Aansh Jha Homework 5

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os

csv_file = r'C:\Users\aansh\OneDrive\Desktop\Senior Year\STAT 3255\aanshjha-idsf24\data\nycc
feather_file = r'C:\Users\aansh\OneDrive\Desktop\Senior Year\STAT 3255\aanshjha-idsf24\data\

df = pd.read_csv(csv_file)

df.to_feather(feather_file)

csv_size = os.path.getsize(csv_file)
feather_size = os.path.getsize(feather_file)

csv_size_mb = csv_size / (1024 * 1024)
feather_size_mb = feather_size / (1024 * 1024)

print(f"CSV file size: {csv_size_mb} MB")
print(f"Feather file size: {feather_size_mb} MB")

dff = pd.read_feather(feather_file)
print(dff.shape)
print(dff.head())

print(dff.head())
```

```
CSV file size: 0.3399038314819336 MB
Feather file size: 0.1900959014892578 MB
(1875, 29)
```

```
   CRASH DATE CRASH TIME   BOROUGH  ZIP CODE  LATITUDE  LONGITUDE  \
0  06/30/2024      17:30      None       NaN   0.00000    0.00000
1  06/30/2024       0:32      None       NaN       NaN        NaN
2  06/30/2024       7:05  BROOKLYN   11235.0  40.58106  -73.96744
3  06/30/2024      20:47      None       NaN  40.76363  -73.95330
4  06/30/2024      10:14  BROOKLYN   11222.0  40.73046  -73.95149

                LOCATION     ON STREET NAME      CROSS STREET NAME  \
0             (0.0, 0.0)               None                   None
1                   None  BELT PARKWAY RAMP                   None
2  (40.58106, -73.96744)               None                   None
3   (40.76363, -73.9533)          FDR DRIVE                   None
4  (40.73046, -73.95149)  GREENPOINT AVENUE  MC GUINNESS BOULEVARD

          OFF STREET NAME  ...  CONTRIBUTING FACTOR VEHICLE 2  \
0              GOLD STREET  ...                    Unspecified
1                    None  ...                    Unspecified
2  2797      OCEAN PARKWAY  ...                           None
3                    None  ...                           None
4                    None  ...                    Unspecified

   CONTRIBUTING FACTOR VEHICLE 3  CONTRIBUTING FACTOR VEHICLE 4  \
0                           None                           None
1                           None                           None
2                           None                           None
3                           None                           None
4                           None                           None

   CONTRIBUTING FACTOR VEHICLE 5  COLLISION_ID  \
0                           None       4736746
1                           None       4736768
2                           None       4737060
3                           None       4737510
4                           None       4736759

                  VEHICLE TYPE CODE 1                  VEHICLE TYPE CODE 2  \
0                               Sedan                                Sedan
1  Station Wagon/Sport Utility Vehicle  Station Wagon/Sport Utility Vehicle
2  Station Wagon/Sport Utility Vehicle                                 None
3                               Sedan                                 None
4                                 Bus                            Box Truck

   VEHICLE TYPE CODE 3 VEHICLE TYPE CODE 4 VEHICLE TYPE CODE 5
```

```
0                None                None                None
1                None                None                None
2                None                None                None
3                None                None                None
4                None                None                None

[5 rows x 29 columns]
   CRASH DATE CRASH TIME   BOROUGH  ZIP CODE  LATITUDE  LONGITUDE  \
0  06/30/2024      17:30      None       NaN   0.00000    0.00000
1  06/30/2024       0:32      None       NaN       NaN        NaN
2  06/30/2024       7:05  BROOKLYN   11235.0  40.58106  -73.96744
3  06/30/2024      20:47      None       NaN  40.76363  -73.95330
4  06/30/2024      10:14  BROOKLYN   11222.0  40.73046  -73.95149

                LOCATION     ON STREET NAME      CROSS STREET NAME  \
0             (0.0, 0.0)               None                   None
1                   None  BELT PARKWAY RAMP                   None
2  (40.58106, -73.96744)               None                   None
3   (40.76363, -73.9533)          FDR DRIVE                   None
4  (40.73046, -73.95149)  GREENPOINT AVENUE  MC GUINNESS BOULEVARD

          OFF STREET NAME  ...  CONTRIBUTING FACTOR VEHICLE 2  \
0            GOLD STREET  ...                     Unspecified
1                   None  ...                     Unspecified
2  2797      OCEAN PARKWAY  ...                            None
3                   None  ...                            None
4                   None  ...                     Unspecified

