Exploratory Data Analysis Report

1. Initial Question:

The initial question that guided the selection of this dataset was;

How does the toxicity of small molecules affect their ability to regulate the circadian rhythm?

This was the motivating research question before starting the data analysis. However, during the analysis, the following additional questions emerged:

- What is the class distribution of toxic versus non-toxic molecules?- Which molecular descriptors are most correlated with the toxicity of the molecules?

- Are there any significant outliers in the molecular descriptor data?

- How can we visualize and understand the distribution of the molecular descriptors?

2. Data Wrangling

Handling Missing Values

The dataset contained no missing values. Each column had complete data for all 171 entries, making it unnecessary to impute or drop data based on missing values.

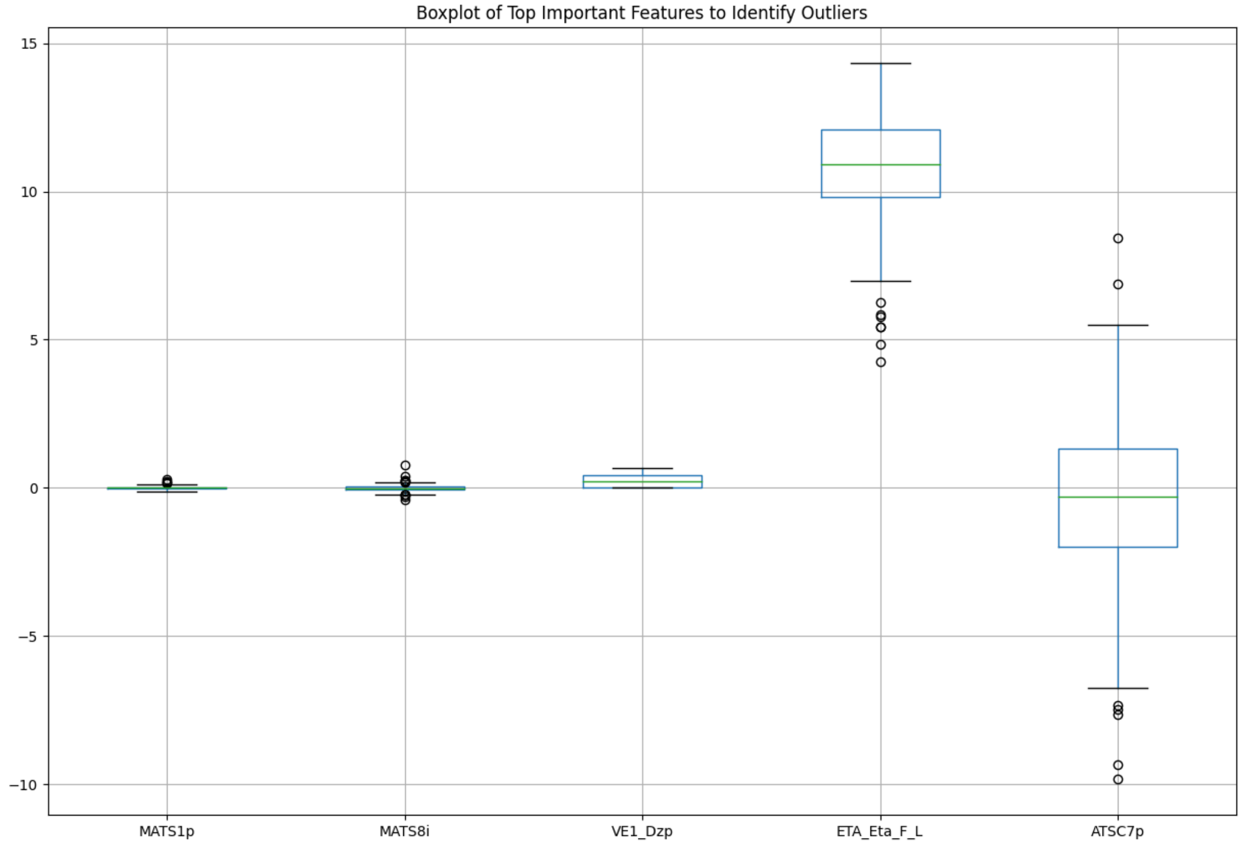
Handling Outliers

To identify outliers, the **Interquartile Range (IQR)** method was used. This method detects data points that are significantly higher or lower than the majority of the data. Several molecular descriptors showed the presence of outliers, and these were explored through box plots. Outliers were prevalent in descriptors such as `EE_Dt`, `C2SP2`, and `AATSC7p`, among others.

Outliers were detected using the IQR method across the numerical columns. Box plots were generated to visually inspect the distribution of these features and identify outliers. For example:

- The descriptor `EE_Dt` had several high outliers, suggesting that a few molecules deviate significantly in this measurement.

