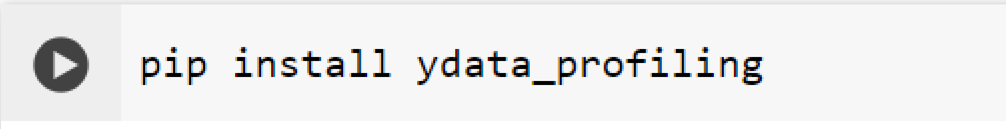
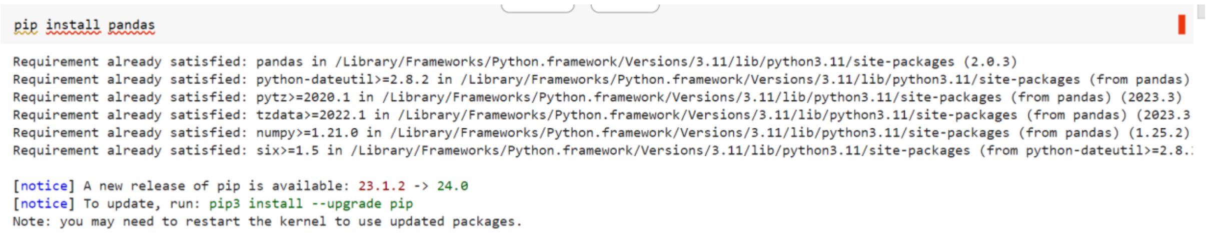
Austin Analysis Report

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



```
pip install ydata_profiling
```

2) Installed Pandas Library

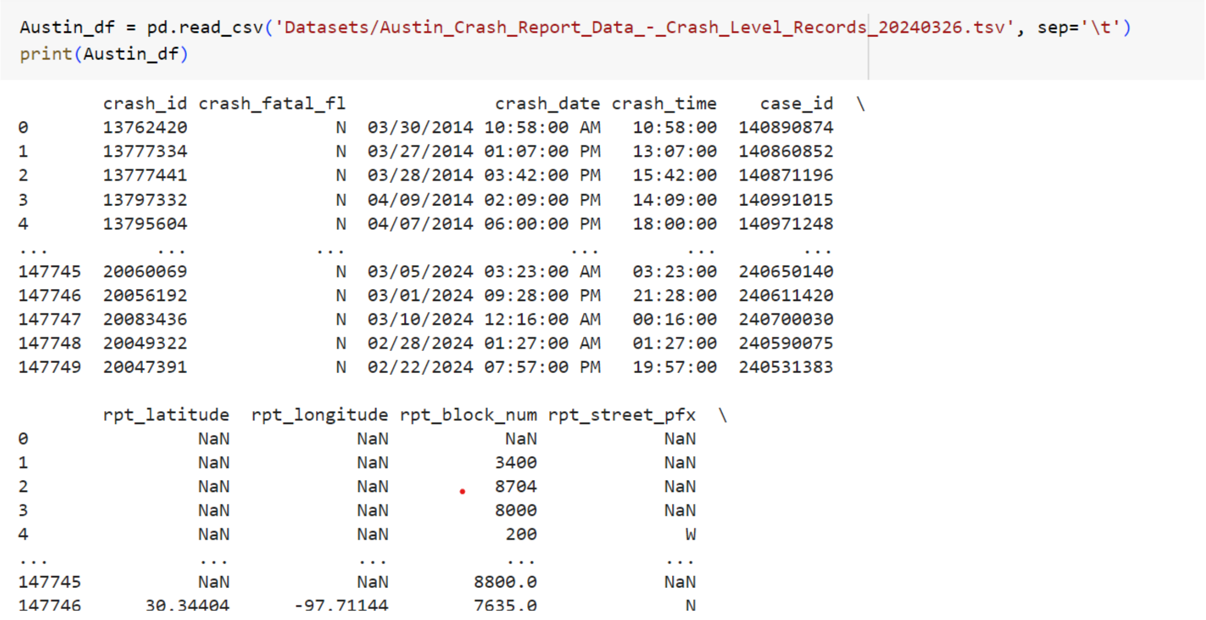


```
pip install pandas

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 Austin_df dataframe we read the Austin dataset and given below is the output

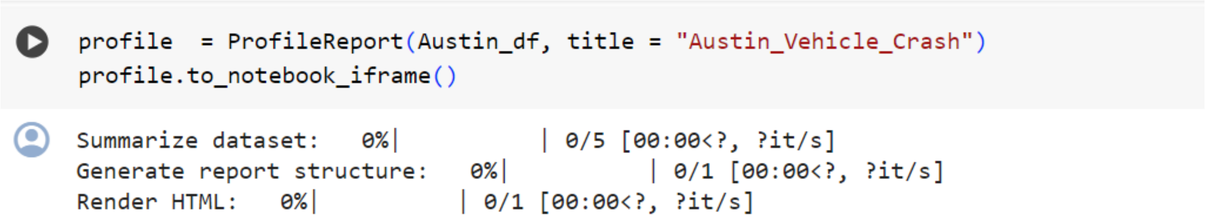


```
Austin_df = pd.read_csv('Datasets/Austin_Crash_Report_Data_-_Crash_Level_Records_20240326.tsv', sep='\t')
print(Austin_df)
```

```
        crash_id crash_fatal_fl              crash_date crash_time    case_id  \
0       13762420              N  03/30/2014 10:58:00 AM   10:58:00  140890874
1       13777334              N  03/27/2014 01:07:00 PM   13:07:00  140860852
2       13777441              N  03/28/2014 03:42:00 PM   15:42:00  140871196
3       13797332              N  04/09/2014 02:09:00 PM   14:09:00  140991015
4       13795604              N  04/07/2014 06:00:00 PM   18:00:00  140971248
...          ...            ...                     ...        ...        ...
147745  20060069              N  03/05/2024 03:23:00 AM   03:23:00  240650140
147746  20056192              N  03/01/2024 09:28:00 PM   21:28:00  240611420
147747  20083436              N  03/10/2024 12:16:00 AM   00:16:00  240700030
147748  20049322              N  02/28/2024 01:27:00 AM   01:27:00  240590075
147749  20047391              N  02/22/2024 07:57:00 PM   19:57:00  240531383

        rpt_latitude  rpt_longitude rpt_block_num rpt_street_pfx  \
