**Analysis Document New York**

**A screenshot of a computer

Description automatically generated**

**Dataset Composition**: The dataset contains 29 variables across 2,075,427 observations. There is a significant amount of missing data, approximately 29.5%, with no duplicate rows present.

**Variable Types**: There are a mixture of DateTime, Categorical, Text, and Numeric variables in the dataset.

**Correlations and Imbalances**:

* 1. Certain variables such as the number of persons injured and killed show high correlations with corresponding motorist and pedestrian figures, indicating that most injuries or fatalities in crashes involve these groups.
  2. There are significant imbalances in the data, particularly in variables recording the number of pedestrians and cyclists killed, and in the contributing factors for vehicles 4 and 5, with most records being zeros or missing.

**Missing Data**:

* 1. Geographic information such as borough, zip code, latitude, and longitude have a high percentage of missing values.
  2. Details about the streets (on street name, cross street name, off street name) also have substantial missing data.
  3. Factors contributing to crashes and vehicle types for the second, third, fourth, and fifth vehicles involved have missing values up to 99.6%.

**Skewness**:

* 1. The variables indicating the number of persons killed and the number of motorist killed are highly skewed, meaning most of the data points are zeros, indicating crashes without fatalities are far more common.

**Uniqueness**:

* 1. The Collision ID variable is unique across all observations, which could indicate that it serves as a primary key in the dataset.

**Samples and Time Frame**:

* 1. The crashes recorded span a specific time frame with time stamps indicating the precise time of each crash. The data appears to be current up to 2024, which matches the report's generation date.
  2. The sample records indicate variability in the completeness of data captured for each incident.

**General Observations**:

* 1. Most crashes did not result in fatalities.
  2. Geographic data is incomplete, which may affect spatial analysis.
  3. Detailed crash factors and vehicle types are often not recorded, especially for vehicles beyond the first one involved in the crash.

**Dataset New York**

Overview of the dataset

A screenshot of a graph

Description automatically generated

**New York Crash Report:**

**Data Range:** The date range for the crashes extends from July 1, 2012, to March 22, 2024.

**Unique Dates:** There are 4,283 distinct crash dates, which is only 0.2% of the total entries, indicating a high number of crashes on the same dates.

**Data Completeness:** The CRASH\_DATE column is complete, with no missing entries.

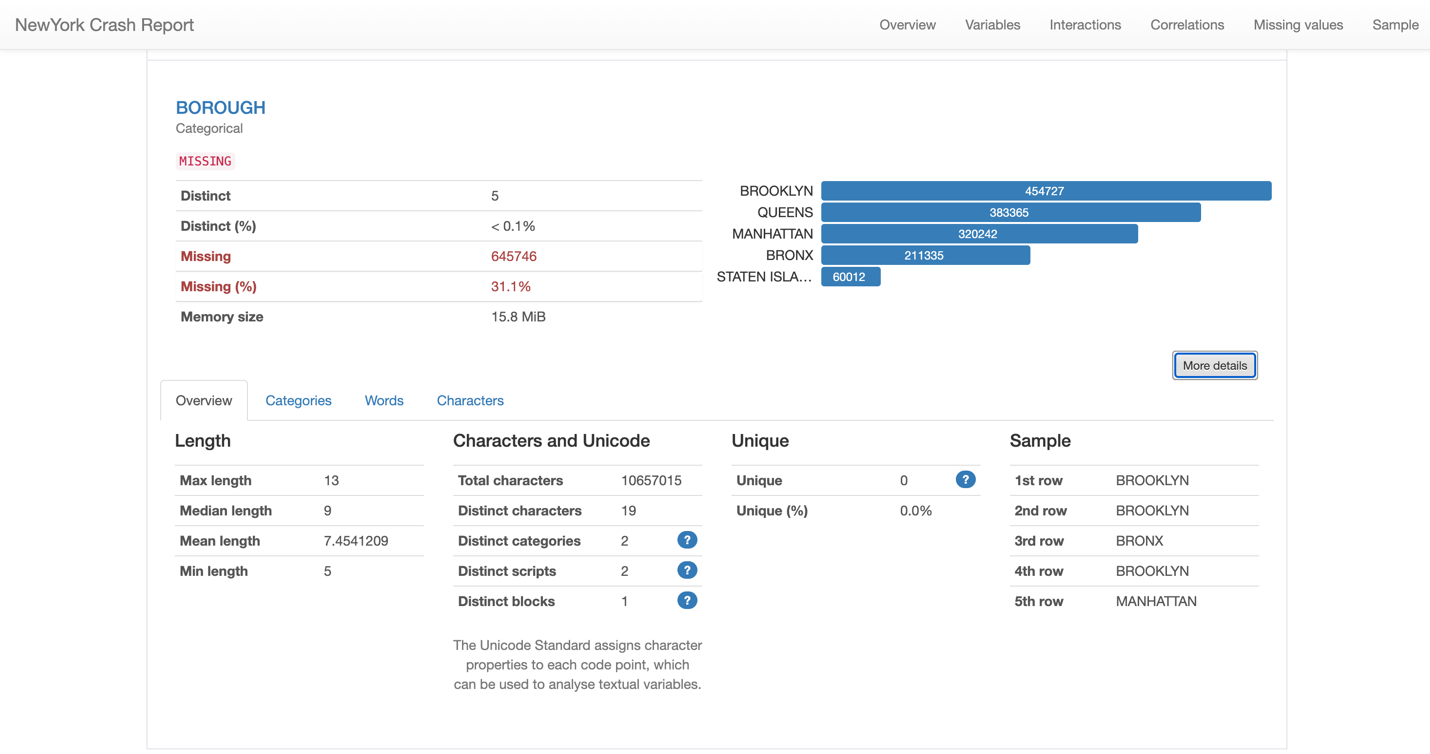
A screenshot of a computer

Description automatically generated

**For the New York Crash Report:**

The variable CRASH\_TIME also shows a distribution within a single day, with its own distinct count and memory size.

