Background

The fitness industry is one of the fastest growing industries today. It continues to grow exponentially every year, generating more than 30 billion dollars in the United States as of 2021. The fitness industry is one of the biggest industries in the world. With an immense growing of popularity and people are getting into wellness, is there a way we can attract more people to subscribe into fitness classes or gym? A data set obtained from Kaggle from a group of people ages 20-64. The data was collected by the National Sports Promotion at the Seoul Olympics.  Are we able to predict which age group and gender have a poor body performance? With the information, we can pitch this to the fitness industry to focus on that specific group to help attract them to get a better body performance for their age.

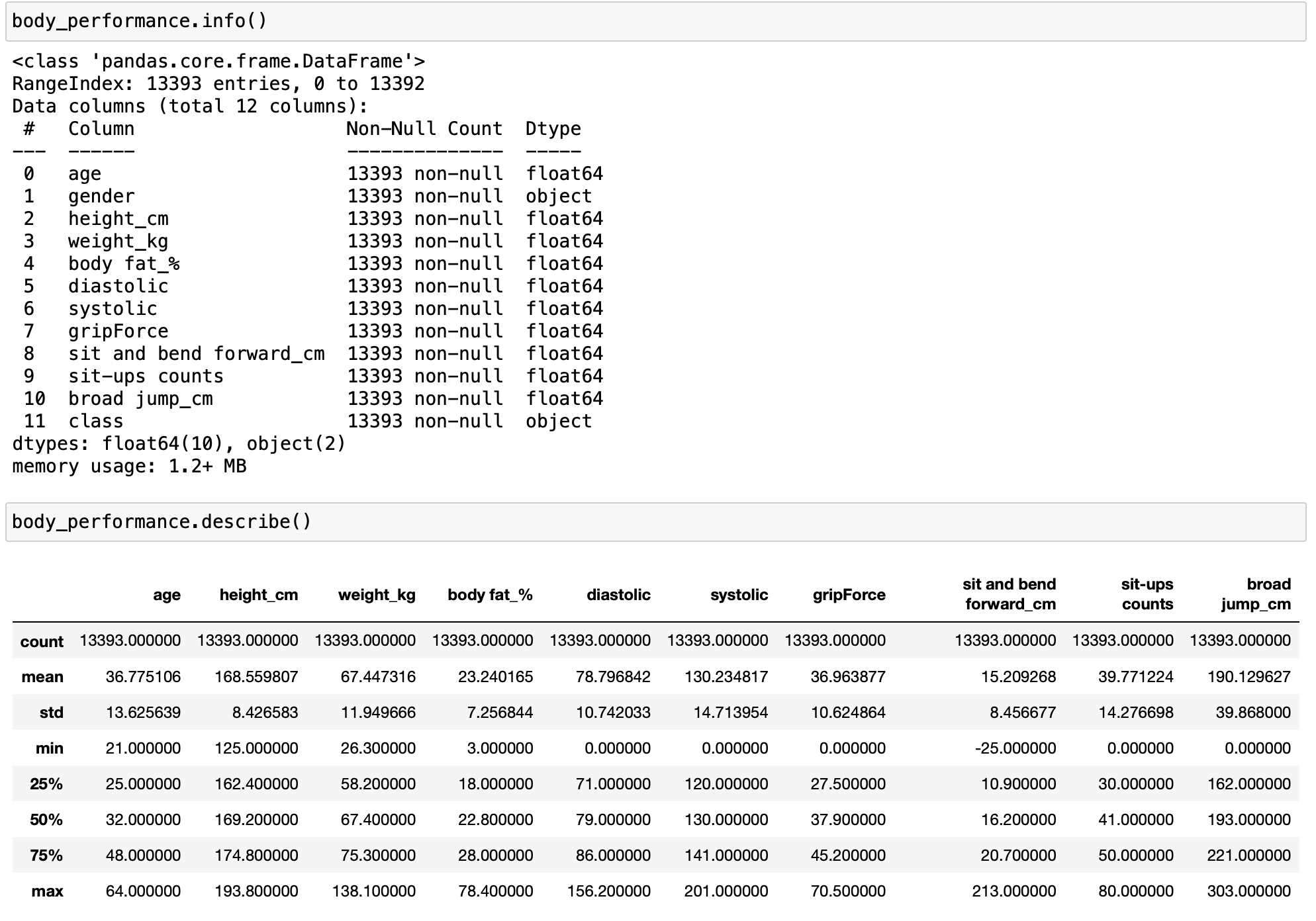
Outline

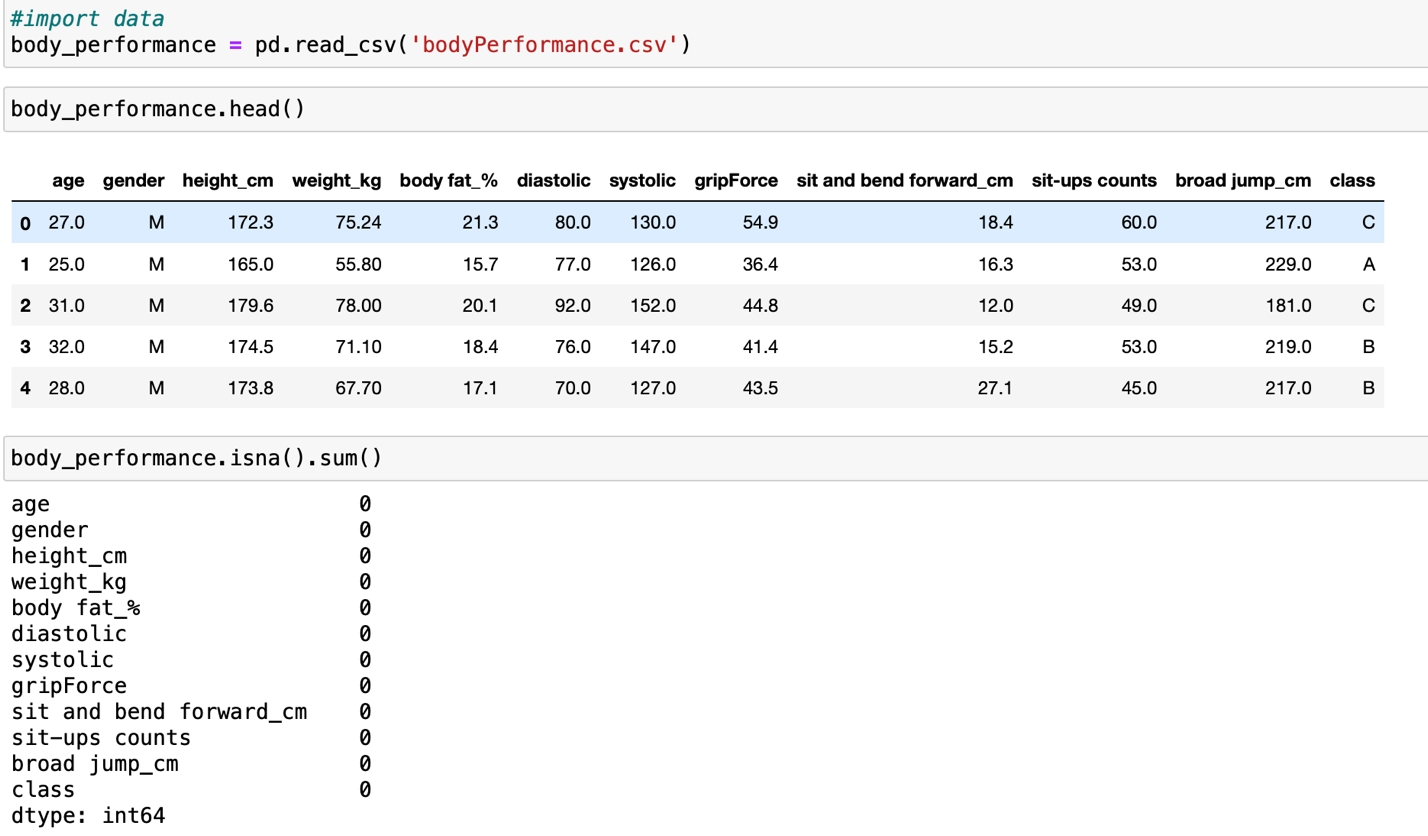
The first step was to obtain a set of data that contains people’s performances. I was able to obtain this data from Kaggle (<https://www.kaggle.com/kukuroo3/body-performance-data>). The data was collected by the National Sports Promotion at the Seoul Olympics. The data contained people of ages 20 to 64. The data also consisted of height, weight, body fat %, blood pressure (systolic and diastolic), workouts (grip force, sit and bend forward, sit-ups and broad jumps) and class (grade of performance set by the institution).