- Similarly, `C2SP2` and `AATSC7p` displayed outliers that fell outside the typical range, which might influence further analysis**.**



3. Exploratory Data Analysis (EDA)

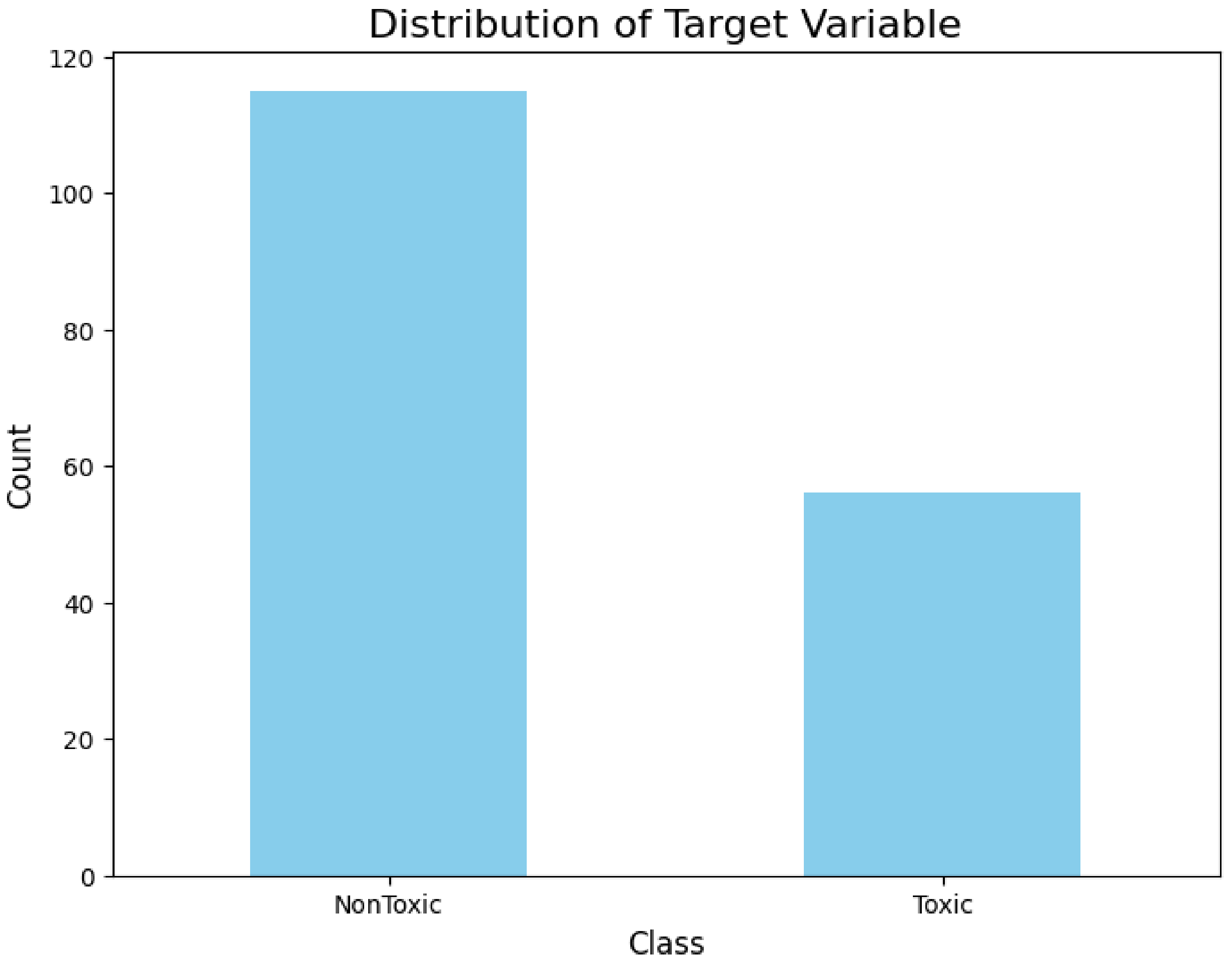
Class Distribution Analysis

One of the first steps in EDA was to examine the class distribution of the "Class" column, which indicates the toxicity of the molecules. The analysis showed a class imbalance:

- 115 instances of "NonToxic"

