CMPT 353: Computational Data Science

Summer 2018

Final Project: Walking Sensor Data

Brittany Hewitson

301249491

Greyson Wang

#########

Philip LeBlanc

########

# Introduction

# Methods

## Data Collection

Data was collected on a range of individuals using the PhysicsToolbox Suite by Vieyra Software (insert reference here: <https://www.vieyrasoftware.net/sensors-sensor-modes>) for Android cell phones. The cell phone was tied to each subject’s ankle while they walked on flat ground for approximately 20 seconds. During this time, the g-force, linear acceleration, and angular velocity in the x, y, and z directions was recorded by the app. The process was then repeated on the subject’s other ankle.

A random sample of subjects were chosen who all possessed different qualities. To accurately model the relationship between step frequency and the physical attributes of an individual, 18 subjects of varying heights, weights, ages, levels of activity, and activity of choice were chosen for data collection. The most important of these categories were height, levels of activity, and activity of choice. This was because our main questions were centered around determining whether an individual’s height affected their stride length, therefore altering their step frequency; as well as investigating whether individuals who were highly active had different step frequencies than those who were less active. Additionally, we wanted to determine whether individuals who participated in different activities had different step frequencies, such as if runners had a higher step frequency than cyclists.

In addition to comparing different subjects we also performed a series of tests on the same individual, collecting multiple datasets of them walking on flat ground and then again walking on stairs. The data was used to create a model aimed at predicting whether the individual was walking on flat ground or on stairs based on characteristics such as the step frequency and angular velocity.

## Data Processing

Various libraries in Python were used to process the data, namely pandas, NumPy, SciPy, matplotlib, and scikit-learn. The data processing can be split up into two categories: data amongst different subjects where the relationship between physical attributes and step frequency were investigated, and data amongst the same subject where a model was created to determine whether the data represented the individual walking on flat ground or on stairs.

### Data Processing Among Different Subjects

#### Extract-Transform-Load

Before any of the data could be used for analysis, a series of extract-transform-load (ETL) operations were performed on the raw data. The two main data sources for this portion of the project were the walking data collected on the ankles of each subject and the information about the subjects, such as their age, weight, height, etc. The walking data consisted of g-force, linear acceleration, and angular velocity measurements. For the purpose of this project we decided to extract only the linear acceleration and angular velocity measurements in the x-, y-, and z-directions. Each component was then used to calculate the vector sum, producing the total linear acceleration and the total angular velocity measured at each point in time. The equation used to calculate the vector sum is shown below.

The data was then further transformed by filtering with a low pass Butterworth filter to eliminate high frequency noise present from the sensor readings. This was accomplished through the use of SciPy’s signal.butter() method, with an order of 3 and a cutoff frequency of 0.1 half-cycles per sample. Thus, this function designed an 3rd order digital Butterworth filter and returned the filter coefficients. The coefficients were then used in SciPy’s signal.filtfilt() method, which used the coefficients to apply the filter on the total linear acceleration and angular velocity data.

Once filtered, the data for each subject was divided in half, providing approximately 10 seconds of walking data for each of the subject’s left and right ankles. We chose to divide the data in an attempt to provide more points for training the machine learning models. The next transformation was to take the Fourier transform of the divided linear acceleration data using NumPy’s fft.fft() method in order to determine the step frequency for each subject. Because the linear acceleration of the ankle as a function of time reflects a sinusoidal shape, the Fourier transform of the data consisted of a series of impulse functions located at the various frequencies found within the time-domain data. The largest peak was representative of the frequency most commonly found within the data, thus we extracted the frequency at which this peak was located and labeled it as the average step frequency for the subject.

The frequency spectrums of each subject’s ankles was saved into a dictionary to be used later in the analysis. The average step frequency was saved to the dataframe holding the physical attribute information about each subject.

Once the data was transformed and stored in the appropriate locations, it was used to analyze the relationships between the step frequency of each subject and various attributes of the subject. This was accomplished through the use of machine learning classification and regression.

#### Machine Learning Classification

Classification was used to investigate the relationships between the step frequency of each subject with their level of activity, activity of choice, and gender. Various classifiers were used from Python’s sci-kit learn library, including the naive Bayes classifier, nearest neighbour’s classifier, and support vector machines. The data was split into training data to train the models and testing data to test the model on data it had not seen before. The scores of the resulting test outputs were then printed and compared with each other.