The timeframe for CRASH\_TIME is also just for April 7, 2024.



**Distinct Categories**: There are five distinct boroughs represented in the dataset, which is expected as New York City comprises five boroughs: Brooklyn, Queens, Manhattan, Bronx, and Staten Island.

**Missing Data:** There are 645,746 missing entries, which constitute 31.1% of the dataset. This is a significant proportion and might affect the accuracy of any analysis concerning boroughs unless the missing data can be accounted for or imputed.

**Length Statistics:**

* The maximum length of the borough names in the dataset is 13 characters.
* The median length is 9 characters.
* The mean (average) length is approximately 7.45 characters.
* The minimum length is 5 characters.

**Character and Unicode Analysis:**

There are a total of 10,657,015 characters across all the entries for the borough variable.

The data contains 19 distinct characters.

There are two distinct categories and two distinct scripts present, likely referring to different ways borough names are formatted or variations in data entry.

Only one distinct Unicode block is used, which suggests that there are no special or foreign characters used in the borough names.

**A screenshot of a computer

Description automatically generated**

**Zip Code:**

Based on the image provided for the "ZIP CODE" variable from the New York Crash Report:

* **Data Completeness**: 31.1% of the entries for the ZIP CODE variable are missing, which indicates that almost a third of the data for this variable is not recorded. This could impact analyses that rely on geographical segmentation or location-based trends.
* **Distinct Values**: There are 235 distinct ZIP codes present in the dataset. The distinctness of less than 0.1% suggests that the ZIP codes have a very high repetition rate, indicating that crashes are reported across a range of ZIP codes.
* **Length of ZIP Codes**: The ZIP codes have a consistent length, with a maximum, median, mean, and minimum length of 5 characters. This consistency is typical for ZIP codes, as they are standardized in length.
* **Characteristics of Data**:
  1. There are 11 distinct characters used within the ZIP codes, likely the numbers 0-9 plus a missing data placeholder or space.
  2. The data is categorized into 2 distinct categories, which might differentiate between regular ZIP codes and those that are missing or otherwise marked.
  3. All the ZIP codes use a single script and block in Unicode, which is expected as they are numerical and standardized for addresses.

**ZIP codes such as “11207”, “10467”, “10453”, and others appear prominently, suggesting a higher number of crash reports in these areas. These could be areas of interest for further investigation into crash causes or for targeted safety improvements.**

**Latitude:**

**A screenshot of a computer

Description automatically generated**

* **Data Completeness**: 11.3% of the entries for this variable are missing, which is a considerable amount but less than what we observed for the ZIP code data.
* **Distinct Values**: There are 126,594 distinct values, making up 6.9% of the total entries, suggesting that while there is some repetition, there's also a significant amount of variation.
* **Zeros**: There are 4,360 entries with a value of zero, but they only make up 0.2% of the data, indicating that zeros are not a common value for this variable.

**Longitude:A screenshot of a computer

Description automatically generated**

* **Data Completeness**: 11.3% of the data points for longitude are missing, which is substantial and could impact spatial analyses.
* **Distinct Values**: There are 98,351 distinct longitude values, making up 5.3% of the data, indicating a variety of crash locations.
* **Negative Values**: All longitude values are negative, which is appropriate since New York is in the Western Hemisphere, but there are also 4,360 zero values which are not valid for longitude and may indicate missing or improperly coded data.
* **Data Range and Central Tendency**:
  + The mean longitude is -73.752129, which is in the expected range for New York.
  + The minimum value of -201.35999 is an error since it is outside the possible range for longitude values, suggesting data quality issues.
  + The median value of -73.92726 is close to the mean, which is consistent for longitude data without extreme outliers.

