Internship Report

27.05.2024 – 27.07.2024

AMPLY INNOVATIONS PRIVATE LIMITED

SUBMITTED BY

SOWMIYA P

EXERCISE DATA SCIENTIST

Task 1:

Choose your Own Dataset related to Exercise and Physical Therapy and Try to Analyse that Dataset based on this Parameters.

Your Analysis Metrics will be:

1. Finding Duplicates
2. Finding Irrelevant data
3. Data type Mismatch
4. Data Formatting
5. Finding Missing Value
6. Finding Null Values
7. Finding Outliers
8. Finding any Bias and Mitigate it.

Title:

Diabetes Prediction

Dataset Link:

https://www.kaggle.com/code/ihabsherbiny/diabetes-prediction-accuracy-99/input

Task Description:

The task involves developing a predictive model for Diabetes Prediction using a given dataset. The aim is to analyse the data, preprocess it, and apply machine learning techniques to predict the likelihood of Diabetes in individuals based on various health and demographic factors.

Description:

Welcome to the Diabetes Prediction Dataset, a valuable resource for researchers, data scientists, and medical professionals interested in the field of diabetes risk assessment and prediction. This dataset contains a diverse range of health-related attributes, meticulously collected to aid in the development of predictive models for identifying individuals at risk of diabetes. By sharing this dataset, we aim to foster collaboration and

innovation within the data science community, leading to improved early diagnosis and personalized treatment strategies for diabetes. Columns:

- Id: Unique identifier for each data entry.
- Pregnancies: Number of times pregnant.
- Glucose: Plasma glucose concentration over 2 hours in an oral glucose tolerance test.
- BloodPressure: Diastolic blood pressure (mm Hg).
- SkinThickness: Triceps skinfold thickness (mm).
- Insulin: 2-Hour serum insulin (mu U/ml). BMI: Body mass index (weight in kg / height in m^2).
- DiabetesPedigreeFunction: Diabetes pedigree function, a genetic score of diabetes.
- Age: Age in years.
- Outcome: Binary classification indicating the presence (1) or absence (0) of diabetes.

Objective:

The primary objective of this task is to create an accurate and reliable predictive model that can identify individuals at high risk of Diabetes. This model aims to assist healthcare professionals in early detection and prevention strategies, thereby reducing the incidence and impact of Diabetes.

Process:

1. Data Collection and Understanding:

- Import the dataset and understand its structure.
- Identify key features and the target variable.

2. Data Preprocessing:

- Handle missing values and outliers.
- Encode categorical variables.
- Scale numerical features for consistency.

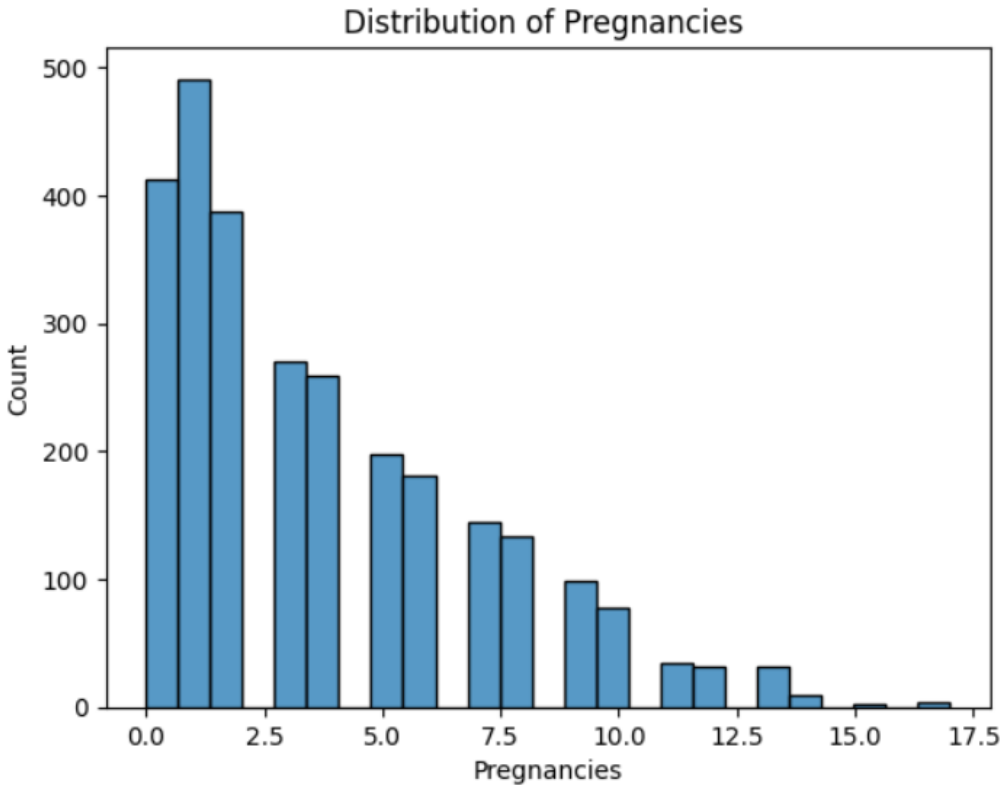
3. Exploratory Data Analysis (EDA):

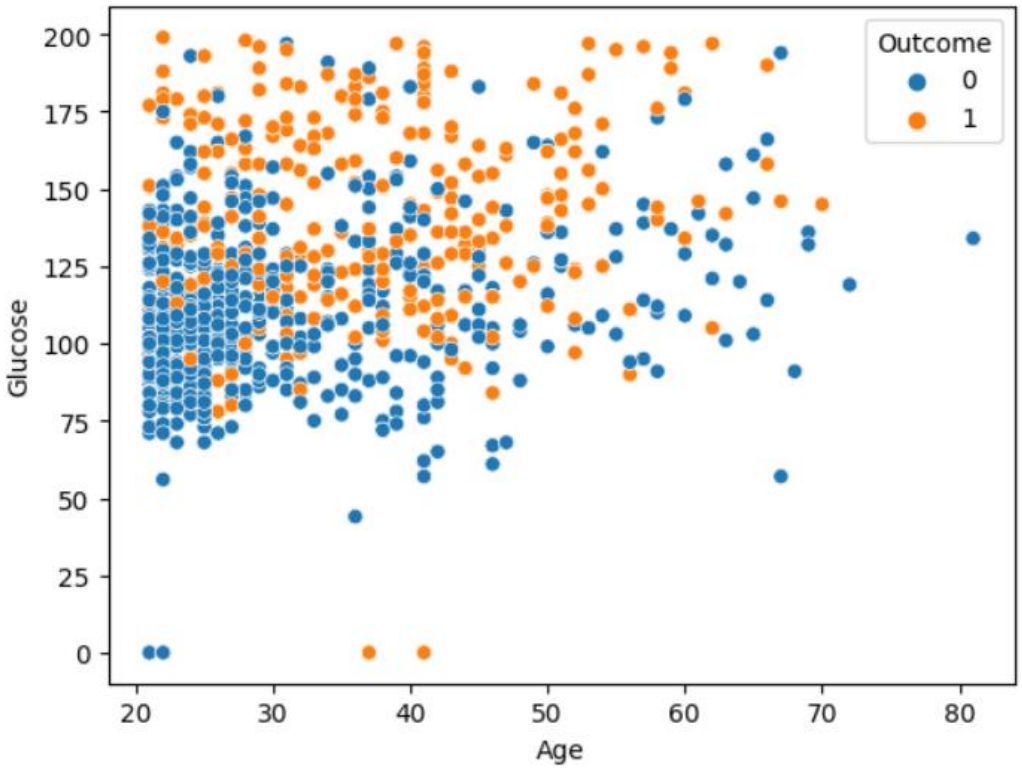
- Visualize data distributions and relationships between features.
- Analyze correlations and patterns

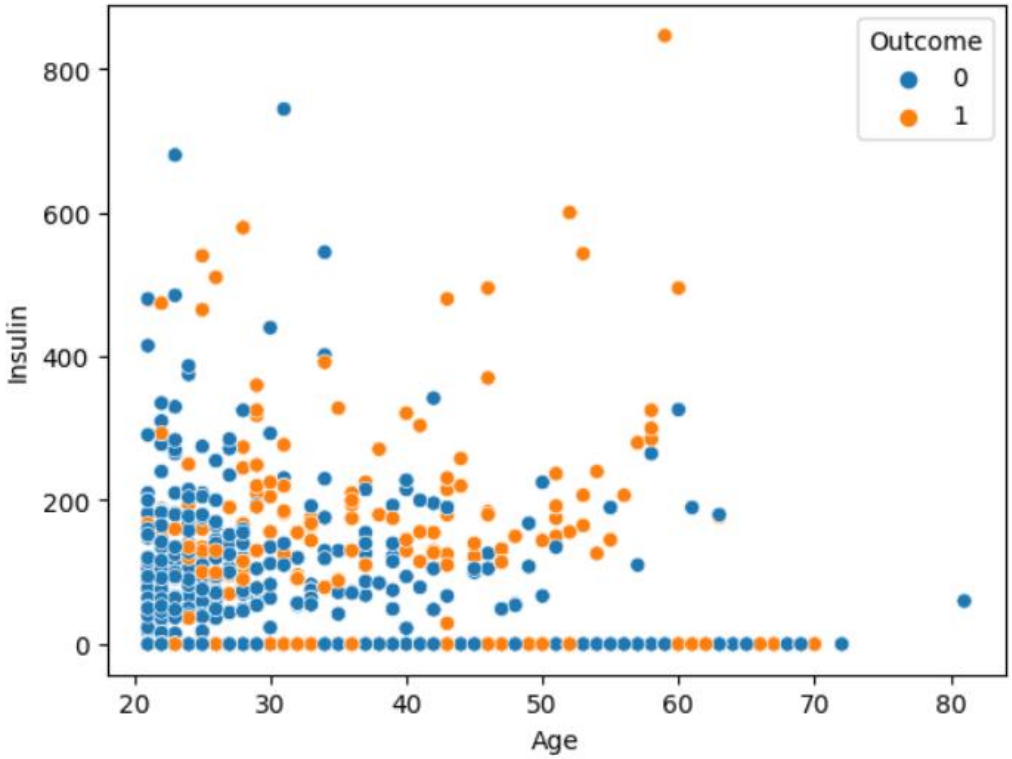
Outcome:

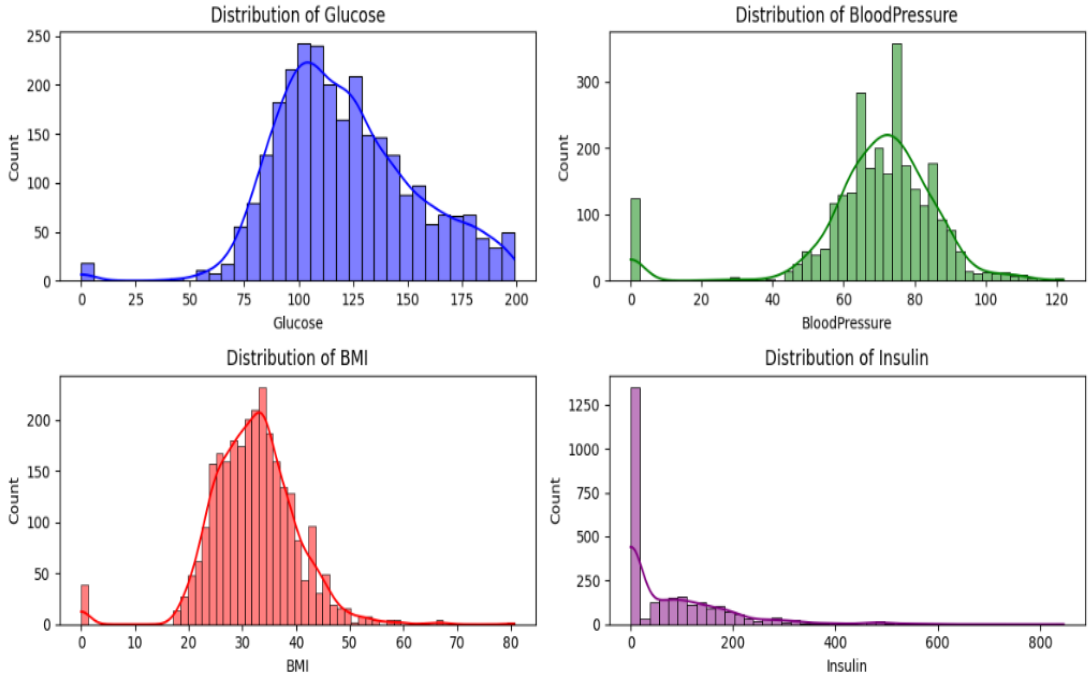
- Binary classification indicating the presence (1) or absence (0) of diabetes.
- The outcome of this task is predicting the Diabetes based on various factors such as age, gender, BloodPressure, Insulin, BMI, Pregnancies using various visualization plots.

