# **Austin Crash Data Profiling Report**

The profiling report of Austin's traffic data provides a detailed look into various aspects of traffic incidents, including crash severity, locations, and times. This dataset encompasses a wide array of variables (54 in total) across 147,750 observations, shedding light on the city's traffic dynamics. Here's a comprehensive overview based on the provided data characteristics

**Dataset overview:**  
1. Number of Variables and Observations: The dataset consists of 54 variables and 147,750 observations. This breadth suggests a comprehensive data collection effort, capturing a wide range of information related to traffic incidents.

2. Missing Cells: There are 1,725,084 missing cells, accounting for 21.6% of the dataset. This high proportion of missing data indicates that a significant amount of information is absent across the dataset.

3. Duplicate Rows: The dataset contains no duplicate rows (0.0%), which is crucial for ensuring the uniqueness of each traffic incident report. This ensures that each observation in the dataset represents a distinct event, eliminating concerns over data redundancy that could skew analysis results.

4. Total Size in Memory: The total size of the dataset is 60.9 MiB, with an average record size of 432.0 bytes

5. Variable Types: The breakdown of variable types is as follows: 17 Numeric, 11 Boolean, 2 DateTime, 7 Text, 2 Unsupported, and 15 Categorical. This diversity in variable types indicates a rich dataset that includes quantitative data (Numeric), binary data (Boolean), temporal data (DateTime), descriptive data (Text and Categorical), and some Unsupported types that may require special handling or conversion to be usable

**Following are the variables in the Austin dataset with their observations:**

1.Crash id -  
Crash Id contains 100% distinct values which are 147750   
This variable contains 0 missing values

2. crash fatal fl-  
It’s a Boolean variable which has high correlation with apd confirmed death count, apd confirmed fatality, death count, motor vehicle death count and pedestrian death count.  
There are 2 distinct values - true and false with 146884 false values and 866 true values  
There are no missing values present

3. Crash date-  
There are 144667 distinct values which is 97.9%   
There are 0 missing values on the crash date variable

4. Crash time-  
This variable has 1440 distinct values   
There are 0 missing values for crash time

5. case id-  
There are 145678 unique values for case id variable  
The missing values constitute 1.3%5 and have a count of 1858

6. rpt latitude-  
rpt latitude has a high correlation with latitude which has 7976 distinct values   
The no o missing values present is 137456 which is 93%

7. rpt longitude-  
rpt longitude has a high correlation with longitude which has 7264 distinct values which is 70%  
The no of missing values present is 137456 which is 93%

8. rpt street pfx  
It is a categorical variable with 8 distinct values   
This variable contains 67805 missing values which is a percentage of 45.9

9. rpt street name  
This variable has 9794 distinct values   
The no of missing values in this variable are 3

10. rpt street sfx  
This is a categorical variable with 18 unique records in the data   
It has a high correlation with onsys fl  
The no of missing values reported are 50760

11. crash speed limit-  
The crash speed limit variable has 28 distinct values  
The number of missing values are 2  
A high correlation is reported with onsys fl

12. road constr zone fl  
It is a Boolean variable with two distinct values true and false  
the count for false values is 139901  
the count for true values is 7847

13. Latitude  
It has a high correlation with rpt latitude variable  
There are 96357 distinct values which is 66.2%  
Missing values reported are 2243

14. Longitude   
It has a high correlation with rpt longitude variable  
There are 96230 distinct values which is 66.1%  
Missing values reported are 2243

15. Street name  
Street name variable has 4630 unique records which is 3.1%  
The number of missing values are 2

16. street nbr  
This variable has 9826 unique values present  
There are 87038 missing values which is 58.9%

17. Street name 2  
Street name 2 has 3396 distinct values present   
There are 81494 missing records present for this data which is 55.1%

18. crash sev id  
Crash sev id has high correlation with nonincap injurt cnt and total injury cnt  
There are 8 unique records for crash sev id and 0 missing values

19. sus serious injury cnt  
This variable has high correlation with motor vehicle injury count  
It has 7 distinct and 0 missing values reported

20.non cap injury count  
This variable has high correlation with crash sev id and total injury count  
The number of distinct records are 14  
Missing values reported are 1 on the dataset

21. poss injury cnt  
This variable has 18 distinct values and 1 missing value reported on the dataset  
It has high correlation with total injury cnt

22. non injury cnt-  
Non injury cnt has 46 unique records and 1 missing value

23. unknwn injury cnt-   
This variable has 16 distinct values   
The most occurred value out of the distinct values is 0 and has a count of 133316  
The missing values on the variable are 2