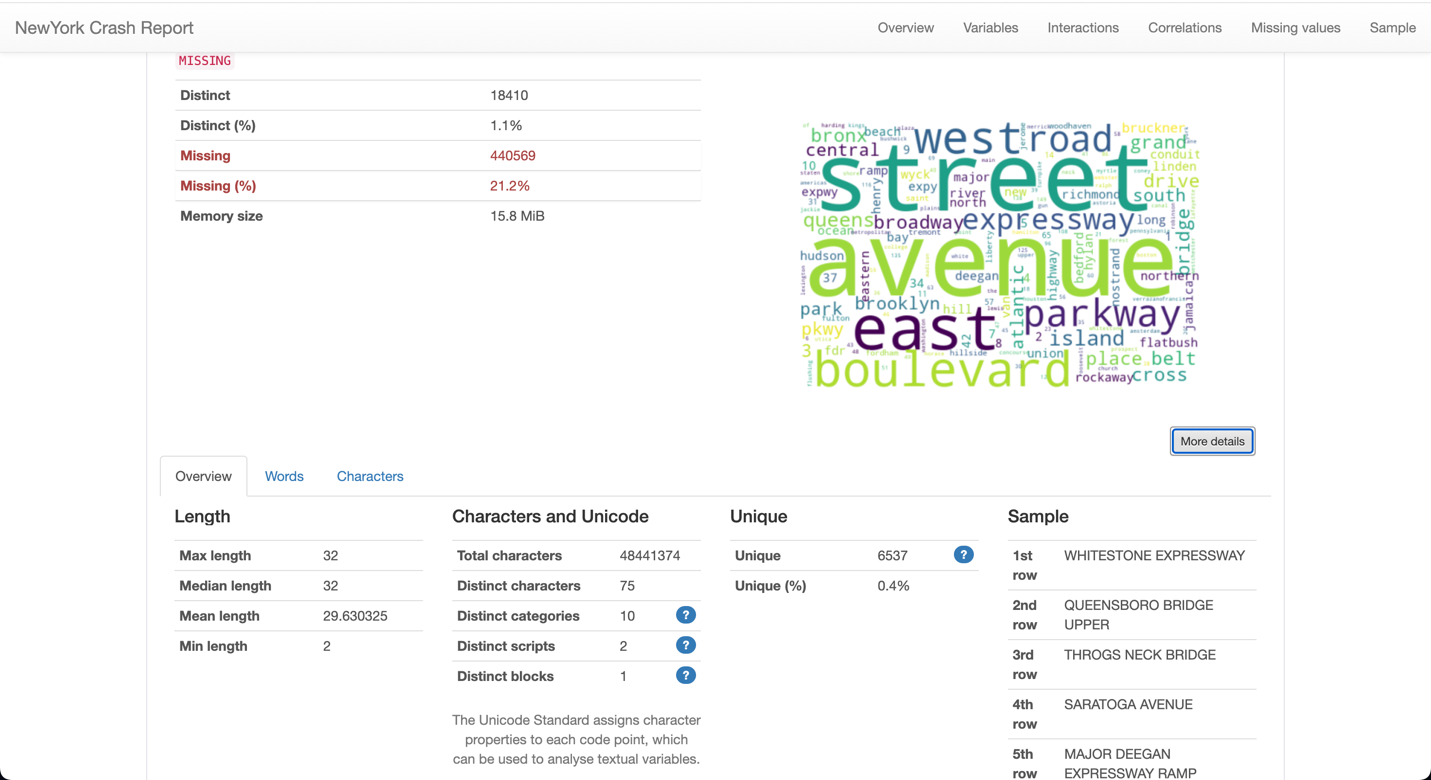
**Location:**

**A screenshot of a computer

Description automatically generated**

* **Distinct Locations**: There are 283,006 distinct entries for locations, which is 15.4% of the total data. This suggests a wide variety of crash locations across the dataset.
* **Missing Data**: 11.3% of the location data is missing, which could pose challenges for spatial analysis and may require data imputation.
* **Data Length**: The length of the location entries varies, with a minimum of 10 characters and a maximum of 25 characters. The median and mean lengths are 24 and approximately 22.78 characters, respectively, which indicates that most location entries are quite detailed.

**On Street Name:**

****

* **Distinct Values**: There are 18,410 distinct "ON STREET NAME" entries, representing 1.1% of the data. This indicates a variety of streets where crashes have occurred.
* **Missing Data**: A significant 21.2% of the data for this variable is missing. This could be due to crashes occurring at intersections without clear "on street" demarcations or reporting inconsistencies.
* **Data Length**: The entries vary in length with a minimum of 2 characters and a maximum of 32 characters. Both the median and mean lengths are close to the maximum, which suggests that most of the street names are fairly long, potentially including descriptors or multiple words.