#### Regression

Regression was used to investigate the relationships between the step frequency of each subject with their height, age, and weight. Multiple regression techniques were used to investigate these relationships, including SciPy’s stats.linregress()and sci-kit learn’s .LinearRegression() for linear regression, NumPy’s .polyfit() and .polyval() for polynomial regression, and statsmodel’s .OLS() for an ordinary least squares tests. Similar to the machine learning classification, the data was split into training and testing data, where the training data was used to produce each of the regression models. The models were then used on the testing data to predict the output, and the resulting graphs were compared with each other.

Before the data was used in any of the models, the frequency spectrums for each subject were compared with each other in an analysis of variance (ANOVA) test to determine whether they had the same means. The ANOVA test used to determine this result was the stats.f\_oneway() method.

### Data Processing Among the Same Subject

# Results and Analysis

## Frequency Analysis

The results of the frequency analysis were used throughout the rest of the project. The figures below show a typical frequency spectrum for the total linear acceleration.

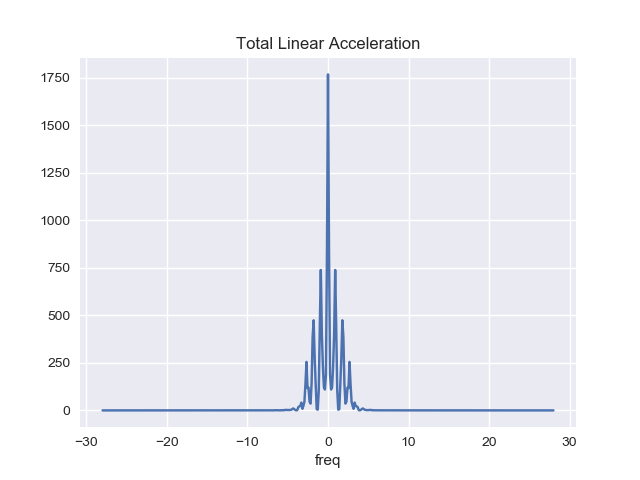


Figure 1: Frequency spectrum of the total acceleration untransformed

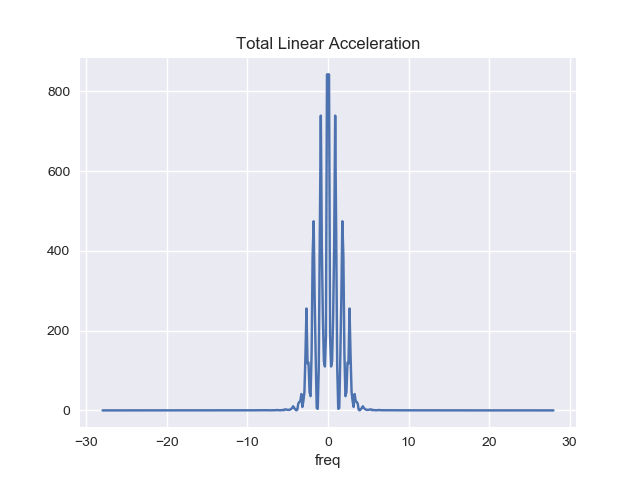


Figure 2: Transformed frequency spectrum of the total acceleration

As shown above, the frequency at 0 steps per second is the largest. As this is not a possible average step frequency while an individual is walking, the frequency spectra were transformed to eliminate this DC offset value. Thus, the highest frequency greater than 0 steps were second was chosen as the average step frequency.  The DC offset may be caused by our measurement, since there is a one or two second delay from when the subject stops walking and when the data measurement is stopped. This period where the acceleration and velocity are not measured could contribute to an extremely low frequency component in the frequency spectrum.

The above figures appear to have a distribution close enough to that of a normal distribution. However, the results of the normal test on each spectrum produced a p-value much less than 0.05, thus implying they are non-normal. The Central Limit Theorem states that when the number of samples are large enough we can approximate the data as having a normal distribution. In this case, there were 72 data sets with approximately 200-500 data points within each set, thus for the remainder of the project we will assume the frequency spectra are normally distributed.

