RV College of Engineering®

(Autonomous Institution Affiliated to VTU, Belagavi)

Exploring Near Earth Objects

Experiential Learning Report

Submitted by

Sarvagya Kumar (RVCE22RCD029)

K. M. S. Siddharth (RVCE22BCD027)

Vishal H (RVCE22BBT027)

DEPARTMENT OF ARTIFICIAL INTELLIGENCE

Submitted to

PROF. SOMESH NANDI

Contents

Introduction

- **Numpy:**

  Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

- **Pandas:**

  Pandas is an open-source library that is built on top of NumPy library. It is a Python package that offers various data structures and operations for manipulating numerical data and time series. It is mainly popular for importing and analyzing data much easier.

- **Matplotlib:**

  Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy.

- **Seaborn:**

  Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Objectives :-

The objectives of this EDA are to explore and gain insights into the NASA Near Earth Objects dataset. Specifically, the code aims to:

- Read in the data from a CSV file (neo.csv).
- Check the head of the data to ensure proper loading.
- Check for any missing values in the data.
- Plot the distribution of the estimated minimum and maximum diameters of the near-earth objects.
- Plot the correlation between selected variables, including the estimated diameter and relative velocity.
- Plot the number of hazardous and non-hazardous near-earth objects.
- Plot the relationship between the hazardous variable and the absolute magnitude of the near-earth objects.
- To explore and visualize the data in different ways, which can help us identify patterns, trends, and potential issues or outliers in the data.
- To form graphical and statistical representation of the obtained data.
- The data set used in this EDA is: https://www.kaggle.com/datasets/sameepvani/nasa-nearest-earth-objects

The Dataset

- The used dataset (neo.csv) contains information about the near-Earth objects (NEOs) that have been discovered by NASA.
- NEOs are comets and asteroids that orbit the Sun and come close to the Earth's orbit. This dataset includes the information on these objects that have come within 0.05 astronomical units (7.5 million km) of the Earth's orbit.
- The dataset contains 90836 observations and 10 columns with information on the objects' physical characteristics, such as their estimated diameter, absolute magnitude, and their closest approach to Earth (miss distance). It also includes information on the objects' orbit, such as their eccentricity and inclination. Additionally, it includes information on whether the object is considered hazardous or not.
- This dataset can be used for various purposes, including studying the physical characteristics and behavior of NEOs, predicting the potential hazards associated with these objects, and developing strategies to mitigate the risks.

Id	Name	Est_diameter	Est_diameter	Relative_velo	Miss_distanc	Orbiting_bod	Sentry_objec	Absolute_ma	Hazardous
3005806	(1983 LC)	0.34	0.77	70007.69	10334476.51	Earth	False	19.45	True
3005806	(1983 LC)	0.34	0.77	50377.61	26618810.03	Earth	False	19.45	True
3092100	(1986 NA)	0.28	0.64	51810.08	70052459.21	Earth	False	19.85	False
3092101	(1988 NE)	0.4	0.9	22815.74	65063306.32	Earth	False	19.1	False
3001703	(1989 AZ)	0.32	0.71	54203.68	30989510.07	Earth	False	19.6	False
3001703	(1989 AZ)	0.32	0.71	56164.41	20219801.81	Earth	False	19.6	False
3002856	(1991 GO)	0.27	0.59	85339.3	27025626.03	Earth	False	20	True
3002856	(1991 GO)	0.27	0.59	99664.49	46888927.92	Earth	False	20	True
3003147	(1991 TF3)	0.37	0.82	44613.81	14798062.04	Earth	False	19.29	False
3003147	(1991 TF3)	0.37	0.82	49837.22	12823591.5	Earth	False	19.29	False
3003147	(1991 TF3)	0.37	0.82	51071.38	13057764.79	Earth	False	19.29	False
3005816	(1991 VG)	0.01	0.01	21056.79	25188081.35	Earth	False	28.3	False
3005816	(1991 VG)	0.01	0.01	4239.75	8498417.07	Earth	False	28.3	False
3005816	(1991 VG)	0.01	0.01	7381.97	7047199.07	Earth	False	28.3	False
3005816	(1991 VG)	0.01	0.01	40561.33	52151023.48	Earth	False	28.3	False
3005816	(1991 VG)	0.01	0.01	48107.01	62939243.36	Earth	False	28.3	False
3005831	(1992 JD)	0.03	0.06	47458.13	58091517.12	Earth	False	25	False
3005831	(1992 JD)	0.03	0.06	39088.29	46449877.95	Earth	False	25	False
3005831	(1992 JD)	0.03	0.06	26867.13	13484205.97	Earth	False	25	False
3005831	(1992 JD)	0.03	0.06	25339.08	2674190.4	Earth	False	25	False
3005831	(1992 JD)	0.03	0.06	31071.82	34390590.57	Earth	False	25	False
3005831	(1992 JD)	0.03	0.06	45101.64	51355538.32	Earth	False	25	False
3005851	(1992 SZ)	0.26	0.59	46463.36	67973572.18	Earth	False	20.01	False
3092114	(1993 DA)	0.01	0.03	36440.92	39809124.32	Earth	False	26.4	False
3092114	(1993 DA)	0.01	0.03	29039.89	25513924.76	Earth	False	26.4	False
3092114	(1993 DA)	0.01	0.03	39684.38	38626837.63	Earth	False	26.4	False
3092114	(1993 DA)	0.01	0.03	26295.18	29080058.92	Earth	False	26.4	False
3092114	(1993 DA)	0.01	0.03	23860.24	5811568.4	Earth	False	26.4	False
3092114	(1993 DA)	0.01	0.03	42087.87	48402119.36	Earth	False	26.4	False

The Dataset (neo.csv)

Code & Methodology

- **Importing the Required Libraries :** These are the necessary libraries being imported for data analysis and visualization.

```
# import necessary libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

- **Read the dataset :** We will first read the dataset (neo.csv) which a .csv file i.e comma separated values. A CSV (comma-separated values) file is a text file that has a specific format which allows data to be saved in a table structured format.