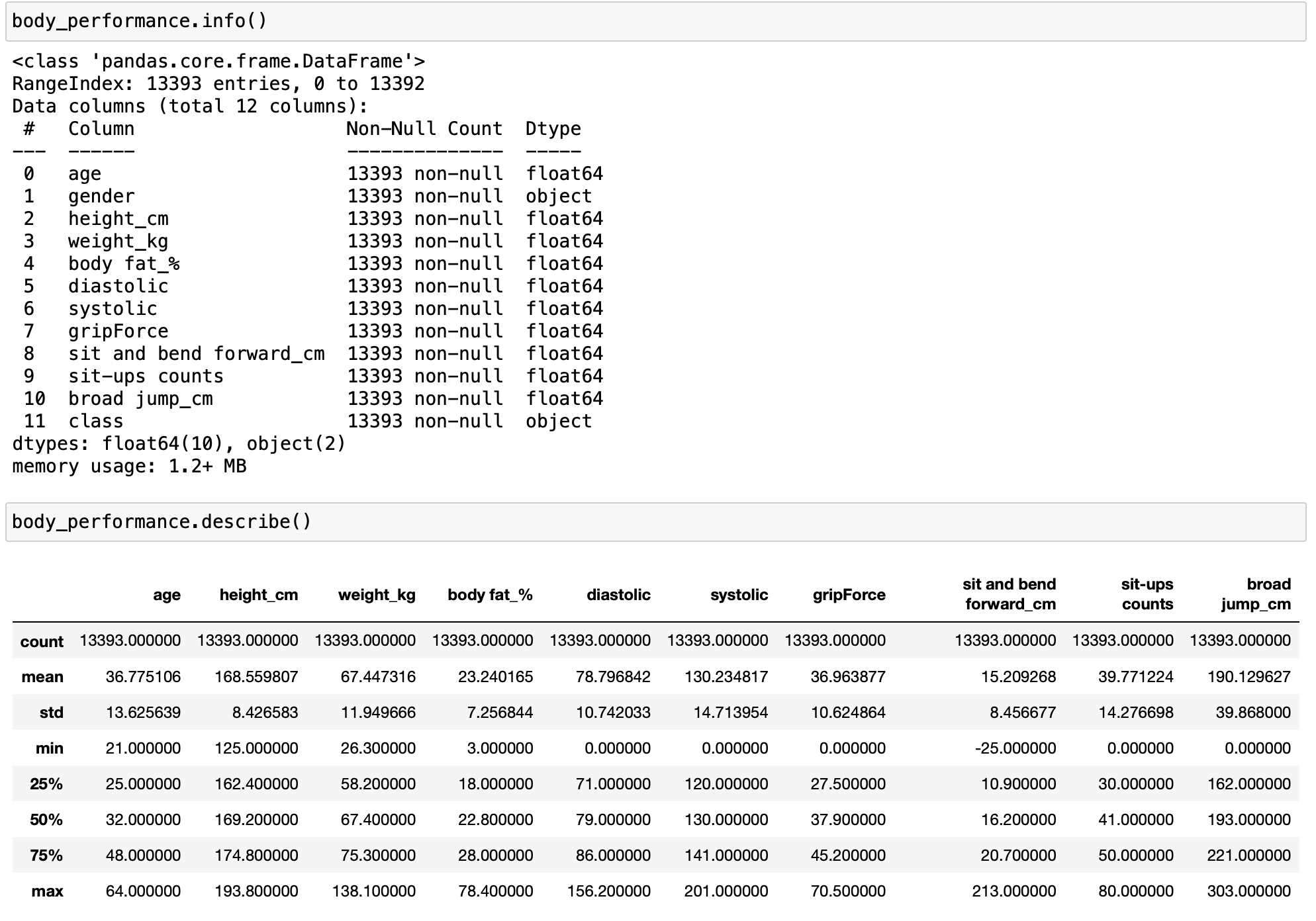
The data is pulled using python and I used jupyter notebook to explore the data and trained classification model to predict which class has a poor body performance. Using this information will help us create a business model to attract more people into fitness.

Data

Using simple Python API to pull the data and downloading the data from Kaggle in a csv format. There are no null value from the data after I use the isna() code. There wasn’t much cleaning to do with this data as everything looked pretty cleaned. They only thing I did was I grouped the ages in groups of 10; ages 20-30, ages 30-40 etc. This gave me a generalization of which age group have a better or worst performance as there are way too many different ages in this data. This will help me later when I explore the data using visualization. Here we see that the average age was around 36 with body fat % of 23, diastolic of 78 and systolic of 130. The normal blood pressure is defined as a systolic pressure of less than 120 and a diastolic pressure of less than 80. The acceptable body fat % for men are 18-25% and 25-31% for women. From this I can tell that we will probably be looking to target people of ages less than 30 or greater than 40.

