

# NEOs CLASSIFICATION

---

Francesco Di Stefano  
Gabriele Di Palma  
Francesco Tramontano

# Table of content



## Analyzing dataset

- Dropped imperial u.m.
- Check NAs
- Check duplicates
- Converting variables

## EDA

- Distributions
- Balance
- Correlations
- Outliers

## Models

- Logistic regression
- Random Forest
- Neural Network

## Logistic Regression

- Undersampling
- Model evaluation

# Analyzing dataset

01

## Dropped imperial u.m.

Our dataset presented a series of the same features expressed in different units of measurement, so we kept the units of measurement used in Europe.

02

## Check NAs

There are no NAs

03

## Check duplicates

There are no repeated rows but an analysis of the NEOs id revealed that the same object was detected several times in the search span.

04

## Converting variables

In this section we have converted all the features that refer to a distance into 'astronomical units' (the distance between the earth and the sun) in order to have the same scale.

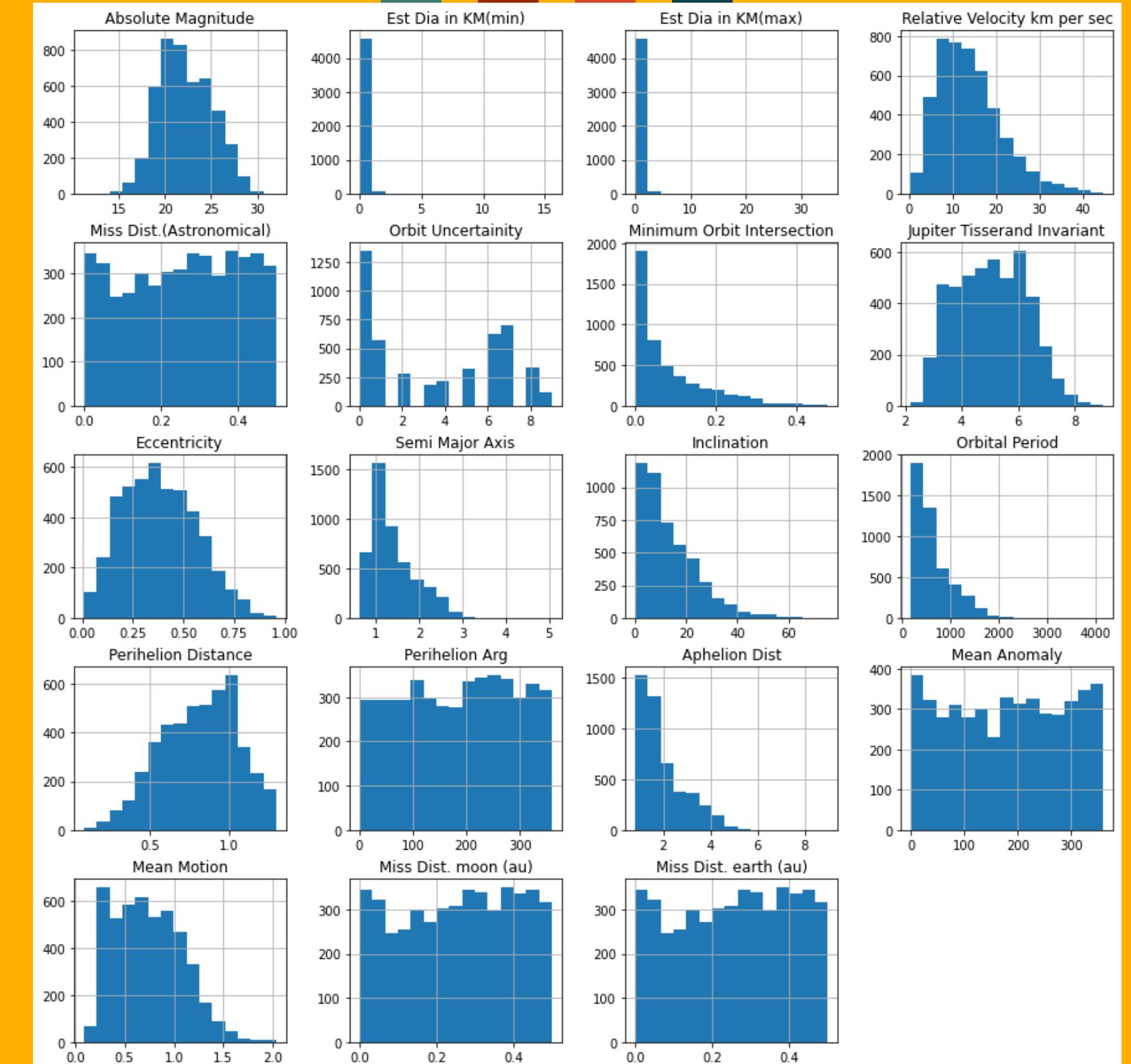


# EXPLORATORY DATA ANALYSIS



# Distributions

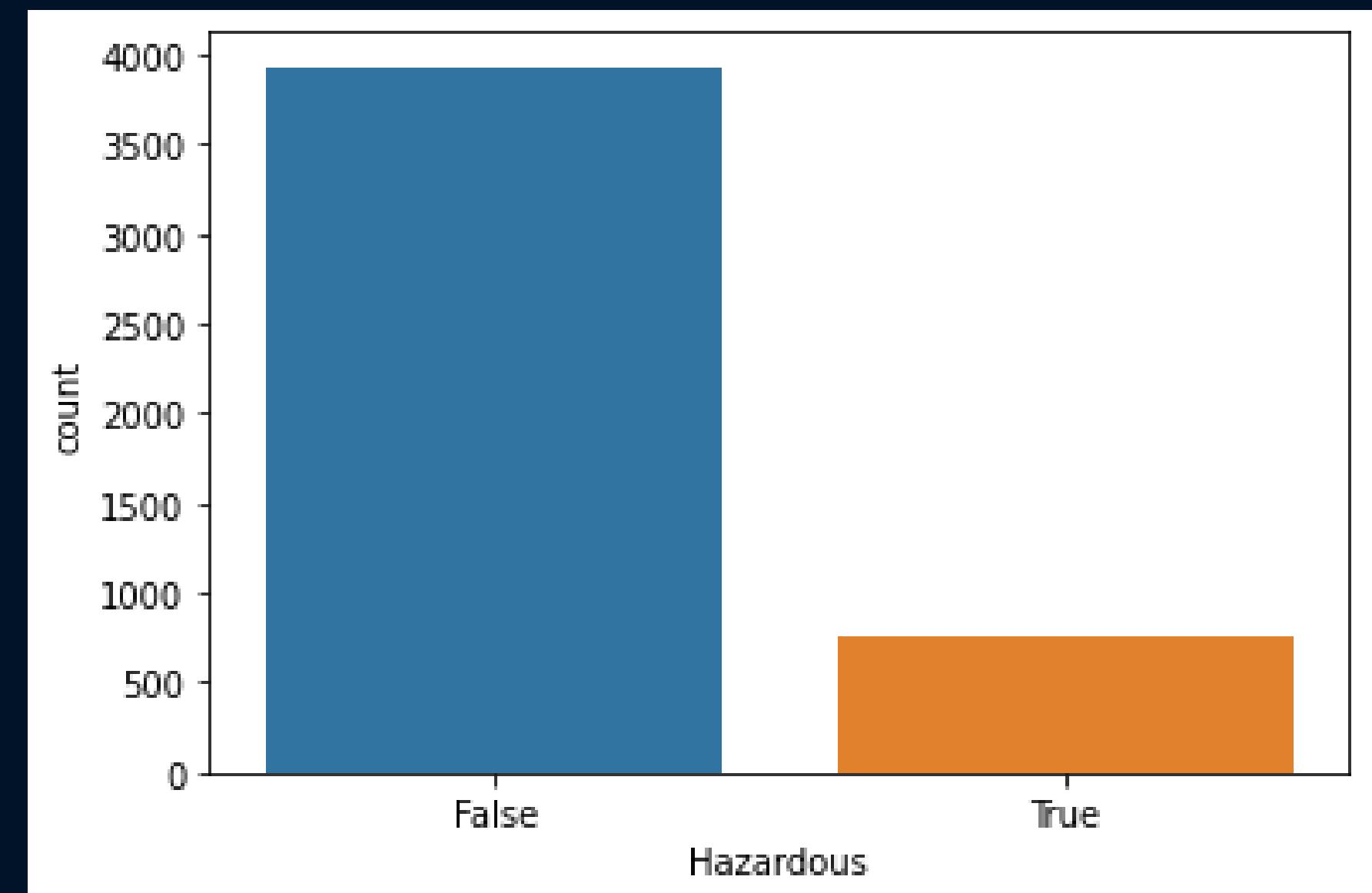
As can be seen, the distributions of the variables are not very usable for some models due to the lack of Gaussianity.