```
# read in the data
neo = pd.read_csv('neo.csv')
```

- **Checking the head of the data :** It is a common step in exploratory data analysis (EDA) and is a quick and easy way to get a preliminary understanding of the dataset and identify any initial patterns or trends.

```
# check the head of the data
print(neo.head())
```

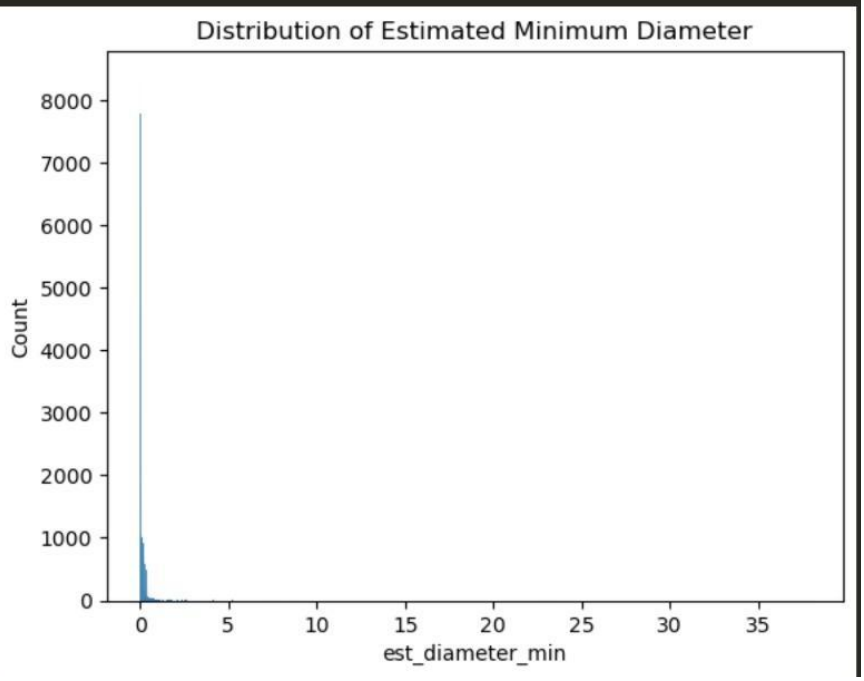
- **Checking for Missing Values :** By checking for **missing values** in EDA, we can get an idea of the completeness and quality of the dataset, and determine how to handle missing values approximately to ensure accurate and reliable analysis.

```
print(neo.isnull().sum())
✓ 0.0s
```

- We are creating a *histogram plot of the distribution of the estimated Minimum diameter* of the near-earth objects (NEOs) dataset. We plotted the graph to determine how many objects have a small diameter, how many have a large diameter, and where the majority of objects fall in terms of diameter.

