CSE 587 DATA INTENSIVE COMPUTING

PHASE-I

DETECTION OF IMPACT ZONE ON EARTH BY SPACE OBJECTS

Team members

Jahnavi Rudraraju

UBID: 50464467

jahnavir@buffalo.edu

Sujith Varma Vatsavai

UBID: 50442317

sujithva@buffalo.edu

Dataset Link: https://www.kaggle.com/datasets/sameepvani/nasa-nearest-earth-objects

Motivation

Earth is greeted everyday by asteroids, comets, meteors, space junk such as detached rocket parts, tiny components from shuttles, shattered space shuttles in space. These particles revolve around the earth's orbit with a minimum speed of 29.78km/s. We cannot find all the tiny particles with our current technology but we can analyze and keep track of the large particles like comets, asteroids, meteors based on their closest approach and speed of orbits. Most of the large particles gets burnt away while entering earth's atmosphere, but some particles survive and impacts everyday life. We cannot physically stop the particles from entering our atmosphere. The best we can do is to analyze the impact zone and try to evacuate the vicinity before collision, according to our dataset the collision course of each object can be measured by their rarity and the closest approach distance, date and time. We can also measure the impact of the collision based on their size in the data provided.

Problem statement

The main problem we are going to focus is the detection of objects hitting earth, some object with diameter of at-least 5kms can be detected by using 'Radar tracking data', 'NEOWISE spacecraft' to analyze the orbit of the particles, our problem statement will focus on small particles of a diameter below 1-2 kms. We analyze the data given by NEO observers

Analyzing whether the object is on collision course with earth, the possible impact zone and the power of impact of all the space objects based on their rarity, last closest approach to earth, magnitude and diameter of the object.

Dataset Description

Dataset contains all the objects that passed by the earth, described with 90836 rows (Number of objects) and 10 columns which includes the name of the object, closest approach data, the distance between earth and the object, the speed of the object, size of the object and is the object dangerous based on the magnitude.

Implementation of code

We pre-process the code using data cleaning methods, here we just renamed a column and created 2 new columns with it. Our dataset has all the details necessary for future classification so we did not remove or replace any values because our ideology of tracking all the objects depend on these values. EDA (Exploratory data analysis) is used for visualization of the data.

Data cleaning methods

In order to process the data of the csv file (or) the dataset we need to perform certain operations to clean and the filter the mentioned data to fit in our code properly, we used various kinds of operations in our dataset and all the methods are described below.

1. **Outliners:**
   Outliners in the data are a separate set of data points which stays with the rest of the values, within the dataset the outliners can fluctuate the output. Outliners can be detected when abnormal distribution of data points occur (or) we can use boxplot function to detect the outliners. Here in our data we are removing 75% of the outliners because the data distribution is disturbed and the output graphs are being fluctuated.

   We used IQR method to split the outliners using quantile function

2. **Splitting Columns:**
   In our data-frame we are splitting columns into 2 and adding them to our dataset in order to perform further classification, categorical columns are being split into 2. By using this method we can understand the data properly.

3. **Renaming (or) Clear formatting:**
   We used clear formatting in our code to make the columns more readable and simplified to use in the EDA process, this process is also used to implement another data cleaning method (splitting).

4. **Duplicate Entries:**
   Our dataset does not have any duplicated values and if we had duplicate values they are important because our dataset is based on the micro details, even the decimals are important for the classification process because our dataset revolves around all the

object that might hit the earth, so due to lack of visual observation of the tiny particles we rely on this data.

5. **Missing values:**
   This dataset is report (or) is a record of all the asteroids, meteors, detached rocket particles drifting in space and revolving around earth's orbit. There are no null, missing, and NaN values in this dataset. Some null are present in 'Estimated diameter max' column but these values are used for classification.

6. **Inconsistent values:**
   Our dataset has many inconsistent data but these values are useful for further processing, for example in the column 'Estimated diameter max' there are similar values, NaN values all the values are replaced to 0 because the all the information missing in this column can be compensated from another column.

7. **Data transformation:**
   This process is essential because our dataset contains over 90000 rows, all the EDA graphs should be plotted in a large scale all the graphs are skewed to the left and some are skewed to the right so we use log transformation to normalize the data and use this normalized data to plot graphs.

EDA (Exploratory Data Analysis) Output graphs

Exploratory Data Analysis is the process of analyzing, summarizing, and visualizing a dataset to gain insights and understanding of the data. EDA is done to understand big data before using expensive big data methodology. EDA is the first step in data analysis that identifies the patterns and anomalies in the data. Basic tools of EDA are plots, graphs, and summary stats. It is a method for systematically going through data, plotting distributions, plotting time series, looking at pairwise relationships using scatter plots, generating summary stats. Eg: min, max, mean, upper, lower quartiles, identifying outliers.

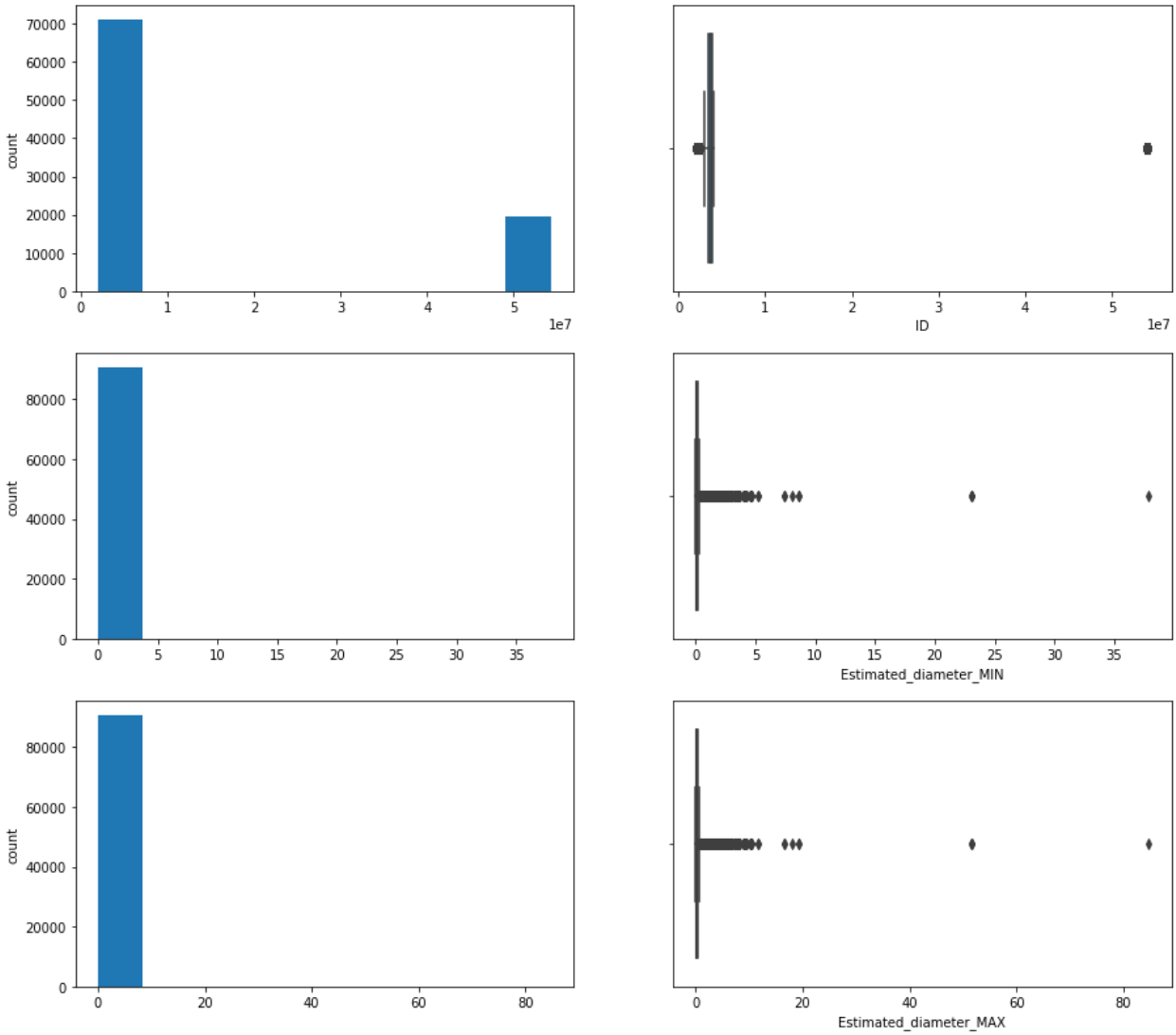
Univariate, bivariate and multivariate are used to describe the number of variables that are being analyzed in any particular analysis. Choosing any of these analysis will depend on the type of data that needs to be analyzed.

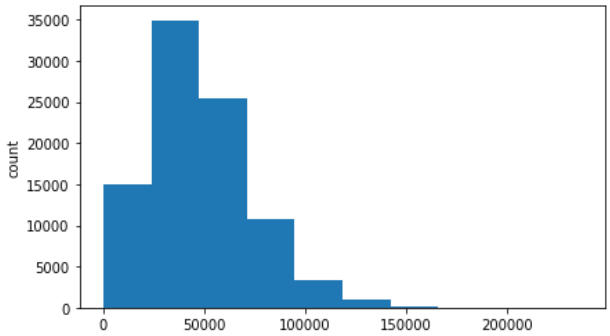
EDA Univariate Analysis:

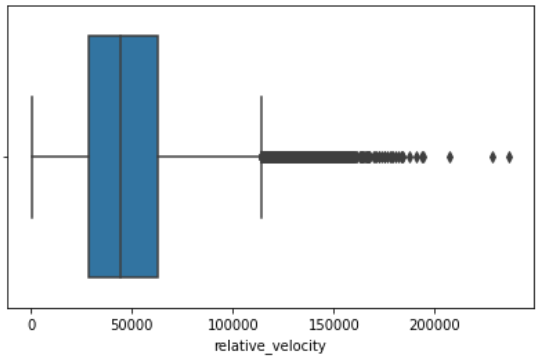
Univariate analysis is based on analyzing a single variable at a time. This is used for understanding the distribution and spread of a variable. Most used approaches of univariate are histograms, box plots, and summary statistics.

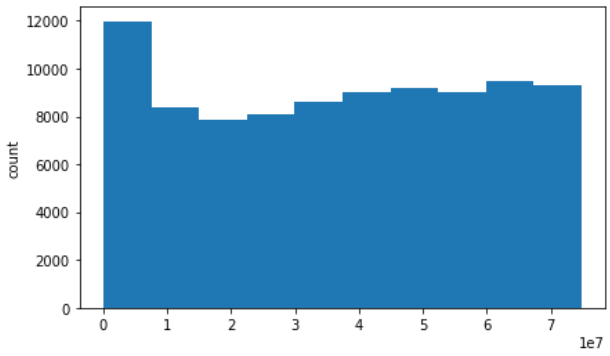
1. **Boxplot:**

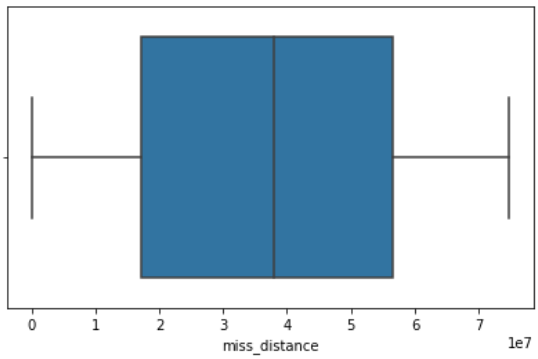
   Box plot is used to show the distribution of the data in boxes, only numerical values can be plotted in this method. This type of graph has shows us whiskers like structures to represent the data and outliners for the rest of the data.