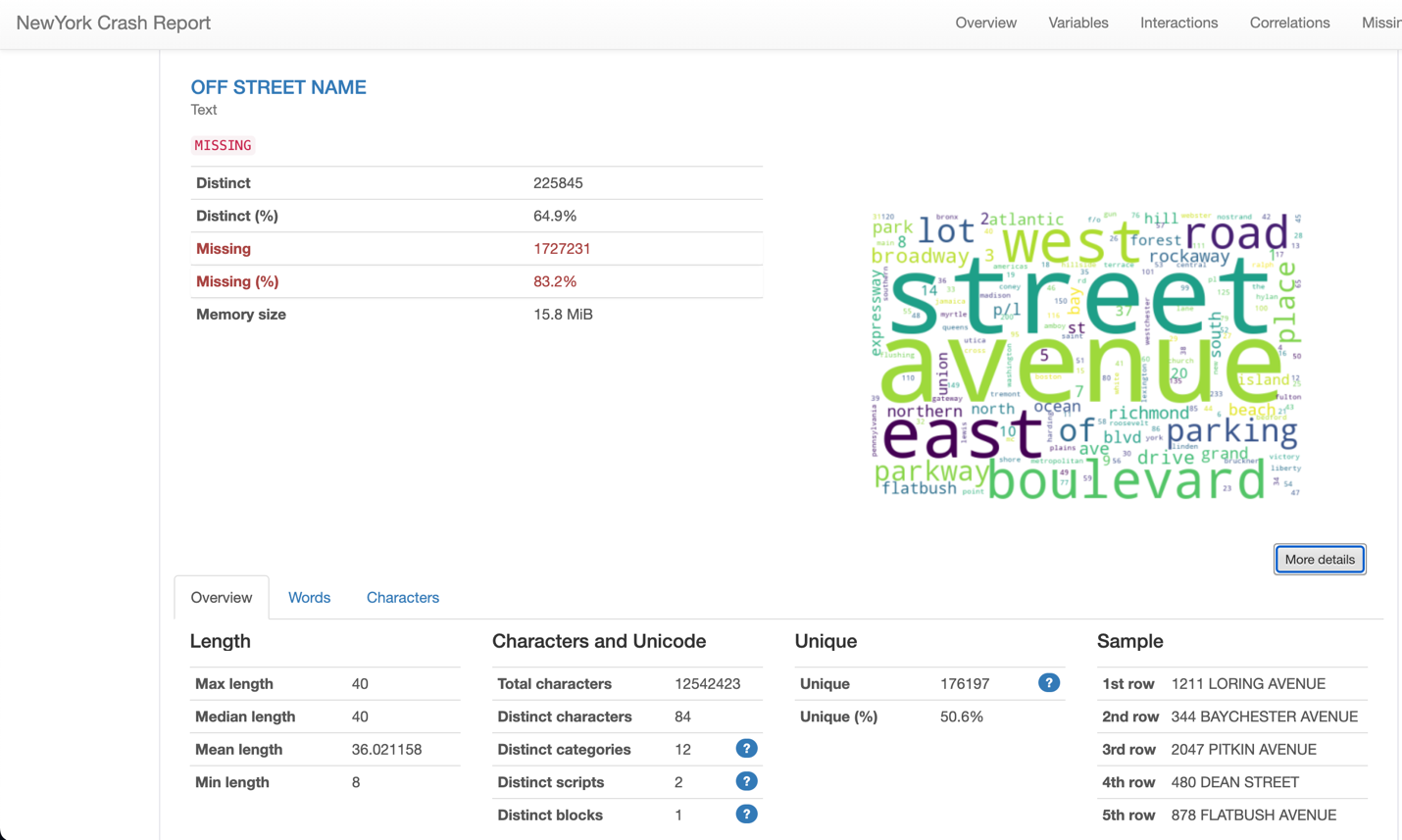
**Cross Street Name:**

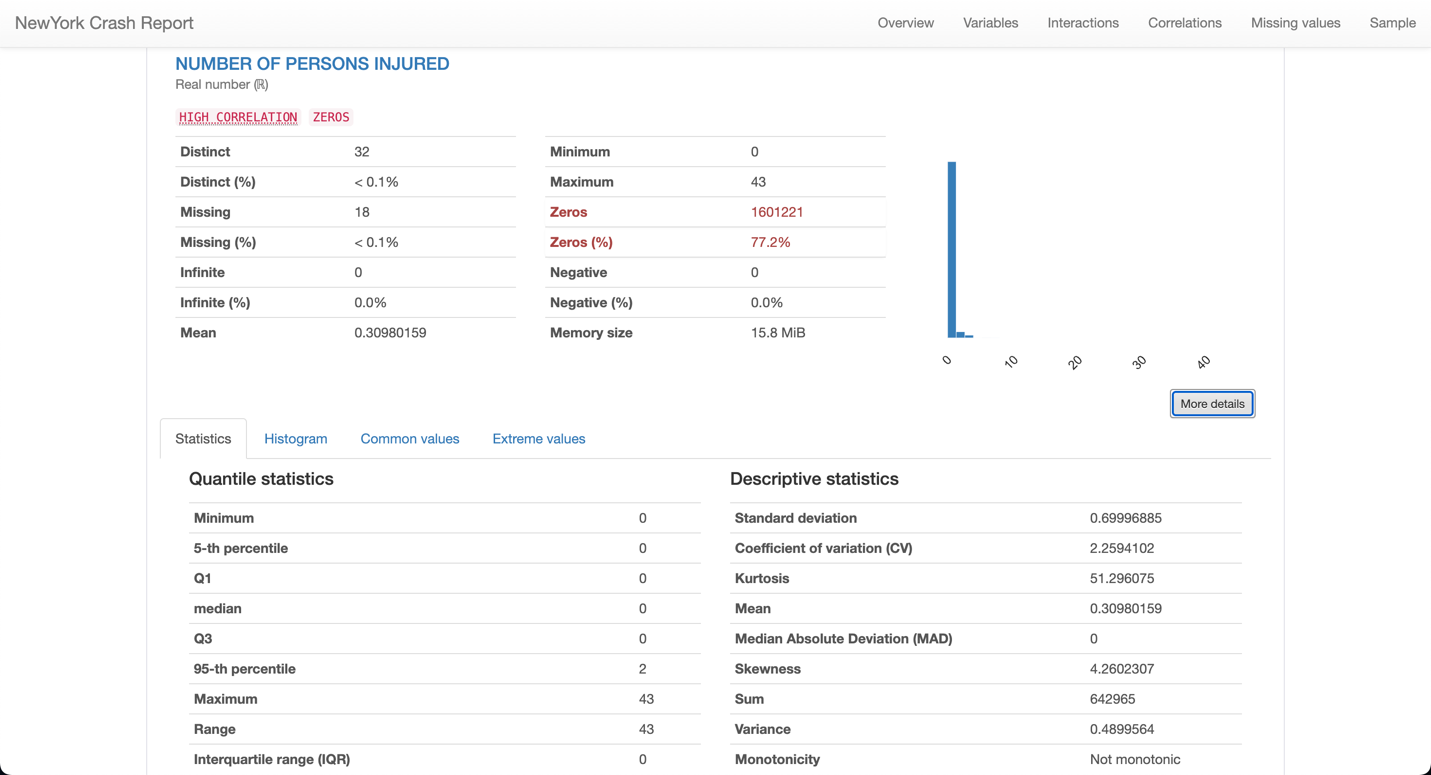
A screenshot of a computer

Description automatically generated

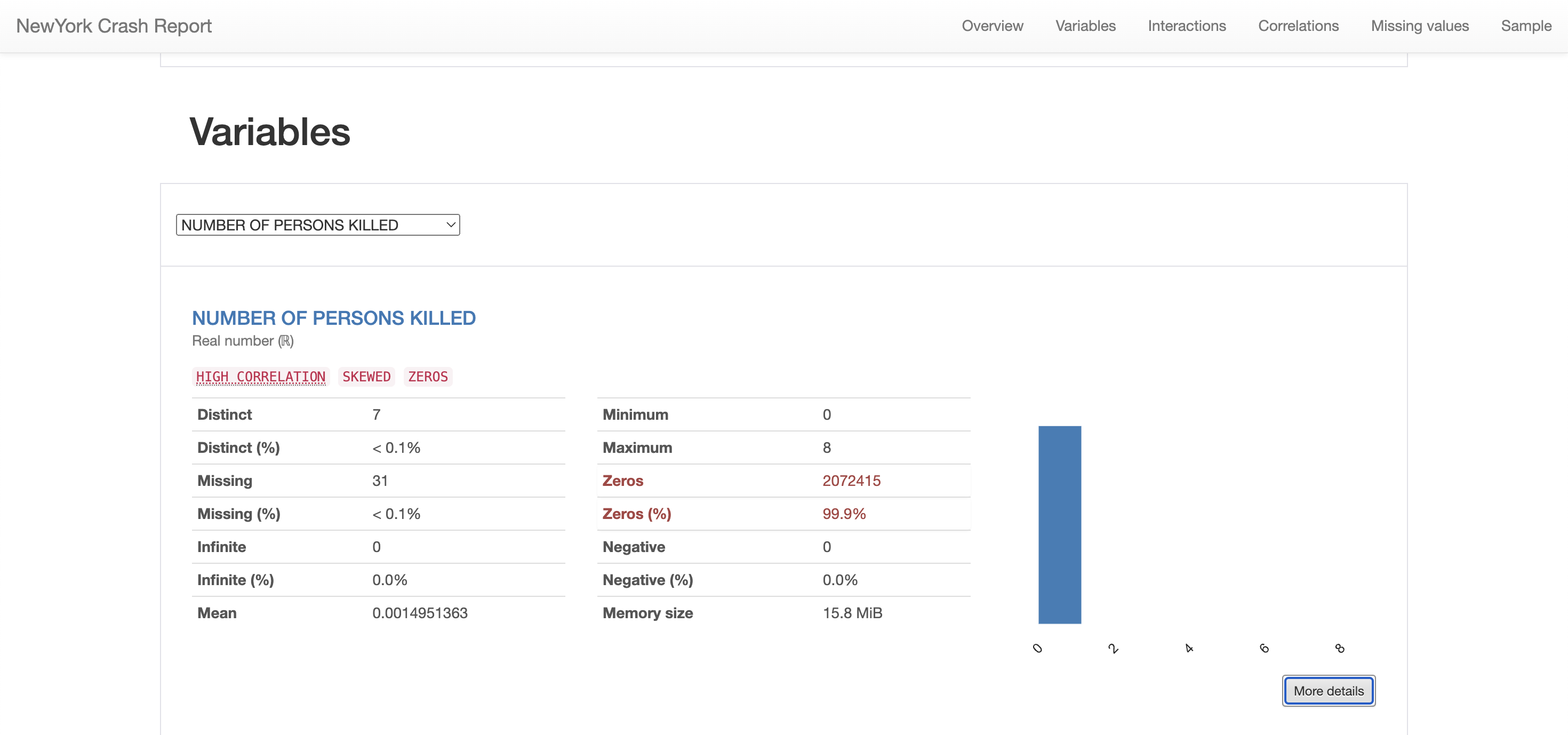
* **Distinct Values**: There are 20,236 distinct "CROSS STREET NAME" entries, representing 1.6% of the data. This suggests crashes occur at a variety of intersections across New York.
* **Missing Data**: A significant 37.8% of the data for this variable is missing, indicating that for many crashes, the cross street was not reported. This high percentage of missing data could hinder location-specific analysis.
* **Data Length**: The length of the cross street names varies, with a minimum of 1 character and a maximum of 32 characters. The median and mean lengths are 32 and approximately 22.70 characters, respectively.

**Cross Street Name:**

****

* **Distinct Values**: There are 225,845 distinct entries, which is a substantial 64.9% of the data. This suggests a very high diversity in the off-street locations where crashes have occurred.
* **Missing Data**: A very high 83.2% of the "OFF STREET NAME" data is missing, which could greatly impact the usefulness of this variable for location-specific analysis.
* **Data Length**: The length of the entries ranges from 8 to 40 characters, with the median and mean lengths both close to the maximum, indicating that when "OFF STREET NAME" data is provided, it tends to be quite detailed.
* **Number of Persons Injured:**
* **Distinct Values**: There are 32 distinct values, which indicate the number of persons injured in crashes ranges from 0 to a certain number (maximum provided in the data).
* **Missing Data**: There are only 18 cases missing, which is less than 0.1% of the total, indicating that this variable is almost completely reported.
* **Zeros**: A large majority of crashes, 77.2%, did not result in any reported injuries. This could indicate a high number of non-injury incidents or minor crashes.
* **Range**: The maximum number of persons injured in a single crash is 43, which suggests at least one major incident.
* **Central Tendency**: The mean number of persons injured per crash is approximately 0.31, indicating that on average, there are very few injuries per crash.

