Aansh Jha Homework 4

1 - Use the filter from the website to download the crash data of the week of June 30, 2024 in CSV format; save it under a directory `data` with an informative name (e.g., `nyccrashes_2024w0630_by20240916.csv`); read the data into a Panda data frame with careful handling of the date time variables.

```
import pandas as pd
import numpy as np
import matplotlib


df = pd.read_csv('nyccrashes_2024w0630_by20240916.csv')
```

2 - Clean up the variable names. Use lower cases and replace spaces with underscores.

```
df.columns = df.columns.str.lower().str.replace(' ', '_')
```

3 - Get the basic summaries of each variables: missing percentage; descriptive statistics for continuous variables; frequency tables for discrete variables.

```
missing_percentage = df.isna().mean() * 100
print("Missing Percentage:", missing_percentage)
descriptive_stats = df.describe()
print("Descriptive Statistics:", descriptive_stats)
freq_tables = {}
for col in df.select_dtypes(include=['object', 'category']).columns:
    value_counts = df[col].value_counts()
    freq_tables[col] = value_counts
print("Frequency Tables:", freq_tables)
```

```
Missing Percentage: crash_date                      0.000000
crash_time                          0.000000
borough                            28.434886
zip_code                           28.434886
latitude                            7.885305
longitude                           7.885305
location                            7.885305
on_street_name                     28.853047
cross_street_name                  49.223417
off_street_name                    71.146953
number_of_persons_injured           0.000000
number_of_persons_killed            0.000000
number_of_pedestrians_injured       0.000000
number_of_pedestrians_killed        0.000000
number_of_cyclist_injured           0.000000
number_of_cyclist_killed            0.000000
number_of_motorist_injured          0.000000
number_of_motorist_killed           0.000000
contributing_factor_vehicle_1       0.477897
contributing_factor_vehicle_2      23.416965
contributing_factor_vehicle_3      90.860215
contributing_factor_vehicle_4      97.252091
contributing_factor_vehicle_5      99.163680
collision_id                        0.000000
vehicle_type_code_1                 1.612903
vehicle_type_code_2                33.512545
vehicle_type_code_3                91.517324
vehicle_type_code_4                97.431302
vehicle_type_code_5                99.163680
dtype: float64
Descriptive Statistics:            zip_code     latitude    longitude  number_of_persons_inju
```

```
       1198.000000  1542.000000  1542.000000                1674.000000
mean  10895.911519    40.533450   -73.585296                   0.625448
std     530.386669     2.739299     4.971539                   0.929093
min   10001.000000     0.000000   -74.237366                   0.000000
25%   10456.000000    40.661164   -73.967523                   0.000000
50%   11208.000000    40.712260   -73.923765                   0.000000
75%   11237.000000    40.766286   -73.869988                   1.000000
max   11694.000000    40.907246     0.000000                  11.000000

       number_of_persons_killed  number_of_pedestrians_injured  \
count               1674.000000                    1674.000000
mean                   0.004779                       0.093190
std                    0.109232                       0.343553
min                    0.000000                       0.000000
25%                    0.000000                       0.000000
50%                    0.000000                       0.000000
75%                    0.000000                       0.000000
max                    4.000000                       7.000000

       number_of_pedestrians_killed  number_of_cyclist_injured  \
count                   1674.000000                1674.000000
mean                       0.002987                   0.067503
std                        0.100759                   0.250966
min                        0.000000                   0.000000
25%                        0.000000                   0.000000
50%                        0.000000                   0.000000
75%                        0.000000                   0.000000
max                        4.000000                   1.000000

       number_of_cyclist_killed  number_of_motorist_injured  \
count                    1674.0                 1674.000000
mean                        0.0                    0.439068
std                         0.0                    0.903035
min                         0.0                    0.000000
25%                         0.0                    0.000000
50%                         0.0                    0.000000
75%                         0.0                    1.000000
max                         0.0                   11.000000

       number_of_motorist_killed  collision_id
count                1674.000000  1.674000e+03
mean                    0.001792  4.738487e+06
std                     0.042308  1.631902e+03
```

```
min                     0.000000  4.736561e+06
25%                     0.000000  4.737587e+06
50%                     0.000000  4.738146e+06
75%                     0.000000  4.738776e+06
max                     1.000000  4.752444e+06
Frequency Tables: {'crash_date': crash_date
07/03/2024    258
07/05/2024    254
07/02/2024    251
06/30/2024    249
07/01/2024    247
07/06/2024    212
07/04/2024    203
Name: count, dtype: int64, 'crash_time': crash_time
