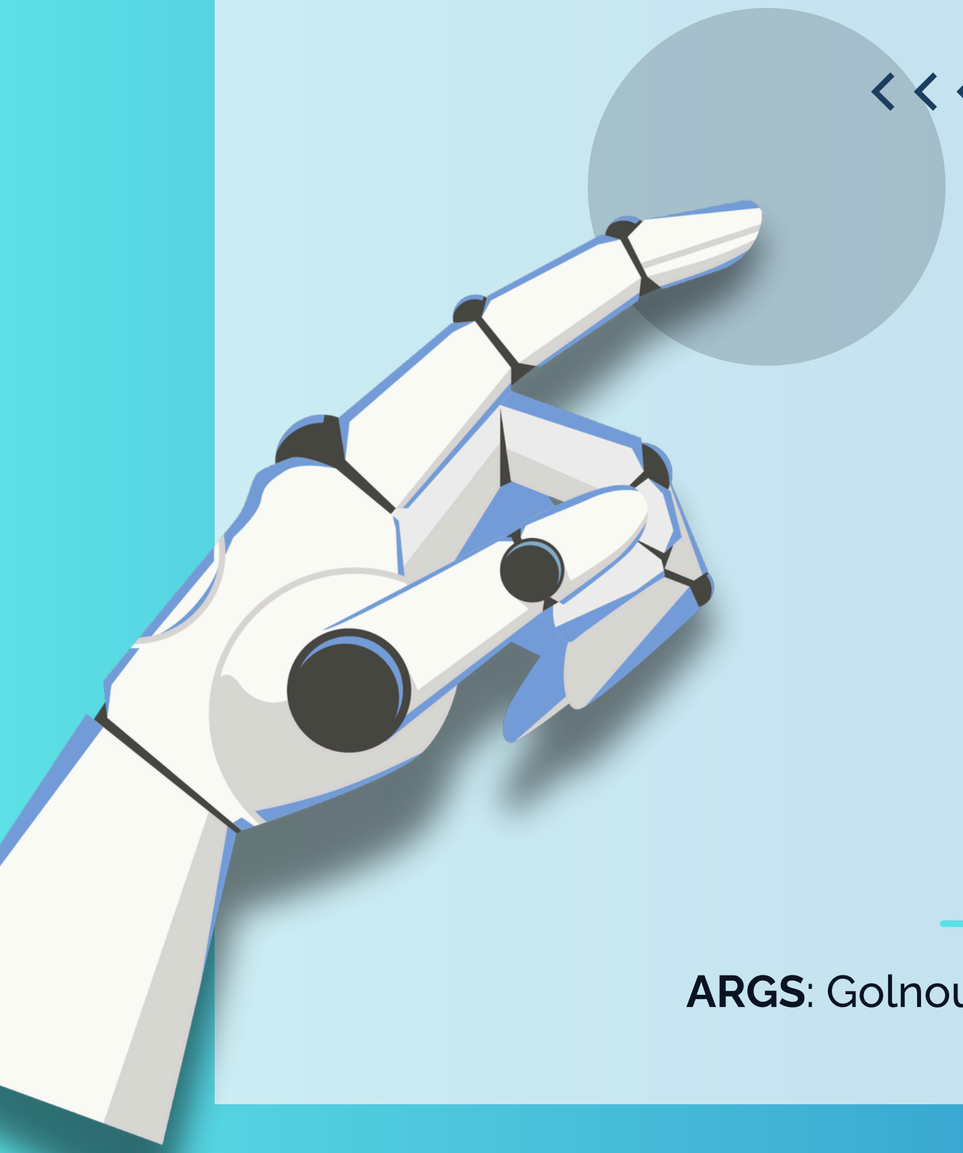


BEHAVIOR DETECTION & MOBILE ANALYSIS

ARGS: Golnoush Nematbakhsh, Arghavan Azhir, Ramtin Karbaschi, Sam Rasouli

CONTENTS

Machine Learning

- Preprocessing
- Training
- Evaluation

Statistical Analysis

- Descriptive Statistics
- Hypothesis Testing

Diagnosis of disorders caused by inappropriate use of the Internet

Price Prediction

- Preprocessing
- Training
- Evaluation

Clustering

- K-means
- DBScan

Review of mobile phone market data

01.

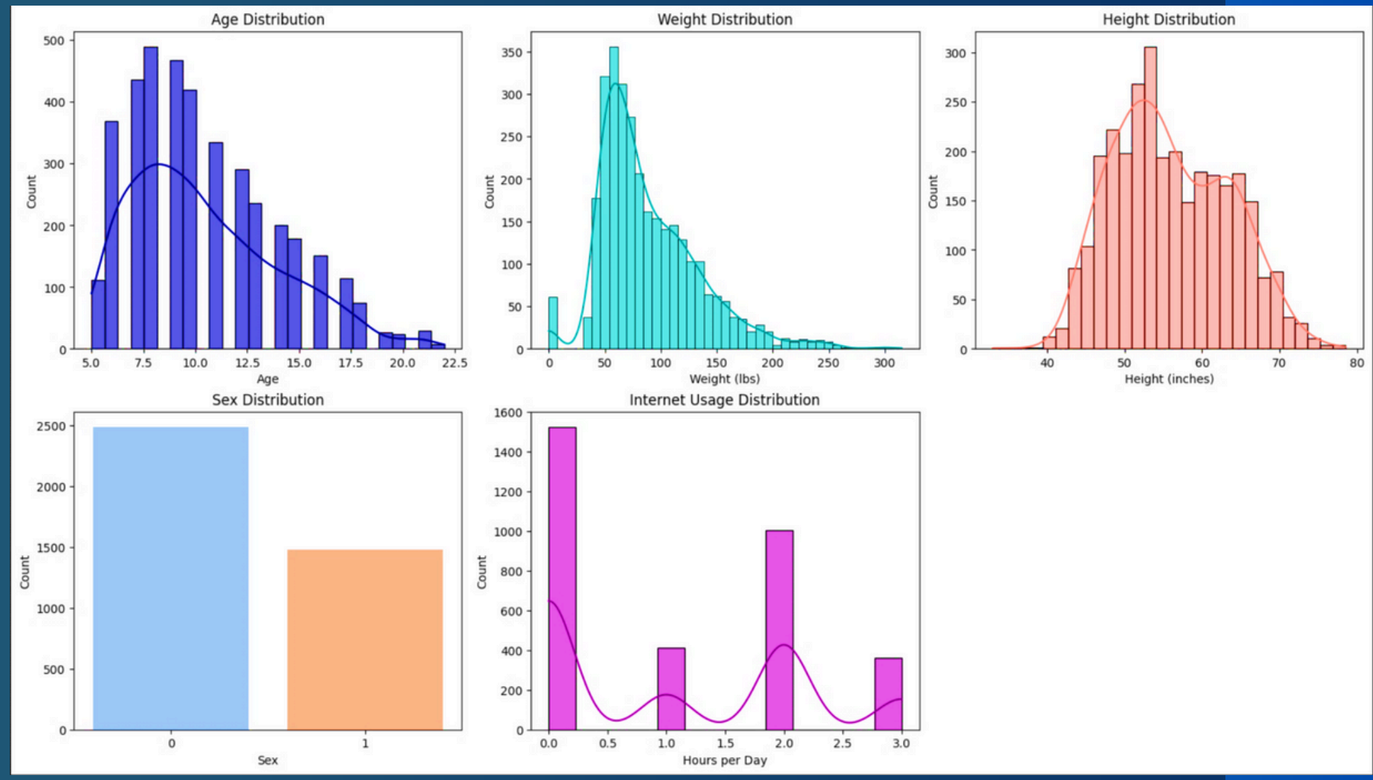
DIAGNOSIS OF DISORDERS CAUSED BY INAPPROPRIATE USE OF THE INTERNET

STATISTICAL ANALYSIS

DESCRIPTIVE STATISTICS

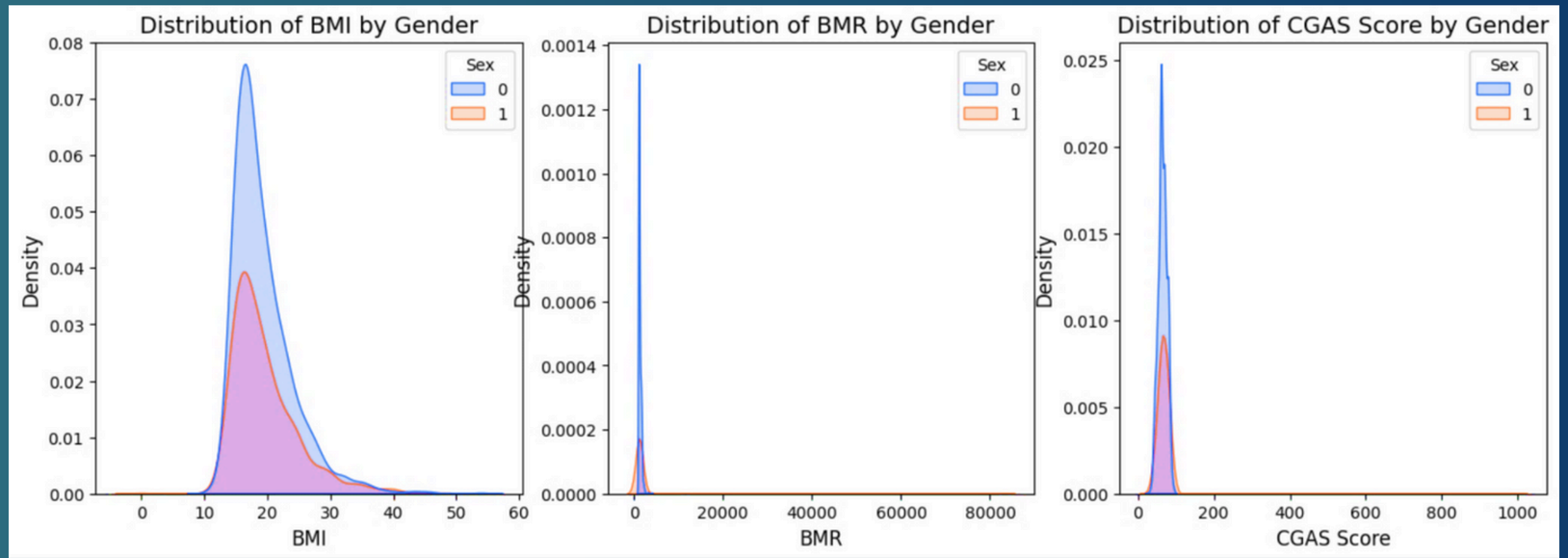
Using histograms, the distribution of Age, Wight, Height, Gender, and Internet use variables was examined.

- **Age**: *Most people between the ages of 7 and 12;*
- **Weight**: *Most people weigh less, and a few weigh more;*
- **Height**: *About normal, with an average height between 50 and 60 inches;*
- **Gender**: *More men than women;*
- **Internet use**: *Non-normal and intermittent distribution*



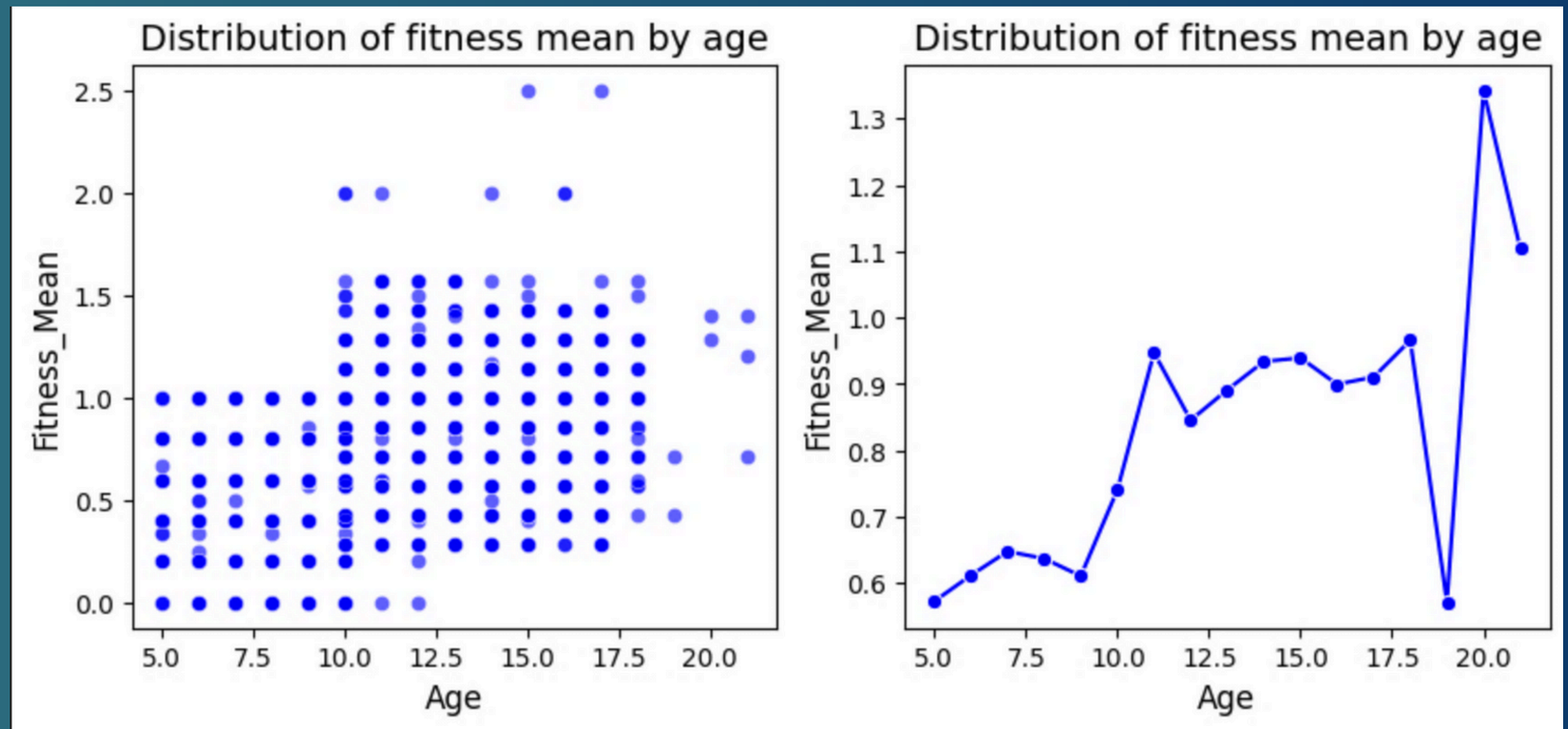
DESCRIPTIVE STATISTICS

- The distribution of muscle mass, basal metabolic rate, and mental performance were examined by gender.
- The distribution of muscle mass is similar for both genders.
- The peak of the distribution is higher for men, indicating that more people are in the normal BMI range.
- The density of basal metabolic rate is almost the same for both groups.
- Most people in both genders have values close to zero for CGAS, and no significant difference is observed between the two genders.