```
# plot the distribution of the estimated diameter
sns.histplot(data=neo, x='est_diameter_min')
plt.title('Distribution of Estimated Minimum Diameter')
plt.show()
```

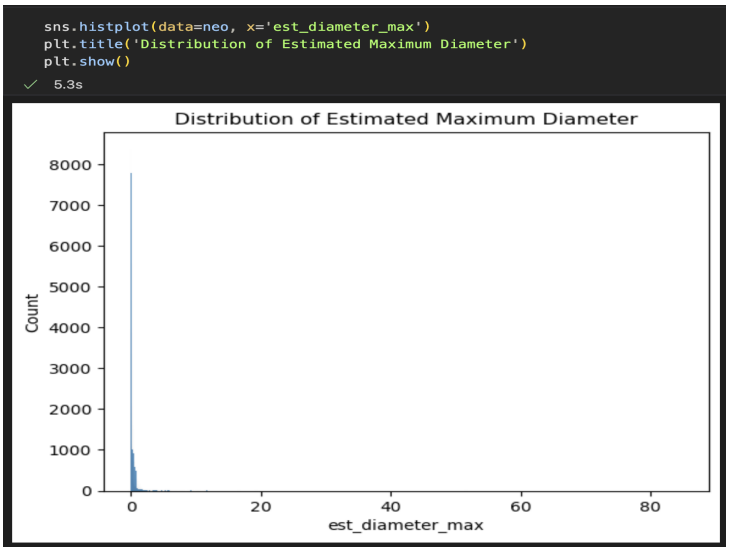
✓ 5.3s



- Similarly we plot the distribution of the estimated Maximum diameter of the near-earth objects (NEOs) dataset

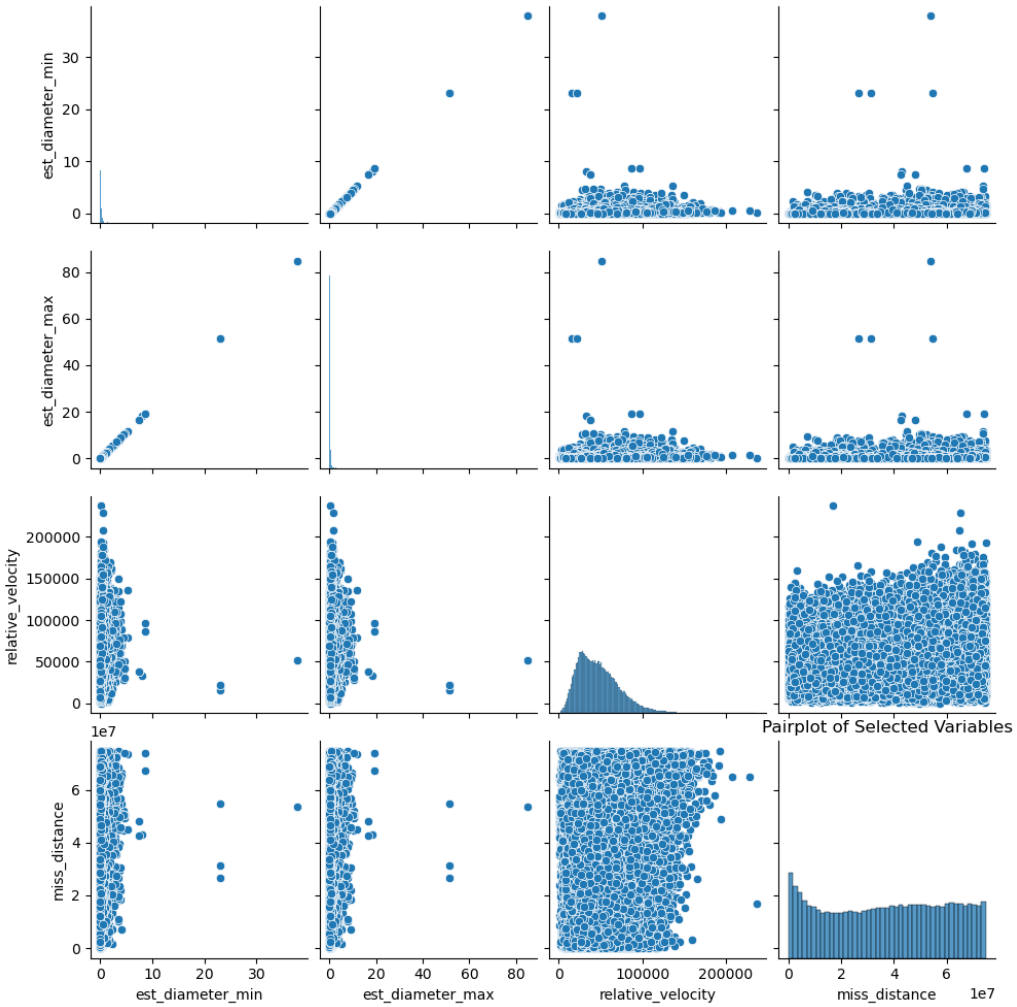
```
sns.histplot(data=neo, x='est_diameter_max')
plt.title('Distribution of Estimated Maximum Diameter')
plt.show()
```

✓ 5.3s



- We are now plotting a Pairplot between the selected variables in the "neo.csv" dataset. The pairplot is a grid of scatterplots that shows the relationship between each variable and the other variables in the dataset. This visualization helps to identify patterns and relationships between variables, such as positive or negative correlation. It can also help to identify potential outliers or anomalies in the data.

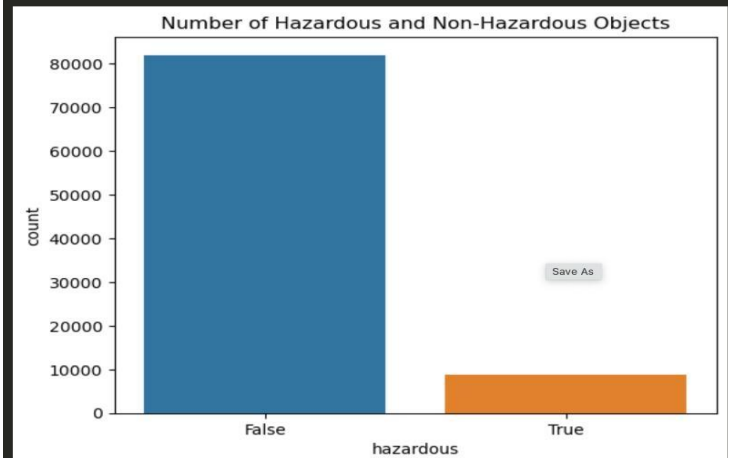
```
# plot the correlation between variables
sns.pairplot(data=neo[['est_diameter_min', 'est_diameter_max', 'relative_velocity', 'miss_distance']])
plt.title('Pairplot of Selected Variables')
plt.show()
#This code generates a pair plot using the Seaborn library to visualize the pairwise relationships between four
selected variables from the "near_earth_objects.csv" dataset: est_diameter_min, est_diameter_max, relative_velocity,
and miss_distance.
```



- We are now Plotting a *count plot of hazardous and non-hazardous objects*.It is significant because it helps to visualize the balance or imbalance of the classes in the dataset. It can help us understand the proportion of hazardous objects in the dataset, which may be useful information for decision making.

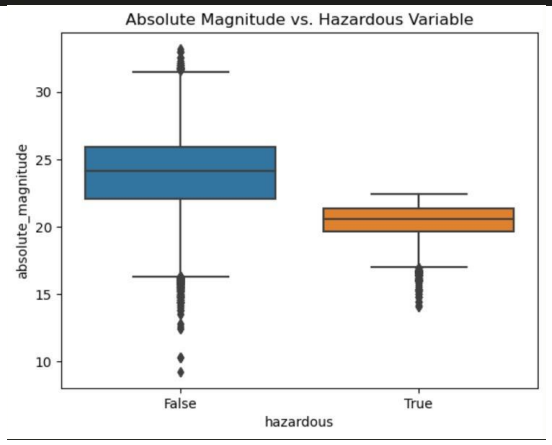
```
# plot the number of hazardous and non-hazardous objects
sns.countplot(data=neo, x='hazardous')
plt.title('Number of Hazardous and Non-Hazardous Objects')
plt.show()
```

✓ 0.0s



- The Box plot of absolute magnitude vs. hazardous variable is significant because it helps to understand the relationship between these two variables. By plotting these two variables, we can see if there is any **correlation** between the brightness of an object and its **potential hazard** to Earth. Additionally, this plot can help in identifying any outliers or **unusual patterns** in the data.