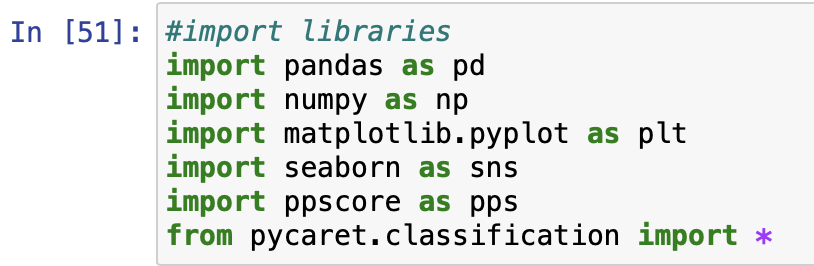




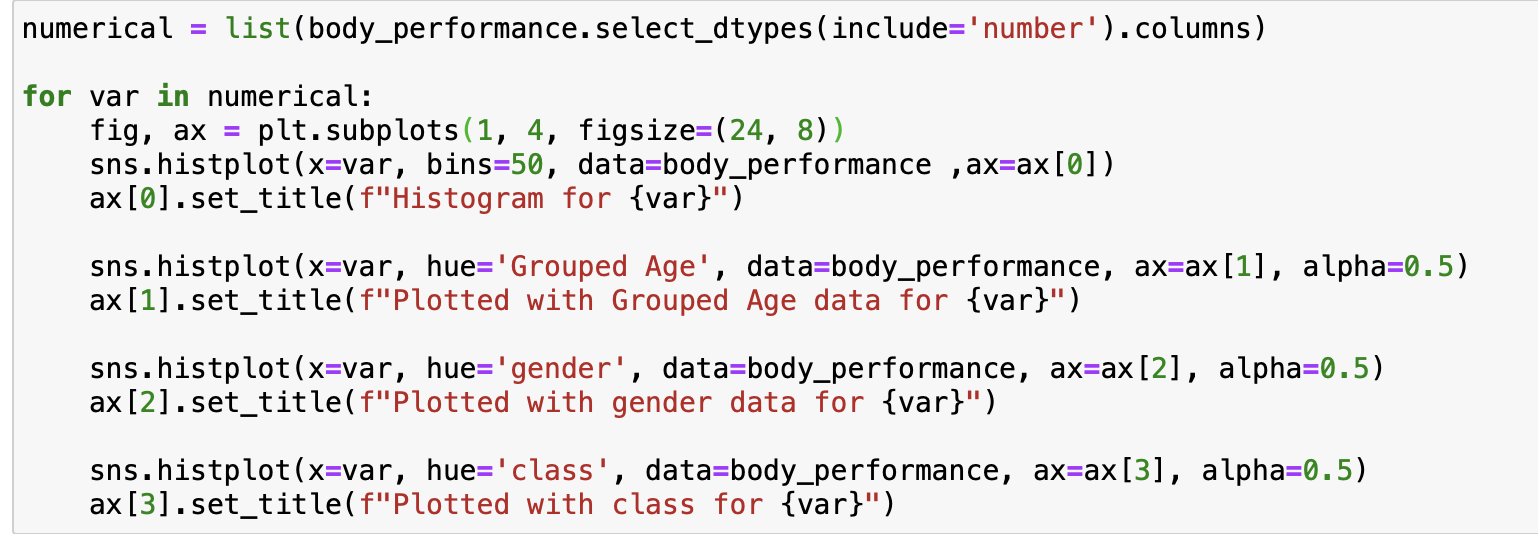


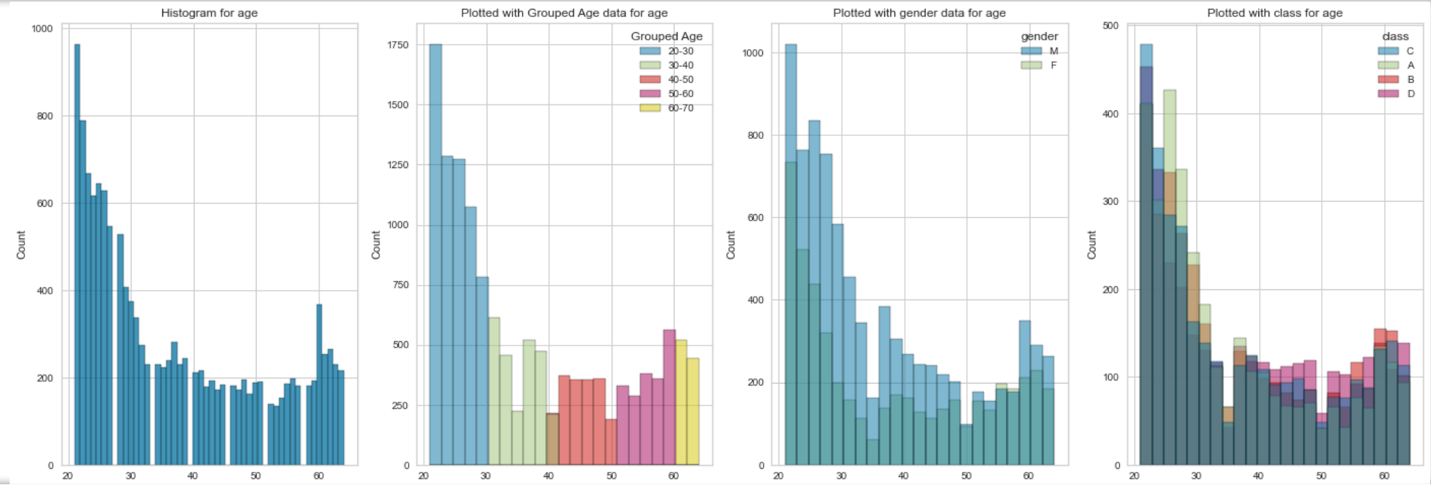
Data Exploration

First step we need to import the libraries in order for data wrangling and data exploring. This was done using python in jupyter notebook. Data wrangling was done beforehand and can be seen above or at Data section.

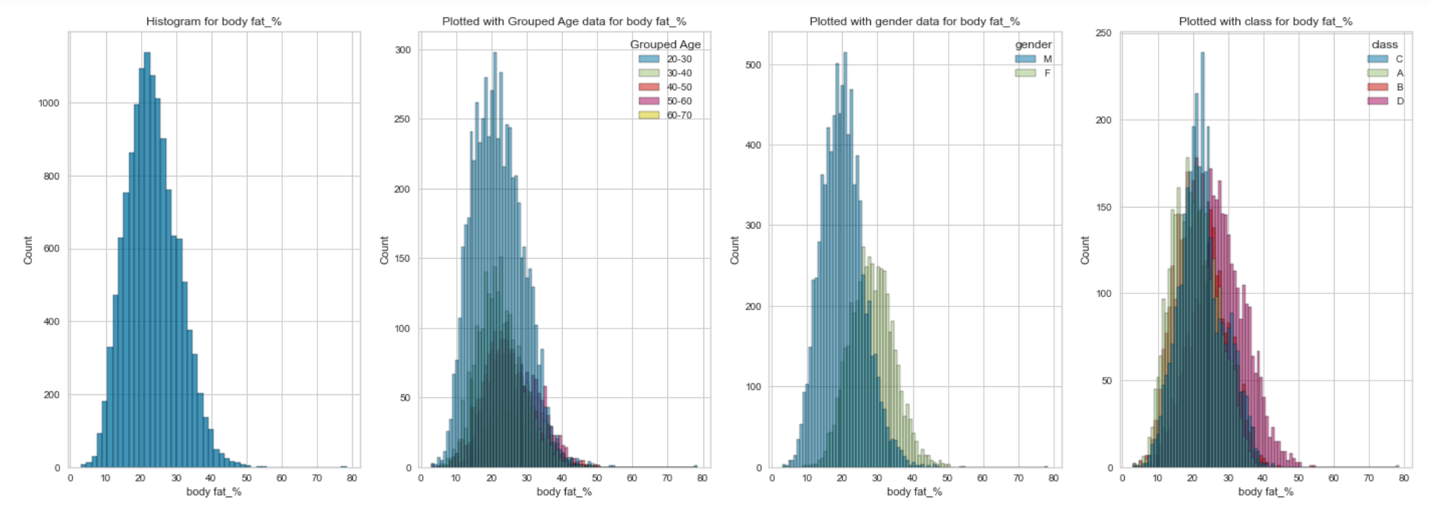
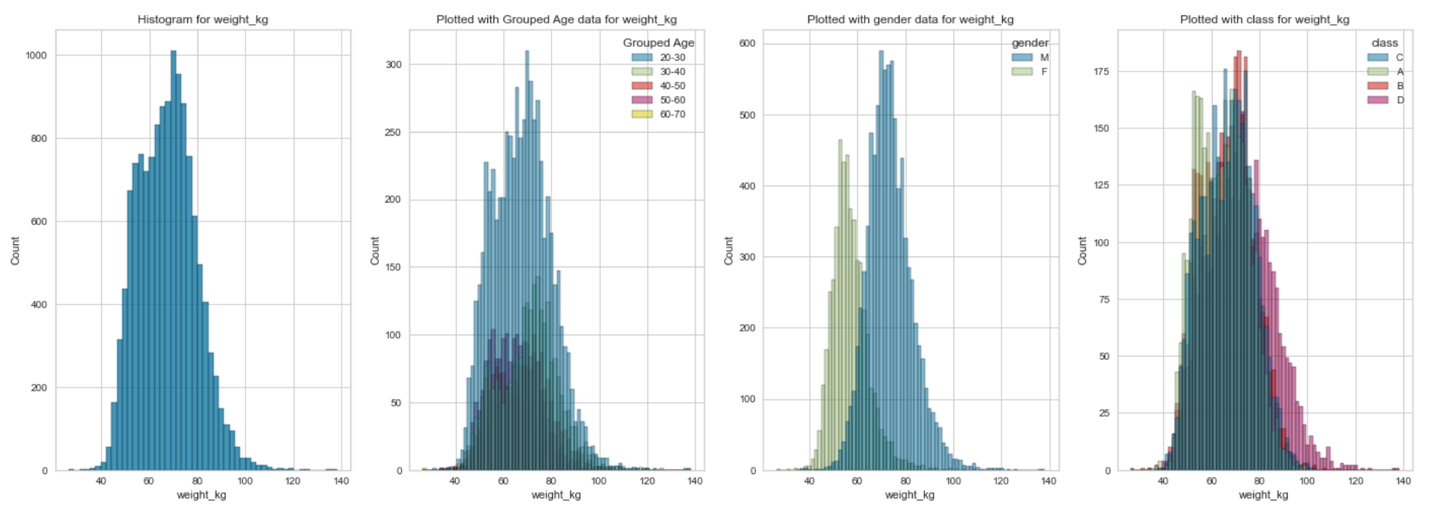
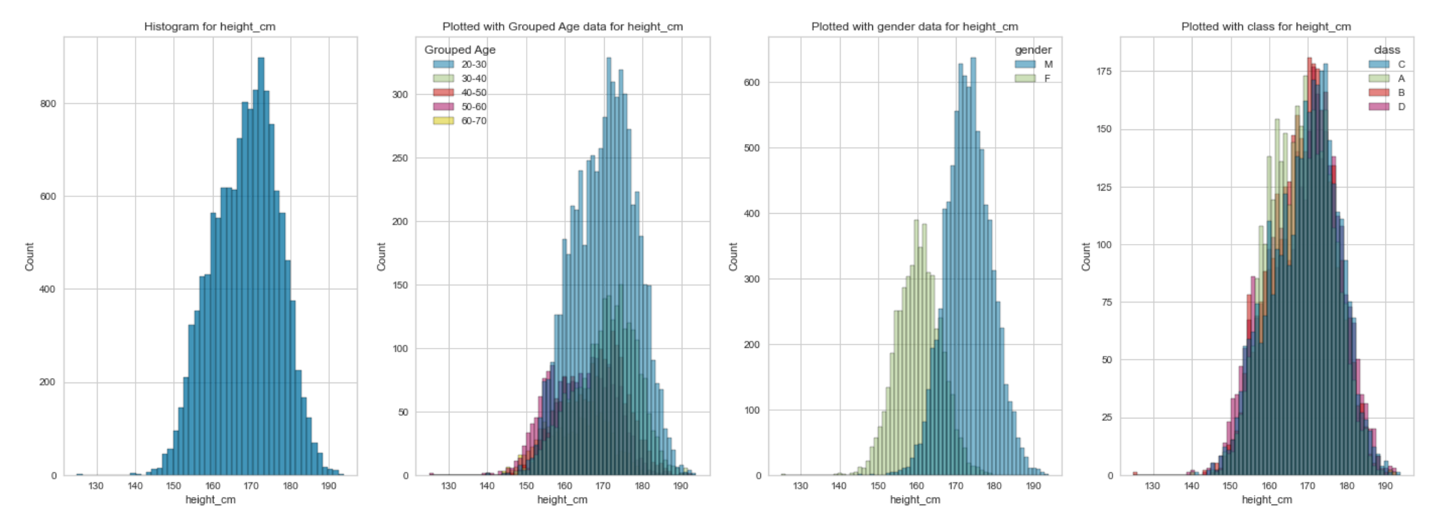