# BALANCE

## False-Positive ratio

Out of 4687 observations 755 are classified as potentially dangerous, representing 16.11% of the total observations



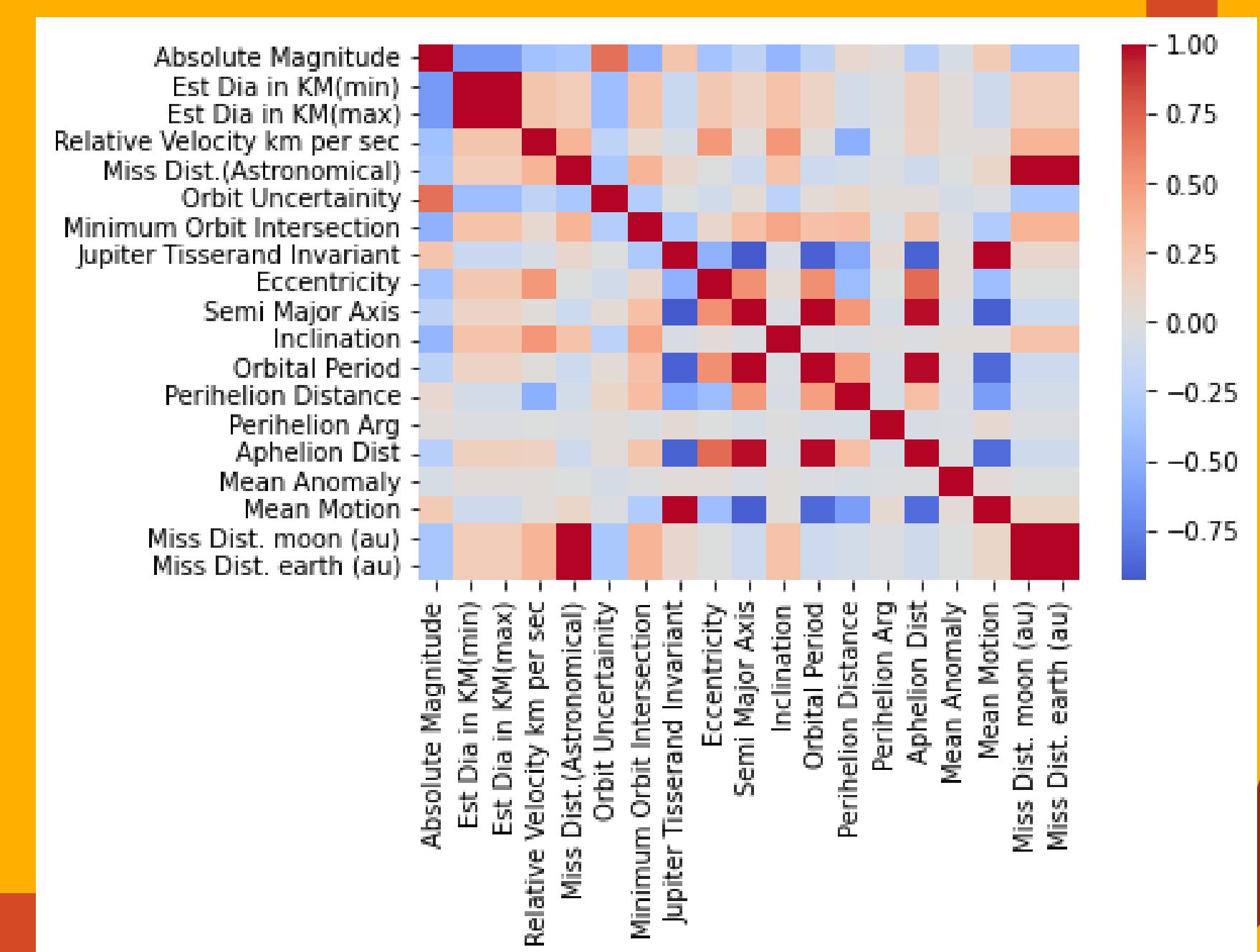
# CORRELATIONS

## Heatmap

This heatmap summarizes all correlations within the dataset..

