NYC Analysis Report

1) For generating the analysis report , Started with installing ydata_profiling python library

```
pip install ydata_profiling
```

2) Installed Pandas Library

```
pip install pandas
```

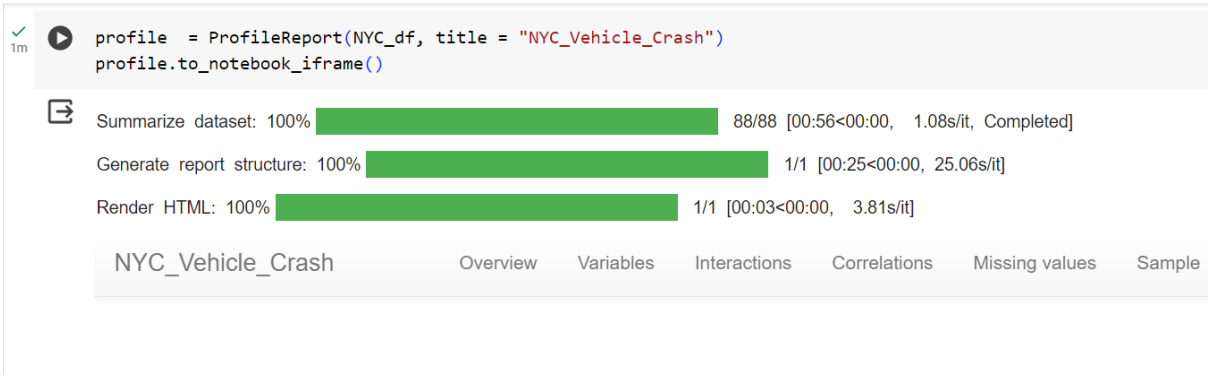
```
Requirement already satisfied: pandas in /Library/Frameworks/Python.framework/Versions/3.11/lib/python3.11/site-packages (2.0.3)
Requirement already satisfied: python-dateutil>=2.8.2 in /Library/Frameworks/Python.framework/Versions/3.11/lib/python3.11/site-packages (from pandas)
Requirement already satisfied: pytz>=2020.1 in /Library/Frameworks/Python.framework/Versions/3.11/lib/python3.11/site-packages (from pandas) (2023.3)
Requirement already satisfied: tzdata>=2022.1 in /Library/Frameworks/Python.framework/Versions/3.11/lib/python3.11/site-packages (from pandas) (2023.3
Requirement already satisfied: numpy>=1.21.0 in /Library/Frameworks/Python.framework/Versions/3.11/lib/python3.11/site-packages (from pandas) (1.25.2)
Requirement already satisfied: six>=1.5 in /Library/Frameworks/Python.framework/Versions/3.11/lib/python3.11/site-packages (from python-dateutil>=2.8.

[notice] A new release of pip is available: 23.1.2 -> 24.0
[notice] To update, run: pip3 install --upgrade pip
Note: you may need to restart the kernel to use updated packages.
```

3) By taking NYC_df dataframe we read the NYC dataset and given below is the output