## Classification with Multiple Subjects

The three classifiers used to identify a subject’s level of activity based on their step frequency all performed around the same. The scores of each classifier are tabulated below for their accuracy in identifying a subject’s level of activity, activity of choice, and gender.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Summary of the performance of the classifiers for level of activity   |  |  | | --- | --- | | Classifier | Score | | Bayesian | 0.389 | | kNN | 0.444 | | SVM | 0.333 | | Table 2: Summary of the performance of the classifiers for activity of choice   |  |  | | --- | --- | | Classifier | Score | | Bayesian | 0.444 | | kNN | 0.111 | | SVM | 0.167 | |

Table 3: Summary of the performance of the classifiers for gender

|  |  |
| --- | --- |
| Classifier | Score |
| Bayesian | 0.556 |
| kNN | 0.722 |
| SVM | 0.556 |

From the above tables, the k-nearest neighbours classifier performed the best when identifying gender and level of activity, with the Bayesian and support vector machine (SVM) classifiers producing similar accuracy scores. When identifying the activity of choice, the Bayesian classifier performed the best with the k-nearest neighbours and SVM performing rather inadequately in comparison. It can be expected that the k-nearest neighbours perform the best in most cases because of the nature of the data. Most of the step frequencies are relatively close together, averaging around 0.768 steps per second, with a minimum frequency of 0.101 steps per second and a maximum frequency of 2.080 steps per second. If most of the frequencies are relatively similar, the probability of the different categorical groups having large differences between their frequencies is very small, thus the Bayesian classifier will have a difficult time distinguishing between whether a frequency is representative of say a highly active person compared to a moderately active person because the data points overlap too much. This can also explain why the SVM is consistently performing poorly, as this method relies of predicting boundaries between the groups of data.

A common trend send amongst all the classifiers however is that they consistently produced fairly low scores in each case. The low scores could be a result of several possible factors. The first being that there is no relationship between the constraint being analyzed and the step frequency, therefore it is difficult to generate a model that predicts these attributes given the step frequency. From these results it appears as though there is no relationship between how active someone is and how fast they walk, or whether the activity of choice plays a role in the walking speed. There does appear to be a relationship the gender and walking speed, as the classifiers consistently produced the highest accuracy scores for this category. After investigating further, it appears that the male subjects had a slightly higher average step frequency of 0.76 steps per second compared to the female average at 0.74.

The second factor is that we may not have provided enough data to effectively train the models on. Without enough data points for training, the model would not be able to accurately identify a relationship between the input and the output, and therefore it would not perform well when presented with the testing data. With only 72 values for the average frequency, it is highly possible that there was not enough data to train an accurate model in this project.

## Regression with Multiple Subjects

Both linear and polynomial regression were used to investigate the relationships between the average step frequency and the subject’s height, weight, and age. Statistical linear regression and machine learning linear regression were both used on the data, and both produced similar outputs. From the figures below we see that the data is not linear, thus the linear regression model produces poor results. The p-value for the statistical linear regression was 0.479 for height, 0.999 for age, and 0.578 for weight, which are all greater than 0.05. Because of this, we do not have enough information to reject the null hypotheses, thus we cannot say the slope of the linear regression is non-zero for any of the regressions. As a result, the linear regression model does a poor job of predicting the average step frequency based on a subject’s age, height, or weight.

