PROFILING DOCUMENTATION FOR AUSTIN DATASET

The Austin dataset comprises various data points that are critical to answering the business requirements. Following are the key observations about the dataset:

1. Number of fields: The dataset consists of 54 distinct fields.
2. Number of Observations: There are a total of 147.7k (147750) records, where each record corresponds to a crash-level record.
3. Missing Cells: A count of 1.7M (1725084) cells within the dataset have missing data, which constitutes approximately 21.6% of the total data. This is relatively high and can impact analysis.
4. Duplicate Rows: There are no duplicate rows, as indicated by a count of 0, which means that each observation is unique.
5. Total Size in Memory: The dataset occupies 60.9 MiB (Megabytes) in memory, providing an idea of the data's footprint in terms of storage.
6. Average Record Size in Memory: Each record, on average, occupies 432.0 bytes in memory, which can be useful for understanding the load and performance implications for data processing tasks.

COLUMN WISE ANALYSIS:

crash_id:

- The data field contains 147750 unique values.
- There are zero missing values for the crash_id as indicated by a count of 0.
- No. of Distinct: 147750
- Min Value: 1001
- Max Value: 180290542
- This is the unique identifier for the crashes in the Austin city.

crash_fatal_fl:

- The data field contains 2 unique values.
- There are zero missing values for the crash_fatal_fl as indicated by a count of 0.
- No. of Distinct values: 2
- As the column has a Boolean data type, it accepts Y or N as input.
- The missing values would be replaced by “NA” and hence the column would be changed to datatype varchar.
- This column identifies if the crash involved one or more fatalities.

crash_date (date and time):

- The data field contains 144667 unique values for the crash_date
- There are no missing values for the crash_date variable as indicated by a count of 0.
- Percentage of Distinct: 97.9%.
- Min Value: 2014-03-26 06:41:00
- Max Value: 2024-03-11 22:05:00
- The min and max values indicate that the crash dates range from the year 2014 to 2024.

crash_time:

- The time field contains 1440 unique values for the crash_time.
- There are no missing values for the crash_time variable as indicated by a count of 0.
- Percentage of Distinct: 1.0%.
- The datatype of this column is varchar in Talend jobs.
- Min Value: 00:00:00
- Max Value: 23:59:00
- This column indicates the time at which the crash occurred.

case_id :

- The data field contains 145678 unique values for the case_id.
- There are missing values for the case_id variable as indicated by a count of 1858.
- Percentage of Distinct: 99.9%.
- The column consists of highly inconsistent values having special characters with varied lengths for this column.
- Examples of inconsistent data having special characters (-, `): 16-16788, 1607-0030, 161701669`.
- Due to inconsistencies present in this column, we had to take the datatype of this column as varchar.
- Handled null values for this column by substituting “-1”.
- Min Value: -1.
- Max Value: TXC231423139.

- There is no specific inference given in the original dataset about what information is conveyed through this column.

rpt_latitude:

- The data field contains 7976 unique values for the rpt_latitude.
- There are missing values for the rpt_latitude variable as indicated by a count of 137456 which is almost 93% of the data missing for this column.
- Percentage of Distinct: 77.5%.
- Min Value: 0.0
- Max Value: 36.5005
- The column consists of the reported latitude of the crash.

rpt_longitude:

- The data field contains 7264 unique values for the rpt_longitude.
- There are missing values for the rpt_longitude variable as indicated by a count of 137456 which is almost 93% of the data missing for this column.
- Percentage of Distinct: 70.5%.
- Min Value: -0.0
- Max Value: -93.5079
- The column consists of the reported longitude of the crash.

Both rpt_latitude and rpt_longitude would be converted to datatype float at the cleaning stage and the missing values would be replaced by -1.

rpt_block_num:

- The data field contains 4789 unique values for the rpt_block_num.
- There are missing values for the rpt_block_num variable as indicated by a count of 19611 which is almost 13.3% of the data.
- This column was treated as having a varchar datatype due to empty strings present in it which was later removed from the data during the data cleaning process and then converted as an integer in the staging table.
- Min Value: 0
- Max Value: 108010800
- The missing values are handled by substituting -1.