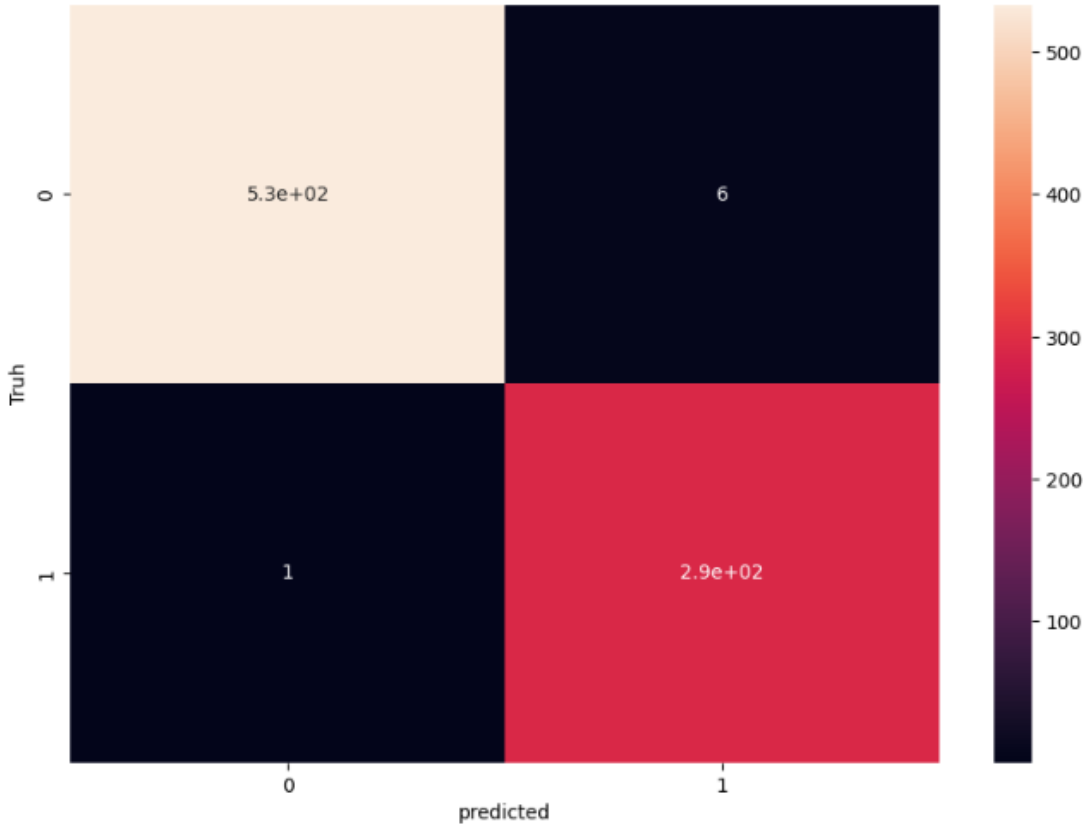
Output Screenshots:

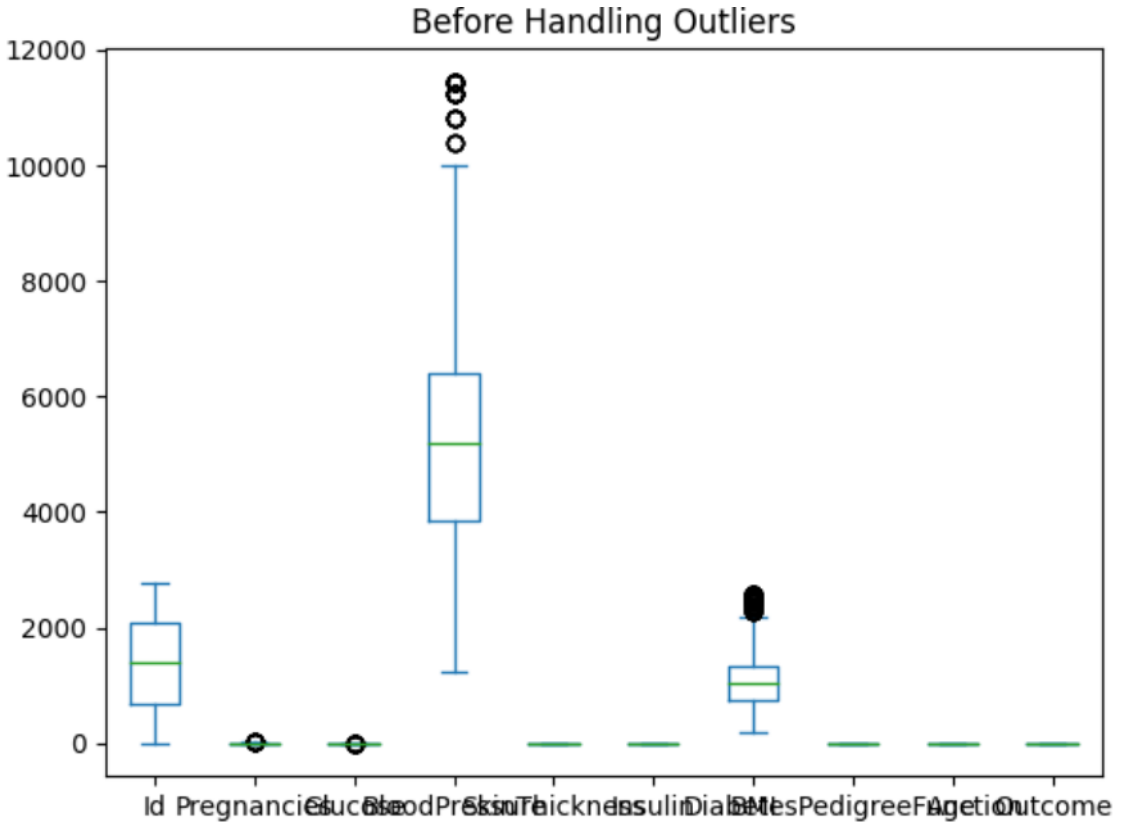


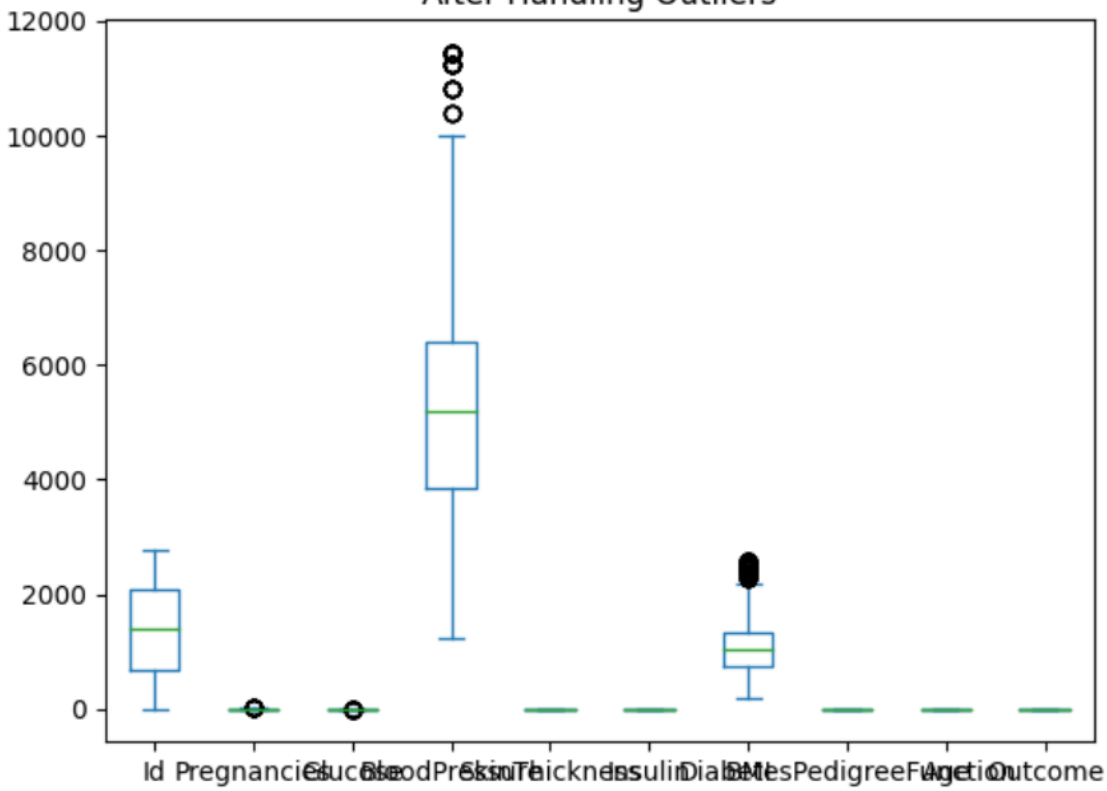

Outcome
0
1
Glucose
Age
200
175
150
125
100
75
50
25
0
20
30
40
50
60
70
80


Outcome
0
1
Insulin
Age
800
600
400
200
0
20
30
40
50
60
70
80

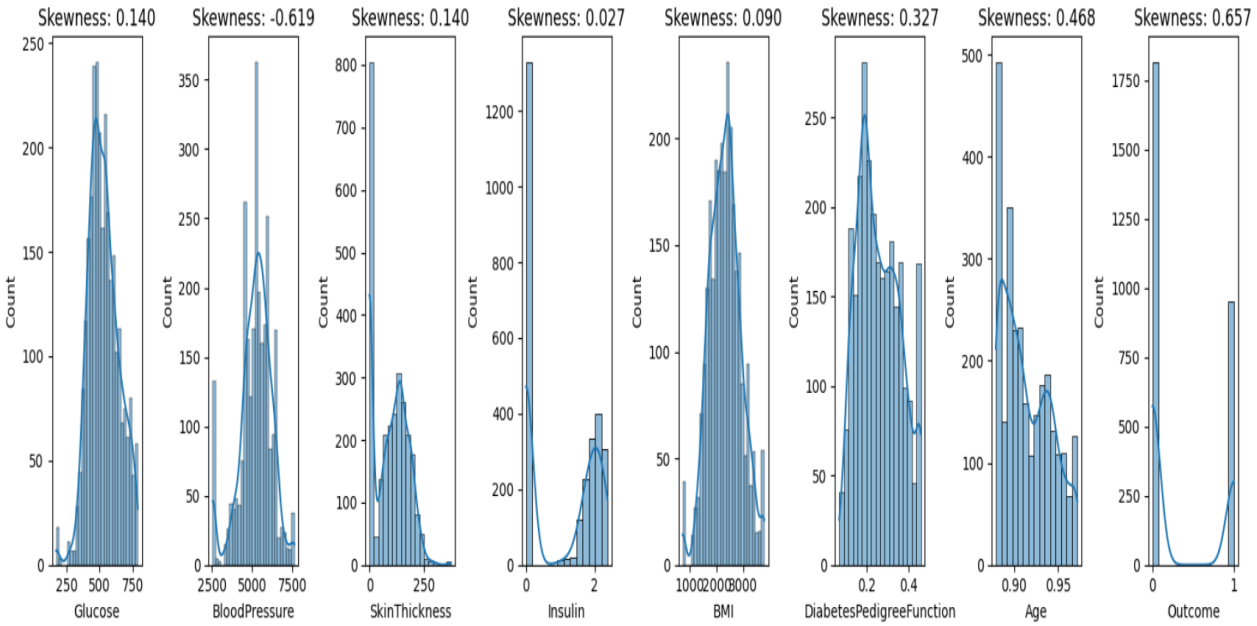

Distribution of Glucose
Count
Glucose
Distribution of BloodPressure
Count
BloodPressure
Distribution of BMI
Count
BMI
Distribution of Insulin
Count
Insulin