## Considerations

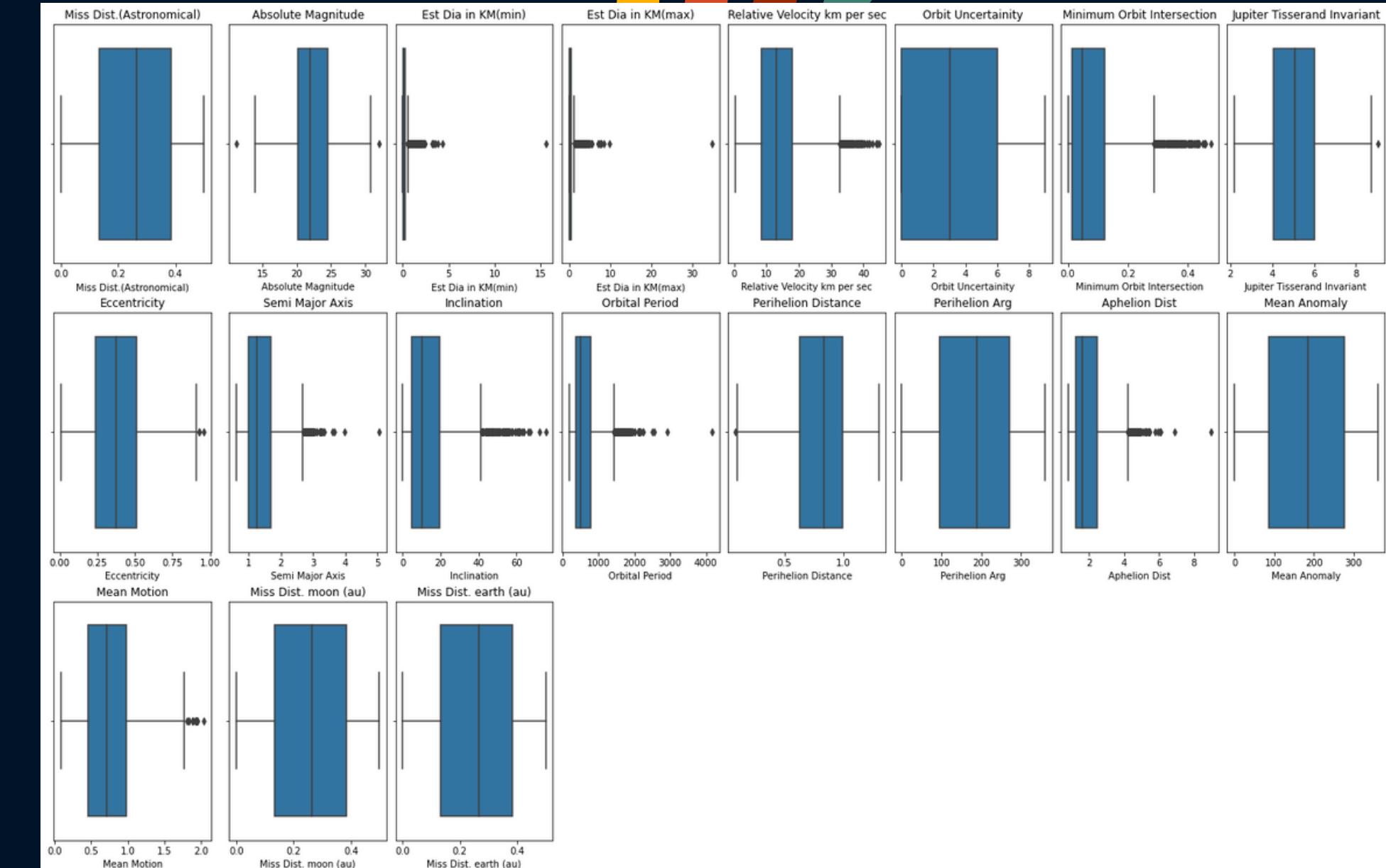
- All features that consider distances have a very high correlation.
- There is a negative correlation between 'Jupiter Tisserand invariant' and 'Orbital period'