- This column gives an inference of the reported block number on which the crash happened.

rpt_street_pfx:

- The data field contains 8 unique values for the rpt_street_pfx.
- There are missing values for the rpt_street_pfx variable as indicated by a count of 67805 which is almost 45.9% of the data.
- No. of Distinct: 8
- The length of this column is set as 10 having a varchar datatype.
- The missing values would be replaced by “NA”.
- This column gives an inference of on which road the crash occurred by specifying the street prefix.

rpt_street_name:

- The data field contains 9794 unique values for the rpt_street_name.
- There are missing values for the rpt_street_name variable as indicated by a count of 3 which is almost 0.1% of the data.
- No. of Distinct: 9794
  The column consists of categorical data and the length of the data in this column ranges from 0 to 50.
- The missing values would be replaced by “NA”.
- This column gives an inference of on which road the crash occurred by specifying the street name.

rpt_street_sfx:

- The data field contains 18 unique values for the rpt_street_sfx.
- There are missing values for the rpt_street_sfx variable as indicated by a count of 50340 which is almost 34.1% of the data.
- No. of Distinct: 18
- The length of this column is set as 4 having a varchar datatype.
- The missing values would be replaced by “NA”.
- This column gives an inference of on which road the crash occurred by specifying the street suffix.

crash_speed_limit:

- The data field contains 28 unique values for the crash_speed_limit.
- There are missing values for the crash_speed_limit variable as indicated by a count of 2 which is almost 0.1% of the data.
- Min Value: 0
- Max Value: 85
- The missing values are handled by substituting -1.
- This column gives an inference of the speed limit involved in the crash.

road_constr_zone_fl :

- The data field contains 2 unique values.
- There are missing values for the road_constr_zone_fl variable as indicated by a count of 2 which is almost 0.1% of the data.
- As the column has a Boolean data type, it accepts Y or N as input.
- The missing values would be replaced by "NA" and hence the column would be changed to datatype varchar.
- This column checks if the crash occurred in a construction, maintenance, or utility work zone.

latitude:

- The data field contains 96357 unique values for latitude.
- There are 2243 missing values for the latitude.
- No. of Distinct: 96357
- Minimum Value: 0
- Maximum Value: 30.5116

longitude :

- The data field contains 96230 unique values for longitude.
- There are 2243 missing values for the longitude.
- No. of Distinct: 96230
- Minimum Value : -97.9268
- Maximum Value: 0.0

Both latitude and longitude would be converted to datatype float at the cleaning stage and the missing values would be replaced by -1.

street_name:

- The data field contains 4630 unique values for street_name column.
- There are 2 missing values for the street_name column.
- No. of Distinct: 4630
- The column consists of categorical data and the length of the data in this column ranges from 2 to 50. This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The missing values are replaced by “NA”.
- This column specifies the street name on which the crash occurred.

street_nbr:

- The data field contains 9826 unique values for the street_nbr.
- There are 87038 missing values for the street_nbr variable.
- No. of Distinct: 9826
- This column was treated as having a varchar datatype due to empty strings present in it which were later removed from the data during the data cleaning process and then converted as an integer in the staging table.
- Min Value : 0
- Max Value: 21146
- The missing values are handled by substituting -1.
- This column specifies the block number of the primary street where crash occurred.

street_name_2:

- The data field contains 3396 unique values for the street_name_2 column.
- There are 81474 missing values for the street_name_2 column.
- No. of Distinct: 3396
- The column consists of categorical data and the length of the data in this column ranges from 2 to 50. This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The missing values are replaced by “NA”.
- This column specifies the secondary road’s street name on which the crash occurred.

street_nbr_2:

- This data field consists of only missing values of count 147750.
- There are no distinct values present in this column.
- This column was treated as having a varchar datatype due to empty strings present in it.
- The missing values are replaced by “NA”.
- This column specifies the block number of the secondary street where crash occurred.

crash_sev_id :

- The data field contains 8 unique values.
- There are 0 missing values for this column.
- No. of Distinct: 8
- Min Value : 0
- Max Value: 99
- This column describes the crash severity based on the types of conditions of injuries and fatalities that occurred in the crash.
- According to the dataset, values like 0 = UNKNOWN, 1 = INCAPICITATING INJURY, 2 = NON INCAPICITATING INJURY, 3 = POSSIBLE INJURY and 4 = KILLED, 5 = NOT INJURED
- But the data for this column has various other values other than the one specified such as 99 and 94.

sus_serious_injry_cnt:

- The data field contains 7 unique values.
- There are 0 missing values for this column.
- No. of Distinct: 7
- Min Value : 0
- Max Value: 10
- This column describes the Total Suspected Serious Injury Count for a particular crash.

nonincap_injry_cnt:

- The data field contains 14 unique values.
- There is 1 missing value for this column.
- No. of Distinct: 14
- The missing values are handled by substituting -1.
- Min Value: 0
- Max Value: 14
- This column describes the Total Non-incapacitating Injury Count for a particular crash.

poss_injry_cnt:

- The data field contains 16 unique values.
- There is 1 missing value for this column.
- No. of Distinct: 16
- The missing values are handled by substituting -1.
- Min Value: 0
- Max Value: 20
- This column describes the Total Possible Injury Count for a particular crash.

non_injry_cnt:

- The data field contains 46 unique values.
- There is 1 missing value for this column.
- No. of Distinct: 46
- The missing values are handled by substituting -1.
- Min Value: 0
- Max Value: 56
- This column describes the Total Not Injured Count for a particular crash.