Next step we create a visualization for the data. In this case we created a histogram to visualized our data.

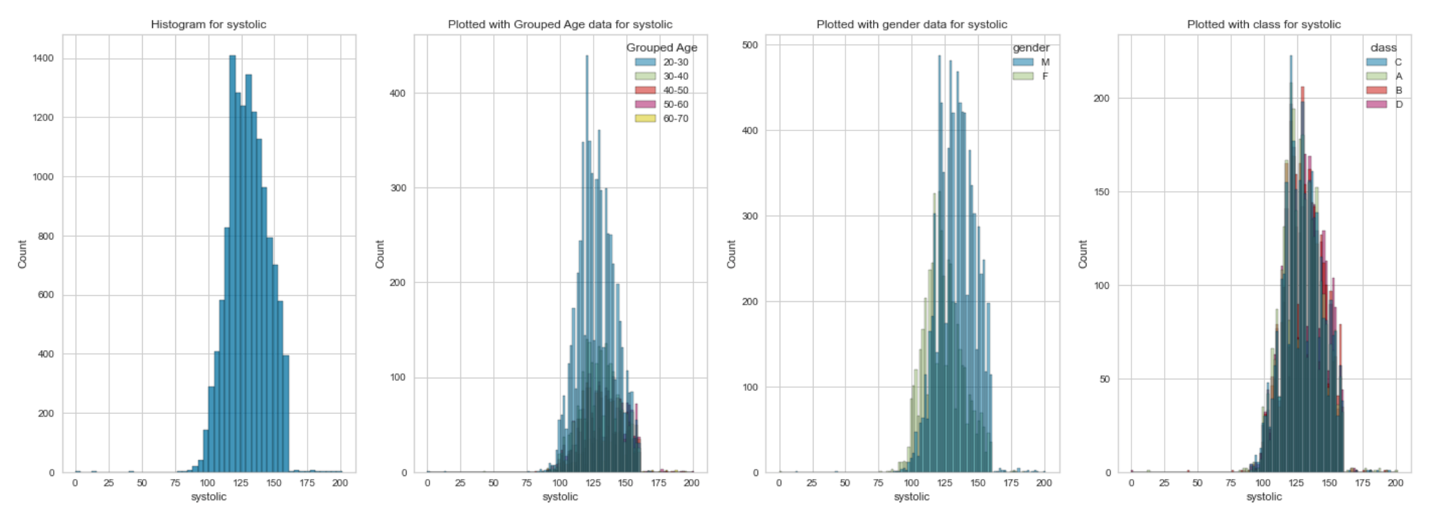
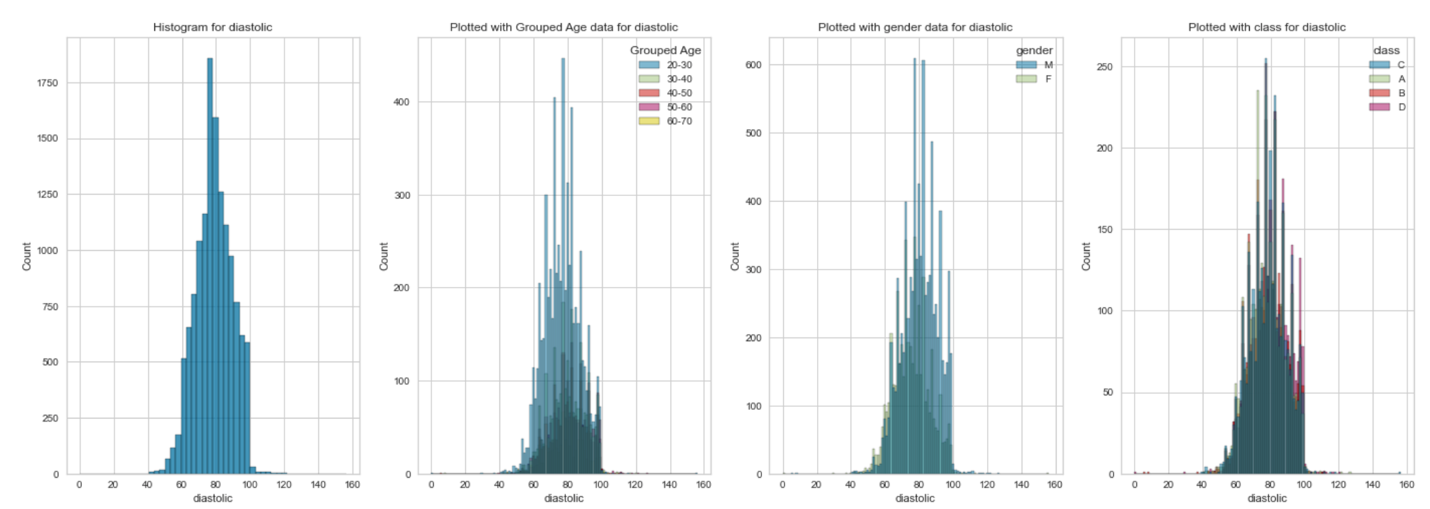