0:00     35
15:00    19
19:00    18
13:00    17
14:30    15
         ..
20:28     1
22:28     1
15:03     1
22:23     1
18:39     1
Name: count, Length: 735, dtype: int64, 'borough': borough
BROOKLYN         416
QUEENS           336
MANHATTAN        212
BRONX            189
STATEN ISLAND     45
Name: count, dtype: int64, 'location': location
(0.0, 0.0)                7
(40.668507, -73.92561)    4
(40.848038, -73.93285)    3
(40.762856, -73.98942)    3
(40.753326, -73.8718)     3
                         ..
(40.731968, -73.88478)    1
(40.733383, -73.95216)    1
(40.642082, -73.94963)    1
(40.72565, -73.93205)     1
(40.67989, -73.94034)     1
```

```
Name: count, Length: 1476, dtype: int64, 'on_street_name': on_street_name
BELT PARKWAY                  28
FDR DRIVE                     15
LONG ISLAND EXPRESSWAY        14
BROOKLYN QUEENS EXPRESSWAY    12
BROADWAY                      12
                              ..
METCALF AVENUE                 1
WEST 162 STREET                1
PEARL STREET                   1
CAMP ROAD                      1
WEST 157 STREET                1
Name: count, Length: 651, dtype: int64, 'cross_street_name': cross_street_name
BROADWAY              16
BRUCKNER BOULEVARD    10
3 AVENUE               8
2 AVENUE               7
PARK AVENUE            6
                      ..
CLINTON AVENUE         1
FRANKFORT STREET       1
SEAGIRT BOULEVARD      1
RICHARDSON STREET      1
AVENUE L               1
Name: count, Length: 604, dtype: int64, 'off_street_name': off_street_name
369       HYLAN BOULEVARD      2
2724      UNIVERSITY AVENUE    2
1825      EASTCHESTER ROAD     2
682       ROGERS AVENUE        1
217       WALKER STREET        1
                              ..
1466      53 STREET            1
24        EAST 120 STREET      1
5218      13 AVENUE            1
62-60     99 STREET            1
2797      OCEAN PARKWAY        1
Name: count, Length: 480, dtype: int64, 'contributing_factor_vehicle_1': contributing_factor
Unspecified                                    423
Driver Inattention/Distraction                 404
Failure to Yield Right-of-Way                  109
Following Too Closely                           89
Unsafe Speed                                    74
Passing or Lane Usage Improper                  66
```

```
Other Vehicular                                          60
Traffic Control Disregarded                              51
Alcohol Involvement                                      50
Passing Too Closely                                      50
Driver Inexperience                                      45
Backing Unsafely                                         45
Turning Improperly                                       40
Pedestrian/Bicyclist/Other Pedestrian Error/Confusion    22
Unsafe Lane Changing                                     22
Reaction to Uninvolved Vehicle                           14
Aggressive Driving/Road Rage                             13
View Obstructed/Limited                                  13
Pavement Slippery                                         9
Fell Asleep                                               8
Oversized Vehicle                                         8
Brakes Defective                                          6
Lost Consciousness                                        5
Tire Failure/Inadequate                                   5
Outside Car Distraction                                   4
Glare                                                     3
Illnes                                                    3
Failure to Keep Right                                     3
Cell Phone (hand-Held)                                    2
Passenger Distraction                                     2
Driverless/Runaway Vehicle                                2
Tinted Windows                                            2
Steering Failure                                          2
Accelerator Defective                                     2
Tow Hitch Defective                                       2
Pavement Defective                                        2
Obstruction/Debris                                        2
Fatigued/Drowsy                                           2
Lane Marking Improper/Inadequate                          1
Cell Phone (hands-free)                                   1
Name: count, dtype: int64, 'contributing_factor_vehicle_2': contributing_factor_vehicle_2
Unspecified                                           1086
Driver Inattention/Distraction                          63
Other Vehicular                                         26
Unsafe Speed                                            19
Following Too Closely                                   13
Failure to Yield Right-of-Way                           13
Passing or Lane Usage Improper                          10
Traffic Control Disregarded                              9
```

```
Pedestrian/Bicyclist/Other Pedestrian Error/Confusion       9
Driver Inexperience                                         4
Unsafe Lane Changing                                        4
Aggressive Driving/Road Rage                                4
Turning Improperly                                          4
Passing Too Closely                                         4
View Obstructed/Limited                                     3
Alcohol Involvement                                         3
Backing Unsafely                                            2
