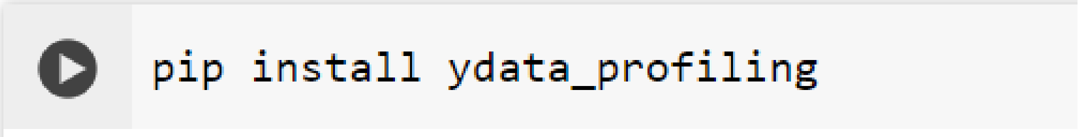
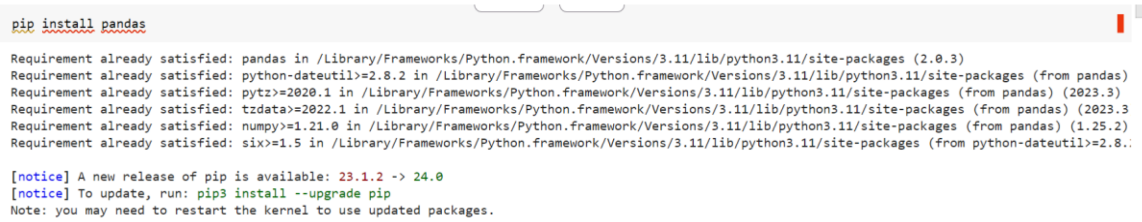
Chicago Analysis Report

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



```
pip install ydata_profiling
```

2) Installed Pandas Library

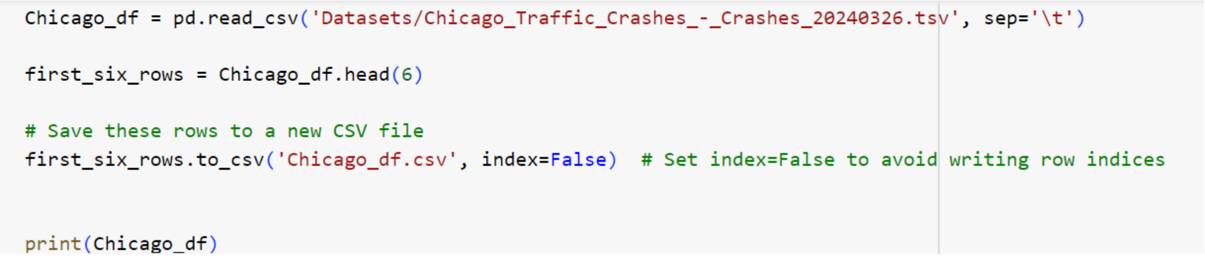


```
pip install pandas

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 Chicago_df dataframe we read the Chicago dataset and given below is the output



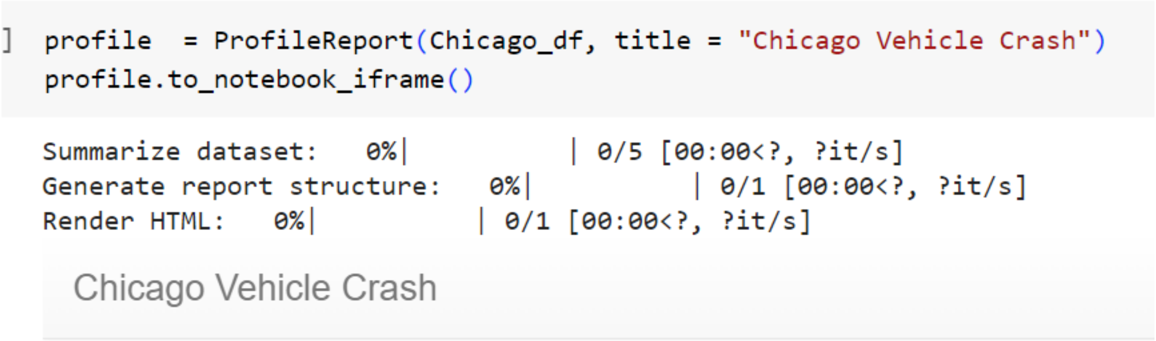
```
Chicago_df = pd.read_csv('Datasets/Chicago_Traffic_Crashes_-_Crashes_20240326.tsv', sep='\t')

first_six_rows = Chicago_df.head(6)

# Save these rows to a new CSV file
first_six_rows.to_csv('Chicago_df.csv', index=False)  # Set index=False to avoid writing row indices


print(Chicago_df)
```

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

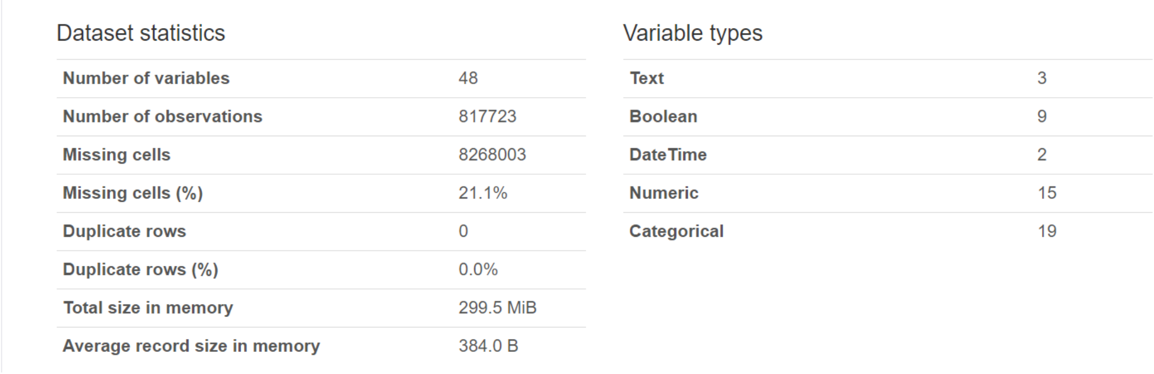


```
] profile  = ProfileReport(Chicago_df, title = "Chicago Vehicle Crash")
  profile.to_notebook_iframe()
```

```
Summarize dataset:   0%|            | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|            | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|            | 0/1 [00:00<?, ?it/s]
```

Chicago Vehicle Crash

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