```
# plot the absolute magnitude vs. hazardous variable
sns.boxplot(data=neo, x='hazardous', y='absolute_magnitude')
plt.title('Absolute Magnitude vs. Hazardous Variable')
plt.show()
```

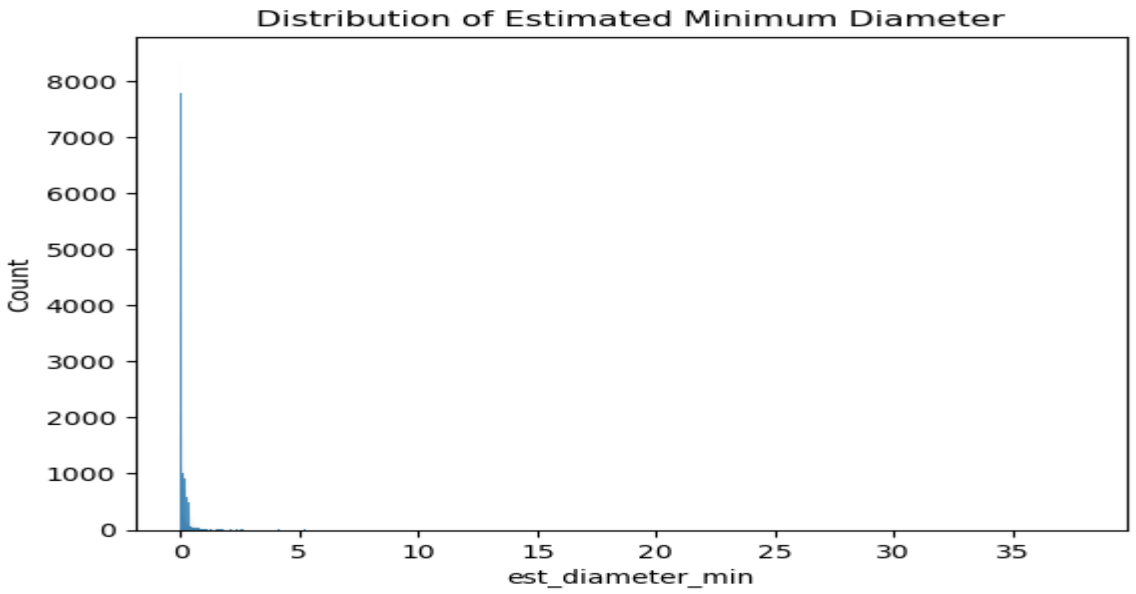


Complete Code :-

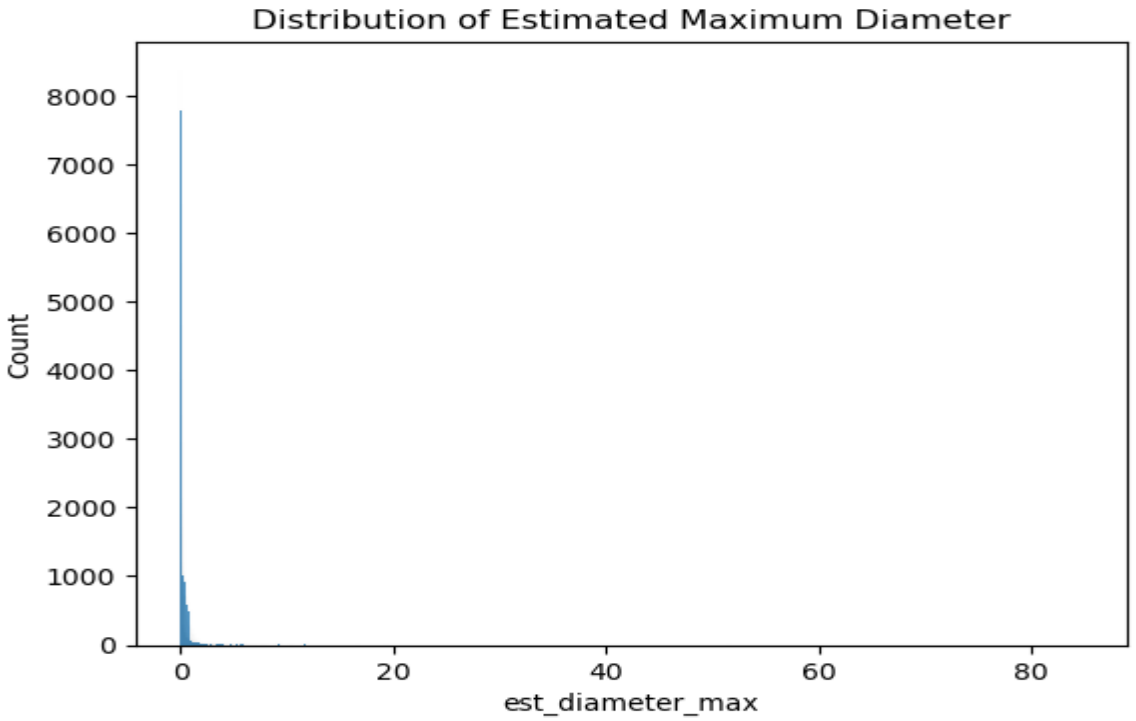
```
# import necessary libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# read in the data
neo = pd.read_csv('neo.csv')