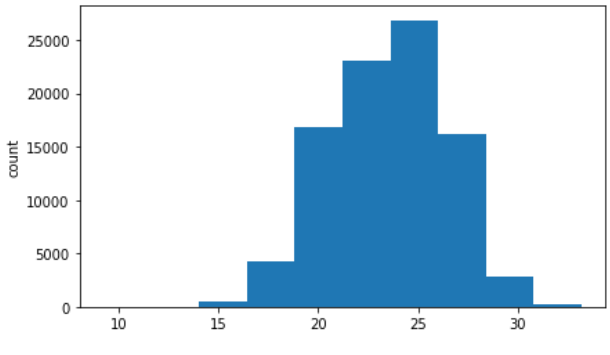


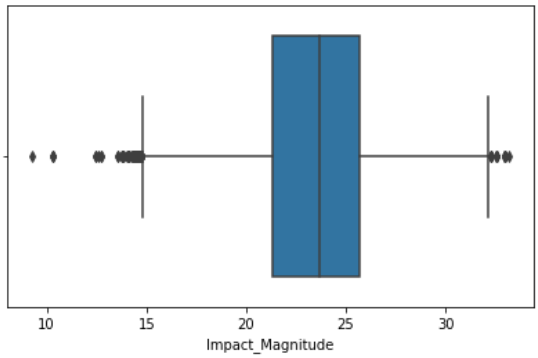

count
0
5000
10000
15000
20000
25000
30000
35000
0
50000
100000
150000
200000

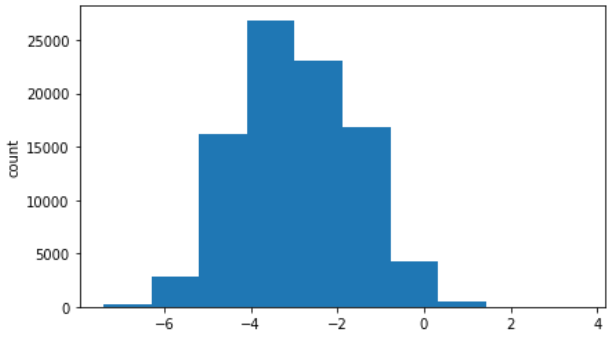

0
50000
100000
150000
200000
relative_velocity

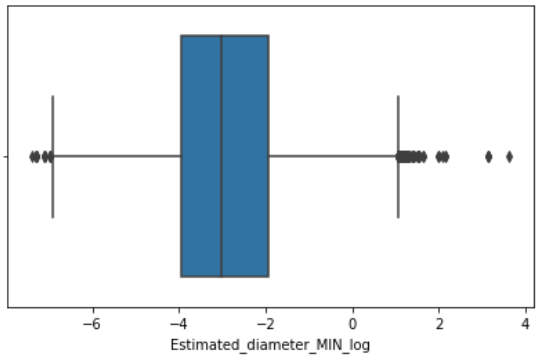

count
0
2000
4000
6000
8000
10000
12000
0
1
2
3
4
5
6
7
1e7

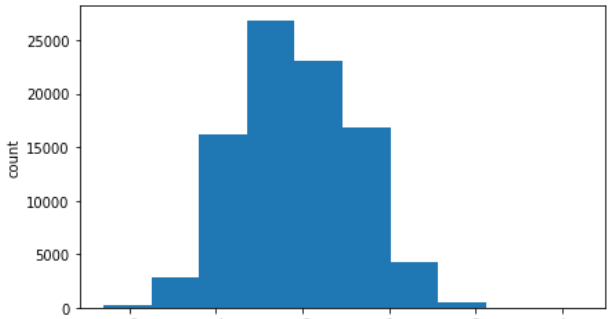

0
1
2
3
4
5
6
7
miss_distance
1e7

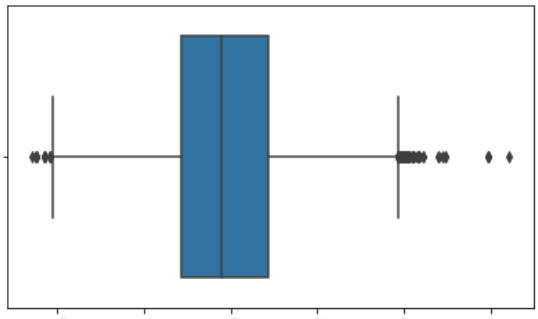

count
0
5000
10000
15000
20000
25000
10
15
20
25
30

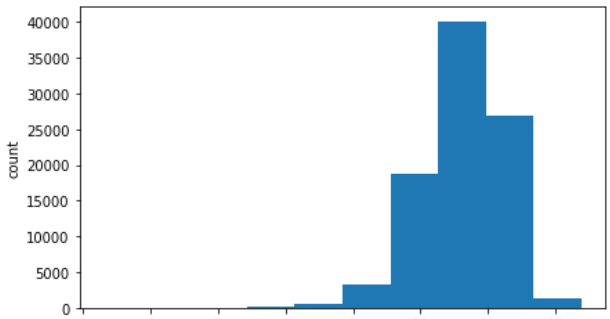

10
15
20
25
30
Impact_Magnitude

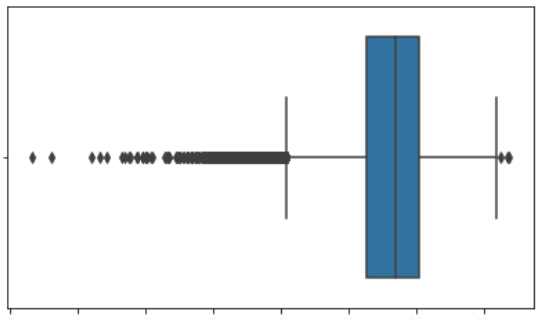

count
0
5000
10000
15000
20000
25000
−6
−4
−2
0
2
4

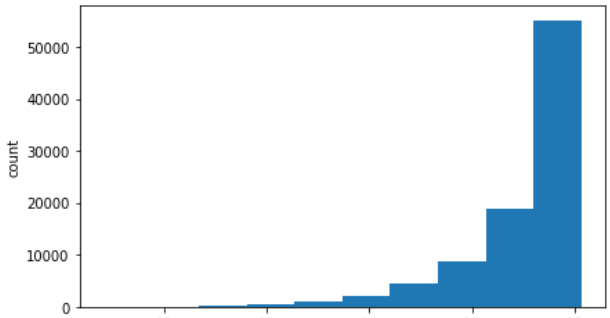

−6
−4
−2
0
2
4
Estimated_diameter_MIN_log

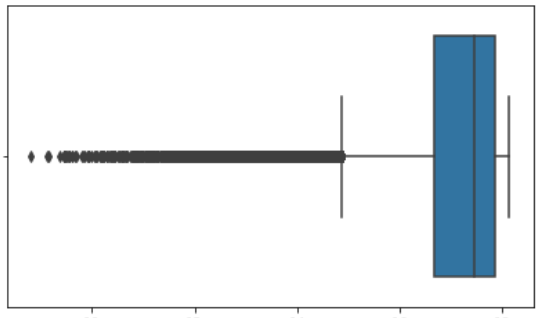

count
25000
20000
15000
10000
5000
0
−6
−4
−2
0
2
4

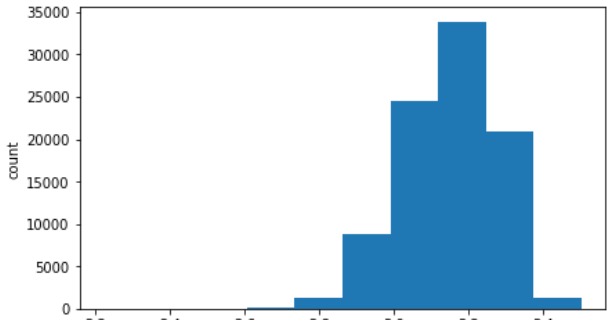

−6
−4
−2
0
2
4
Estimated_diameter_MAX_log

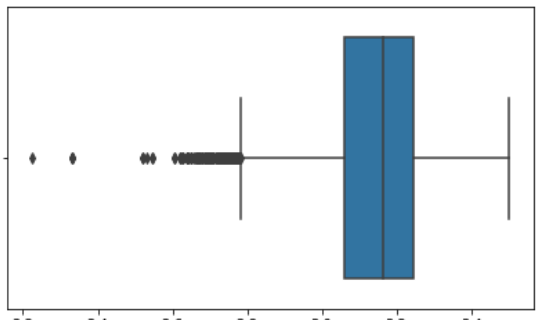

count
40000
35000
30000
25000
20000
15000
10000
5000
0
5
6
7
8
9
10
11
12


5
6
7
8
9
10
11
12
relative_velocity_log


count
50000
40000
30000
20000
10000
0
10
12
14
16
18

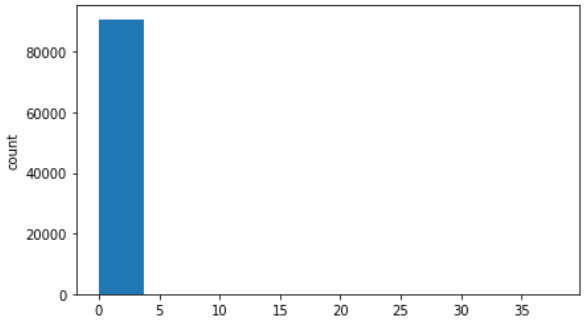

10
12
14
16
18
miss_distance_log

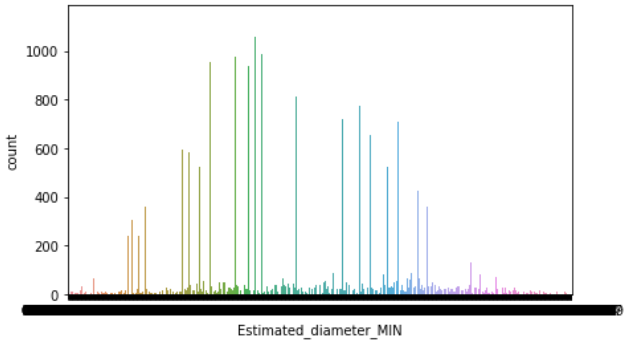

count
35000
30000
25000
20000
15000
10000
5000
0
2.2
2.4
2.6
2.8
3.0
3.2
3.4

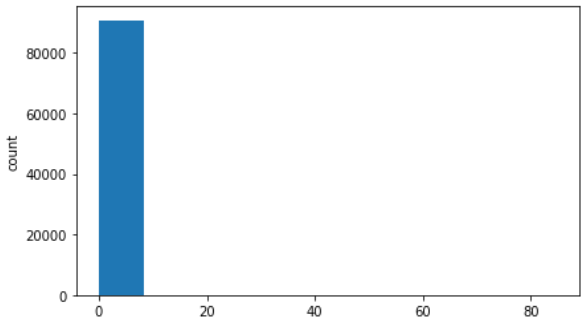

2.2
2.4
2.6
2.8
3.0
3.2
3.4
Impact_Magnitude_log

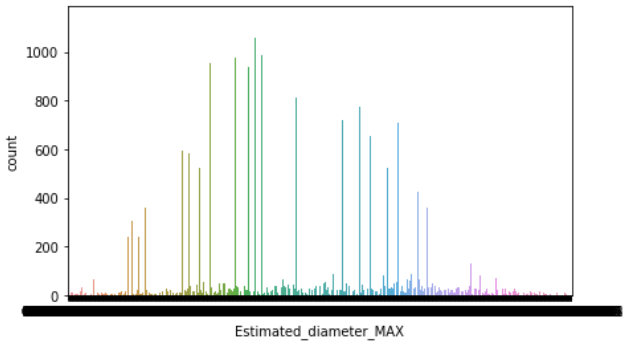
2. CountPlot:

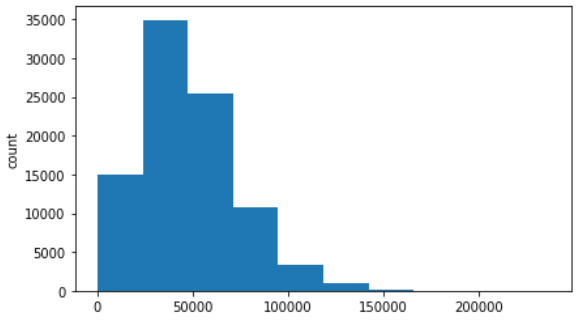
It represents the counts of the observation present in the categorical variable. It shows the visual depiction in the bar chart.