5.3e+02
6
1
2.9e+02
Truh
0
1
0
1
predicted
500
400
300
200
100

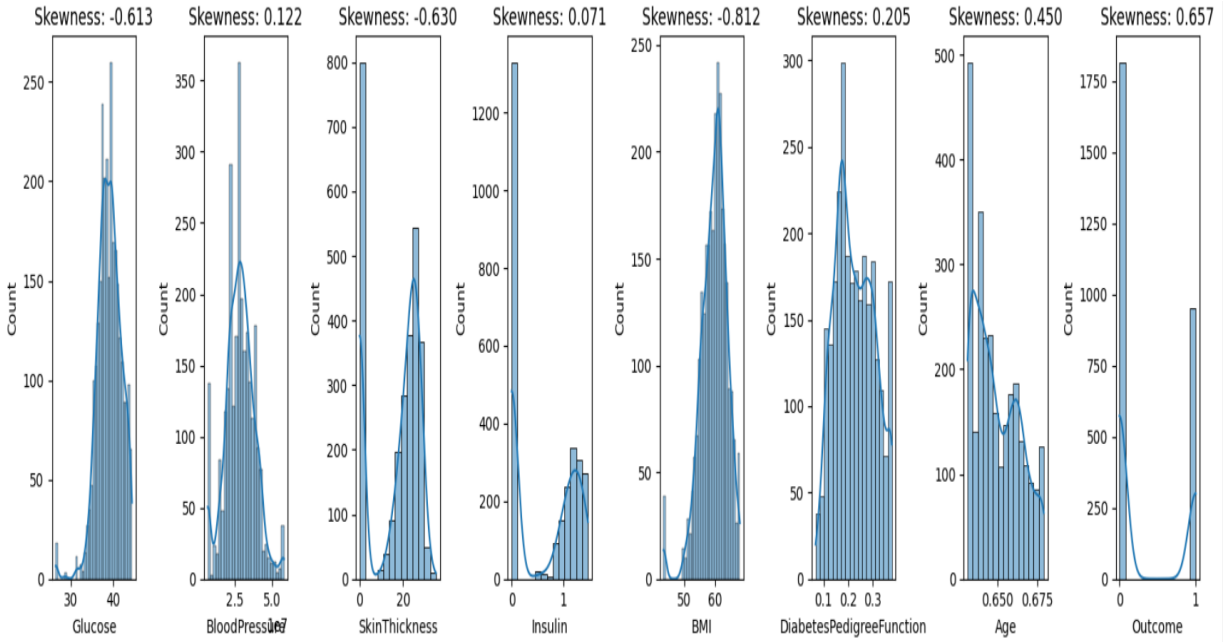

Before Handling Outliers
12000
10000
8000
6000
4000
2000
0
Id
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
Diabetes
BMI
PedigreeFunction
Age
Outcome


After Handling Outliers
12000
10000
8000
6000
4000
2000
0
Id
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
Diabetes
BMI
PedigreeFunction
Age
Outcome

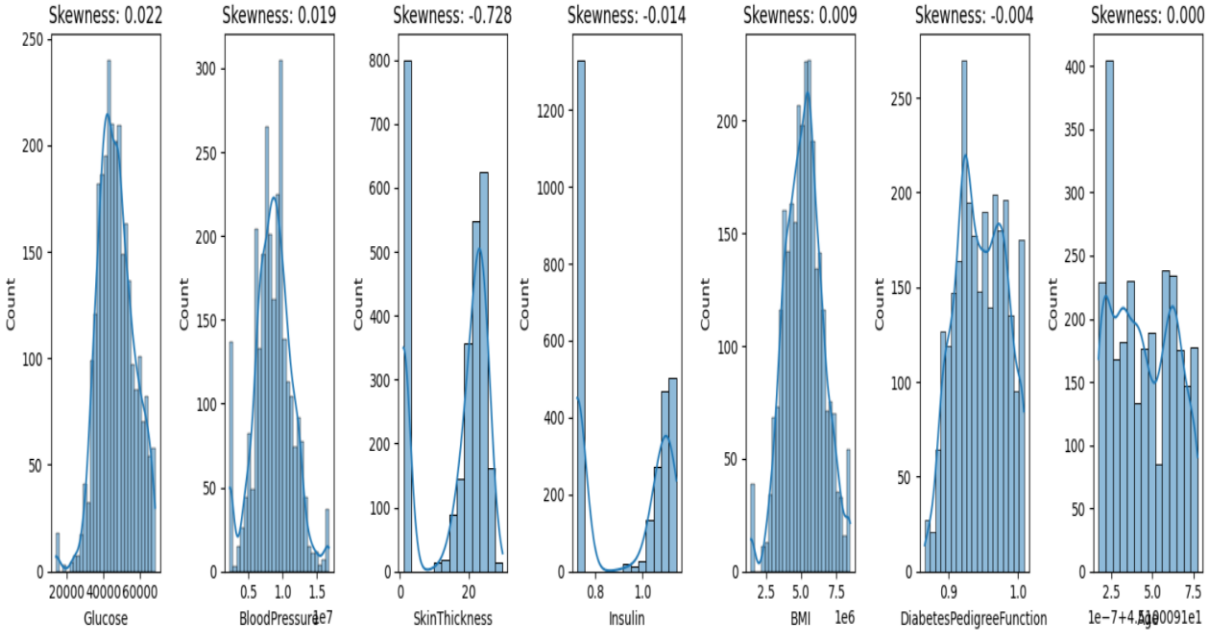
BEFORE HANDLING SKEWNESS:



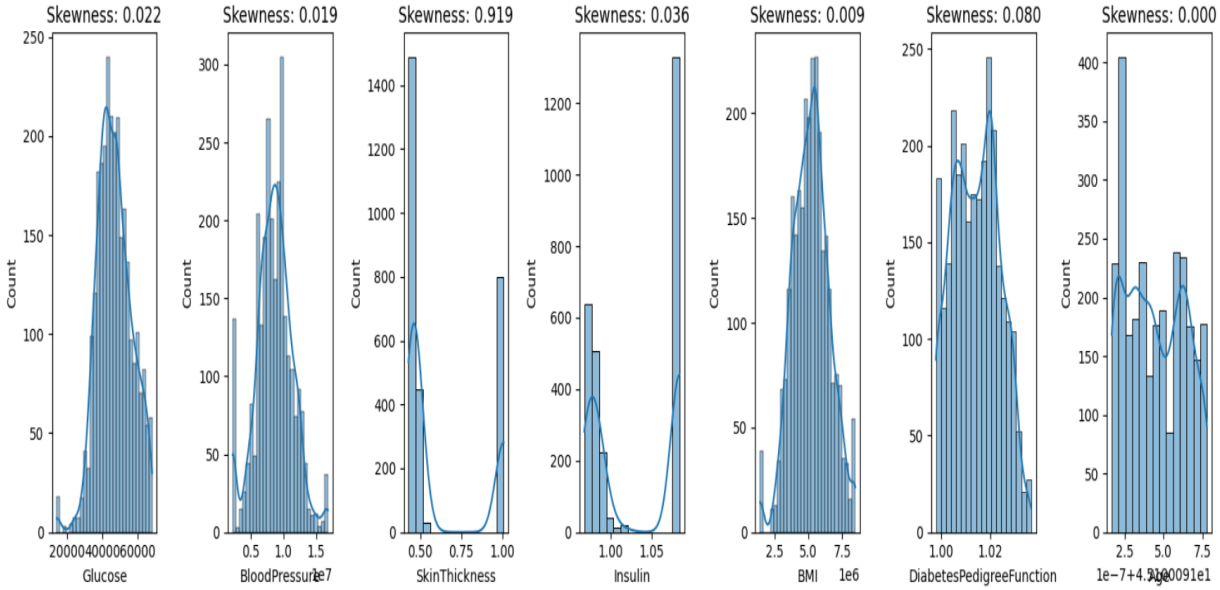
AFTER TRANSFORMATION & POWER TRANSFORMATION:



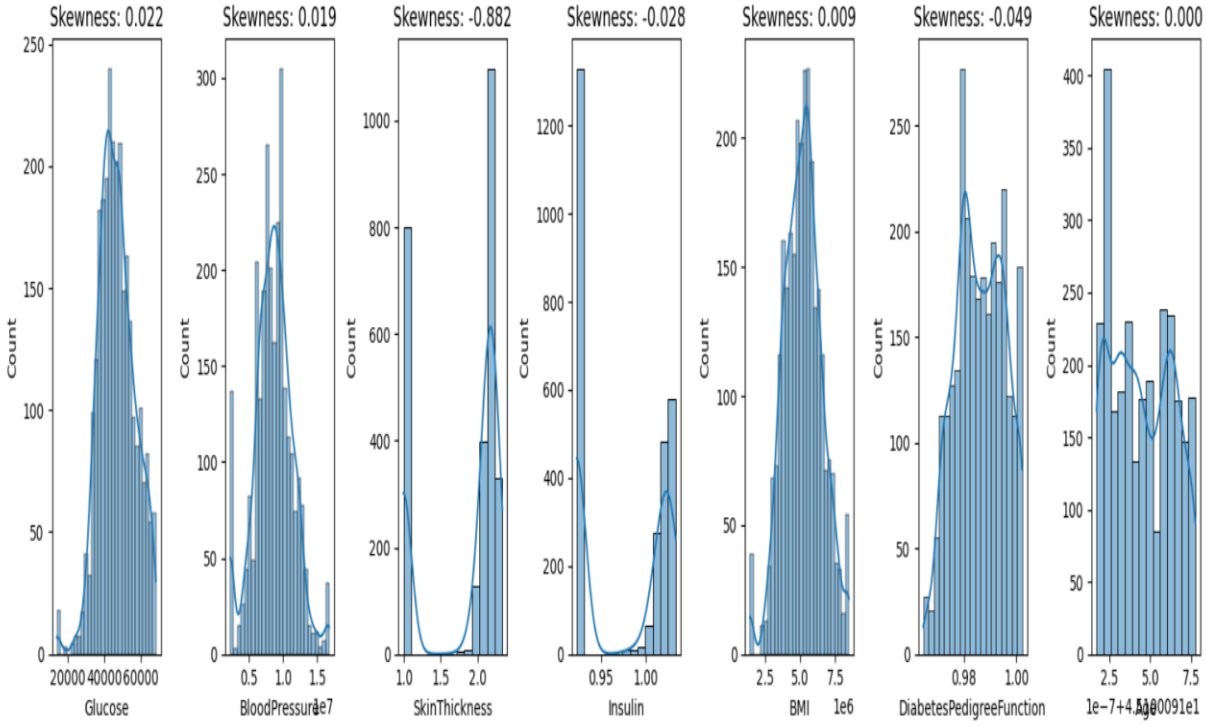
AFTER YEO-JOHNSON TRANSFORMATION



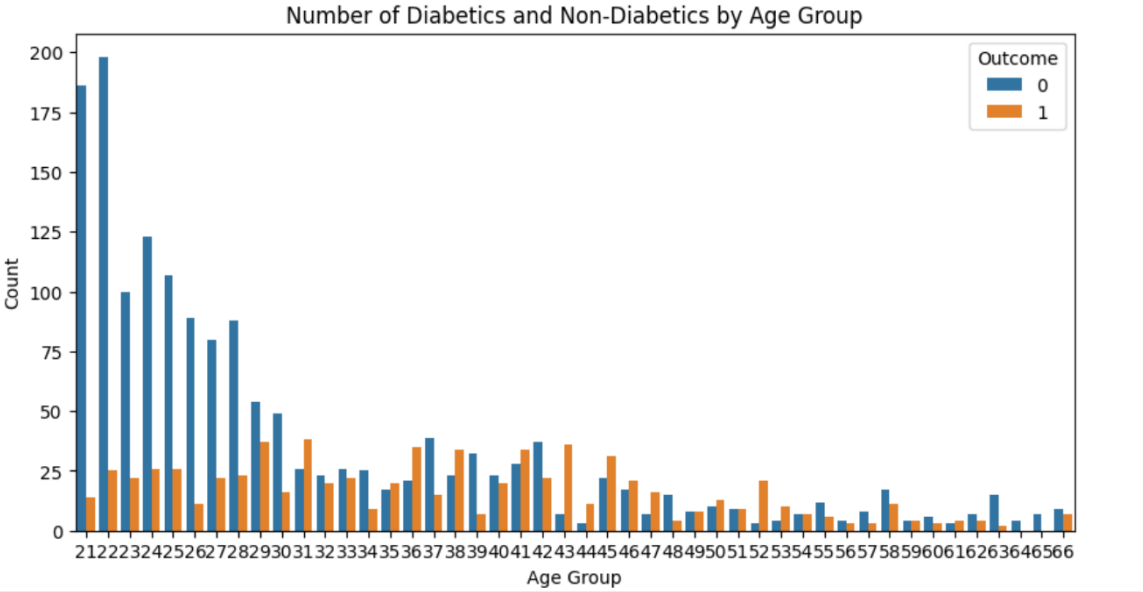
AFTER RECIPROCAL TRANSFORMATION:

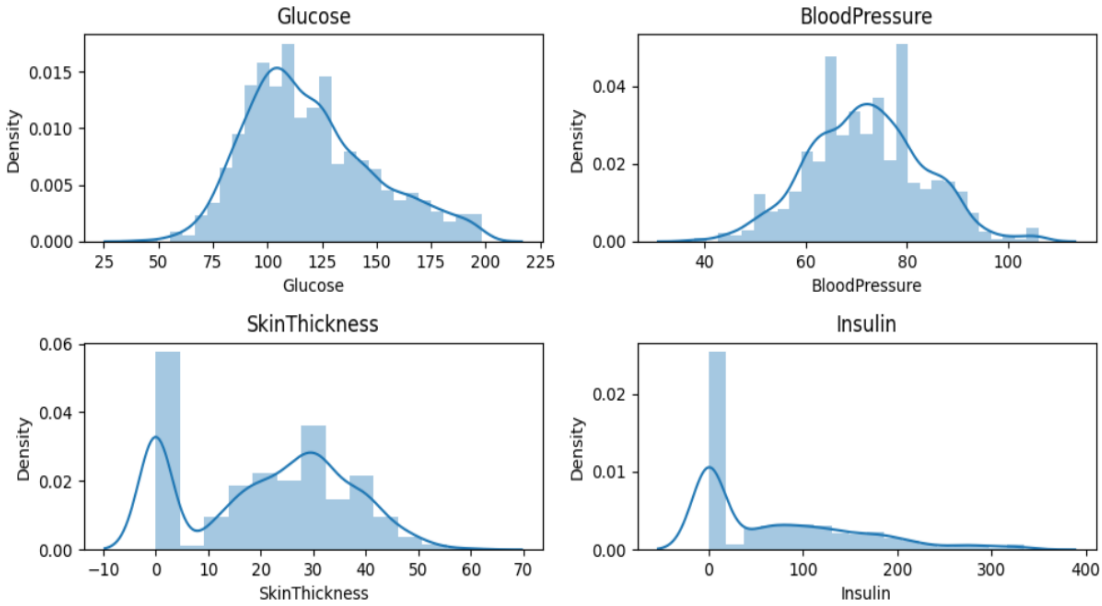


AFTER ROOT TRANSFORMATION:

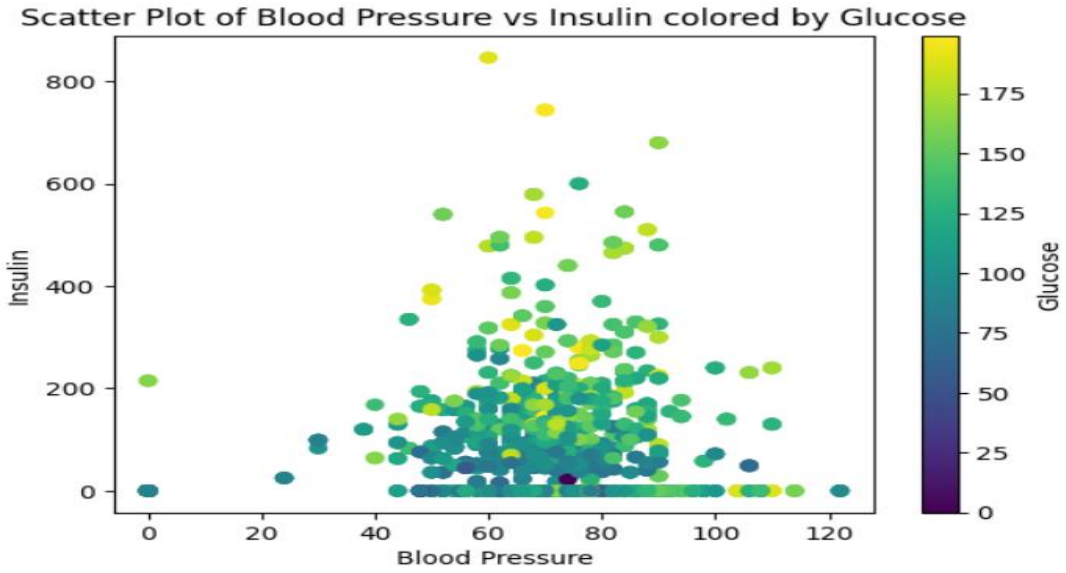


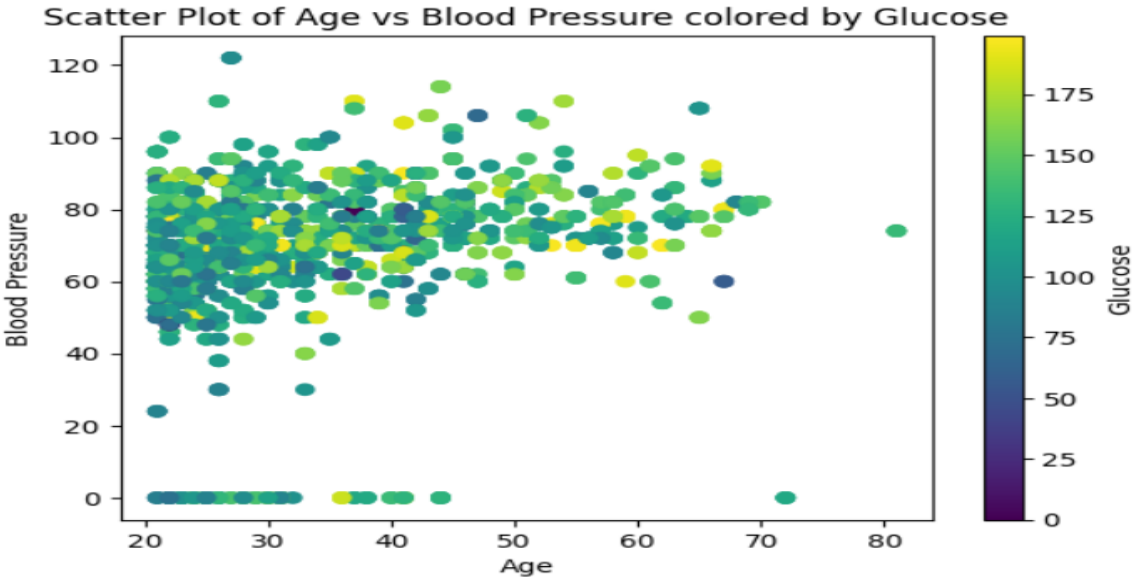
VISUALIZATION:

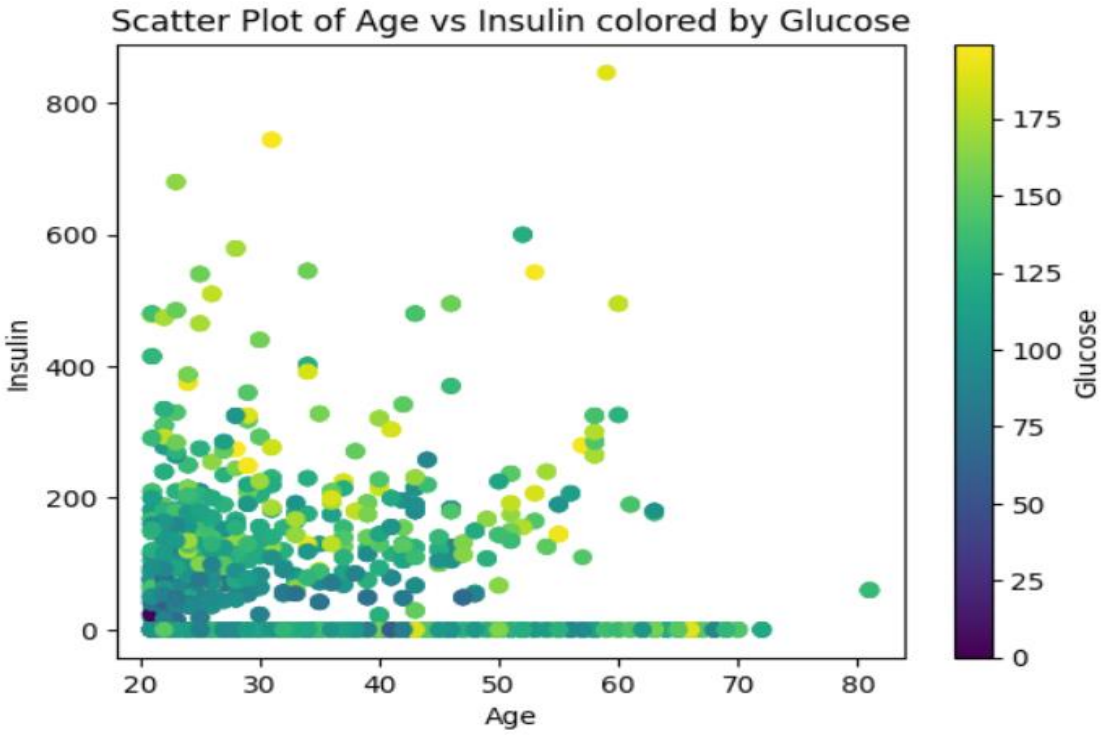


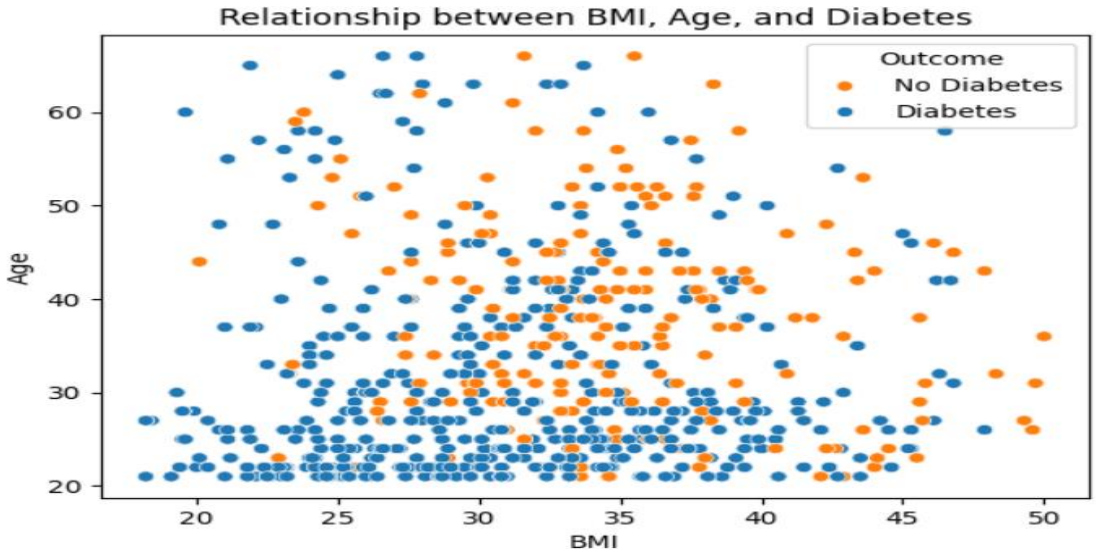


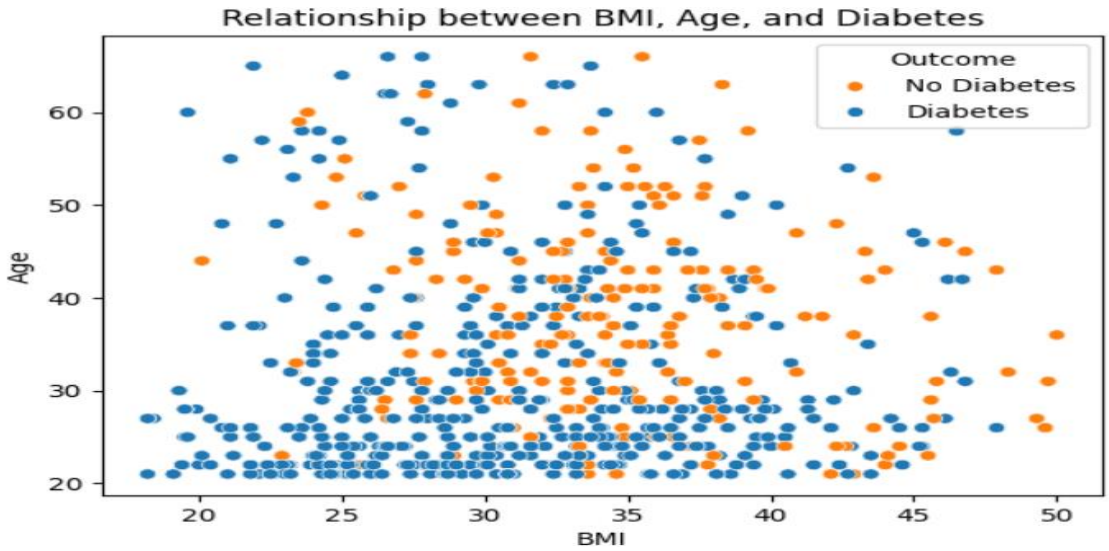
VISUALIZATION USING SCATTER PLOT:




Scatter Plot of Age vs Blood Pressure colored by Glucose
Blood Pressure
Age
Glucose
0
20
40
60
80
100
120
20
30
40
50
60
70
80
0
25
50
75
100
125
150
175


Scatter Plot of Age vs Insulin colored by Glucose
Insulin
Age
Glucose
0
200
400
600
800
20
30
40
50
60
70
80
0
25
50
75
100
125
150
175





Github link:
https://github.com/cheersbuddy/Dataset_Analysis/blob/main/sowmiya.ipynb

Google Colab link:

https://colab.research.google.com/drive/1qobiJZ_HZN-S2xStWDrU1Sb8GHsUsz2W#scrollTo=RdMfhBsD4vAj

Conclusion:

Thus the outcome and the relationship between all the parameters are visualized and all the tasks are completed.