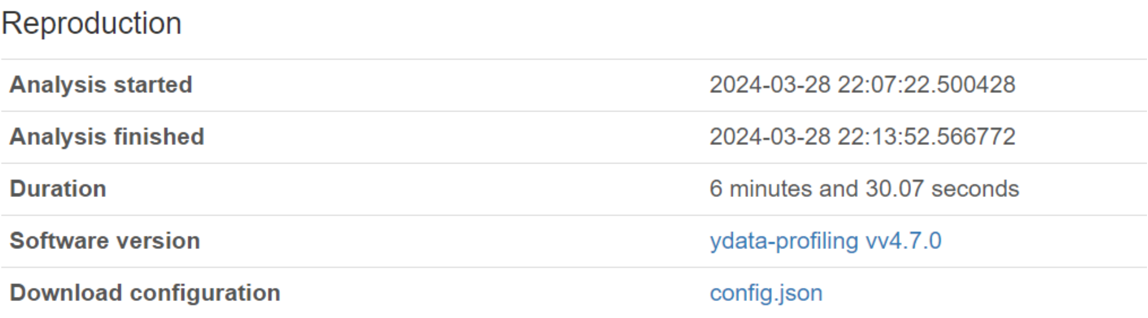
Dataset statistics

Number of variables	48
Number of observations	817723
Missing cells	8268003
Missing cells (%)	21.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	299.5 MiB
Average record size in memory	384.0 B

Variable types

Text	3
Boolean	9
DateTime	2
Numeric	15
Categorical	19

6) Complete report generation analysis:



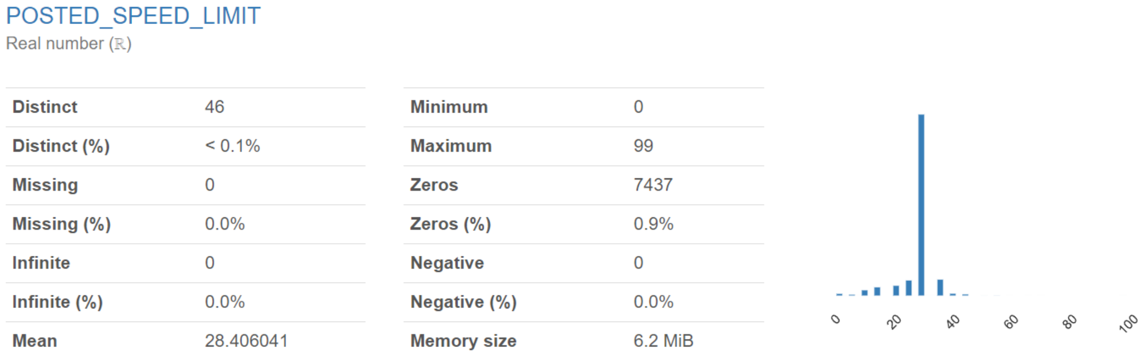
Reproduction

Analysis started	2024-03-28 22:07:22.500428
Analysis finished	2024-03-28 22:13:52.566772
Duration	6 minutes and 30.07 seconds
Software version	ydata-profiling vv4.7.0
Download configuration	config.json

➔Variable Interpretation

There are in total 48 columns in Chicago Dataset and here we have inferred about one column i.e. POSTED_SPEED_LIMIT

Interpretation of this column is given below



POSTED_SPEED_LIMIT

Real number (ℝ)

Distinct	46
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	28.406041

Minimum	0
Maximum	99
Zeros	7437
Zeros (%)	0.9%
Negative	0
Negative (%)	0.0%
Memory size	6.2 MiB

Statistical Summary:

1. Type of Variable:

• The POSTED_SPEED_LIMIT is classified as a real number (ℝ), indicating that it can take any value within the range of real numbers, which includes integers and fractions.