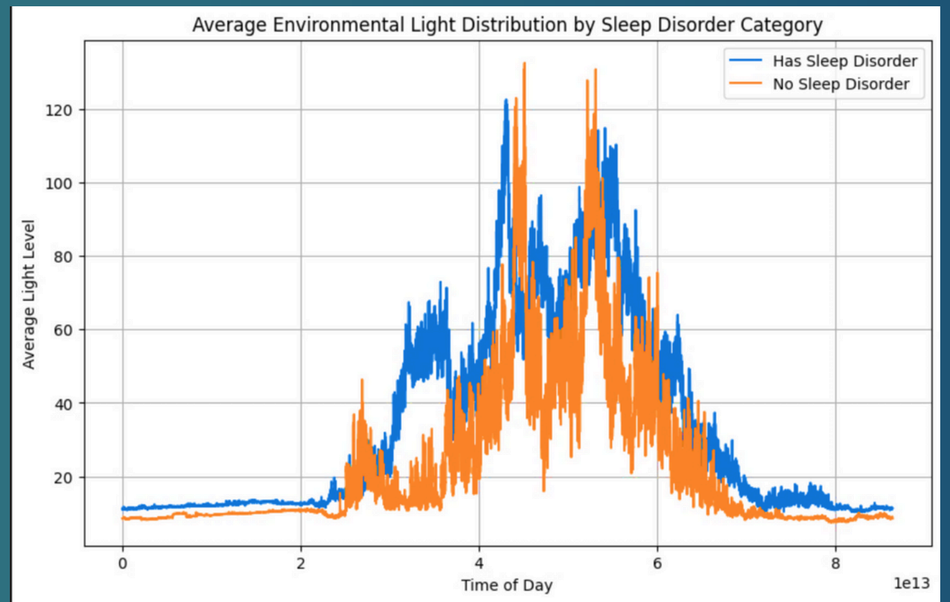
DESCRIPTIVE STATISTICS

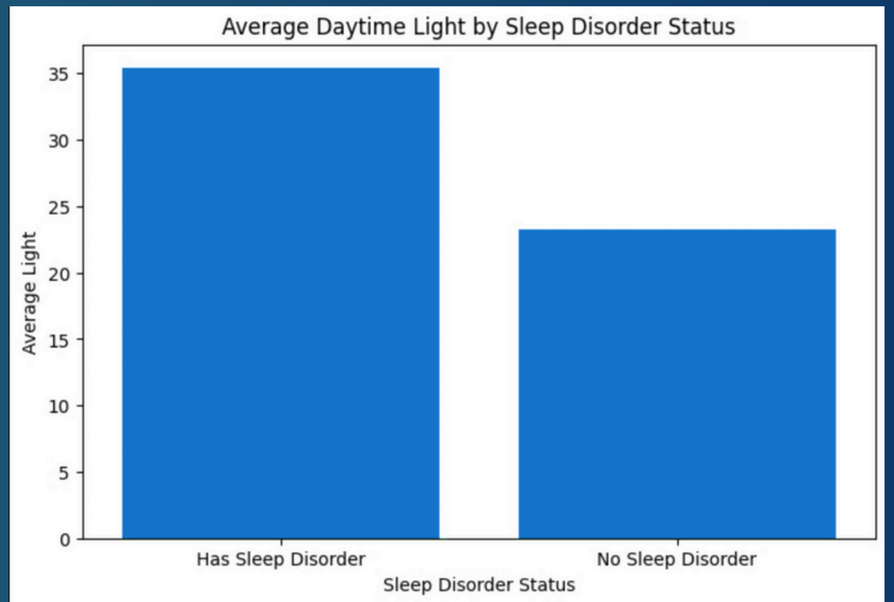
- The blue dots represent the data scatter plot.
- Most of the data are in the age range of 5 to 20 years.
- The fitness generally appears to increase with age, especially from age 5 to around 12 years.
- In the age range of 12-20 years, the fitness fluctuates, peaking at age 20 but then decreasing again.



DESCRIPTIVE STATISTICS

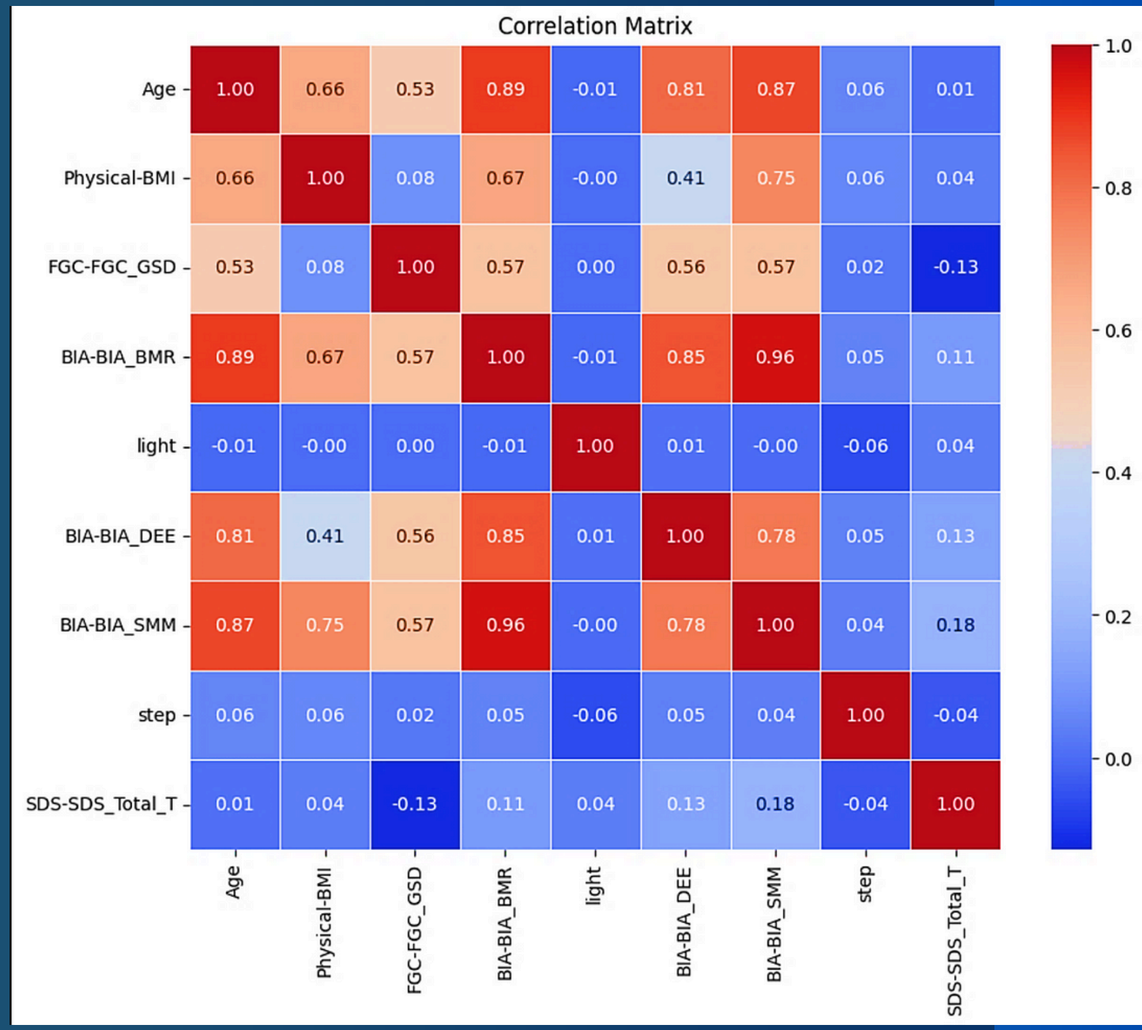
- Changes in the average ambient light level during the day for two groups of people with sleep disorders and without sleep disorders have been examined.
- People with sleep disorders experience higher average light levels than people without sleep disorders.
- Both groups have similar, lower light levels at the beginning (0 to 2) and end of the day (7 to 8).
- During the day (especially at the peak between points 4 and 6), the differences are more pronounced.





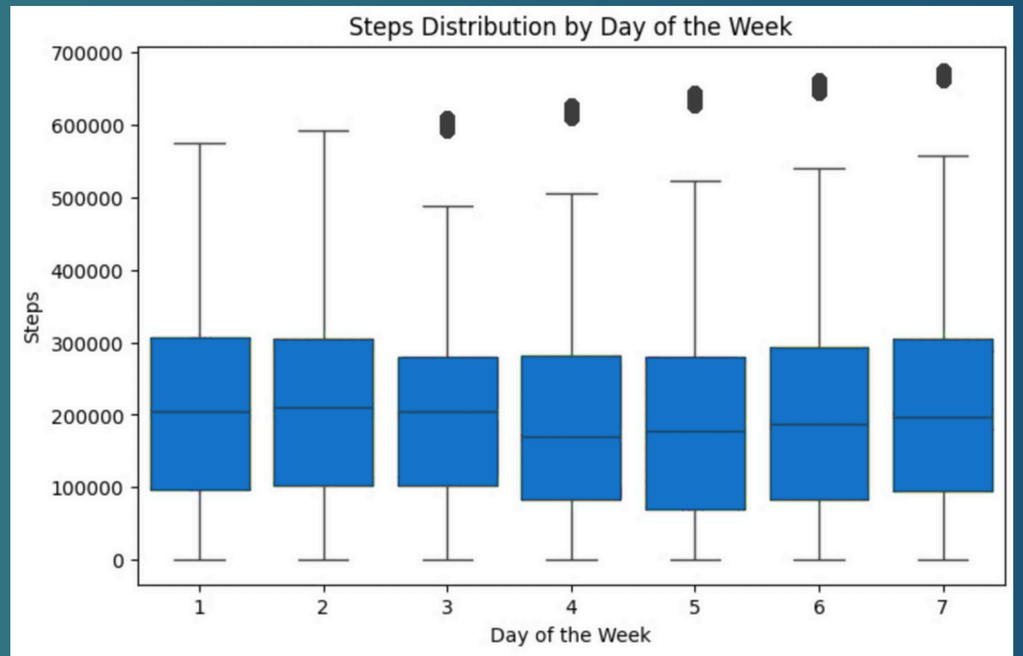
DESCRIPTIVE STATISTICS

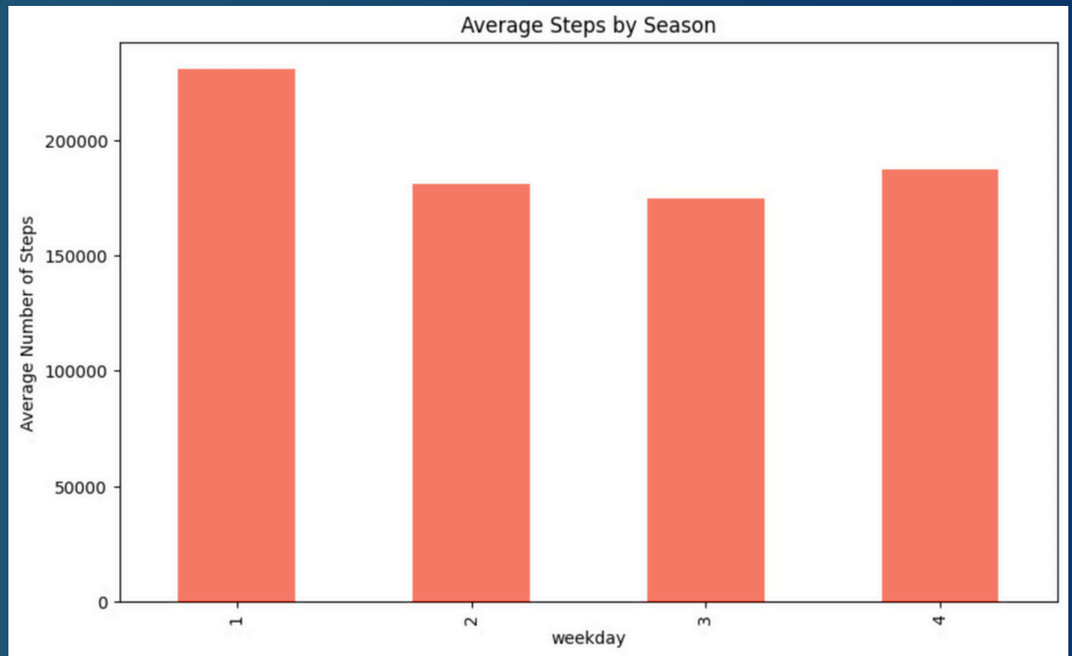
- The correlation matrix for the various variables has been examined.
- Age has a relatively strong and direct relationship with basal metabolic rate and lean muscle mass.
- There is a very strong and direct relationship between basal metabolic rate and lean muscle mass.
- Basal metabolic rate and daily energy expenditure (DEE) also have a strong positive correlation.
- Variables such as ambient light and step count have little linear effect in this data set.



DESCRIPTIVE STATISTICS

- The total number of steps was examined across the days of the week and then the seasons of the year.
- In general, the total number of steps seems to be almost the same across all days of the week.
- Across the seasons of the year, the highest value for the average number of steps was reported in the first season.





HYPOTHESIS TESTING

- Sleep disorders can lead to a decrease in body metabolism. Divide people into two categories with sleep disorder and without sleep disorder, do you think the average metabolic rate of these two groups is equal?

HYPOTHESIS TESTING

Average metabolic rate:

- **The first group:** *people with sleep disorders;*
- **The second group:** *people without sleep disorders;*
- **Zero hypothesis:** *the average metabolic rate of these two groups is equal;*

- **Data type check:** *The data of the two groups are not normal. (Shapiro test);*
- **Select the appropriate test:** *Mann Whiney U test;*

Result:

- The null hypothesis is rejected and the average metabolic rate of these two groups is unequal.

HYPOTHESIS TESTING

- Is there a noticeable difference in lean muscle mass on average between girls and boys?

Divide the participants into two groups > > > *children (ages 5 to 13)* & *youth (ages 14 to 22).*

- Is there a noticeable difference in lean muscle mass on average between children and youth?

HYPOTHESIS TESTING

Pure muscle mass:

- **The first group:** *pure muscle mass among girls;*
- **The second group:** *pure muscle mass among boys;*
- **Zero hypothesis:** *There is no visible difference in the mean of pure muscle mass of girls and boys and it is equal;*

- **Data type check:** *The data of the two groups are not normal. (Shapiro test);*
- **Select the appropriate test:** *Mann Whitney U test;*

Result:

- Pure muscle mass between girls and boys is equal. So there is no significant difference.

HYPOTHESIS TESTING

Pure muscle mass:

- **The first group:** *Children (5 to 13 years old);*
- **The second group:** *Young (from 14 to 22 years old);*
- **Zero hypothesis:** *There is no visible difference in the average of the pure muscle mass of children and teenagers and it is equal;*

- **Data type check:** *The data of the two groups are not normal. (Shapiro test);*
- **Select the appropriate test:** *Mann Whitney U test;*

Result:

- Pure muscle mass is unequal between children and adolescents.

HYPOTHESIS TESTING

- Examine the following statement not only using hypothesis testing; but also with appropriate explanations and further analysis.

- Scientists claim that children and adults who use the Internet for more than two hours are more physically active than children and adults who use the Internet for less than this amount.

HYPOTHESIS TESTING

Physical activity level:

- **The first group:** *Children and adults who use the Internet for more than two hours;*
- **The second group:** *Children and adults who use the Internet for less than two hours;*
- **Zero hypothesis:** *The level of physical activity in children and adults who use the Internet for more than two hours is equal to that in children and adults who use the Internet for less than this amount;*

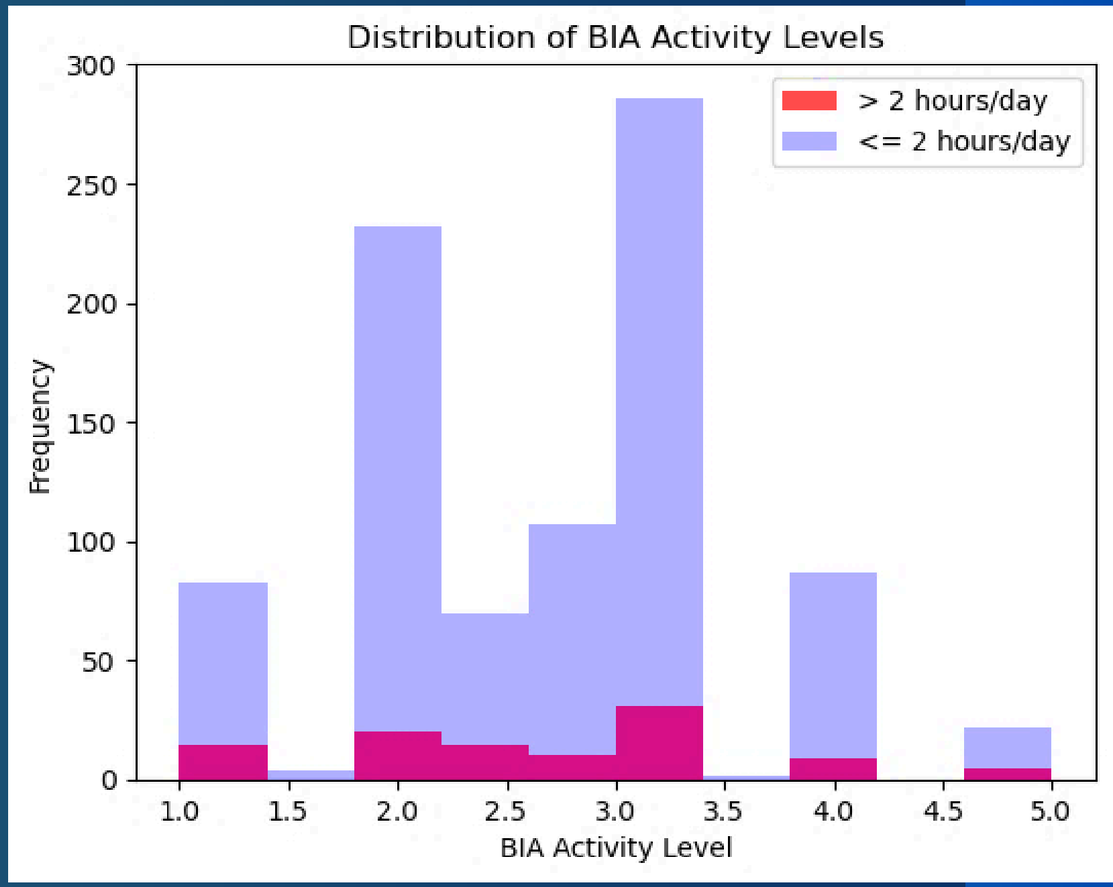
HYPOTHESIS TESTING

Physical activity level:

- The first check was done with the column “BIA BIA ACTIVITY LEVEL” which:
  - **Data type check:** *The data of the two groups are not normal. (Shapiro test);*
  - **Select the appropriate test:** *Mann Whitney U test;*

Result:

Here and using this column we only conclude that the physical activity of the two groups is not equal.