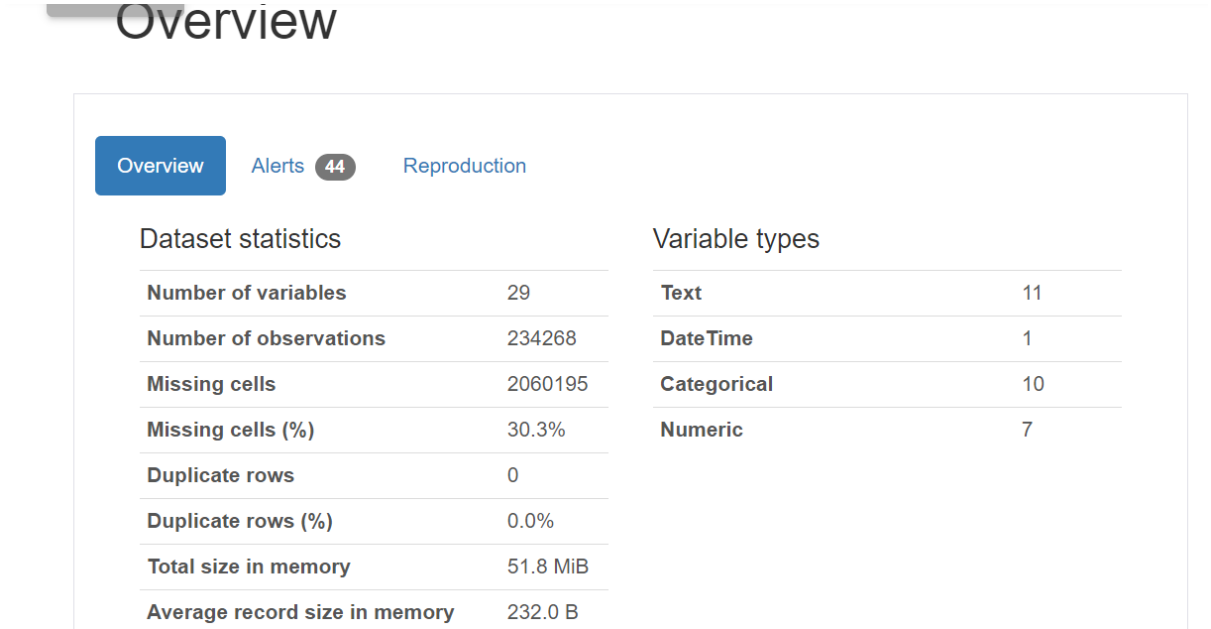
```
NYC_df = pd.read_csv('NY_Motor_Vehicle_Collisions_-_Crashes_20240326.tsv', sep='\t')
print(NYC_df)
```

```
234264                    Unspecified                          NaN
234265  Driver Inattention/Distraction                         NaN
234266                    Unspecified                          NaN
234267                            NaN                          NaN

        CONTRIBUTING FACTOR VEHICLE 4  CONTRIBUTING FACTOR VEHICLE 5  \
0                                 NaN                            NaN
1                                 NaN                            NaN
2                                 NaN                            NaN
3                                 NaN                            NaN
4                                 NaN                            NaN
...                               ...                            ...
234263                            NaN                            NaN
234264                            NaN                            NaN
234265                            NaN                            NaN
234266                            NaN                            NaN
234267                            NaN                            NaN

        COLLISION_ID                VEHICLE TYPE CODE 1  \
0          4455765.0                              Sedan
1          4513547.0                              Sedan
2          4541903.0                              Sedan
```

4) Here we got the complete analysis report of the dataset by executing this code



5) Given below is the primary statistics of the NYC dataset



6) Complete report generation analysis:

Reproduction

Analysis started	2024-04-05 04:22:02.169069
Analysis finished	2024-04-05 04:22:58.842302
Duration	56.67 seconds
Software version	ydata-profiling vv4.7.0
Download configuration	config.json

➔Variable Interpretation

There are in total 29 columns in NYC Dataset and here we have inferred about one column i.e. ZIP CODE

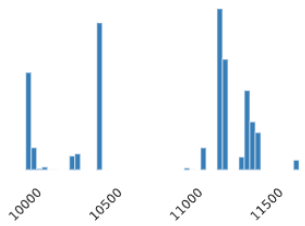
Interpretation of this column is given below

ZIP CODE

Real number (ℝ)

HIGH_CORRELATION MISSING

Distinct	217	**Minimum**	10000
Distinct (%)	0.1%	**Maximum**	11697
Missing	80914	**Zeros**	0
Missing (%)	34.5%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	10895.993	**Memory size**	1.8 MiB



Given above is the histogram for the ZIP CODE variable from a dataset, which is categorized as a real number .

Statistical Summary:

1. Uniqueness:

• The ZIP CODE variable has 217 distinct values, which is a small fraction (0.1%) of the dataset. This indicates that there are many repeated zip code values, as one would expect in a large dataset covering multiple entries per geographical area.

2. Missing Data:

• A significant portion, 34.5%, of the ZIP CODE data is missing. This high percentage of missing data could impact the reliability of any geographical analysis and may necessitate data cleaning or imputation efforts.

3. Range of Data:

• The minimum and maximum values are 10000 and 11697, respectively, which seem to fall within the range of valid U.S. zip codes.