Task-2

1. Understand the IMU Dataset and make it to a standardise format like CSV

2. Do some Cleaning and Preprocessing task in IMU Dataset

3. Give inference like Correlation and Data Distribution of Each column and your own Taste of Analytics

4. Implement the dimensionality reduction techniques for IMU Dataset

Title:

IMU Dataset

Dataset Link: https://drive.google.com/drive/folders/1h78miqhBxy_cO1TWBg-3I2fSkm--_18e?usp=sharing

Dataset Description:

WithKinect:

- **IMU folder:** Contains text files with accelerometer readings (rows 1, 2, and 3), gyroscope readings (rows 4, 5, and 6), and magnetometer readings (rows 7, 8, and 9), all delimited by commas.
- **Kinect folder:** Contains CSV files documenting Kinect marker readings for shoulder, elbow, and angle during forward, backward, and side exercises.

WithVicon:

- **IMU folder:** Similar to the WithKinect IMU folder, it contains text files with accelerometer, gyroscope, and magnetometer readings.
- **VICON folder:** Contains a text file detailing marker readings for shoulder (rows 1, 2, and 3), elbow (rows 4, 5, and 6), and wrist (rows 7, 8, and 9).

Objective:

- Integrate Sensor Data (IMU): Combine and analyze accelerometer, gyroscope, and magnetometer readings from both Kinect and Vicon folders to understand movement patterns during various exercises.
- Compare Marker Data (VICON and Kinect): Analyze marker readings from VICON and Kinect folders to compare how different motion tracking technologies capture and interpret movements such as shoulder, elbow, and angle positions during exercises.

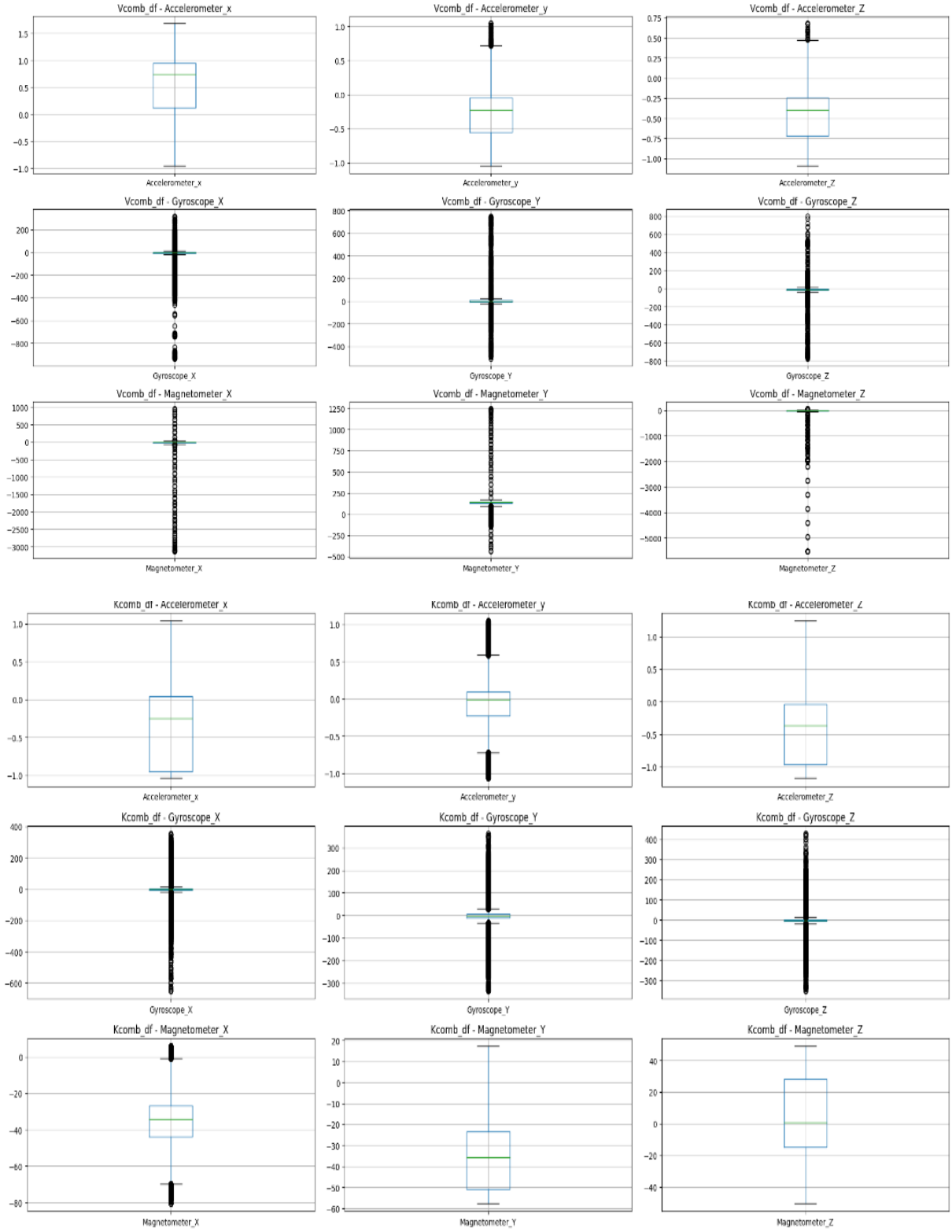
Implementation:

- Loading the Dataset: The dataset was loaded into the colab using pandas.
- Exploratory Data Analysis (EDA): Initial data exploration to understand the distribution and relationships between features.
- Feature Engineering: Selection and transformation of features to enhance model performance.
- Visualization: Creating visualizations to illustrate key metrics and patterns in the data using Matplotlib and seaborn.

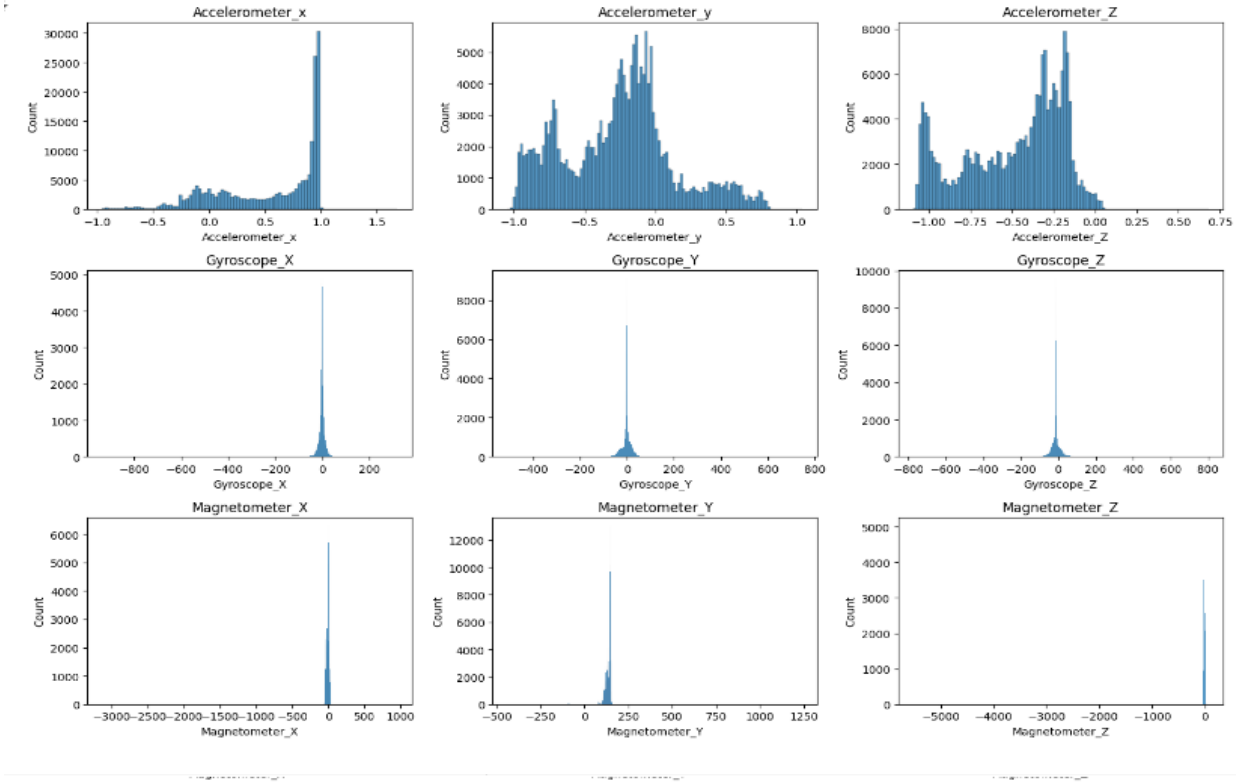
Process:

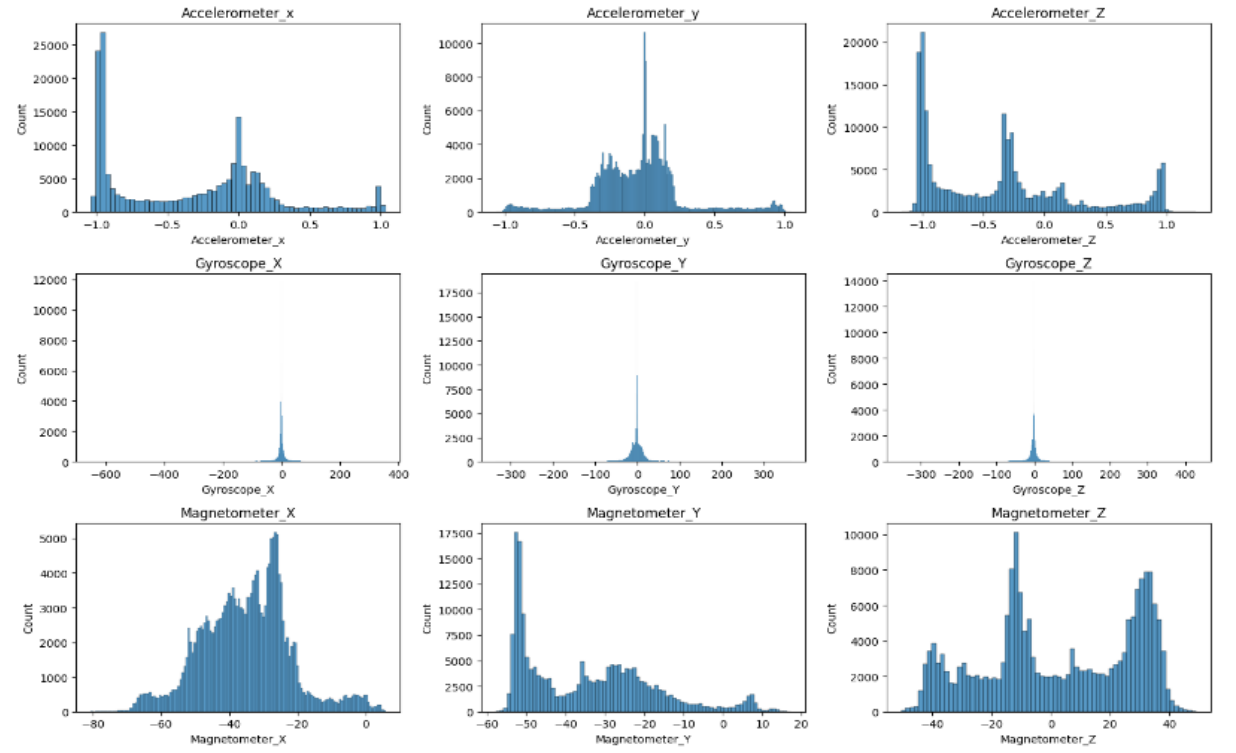
- Load CSV Files of Vicon device IMU readings
- Load CSV Files of Kinect device IMU readings
- Combining Dataframes of Vicon
- Combining Dataframes Of Kinect
- Droping the 10th coloumn because there is no information about that, and adding headings to the dataframe
- EDA
- Checking for null values
- Checking for Duplicates
- Removing Duplicates
- Checking for outliers