2. Distinct Values:

• There are 46 distinct speed limit values within the dataset, which represent less than 0.1% of the observations. This suggests a dataset with a large number of records.

3. Missing Values:

• There are no missing values for this variable, as indicated by 0 missing counts and 0% missing percentage. This suggests that the data collection for the speed limit was complete with no gaps.

4. Infinite Values:

• There are no infinite values, which confirms that all data points for this variable are valid real numbers.

5. Descriptive Statistics:

• The minimum posted speed limit recorded is 0, which could indicate areas where vehicles are not allowed to move or possibly data entry errors.

• The maximum value is 99, which seems unusually high for a posted speed limit and might indicate outliers or incorrect data.

• The mean (average) posted speed limit across all observations is approximately 28.41, which may be indicative of typical urban speed limits.

6. Zero Values:

• There are 7,437 instances of a zero value for the posted speed limit, making up about 0.9% of the data. This might require further investigation to ensure they are accurate and not data entry errors.

7. Negative Values:

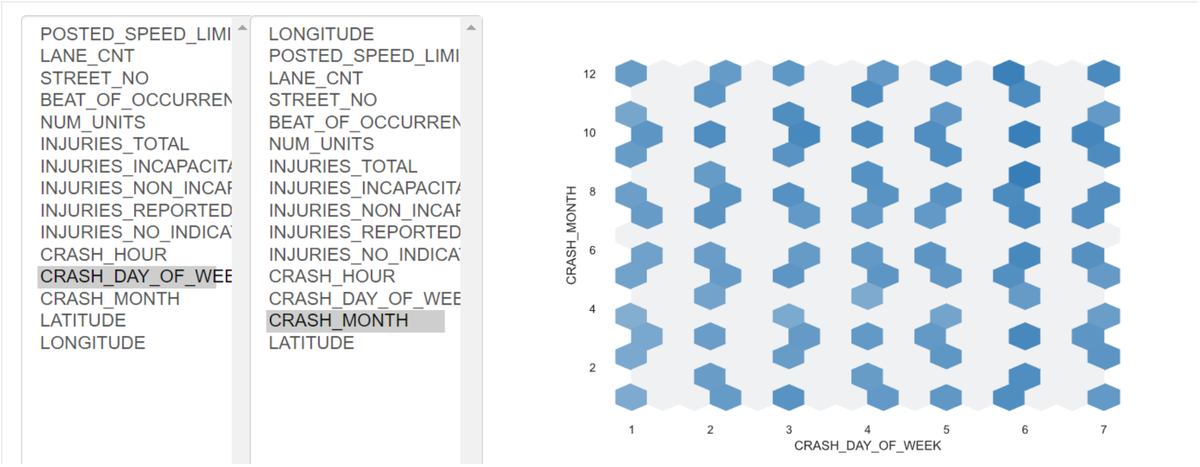
• There are no negative values, which is expected as speed limits cannot be negative.

8. Memory Size:

- The memory size occupied by this variable is 6.2 MiB, which provides information about the data footprint and may be relevant for data processing and analysis considerations.

