dataset that focuses on the vehicles involved in the crashes, and Motor Vehicle Collisions - Person dataset outlining any injuries or casualties occurred in the crashes. The team was able to join all three datasets totalling 8 million rows into one cohesive superset using **collision_id**. Thereafter, the team created a dashboard with a map of the precincts in NYC, where the user can hover their cursor on each precinct to navigate through the dataset specific to the current selection. In order to visualize and conceptualize the data, in addition to the main heatmap, two secondary visualizations next to the heatmap were implemented. The first one displays the outline of the precinct that the user's cursor is hovering over, and indicates the distribution of the crashes within. The other is a radar chart, which intuitively indicates the top 6 causes to crash in each precinct.

To ensure the dataset is relevant and reliable, the team performed extensive cleaning and preprocessing. For instance, rows with missing or invalid values in critical fields such as latitude, longitude, and crash time were removed. Additionally, the data was filtered to include crashes from the past five years (2019–2024). The team also applied aggregation techniques to simplify this complex dataset, making it easier to interpret and visualize. For example, crash occurrences were grouped by time intervals (hourly or monthly) and locations (latitude and longitude clusters) to provide meaningful insights.

Data Processing and Join

The team started the data cleaning and joining process by mounting a Google Drive folder containing the necessary datasets to a Colab environment. Three separate CSV files are loaded into pandas dataframes:

- Motor_Vehicle_Collisions_Vehicles
- Motor_Vehicle_Collisions_Person
- Motor_Vehicle_Collisions_Crashes

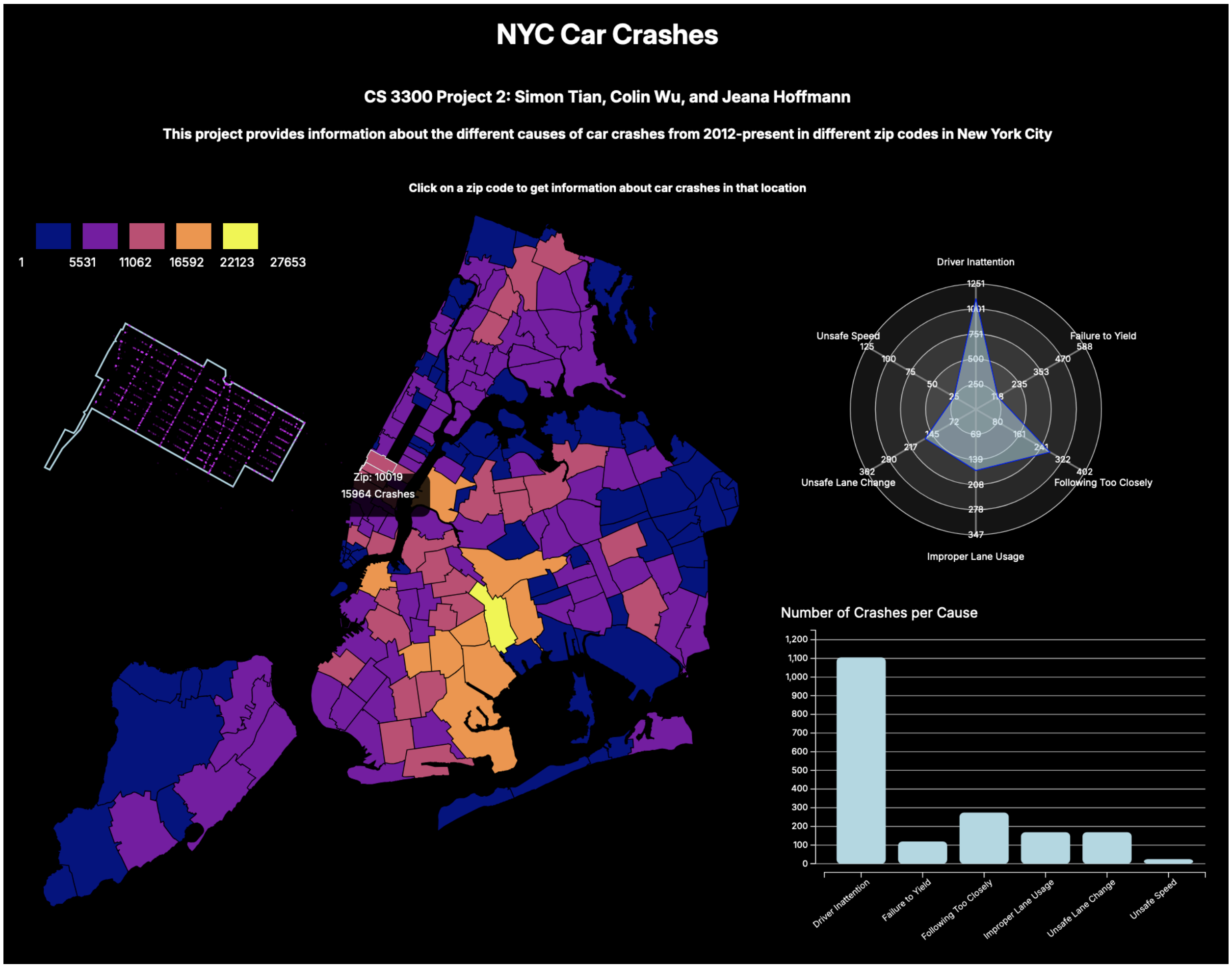
Each dataset contains a variety of columns, which many are not applicable or with limited information to be extracted for the intended visualizations. To reduce the magnitude and streamline the dataset, the team dropped multiple columns from each dataset.