# check the head of the data
print(neo.head())

# check for missing values : This checks if there are any missing values in the
DataFrame 'neo' and prints the sum of missing values for each column.
print(neo.isnull().sum())

# plot the distribution of the estimated diameter
sns.histplot(data=neo, x='est_diameter_min')
plt.title('Distribution of Estimated Minimum Diameter')
plt.show()

neo.describe()

sns.histplot(data=neo, x='est_diameter_max')
plt.title('Distribution of Estimated Maximum Diameter')
plt.show()

# plot the correlation between variables
sns.pairplot(data=neo[['est_diameter_min', 'est_diameter_max',
'relative_velocity', 'miss_distance']])
plt.title('Pairplot of Selected Variables')
plt.show()
#This code generates a pair plot using the Seaborn library to visualize the
pairwise relationships between four selected variables from the
"near_earth_objects.csv" dataset: est_diameter_min, est_diameter_max,
relative_velocity, and miss_distance.

# plot the number of hazardous and non-hazardous objects
sns.countplot(data=neo, x='hazardous')
plt.title('Number of Hazardous and Non-Hazardous Objects')
plt.show()

# plot the absolute magnitude vs. hazardous variable
sns.boxplot(data=neo, x='hazardous', y='absolute_magnitude')
plt.title('Absolute Magnitude vs. Hazardous Variable')
plt.show()
```

Plots :-

- *Histogram plot of the distribution of the estimated Minimum diameter* of the near-earth objects (NEOs) dataset.



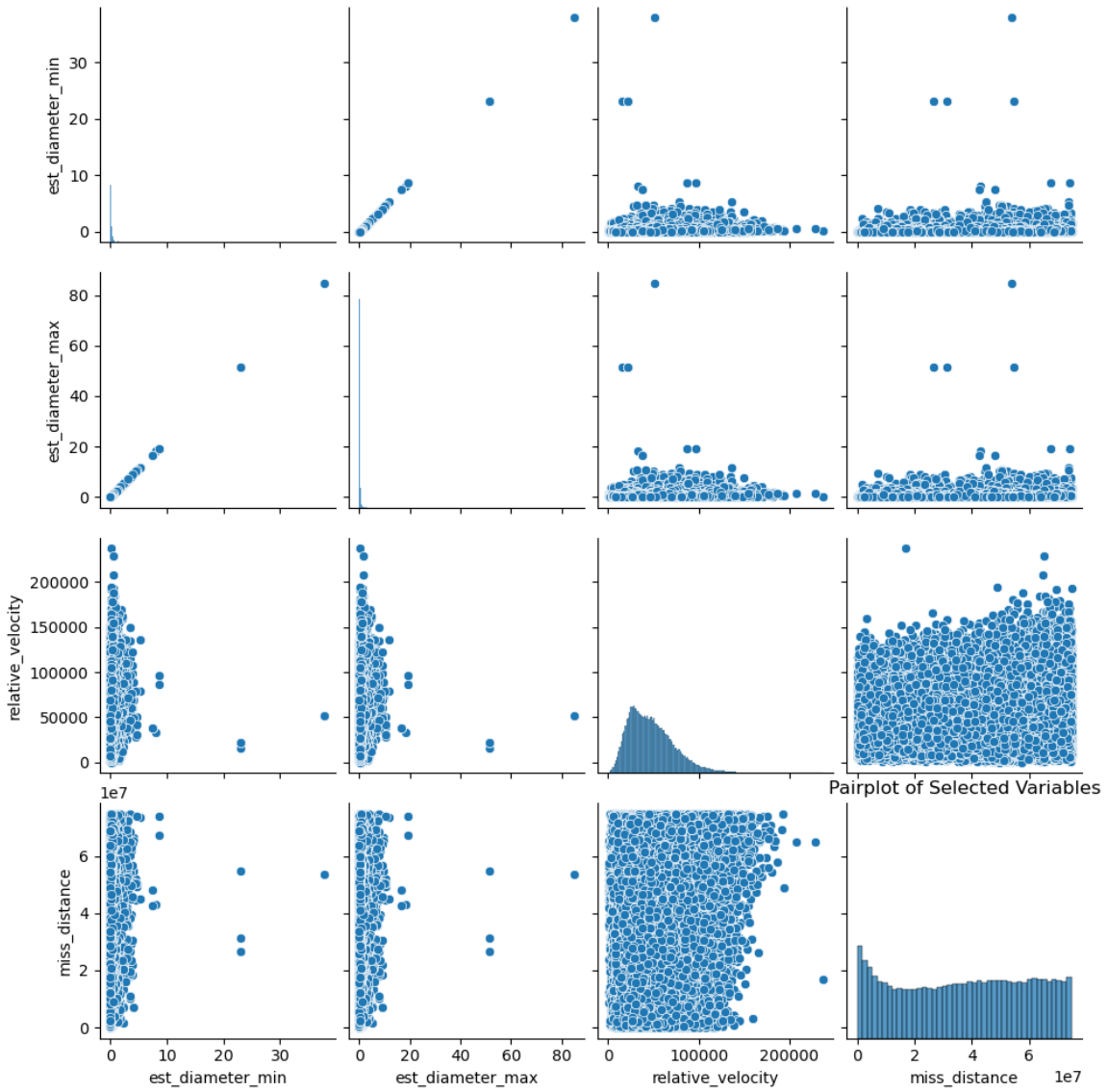
- *Histogram plot of the distribution of the estimated Maximum diameter* of the near-earth objects (NEOs) dataset.