0                NaN            NaN           NaN            NaN
1                NaN            NaN          3400            NaN
2                NaN            NaN          8704            NaN
3                NaN            NaN          8000            NaN
4                NaN            NaN           200              W
...              ...            ...           ...            ...
147745           NaN            NaN        8800.0            NaN
147746      30.34404      -97.71144        7635.0              N
```

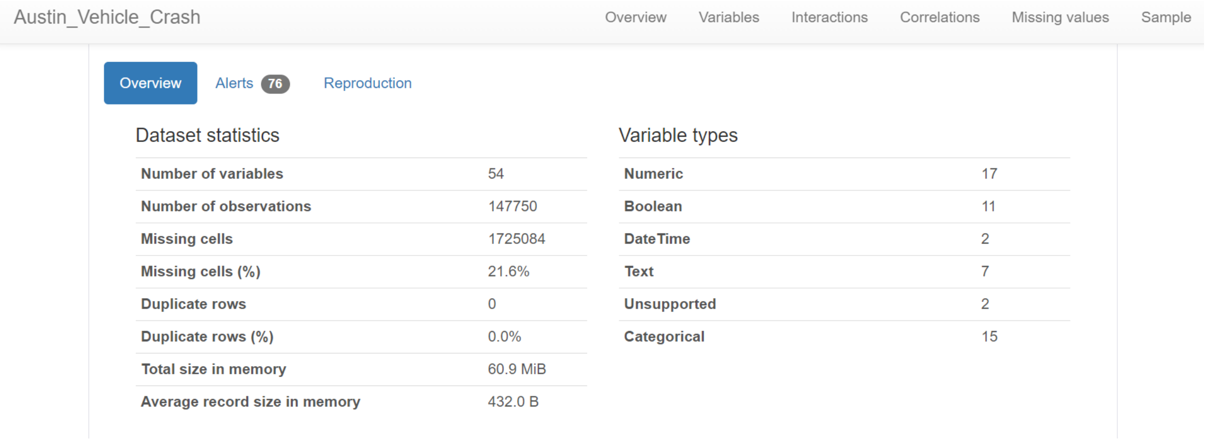
4) Here we got the complete analysis report of the dataset by executing this code



```
profile  = ProfileReport(Austin_df, title = "Austin_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]
```

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



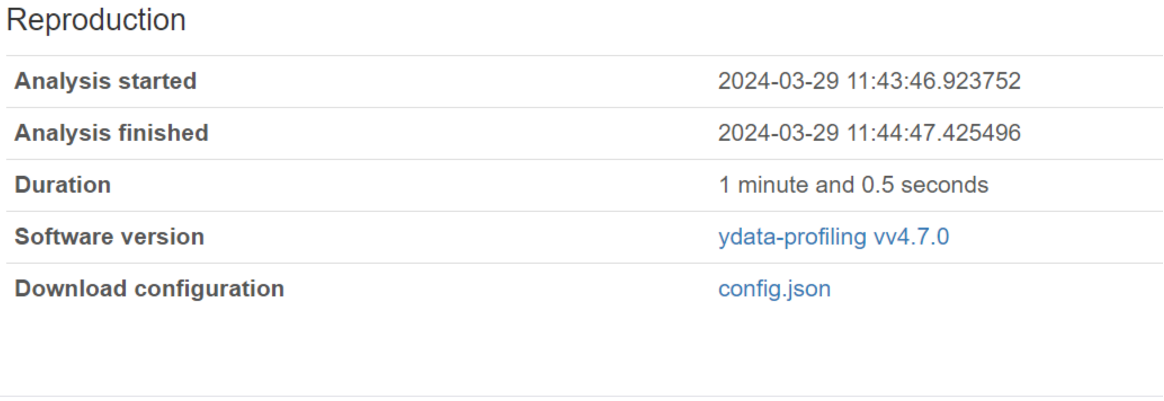
Dataset statistics

Number of variables	54
Number of observations	147750
Missing cells	1725084
Missing cells (%)	21.6%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	60.9 MiB
Average record size in memory	432.0 B

Variable types

Numeric	17
Boolean	11
DateTime	2
Text	7
Unsupported	2
Categorical	15

6) Complete report generation analysis:



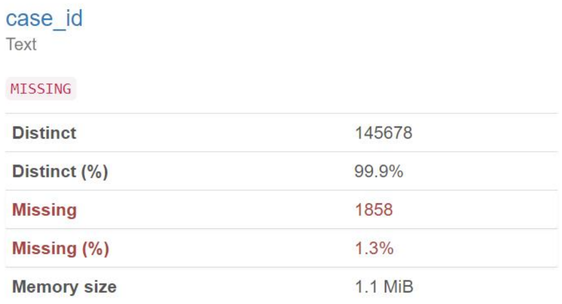
Reproduction

Analysis started	2024-03-29 11:43:46.923752
Analysis finished	2024-03-29 11:44:47.425496
Duration	1 minute and 0.5 seconds
Software version	ydata-profiling vv4.7.0
Download configuration	config.json

➔Variable Interpretation

There are in total 54 columns in Austin Dataset and here we have inferred about one column i.e. case_id

Interpretation of this column is given below