Reaction to Uninvolved Vehicle                              2
Passenger Distraction                                       1
Drugs (illegal)                                             1
Failure to Keep Right                                       1
Fatigued/Drowsy                                             1
Name: count, dtype: int64, 'contributing_factor_vehicle_3': contributing_factor_vehicle_3
Unspecified                      147
Other Vehicular                    3
Unsafe Speed                       2
Aggressive Driving/Road Rage       1
Name: count, dtype: int64, 'contributing_factor_vehicle_4': contributing_factor_vehicle_4
Unspecified                      45
Aggressive Driving/Road Rage      1
Name: count, dtype: int64, 'contributing_factor_vehicle_5': contributing_factor_vehicle_5
Unspecified    14
Name: count, dtype: int64, 'vehicle_type_code_1': vehicle_type_code_1
Sedan                                   769
Station Wagon/Sport Utility Vehicle     569
Taxi                                     49
Bike                                     37
Pick-up Truck                            35
Motorcycle                               24
Box Truck                                23
Bus                                      22
E-Bike                                   14
Moped                                    13
Tractor Truck Diesel                     11
E-Scooter                                11
Ambulance                                10
Van                                       6
Convertible                               5
Dump                                      5
Motorscooter                              4
PK                                        3
```

```
Chassis Cab                              3
Tractor Truck Gasoline                   3
Garbage or Refuse                        3
subn                                     3
Flat Bed                                 3
Beverage Truck                           2
MOPED                                    2
Armored Truck                            2
Motorbike                                2
TRUCK                                    1
U-Haul                                   1
AMBULANCE                                1
moped                                    1
Multi-Wheeled Vehicle                    1
LIMO                                     1
FDNY FIRE                                1
Flat Rack                                1
PICK UP                                  1
R/V                                      1
MTA BUS                                  1
Tow Truck / Wrecker                      1
Pedicab                                  1
UNK                                      1
Name: count, dtype: int64, 'vehicle_type_code_2': vehicle_type_code_2
Sedan                                  447
Station Wagon/Sport Utility Vehicle    319
Bike                                    76
Box Truck                               34
Moped                                   32
Pick-up Truck                           27
E-Scooter                               24
Taxi                                    22
E-Bike                                  22
Bus                                     22
Motorcycle                              20
Tractor Truck Diesel                    13
Van                                      6
Garbage or Refuse                        5
Carry All                                4
PK                                       4
Chassis Cab                              4
Dump                                     3
Motorbike                                3
```

```
Ambulance                                3
Flat Bed                                 2
Convertible                              2
UNKNOWN                                  2
Unknown                                  2
Van Camper                               1
Tow Truck / Wrecker                      1
Scooter                                  1
TRUCK                                    1
Power shov                               1
SCOOTER                                  1
MOPED                                    1
COURIER VA                               1
UK                                       1
FORKLIFT                                 1
PASSENGER                                1
Sprinter V                               1
3-Door                                   1
UHAUL VAN                                1
Motorscooter                             1
Name: count, dtype: int64, 'vehicle_type_code_3': vehicle_type_code_3
Sedan                                   73
Station Wagon/Sport Utility Vehicle     51
Bus                                      5
Pick-up Truck                            5
Taxi                                     2
Tractor Truck Diesel                     2
Bike                                     1
Moped                                    1
Van                                      1
Motorcycle                               1
Name: count, dtype: int64, 'vehicle_type_code_4': vehicle_type_code_4
Station Wagon/Sport Utility Vehicle     21
Sedan                                   19
Convertible                              1
Pick-up Truck                            1
Motorcycle                               1
Name: count, dtype: int64, 'vehicle_type_code_5': vehicle_type_code_5
Sedan                                   9
Station Wagon/Sport Utility Vehicle     3
Bike                                    1
Box Truck                               1
Name: count, dtype: int64}
```