**Number of Persons Killed:**



* **Distinct Values**: There are 7 distinct values for the number of persons killed in crashes, indicating that the vast majority of crashes did not result in fatalities.
* **Missing Data**: There are 31 cases missing, which is an extremely small fraction of the data (< 0.1%), indicating that this variable is nearly completely reported.
* **Zeros**: A significant majority of crashes, 99.9%, had no fatalities. This high percentage of zero values is typical for traffic crash data, where fatalities are fortunately less common than non-fatal injuries.
* **Range**: The number of fatalities in crashes ranges from 0 to 8, with the maximum value indicating at least one extremely severe crash.
* **Mean**: The mean number of fatalities per crash is very low (approximately 0.0015), reinforcing the observation that fatalities are rare.

**Number of Pedestrians Injured:**

**A screenshot of a computer

Description automatically generated**

* **Distinct Values**: There are 14 distinct values for the number of pedestrians injured in crashes, indicating the count of injured pedestrians in an incident varies from 0 up to a maximum observed in the data.
* **Zeros**: The vast majority of the data, 94.6%, recorded zero pedestrians injured, which suggests that pedestrian injuries are relatively uncommon in the reported crashes.
* **Range**: The number of pedestrians injured ranges from 0 to 27, with the maximum indicating a particularly serious incident involving numerous pedestrians.

**Number of Pedestrians Killed:**

**A screenshot of a computer

Description automatically generated**

* **Distinct Values**: There are 4 distinct values for the number of pedestrians killed in the reported crashes.
* **Data Distribution**: The majority of crashes (207,3905 instances) reported zero pedestrian fatalities, which suggests that fatal pedestrian incidents are relatively infrequent.
* **Reported Fatalities**: There are 1,509 instances of a single pedestrian being killed, 12 instances of two pedestrians killed, and 1 instance of six pedestrians killed in a crash.
* **Missing Data**: There are no missing entries for this variable, indicating that the dataset is complete with respect to reporting pedestrian fatalities.

**Number of Cyclist Injured:**

A screenshot of a computer

Description automatically generated

* **Distinct Values**: There are 4 distinct values for the number of pedestrians killed in the reported crashes.
* **Data Distribution**: The majority of crashes (207,3905 instances) reported zero pedestrian fatalities, which suggests that fatal pedestrian incidents are relatively infrequent.
* **Reported Fatalities**: There are 1,509 instances of a single pedestrian being killed, 12 instances of two pedestrians killed, and 1 instance of six pedestrians killed in a crash.
* **Missing Data**: There are no missing entries for this variable, indicating that the dataset is complete with respect to reporting pedestrian fatalities.

**• Data Distribution**: A majority of the data, 2,020,463 instances, reported zero cyclists injured, indicating that injuries to cyclists are relatively uncommon in the reported crashes.

* **Reported Injuries**: There are 54,340 instances of one cyclist being injured, 600 instances of two cyclists injured, 23 instances of three cyclists injured, and 1 instance of four cyclists injured in a crash.
* **Missing Data**: There are no missing entries for this variable, indicating that the dataset is complete regarding reporting cyclist injuries.

**Number of Cyclist Killed:**

A screenshot of a computer

Description automatically generated

* **Distinct Values**: There are 3 distinct values for the number of cyclists killed in the crashes.
* **Data Distribution**: The overwhelming majority of the data, 2,075,189 instances, reported zero cyclists killed, which suggests that cyclist fatalities are quite rare in the reported crashes.
* **Reported Fatalities**: There are 237 instances of a single cyclist being killed and 1 instance of two cyclists killed in a crash.
* **Missing Data**: There are no missing entries for this variable, indicating that the dataset is complete with respect to reporting cyclist fatalities.