   CONTRIBUTING FACTOR VEHICLE 3  CONTRIBUTING FACTOR VEHICLE 4  \
0                           None                           None
1                           None                           None
2                           None                           None
3                           None                           None
4                           None                           None

   CONTRIBUTING FACTOR VEHICLE 5  COLLISION_ID  \
0                           None       4736746
1                           None       4736768
2                           None       4737060
3                           None       4737510
4                           None       4736759

                 VEHICLE TYPE CODE 1                VEHICLE TYPE CODE 2  \
```

```
0                                 Sedan                                   Sedan
1  Station Wagon/Sport Utility Vehicle  Station Wagon/Sport Utility Vehicle
2  Station Wagon/Sport Utility Vehicle                                    None
3                                 Sedan                                   None
4                                   Bus                              Box Truck

   VEHICLE TYPE CODE 3 VEHICLE TYPE CODE 4 VEHICLE TYPE CODE 5
0                 None                None                 None
1                 None                None                 None
2                 None                None                 None
3                 None                None                 None
4                 None                None                 None

[5 rows x 29 columns]
```

1. Construct a contigency table for missing in geocode (latitude and longitude) by borough. Is the missing pattern the same across boroughs? Formulate a hypothesis and test it.

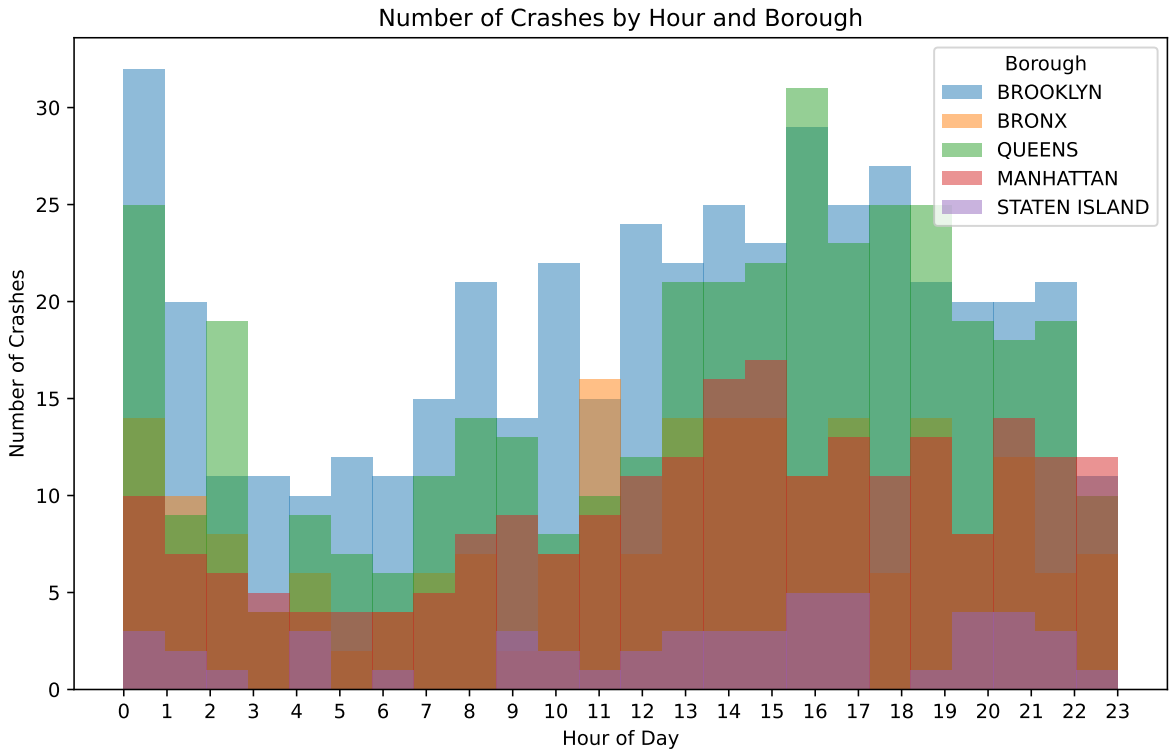
```
dff['Missing_Geocode'] = dff['LATITUDE'].isnull() | dff['LONGITUDE'].isnull()

# Create the contingency table
contingency_table = pd.crosstab(dff['BOROUGH'], dff['Missing_Geocode'])
print(contingency_table)
```

```
Missing_Geocode  False  True
BOROUGH
BRONX              208      5
BROOKLYN           455      7
MANHATTAN          221      7
QUEENS             375      6
STATEN ISLAND       48      2
```

2. Construct a hour variable with integer values from 0 to 23. Plot the histogram of the number of crashes by hour. Plot it by borough.