➔ Interaction between the columns of the data:

Interactions

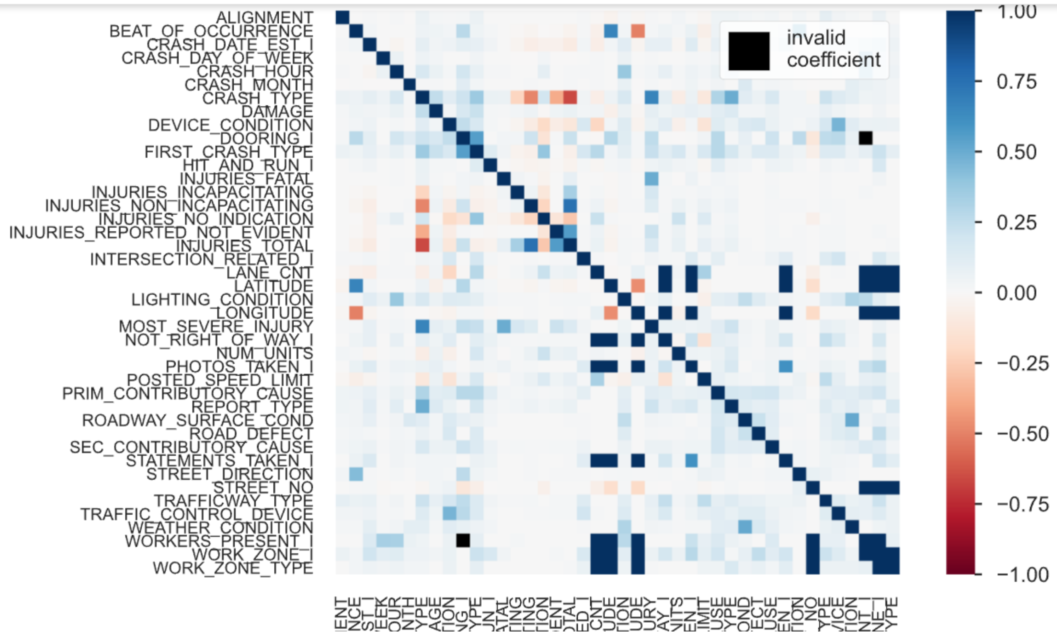


The image displays a matrix that explores the relationship between CRASH_MONTH and CRASH_DAY_OF_WEEK.

In the context of this report, an interaction indicates how two factors, such as the month of the year and the day of the week, may combine to affect the frequency or characteristics of traffic crashes.

From the visualization, it appears that the data points are uniformly distributed across the matrix, which suggests that there is no immediate, visually evident pattern linking the day of the week with the month in terms of crash frequency or severity. Each hexagon represents the intersection of these two factors and their count of incidents.

➔Analysis of Heatmap of the Correlations:



In the heatmap, the black squares labeled as 'invalid coefficient' suggest that for those pairs of variables, the correlation could not be computed. This could be due to a lack of variance in one of the variables or other issues.

Analyzing the visible parts of the heatmap, we can infer several points:

1. Injuries-Related Variables:

• There seem to be strong positive correlations among the different injuries-related variables (like INJURIES_TOTAL, INJURIES_INCAPACITATING, and INJURIES_NON_INCAPACITATING). This is expected as cases with higher total injuries will naturally have higher numbers in specific injury categories.

2. Crash Time Variables:

• Variables like CRASH_HOUR, CRASH_DAY_OF_WEEK, and CRASH_MONTH may show weaker correlations with each other, which is typical since the time of a crash is often independent of these factors.

3. Environmental and Road Conditions:

• There are likely to be various correlations between environmental conditions (like WEATHER_CONDITION, LIGHTING_CONDITION) and crash outcomes (like INJURIES_TOTAL). This suggests that poor weather or lighting could have an impact on the severity or frequency of crashes.