case_id

Text

MISSING

Distinct	145678
Distinct (%)	99.9%
Missing	1858
Missing (%)	1.3%
Memory size	1.1 MiB

tx
go
23
2023
190870090
152200267
212680115
191650917
180771233
2019217897
202010169
180722047
173650232
152471142
2002-0254
161290062
161411162
23-945608
c315-0226-005
182561328
172880772
2100105605
180130881

1. Column Name: case_id

• This column is of the type 'Text', which means it contains non-numeric values, likely alphanumeric strings that serve as identifiers for cases.

2. Data Quality Metrics:

• Distinct: There are 145,678 unique values in the case_id column, which make up 99.9% of the entries. This high percentage of distinct values suggests that almost every entry in the case_id column is unique, as one would expect from a field used to uniquely identify records.

• Missing: There are 1,858 missing entries, which constitute 1.3% of the data. This indicates that a small fraction of the records do not have a case_id assigned, which could be an area of concern depending on the importance of this field for analysis or reporting purposes.

• Memory Size: The case_id column occupies 1.1 MiB (Mebibytes) in memory. This information is useful for understanding the data storage and memory requirements, which can be important when processing large datasets.

3. Visualization - Word Cloud:

• On the right side of the report, there's a word cloud visualization that represents the frequency of the case_id values in the dataset. In a word cloud, the size of the number indicates how frequently that particular value occurs. Larger numbers occur more frequently, and smaller numbers occur less frequently.