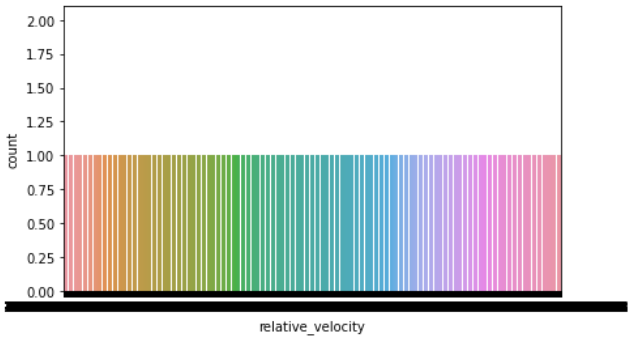


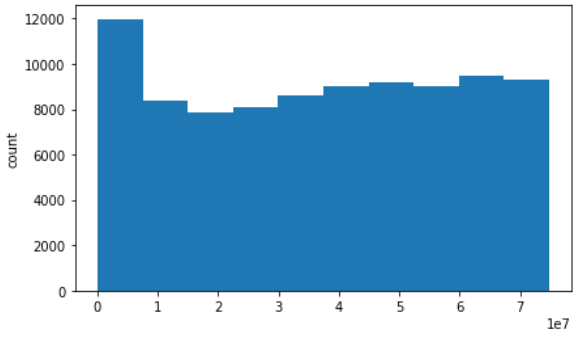


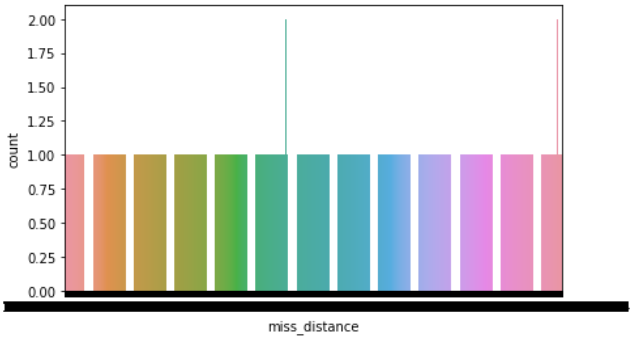


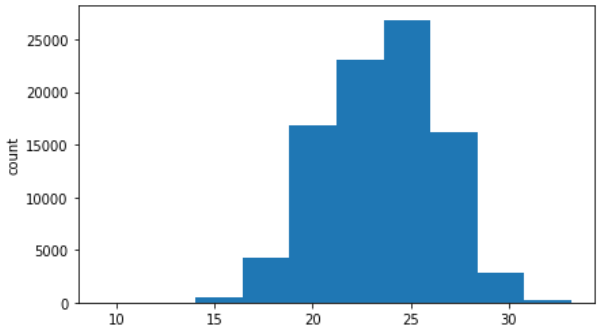


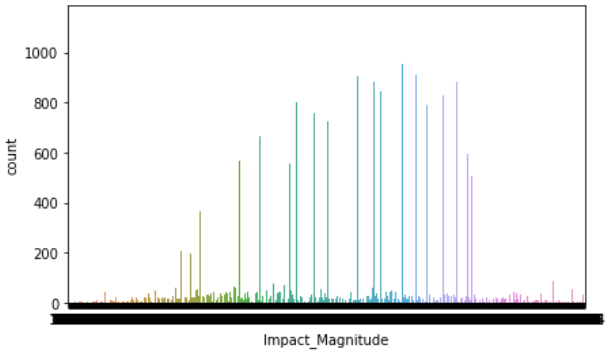


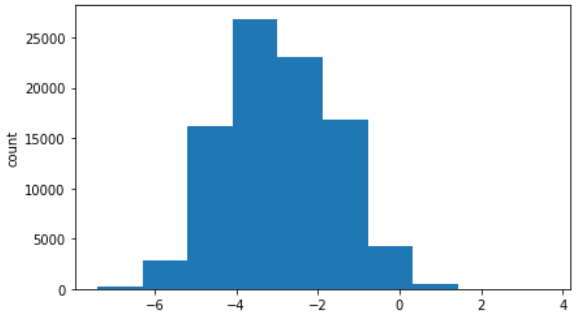


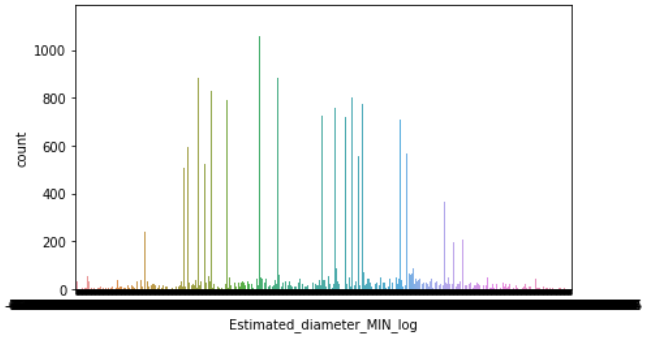


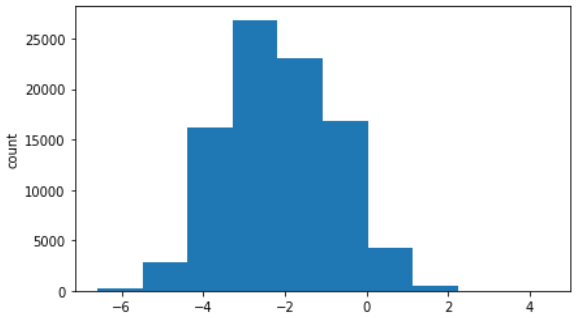


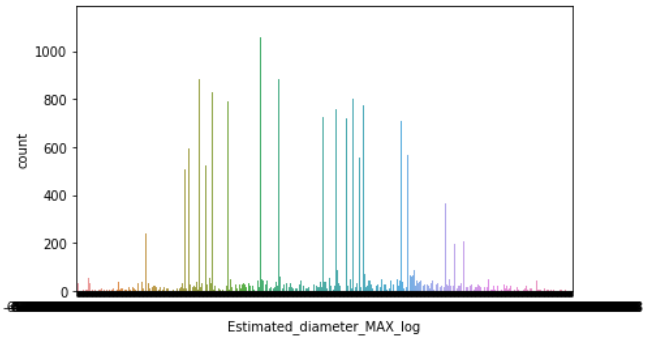


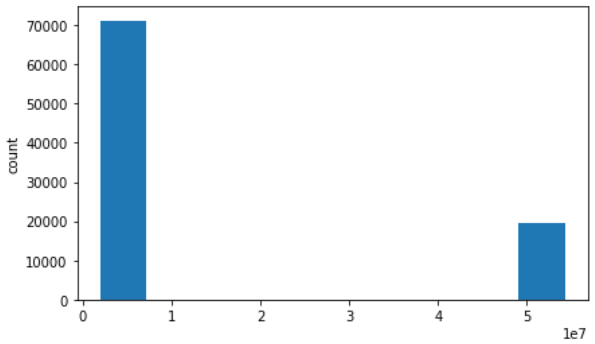


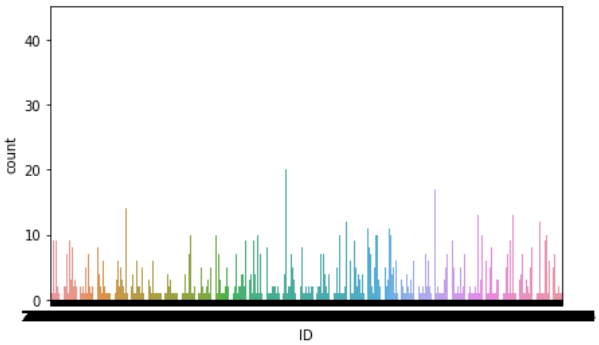






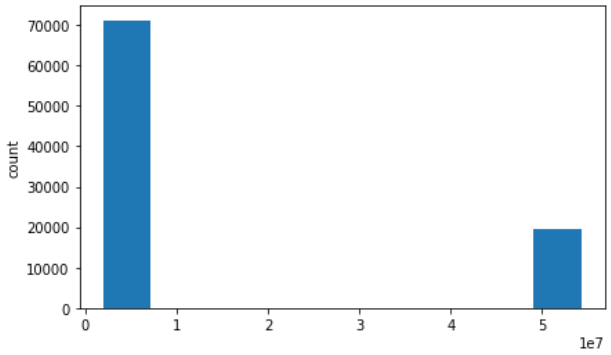


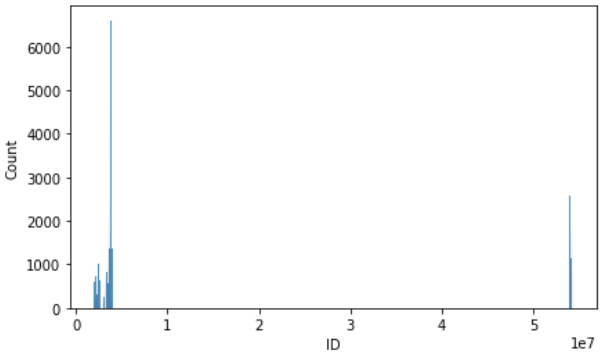


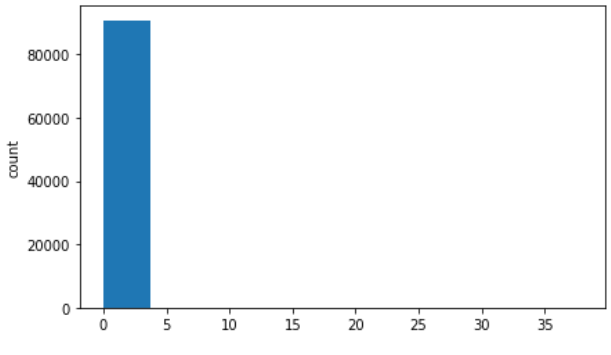


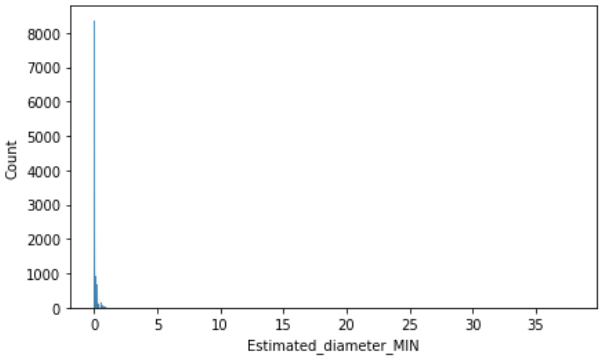
3. <u>Histogram:</u>

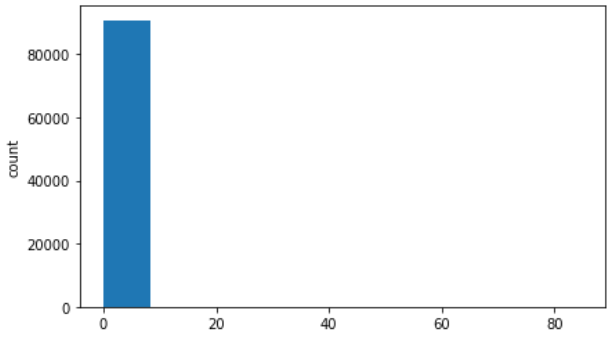
It helps in analyzing the distribution of one or more variables by counting the observations that fall within the discrete bins. It helps in finding the median and frequency of the data. If there are any gaps or outliers in the data they can be easily identified.