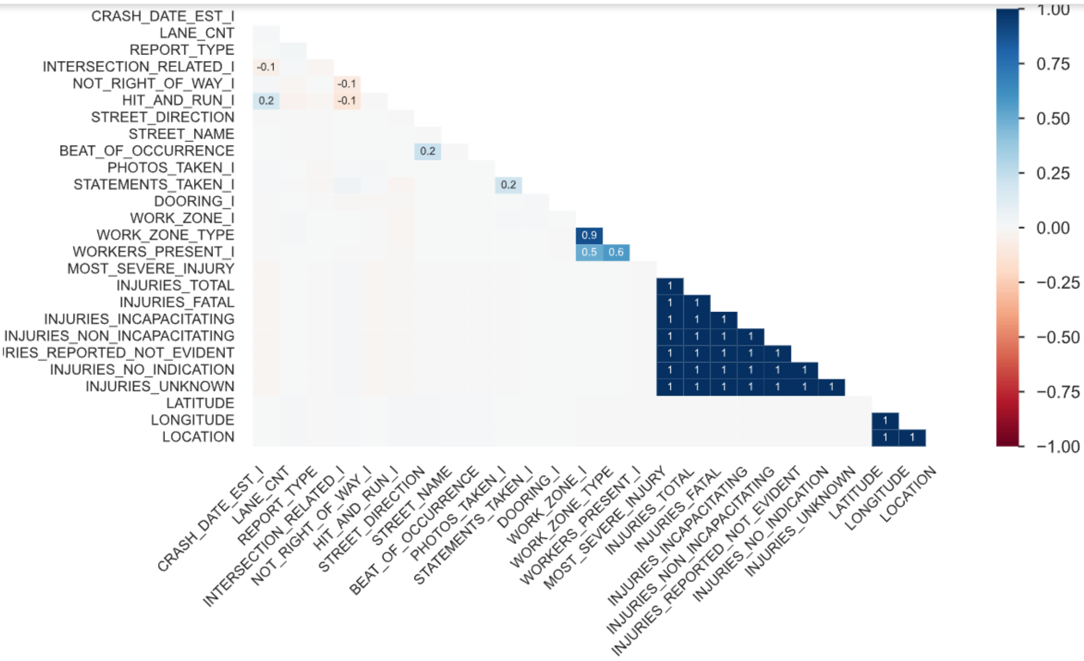
4. Contributing Factors:

• PRIM_CONTRIBUTORY_CAUSE and SEC_CONTRIBUTORY_CAUSE are likely correlated, indicating that when a primary cause is identified, there is often a secondary cause noted as well.

5. Posted Speed Limit:

• The variable POSTED_SPEED_LIMIT may have correlations with injury variables, which would be useful to explore. Higher speed limits could potentially correlate with more severe crashes.

➔Analysis for the Missing Values:



This heatmap that seems to illustrate the patterns of missing data across various variables in a traffic incidents dataset. Heatmaps of missing data can help identify if the absence of data in one variable is related to the absence in another, which can be crucial for understanding data quality and for planning data cleaning strategies.

From the heatmap, we can observe the following:

1. Strong Positive Correlations (Dark Blue Cells):

- A series of 1's along a diagonal line where a variable intersects with itself, which is expected since a variable will always perfectly correlate with missing values in itself.
- A block of cells with values of 0.9, 0.6, and 0.5 indicates that there are strong positive correlations between missing values in several variables related to injuries. This suggests that when one injury-related variable has missing data, it's quite likely that the others do

too. For example, if INJURIES_TOTAL is missing, INJURIES_FATAL and INJURIES_INCAPACITATING are also likely to be missing.

2. Weak Correlations (Light Blue and White Cells):

• Several variables show weak correlations with each other regarding missing data. These pale cells suggest that the presence (or absence) of data in one variable doesn't strongly imply the same in another. For example, the missingness of STREET_DIRECTION doesn't seem strongly linked with the missingness of INJURIES_TOTAL.

3. Negative Correlations (Light Red Cells):

• There are few light red cells indicating a slight negative correlation in missing data between some variables. This pattern implies that if data is missing in one variable, it is less likely to be missing in the other. However, the correlations are weak, as indicated by the light color.

Some general Interpretation for the heatmap of missing values-

1) There are distinct blocks of missing data correlations, particularly among variables related to injuries. This suggests that when one injury-related variable is missing, others are likely missing too, indicating a pattern or systemic issue in the data collection process for these variables.

2) Outside of these injury-related variables, there doesn't seem to be a strong pattern of missing data. This might indicate that for the majority of variables, missing data is random or non-systemic.

3) The presence of light red and light blue cells in the heatmap indicates both weak positive and weak negative correlations between the missingness of different variables. However, these correlations are not strong, suggesting that for many variable pairs, the presence or absence of data in one does not strongly predict the presence or absence of data in the other.

4) Some variables appear to be complete with no missing data (as indicated by the absence of dark blue or red colors correlating with other variables), which is a positive sign for the overall data quality.

5) There are some invalid coefficients noted, which may indicate variables with no variance (a single value throughout) or other issues that prevent calculation of a correlation coefficient