• The word cloud highlights certain case_id values that are more prominent than others. For example, we see numbers like "2020", "2021", and "2019" in larger fonts, which may indicate these are common prefixes or years associated with the case IDs, and possibly occur multiple times.

• The color-coding and specific numbers mentioned in the word cloud don't have a clear meaning without additional context but generally, the variety of colors could represent different subsets or categories within the data.

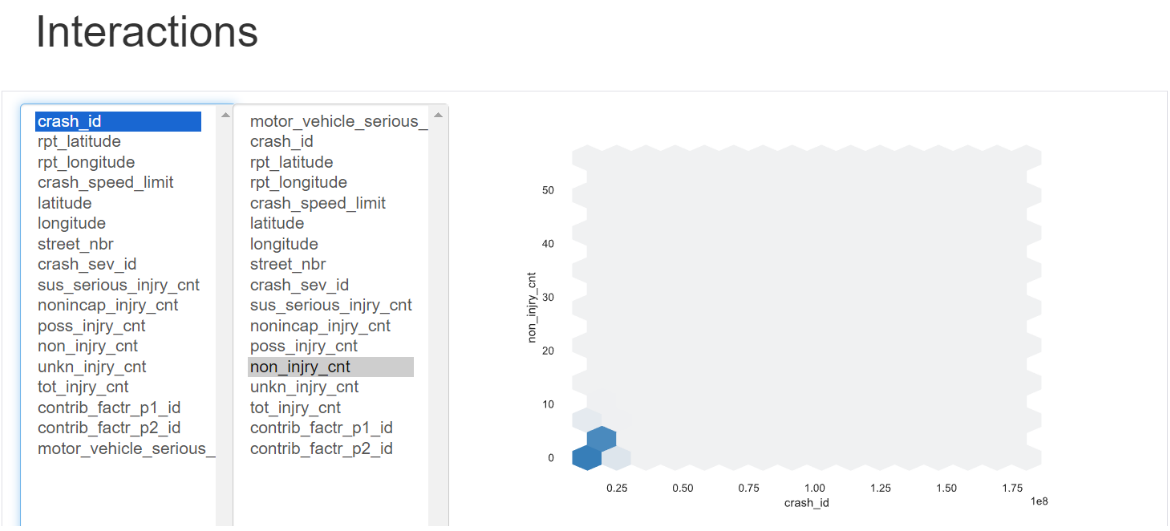
Also, The report indicates that the case_id column is primarily composed of unique identifiers for cases, with a very high uniqueness ratio and a small percentage of missing values. The word cloud visualizes the distribution of these identifiers, with some numbers appearing more frequently than others.

To improve data quality:

• Missing Values: Attention may be needed to address the 1.3% missing case_id values. Depending on the significance of the case_id in the dataset, it may be important to either fill in these missing values or understand why they are absent.

- Data Cleaning: If the case_id values should all be unique (for example, in a situation where each case must have a unique identifier), the fact that 0.1% of the values are not distinct could require investigation. It could be due to duplicates or data entry errors.

➔ Interaction between the columns of the data:



1. Variables List:

- On the left, there is a column of variable names, starting with crash_id, followed by various other attributes like rpt_latitude, rpt_longitude, crash_speed_limit, etc.

- These attributes may represent different data collected about each crash, such as the report's geographic location, the speed limit where the crash occurred, and the number of various types of injuries associated with the crash.

2. Interaction Analysis:

- The central part of the image is a list, presumably a dropdown menu, with the selected variable motor_vehicle_serious_.

- Next to it, there is a partially visible list that mirrors the variables on the left. This suggests that the report is examining interactions between motor_vehicle_serious_ and other variables in the dataset.

3. Graph - Interaction Plot:

• The graph on the right appears to be an interaction plot, although the exact type of graph is not fully visible due to the image cropping. It appears to plot non_injry_cnt against crash_id, which may be an index or unique identifier for each crash.