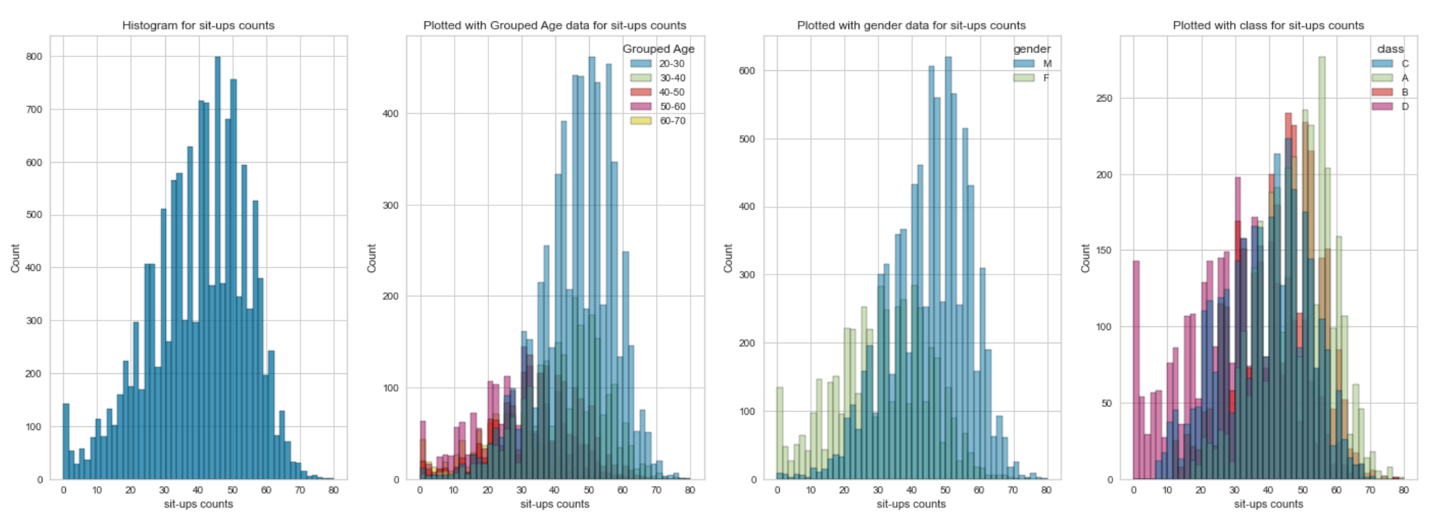
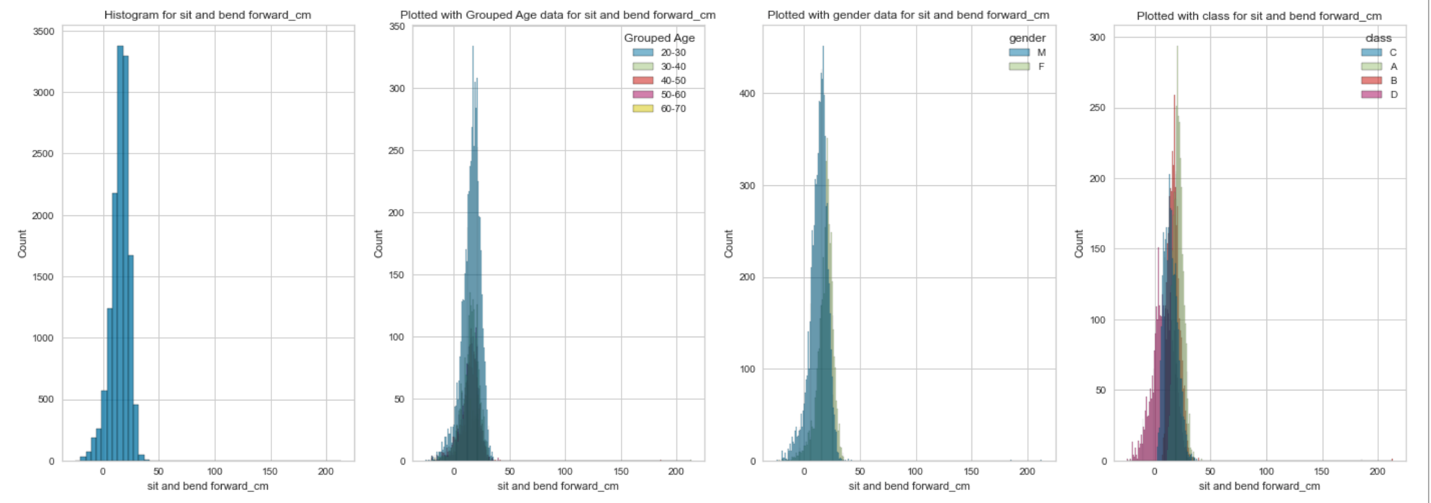
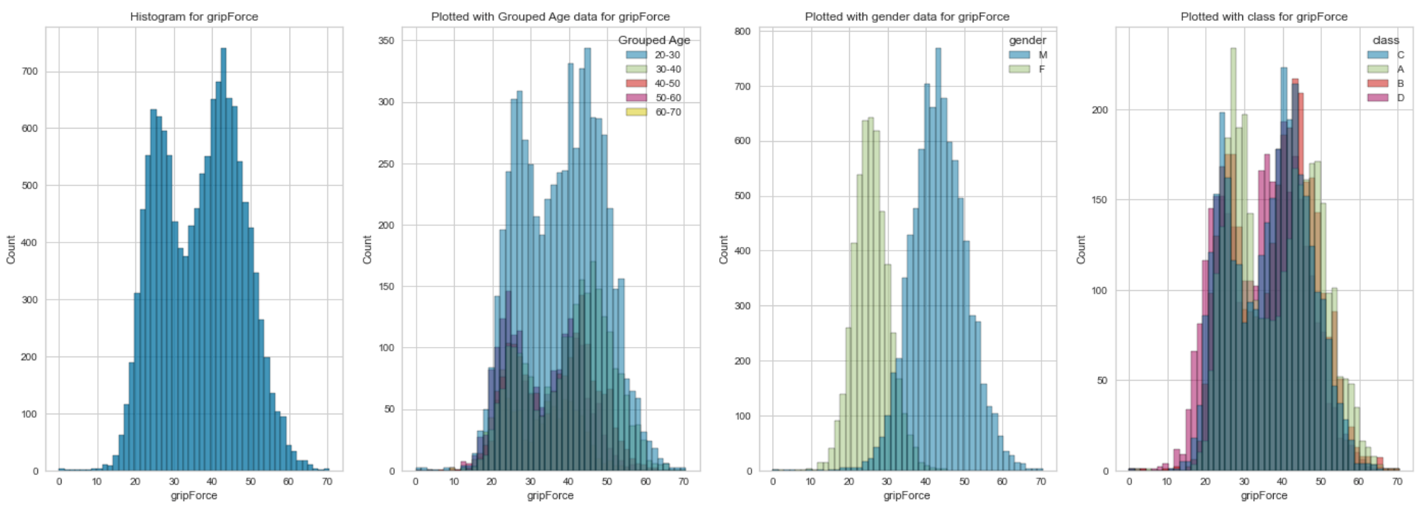
Here we can see majority of the people in this data is around the ages of 20 to 30.

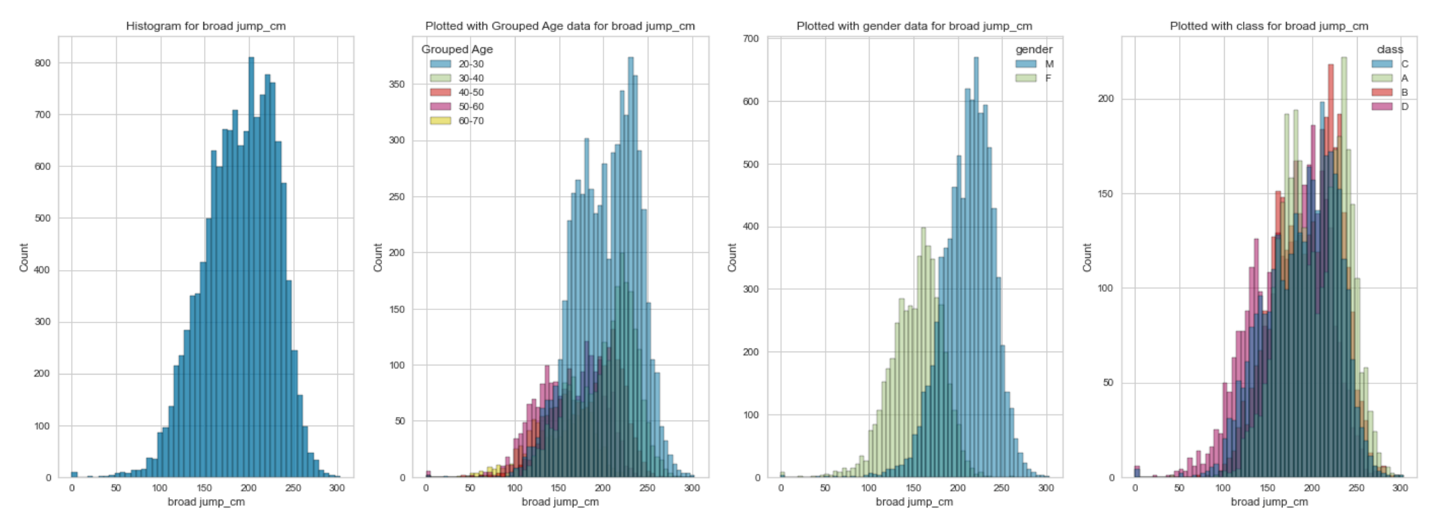


Class C have the highest body fat % compared to the other classes. Another interesting thing about this was that people of ages 20 to 30 have a significantly higher body fat % than the rest.