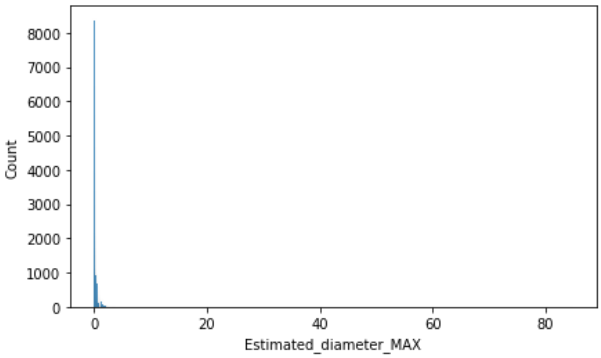

count
1e7

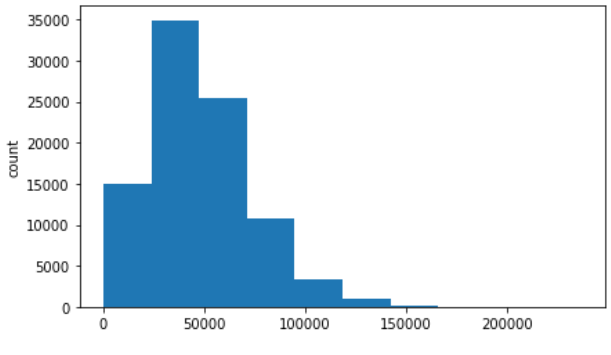

Count
ID
1e7

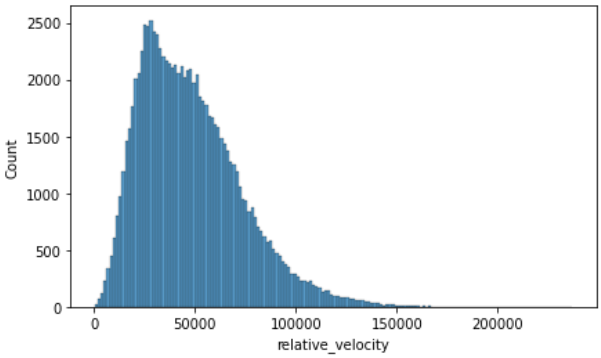

count

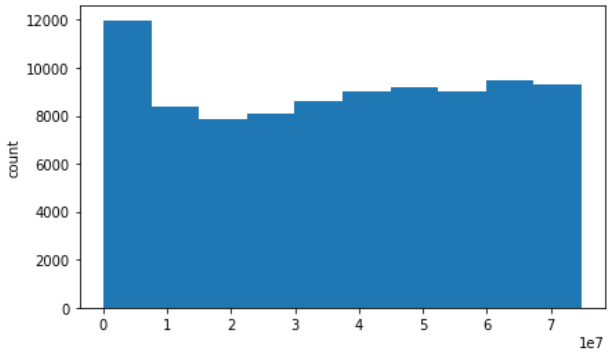

Count
Estimated_diameter_MIN

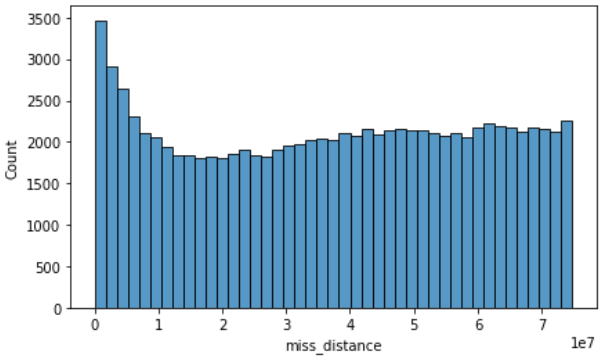

count

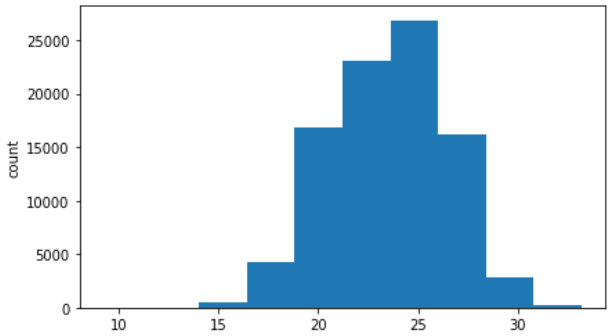

Count
Estimated_diameter_MAX

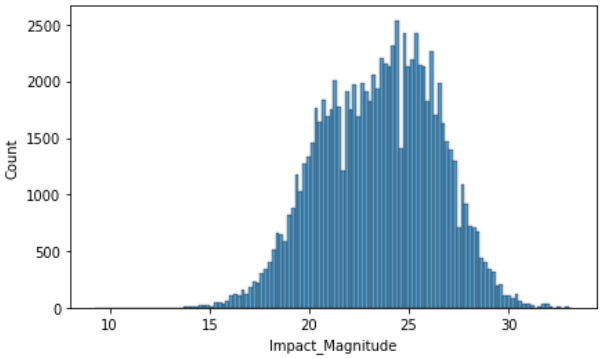

count

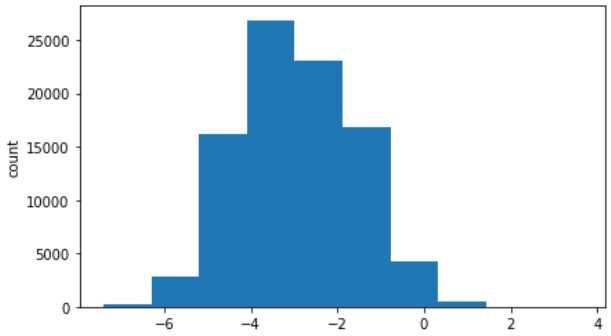

Count
relative_velocity

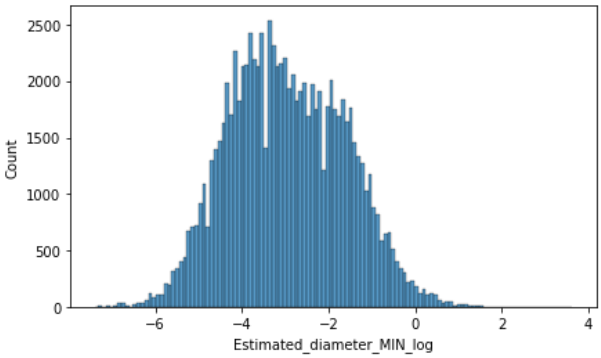

count
1e7

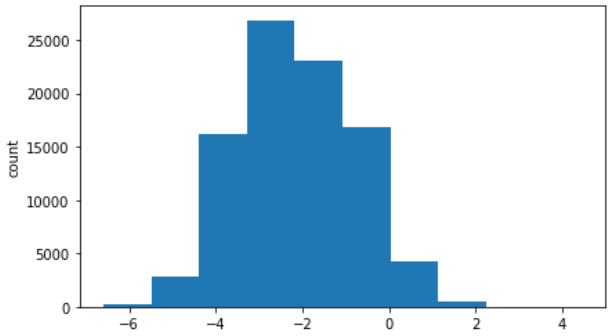

Count
miss_distance
1e7

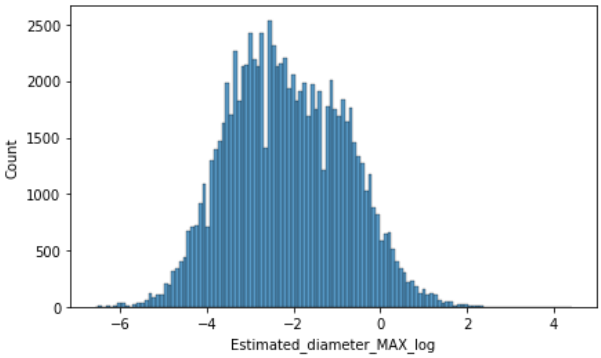

count

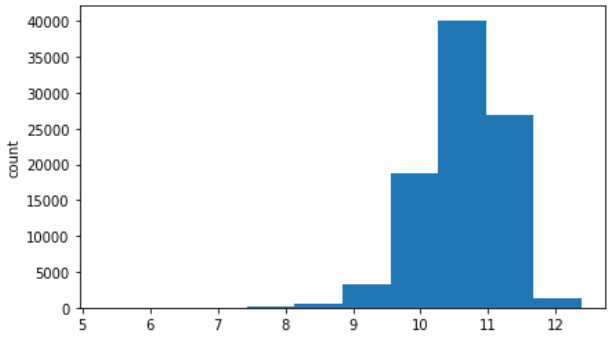

Count
Impact_Magnitude

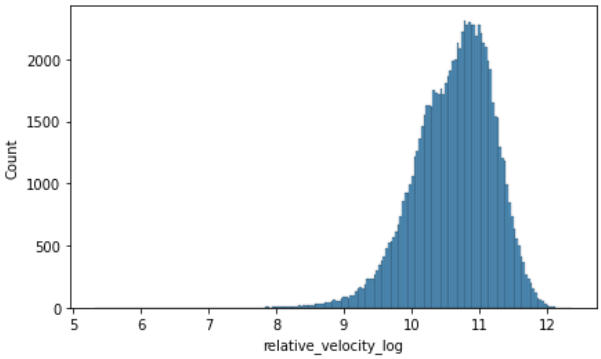

count

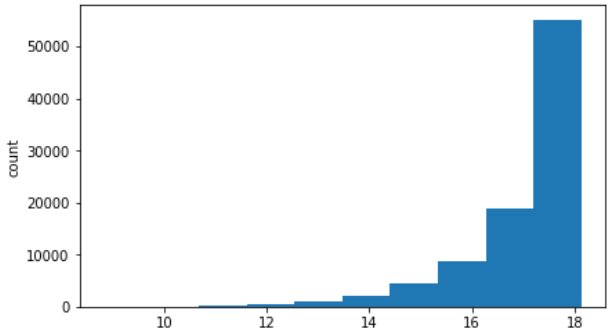

Count
Estimated_diameter_MIN_log

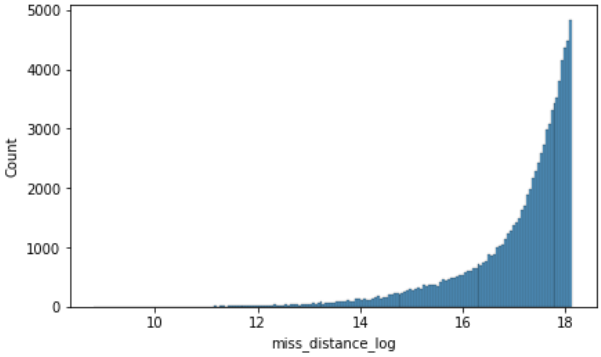

count

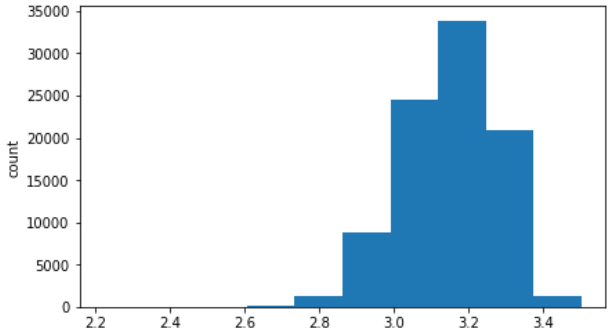