For instance, in the vehicle dataset, fields like **VEHICLE_ID, VEHICLE_MAKE,** and **DRIVER_LICENSE_STATUS** are removed, retaining essential columns such as **COLLISION_ID, STATE_REGISTRATION, VEHICLE_TYPE**, and **DRIVER_SEX**.

Similarly, the "person" dataset is consolidated to focus on columns like **COLLISION_ID, PERSON_TYPE,** and **PERSON_INJURY**, while removing many other attributes such as **POSITION_IN_VEHICLE** and **EJECTION**.

For the crash dataset, fields related to location details (e.g., **ON_STREET_NAME** and **CROSS_STREET_NAME**) and vehicle types are dropped, preserving data such as **ZIP_CODE**, **LATITUDE**, and the number of injuries or fatalities.

Once the team has cleaned and reduced the datasets down to only the essential components, they are merged into a single dataframe using the common key, **COLLISION_ID**. This is an essential element that was previously overlooked, as the team struggled to find good data that incorporates location and coordinates. This step combines the vehicle, person, and crash-level information into a unified dataset. After merging, redundant columns such as multiple copies of **CRASH_DATE** and **CRASH_TIME** are removed to avoid duplication. The merged dataset now contains a consolidated view of the data, including crash details (e.g., location and injuries), vehicle attributes (e.g., type and registration), and human-related factors (e.g., injury severity and sex).