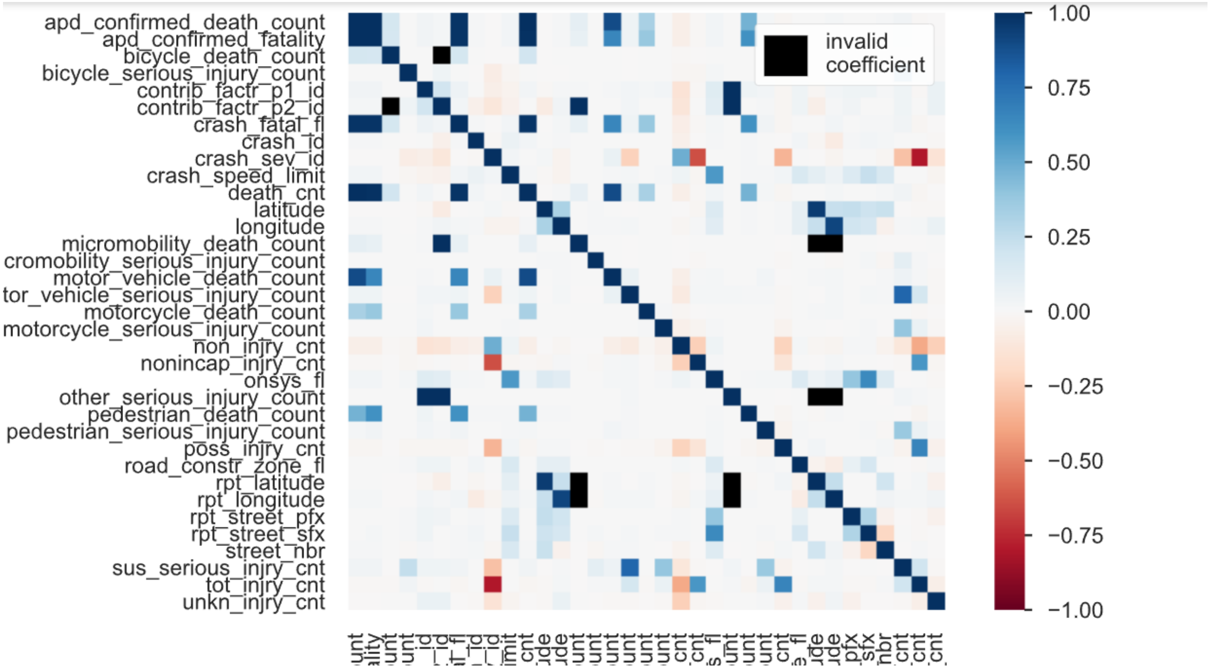
• There are data points represented as hexagons on the graph, which likely indicate the frequency or count of crashes at different levels of non_injry_cnt.

• The axis labeled non_injry_cnt suggests the graph is showing the distribution or count of non-injury incidents. The other axis is labeled crash_id, but given that IDs are unique, it might actually be representing another variable, such as the count of crashes or another metric related to crash_id.

• The crash_id axis scale is large, going up to 1e8 (100 million), which might indicate that this is not a simple count but could be a transformed or indexed value.

Summary: The report is analyzing the data regarding vehicle crashes, focusing on the interactions between a variable related to serious motor vehicle involvement and other factors like injury counts and location data. The interaction plot shows the distribution of non-injury counts against a large-scale variable related to crash_id. The actual insights from the interaction plot would require full visibility of the graph and understanding the data points represented.

➔ Analysis of Heatmap of the Correlations:



For this heatmap:

- Dark blue indicates a strong negative correlation (close to -1).
- Dark red indicates a strong positive correlation (close to 1).
- Lighter colors or white indicate no or a very weak correlation (close to 0).

Invalid Coefficients: The black squares, labeled as 'invalid coefficient', suggest that for those pairs of variables, the correlation couldn't be calculated. This might happen if one or both of the variables have a constant value or insufficient variation across observations, or if there is missing data.

Analysis of Specific Correlations: From the portion of the heatmap that is visible, we can infer the following:

1. non_injry_cnt (Non-Injury Count) and nonincap_injry_cnt (Non-incapacitating Injury Count):

• There seems to be a weak negative correlation between these two variables, indicated by a light blue color. This could imply that in incidents where there are non-incapacitating injuries, the number of non-injuries is slightly less frequent. However, the correlation is not strong, so we shouldn't infer a definitive inverse relationship.

2. motorcycle_death_count and motorcycle_serious_injury_count:

• There is a strong positive correlation between these variables, as expected. Incidents involving motorcycles that result in death are likely to also involve serious injuries.