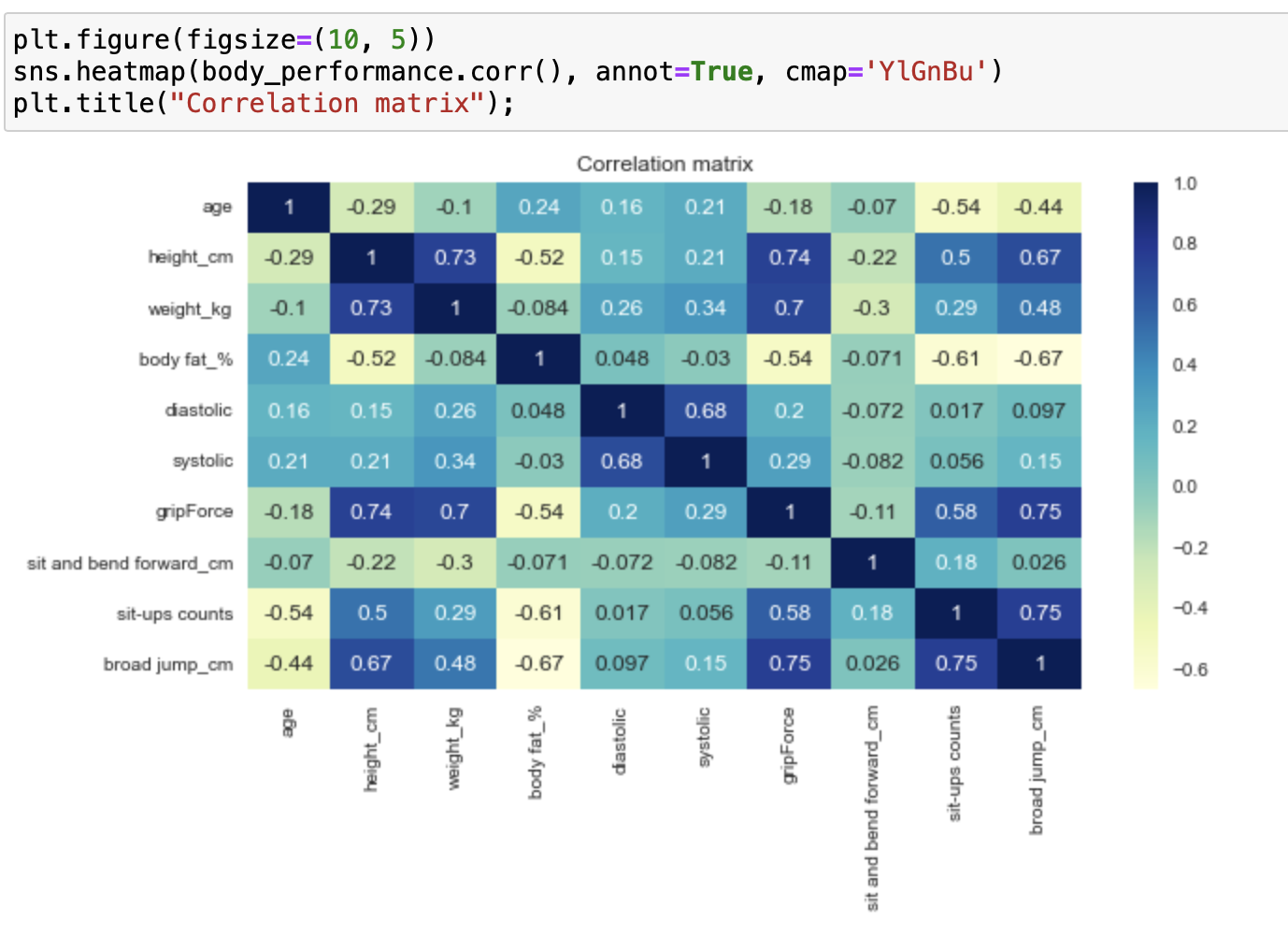


The males seemed to have both a higher blood pressure readings of diastolic and systolic.

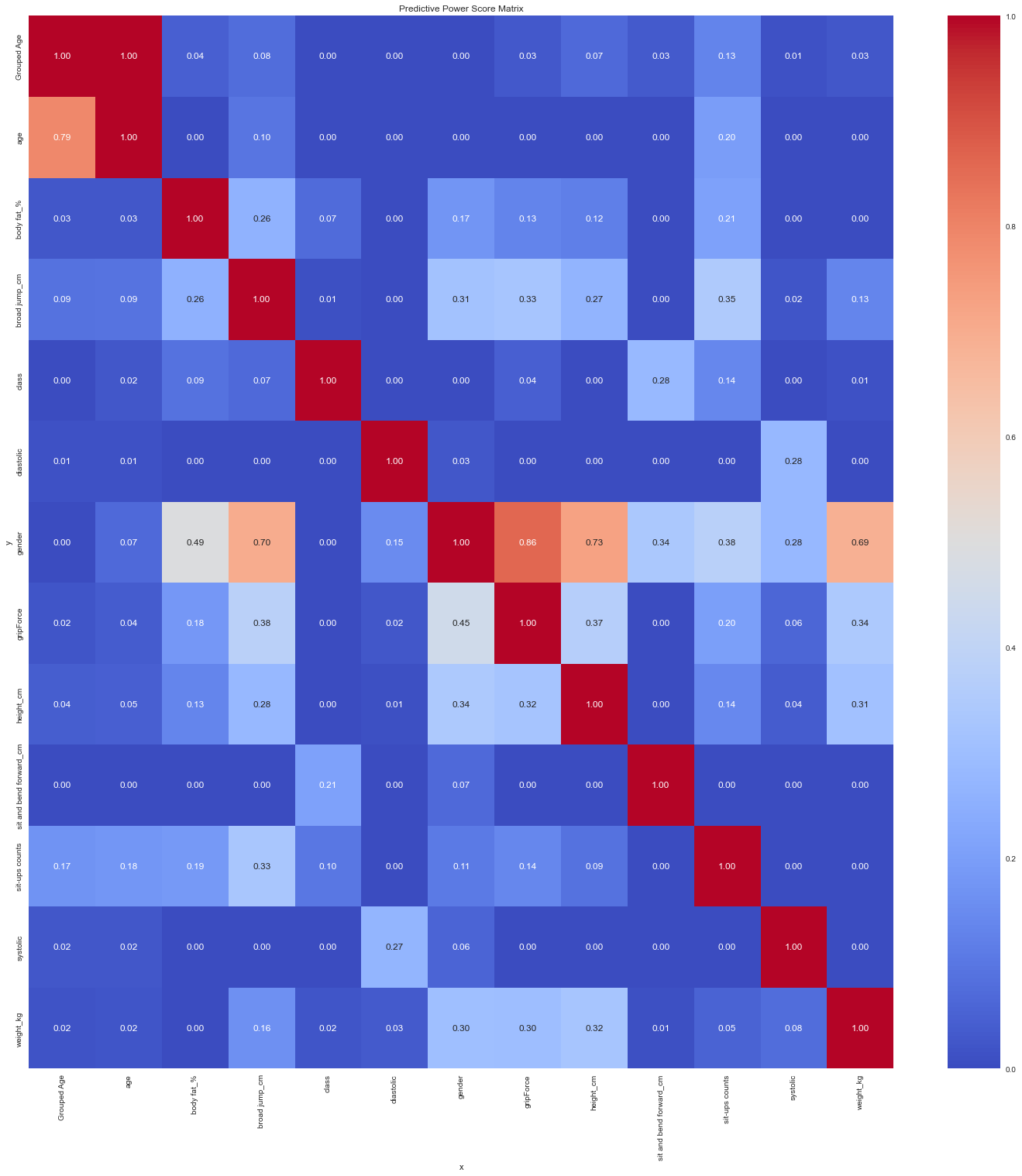




To get more readings out of this data I looked if there are any correlations between any variables. I generated two matrixes, the first one I looked into is the correlation matrix. We can see here that height and weight have a 0.73 correlation, height and grip force (0.74), height and broad jump (0.67), weight and height (0.73), weight and grip force (0.7), diastolic and systolic (0.68), grip force and broad jump (0.75), sit up and broad jump (0.75). All of these seemed to have a relatively high correlation. However, some of the of them did not make some sense to me. I realized that correlation matrix works only with numerical variables. In this case we are dealing with categorical data.