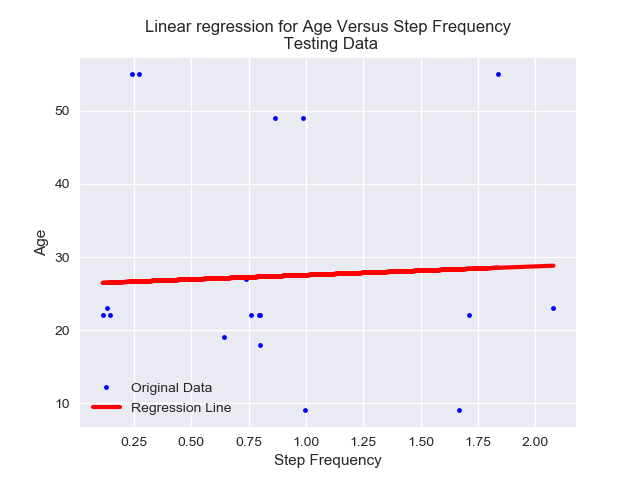


Figure 3: Linear regression using statistical modelling for subject age

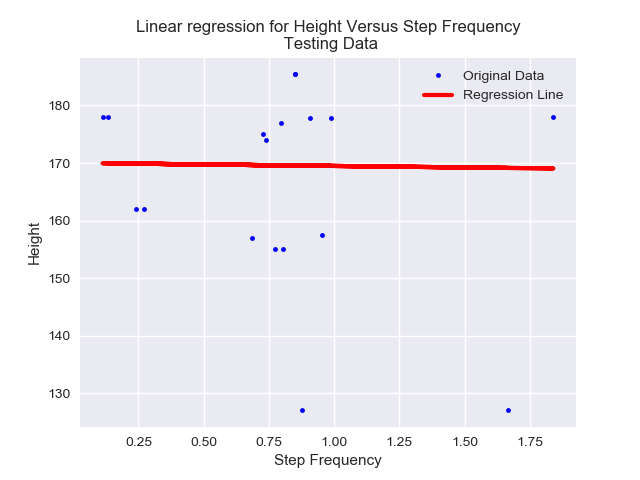


Figure 4: Linear regression using statistical modelling for subject height

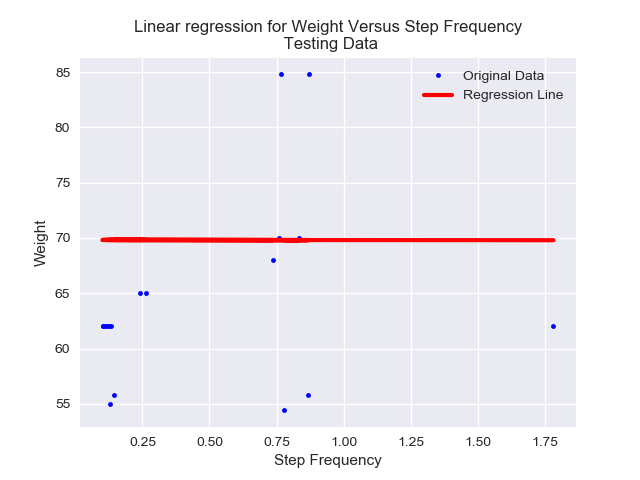


Figure 5: Linear regression using statistical modelling for subject weight

The results of the polynomial regression were much better when compared to the linear regression, although still not entirely accurate. As shown by the graph below, the polynomial regression fitted closely to the testing points.