Count
Estimated_diameter_MAX_log

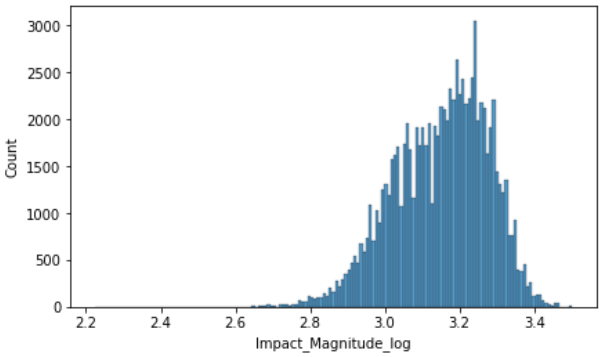

count
40000
35000
30000
25000
20000
15000
10000
5000
0
5
6
7
8
9
10
11
12


Count
2000
1500
1000
500
0
5
6
7
8
9
10
11
12
relative_velocity_log


count
50000
40000
30000
20000
10000
0
10
12
14
16
18

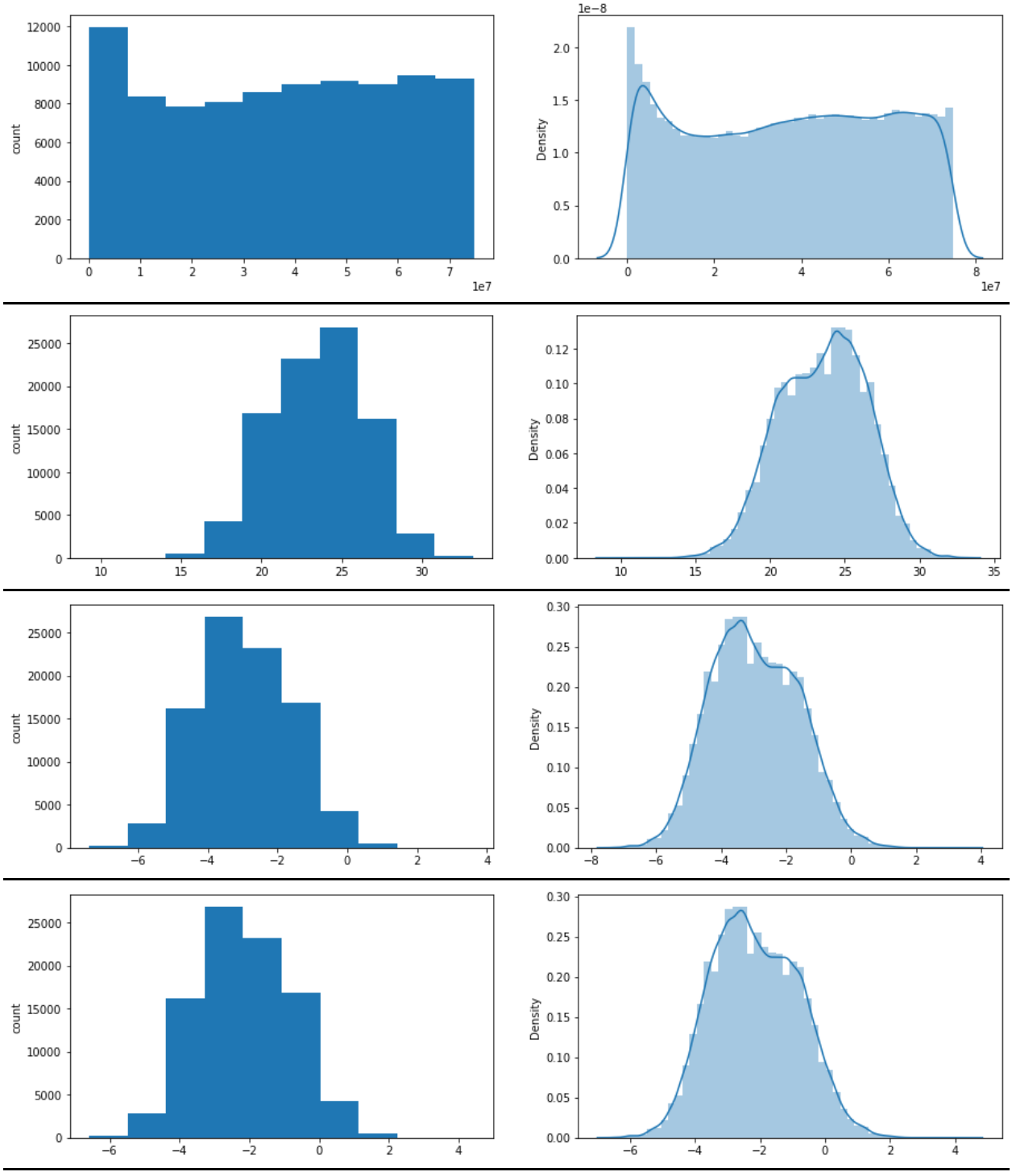

Count
5000
4000
3000
2000
1000
0
10
12
14
16
18
miss_distance_log

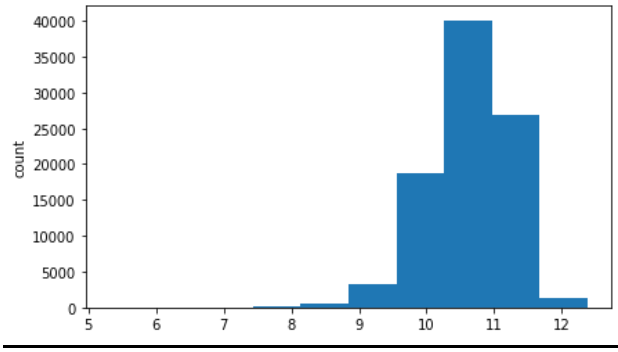

count
35000
30000
25000
20000
15000
10000
5000
0
2.2
2.4
2.6
2.8
3.0
3.2
3.4

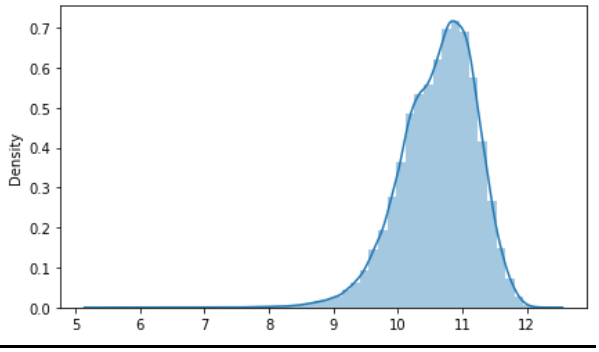

Count
3000
2500
2000
1500
1000
500
0
2.2
2.4
2.6
2.8
3.0
3.2
3.4
Impact_Magnitude_log

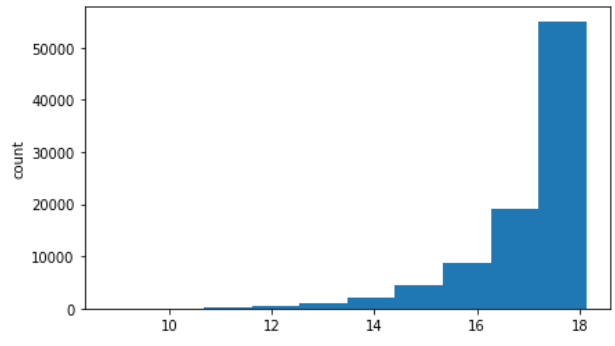
4. <u>Density plot:</u>

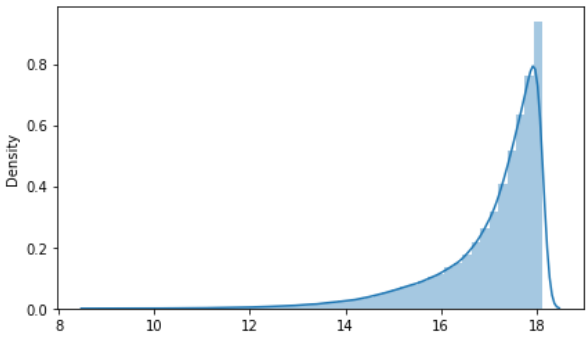
In this plot we can visualize the data distribution of numeric values over a given time period. It will smoothen the distribution values and also noise can be reduced.