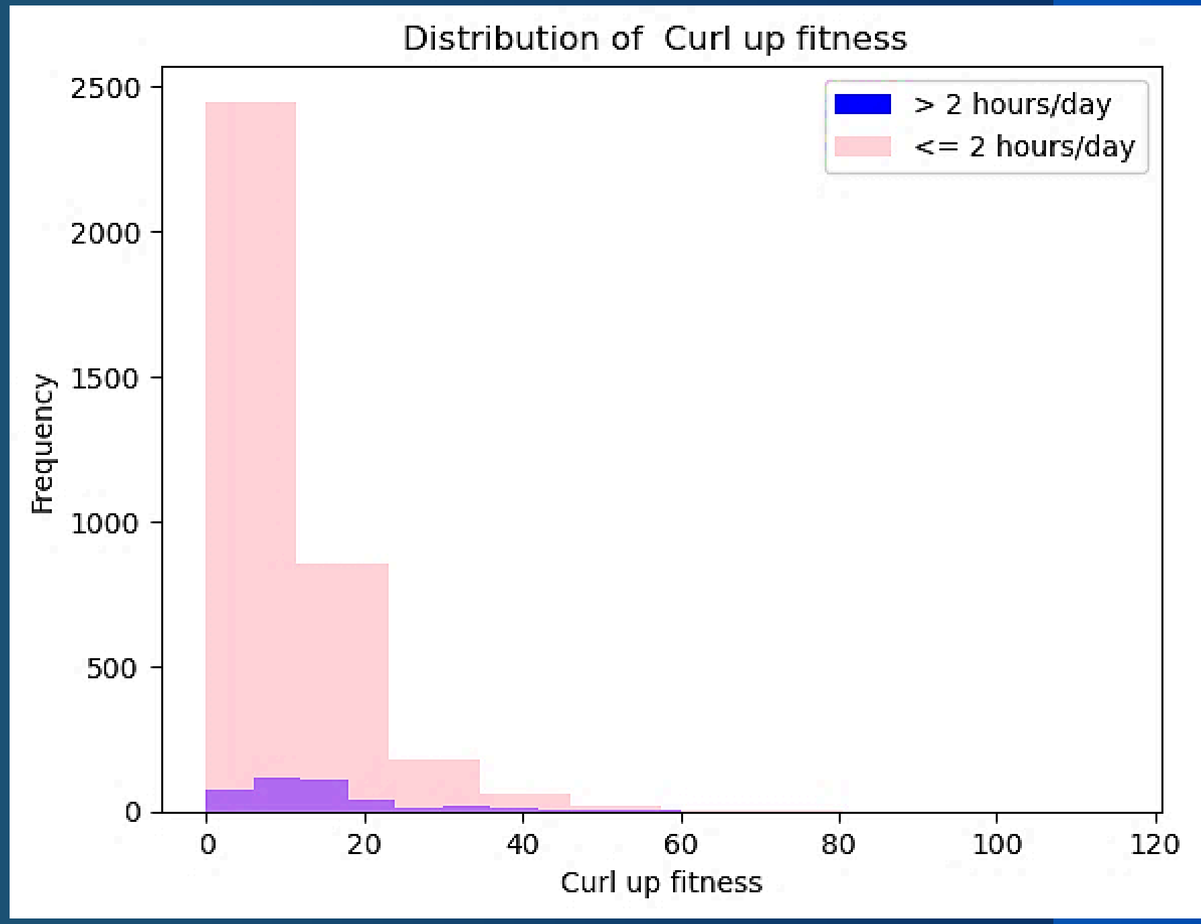
HYPOTHESIS TESTING

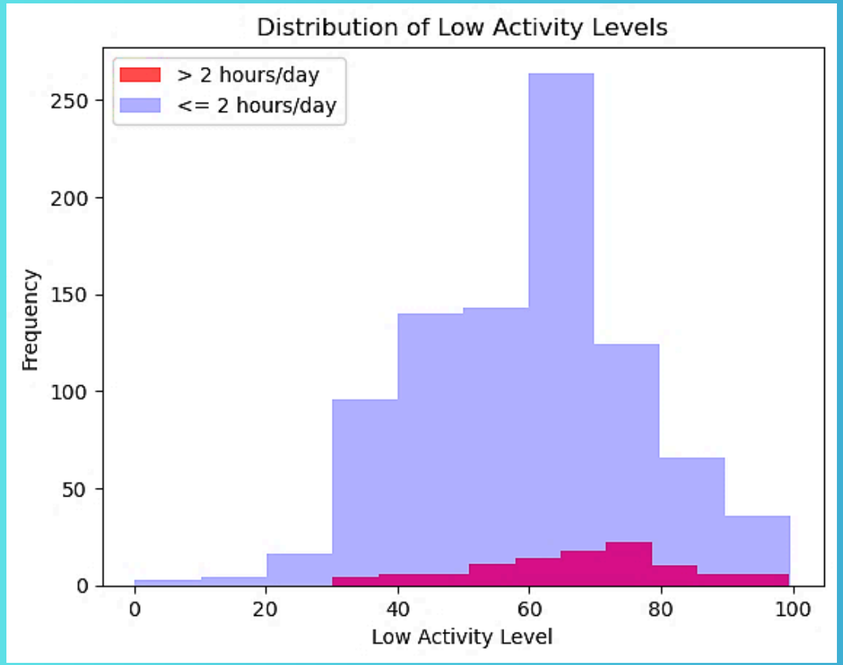
Physical activity:

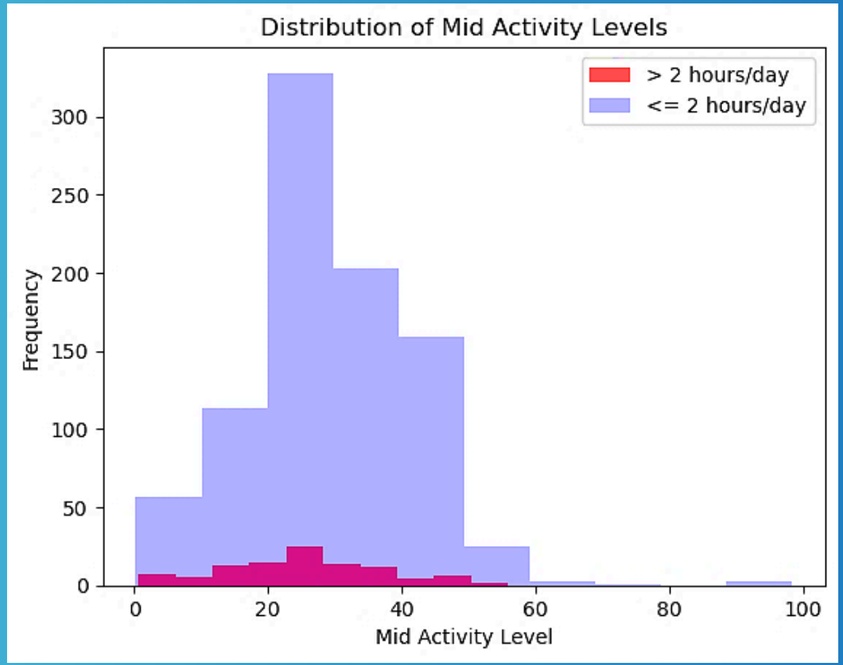
- This time, the number of long sessions they spent on the test was examined:

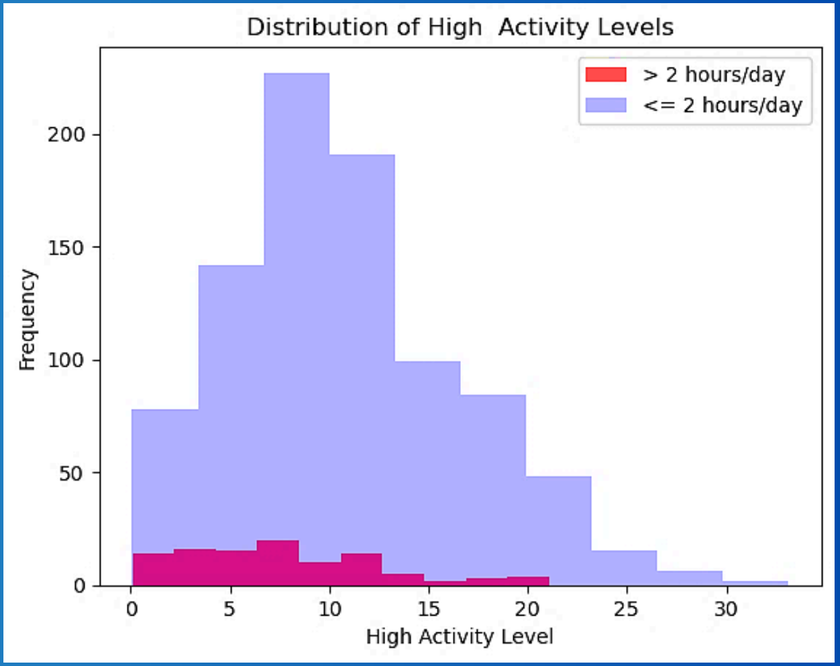
Result:

- *The group that used the internet for **less** than two hours clearly performed **better** on the test.*









It can be seen that the activity of those who use the Internet less is more.

More activities

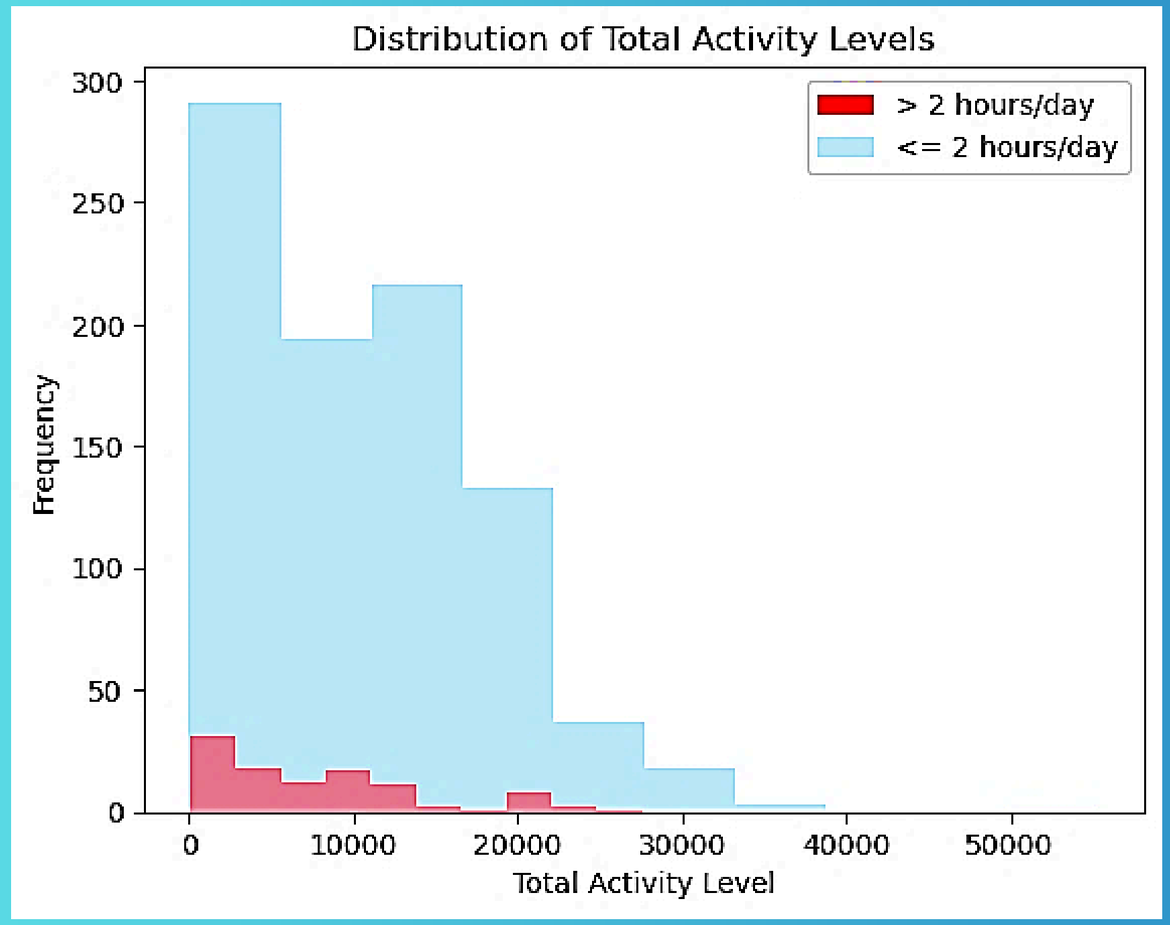
Both groups did not have a normal distribution and the Mann-Whitney test rejected the null hypothesis.

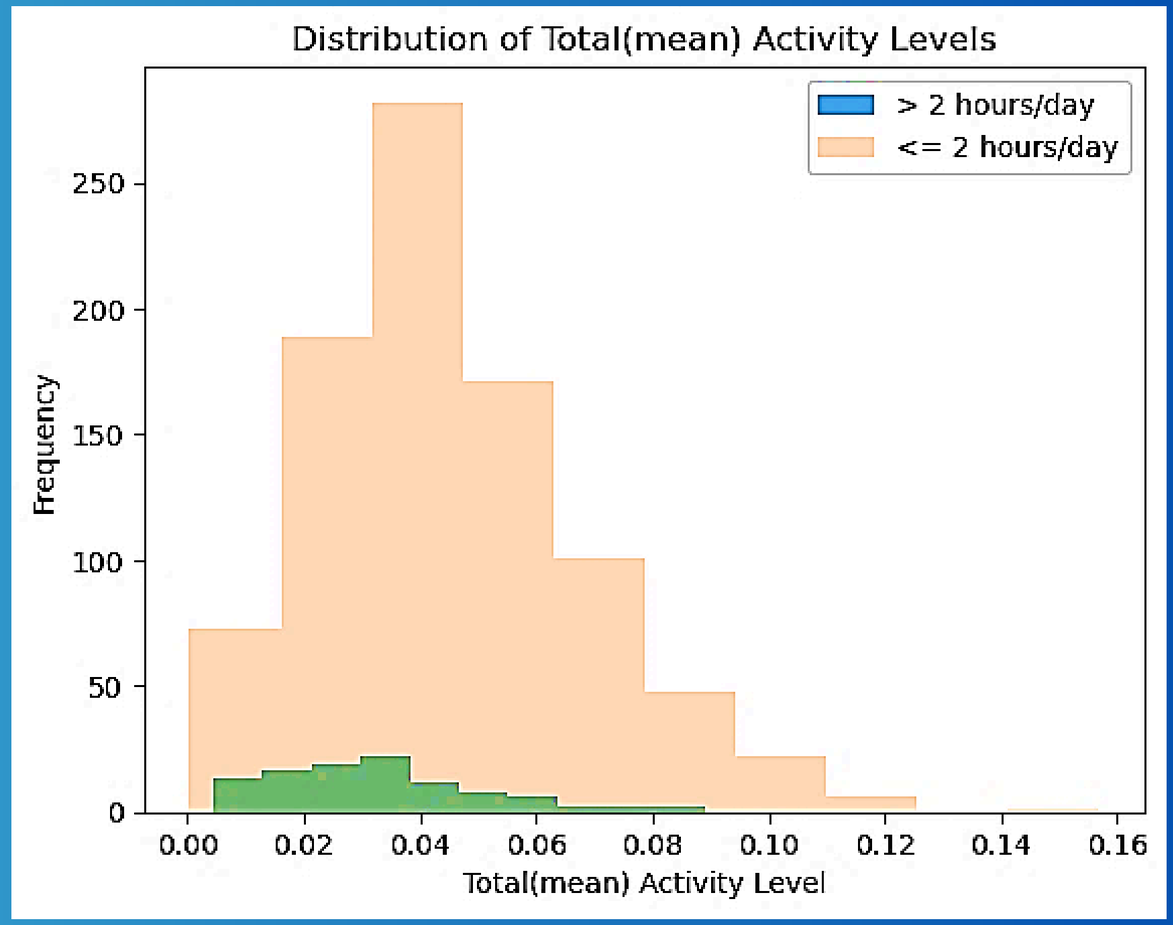
Normal activities

Fewer Activities

A group of data was normal, the result of the T-test and mann-withney U both rejected hypotheses.