4. Data Integrity:

• There are no zero, infinite, or negative values, which is appropriate for zip codes as they are positive integers and typically five digits long in the context of U.S. addresses.

5. Mean Value:

• The mean value of the ZIP CODE variable is approximately 10895.993, which suggests that the central tendency of the available zip code data falls in this range.

6. Memory Size:

• The memory size occupied by the ZIP CODE data is 1.8 MiB, indicating the variable's size in the dataset.

Histogram Analysis:

• The histogram shows the frequency distribution of ZIP CODE values. There are a few peaks where certain zip codes appear more frequently. However, without the counts on the y-axis, the exact frequency of these zip codes cannot be determined.

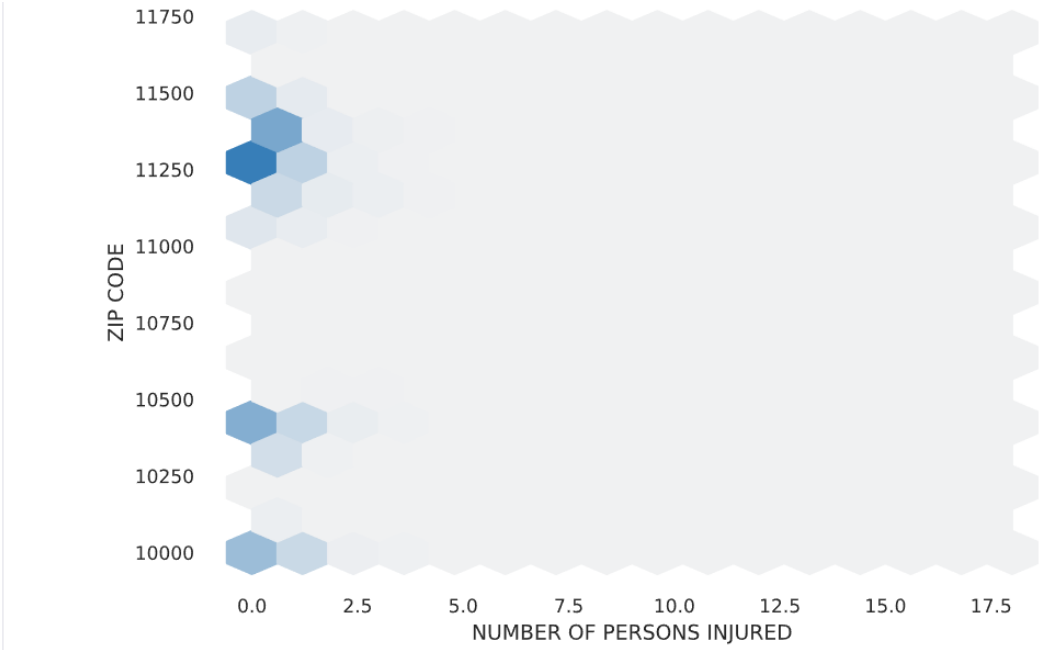
• ZIP Code Distribution: The ZIP codes represented on the vertical axis span from around 10000 to approximately 11750. This range likely covers a specific geographical region.

• Injury Counts: The number of persons injured, shown on the horizontal axis, ranges from 0 to over 17. The plot points suggest that most recorded incidents involve fewer than 5 persons injured.

• Data Density: The darker and larger hexagons around lower injury counts (near 0 to 5) indicate a higher number of incidents in these ranges, suggesting that incidents with fewer injuries are more common.

• Geographical Spread: There is no apparent clustering of hexagons at any specific ZIP code, which could imply that no single ZIP code area is significantly more prone to incidents with high numbers of injuries.