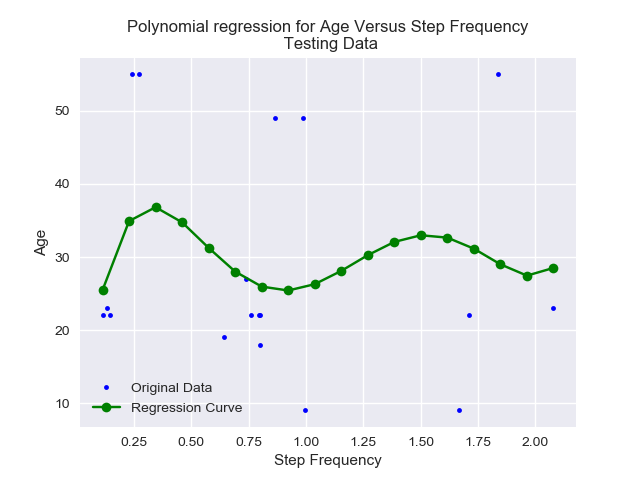


Figure 6: Polynomial regression for subject age

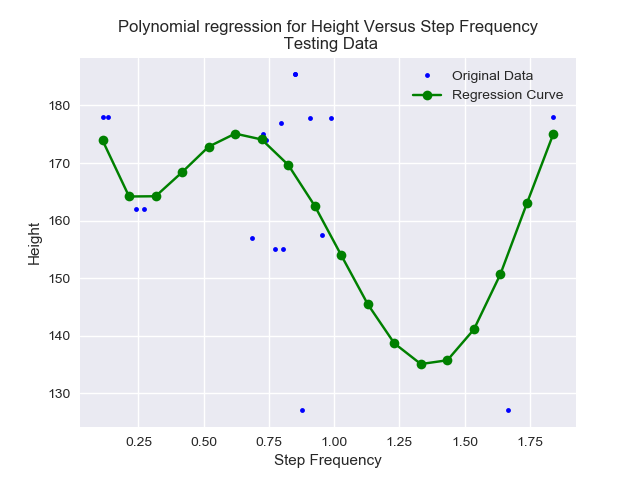


Figure 7: Polynomial regression for subject height

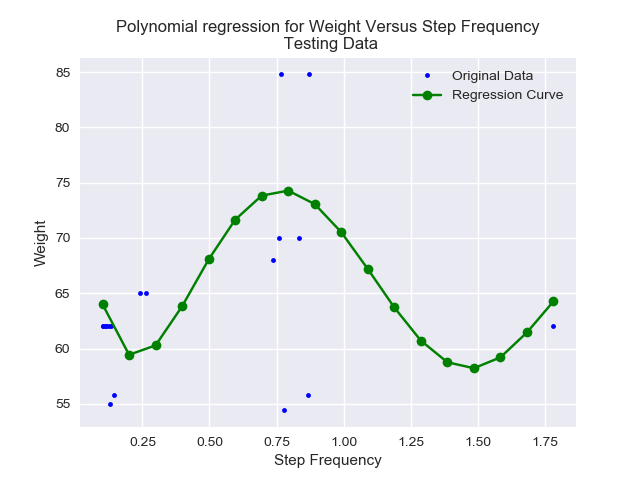


Figure 8: Polynomial regression for subject weight

## Analysis of Variance Test

An ANOVA test was performed on the frequency spectra for each subject to determine whether the means were the same. The resulting p-value was much less than 0.05, which provides us with sufficient evidence to say the means of the frequency spectra were not equal across subjects. The figure below shows the distribution of average frequencies for all of the subjects. This supports the theory that different individuals have different step frequencies, possibly affected by characteristics such as their height and level of activity.

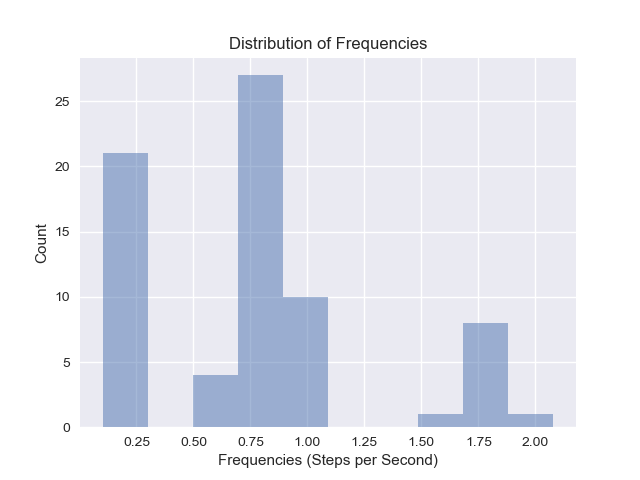


Figure 9: Distribution of the average step frequency for subjects

## General Conclusions

The reason for the poor results for the classification of a subject’s qualitative attributes and the prediction of their quantitative attributes are very similar. Firstly, there may not be a relationship between the step frequency and these attributes. This is very likely in the case of classification, as the step frequency may not vary much between genders, or the preferred activity of an individual. However, this reasoning is less likely in the case of regression. This is because the height of a person can be linked to their step frequencies, as individuals with longer legs may take less steps to travel the same distance as someone with shorter legs.

Additionally, there may not be enough data in each case to create an accurate model. This is the most likely cause of the poor results, as there were very few points to train the models on. As a result, the models could not accurately predict the outputs. Future considerations for this project should include collecting more data on individuals, whether data is collected on a larger number of subjects or if several data sets are collected on each subject.

Another possible explanation for the results could be that the data collected was not entirely accurate. Because the data was collected on a cell phone that was tied to the subject’s leg, it may not have accurately captured the walking pattern for each person. Instead, the sensors could have measured periods where the phone was loose on the subject’s ankle and therefore was “bouncing” around during the walking phase. This would correspond to the higher frequencies, such as the 2 steps per second; as this corresponds to a running pace rather than a walking pace.

Furthermore, when collecting the subject’s quantitative data, they were not directly measured but rather had to recall their heights and weights. This could have affected the training of the model, because if an individual did not record their correct measurement, the model would be training on incorrect data. Thus, when the model was introduced to training data it would not accurately classify the subject’s qualitative characteristics nor predict the quantitative measures.

Finally, the subjects chosen to participate in this experiment could have affected the results. As shown by the figures below, there was not a large variation in the quantitative and qualitative measures of the subjects who participated in the experiment. Most subjects had similar ages and heights and participated in similar activities. Because of this, the models may not be able to accurate classify or predict the output of the testing data.