T-test , Mann-Whitney U test



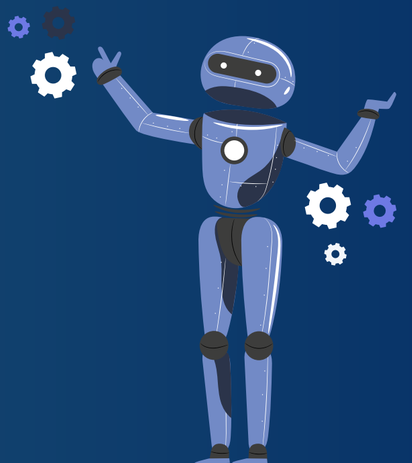


Overall, and on average, those who use the Internet less are more active.
The null hypothesis was also rejected by statistical tests and their activity levels are not equal.

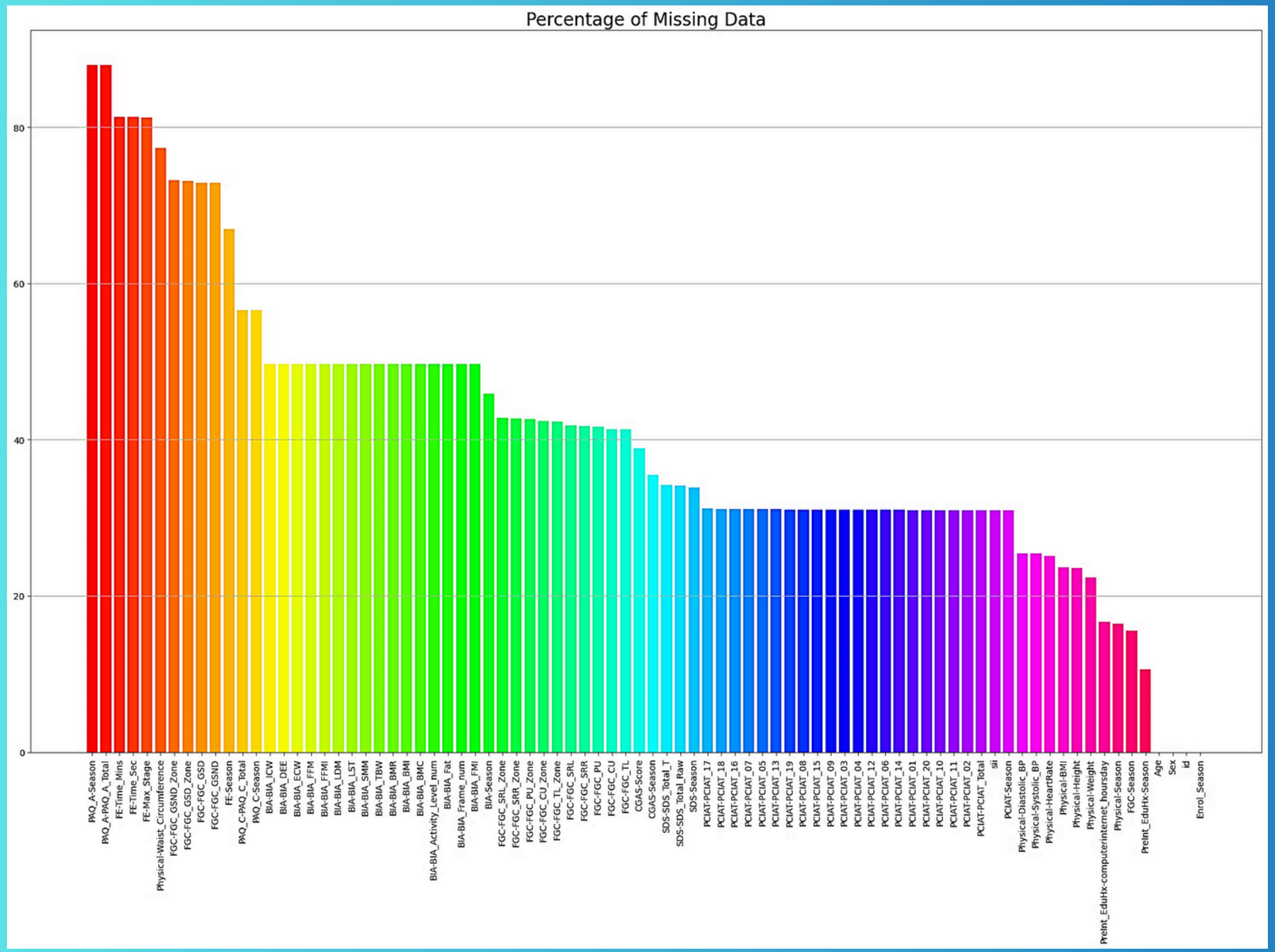
MACHINE LEARNING

PREPROCESSING

- **Objective**
  - predicting the severity of disorders caused by improper use of the Internet in children and adolescents
- **Statement of the problem:**
  - Analysis and evaluation of behavioral patterns
  - Assign the appropriate category to each participant
- **Importance of the issue:**
  - Knowing the factors affecting the occurrence of diseases such as anxiety, depression, ADHD
  - Help professionals to prevent
- **Data and information sources:**
  - Information about 4000 participants (5 to 22 years old)
    - Physical data: height, weight, blood pressure, heart rate, etc.
    - Behavioral data: Internet usage hours and...
    - Wristband data: acceleration, ambient light, battery voltage, etc.



PREPROCESSING



Exploratory Data Analysis (EDA)

PREPROCESSING

- **Data Pre-processing:**
  - Deleting unnecessary columns
  - Fill in the missing data
  - Converting object data to numerical data

- **Engineering Features:**
  - Health Index
  - Activity-to-Age Ratio
  - Average Weekly Activity

TRAINING

- **Pipeline:**
  - Filling in missing data
  - Standardization of the data
- **Feature selection and dimension reduction:**
  - RFE - *Recursive Feature Elimination*
  - PCA - *Principal Component Analysis*
- **Data balancing:**
  - SMOTE
  - SMOTE Tomek
- **Model:**
  - Random Forest
  - Classifier
- **Hyperparameter tunning:**
  - GridSearchCV

```
# Tuning process of hyperparameters
param_grid = {
    'n_estimators': [250],
    'max_depth': [9],
    'min_samples_split': [2],
    'min_samples_leaf': [4],
    'max_leaf_nodes': [None],
}
```

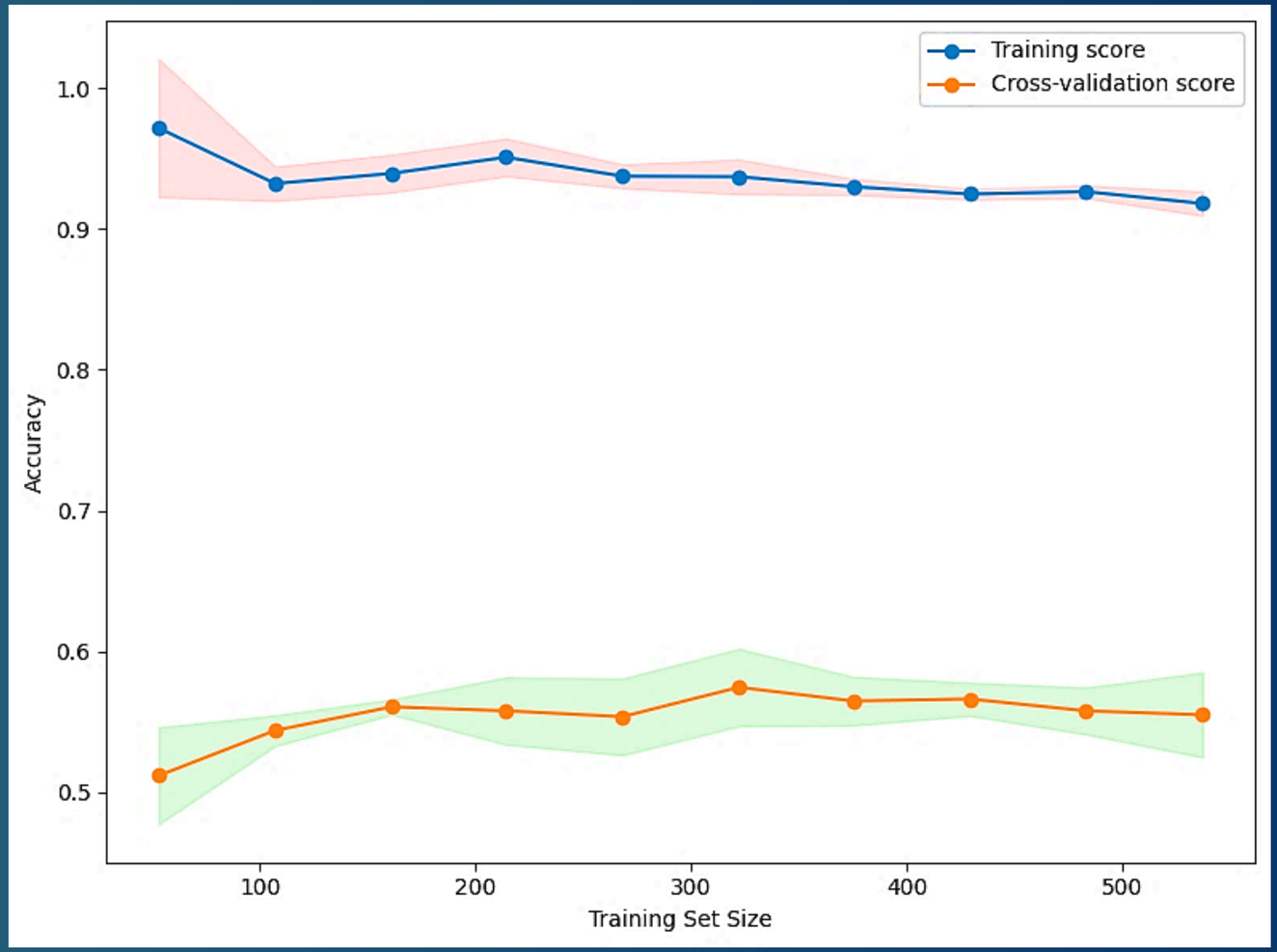
EVALUATION

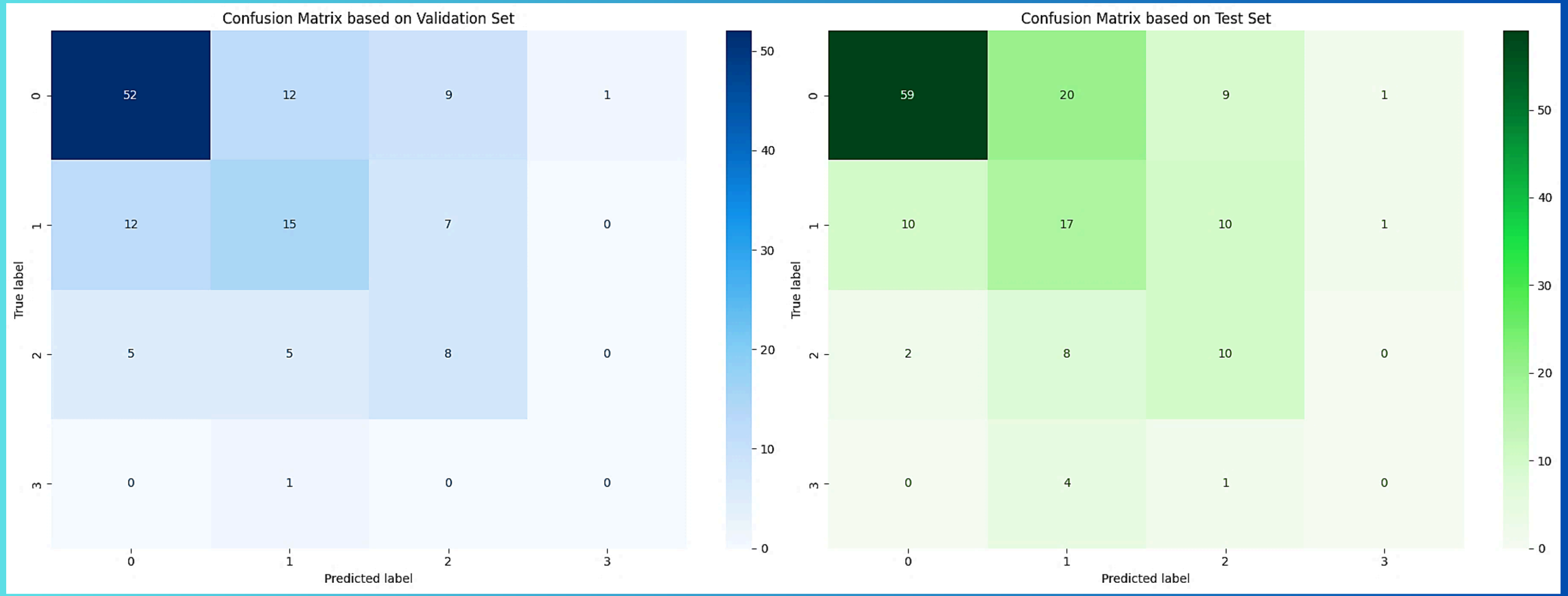
Evaluation metrics:

- Accuracy
- Precision
- AUC
- Recall
- F1-Score

Validation Accuracy	0.59
Validation Precision	0.60
Validation Recall	0.59
Validation F1 score	0.59
Validation AUC	0.75

Test Accuracy	0.56
Test Precision	0.61
Test Recall	0.56
Test F1 score	0.58
Test AUC	0.69





Confusion Matrix

02.

