# Statistical Analysis and Machine Learning for the Relationship between Physical Fitness and Academic Performance using Python

Under supervision of Professor: **Ghada Hamed**

|  |  |  |
| --- | --- | --- |
| Name | Section Number | Seat Number |
| عمر محمد محمود اسماعيل | 19 | 20201700519 |
| مينا اشرف ميخائيل صالح | 30 | 20201700888 |
| دعاء يحيي اسماعيل حسن | 10 | 20201700243 |
| مصطفي محمد بيومي محمد | 28 | 20201700837 |

for the year: 2021/2022

Team ID: 80

## Introduction

## Are physically active people more productive? Are they better at learning than the average person? This question will be answered thoroughly in this research.

## This research was conducted to discover the relation between physical fitness and academic performance and achievements, and whether they are related or not. And what type of correlation they have in case it exists

## We chose this topic because it is often not discussed although it is quite an important one, by answering this question it’ll help us understand more about the human capabilities and what we can do to use them to their maximum potential, we also chose this topic to answer personal questions, simply because we were wondering if fitness has a positive impact on productivity or not, and if it does then by how much does it improve it?.

## Our goal is to find exactly how much fitness affects academic performance and explain their correlation with each other numerically and graphically, this will be done with the aid of the powerful programming language Python and the data analysis libraries: numpy, pandas, matplotlib and seaborn, first we will visualize the data we collected in our dataset, then clean the data and analyze it, finally we will build a machine learning model.

Research Summary

## Data

This data was collected during the fall semester of 2017 at Oral Roberts University, it has anonymous records of 581 students, each record contains data regarding the physical activity of each student using Fitbit, these data include Steps, Cardio, etc.… the students GPA is also added to the records for data analysis purposes.

## Statistics

Table

Description automatically generatedUpon loading the dataset in a pandas dataframe, the data description can be retrieved, through it we can see the statistical analysis of the one dimensional data in the data set, the mean, standard deviation, the quartiles, minimum and maximum values and the number of rows.

Description of data before cleaning

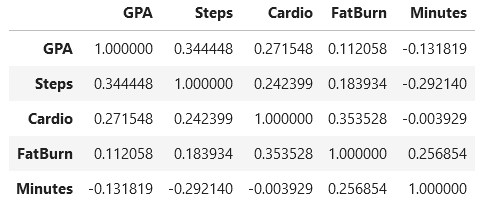
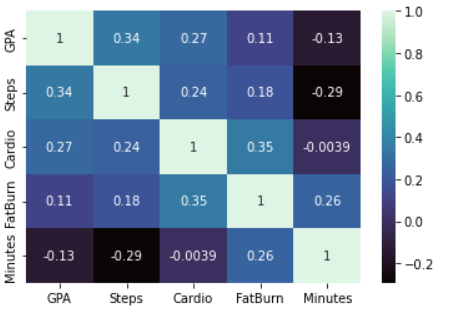
Data Cleaning

## Table Description automatically generatedThis dataset is not suitable for data analysis, this can be seen in *Fig (1),* the dataset contains lots of outliers, which will affect the accuracy of our calculations drastically, this is not good as it’ll lead to the sample data not being representative of the actual population, so to fix such problem, all we have to do is remove the outliers, this can be done simply using the user defined removeOutliers() function, this function simply sets the values of the data which are either higher than the higher bound or lower than the lower bound to null, so that later we can delete all rows containing null values, by doing so the dataset will be clean are ready for analysis. *Fig (2)* shows the box plot of the data columns after removing outliers.

## Description of data after cleaning

## It can be seen that the number of rows after cleaning is 476 which means 104 rows were deleted because they had extreme values.

## Correlation and Regression

To study the correlation between the different fields, the .corr() function will be used, this function returns a pairwise correlation of all the fields in the dataframe, from the the 5\*5 generated table we can see the value of Pearson’s correlation coefficient of any two variables, on the diagonal r is always 1 because the correlation coefficient of a variable and itself is always perfect. Most correlations in this data set are weak which means that the effect of the various physical fitness activities is low on the GPA, which answers our research question.

A picture containing text, crossword puzzle

Description automatically generated Heat map to visualize correlations pairwise correlation between all columns

Pairwise scatter plots to visualize regression between all variables

It is clear that most of

The variables in this data

Set have a weak

Correlation with regards

to one another.