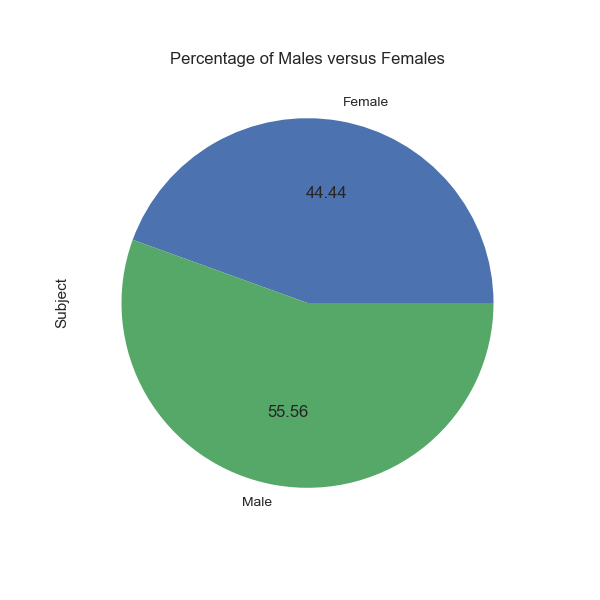


Figure 10: Distribution of genders for involved subjects

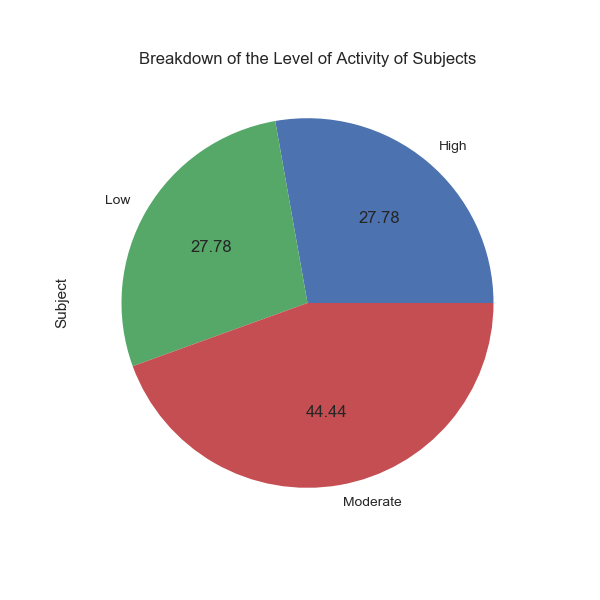


Figure 11: Distribution of the level of activity for involved subjects

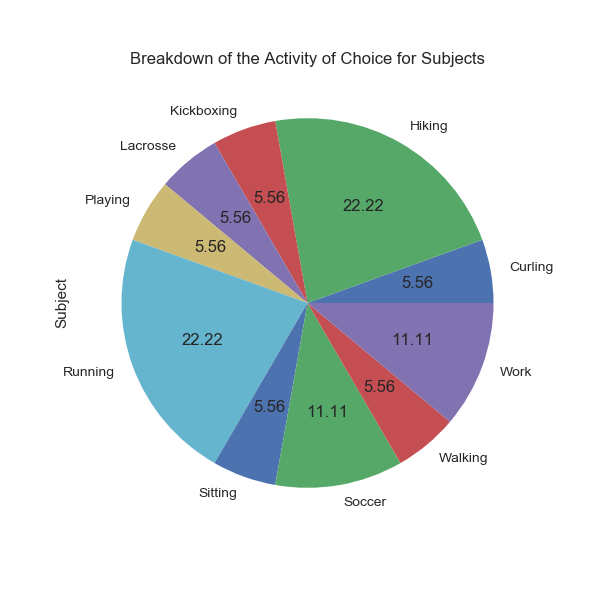


Figure 12: Distribution of the activity of choice for involved subjects

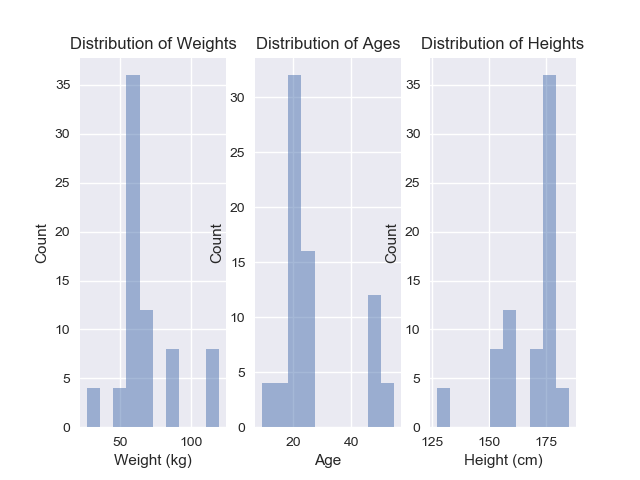


Figure 13: Distributon of weights, ages, and heights for involved subjects

# Conclusion