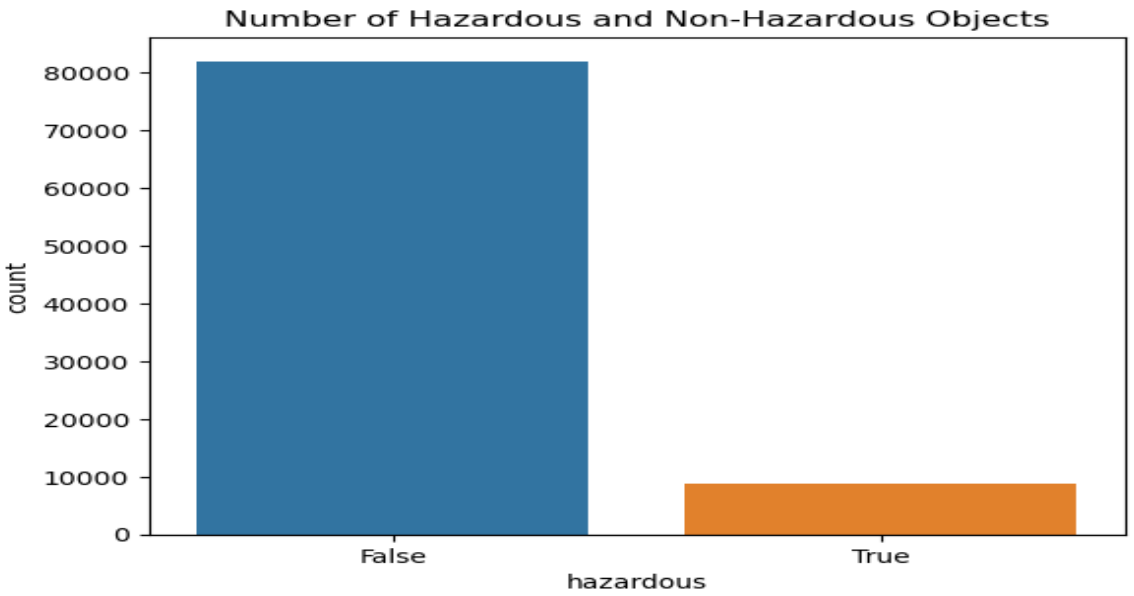
- **Pairplot :** This plot shows the pairwise relationships and distributions between the selected variables, including 'est_diameter_min', 'est_diameter_max', 'relative_velocity', and 'miss_distance'. It allows us to visually identify any correlations or patterns between the variables, and potentially identify any outliers or anomalies in the data.

  ➢ The scatter plot between the estimated minimum diameter and the miss distance shows a weak negative correlation, indicating that as the estimated minimum diameter of the object increases, the miss distance decreases slightly.

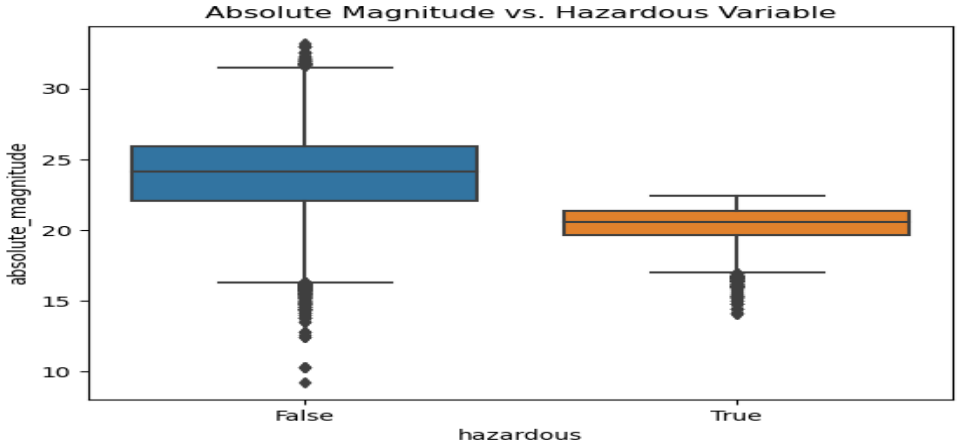
  ➢ The scatter plot between the relative velocity and the miss distance shows a weak positive correlation, indicating that as the relative velocity of the object increases, the miss distance also tends to increase slightly.



- **Count plot of Hazardous and Non - Hazardous NEOs** :- From the plot, we can see that the majority of the objects are non-hazardous. This is important information for researchers and policymakers as they can prioritize their efforts and resources towards monitoring and tracking the more dangerous objects. Additionally, the plot highlights the imbalanced nature of the dataset, which is an important consideration for machine learning models that will be trained on this data.

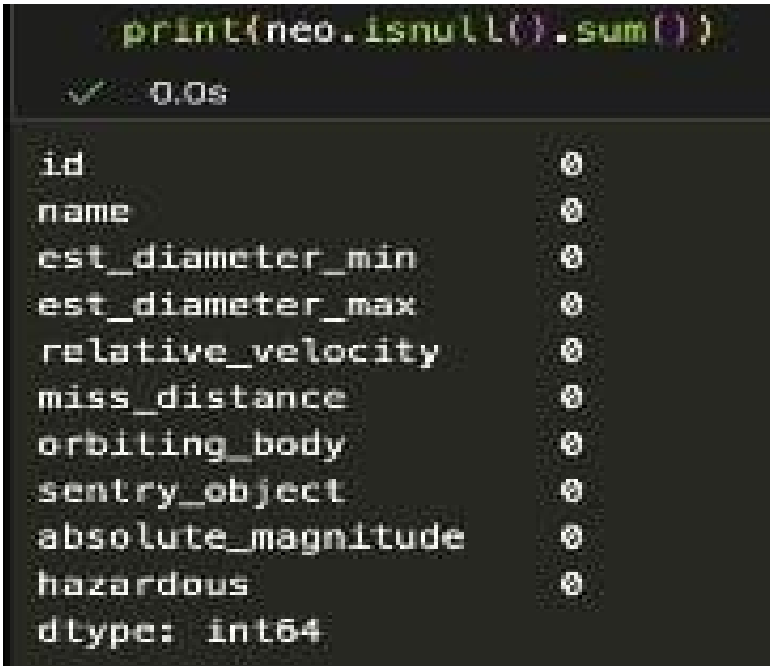