```
dff['CRASH TIME'] = pd.to_datetime(dff['CRASH TIME'])
dff['hour'] = dff['CRASH TIME'].dt.hour

not_missing_borough = dff.dropna(subset=['BOROUGH'])

plt.figure(figsize=(10, 6))
for borough in not_missing_borough['BOROUGH'].unique():
    borough_data = not_missing_borough[not_missing_borough['BOROUGH'] == borough]
    plt.hist(borough_data['hour'], bins=24, alpha=0.5, label=borough)

plt.xlabel('Hour of Day')
plt.ylabel('Number of Crashes')
plt.title('Number of Crashes by Hour and Borough')
plt.xticks(range(24))
plt.legend(title='Borough')
plt.show()
```

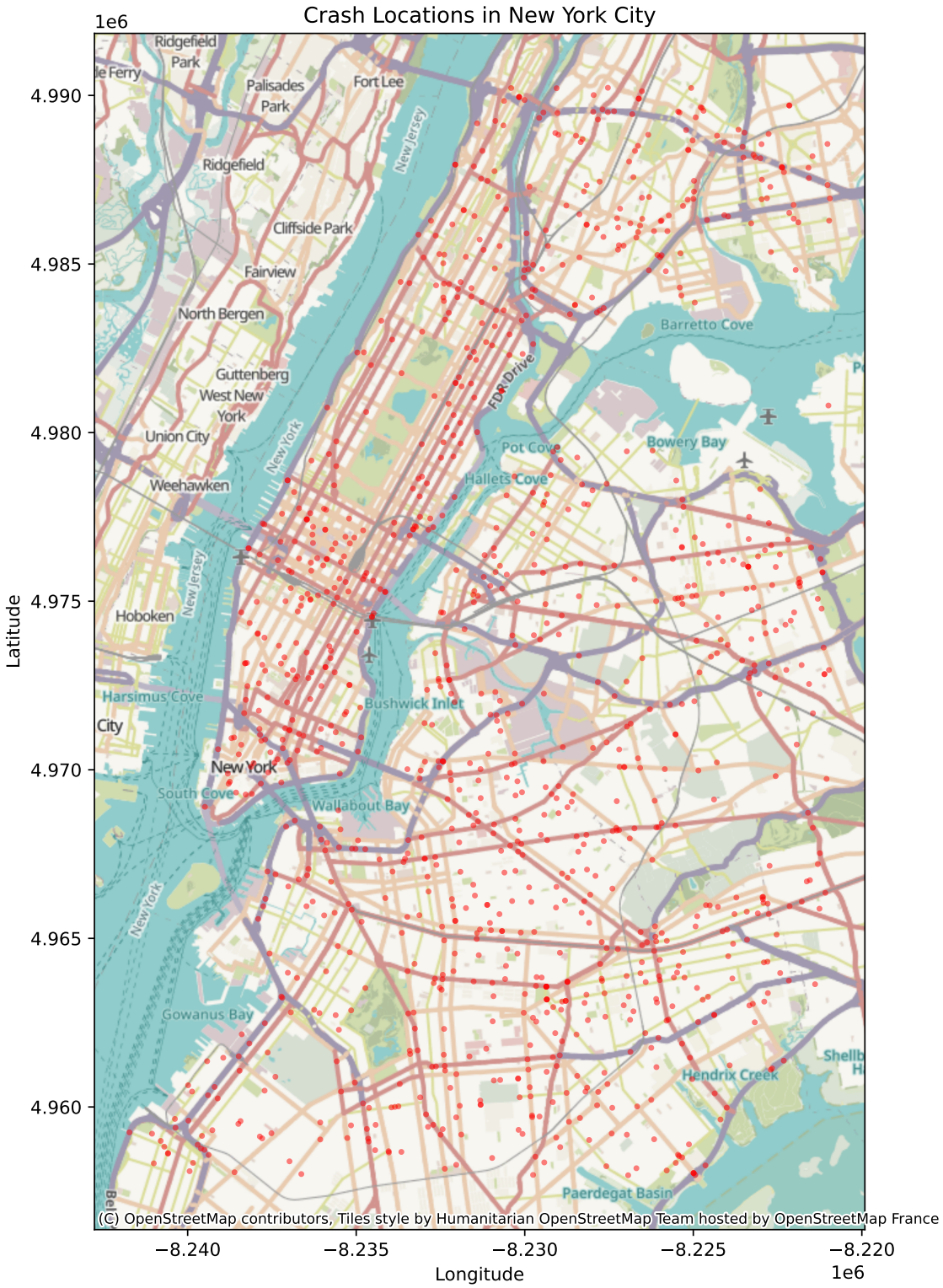
```
C:\Users\aansh\AppData\Local\Temp\ipykernel_10380\4061894526.py:1: UserWarning:

Could not infer format, so each element will be parsed individually, falling back to `dateuti
```



3. Overlay the locations of the crashes on a map of NYC. The map could be a static map or Google map.