3. bicycle_death_count and bicycle_serious_injury_count:

• A strong positive correlation is seen here as well, suggesting that fatal bicycle crashes are often associated with serious injuries.

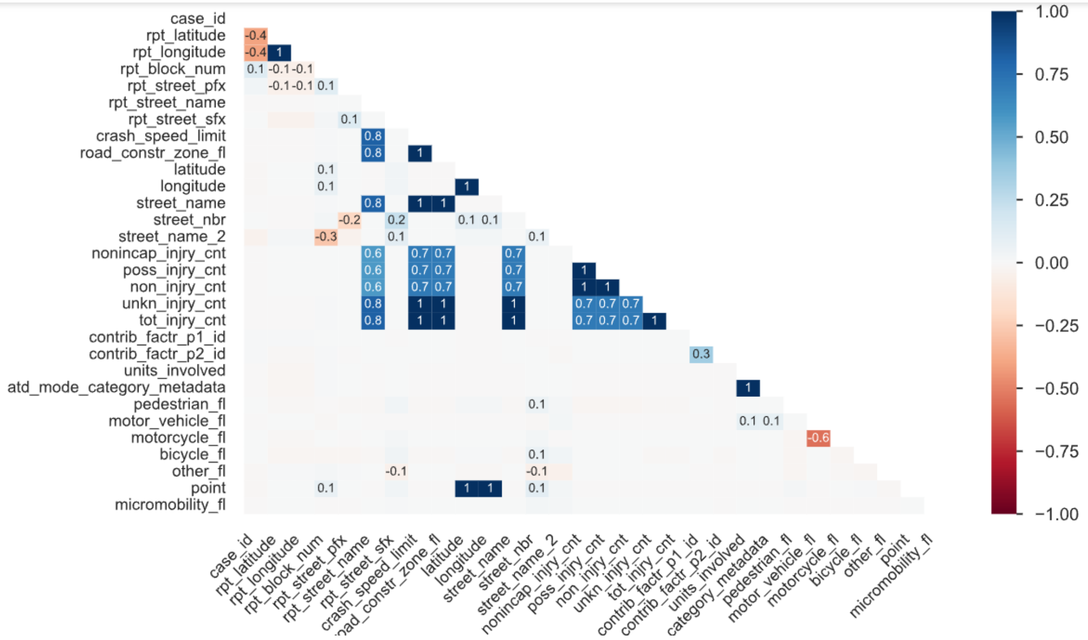
4. pedestrian_death_count and pedestrian_serious_injury_count:

• Another strong positive correlation, indicating that incidents resulting in pedestrian fatalities also commonly involve serious injuries to pedestrians.

5. apd_confirmed_death_count and apd_confirmed_fatality:

- This shows a perfect or nearly perfect positive correlation, represented by dark red squares. This suggests redundancy or that these two columns possibly represent the same underlying data regarding confirmed deaths by the Austin Police Department (APD), based on typical naming conventions in datasets.

➔Analysis for the Missing Values:



1. High Positive Correlation (Dark Red Squares):

- Several injury-related variables such as nonincap_injry_cnt, poss_injry_cnt, non_injry_cnt, unkn_injry_cnt, and tot_injry_cnt all show a high positive correlation (0.6 to 0.8 and some 1s), suggesting that if one of these has a missing value, the others are likely to have missing values as well. This could indicate a systematic issue in how injury data is collected or recorded.

2. High Negative Correlation (Dark Blue Squares):

- rpt_latitude and rpt_longitude have a strong negative correlation with case_id (-0.4). This could suggest that when there is a case_id present, there is less likely to be missing geographic information and vice versa.

3. No or Weak Correlation (Light Colored Squares):

- The atd_mode_category_metadata appears to have little to no correlation with most other variables, as indicated by the light-colored cells around it. This might mean that the presence of missing data in this column occurs independently of other variables.

4. Interesting Observations:

- The units_involved variable shows a notable positive correlation of 0.3 with tot_injry_cnt, which could imply that in records where the total injury count is not recorded, the number of units involved is also more likely to be missing.
- The variable pedestrian_fl shows a negative correlation of -0.1 with contrib_factr_p1_id and contrib_factr_p2_id. This suggests that when pedestrian involvement flag data is missing, there might be more complete data on contributing factors and vice versa.