- **BoxPlot of Abs. Magnitude and Hazardous Variable** :- The box plot comparing the hazardous and non-hazardous near earth objects by their absolute magnitude shows that the hazardous objects tend to have a higher absolute magnitude on average than the non-hazardous ones. This suggests that there may be a correlation between the object's size and its potential hazard to Earth.

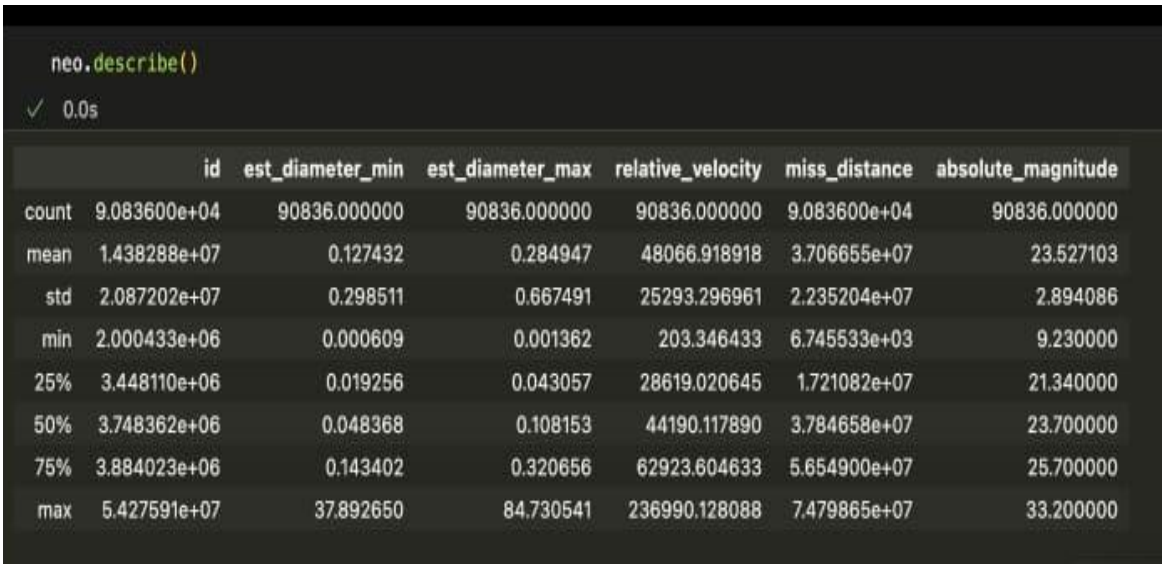


Findings

- No data is **missing** i.e. all the data whether correct or incorrect is placed in the dataset. Found using (neo.isnull().sum()).



- Some asteroids with very large diameter (84 Km) appear in the histogram which could be due to error in the measurement/entry of data. Found using neo.describe().

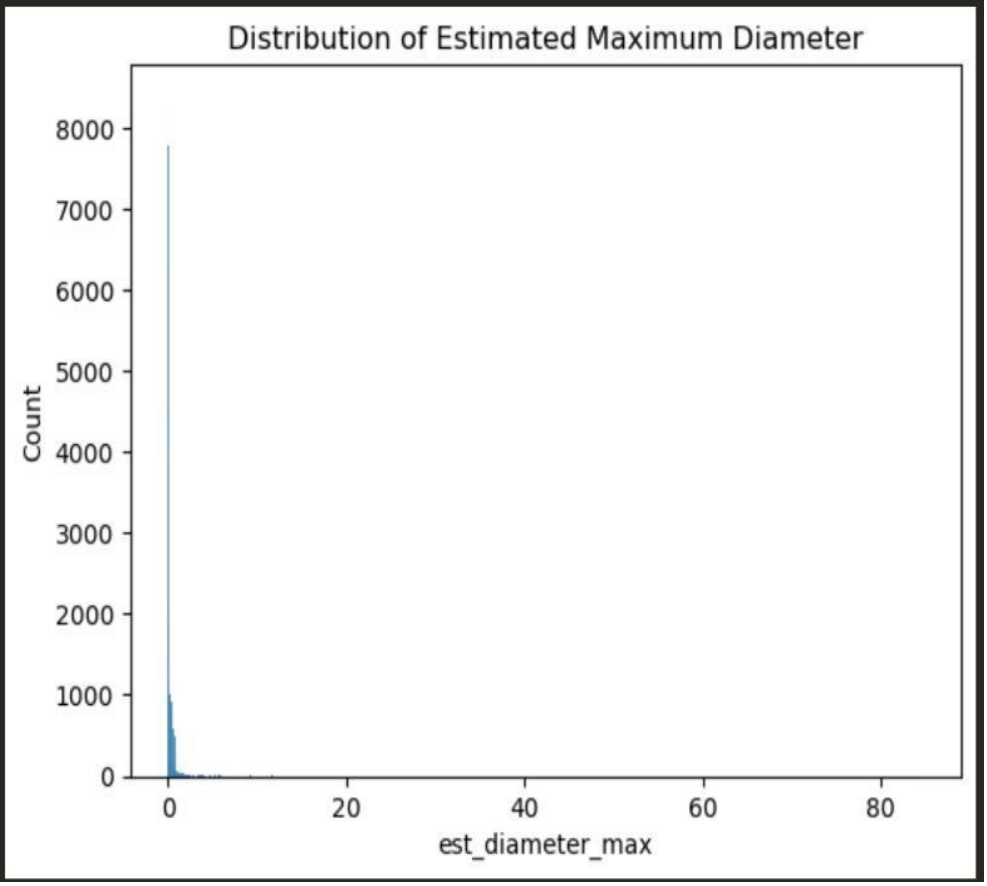


	id	est_diameter_min	est_diameter_max	relative_velocity	miss_distance	absolute_magnitude
count	9.083600e+04	90836.000000	90836.000000	90836.000000	9.083600e+04	90836.000000
mean	1.438288e+07	0.127432	0.284947	48066.918918	3.706655e+07	23.527103
std	2.087202e+07	0.298511	0.667491	25293.296961	2.235204e+07	2.894086
min	2.000433e+06	0.000609	0.001362	203.346433	6.745533e+03	9.230000
25%	3.448110e+06	0.019256	0.043057	28619.020645	1.721082e+07	21.340000
50%	3.748362e+06	0.048368	0.108153	44190.117890	3.784658e+07	23.700000
75%	3.884023e+06	0.143402	0.320656	62923.604633	5.654900e+07	25.700000
max	5.427591e+07	37.892650	84.730541	236990.128088	7.479865e+07	33.200000

- **Skewed distribution:** A skewed distribution is neither symmetric nor normal because the data values trail off more sharply on one side than on the other. The distribution of minimum diameter variable is right skewed which means there are more small asteroids than larger ones.