24. tot inury cnt-  
The tot injury cnt variable has 18 distinct values and 2 missing values

25. Death count-  
Death count has high correlation with apd confirmed death count , apd confirmed fatality, crash fata fl, motor vehicle death count  
The no of missing values in this are 0 and it has 5 unique values

26. contrib\_factor\_p1\_id-  
It has 70 distinct values present  
The number of missing values are 119143 which is 80.6%

27. contrib\_factor\_p2\_id-  
It has 65 distinct values present  
The number of missing values are 143235 which is 96.9%

28. units involved  
The no of unique values present in this variable are 1112  
Missing values present in this variable are 7 which is less than 0.1%

29. atd\_mode\_category\_metadata   
There are 100% distinct values present in this variable that is 147743  
Missing values present in this variable are 7 which is less than 0.1%

30. pedestrian fl  
This variable contains only one distinct value that is true which has a count of 3505  
There are 144245 missing values present which is 97.6% of missing values

31. motor vehicle fl  
This variable contains only one distinct value that is true which has a count of 146634  
There are 1116 missing values present which is 0.8% of missing values

32. motorcycle fl  
This variable contains only one distinct value that is true which has a count of 3602  
There are 144148 missing values present which is 97.6% of missing values

33. bicycle fl  
This variable contains only one distinct value that is true which has a count of 2444  
There are 145306 missing values present which is 98.3% of missing values

34. other fl  
This variable contains only one distinct value that is true which has a count of 4845  
There are 1429055 missing values present which is 96.7% of missing values

35.point  
Point variable has 97739 unique values present  
The number of missing values present are 2243 which is 1.5%

36. apd\_confirmed\_fatality  
This variable contains 2 distinct value that is True and False which has a count of 842 and 146908  
There are 0 missing values present in this variable

37. apd\_confirmed\_death\_count  
This is a categorical variable with 5 unique values (0,1,2,3,4)  
It has 0 missing values

38. motor\_vehicle\_death\_count  
This is a categorical variable with 5 unique values (0,1,2,3,4)  
It has 0 missing values

39. motor\_vehicle\_serious\_injury\_cnt  
This variable has 6 unique values as shown   
The number of missing values present are 0

40.bicycle\_death\_count  
This is a categorical variable with two distinct values (0,1) with a count of 147721 and 29 respectively  
It has 0 missing values present

41.bicycle\_serious\_injury\_count  
This is a categorical variable with 4 distinct values (0,1,2,3) with a count of 147487, 260, 2,1 respectively  
It has 0 missing values present

42. pedestrian\_death\_count  
This is a categorical variable with 3 distinct values (0,1,2) with a count of 147431, 317, 2 respectively  
It has 0 missing values present

43. pedestrian\_serious\_injury\_count  
This is a categorical variable with 5 distinct values (0,1,2,3,9) with a count of 147131, 606, 11,1,1 respectively  
It has 0 missing values present

44.motorcycle\_death\_count  
This is a categorical variable with 5 distinct values (0,1,2) with a count of 146137, 121,2 respectively  
It has 0 missing values present

45. motorcycle\_serious\_injury\_count  
This is a categorical variable with 3 distinct values (0,1,2) with a count of 147064, 664, 24 respectively  
It has 0 missing values present

46. other death count  
It is a categorical variable with 1 distinct value which has a count of 147750  
There are no missing values present

47. other\_serious\_injury\_count  
This is a categorical variable with 3 distinct values (0,1,3) with a count of 147746,3,1 respectively  
It has 0 missing values present

48. onsys\_fl  
This variable has two unique values true and false with count of 73041 and 74709 respectively  
It has 0 missing values

49. private\_dr\_fl  
This variable has 1 distinct value False with a count 147750  
It has 0 missing values present

50. micromobility\_serious\_injury\_count  
This is a categorical variable with 3 distinct values (0,1,2) with a count of 147708,40,12 respectively  
It has 0 missing values present

51. micromobility\_death\_count  
This is a categorical variable with 3 distinct values (0,1) with a count of 147744 ,6 respectively  
It has 0 missing values present

52. micromobility\_fl  
This variable has only one distinct value True with a count of 311  
The missing values reported in the variable are 147439 which is 99.8%

**Overall insights:**

This profiling report highlights the dataset's rich potential for analyzing traffic incidents in Austin, offering insights into when, where, and how crashes occur. However, the significant amount of missing data, especially regarding location and contributing factors, presents a challenge that needs to be addressed to unlock the full potential of the dataset for comprehensive analysis