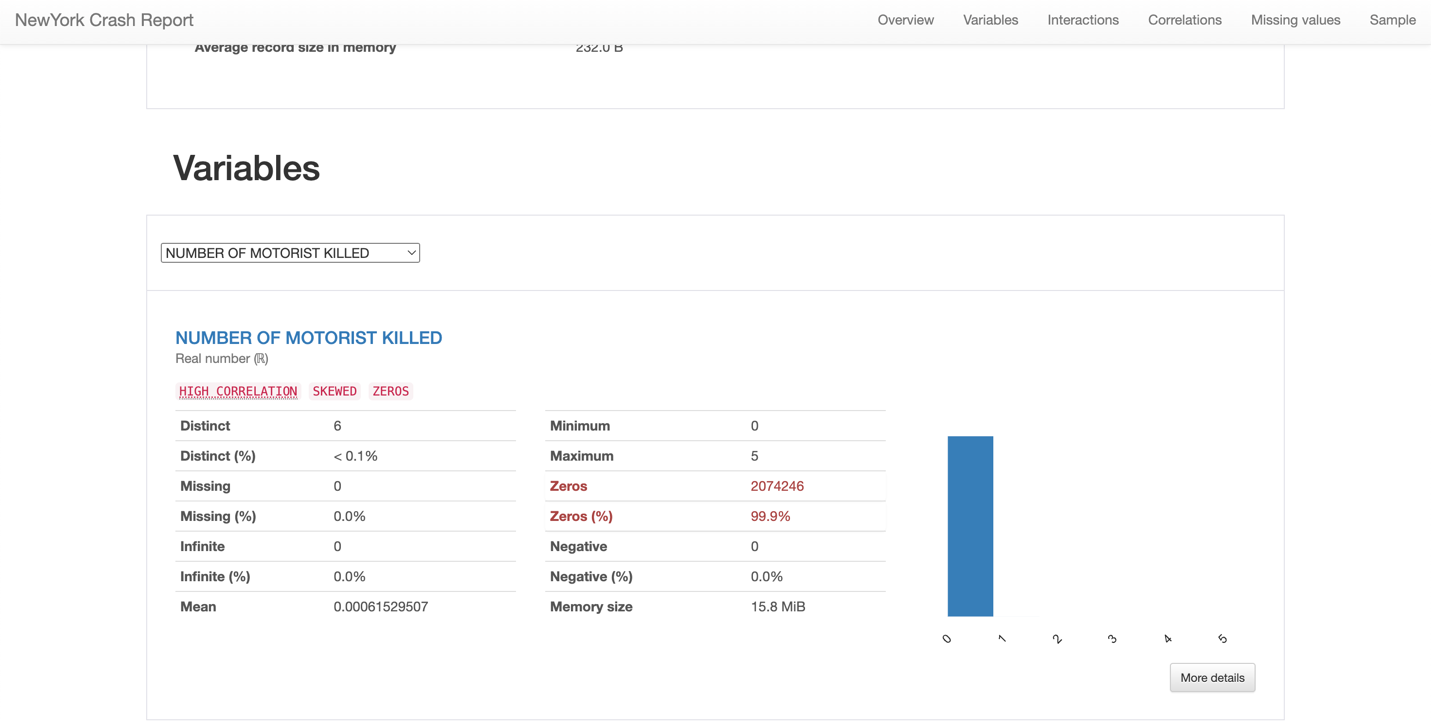
**Number of Motorist Injured:**

A screenshot of a computer

Description automatically generated

* **Distinct Values**: There are 31 distinct values, indicating the count of injured motorists in an incident ranges from 0 up to a maximum observed in the data.
* **Data Distribution**: A majority of the data, 1,772,939 instances, reported zero motorists injured, which suggests that a significant number of crashes do not result in motorist injuries.
* **Range**: The maximum number of motorists injured in a single crash is 43, which is quite high and indicates at least one very serious incident.
* **Missing Data**: There are no missing entries for this variable, indicating that the dataset is complete with respect to reporting motorist injuries.
* **Mean**: The mean number of motorists injured per crash is approximately 0.22, suggesting that when motorist injuries occur, they are usually low in number.

**Number of Motorist Killed:**



* **Distinct Values**: There are 6 distinct values for the number of motorists killed in crashes.
* **Data Distribution**: The vast majority of crashes, 2,074,246 instances, reported zero motorists killed, indicating that most crashes do not result in motorist fatalities.
* **Range**: The number of fatalities among motorists in crashes ranges from 0 to 5, with the maximum suggesting at least one extremely severe crash.
* **Missing Data**: There are no missing entries for this variable, which means the dataset is complete with respect to reporting motorist fatalities.
* **Mean**: The mean number of motorists killed per crash is very low (approximately 0.000615), reinforcing the observation that fatalities among motorists are rare.

**CONTRIBUTING FACTOR VEHICLE 1**

**A screenshot of a computer

Description automatically generated**

* **Distinct Values**: There are 61 distinct contributing factors listed for the first vehicle involved in crashes.
* **Missing Data**: There are 6,802 missing entries, which represent 0.3% of the total data.
* **Data Length**: The descriptions for contributing factors range in length, with some as short as 1 character and others as long as 53 characters.

**Common terms include "Unspecified", "Inattention/Distraction", "Failure", "Driver", "Lane", "Passing", "Speed", "Fatigued/Drowsy", and "Traffic Control Disregarded".**

**"Unspecified" appears to be the most common entry, which may indicate many reports do not have a clear contributing factor or the data was not fully captured.**

**CONTRIBUTING FACTOR VEHICLE 2A screenshot of a computer

Description automatically generated**

* **Distinct Values**: There are 61 distinct contributing factors listed for the second vehicle involved in crashes.
* **Missing Data**: There are 321,736 missing entries, which represent 15.5% of the total data.
* **Memory Size**: The data for this variable occupies 15.8 MiB in memory.
* **Data Length**: The contributing factors descriptions vary in length, similar to the contributing factors for vehicle 1.

**Common terms include "Unspecified", "Inattention/Distraction", "Following Too Closely", "Failure to Yield Right-of-Way", "Passing or Lane Usage Improper", and "Unsafe Speed".**

**As with vehicle 1, "Unspecified" appears to be a common entry, suggesting that a clear contributing factor is not always recorded for the second vehicle involved in a crash.**

**CONTRIBUTING FACTOR VEHICLE 3**

**A screenshot of a computer

Description automatically generated**

* **Distinct Values**: There are 51 distinct contributing factors listed for the third vehicle involved in crashes.
* **Missing Data**: A significant portion, 192,7163 instances (92.9% of the data), are missing, indicating that for the majority of crashes, the contributing factor for a third vehicle is not reported or applicable.
* **Memory Size**: The data for this variable occupies 15.8 MiB in memory.

**Word Cloud Analysis**:

* The word cloud shows the frequency of words in the contributing factor descriptions, with "Unspecified" being overwhelmingly the most common term. Other noticeable words include "Inattention/Distraction", "Following Too Closely", "Failure to Yield Right-of-Way", and various others that are also present in the contributing factors for the first and second vehicles.