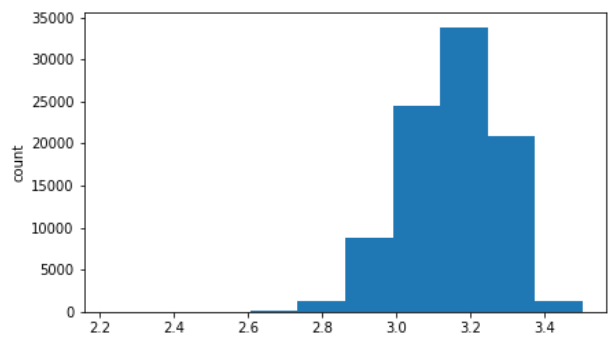


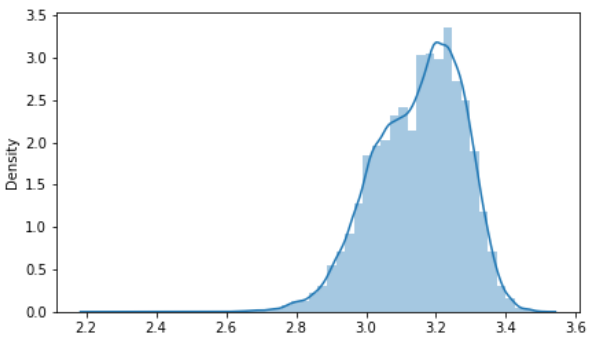

count
40000
35000
30000
25000
20000
15000
10000
5000
0
5
6
7
8
9
10
11
12

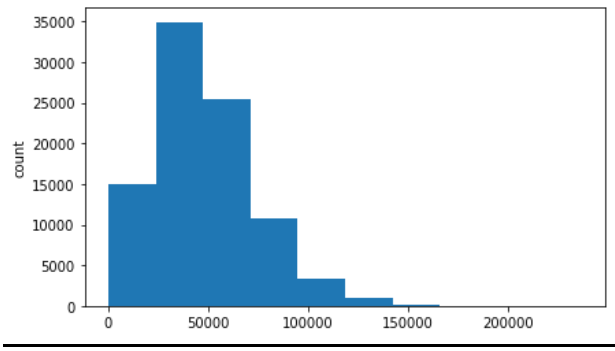

Density
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
5
6
7
8
9
10
11
12

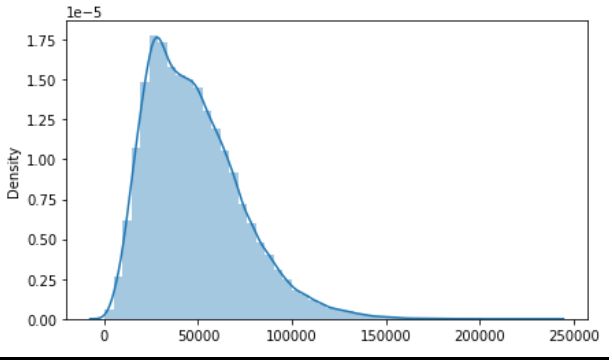

count
50000
40000
30000
20000
10000
0
10
12
14
16
18

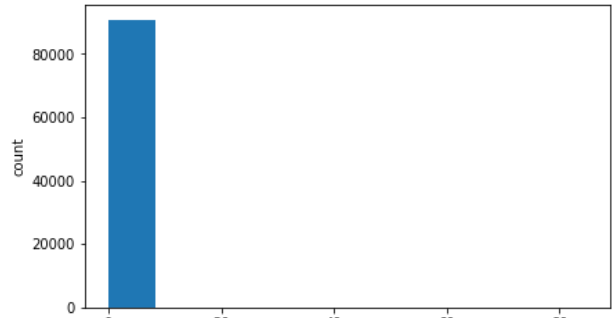

Density
0.8
0.6
0.4
0.2
0.0
8
10
12
14
16
18

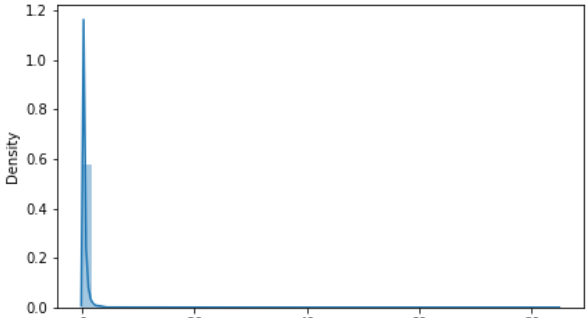

count
35000
30000
25000
20000
15000
10000
5000
0
2.2
2.4
2.6
2.8
3.0
3.2
3.4

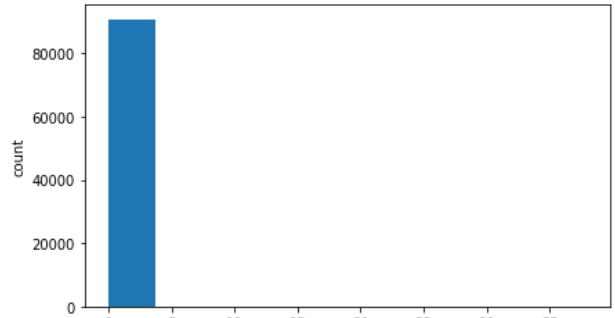

Density
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2.2
2.4
2.6
2.8
3.0
3.2
3.4
3.6

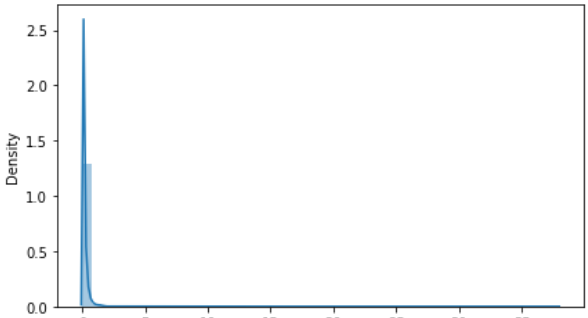

count
35000
30000
25000
20000
15000
10000
5000
0
0
50000
100000
150000
200000

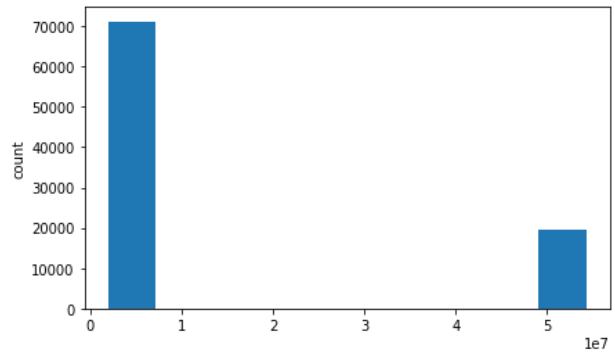

1e−5
Density
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
50000
100000
150000
200000
250000

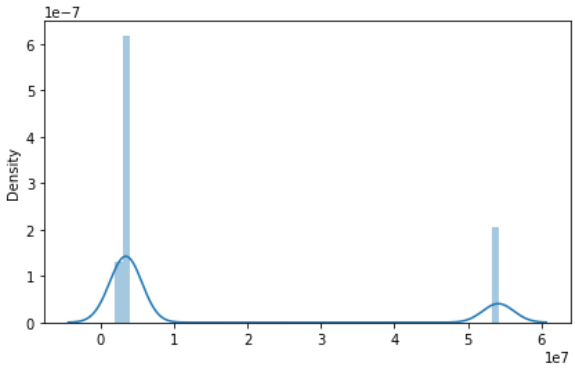

count
0
20000
40000
60000
80000
0
20
40
60
80


Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
20
40
60
80


count
0
20000
40000
60000
80000
0
5
10
15
20
25
30
35


Density
0.0
0.5
1.0
1.5
2.0
2.5
0
5
10
15
20
25
30
35


count
0
10000
20000
30000
40000
50000
60000
70000
0
1
2
3
4
5
1e7

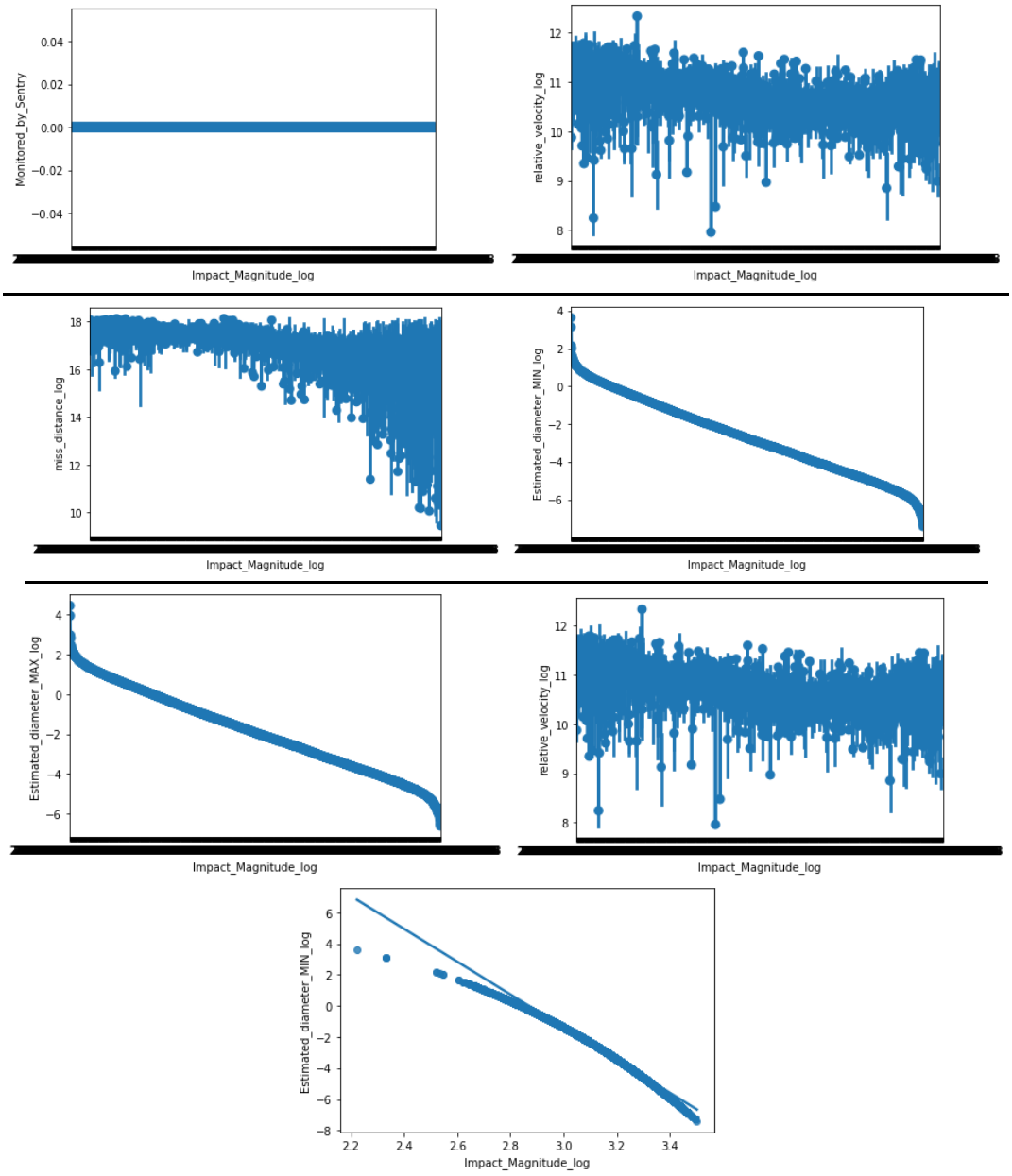

1e−7
Density
0
1
2
3
4
5
6
0
1
2
3
4
5
6
1e7

EDA (Exploratory Data Analysis) Bi-Variate:

Bivariate analysis is based on analyzing the relationship between two variables. This is used for understanding how one variable is related to another variable. Most used approaches are scatter plots, correlation analysis, and contingency tables.

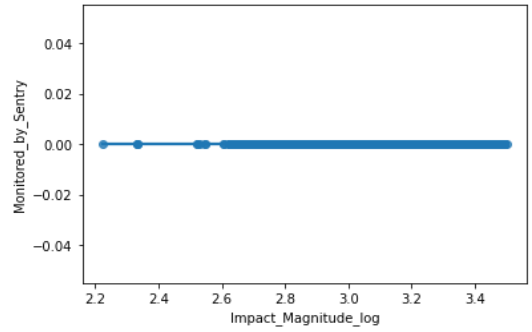
5. **Point-plot:**

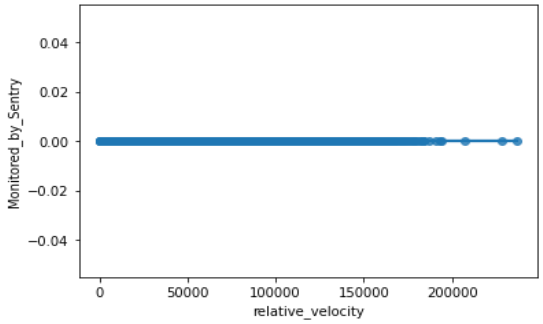
   It estimates the central tendency for a numeric variable by position of the dot. It also provides some indication of the uncertainty around that estimate using error bars. The plot visually groups the number of data points in a data set based on the value of each point. It is useful in focusing comparisons between different levels of one or more categorical variables. Here in the point plot only mean values can be seen.