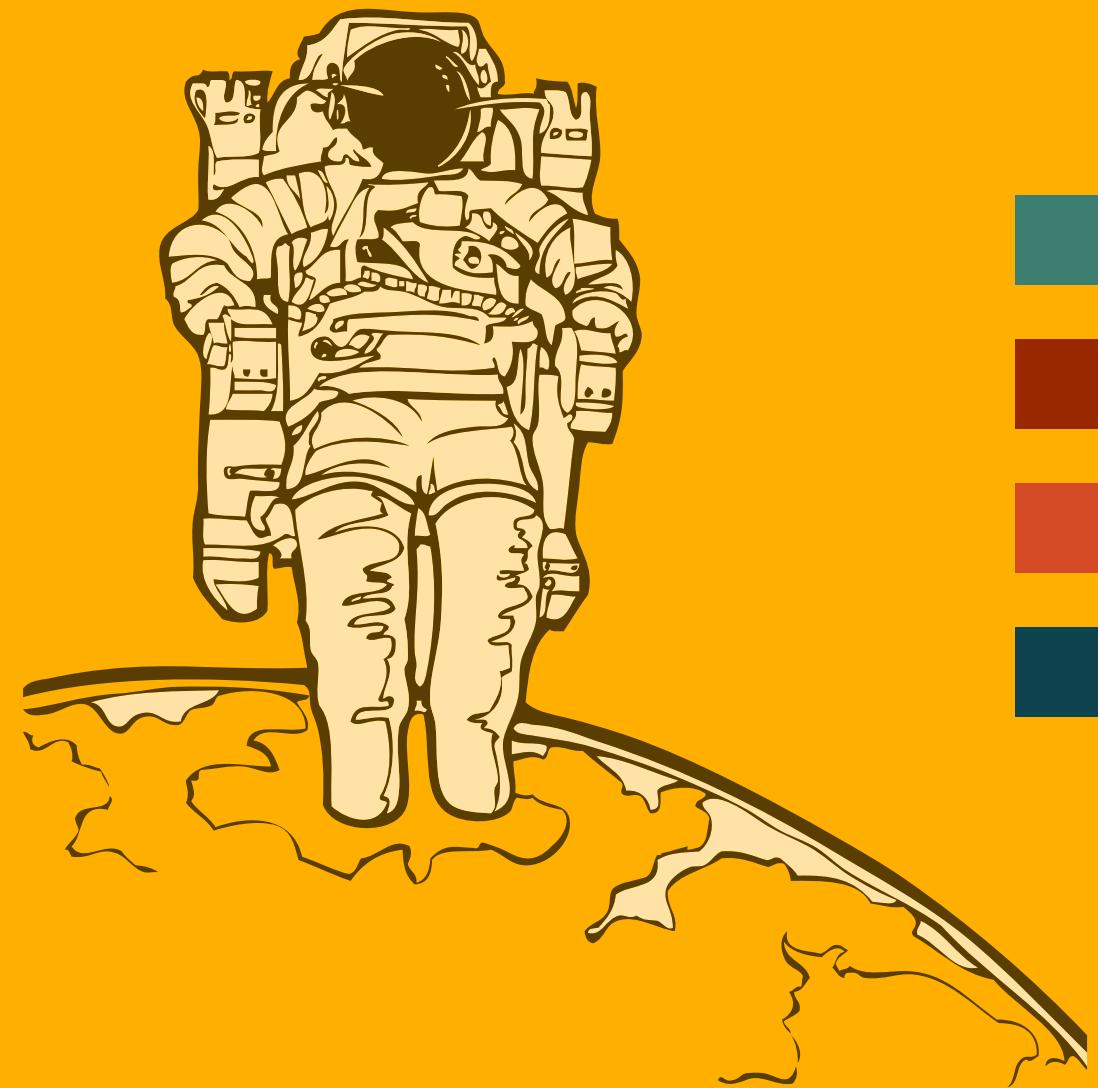
# OUTLIERS

## Boxplots

There are several outliers in some variables, but this does not represent a problem for our analysis because they are the observations that we need to monitor the most. Our task can therefore be compared to an anomaly detection task



# LOGISTIC REGRESSION



# UNDERSAMPLING

| Index | Type      | Size     | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
|-------|-----------|----------|----------------------------------------------------------------------------|
| 0     | DataFrame | (3, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 1     | DataFrame | (6, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 2     | DataFrame | (36, 18) | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 3     | DataFrame | (3, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 4     | DataFrame | (3, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 5     | DataFrame | (4, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 6     | DataFrame | (4, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 7     | DataFrame | (3, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 8     | DataFrame | (3, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 9     | DataFrame | (3, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 10    | DataFrame | (23, 18) | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 11    | DataFrame | (22, 18) | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 12    | DataFrame | (7, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 13    | DataFrame | (5, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 14    | DataFrame | (6, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 15    | DataFrame | (7, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 16    | DataFrame | (8, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 17    | DataFrame | (7, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 18    | DataFrame | (3, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 19    | DataFrame | (6, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 20    | DataFrame | (6, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 21    | DataFrame | (3, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |
| 22    | DataFrame | (3, 18)  | Column names: Absolute Magnitude, Relative Velocity km per sec, Miss D ... |

## Why:

Since logistic regression needs a balanced input of the dataset we opted for an undersampling

| Index | Absolute Magnitude | Relative Velocity km per sec | Epoch(Astronomical Unit) | Sidereal Uncertainty | Orbit Interval | Tisserand Index | Eccentricity | Major Axis | Inclination | Orbital Period | Heliocentric Distance | Period |
|-------|--------------------|------------------------------|--------------------------|----------------------|----------------|-----------------|--------------|------------|-------------|----------------|-----------------------|--------|
| 29    | 23.8               | 11.9306                      | 0.0975441                | 0                    | 0.052677       | 7.48            | 0.386459     | 0.762974   | 20.8338     | 243.424        | 0.468116              | 351    |
| 33    | 23.8               | 11.9306                      | 0.0975438                | 0                    | 0.0515902      | 7.48            | 0.386445     | 0.762954   | 20.8334     | 243.414        | 0.468114              | 351    |
| 316   | 23.8               | 11.917                       | 0.0884576                | 0                    | 0.052677       | 7.48            | 0.386459     | 0.762974   | 20.8338     | 243.424        | 0.468116              | 351    |
| 322   | 23.8               | 11.917                       | 0.0884573                | 0                    | 0.0515902      | 7.48            | 0.386445     | 0.762954   | 20.8334     | 243.414        | 0.468114              | 351    |
| 4195  | 23.8               | 12.0762                      | 0.0562219                | 0                    | 0.052677       | 7.48            | 0.386459     | 0.762974   | 20.8338     | 243.424        | 0.468116              | 351    |
| 4204  | 23.8               | 12.0762                      | 0.0562218                | 0                    | 0.0515902      | 7.48            | 0.386445     | 0.762954   | 20.8334     | 243.414        | 0.468114              | 351    |