**CONTRIBUTING FACTOR VEHICLE 4**

**A screenshot of a computer

Description automatically generated**

**Distinct Values**: There are 41 distinct contributing factors listed for the fourth vehicle involved in crashes.

**Missing Data**: There are 2,041,953 missing entries, which represent 98.4% of the total data. This high percentage indicates that the contributing factor for a fourth vehicle is often not applicable or not reported.

**Data Length**:

* The maximum length of the contributing factor descriptions is 43 characters.
* The median length is 11 characters.
* The mean length is around 11.5 characters.

**The term "Unspecified" is prominently displayed, which indicates that when a fourth vehicle is involved, the contributing factor is often not specified.**

**Other terms that are visible include "Following Too Closely", "Driver Inattention", "Fatigued/Drowsy", "Other Vehicular", and "Failure to Yield Right-of-Way".**

**CONTRIBUTING FACTOR VEHICLE 5**

**A screenshot of a computer

Description automatically generated**

**Distinct Values**: There are 30 distinct contributing factors for the fifth vehicle involved in crashes.

**Missing Data**: A significant number of entries, 2,066,358, are missing, which accounts for 99.6% of the total data. This suggests that in most reported crashes, a fifth vehicle is either not involved or the contributing factors for such a vehicle are not recorded.

**Memory Size**: The data for this variable occupies 15.8 MiB in memory.

**The word cloud features terms like "Unspecified", "Following Too Closely", "Driver Inattention", "Pavement Slippery", among others. However, "Unspecified" is the predominant term, suggesting that when a fifth vehicle is involved, the specific contributing factor is often not detailed.**

**COLLISON\_ID**

**A screenshot of a computer

Description automatically generated**

* **Distinct**: There are 207,5427 distinct collision IDs, which implies each entry in the dataset corresponds to a unique collision event.
* **Distinct (%)**: Every entry (100%) is distinct, affirming the uniqueness of each collision ID.
* **Missing Data**: There is no missing data for this variable; every crash reported has a unique ID.
* **Minimum and Maximum**: The minimum collision ID is 22, and the maximum is 4,712,252, which indicates a wide range and possibly a long time frame for the data collection.

**Vehicle Type Code 1**

**A screenshot of a computer

Description automatically generated**

* **Distinct**: There are 1,631 distinct types of vehicles, suggesting a significant diversity in vehicle types involved in crashes.
* **Missing Data**: There are 13,691 instances where the vehicle type is not reported, making up 0.7% of the total data.
* **Memory Size**: This variable uses 15.8 MiB of memory space in the report.

**The word cloud prominently features "Utility", "Sedan", "Station", "Sport", "Vehicle", indicating these are common vehicle types involved in the crashes.**

**The terms "Passenger", "Pick-up", and "Wagon/Sport" are also visible, which may refer to specific vehicle categories.**

**Vehicle Type Code 2**

**A screenshot of a computer

Description automatically generated**

* **Distinct**: There are 1,819 distinct entries for the second vehicle types, indicating a wide range of vehicles involved.
* **Missing**: A significant number of entries, 396,691 or 19.1% of the dataset, do not have this vehicle type recorded, which could suggest that many collisions involve only one vehicle or the second vehicle type was not captured in the report.

**Unique Entries**: There are 1,080 unique entries which account for less than 0.1% of the total.

**The word cloud displays “Sedan”, “Utility”, “Station”, and “Sport” prominently, similarly to the first vehicle type code variable. This suggests that these types of vehicles are not only primary but also frequently involved as secondary vehicles in collisions.**

**Vehicle Type Code 3**

**A screenshot of a computer

Description automatically generated**

* **Distinct**: There are 260 unique types of tertiary vehicles listed.
* **Missing**: A significant portion, 1,932,530 entries or 93.1%, are missing this information, which likely indicates that most collisions reported do not involve a third vehicle.
* **Unique Entries**: There are 152 unique entries representing less than 0.1% of the data, suggesting a relatively small variety of third vehicles when compared to primary and secondary vehicles.

The word cloud highlights "Sedan" and "Station" as some of the most common entries, similar to the first and second vehicle type codes.

**Vehicle Type Code 4**

**A screenshot of a computer

Description automatically generated**

* **Distinct**: There are 101 unique entries which denote different types of the fourth vehicle involved in the accidents.
* **Missing**: There's a significant number of missing entries (2,043,115), accounting for 98.4% of the data, indicating that a fourth vehicle is rarely involved in collisions.
* **Memory Size**: The data for this variable occupies 15.8 MiB of memory.
* **Unique Entries**: There are 45 unique entries making up less than 0.1% of the total data, which shows there is a very small variation in the types of fourth vehicles involved.

The word cloud highlights the terms "utility," "station," and "sedan" among the most common vehicle types listed, though the vast majority of the reports do not include a fourth vehicle.

**Vehicle Type Code 5**

**A screenshot of a computer

Description automatically generated**

* **Distinct**: There are 70 unique types of vehicles listed as the fifth vehicle involved in collisions.
* **Missing**: A substantial amount of data is missing (2,066,635 entries), accounting for 99.6% of the data, suggesting that a fifth vehicle is very rarely involved in accidents.
* **Memory Size**: The information for this variable takes up 15.8 MiB in memory.
* **Unique Entries**: Out of the total entries, there are 31 unique types of vehicles listed, comprising only 0.4% of the data, indicating minimal variation in this category.

The word cloud in the visualization prominently displays "utility," "station," and "sedan," indicating these are among the more commonly listed types when a fifth vehicle is involved in an accident, though such occurrences are evidently rare.