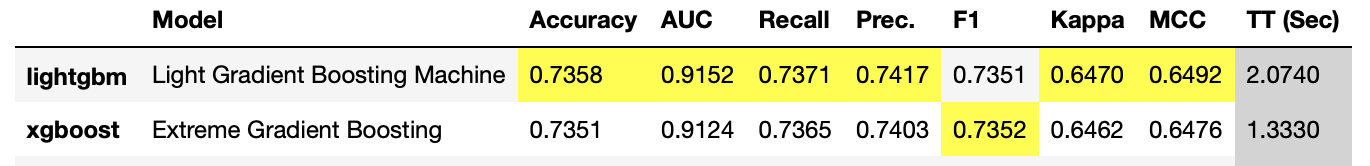
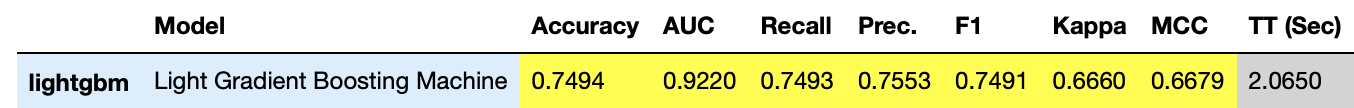
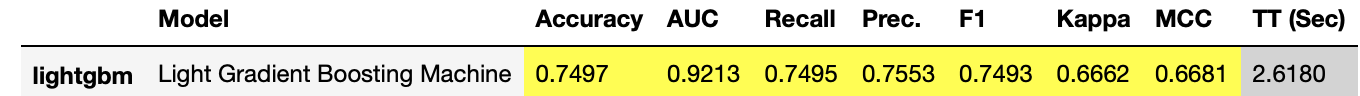
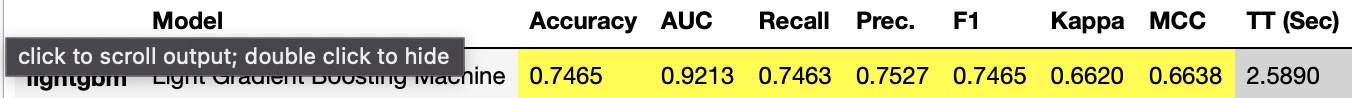
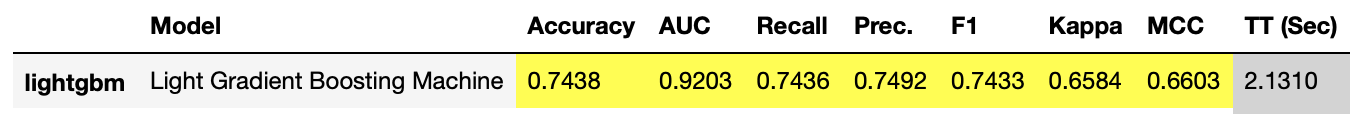
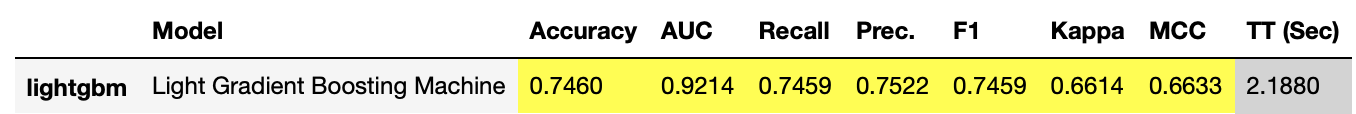
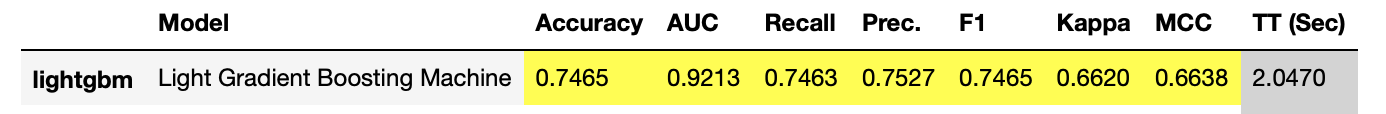


In such case that we are dealing with categorical data, I generated a predictive power score to get another reading for this data. This gave me a better reading of how each category are correlated. Here we can see that gender and broad jump is around 0.70, gender and grip force (0.86), gender and height (0.73), gender and weight (0.69). Here we can see that a lot of these are correlated with the gender. We will use this reading later to help developed a plan to attract more people into fitness.

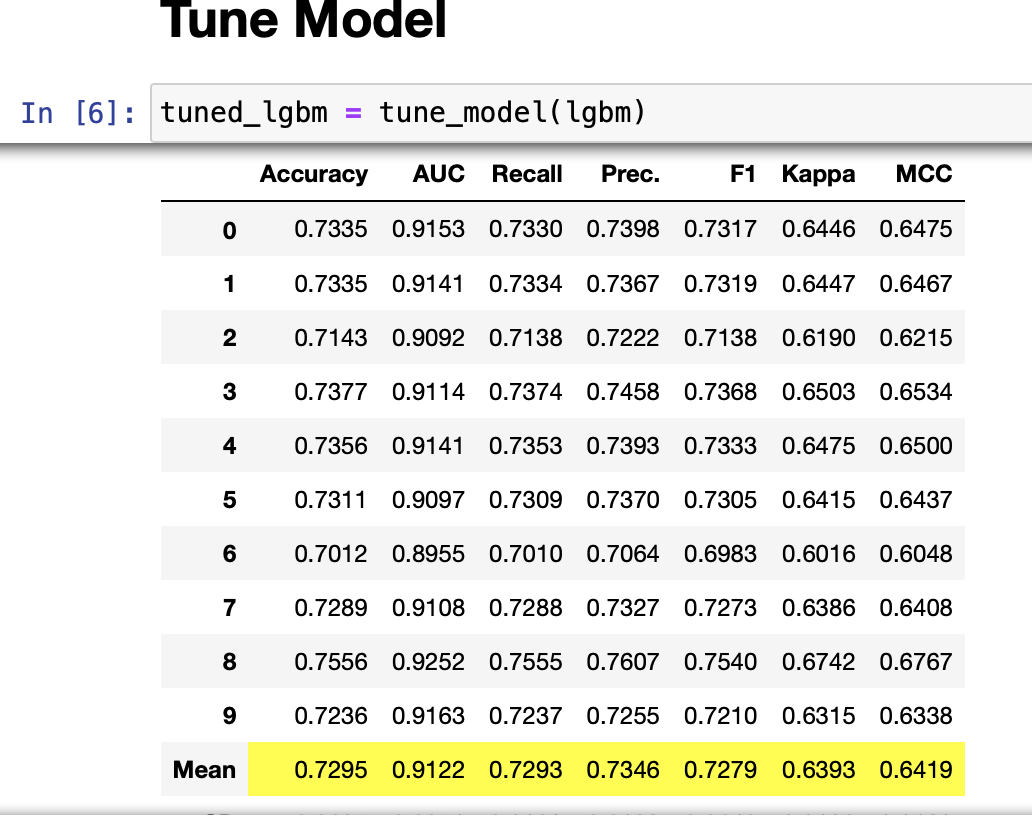


Model Selection

Here we try to compare each model through series of different testing to find the most accurate model we can create for out class. This was also done through jupyter notebook.



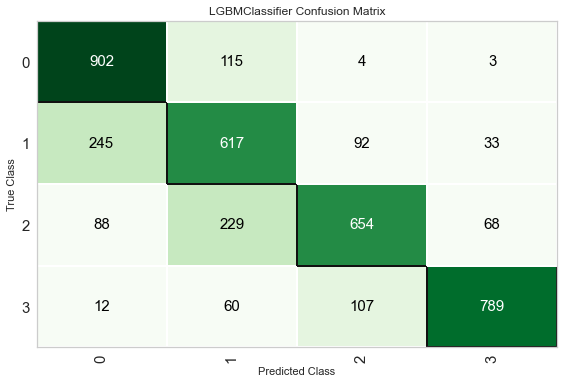
The Light Gradient Boosting Machine has generated a better accuracy for our class model. I used various methods to help find a better accurate one. I tried normalize method minmax, normalize method zscore, normalize method maxabs, normalize method robust, fix imbalance set to True, and removing the outliers. In the end the better one is the robust normalize method with a reading accuracy of 0.7487.