Handling Outlier

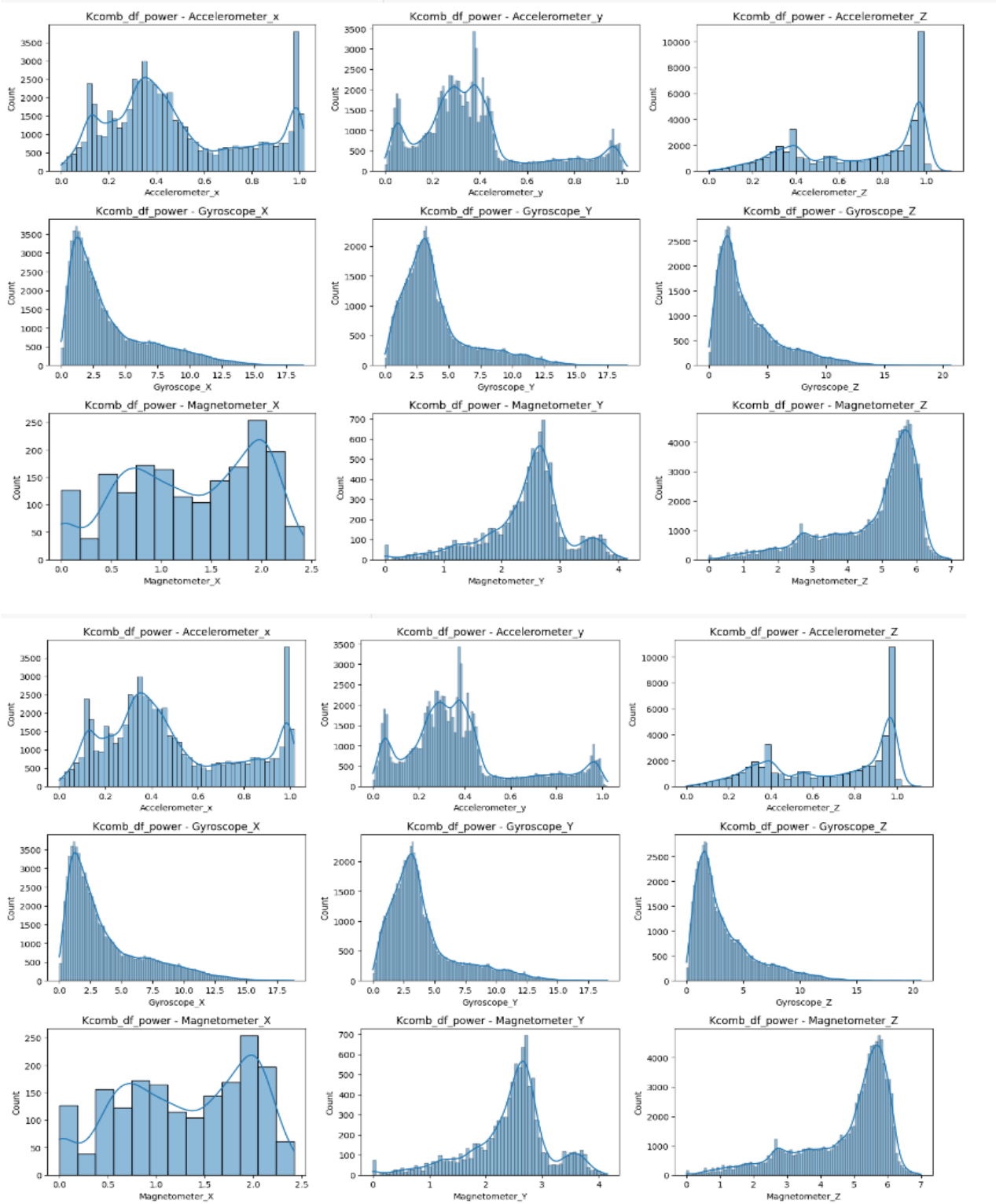


Visualization before making winsorizing, power and Quantile Transformation for Vcomb_df and Kcomb_df:

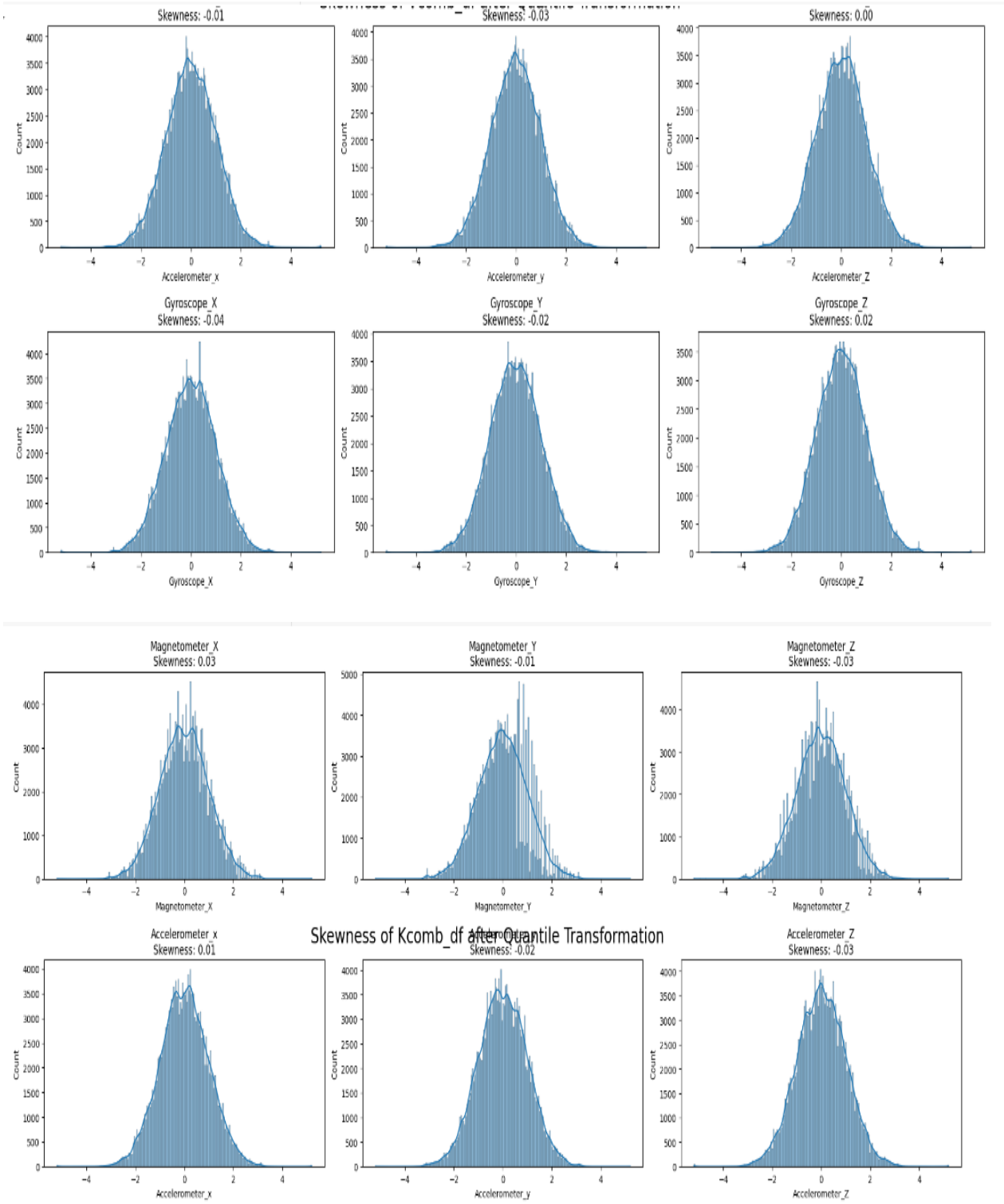


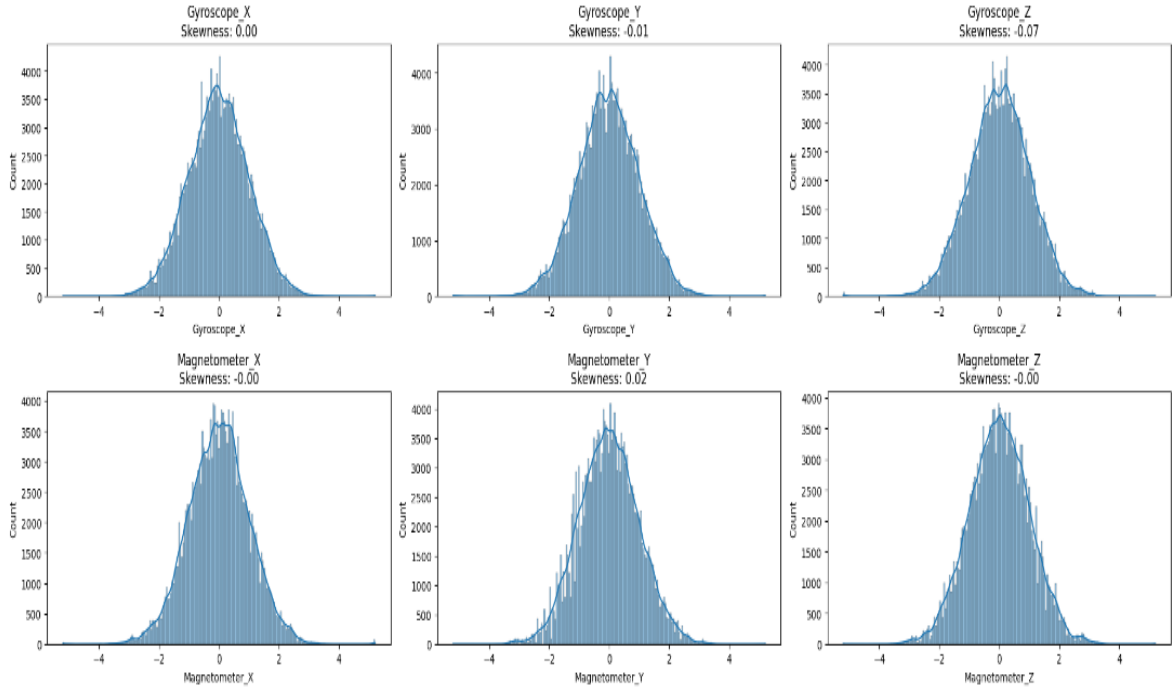


Visualization after making power and Quantile Transformation for Vcomb_df and Kcomb_df:

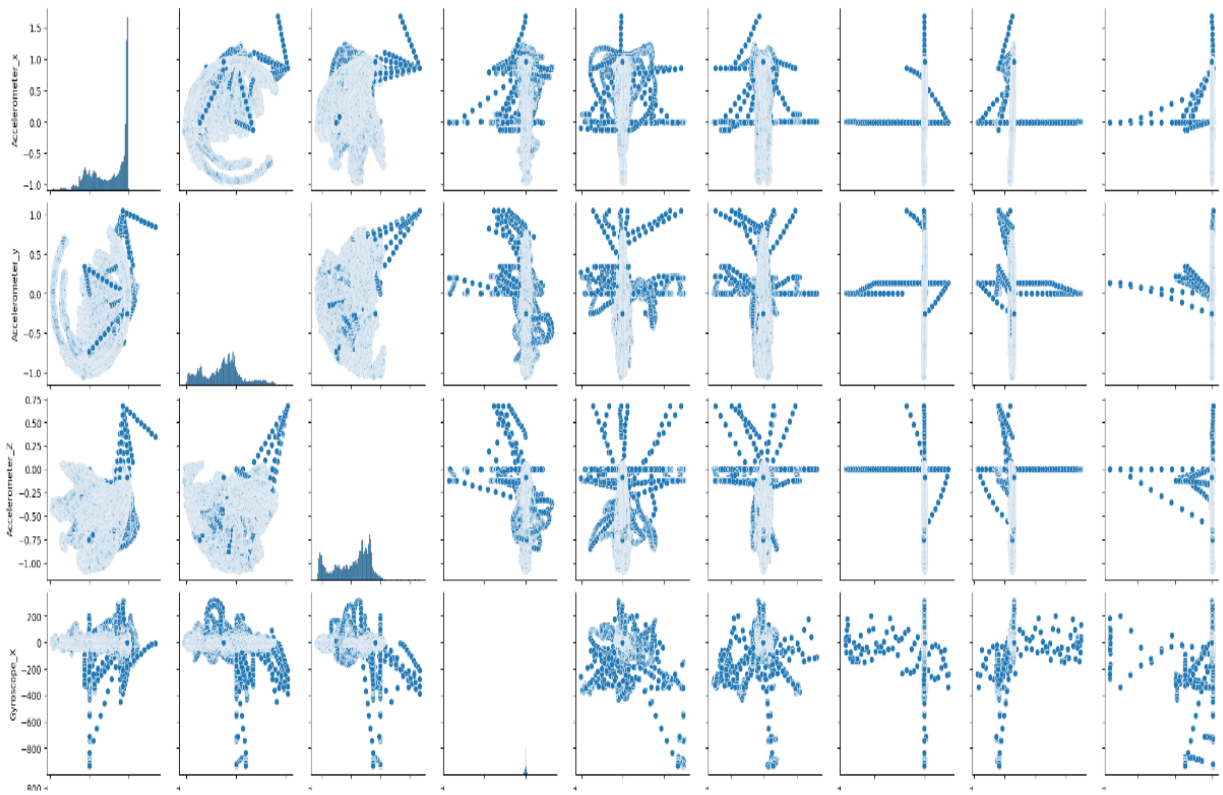


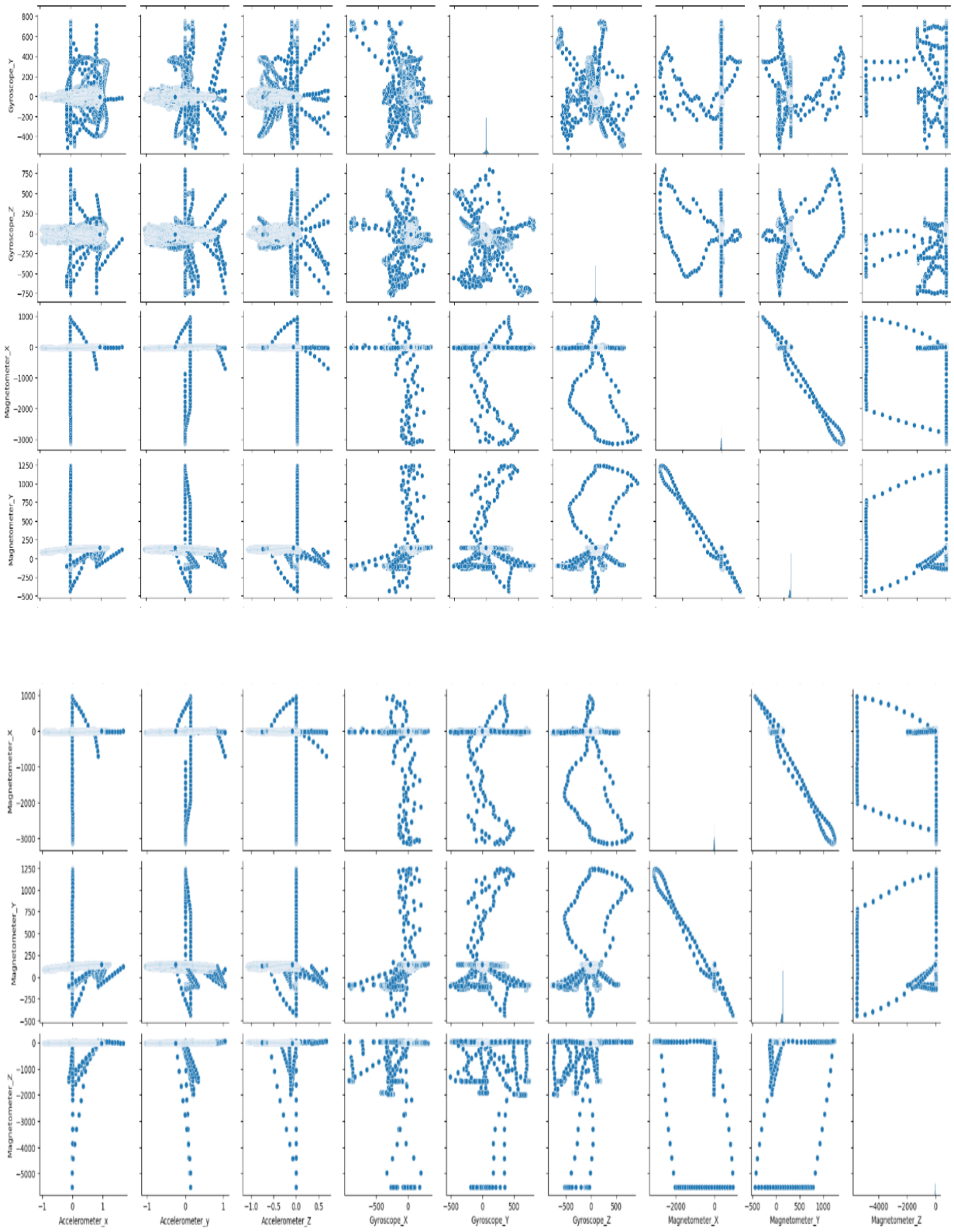
Final Skewness



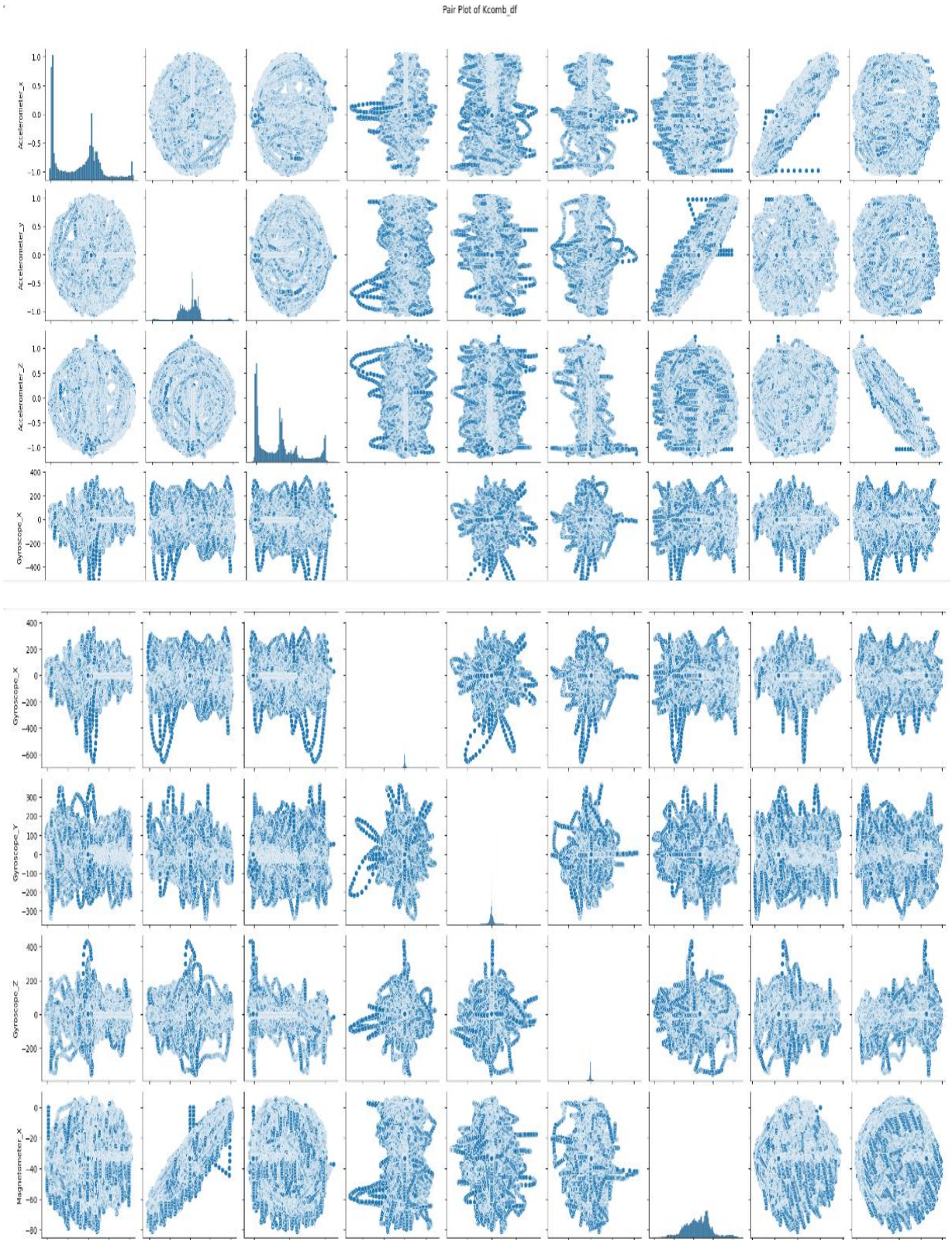

Gyroscope_X
Skewness: 0.00
Gyroscope_Y
Skewness: -0.01
Gyroscope_Z
Skewness: -0.07
Magnetometer_X
Skewness: -0.00
Magnetometer_Y
Skewness: 0.02
Magnetometer_Z
Skewness: -0.00
Count
Gyroscope_X
Gyroscope_Y
Gyroscope_Z
Magnetometer_X
Magnetometer_Y
Magnetometer_Z

PAIR PLOT:

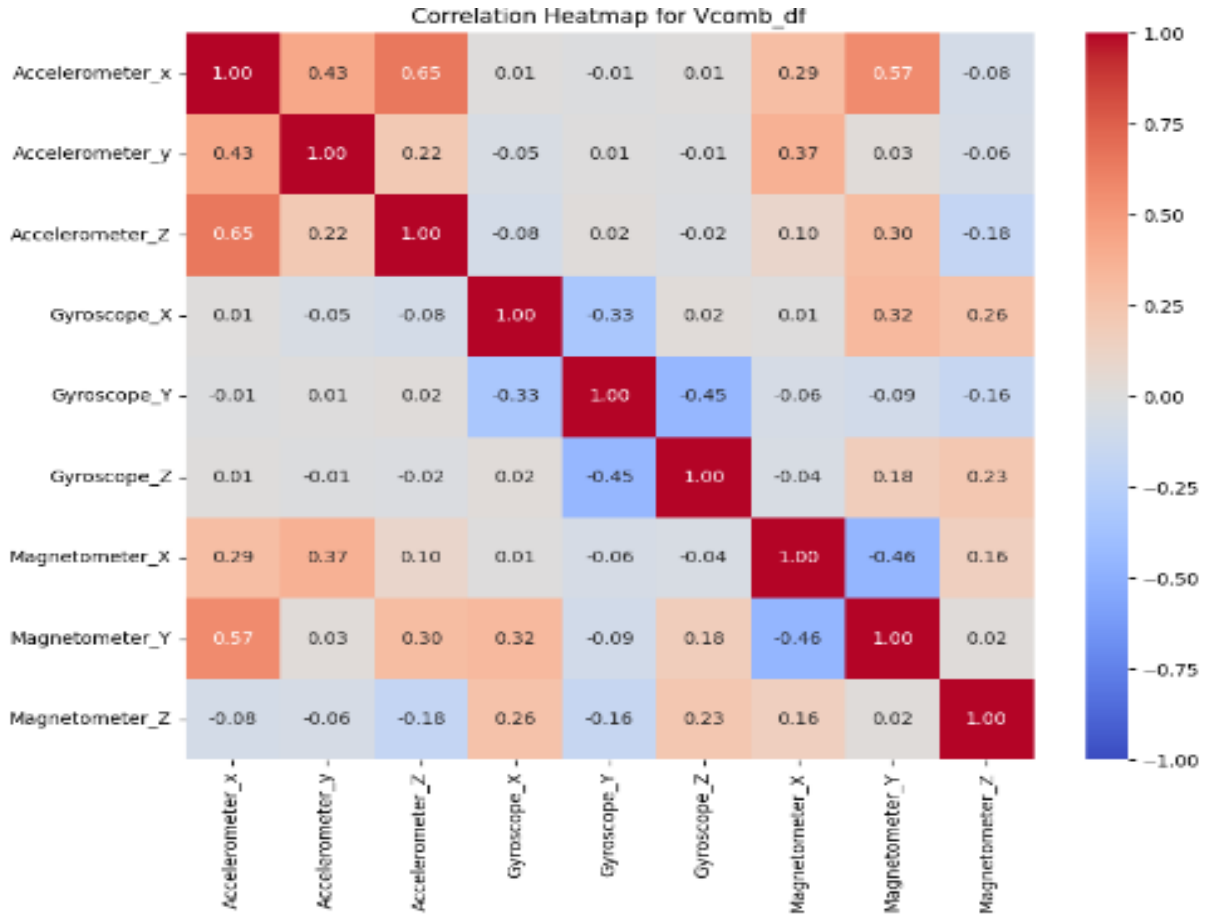

Accelerometer_X
Accelerometer_y
Accelerometer_Z
Gyroscope_X

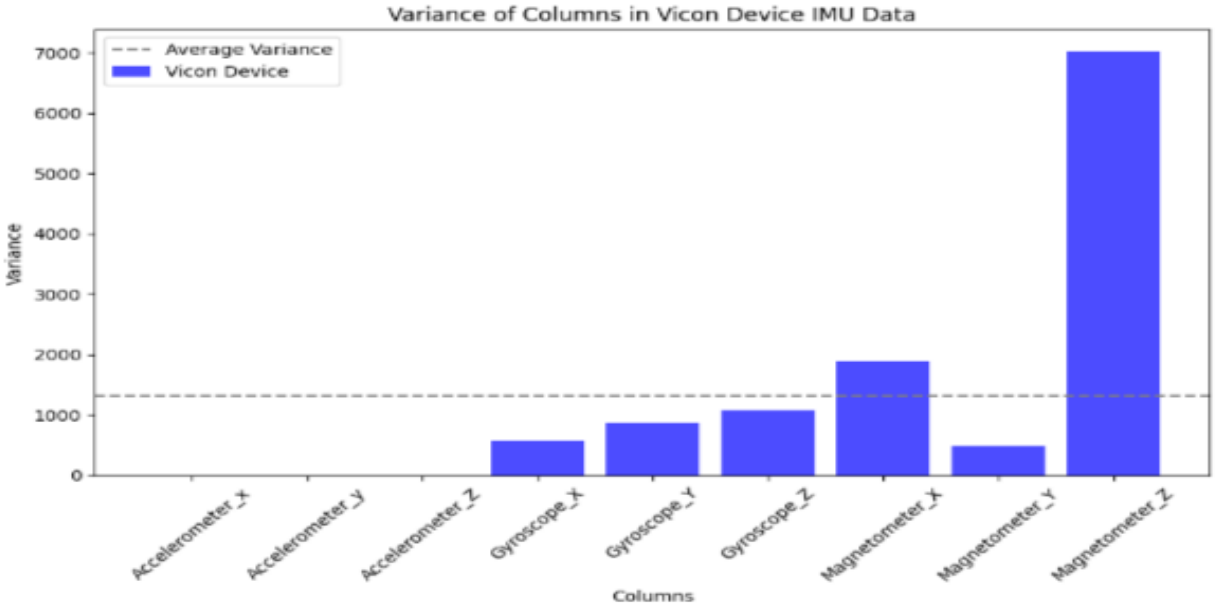

Gyroscope_Y
Gyroscope_Z
Magnetometer_X
Magnetometer_Y
Magnetometer_Z
Accelerometer_x
Accelerometer_y
Accelerometer_Z
Gyroscope_X
Gyroscope_Y
Gyroscope_Z
Magnetometer_X
Magnetometer_Y
Magnetometer_Z

PAIR PLOT:

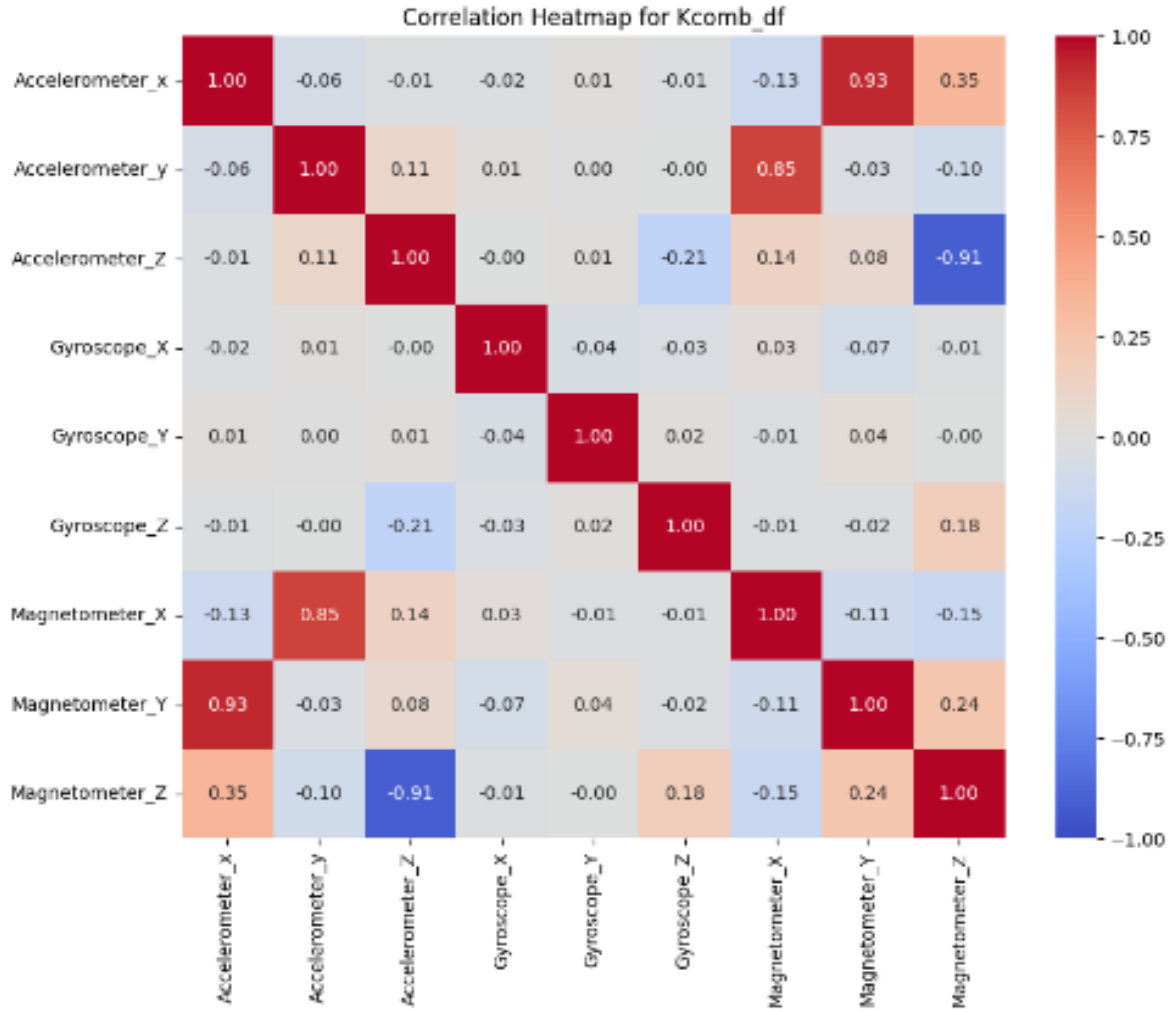

Pair Plot of Kcomb_df
Accelerometer_x
Accelerometer_y
Accelerometer_z
Gyroscope_x
Gyroscope_y
Gyroscope_z
Magnetometer_x

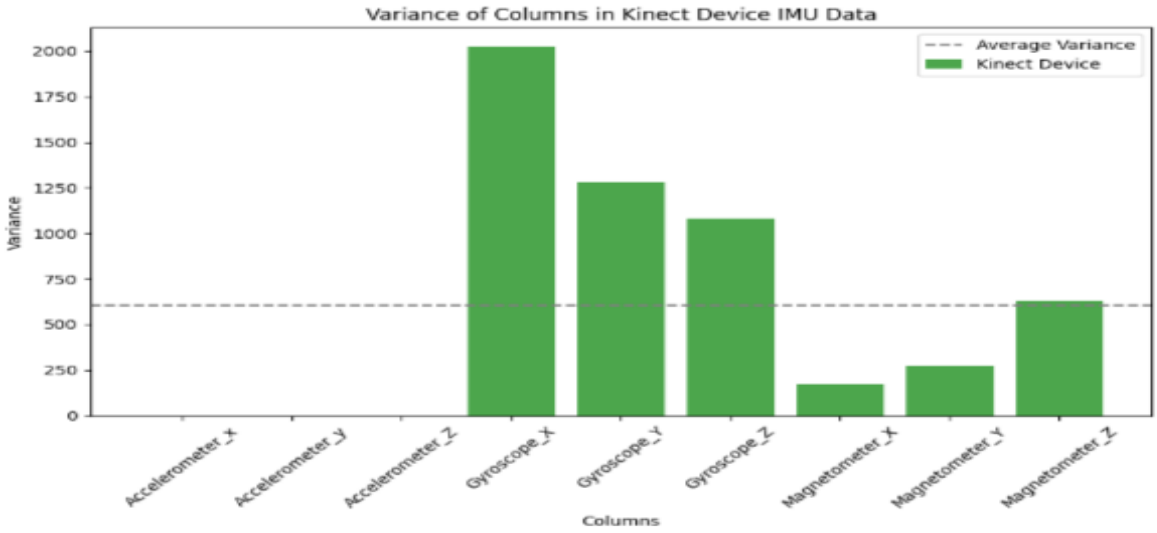
COORELATION HEATMAP FOR VCOM_DF



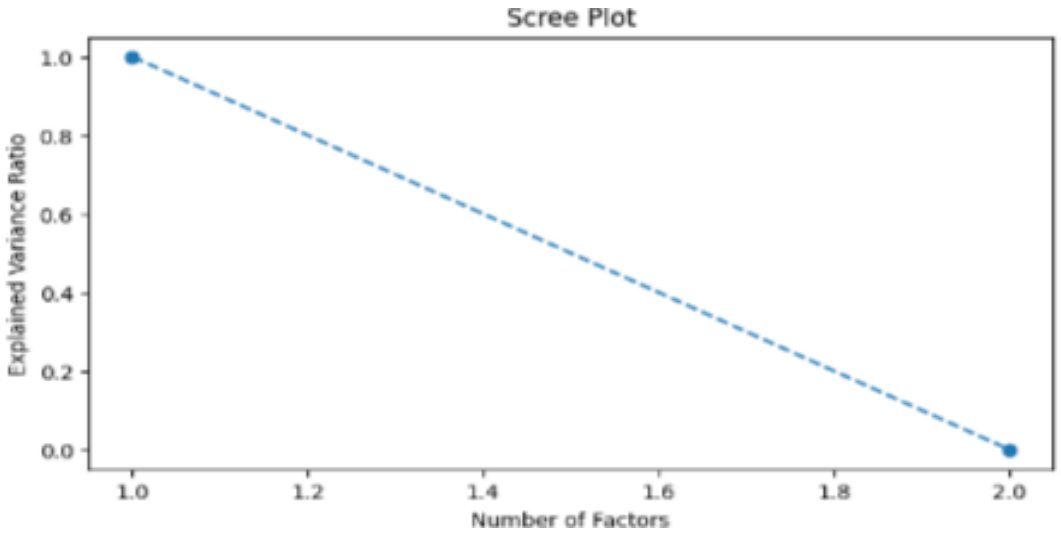


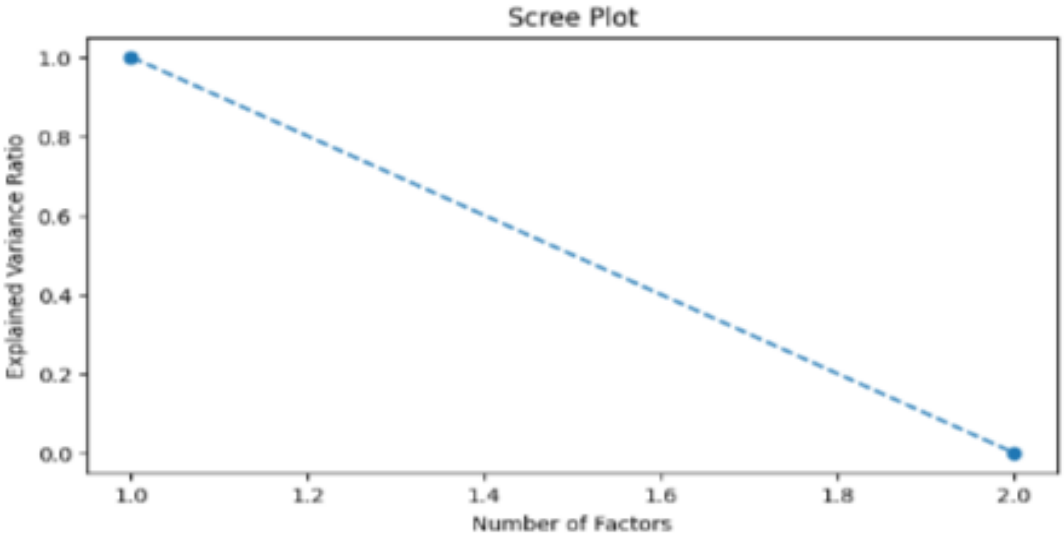
COORELATION HEATMAP FOR VCOM_DF





SCREE PLOT





Google colab link:
https://colab.research.google.com/drive/1wcDDk9SC3vVo5NCYV69Iiq7TX1-P4mIO#scrollTo=5exNLbfX_6AD

Github link:
https://github.com/cheersbuddy/Dataset_Analysis/blob/main/Copy_of_Team_A__IMU_Analysis.ipynb