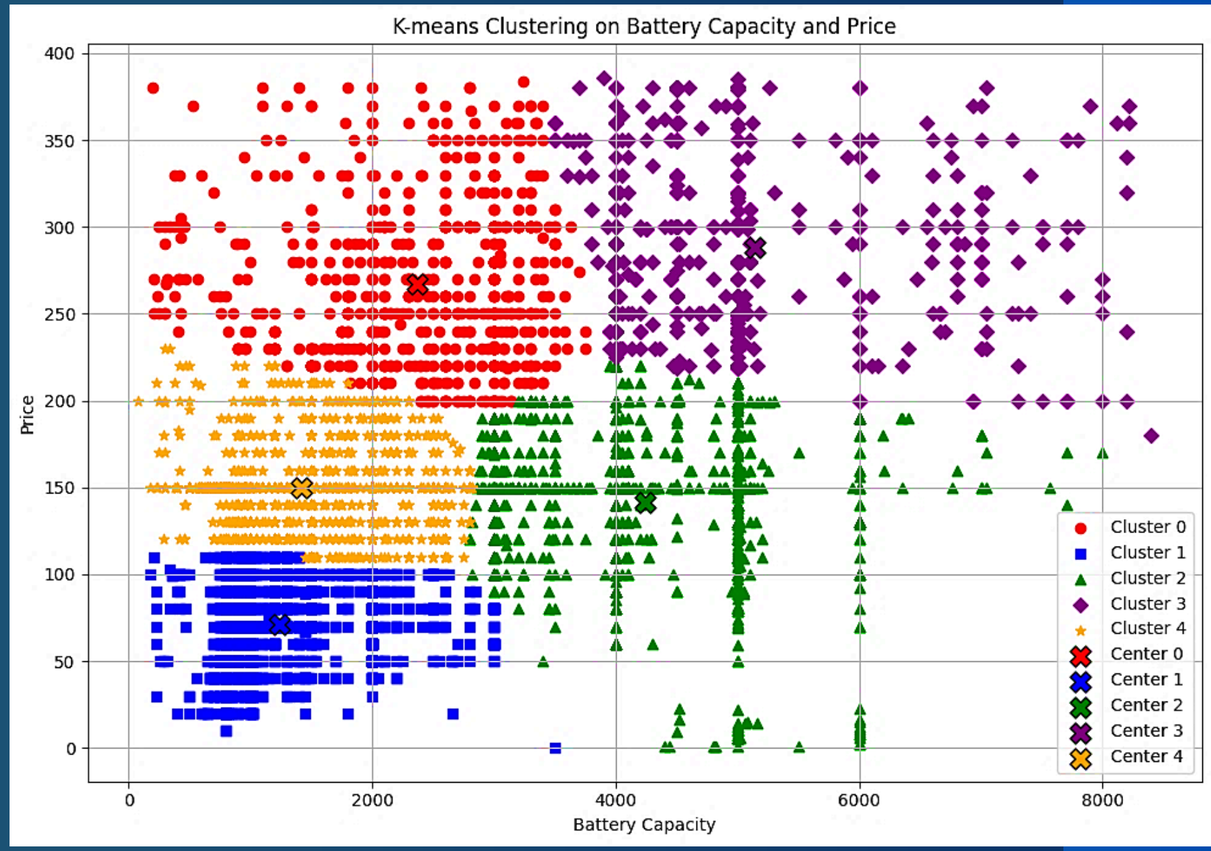
REVIEW OF MOBILE PHONE MARKET DATA

CLUSTERING

K-MEANS

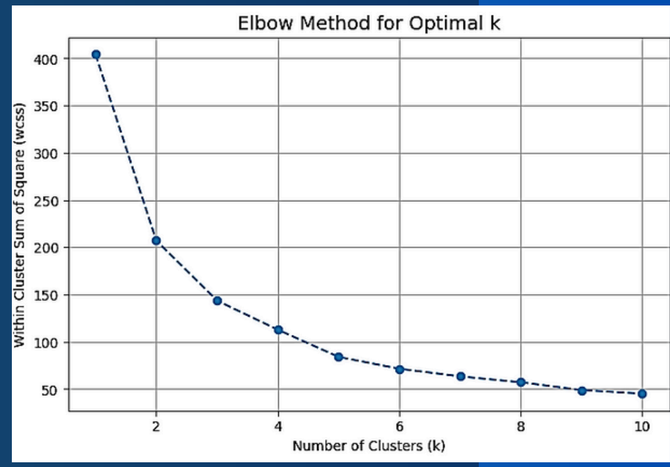
- After extracting numerical values from the two features of battery capacity and price of the devices, missing values are replaced with the median, and outliers are removed.
- The K-Means algorithm is run for 5 clusters on the created dataset.

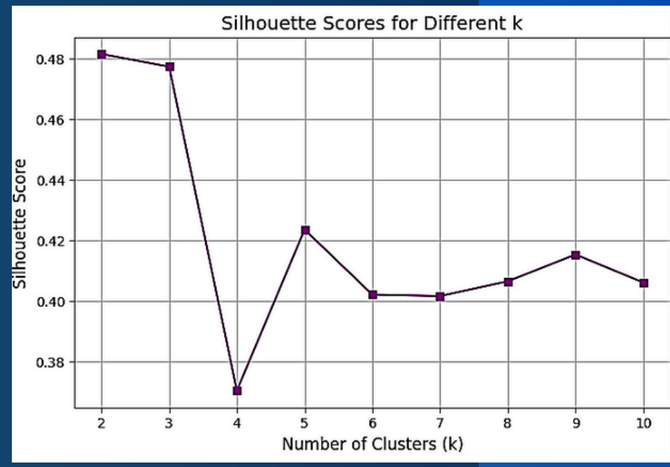
- Data with low battery capacity and low price.
- Indicates simple and economical devices.
- Data with high battery capacity and medium price.
- Devices with large battery capacity but
- are in the middle of the price range.
- Data with very high battery capacity and various prices (from medium to high).
- Indicates advanced devices with excellent battery capacity.



K-MEANS

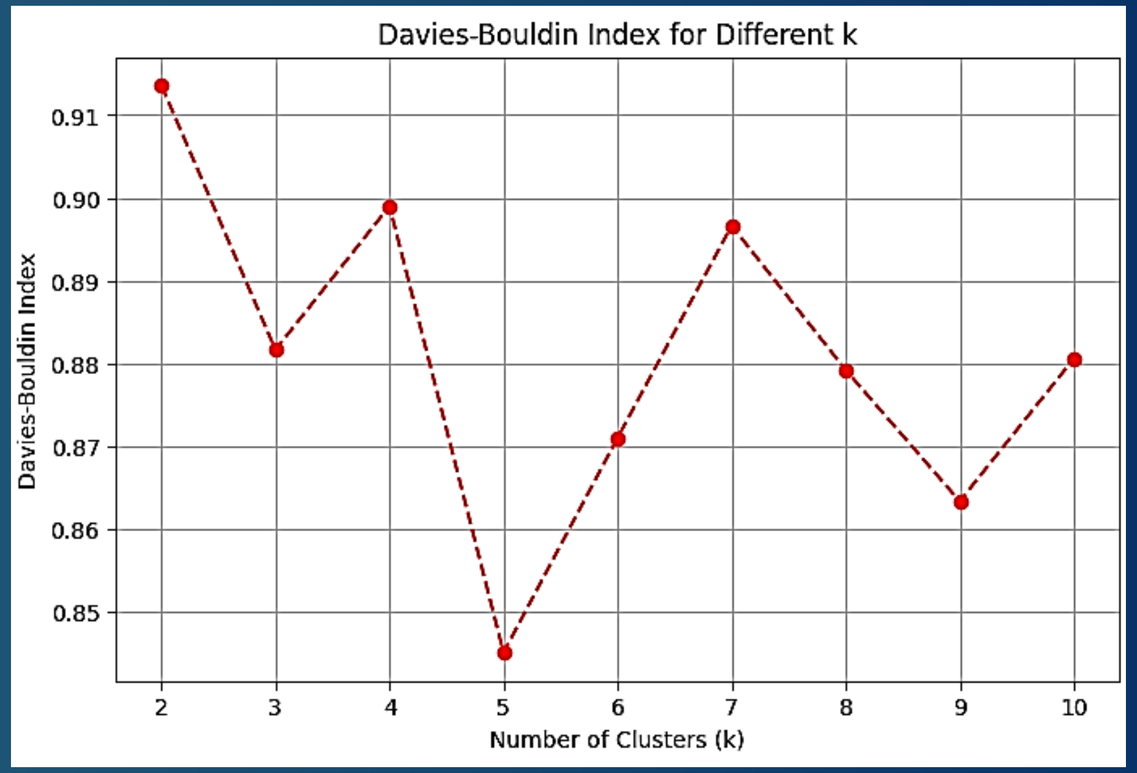
1. The decrease in WCSS value occurs rapidly at first, but from about five clusters onwards, the rate of decrease decreases significantly.
2. A significant decrease is also observed in the three-cluster case, which can be considered, as a suitable option for choosing the number of clusters.
3. The silhouette score in the two-cluster case has the highest value (approximately 0.48), which indicates the desirable quality of clustering in this case.
4. However, the silhouette score in the three-cluster case also remains close to the value of the two-cluster case (approximately 0.48) and, therefore, can be considered, as a significant option for choosing the number of clusters.
5. The five-cluster case with a score of about 0.42 still deserves further investigation.
6. To increase the accuracy and precision of the results, supplementary methods have been used below.





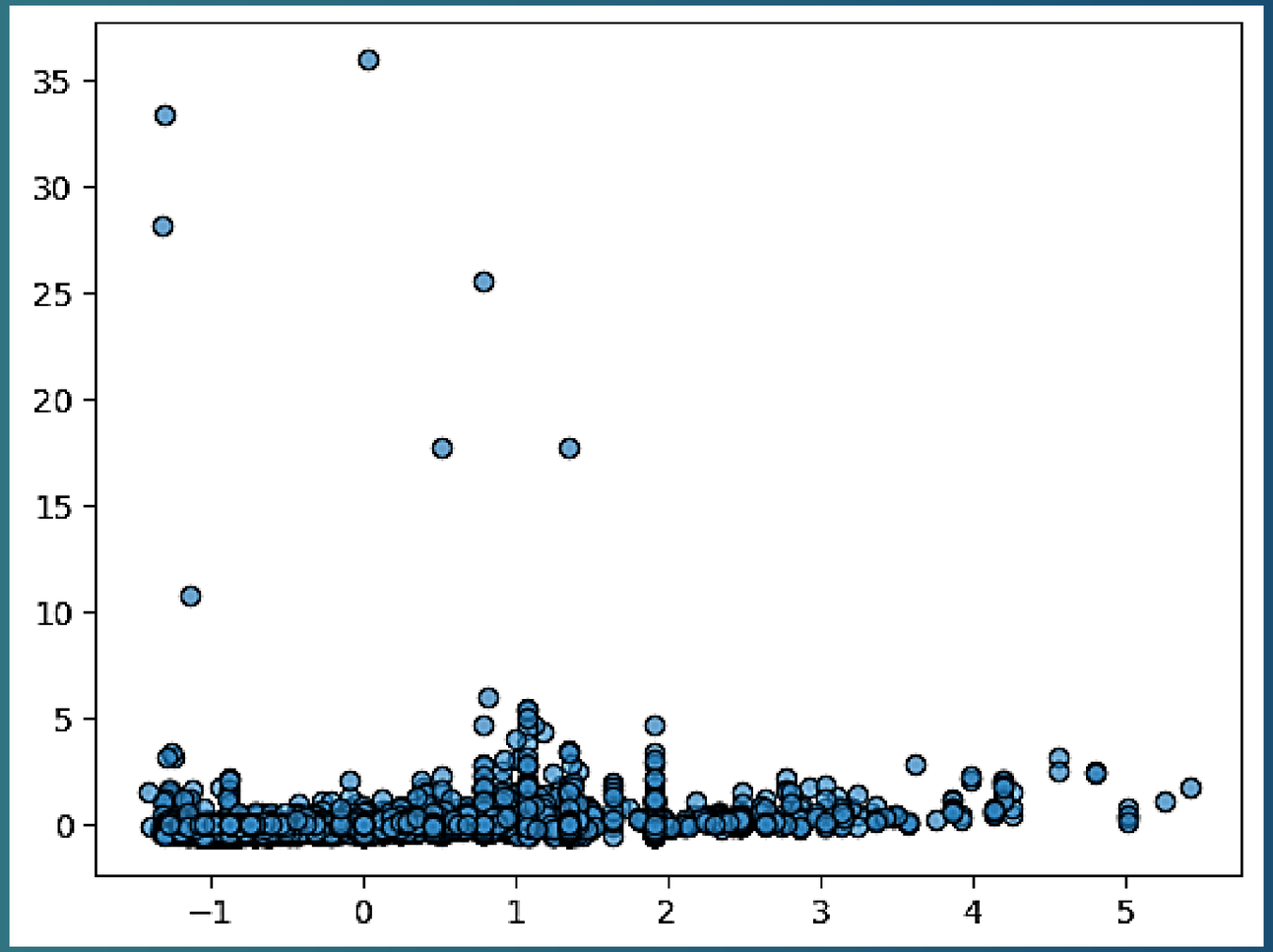
K-MEANS

- One complementary method that can be used to assess the quality of clustering is the Davies-Bouldin Index.
- In general, this index provides the average of the similarity measures of each cluster to the cluster to which it is most similar.
- The best number of clusters is when the average similarity is minimized. Therefore, the lower the value of this index, the more appropriate the clustering performed.



According to the figure above, it can be clearly and confidently concluded that the number of five clusters is the most appropriate choice for clustering.