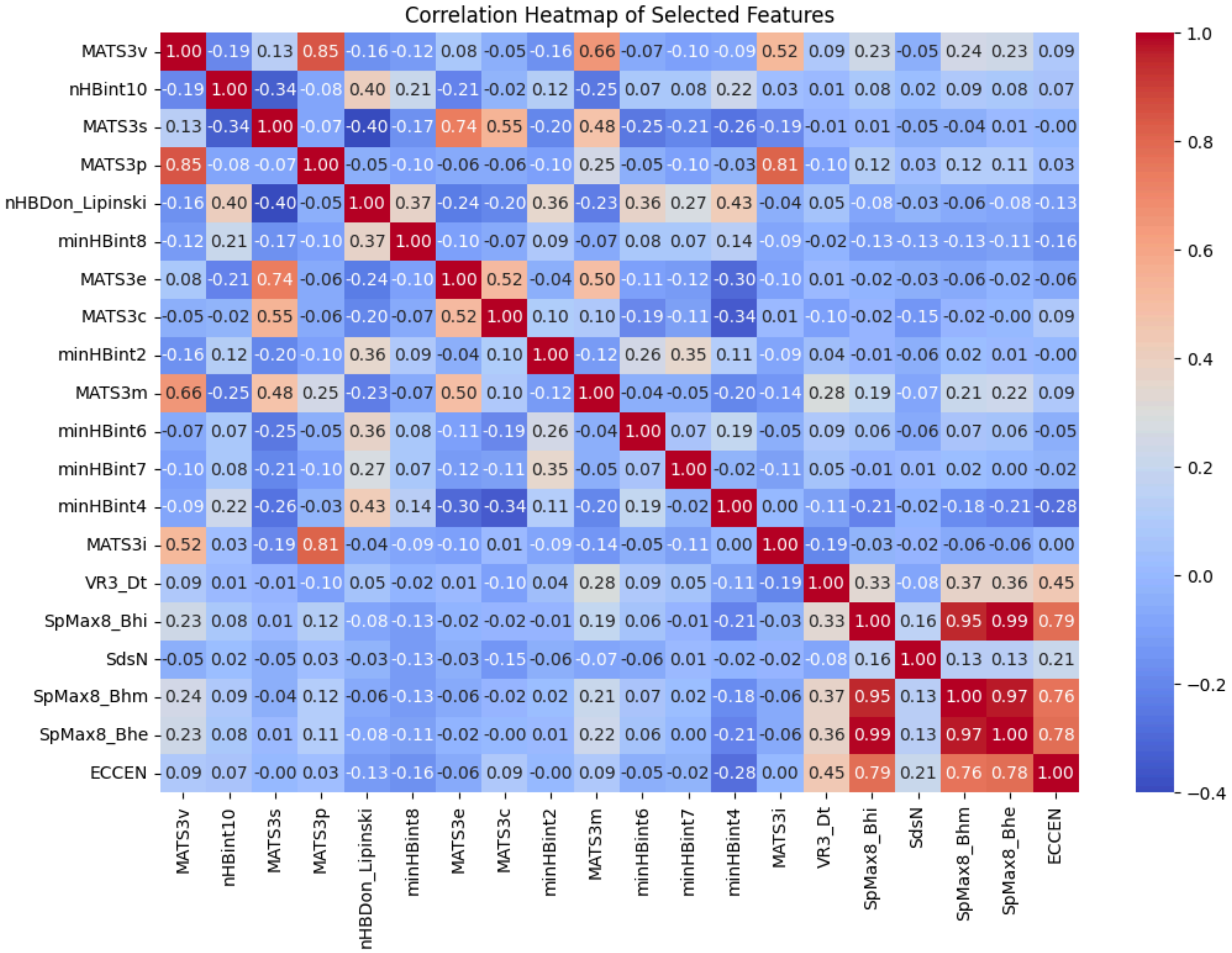
- 56 instances of "Toxic"

This imbalance may influence modeling techniques in the future, as the dataset leans toward non-toxic molecules.



Correlation Analysis

Next, a correlation analysis was performed to identify the molecular descriptors that were most closely related to the toxicity (as measured by the "Class" column). The top 10 molecular descriptors most correlated with toxicity are as follows:



1. **EE_Dt (0.214)**
2. **C2SP2 (0.189)**
3. **AATSC7p (0.165)**
4. **SpDiam_Dt (0.165)**
5. **MLogP (0.164)**
6. **MATS7p (0.164)**
7. **nAcid (0.160)**
8. **nwHBa(0.157)**
9. **GATS7v (0.156)**

10. SpMin4_Bhi(0.156)

These features show a moderate correlation with the toxicity class and may serve as important predictors in further analysis or modeling.

4. Drawing Conclusions

The EDA process yielded several important findings:

- There is a clear class imbalance in the dataset, with far more non-toxic molecules than toxic ones.

- Certain molecular descriptors, such as `EE_Dt` and `C2SP2`, are moderately correlated with toxicity, which may guide future modeling efforts.

- Outliers are present in several molecular descriptors, and they may need to be addressed before any modeling process to avoid skewing results.

Therefore the analysis revealed several key insights about the relationship between molecular descriptors and toxicity:

- The dataset is imbalanced, with more non-toxic molecules, which will need to be accounted for in any predictive modeling.

- The correlation analysis identified features that are potentially important for predicting toxicity, and these should be further investigated.

- The presence of outliers suggests that data cleaning or robust modeling techniques will be necessary to ensure accurate results.

Future steps include addressing the class imbalance through oversampling or other techniques and possibly handling outliers using transformations or robust methods.

https://docs.google.com/spreadsheets/d/1tSh2OlNH0ux_L3SrT0_sXujToDIBlhPV9IAZVuRBJik/edit?usp=sharing