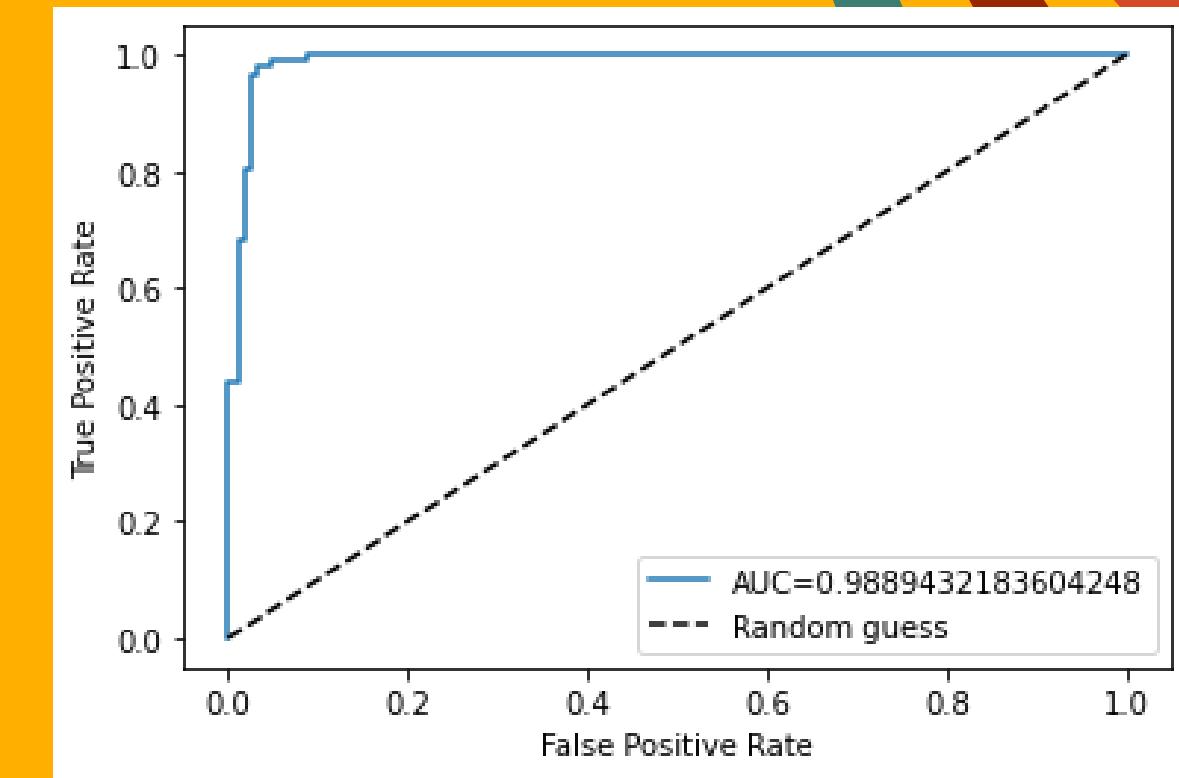
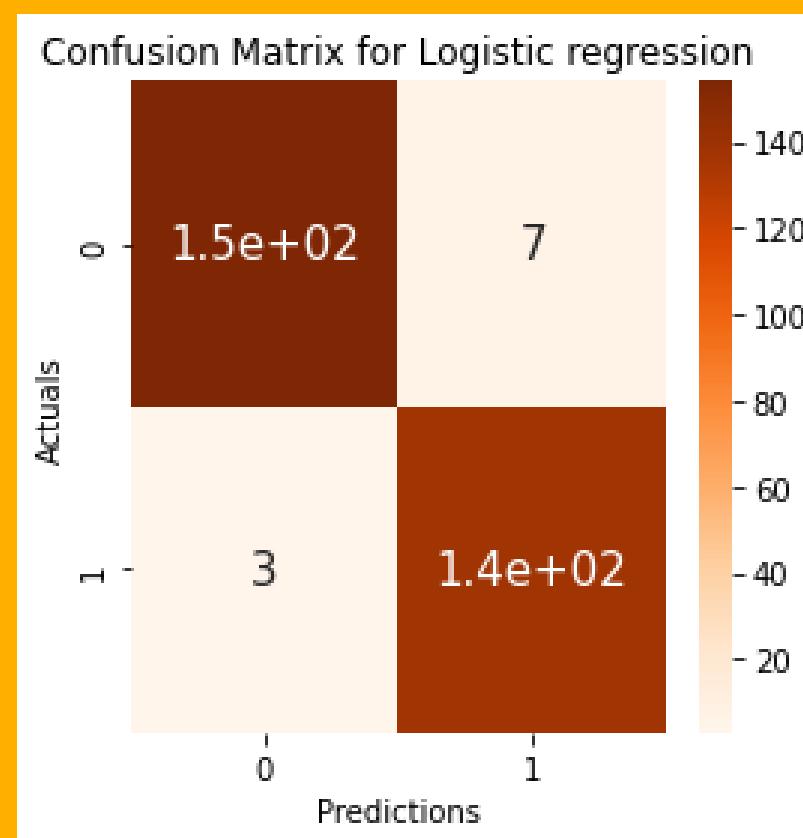
## How:

We performed with a db scan clustering, we then took observations from each of them to arrive at 1510 perfectly balanced observations

# MODEL METRICS

## Confusion Matrix

Logistic regression precision: 0.9517241379310345  
Logistic regression accuracy: 0.9668874172185431  
Logistic regression recall: 0.9787234042553191  
Logistic regression F-1 score: 0.965034965034965  
Logistic regression specificity: 0.9565217391304348



**ROC curve**

The ROC curve is almost perfect with an AUC of: 0.989

# PARAMETERS

Interpretation:

As we can see not all parameters are relevant for the model.

|     |              | 2.5%         | 97.5%        | Odds Ratio   | pvalues      | significant?    |
|-----|--------------|--------------|--------------|--------------|--------------|-----------------|
| x1  | 3.155310e-04 | 5.825838e-03 | 1.355814e-03 | 6.845492e-19 | 6.845492e-19 | significant     |
| x2  | 1.804305e+00 | 1.051243e+01 | 4.355185e+00 | 1.065462e-03 | 1.065462e-03 | significant     |
| x3  | 0.000000e+00 |              | inf          | 1.770546e-37 | 9.999808e-01 | not significant |
| x4  | 3.350756e-01 | 1.084910e+00 | 6.029318e-01 | 9.140011e-02 | 9.140011e-02 | not significant |
| x5  | 1.782153e-03 | 2.396448e-02 | 6.535164e-03 | 3.247812e-14 | 3.247812e-14 | significant     |
| x6  | 8.121680e+02 | 3.591726e+16 | 5.401004e+09 | 5.177060e-03 | 5.177060e-03 | significant     |
| x7  | 1.930467e-01 | 9.055550e+00 | 1.322174e+00 | 7.760404e-01 | 7.760404e-01 | not significant |
| x8  | 0.000000e+00 |              | inf          | 2.015626e+01 | 9.999998e-01 | not significant |
| x9  | 4.942356e-01 | 2.684544e+00 | 1.151867e+00 | 7.432891e-01 | 7.432891e-01 | not significant |
| x10 | 2.754636e-12 | 1.792121e+09 | 7.026124e-02 | 8.280472e-01 | 8.280472e-01 | not significant |
| x11 | 0.000000e+00 |              | inf          | 1.201890e+00 | 9.999999e-01 | not significant |
| x12 | 6.950560e-01 | 1.541571e+00 | 1.035122e+00 | 8.651103e-01 | 8.651103e-01 | not significant |
| x13 | 0.000000e+00 |              | inf          | 2.412068e+01 | 9.999998e-01 | not significant |
| x14 | 1.051653e+00 | 2.426355e+00 | 1.597399e+00 | 2.808427e-02 | 2.808427e-02 | significant     |
| x15 | 4.391807e-14 | 4.087983e-04 | 4.237173e-09 | 9.934457e-04 | 9.934457e-04 | significant     |
| x16 | 0.000000e+00 |              | inf          | 1.323124e+16 | 9.999928e-01 | not significant |
| x17 | 0.000000e+00 |              | inf          | 3.341519e+20 | 9.999895e-01 | not significant |
| x18 | 5.692982e-01 | 7.304492e+00 | 2.039224e+00 | 2.736961e-01 | 2.736961e-01 | not significant |

**THANK  
YOU**