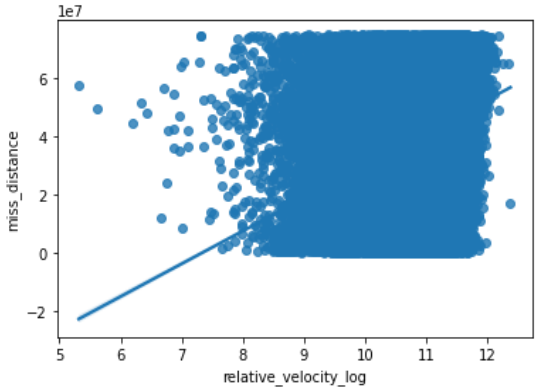


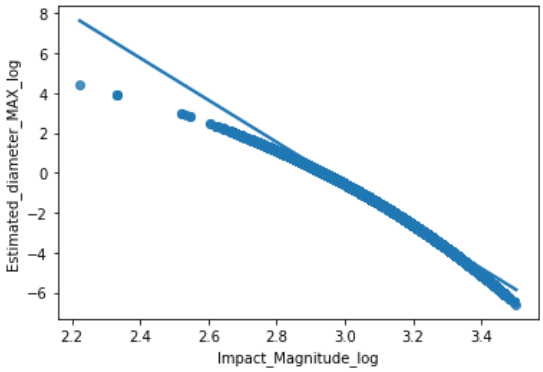
6. Regression plot:

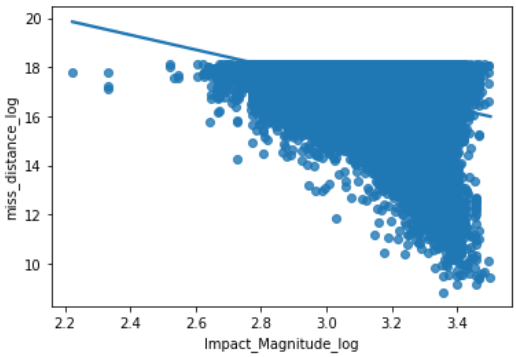
This plot creates a regression line between two parameters which helps to visualize the relationship between them. From this plot we can easily understand which are important factors and which factors can be ignored.

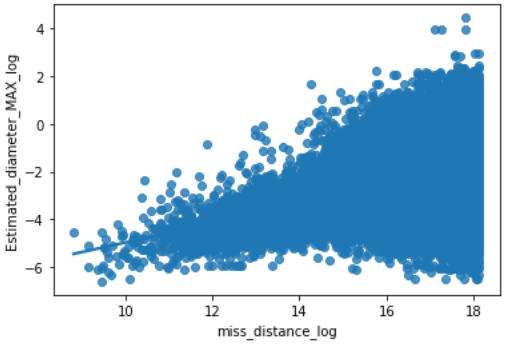


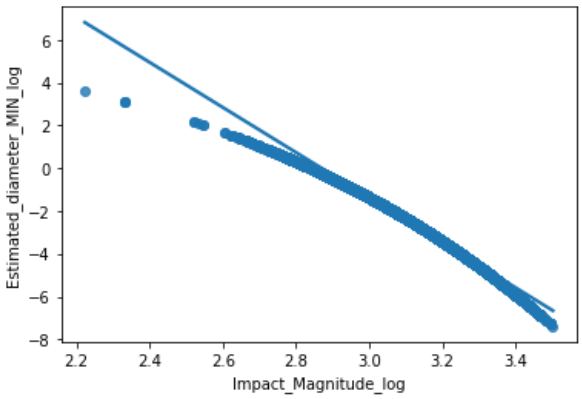






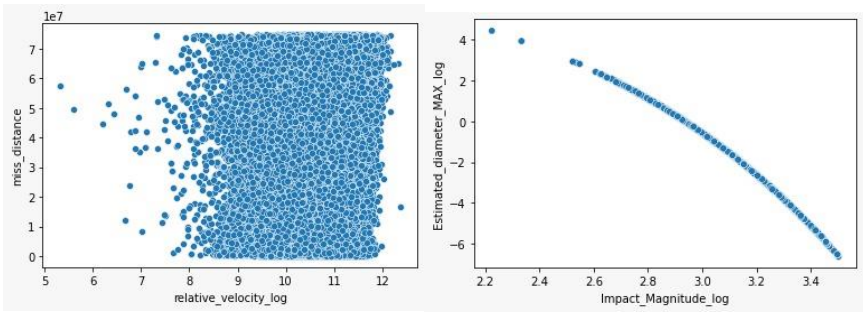


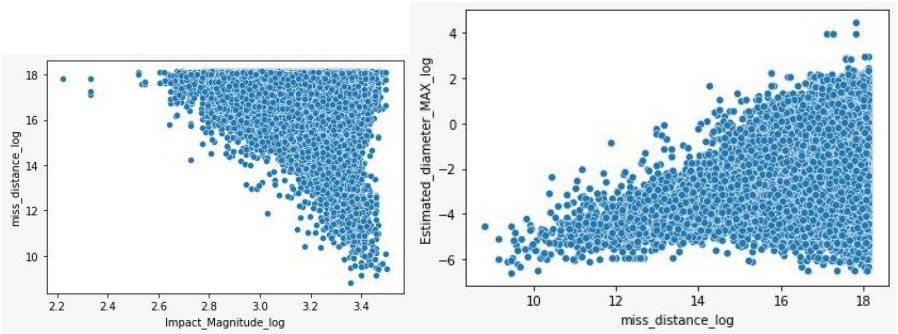


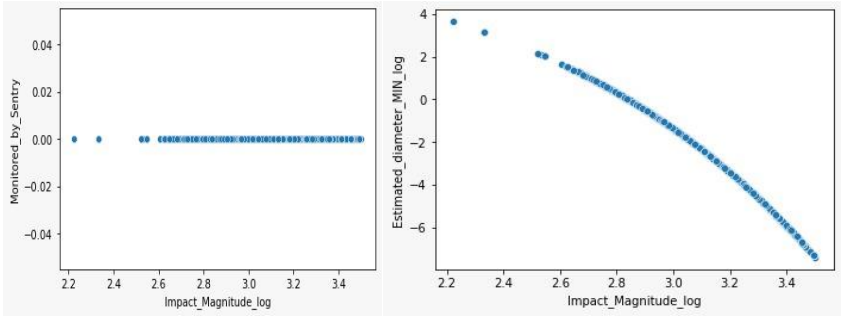


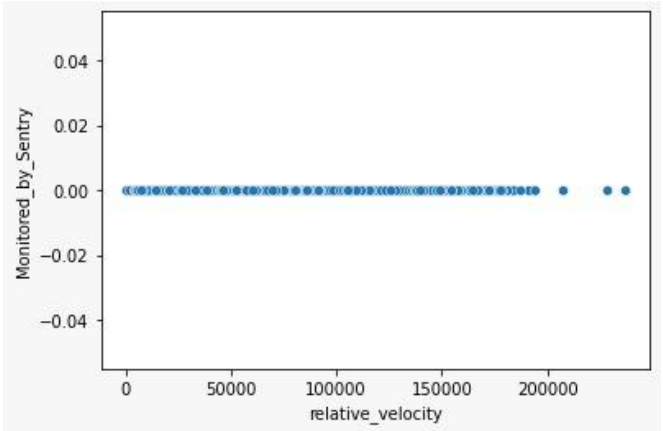
7. **Scatterplot:**

In this plot, dots are used to represent the values for two different numeric variables. Position of each dot indicates values for each individual data point. It helps in observing relationships between variables.



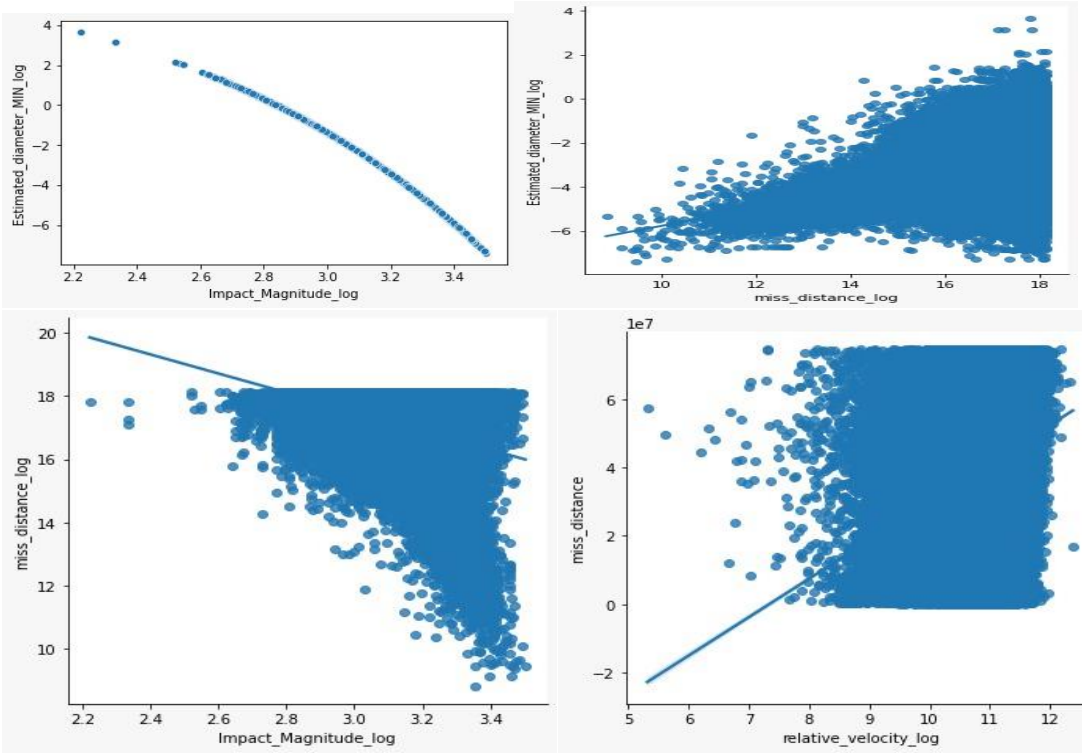


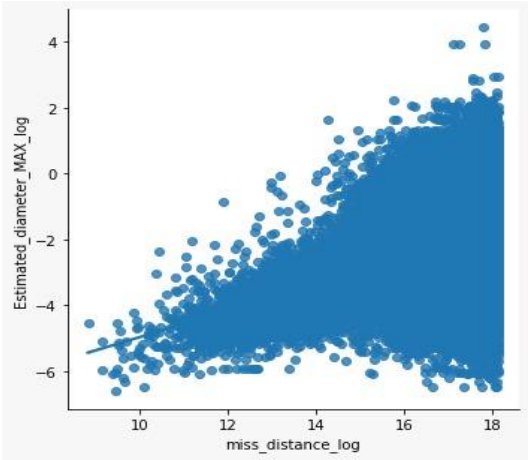


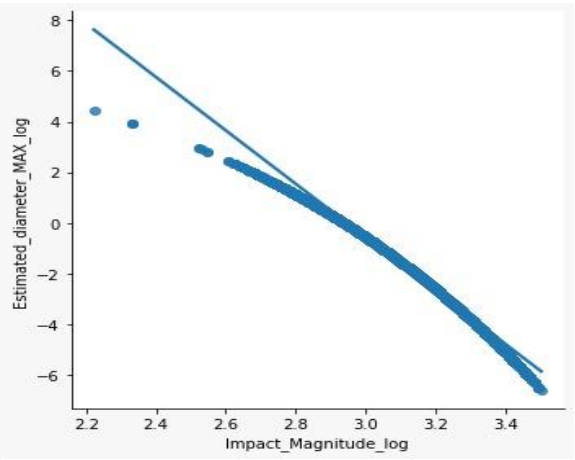


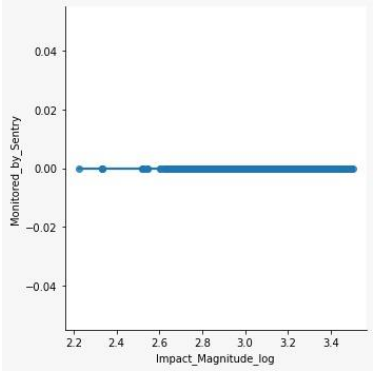
8. Lm plot:

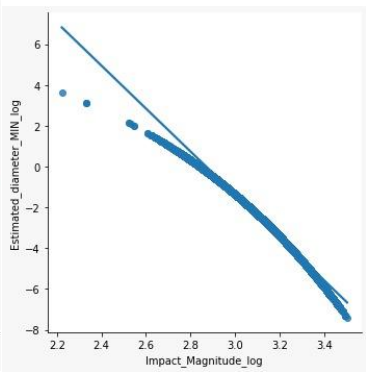
LM-plots are basically scatter plots with overlaid regression lines and combines linear regression model fit (regplot() and FacetGrid()).This interface helps in fitting regression models across conditional subsets of a dataset simple and convenient.

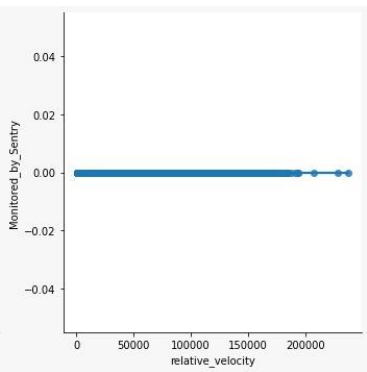



Estimated_diameter_MAX_log
miss_distance_log


Estimated_diameter_MAX_log
Impact_Magnitude_log


Monitored_by_Sentry
Impact_Magnitude_log


Estimated_diameter_MIN_log
Impact_Magnitude_log

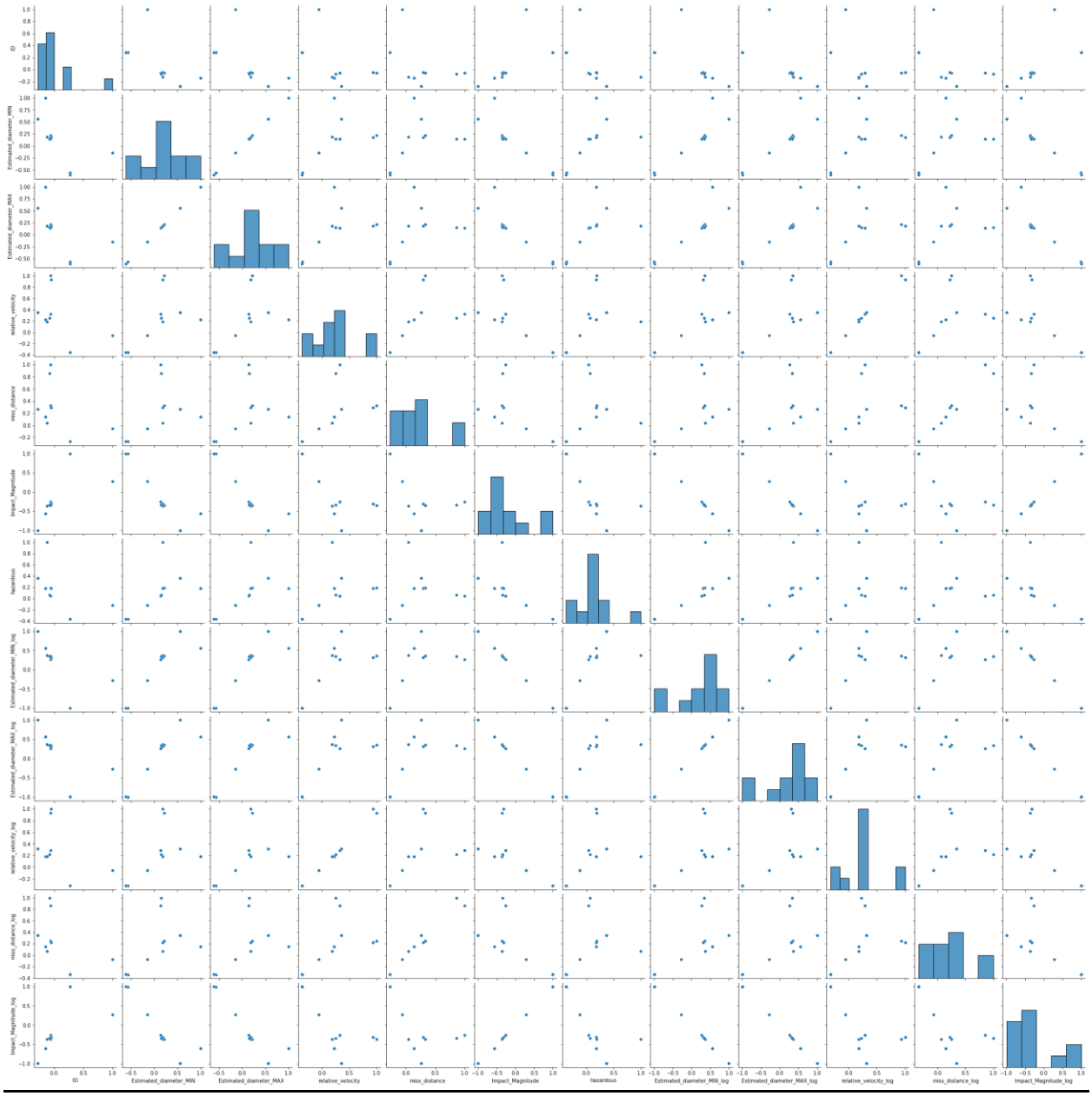

Monitored_by_Sentry
relative_velocity

EDA (Exploratory Data Analysis) Multi-Variate:

Multivariate analysis is based on analyzing more than two variables simultaneously. This is used for understanding the relationships between multiple variables. Most used approaches are factor analysis, cluster analysis, and principal component analysis.

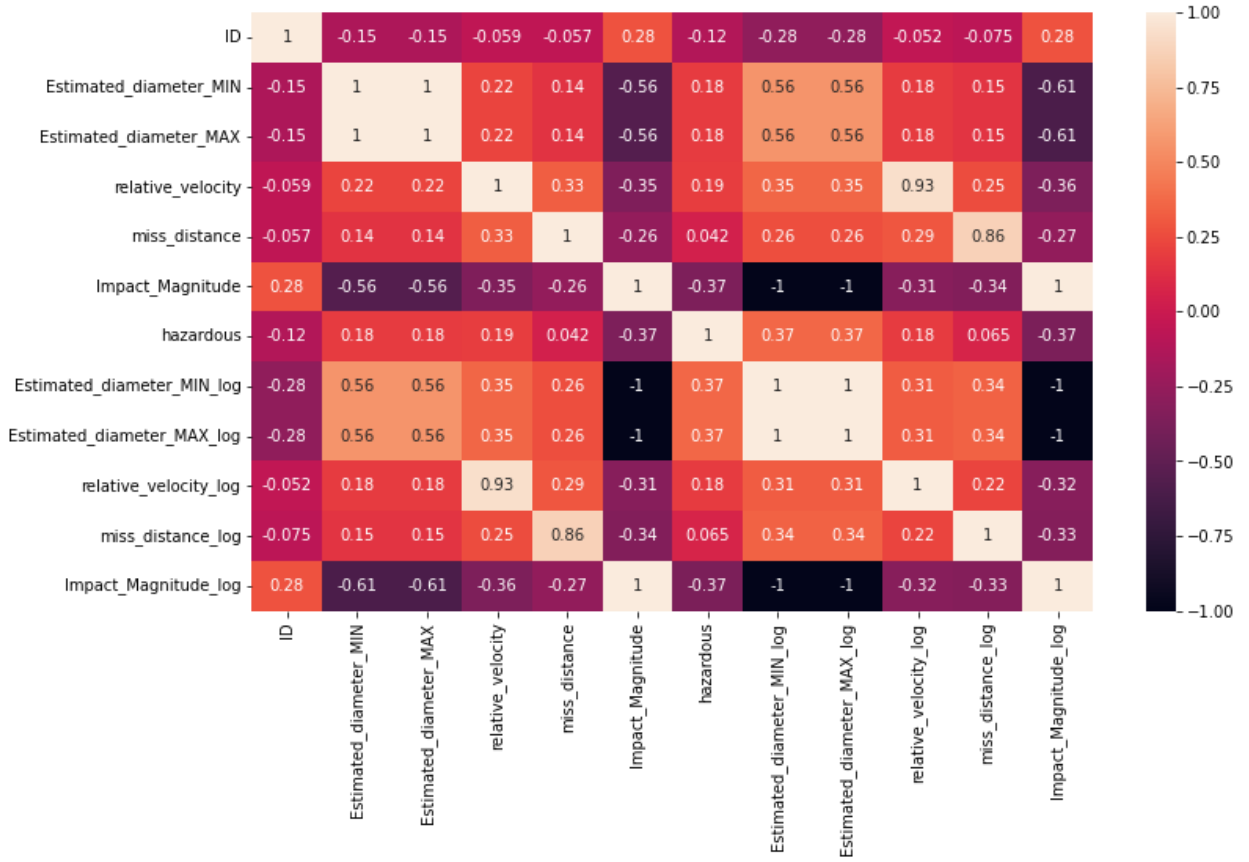
9. Pair-plot:

In this pair-plot, the pairwise relationships between the two variables in the dataset can be plotted. It helps in creating a good visualization to understand the whole data in a single figure. And also univariate distribution can be drawn to show marginal distribution in each column.



10. **Heatmap:**

Plots shows rectangular data as a color-encoded matrix. This heat map helps to show the relationships between two variables, one plotted on each axis. By observing how cell colors change across each axis, you can observe if there are any patterns in value for one or both variables.



Contribution:

The main objective of this project is to determine whether the space objects collide with earth or revolve around the earth's orbit, our goal is to pinpoint the location of impact by any space objects on earth by analyzing this data and determine the number of times a space junk revolves earth, the chances of hitting earth, why the sentry monitoring system is unable to detect some of the junk, etc.

References

https://seaborn.pydata.org/

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

https://www.epa.gov/caddis-vol4/exploratory-data-analysis