➔ Interaction between two columns of the data and Inference about it:



Some of the Observations are:

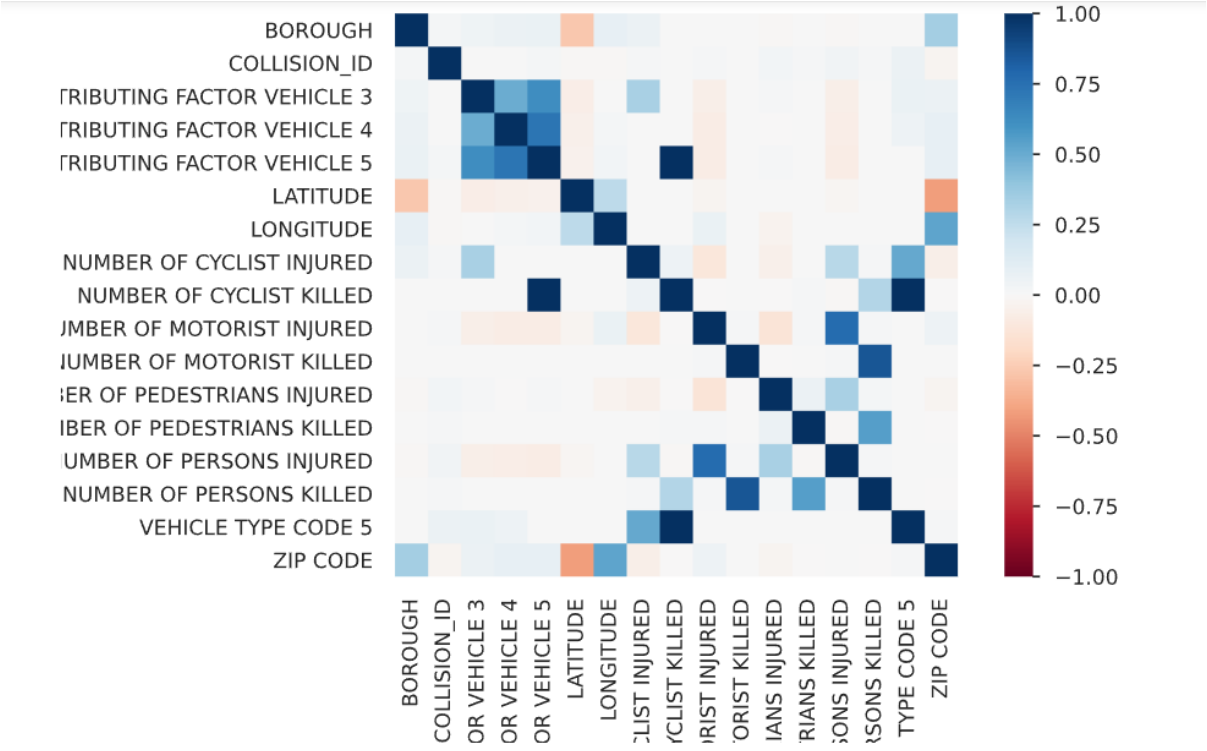
• ZIP Code Distribution: The ZIP codes represented on the vertical axis span from around 10000 to approximately 11750. This range likely covers a specific geographical region.

• Injury Counts: The number of persons injured, shown on the horizontal axis, ranges from 0 to over 17. The plot points suggest that most recorded incidents involve fewer than 5 persons injured.

• Data Density: The darker and larger hexagons around lower injury counts (near 0 to 5) indicate a higher number of incidents in these ranges, suggesting that incidents with fewer injuries are more common.

• Geographical Spread: There is no apparent clustering of hexagons at any specific ZIP code, which could imply that no single ZIP code area is significantly more prone to incidents with high numbers of injuries.

➔ Analysis of Heatmap of the Correlations:



The heatmap provided is a graphical representation of correlation coefficients between various variables typically associated with traffic incidents. The color scheme of the heatmap, ranging from blue (negative correlation) to red (positive correlation), with varying shades in between, helps visualize the strength and direction of the relationships.

For this heatmap key observations are:

1. Latitude and Longitude:
• There appears to be little to no strong correlation between geographic coordinates (latitude and longitude) and other variables. This could suggest that the location data is quite random and not strongly related to the specific factors considered in this dataset.

2. Cyclist and Motorist Injuries and Fatalities:
• There are noticeable correlations within the injury and fatality counts for cyclists, motorists, and pedestrians. This suggests that when an incident results in injury or fatality to one type of road user, there's a likelihood that other types may also be affected.