DBSCAN



Data normalization and plotting

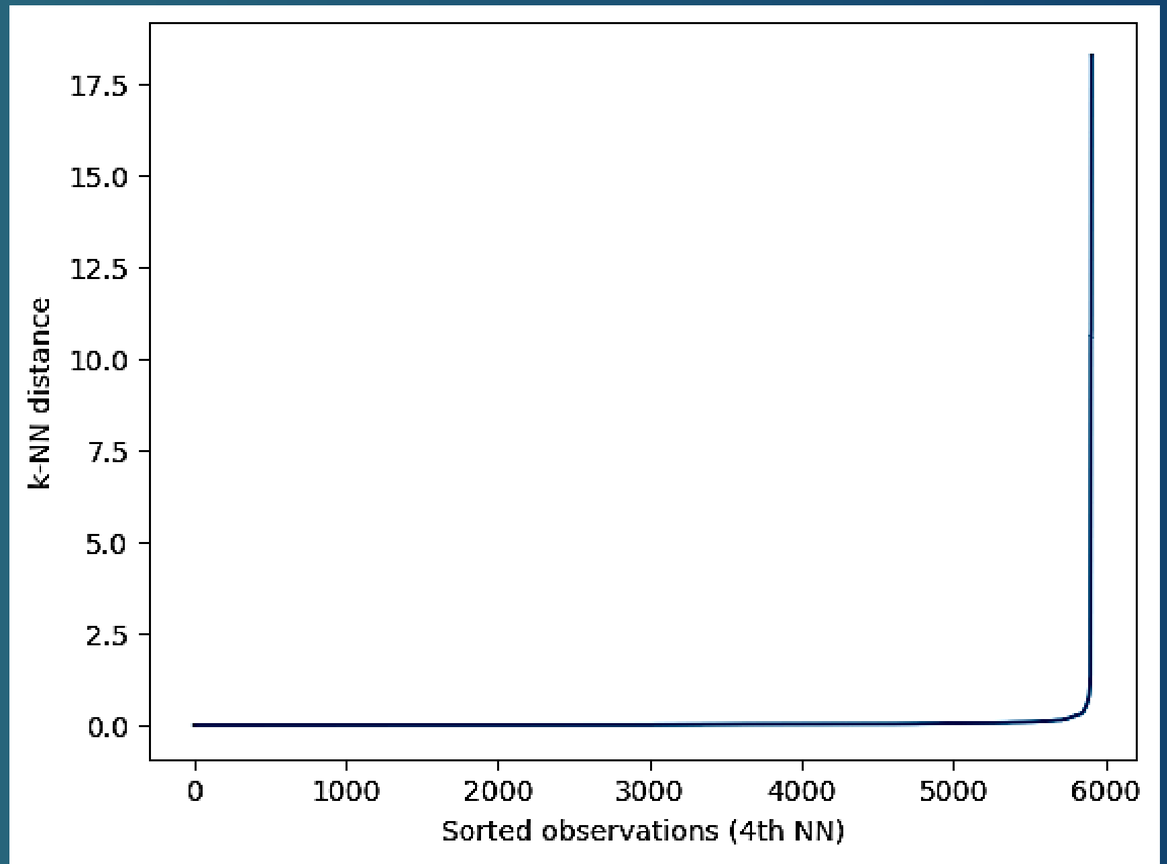
StandardScaler

MinMaxScaler

DBSCAN

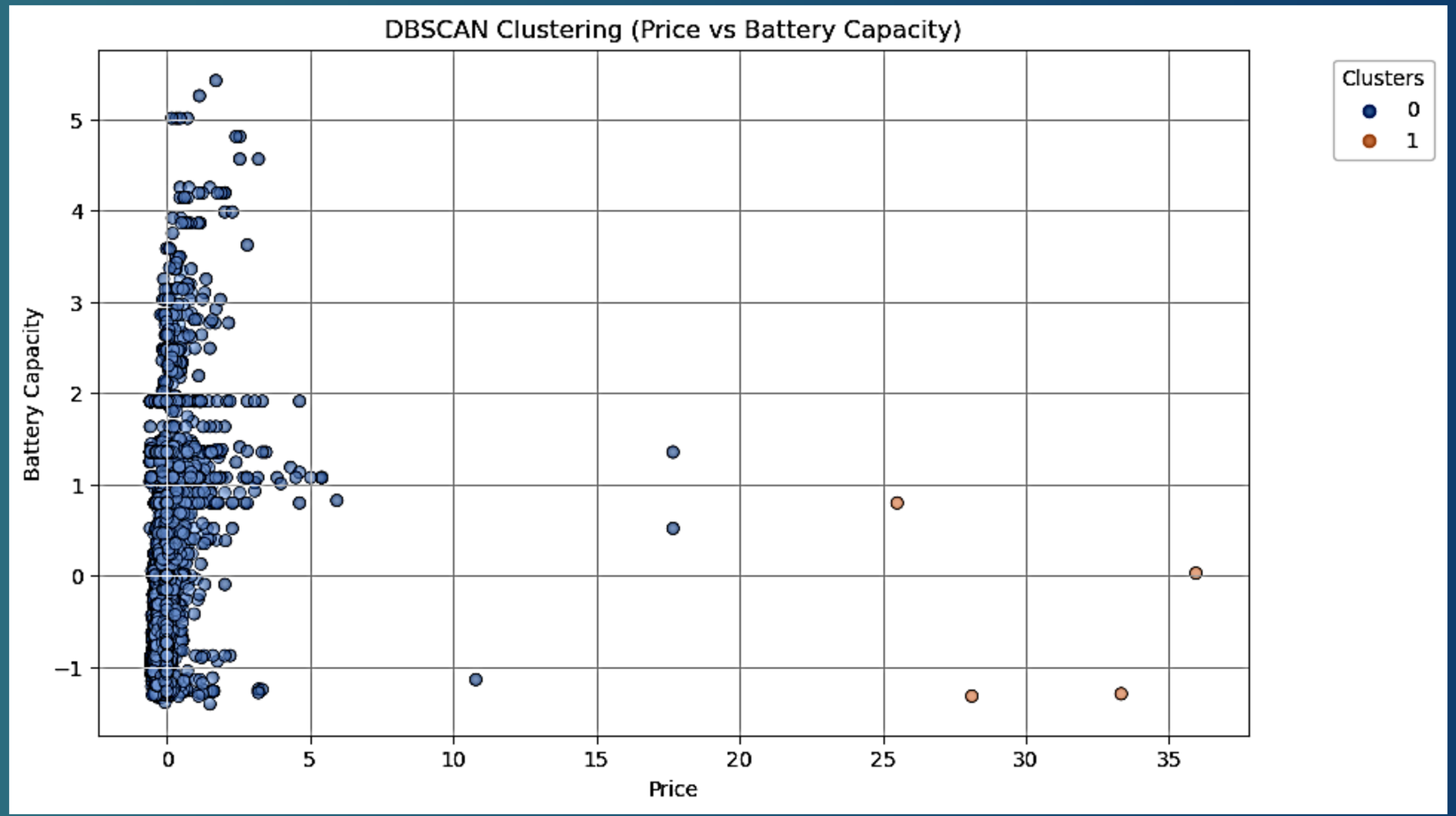
Determining the optimal value of the EPS parameter in the DBSCAN algorithm:

By drawing a nearest neighbor distance plot (k-distance plot), the point where a sudden change (knee) occurs suggests the appropriate value of EPS.



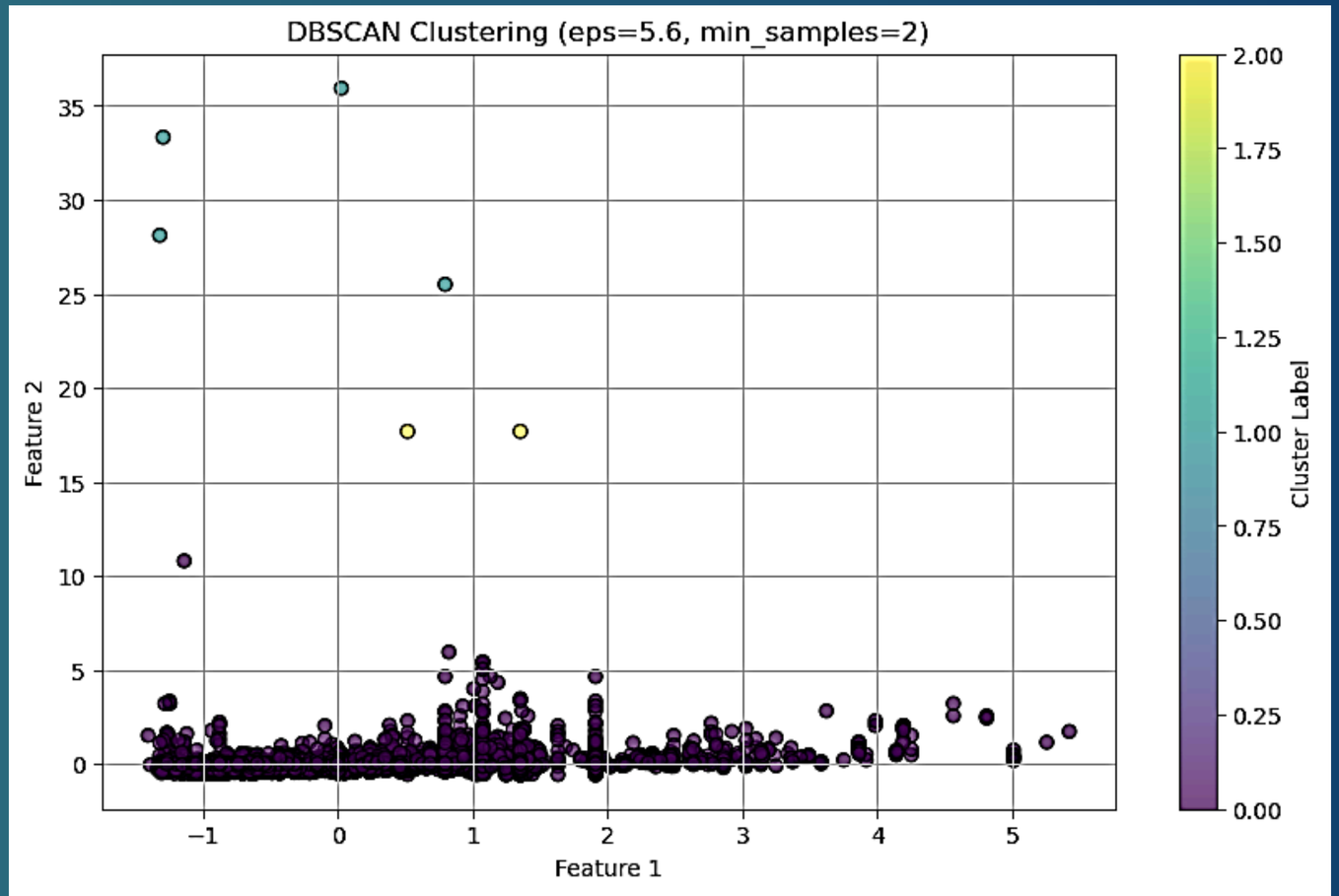
DBSCAN

- **Silhouette Score:** *0.9589856188211109*
- **Davies-Bouldin Index:** *0.16007791871233792*
- **Clusters:** *2*



DBSCAN

- **Best Params - EPS: 5.6, min_samples:** *2*
- **Silhouette Score: 0.9294**
- **Noise Ratio:** *0.00%*
- **Number of Clusters:** *3*



PRICE PREDICTION

WORKFLOW

Understanding the inquiry

Dataset review

Information on more than 5900 mobile phones

Exploratory Data Analysis (EDA)

- *Drawing diagrams*
- *Key insights*

Data preprocessing

- *Missing Value Check*
- *Feature Engineering*
- *Data Partitioning*

Model development

- *Models Used*
- *Workflow*
- *Measurement Criteria*

Analysis of results

Comparison of various models

UNDERSTANDING THE INQUIRY

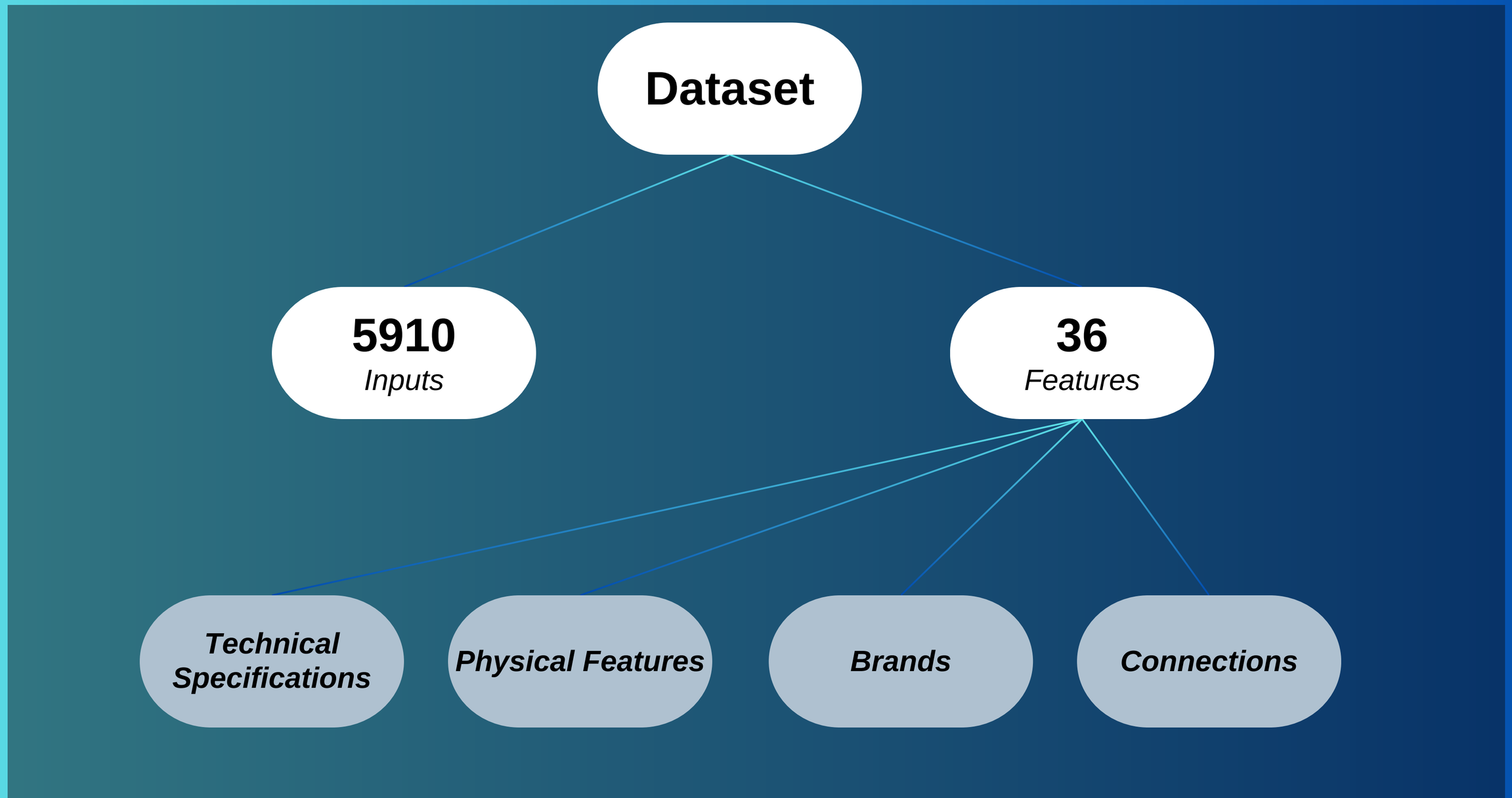
Issue

Difficulty predicting prices with real data

Target

Building and training a machine learning model to predict accurately

EXPLORATORY DATA ANALYSIS (EDA)



DATA PREPROCESSING

Simple imputer

Linear regression

Forward & backfill

One-hot & label encoding

QUESTIONS...?

Which technique for transforming categorical data will give us better results, label encoding or one-hot encoding?

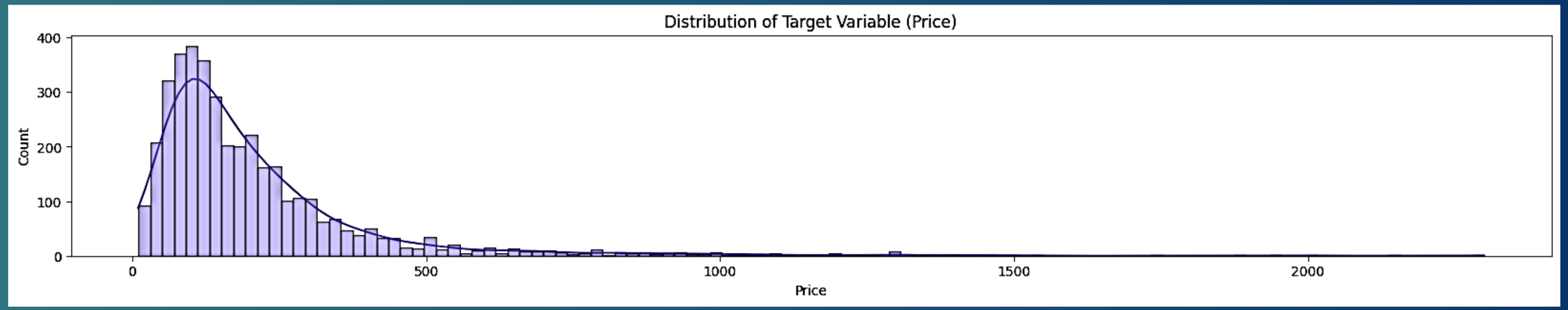
Question

Question

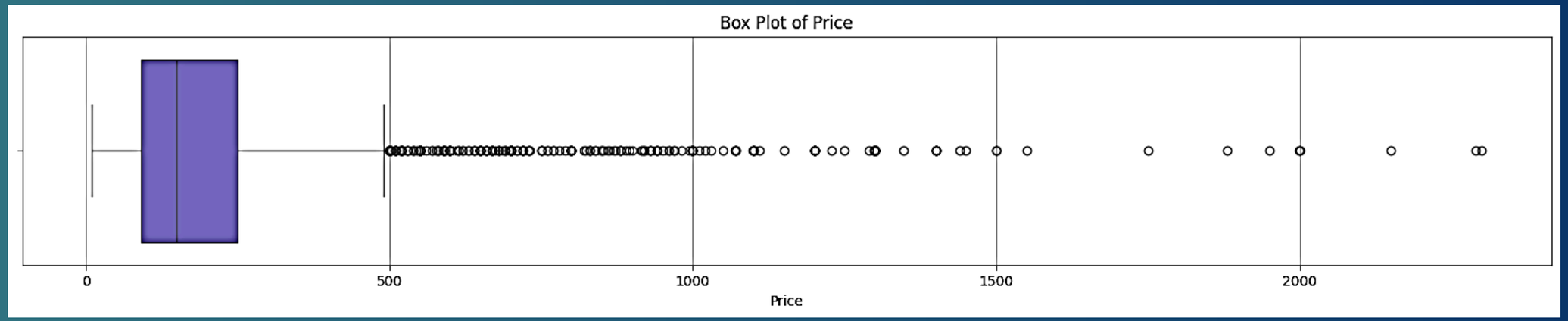
How do we fill in this amount of missing data in important columns like CPU, GPU, and Chipset?