```
import geopandas as gpd
from shapely.geometry import Point
import contextily as ctx

# Create a GeoDataFrame
geometry = [Point(xy) for xy in zip(dff['LONGITUDE'], dff['LATITUDE'])]
gdf = gpd.GeoDataFrame(dff, geometry=geometry)
gdf.crs = "EPSG:4326"  # WGS 84

# Filter for NYC bounds
nyc_bounds = gdf.cx[-74.05:-73.85, 40.63:40.85]
if nyc_bounds.empty:
    print("No crash data within NYC bounds.")
else:
```

```
    # Reproject to Web Mercator
    gdf_nyc = nyc_bounds.to_crs(epsg=3857)

    # Plotting
    fig, ax = plt.subplots(figsize=(12, 12))
    gdf_nyc.plot(ax=ax, marker='o', color='red', markersize=5, alpha=0.5)
    ctx.add_basemap(ax, crs=gdf_nyc.crs.to_string())

    plt.title("Crash Locations in New York City")
    plt.xlabel("Longitude")
    plt.ylabel("Latitude")
    plt.savefig('nyc_crashes_overlay.png', dpi=300)
    plt.show()
```


Crash Locations in New York City
1e6
4.990
4.985
4.980
4.975
4.970
4.965
4.960
Latitude
−8.240
−8.235
−8.230
−8.225
−8.220
1e6
Longitude
Ridgefield Park
Palisades Park
Fort Lee
Ridgefield
Cliffside Park
Fairview
North Bergen
Guttenberg
West New York
Union City
Weehawken
Hoboken
Harsimus Cove
City
New York
South Cove
Wallabout Bay
Bushwick Inlet
Gowanus Bay
Barretto Cove
Pot Cove
Hallets Cove
Bowery Bay
Hendrix Creek
Paerdegat Basin
New Jersey
New York
FDR Drive
(C) OpenStreetMap contributors, Tiles style by Humanitarian OpenStreetMap Team hosted by OpenStreetMap France

4. Create a new variable severe which is one if the number of persons injured of deaths is 1 or more; and zero otherwise. Construct a cross table for severe versus borough. Is the severity of the crashes the same across boroughs? Test the null hypothesis that the two variables are not associated with an appropriate test.

```
import scipy.stats as stats

# Create the 'SEVERE' variable
dff['SEVERE'] = ((dff['NUMBER OF PERSONS INJURED'] > 0) | (dff['NUMBER OF PERSONS KILLED'] >

# Construct the crosstab
severity_borough_table = pd.crosstab(dff['BOROUGH'], dff['SEVERE'])
print("Crosstab of SEVERE vs BOROUGH:")
print(severity_borough_table)

# Perform chi-square test
chi2, p, dof, expected = stats.chi2_contingency(severity_borough_table)
print("\nChi-Square Test Results:")
print(f"Chi-Square Statistic: {chi2}")
print(f"P-value: {p}")
print(f"Degrees of Freedom: {dof}")

# Interpret the result
alpha = 0.05
if p < alpha:
    print("Reject the null hypothesis: There is an association between severity and borough.
else:
    print("Fail to reject the null hypothesis: No association between severity and borough.")
```

```
Crosstab of SEVERE vs BOROUGH:
SEVERE           0    1
BOROUGH
BRONX          115   98
BROOKLYN       262  200
MANHATTAN      119  109
QUEENS         205  176
STATEN ISLAND   35   15

Chi-Square Test Results:
```

```
Chi-Square Statistic: 6.112652162659586
P-value: 0.19089181529657664
Degrees of Freedom: 4
Fail to reject the null hypothesis: No association between severity and borough.
```

5. Merge the crash data with the zip code database.(Unsolvable!?)

6. Fit a logistic model with severe as the outcome variable and covariates that are available in the data or can be engineered from the data. For example, zip code level covariates can be obtained by merging with the zip code database.(Unsolvable!?)