```
sns.histplot(data=neo, x='est_diameter_max')
plt.title('Distribution of Estimated Maximum Diameter')
plt.show()
```

✓ 5.2s



- Size of Object is positively correlated to Relative Velocity and Miss Distance. Therefore we can say that Larger Asteroids have a less probability of hitting earth because of their large miss distance in general. (Exceptions : ‘Apophis’ is a large asteroid with less miss Distance).

- Hazardous Objects have Higher Absolute Magnitude and Less Hazardous Objects have less Absolute Magnitude. This Means that Larger and Hazardous Asteroids shine more and are easily observable from earth compared to Small and Less Hazardous Asteroids.

- Inference : We can Large Asteroid Impacts are predictable and Smaller Impacts are less predictable.

References

- **[1] Sameepvani. (n.d.). NASA Nearest Earth Objects. Kaggle.** https://www.kaggle.com/datasets/sameepvani/nasa-nearest-earth-objects

- **[2] Skewed Distribution** - an overview | ScienceDirect Topics. (n.d.). ScienceDirect. https://www.sciencedirect.com/topics/mathematics/skewed-distribution

- **[3] seaborn.objects.Bar** — seaborn 0.11.2 documentation. (n.d.). Seaborn. https://seaborn.pydata.org/generated/seaborn.objects.Bar.html

- **[4] Pandas - Reading CSV Files.** (n.d.). W3Schools. https://www.w3schools.com/python/pandas/pandas_csv.asp

- **[5] Matplotlib.pyplot.scatter() in Python. (n.d.). GeeksforGeeks.** https://www.geeksforgeeks.org/matplotlib-pyplot-scatter-in-python/?ref=lbp

- **[6] Matplotlib Library:**
  a. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95.
  b. Matplotlib (2021). Homepage. Retrieved from https://matplotlib.org/
  c. Matplotlib (2021). Pyplot tutorial. Retrieved from https://matplotlib.org/stable/tutorials/introductory/pyplot.html

- **[7] Lumpy:**
  a. Molloy, I., Holmes, D., & DeSalvo, G. (2015). Lumpy: a probabilistic framework for structural variant discovery. Genome biology, 16(1), 1-16.
  b. Lumpy (2021). GitHub repository. Retrieved from https://github.com/arq5x/lumpy-sv

- **[8] Near Earth Objects:**
  a. Harris, A. W., & D'Abramo, G. (2015). The NEOSurvey mission: Objects near Earth in the near-infrared. Journal of Astronomical Telescopes, Instruments, and Systems, 1(2), 1-15.
  b. NASA (2021). Near Earth Object Program. Retrieved from https://cneos.jpl.nasa.gov/
  c. Near Earth Object (NEO) Project (2021). NASA Jet Propulsion Laboratory. Retrieved from https://www.jpl.nasa.gov/missions/near-earth-object-neo-project/