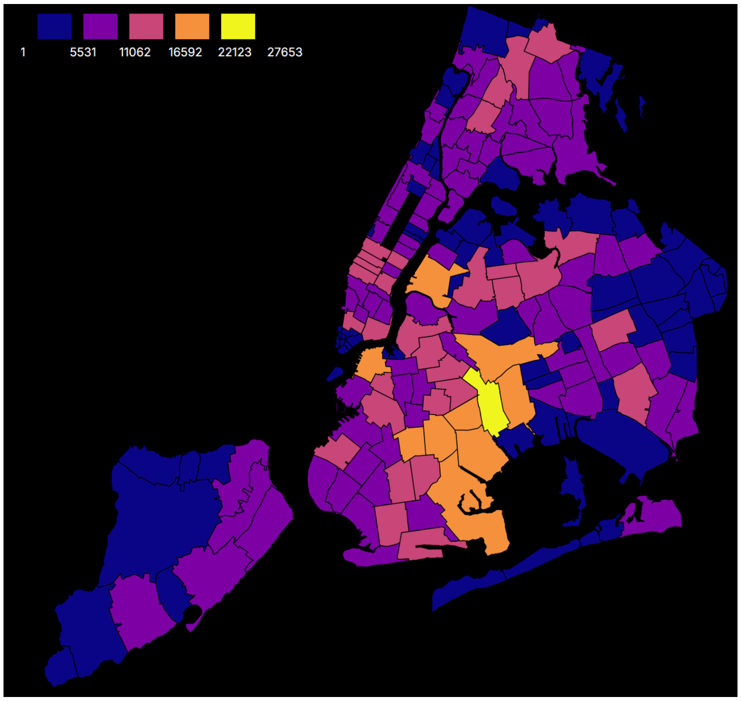
NYC Car Crashes

CS 3300 Project 2: Simon Tian, Colin Wu, and Jeana Hoffmann

This project provides information about the different causes of car crashes from 2012-present in different zip codes in New York City

Click on a zip code to get information about car crashes in that location

Interactive Crash Data Heat Map



Design Rationale

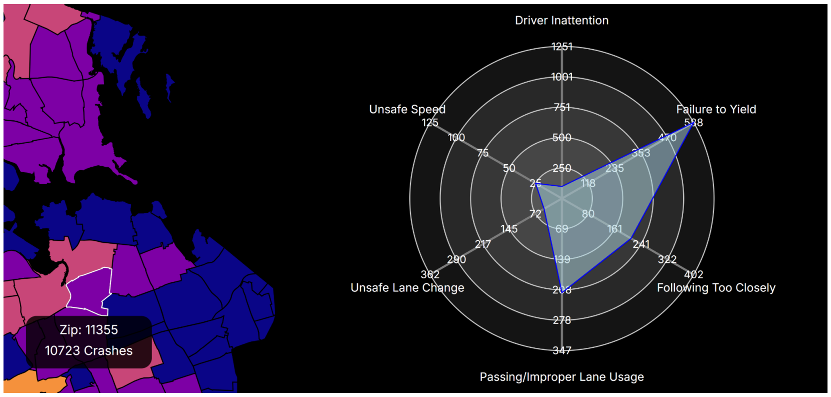
The team decided to improve upon the standard heat map indicating the number of crashes within the past 5 years by adding interactive features to it. The team chose an interactive heat map as this provides both intuitive visualization as well as interactive elements to learn more on a specific area. This also makes it easier for the user to identify patterns and trends. The team used the heatmap to aggregate crash data geographically from NYC Open Data, using color intensity to indicate the frequency of crashes within specific areas. For instance, as shown in the figure above, areas with higher crash rates are shown in brighter colors, while regions with fewer crashes are depicted in darker color schemes.

Additionally, we implemented this heat map for its ability to demonstrate local details and insights. Users can zoom into specific neighborhoods, streets, or intersections to investigate high-crash areas in greater depth. Additionally, clicking any given region reveals exact crash counts and other related details via tooltips, making it easier for users to analyze patterns and trends.

From an implementation standpoint, the interactive heatmap was built using D3.js, complemented by TopoJSON for handling geographic features. The map uses NYC geo-boundary data (GeoJSON source) to define regions and integrates crash data. The team implemented interactivity through D3 event listeners, which respond to user clicks and update tooltips to display reaction to the interaction.

The team also implemented a color scale with d3.scaleQuantile, which can dynamically adjust to map crash frequencies to a visually interpretable range of colors. This approach ensures that both high-crash and low-crash areas are effectively represented in the heatmap. Furthermore, the zooming and panning features, combined with responsive data updates, make the map an accessible and powerful tool for exploring spatial crash trends in NYC.