The highest correlation

with the GPA is Steps and

the lowest is FatBurn,

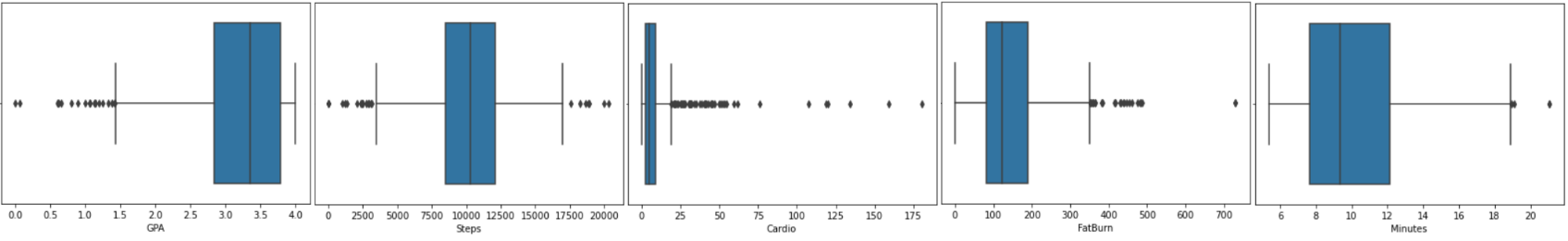
nonetheless both of them

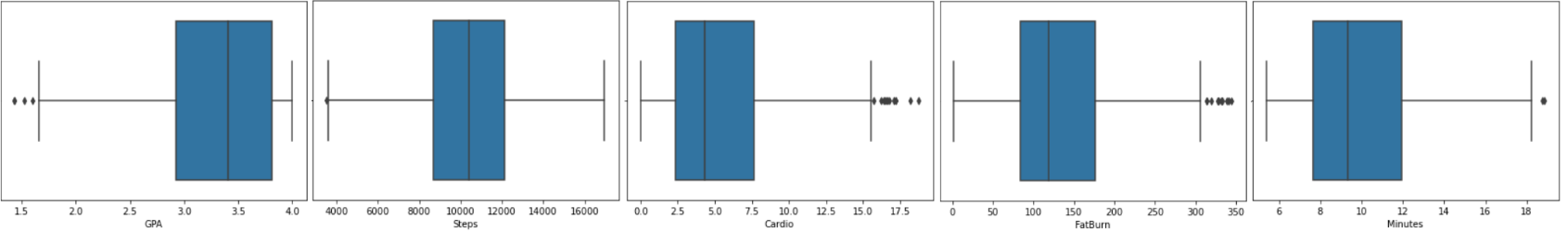
are considered a weak

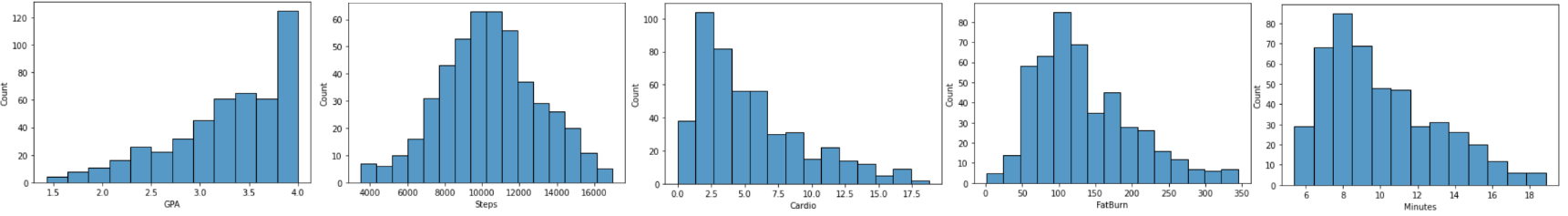
correlation.

Pairwise scatter plot to visualize the regression between all variables of the data set

## Charts and Graphs



Fig(1) box plots of one dimensional data before removing outliers.

Fig(2) box plots of one dimensional data after removing outliers.

Fig(3) Histograms of one dimensional data after removing outliers.

## Machine Learning Model

## Dependent and independent variables

The GPA is the dependent variable, because it is what we want to predict given the other data. The independent variable will be the rest of the data which are: Steps, Cardio, FatBurn and Minute, these will be used to predict the GPA.

Splitting data into training and testing data

In order to make a successful model, we have to split the data into two sections, training section which constitutes 80% of the data, this section will be used to train the LR model on predicting dependent variable given the independent one. The other section will be the testing section, it makes up 20% of the data set, it’ll be used later to check the success rate of the model after it has been trained.

This process is done through the “train\_test\_split” method in the sklearn library.

Standardizing the independent data

In this section, functions from the “StandardScaler” method in the sklearn library will be used to make the independent variable data more standard, meaning it’ll be more towards a bell shaped representation, this is done to achieve better performance later on when we use the linear regression machine learning model.

Training the machine learning model using linear regression

The model is trained here, where it has access to the train data of both the independent and dependent variables, using linear regression methods, the model learns how to predict y given x.

This is done using “LinearRegression” method in the sklearn library

Predicting the output (testing) the model

The model is now given testing values from the 20% that were sectioned for testing from the independent variable, the model should now predict the dependent variable on its own.

Comparing the predicted output to the actual output present in the data set

The (r2\_score) method of the sklearn library is used to determine success rate of the model (coefficient of determination) where it compares the predicted output with the actual one.

Here our model operates with a success rate of 12.48%, this means that 12.48% of the variation in y can be described by x, which is very low and renders it unusable and unreliable.

## Sources

## Data Set:

## <https://figshare.com/articles/dataset/Dataset_Fitbits_field-tests_and_grades_The_effects_of_a_healthy_and_physically_active_lifestyle_on_the_academic_performance_of_first_year_college_students_/7218497>

## Conclusion

Through this research it can be clearly seen that physical fitness has a weak positive correlation with the academic performance, this means that spending more time at the gym or simply doing workouts at home will definitely boost your academic performance, but the difference is not that big and it is barely noticeable in some instances, this also means that by doing more physical activities you won’t automatically have a high GPA, the key is to achieve balance between everything, we have to find the sweet spot, by that I mean that no one should exercise more than often or not at all, everything should be done in moderation, moderation is always key to success.

Since the correlation between the data was weak, the machine learning model we built is not very effective, as this simple model relies on linear regression, basically predicting the output of y through x through the regression equation which would normally be functionable if that data is more correlated. So the use of a linear regression model to predict data in this scenario is not advisable and is definitely not useful in any way.