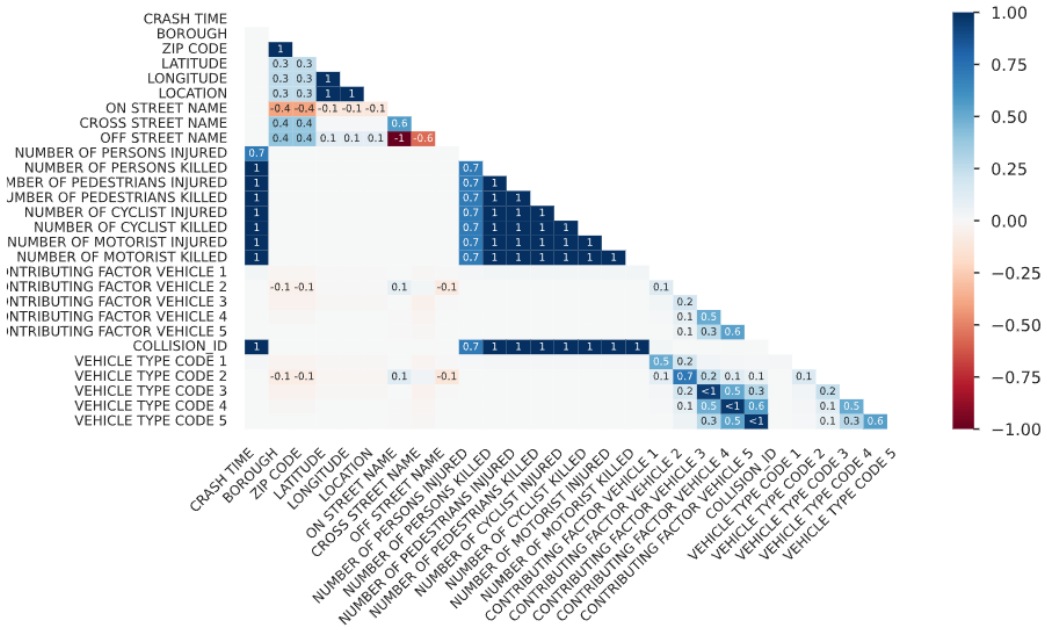
3. Correlation Between Injury and Fatality Counts:
• The dark blue squares indicate a high negative correlation in certain areas. This might be reflecting instances where higher counts in one category (like injuries) might correspond to lower counts in another (like fatalities), or vice versa.

4. Zip Code:
• The 'ZIP CODE' variable seems to have some correlations with other variables, although the exact nature and strength of these correlations are not fully clear from the provided image. The correlation with 'BOROUGH' might suggest that certain boroughs are more prone to incidents resulting in injuries or fatalities.

5. Contributing Factors for Vehicles:

• The contributing factors for vehicles (3, 4, and 5) show some degree of correlation with one another. This might indicate a commonality in reporting or the presence of multiple contributing factors in the same incidents.

➔ Analysis for the Missing Values:



This heatmap shows pattern of missing data among various variables in a dataset related to traffic incidents or collisions. Also it employs a colour scale to indicate the correlation of missing data between pairs of variables, with blue representing a positive correlation, red indicating a negative correlation, and the intensity of the colour denoting the strength of the correlation.

Some of the observations from the heatmap are:

1. Injury and Fatality Variables:

• There is a very strong positive correlation (dark blue and 1 on the scale) among the variables for the number of persons, cyclists, motorists, and pedestrians injured or killed. This implies that if the data

for one of these variables is missing, the data for the others are likely to be missing as well.

2. Location Variables:
• 'ZIP CODE' and 'BOROUGH' show a strong positive correlation with missing data, suggesting that entries missing a 'ZIP CODE' also often lack 'BOROUGH' information, which aligns with these being related geographic identifiers.

3. Vehicle Contributing Factors:
• There are light correlations between the missing data of different contributing factors for vehicles, which indicates that missing data in one of these factors does not always lead to missing data in others.

4. Collision ID:
• 'COLLISION_ID' has strong positive correlations with missing data across several vehicle type codes and contributing factors, suggesting that if a 'COLLISION_ID' is missing, there's a good chance that specific details regarding vehicle types and contributing factors will also be missing.

5. Vehicle Types:
• The vehicle type codes show varying degrees of correlation with missing data, suggesting inconsistencies in how data is recorded or reported for different types of vehicles.