unkn_injry_cnt:

- The data field contains 16 unique values.
- There are 2 missing values for this column.
- No. of Distinct: 14
- The missing values are handled by substituting -1.
- Min Value: 0

- Max Value: 41
- This column describes the Total Unknown Injury Count for a particular crash.

tot_injry_cnt:

- The data field contains 18 unique values.
- There are 2 missing values for this column.
- No. of Distinct: 18
- The missing values are handled by substituting -1.
- Min Value: 0
- Max Value: 21
- This column describes the Total Injury Count for a particular crash.

death_cnt:

- The data field contains 5 unique values.
- There is 0 missing value for this column.
- No. of Distinct: 5
- Min Value: 0
- Max Value: 4
- This column describes the Total Death Count for a particular crash.

contrib_factr_p1_id:

- The data field contains 70 unique values.
- There is a 119143 missing values for this column.
- No. of Distinct: 70
- Min Value: 0
- Max Value: 80
- The missing values are handled by substituting -1.
- This column describes the first factor for a given vehicle which the officer felt possibly contributed to the crash.

contrib_factr_p2_id:

- The data field contains 65 unique values.
- There is a 143235 missing values for this column.

- No. of Distinct: 65
- Min Value: 0
- Max Value: 79
- The missing values are handled by substituting -1.
- This column describes the second factor for a given vehicle which the officer felt possibly contributed to the crash.

According to the business requirement, both contrib_factr_p1_id and contrib_factr_p1_id should be combined to form a common column per crash that occurred – contributing_factor_code which will just have values that are not null and is mapped to its corresponding description – contributing_factor_descriptions from the mapping sheet.

units_involved:

- The data field contains 1112 unique values.
- There are 7 missing values for this column.
- No. of Distinct: 1112
- The column consists of categorical data and the length of the data in this column ranges from 6 to 100. This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column are ‘ / ’ ‘ - ’ ex:- other/unknown, e-scooter
- The missing values are replaced by “NA”.
- This column describes the mode of units involved in the crash.

According to the business requirement mentioned in the change request, this column should be normalized to get all the units involved in separate rows to obtain the results.
Therefore, we have added columns such as Vehicle_Code to map each vehicle/unit involved in the crash to its least granular nature to map to its Vehicle_Description from the vehicle_mapping_sheet.

atd_mode_category_metadata:

- The data field contains 1447743 unique values.
- There are 7 missing values for this column.
- No. of Distinct: 1447743

- The column consists of categorical data consisting of the metadata of each crash associated with a crash_id and the length of the data in this column ranges from 50 to 2000. This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column is ‘ {} ’ ‘ “” ’ ‘ _ ’ ‘ :’ ex:- "[{""mode_id"": 1, ""mode_desc"": ""Passenger car"", ""unit_id"": 2259431, ""death_cnt"": 0, ""sus_serious_injry_cnt"": 0, ""nonincap_injry_cnt"": 0, ""poss_injry_cnt"": 0, ""non_injry_cnt"": 4, ""unkn_injry_cnt"": 0, ""tot_injry_cnt"": 0}]"

pedestrian_fl:

- The data field contains 1 unique value.
- There are 144245 missing values.
- No. of Distinct values: 1
- As the column has a Boolean data type, it accepts Y or N as input.
- The missing values would be replaced by “NA” and hence the column would be changed to datatype varchar.
- This column identifies if the crash involved a Pedestrian.

motor_vehicle_fl:

- The data field contains 1 unique value.
- There are 1116 missing values.
- No. of Distinct values: 1
- As the column has a Boolean data type, it accepts Y or N as input.
- The missing values would be replaced by “NA” and hence the column would be changed to datatype varchar.
- This column identifies if the crash involved a Motor vehicle.

motorcycle_fl:

- The data field contains 1 unique value.
- There are 144148 missing values.
- No. of Distinct values: 1
- As the column has a Boolean data type, it accepts Y or N as input.
- The missing values would be replaced by “NA” and hence the column would be changed to datatype varchar.
- This column identifies if the crash involved a Motorcycle.

bicycle_fl:

- The data field contains 1 unique value.
- There are 145306 missing values.
- No. of Distinct values: 1
- As the column has a Boolean data type, it accepts Y or N as input.
- The missing values would be replaced by "NA" and hence the column would be changed to datatype varchar.
- This column identifies if the crash involved a Bicyclist.

other_fl:

- The data field contains 1 unique value.
- There are 142905 missing values.
- No. of Distinct values: 1
- As the column has a Boolean data type, it accepts Y or N as input.
- The missing values would be replaced by "NA" and hence the column would be changed to datatype varchar.
- This column identifies if the crash involved other factors.

point:

- The data field contains 97739 unique values.
- There are 2243 missing values for this column.
- No. of Distinct: 97739
- The column consists of categorical data and the length of the data in this column ranges from 2 to 45. This information will be helpful at the time of stagging and will help avoid value truncated errors.
- The special characters present in this column are ' (' ') ' ex:- POINT (-97.92678889 30.18993984)
- The missing values are replaced by "NA".
- This column describes the combined data of coordinates of the latitude and longitude of the crash.

apd_confirmed_fatality:

- The data field contains 2 unique values.
- There are zero missing values for the apd_confirmed_fatality as indicated by a count of 0.