Interactive Radar Chart



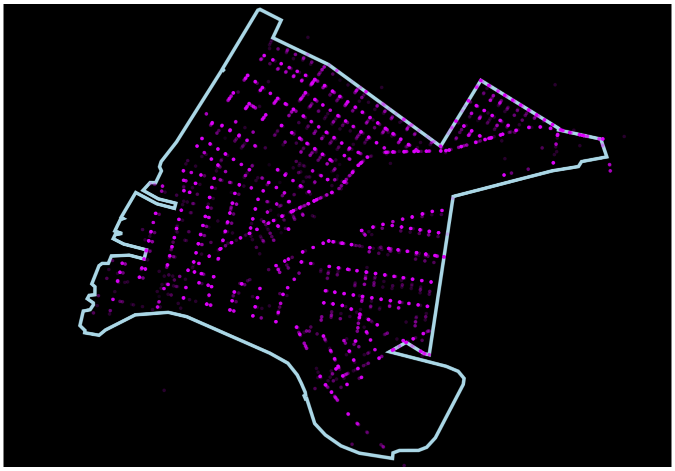
Design Rationale

The team decided to add something unique and effective in this project that is not shown in the class yet. After some research, the team was able to find the boilerplate from this GitHub repository (Link). The team examined the most common factor that contributed to the crashes, and selected 6 of the most common causes, including *"Driver Inattention", "Unsafe Speed", "Unsafe Lane Change", "Passing/Improper lane Usage", "Following Too Closely", and "Failure to Yield*".

The team chose this visualization as this is a useful tool to compare and highlight multiple crash-related factors simultaneously. Radar charts are very efficient in displaying categorical data, an example would be the contributing factors. It would be more challenging to convey the same message using linear or bar charts, as they are limited in dimensions and clarity when trying to display categorical data.

From a technical standpoint, the radar chart implementation in the boilerplate employs D3.js to dynamically generate the chart based on user interactions with the choropleth map. Each zip code selection triggers an update to the radar chart, reflecting the crash factors specific to that area.

Zoomed-In Crash Distribution Chart



Design Rationale

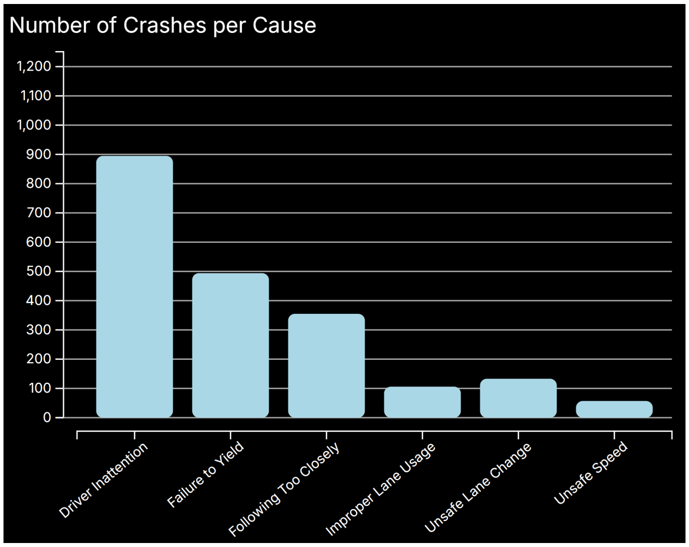
The team decided to add a zoomed-in crash location map in the project to enrich details for the user to learn about the trend of car crashes in NYC. This chart can provide the user with a more detailed understanding of where crashes occur most frequently. When users hover over a specific zip code, this visualization will change to the shape of the area selected, and display the crashes within that area. This is done with small circles for each crash based on their latitude and longitude.

The primary reason for using this zoomed-in map is to provide actionable insights. City planners, traffic engineers, and policymakers can use this detailed spatial visualization to identify high-crash zones, or, hot spots. These insights are useful for implementing projects such as road designs, signage improvement, or enhancing pedestrian safety measures.