EXPLORATORY DATA ANALYSIS (EDA)

Histogram

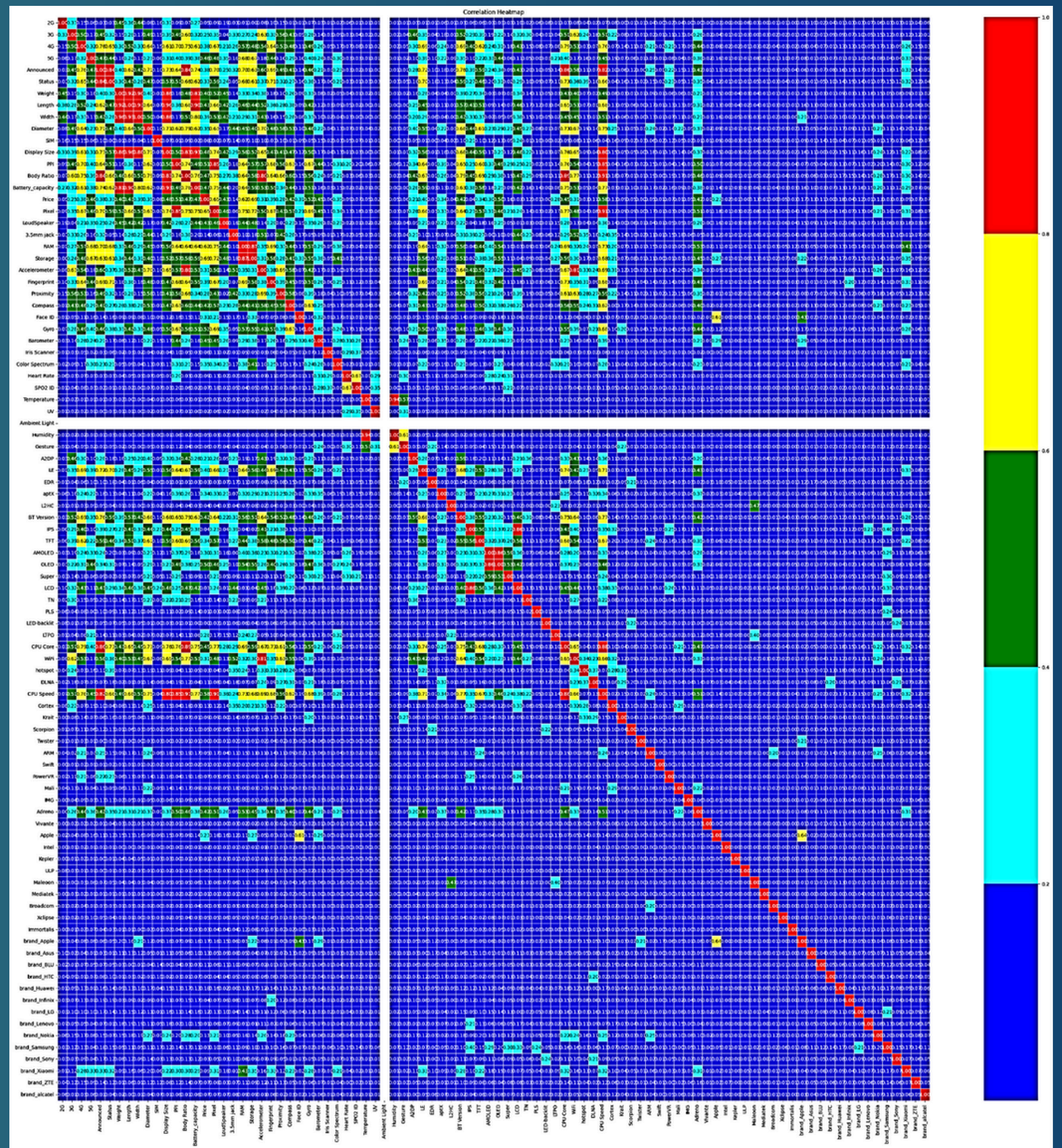


Boxplot



EXPLORATORY DATA ANALYSIS (EDA)

HEATMAP



PROBLEM / SOLUTION

Problem

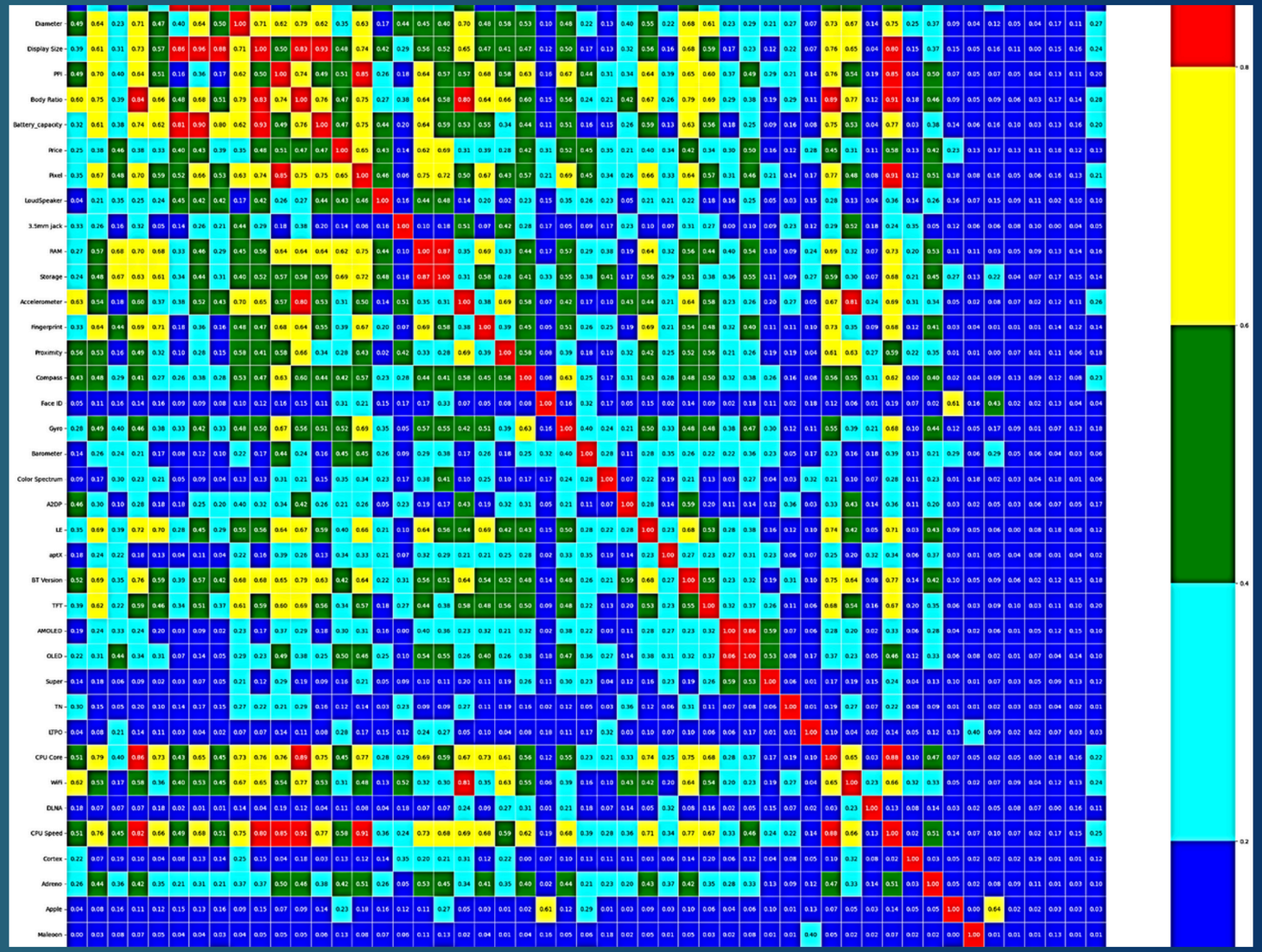
Lots of features to handle

Set a threshold and drop features whose correlation value with the target column is less than the threshold.

Solution

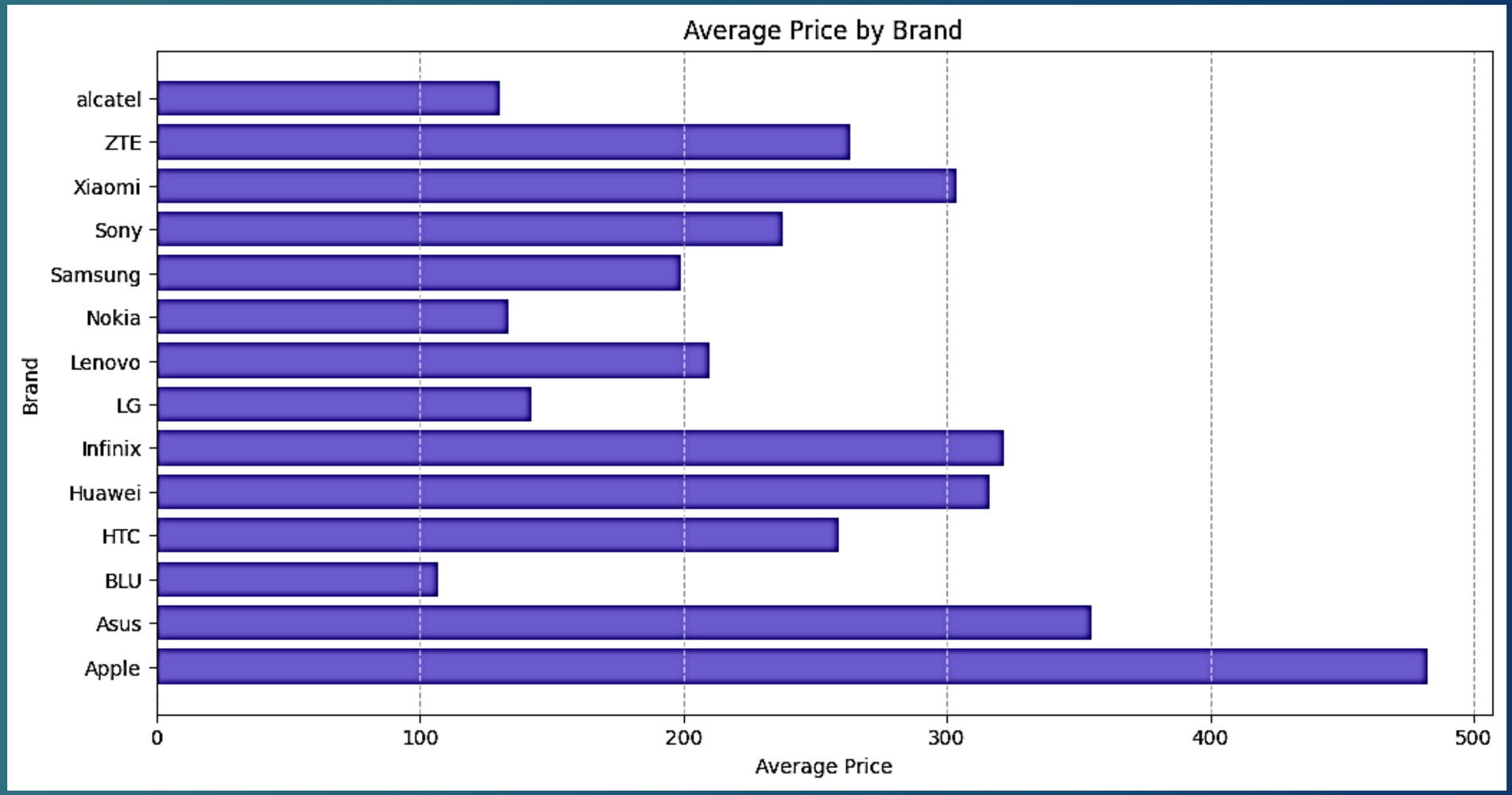
EXPLORATORY DATA ANALYSIS (EDA)

- *Remove features that have a correlation of less than 0.1 with the target.*
- *Remove 40 features (about 50 features remain.)*



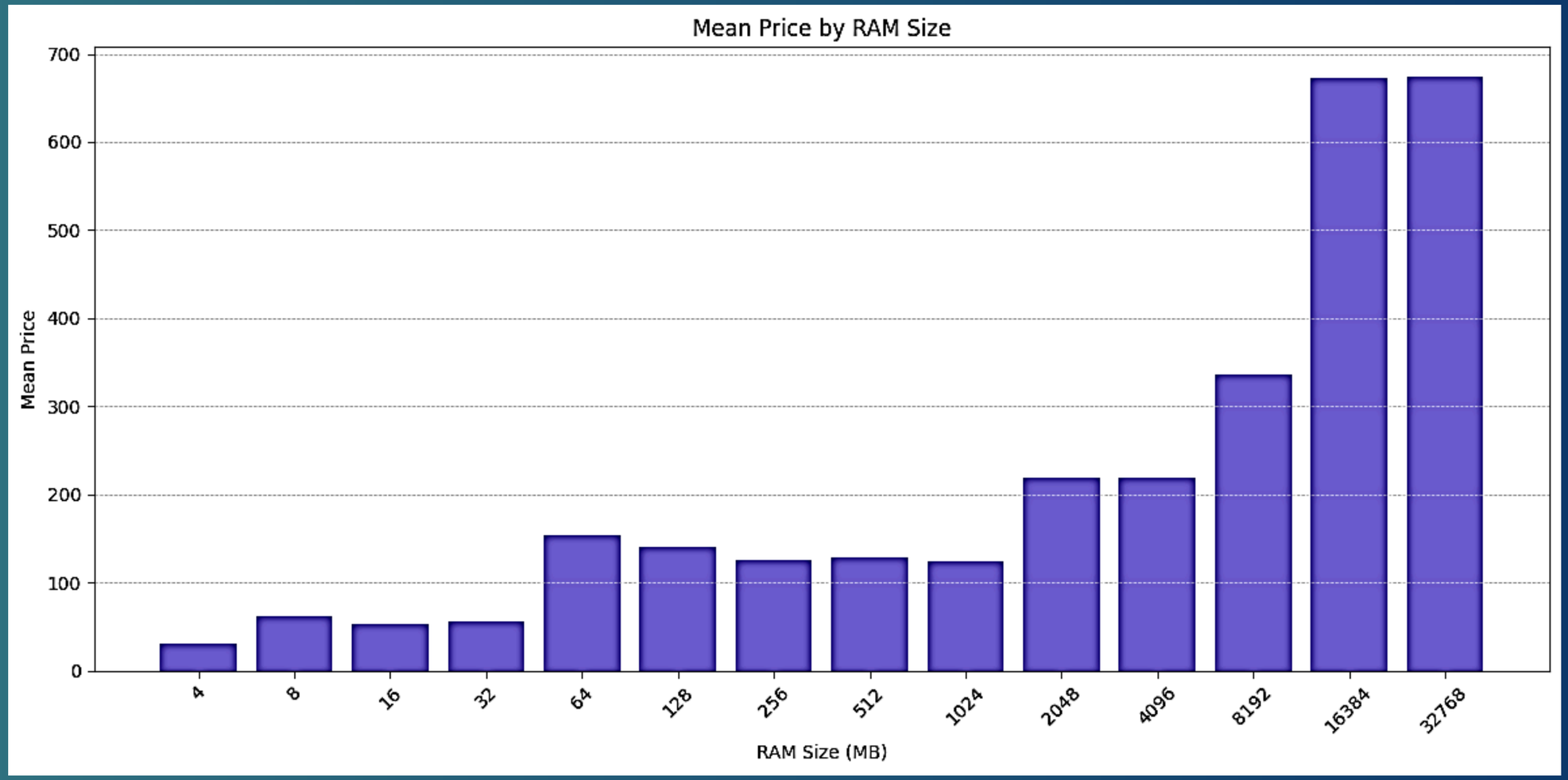
EXPLORATORY DATA ANALYSIS (EDA)

BARPLOT



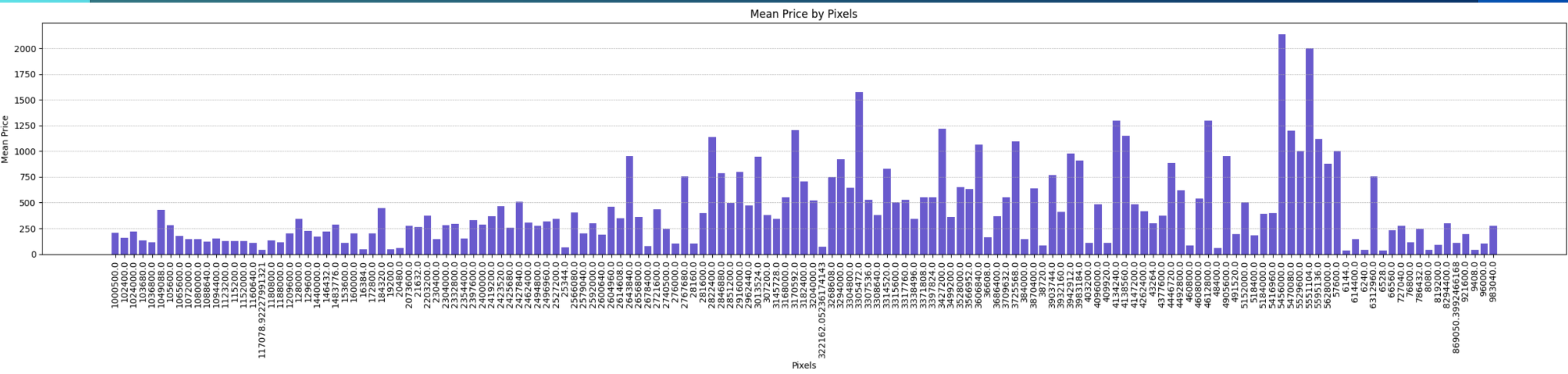
EXPLORATORY DATA ANALYSIS (EDA)