4 - Are their invalid `longitude` and `latitude` in the data? If so, replace them with `NA`.

```
latitude_col = 'latitude'
longitude_col = 'longitude'

df[latitude_col] = np.where((df[latitude_col] < -90) | (df[latitude_col] > 90), np.nan, df[la

df[longitude_col] = np.where((df[longitude_col] < -180) | (df[longitude_col] > 180), np.nan,

print("Latitude sample:", df[latitude_col].head())
print("Longitude sample:", df[longitude_col].head())
```

```
Latitude sample: 0    40.817250
1    40.788795
2    40.666225
3    40.714104
4    40.829002
Name: latitude, dtype: float64
Longitude sample: 0   -73.90649
1   -73.93755
2   -73.80086
3   -73.95322
4   -73.91557
Name: longitude, dtype: float64
```

5 - Are there `zip_code` values that are not legit NYC zip codes? If so, replace them with `NA`.

```
valid_codes = set()

for i in range(10001, 14975):
    valid_codes.add(i)

for index, zip_code in df['zip_code'].items():
    if zip_code in valid_codes:
        df.at[index, 'zip_code'] = zip_code
    else:
```

```
        df.at[index, 'zip_code'] = np.nan

print("Updated ZIP Codes:", df['zip_code'].head())
```

```
Updated ZIP Codes: 0    10455.0
1        NaN
2        NaN
3        NaN
4    10456.0
Name: zip_code, dtype: float64
```

6: Are there missing in `zip_code` and `borough`? Do they always co-occur?

```
missing_zip = df['zip_code'].isna()
missing_borough = df['borough'].isna()

co_occur_missing_count = df[missing_zip & missing_borough].shape[0]
print(f"Number of cases where both ZIP and Borough are missing: {co_occur_missing_count}")
```

```
Number of cases where both ZIP and Borough are missing: 476
```

7 - Are there cases where `zip_code` and `borough` are missing but the geo codes are not missing? If so, fill in `zip_code` and `borough` using the geo codes.

```
missing_geo = df[missing_zip & missing_borough & df['latitude'].notna() & df['longitude'].not

print(f"Number of missing ZIP and Borough with available geolocation: {missing_geo.shape[0]}"

df.loc[missing_zip & missing_borough, 'zip_code'] = df['latitude']
df.loc[missing_zip & missing_borough, 'borough'] = df['longitude']

print("Sample of updated records:", df[['zip_code', 'borough']].head())
```

```
Number of missing ZIP and Borough with available geolocation: 370
Sample of updated records:        zip_code   borough
0  10455.000000     BRONX
1     40.788795 -73.93755
2     40.666225 -73.80086
3     40.714104 -73.95322
4  10456.000000     BRONX
```

8 - Is it redundant to keep both location and the longitude/latitude at the NYC Open Data server?

Longitude and latitude offer precise geolocation for accurate mapping, while zip codes and boroughs are more user-friendly. Keeping both is useful for identifying NYC hotspots through geographic analysis.

9 - Check the frequency of `crash_time` by hour. Is there a matter of bad luck at exactly midnight? How would you interpret this? Is there a matter of bad luck at exactly midnight? How would you interpret this?

```python
df['crash_hour'] = pd.to_datetime(df['crash_time'], format='%H:%M').dt.hour
crash_hourly_frequency = df['crash_hour'].value_counts().sort_index()
print("Crash Frequency by Hour:", crash_hourly_frequency)
midnight_crash_count = crash_hourly_frequency.get(0, 0)
print(f"Crashes at midnight: {midnight_crash_count}")
```

```
Crash Frequency by Hour: crash_hour
0     105
1      55
2      57
3      36
4      40
5      36
6      37
7      47
8      65
9      52
```

```
10     59
11     61
12     72
13     94
14     97
15     92
16    101
17     99
18     86
19     96
20     67
21     81
22     78
23     61
Name: count, dtype: int64
Crashes at midnight: 105
```

Crashes at midnight may be more frequent due to factors like low visibility, driver fatigue, or impaired driving. Not necessarily "bad luck" but a result of riskier driving conditions at that time.

10 - Are the number of persons killed/injured the summation of the numbers of pedestrians, cyclist, and motorists killed/injured? If so, is it redundant to keep these two columns at the NYC Open Data server?

```
df['calculated_injured_total'] = df[['number_of_pedestrians_injured', 'number_of_cyclist_inju
df['calculated_killed_total'] = df[['number_of_pedestrians_killed', 'number_of_cyclist_kille

injured_match = (df['number_of_persons_injured'] == df['calculated_injured_total']).all()
killed_match = (df['number_of_persons_killed'] == df['calculated_killed_total']).all()

print(f"Injury totals match: {injured_match}")
print(f"Fatality totals match: {killed_match}")
```

```
Injury totals match: False
Fatality totals match: True
```

It's redundant to keep both killed and injured columns since discrepancies exist between reported totals and the sum of specific injury categories. The summation of injured columns should be fixed for accuracy.