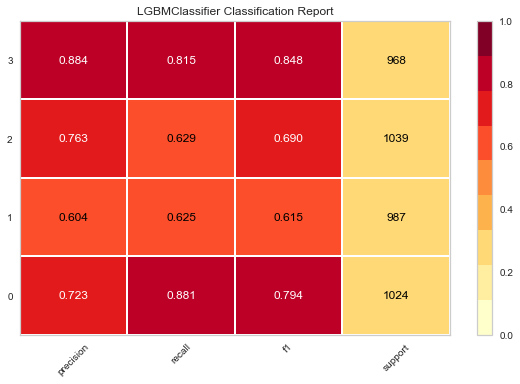


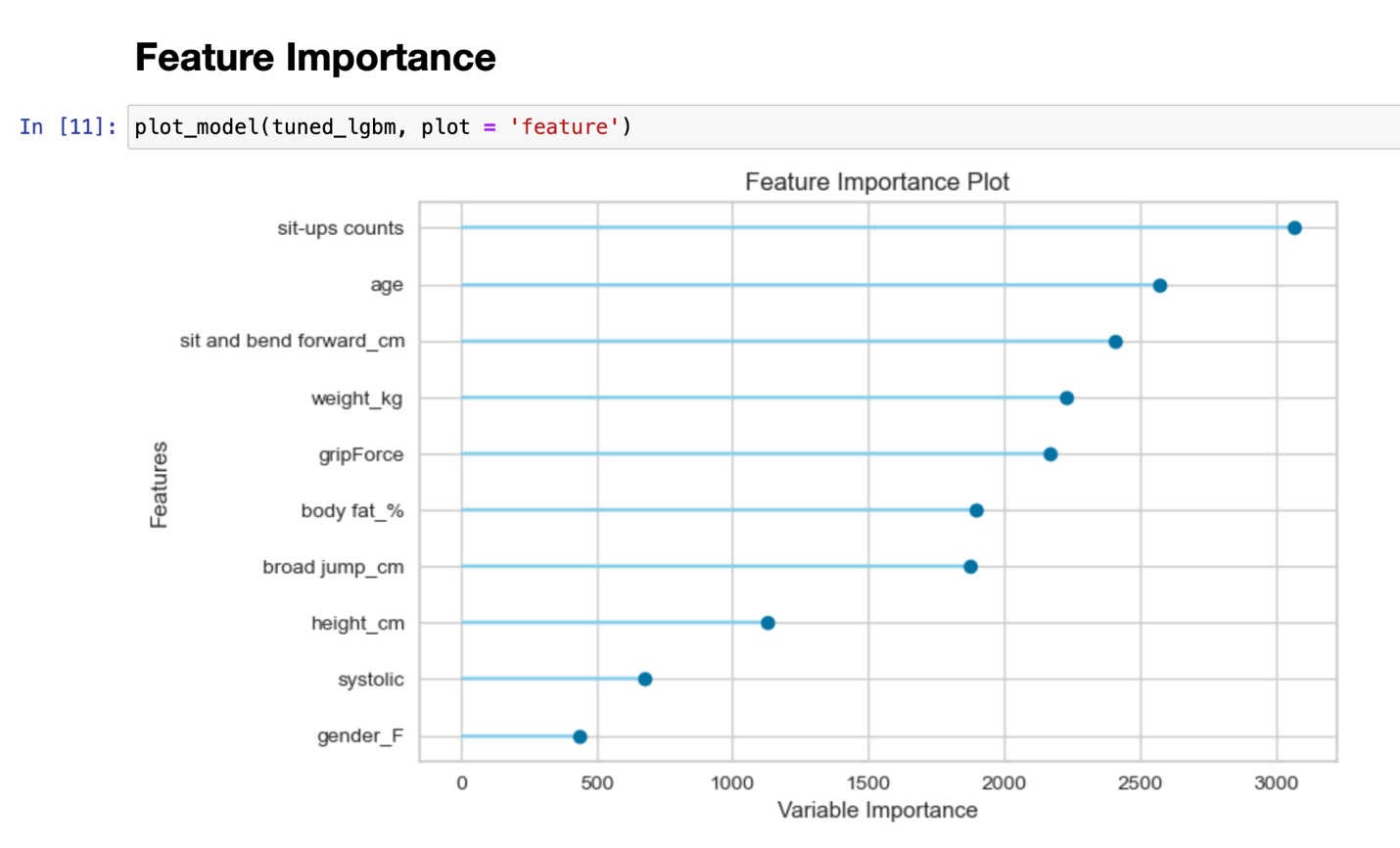
After tuning the model, the accuracy decreased it to 0.7295.

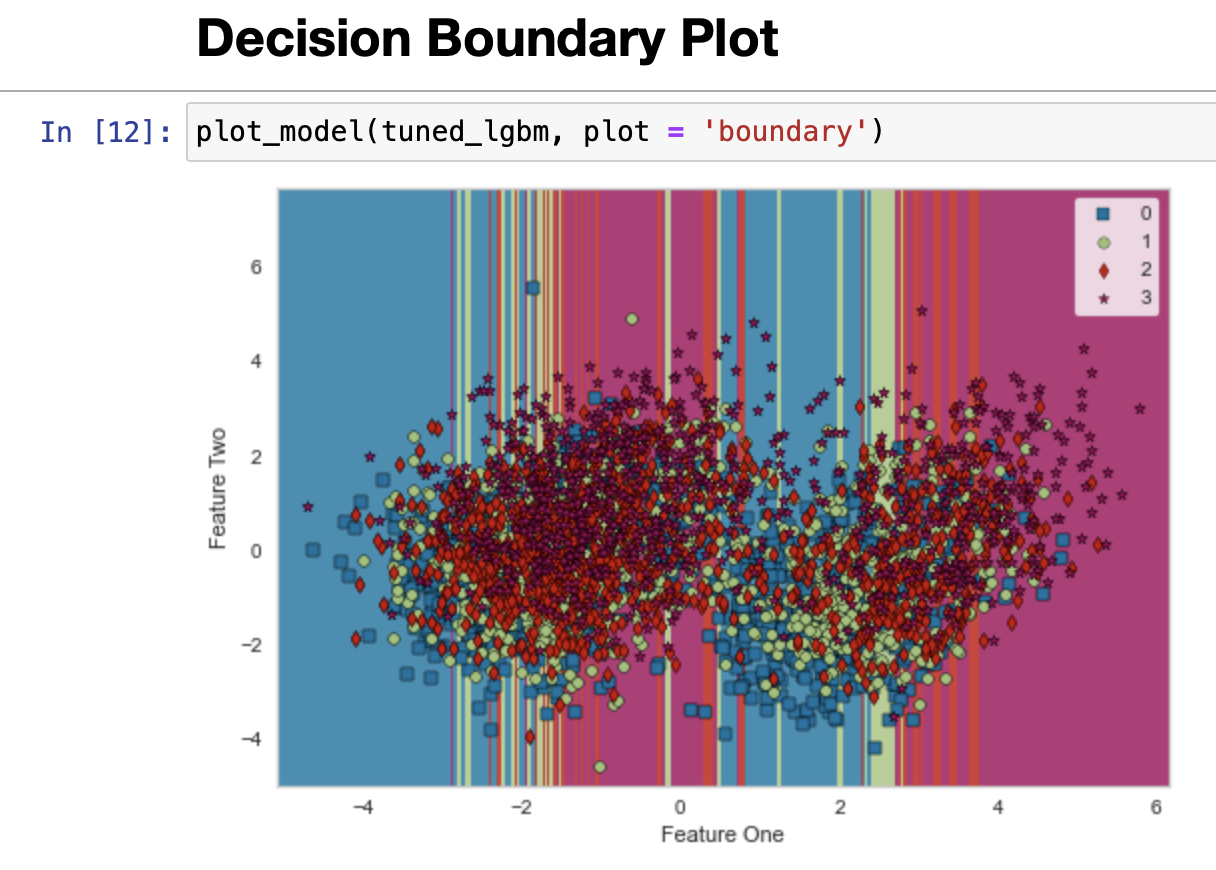
Model Performance

The performance of the model was generated using pycaret in jupyter notebook. The confusion matrix and the classification reports showed that the model is pretty accurate in predicting the body performance each class. The variable importance for the class seemed to show great importance of sit ups, age, sit and bend, and weight.









Recommendations/Future Work

Looking at the histogram we generated from the data we can see that males, ages 20 to 30 and in class C have a higher body fat % and higher diastolic and systolic blood pressure. With this information we could focus on creating courses and attract more people with advertisement that target the group who needed the most to help them get a better health. This will help increase the gym/class subscriptions.

From the model we generated it seemed like the most relevant key features to are sit ups, age, sit and bend and weight. This can help place each individual into the correct class to help improve their performance. Having people placed correctly in their class will help the program director to create specific programs to improve their performance.

For future work we can obtain more body performance data with different ethnicity to help us further improve their performance. Also, we can get more information on what the criteria are for classifying in each class from the institution. Lastly, we can create different programs to target specific body performance.