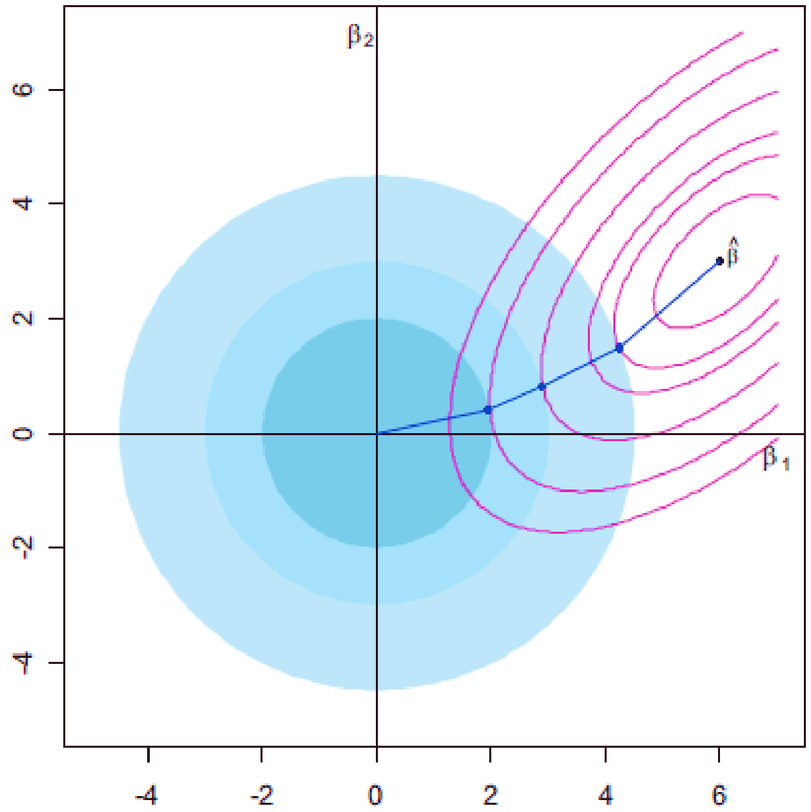
BARPLOT



EXPLORATORY DATA ANALYSIS (EDA)

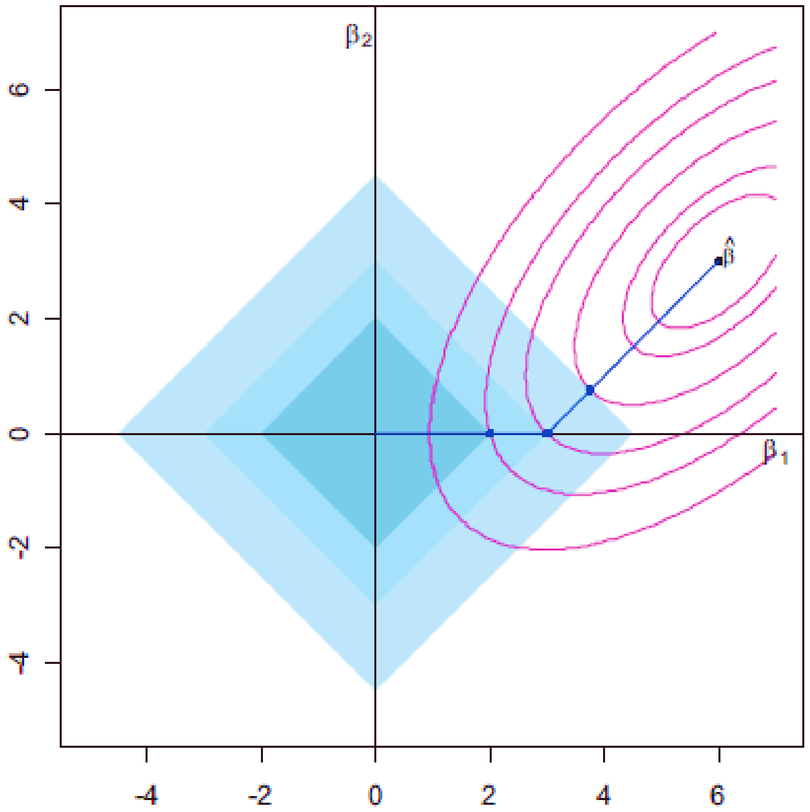
BARPLOT





RIDGE REGRESSION

In Ridge regression, a penalty equal to the square of the coefficient value is added to the .loss function

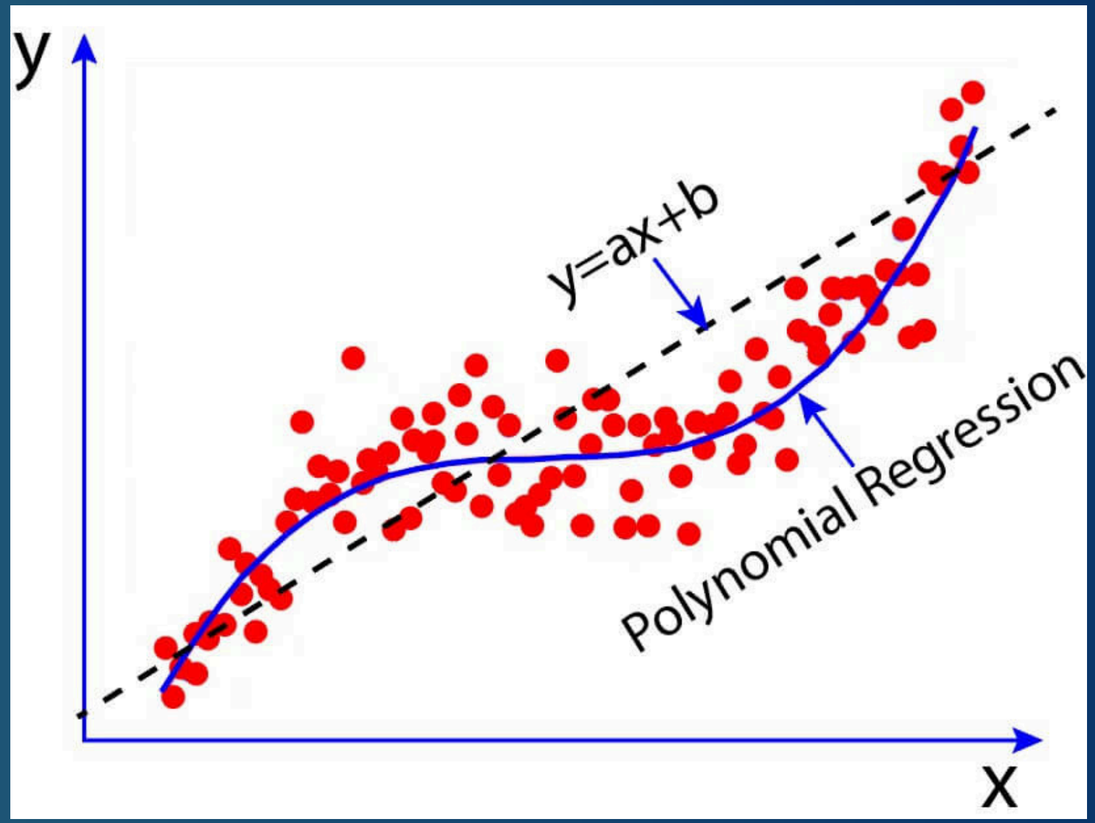


LASSO REGRESSION

Lasso stands for "Least Absolute Shrinkage And Selection Operator", which adds a penalty to the .absolute value of the loss function

POLYNOMIAL FEATURES

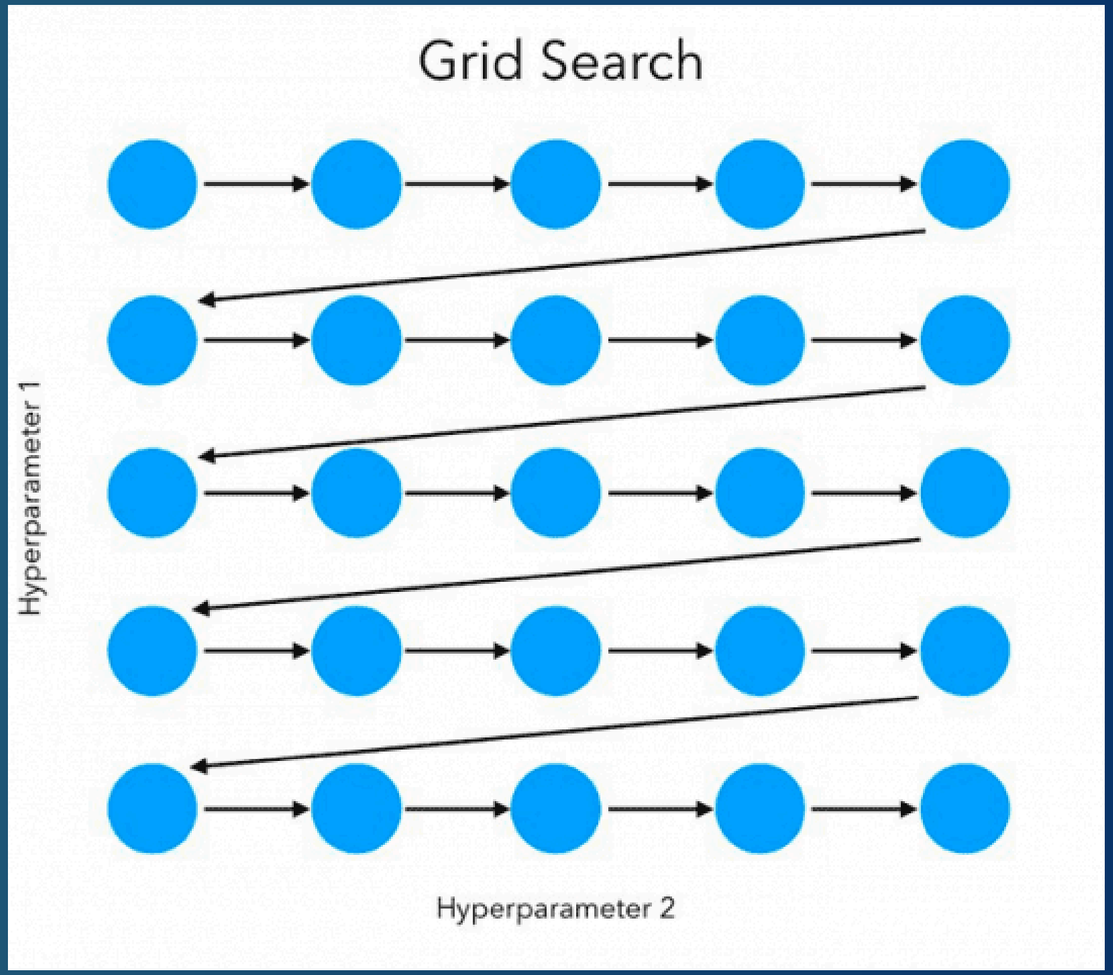
This technique helps to find nonlinear features between input variables and target variables.



HYPERPARAMETER TUNING

GRID SEARCH

Hyperparameter optimization is a technique used in Machine Learning to determine the best combination of hyperparameters for a model.

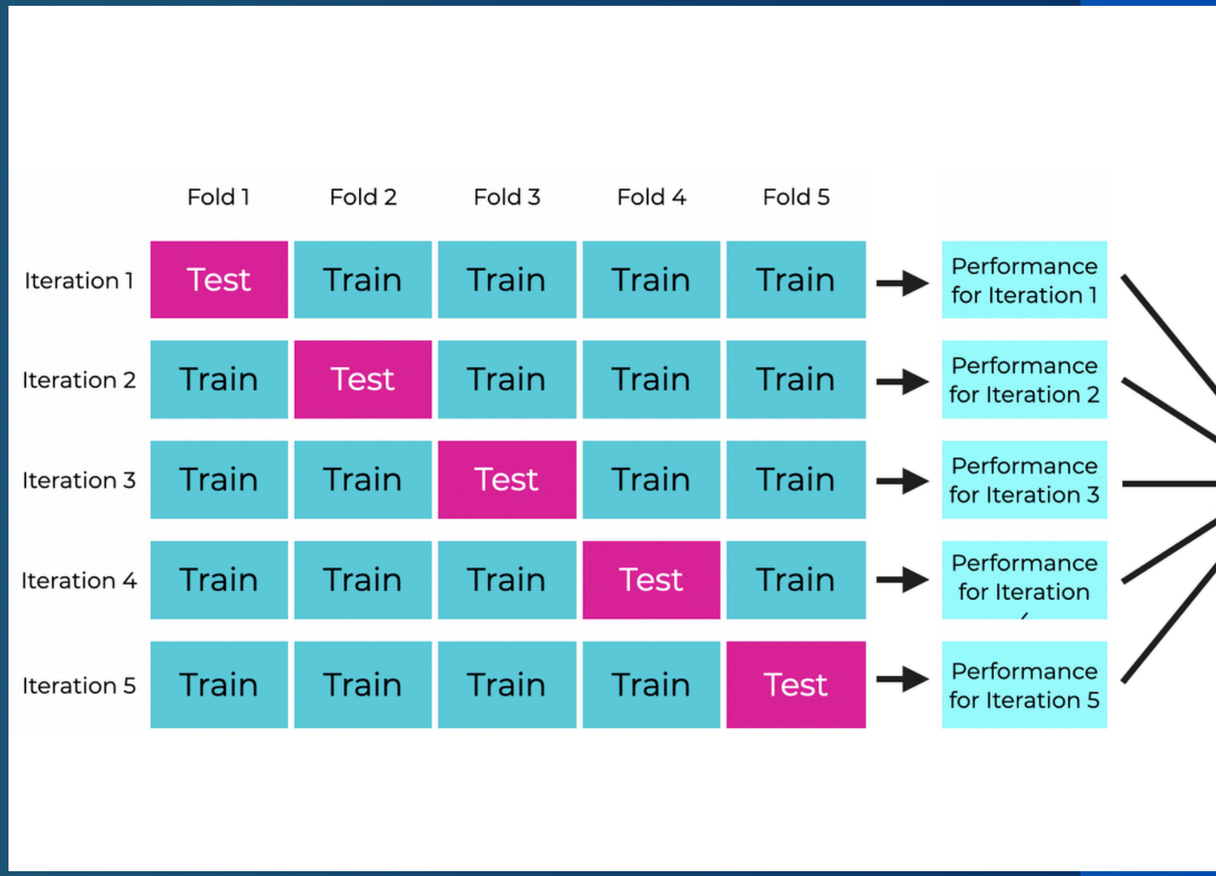


CROSS VALIDATION

It helps evaluate model performance and ensure that, it does not overfit the training data.

K-FOLD

The dataset is divided into 'k' subsets (or folds). The model is trained on k-1 subsets and tested on the remaining subsets.



ANALYSIS OF RESULTS

RIDGE WITHOUT REMOVING COLUMNS

```
MAE: 56.66
MSE: 11265.22
RMSE: 106.13
R-squared: 0.7349
```

RIDGE

```
MAE: 56.74
MSE: 11256.11
RMSE: 106.09
R-squared: 0.7351
```

```
{'poly__degree': 3, 'ridge__alpha': 30}
Cross-Validation Results:
Mean RMSE: 97.62
Standard Deviation of RMSE: 13.78
```

LASSO

```
MAE: 56.73
MSE: 11546.09
RMSE: 107.45
R-squared: 0.7283
```

```
{'lasso__alpha': 0.05, 'poly__degree': 3}
Cross-Validation Results:
Mean RMSE: 95.70
Standard Deviation of RMSE: 15.37
```

THANKS