11 - Print the whole frequency table of `contributing_factor_vehicle_1`. Convert lower cases to uppercases and check the frequencies again.

```
initial_factors_freq = df['contributing_factor_vehicle_1'].value_counts()
print("Initial Frequency Table:", initial_factors_freq)

df['contributing_factor_vehicle_1'] = df['contributing_factor_vehicle_1'].str.upper()

upper_factors_freq = df['contributing_factor_vehicle_1'].value_counts()
print("Uppercase Frequency Table:", upper_factors_freq)
```

```
Initial Frequency Table: contributing_factor_vehicle_1
Unspecified                                              423
Driver Inattention/Distraction                           404
Failure to Yield Right-of-Way                            109
Following Too Closely                                     89
Unsafe Speed                                              74
Passing or Lane Usage Improper                            66
Other Vehicular                                           60
Traffic Control Disregarded                               51
Alcohol Involvement                                       50
Passing Too Closely                                       50
Driver Inexperience                                       45
Backing Unsafely                                          45
Turning Improperly                                        40
Pedestrian/Bicyclist/Other Pedestrian Error/Confusion     22
Unsafe Lane Changing                                      22
Reaction to Uninvolved Vehicle                            14
Aggressive Driving/Road Rage                              13
View Obstructed/Limited                                   13
Pavement Slippery                                          9
Fell Asleep                                                8
Oversized Vehicle                                          8
Brakes Defective                                           6
```

```
Lost Consciousness                                    5
Tire Failure/Inadequate                               5
Outside Car Distraction                               4
Glare                                                 3
Illnes                                                3
Failure to Keep Right                                 3
Cell Phone (hand-Held)                                2
Passenger Distraction                                 2
Driverless/Runaway Vehicle                            2
Tinted Windows                                        2
Steering Failure                                      2
Accelerator Defective                                 2
Tow Hitch Defective                                   2
Pavement Defective                                    2
Obstruction/Debris                                    2
Fatigued/Drowsy                                       2
Lane Marking Improper/Inadequate                      1
Cell Phone (hands-free)                               1
Name: count, dtype: int64
Uppercase Frequency Table: contributing_factor_vehicle_1
UNSPECIFIED                                         423
DRIVER INATTENTION/DISTRACTION                      404
FAILURE TO YIELD RIGHT-OF-WAY                       109
FOLLOWING TOO CLOSELY                                89
UNSAFE SPEED                                         74
PASSING OR LANE USAGE IMPROPER                       66
OTHER VEHICULAR                                      60
TRAFFIC CONTROL DISREGARDED                          51
ALCOHOL INVOLVEMENT                                  50
PASSING TOO CLOSELY                                  50
DRIVER INEXPERIENCE                                  45
BACKING UNSAFELY                                     45
TURNING IMPROPERLY                                   40
PEDESTRIAN/BICYCLIST/OTHER PEDESTRIAN ERROR/CONFUSION    22
UNSAFE LANE CHANGING                                 22
REACTION TO UNINVOLVED VEHICLE                       14
AGGRESSIVE DRIVING/ROAD RAGE                         13
VIEW OBSTRUCTED/LIMITED                              13
PAVEMENT SLIPPERY                                     9
FELL ASLEEP                                           8
OVERSIZED VEHICLE                                     8
BRAKES DEFECTIVE                                      6
LOST CONSCIOUSNESS                                    5
```

```
TIRE FAILURE/INADEQUATE                          5
OUTSIDE CAR DISTRACTION                          4
GLARE                                            3
ILLNES                                           3
FAILURE TO KEEP RIGHT                            3
CELL PHONE (HAND-HELD)                           2
PASSENGER DISTRACTION                            2
DRIVERLESS/RUNAWAY VEHICLE                       2
TINTED WINDOWS                                   2
STEERING FAILURE                                 2
ACCELERATOR DEFECTIVE                            2
TOW HITCH DEFECTIVE                              2
PAVEMENT DEFECTIVE                               2
OBSTRUCTION/DEBRIS                               2
FATIGUED/DROWSY                                  2
LANE MARKING IMPROPER/INADEQUATE                 1
CELL PHONE (HANDS-FREE)                          1
Name: count, dtype: int64
```

12 - Provided an opportunity to meet the data provider, what suggestions would you make based on your data exploration experience?

Fix discrepancies in injury/fatality totals and ensure valid geolocation data