- No. of Distinct values: 2
- As the column has a Boolean data type, it accepts Y or N as input.
- The missing values would be replaced by "NA" and hence the column would be changed to datatype varchar.
- This column identifies the fatalities confirmed for a particular crash.

apd_confirmed_death_count:

- The data field contains 5 unique values.
- There is 0 missing value for this column.
- No. of Distinct: 5
- Min Value: 0
- Max Value: 4
- This column describes the Total Confirmed Death Count for a particular crash.

motor_vehicle_death_count:

- The data field contains 5 unique values.
- There is 0 missing value for this column.
- No. of Distinct: 5
- Min Value: 0
- Max Value: 4
- This column describes the Total Motor Vehicle user Death Count for a particular crash.

motor_vehicle_serious_injury_count:

- The data field contains 6 unique values.
- There is 0 missing value for this column.
- No. of Distinct: 6
- Min Value: 0
- Max Value: 5
- This column describes the Total Motor Vehicle User's Serious Injury Count for a particular crash.

bicycle_death_count:

- The data field contains 2 unique values.
- There is 0 missing value for this column.
- No. of Distinct: 2
- Min Value: 0
- Max Value: 1
- This column describes the Total Bicycle user's Death Count for a particular crash.

bicycle_serious_injury_count:

- The data field contains 4 unique values.
- There is 0 missing value for this column.
- No. of Distinct: 5
- Min Value: 0
- Max Value: 3
- This column describes the Total Bicycle user's Serious Injury Count for a particular crash.

pedestrian_death_count:

- The data field contains 3 unique values.
- There is 0 missing value for this column.
- No. of Distinct: 3
- Min Value: 0
- Max Value: 2
- This column describes the Total Pedestrian Death Count for a particular crash.

pedestrian_serious_injury_count:

- The data field contains 5 unique values.
- There is 0 missing value for this column.
- No. of Distinct: 5
- Min Value: 0
- Max Value: 9
- This column describes the Total Pedestrian Serious Injury Count for a particular crash.

motorcycle_death_count:

- The data field contains 3 unique values.
- There is 0 missing value for this column.
- No. of Distinct: 3
- Min Value: 0
- Max Value: 2
- This column describes the Total Motorcycle user Death Count for a particular crash.

motorcycle_serious_injury_count:

- The data field contains 3 unique values.
- There is 0 missing value for this column.
- No. of Distinct: 3
- Min Value: 0
- Max Value: 2
- This column describes the Total Motorcycle user's Serious Injury Count for a particular crash.

other_death_count:

- The data field contains 1 unique value.
- There is 0 missing value for this column.
- No. of Distinct: 1
- Min Value: 0
- Max Value: 0
- This column describes the Total Other Death Count for a particular crash.

other_serious_injury_count:

- The data field contains 3 unique values.
- There is 0 missing value for this column.
- No. of Distinct: 3
- Min Value: 0
- Max Value: 3

- This column describes the Total Serious Injury Count for a particular crash.

onsys_fl:

- The data field contains 2 unique values.
- There are zero missing values.
- No. of Distinct values: 2
- As the column has a Boolean data type, it accepts Y or N as input.
- This column identifies whether the primary road of the crash was on a highway.

private_dr_fl:

- The data field contains 1 unique value.
- There are zero missing values.
- No. of Distinct values: 2
- As the column has a Boolean data type, it accepts Y or N as input.
- All values described in this column are of N input that is false for the given case.
- This column identifies whether the crash occurred on a private drive / private property or parking lot.

micromobility_serious_injury_count:

- The data field contains 3 unique values.
- There is 0 missing value for this column.
- No. of Distinct: 3
- Min Value: 0
- Max Value: 2
- This column describes the Total Micro mobility user's Serious Injury Count for a particular crash.

micromobility_death_count:

- The data field contains 2 unique values.
- There is 0 missing value for this column.
- No. of Distinct: 2
- Min Value: 0
- Max Value: 1

- This column describes the Total Micro mobility user's Death Count for a particular crash.

micromobility _fl:

- The data field contains 2 unique values.
- There are 147439 missing values.
- No. of Distinct values: 2
- As the column has a Boolean data type, it accepts Y or N as input.
- The missing values would be replaced by "NA" and hence the column would be changed to datatype varchar.
- This column identifies whether the micro-mobility device was involved in the crash.