The map also helps to uncover patterns that might be masked in city-wide overviews, such as specific areas with a high concentration of pedestrian crashes or zones with recurring accidents involving specific types of vehicles. Additionally, the zoom feature enables the visualization of crash density variations within smaller geographic units, making it easier to visualize regional data for safety improvements.

From a technical perspective, the map uses a layered approach, displaying zip code boundaries, crash density data, and specific crash locations. Color scales are applied to the zip code areas based on crash frequencies, allowing users to quickly identify areas with higher crash rates. Tooltip functionality further enriches the experience by providing detailed information, such as the number of crashes and the zip code, whenever a user hovers over an area. This combination of interactive zooming and detailed data display makes the map a powerful tool for analyzing crash data at a local level.

Bar Chart



Design Rationale

The bar chart displays the same categories of causes for crashing in each zip code, but it does a better job of conveying the number of crashes per category. This graph is a supplement to the radar chart for this purpose, as you can clearly see the differences in *distribution* compared to other zip codes through the radar chart, but then you are able to look up the exact values using the bar chart.

Observations and Conclusion

The team was able to make a few interesting observations and conclusions based on the visualization. For instance, we found that:

Number of Crashes by Zip Code:

Zip codes such as 11207, 10019, and 11234 stand out with the highest crash counts, reaching over 27,000, 15,964, and 17,816 crashes respectively, indicating these areas are major hot spots.

Top Contributing Factors:

"Driver Inattention" consistently emerges as the most significant contributing factor, with values often exceeding 1,200 incidents in certain zip codes.

Other key factors include *"Unsafe Speed," "Failure to Yield,"* and *"Following Too Closely,"* although their values are noticeably lower than those for driver inattention, generally ranging between 350 and 750 incidents.

Factors such as *"Passing/Improper Lane Usage"* and *"Unsafe Lane Change"* are less frequent contributors but still significant.

Areas with High Crash Rates:

Manhattan and Brooklyn have the densest crash locations, as reflected by multiple zip codes in bright orange or yellow.

Major highways such as Major Deegan Expressway (in the Bronx) and other major routes connecting Queens and Manhattan have higher than average crash counts.

Crash Density:

The heatmap shows populous areas such as Manhattan and Brooklyn has more crashes compared to other areas, which makes sense. For example, zip code 11207 (Brooklyn) experiences over 27,000 crashes, which has the highest crash count among all zip codes.

Contributing Factors:

Across all zip codes, *"Driver Inattention"* remains the most common cause compared to others in terms of frequency. This highlights the risk of distracted driving and can provide insights to the law enforcements. Speeding and following too closely are secondary factors to accidents, suggesting areas to improve traffic control and road design.

Geographic Trends:

High-crash zones are in the same areas of dense traffic flow, such as downtown Manhattan (**10019**) and sections of Brooklyn. Major highways and intersections also emerge as dangerous zones, emphasizing the need for targeted infrastructure improvements and safety interventions.

One-Liner Conclusion:

Do not drive in Brooklyn, especially highways. Everyone is distracted.

Team Member Contributions

Name	Contributions
Jeana Hoffmann (jmh472)	Looked into potential datasets, ideas, and attended meetings. Created github repo, and interactive hover map by zip code and colored by scaleQuantile function. Did dataset cleaning and made the lat_lon.json file. Implemented the zoom in function where when you click on the zip code, you see where the crashes occurred.
Simon Tian (jt886)	Researched datasets specifically including the collision ID, which turns out to be the key information of joining multiple large datasets from NYC Open data. Found a boilerplate for the interactive heatmap and radar chart component. Further refined the data cleaning process after Jeana to keep useful information.
Colin Wu (cww72)	Cleaned and extracted data for different causes of crashes in different zip codes, created radar chart and bar chart visualizations. Did styling and layout of the dashboard.

Link to Datasets:

1. Motor Vehicle Collisions - Crashes:
https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data

2. Motor Vehicle Collisions - Vehicles:

https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Vehicles/bm4k-52h4/about_data

3. Motor Vehicle Collisions - Person:

https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Person/f55k-p6yu/about_data