WISDM Dataset analysis

Degree: Data Science BSc (Hons)
Module: Big Data & Visualisation 2022 CIS3152

By: Christopher Diaz Montoya
ID: 24707686

Module Lecturer: Muhammad Awais

Contents

Introduction 3
Visualisation tools 3
Implementation 3
Data Check, Upload and Merge 4
Data Pre-processing 4
Data Analysis 4
Interpret Data 5
Data Processing Solution 7
Further Work 7
Bibliography 8
Appendices 9
Appendix A 9
Appendix **B** 9
Appendix C 10
Appendix D 10
Appendix E 11
Appendix F 11
Appendix G 12
Appendix H 13
Appendix I 13
Appendix J 13
Appendix K 14
Appendix L 14
Appendix M 15
Appendix N 15
Appendix O 16
Appendix P 16
Appendix Q 17
Appendix R 17
Appendix S 18
Appendix T 19

Introduction

The aim of this report was to visualise the WISDM dataset, explore reportable insights and to propose a data processing solution. The WISDM dataset holds accelerometer and gyroscope sensor data collected by Fordham University from phones and smart watches while doing eighteen different activities. The data collected was in the following format: Participant ID, Activity, Timestamp, x-axis, y-axis, and z-axis. The final three data points collected allow for a 3D visualisation, a 2D graph uses the x and y-axis, the z-axis adds another dimension to the data.

The report is split into five sections: The first is about visualisation methods and how the data can be turned into insight. The second explains how the data was analysed through the implementation use of python programming language in Jupyter notebooks. Various libraries were used to manipulate and analyse the data set provided. The third explains how the techniques found in visualising tools could provide insight and what could be interpreted from that. The fourth section proposes a data processing solution, this would be an algorithm that the dataset could be used to make predictions with this is called a machine learning algorithm. The final section is the conclusion and critiques the project paired with looking into future work.

Visualisation tools

This section discusses different data visualisation tools found that are in relation to the data in the WISDM data set and what could be done.

The data is given in a structured format within txt files. It was also given in ARFF files which contained a lot of data, that was left aside as it was hard to visualise any meaningful data without more subject matter knowledge.

Another attempt at the WISDM dataset analysis was created by Louly (2019). The analysis visualised the different activities for one participant on a line graph as shown in Appendix A and Appendix B. This seemed to limit the amount of data that could be visualised and would need a large number of graphs to find patterns between different activities. A heat map was also used as shown in Appendix C, but again it was only for one activity and would take a while to sit down and find difference between each individual graph Louly (2019) created.

As the data contains three axis of data, the creation of a 3D graph was researched. A 3D scatter plot was found, and this visualization tool was able to display a large amount scatter plots and looked visually appealing (Geeks for Geeks, 2020). It could also be used to create 3D line charts which could potentially show how the sensors were moved in a 3D space (Seb, 2022)(Koolac,2022) for each individual participant.

Implementation

This section of the report should be read alongside the code as the comments in the .ipynb file help explain what was done.

Data Check, Upload and Merge

After reading and understanding the task at hand the data upload was the first thing that needed to be done to analyse the data. As there were multiple files, the raw text files contained less data so were used.

To upload the data, various libraries were needed and as the project was carried out, more and more libraries were needed. A style for the graphs was also chosen to try make the graph background themed with a horizontal grid line with matplotlib for a quicker to interpret visualisation.

Once the relevant libraries were imported. The raw data set given in a .txt file was uploaded to the algorithm individually. Given more time this could have been done in a function or for loop with the file paths stored in a variable to reuse code for the four raw data sets.

Next up the data frame was checked with df.info and df.shape(). The first allows to view the total amount of rows and columns, while the latter allows to see what data type each column was and if it was as it needed to be. This was done for each of the four raw data sets and will stay like this for the remainder of the project, four separate data sets but all treated the same way.

In the activity column, letters were showing up and not allowing for an easy interpretation, it used a .txt file as a key. This .txt file was uploaded and formatted into a dictionary so that it would replace each letter in the data frame with the respective activity. This could have also been added as a function and reduced code given more time.

As there was now a dictionary which had the key equal to the values in the data frame and the values ready to replace it, adding the actual activities was done with the .replace() learnt from last year. The data frames were then printed to check that the activities were uploaded.

Earlier the data types of each column in the data frame were checked, as the timestamp was and integer data type, this needed to be converted into a datetime data type. This was attempted in python and using online converters with no success. This was done with some code created by Clarke (2013).

Data Pre-processing

The data was then checked for completeness and that it was all in the right format, this is called data pre-processing.

When checking the data, each columns data type was checked. This was to ensure numbers in the z column were read as floats as floats contain decimals. It was of the object data type so the algorithm needed to convert it to a string and then it could be converted into a float.

Data Analysis

Feature exploration was carried out to find with cross tabs to create a 51 x 18 table to view the data briefly and see how many readings each participant had for each activity. This was then visualised in a bar chart, alongside the number of readings each activity had (Data camp, 2021).

Following this a heat map function was created. The data should have been normalised here. The Heat map was able to show which activities had a high correlation with which axis across the four data sets.

This was built on to create a 3D scatter plot graph using the standard deviation, as Louly (2019) attempted 2D graphs it showed a high variance, this means it is good to use the standard deviation over the mean (Statistics Canada, 2021). If the mean was used then some activities may overlap unexpectedly, even when the sensor readings could be as different as Appendix A and Appendix B. This first set of 3D scatter plots contained every row in each data set to spot trends.

This was done for all four data sets as different sensors have different readings, this could have been explored further to find a correlation between sensors, to merge them and get a more accurate reading, for example the watch and phone accelerometer joined together if the activity on both had a strong trend.

The standard deviation was then looked at for each activity as a whole, this plotted one plot per activity, this was done to attempt to further isolate each activity to spot trends better withing the data sets.

Interpret Data

The data set uses two types of motion sensors gyroscope and accelerometer for both phone and watch.

Accelerometer measure acceleration, it measures linear velocity across any axis. For example, a car moved sideways by a thief. Often slow and cannot catch slighter delays and accuracy. It cannot measure the speed of rotation in aircrafts or the stability of flights. Gyroscope sensors are more accurate for rotational measurements. It measures the rotational speed. Gyroscope sensors can measure stability and how fast an object is rotating such as Wii Controllers. Both together give you 6 Degrees of freedom, this means that tracking a device going forwards and backwards, with the speed it is travelling at, along with if the device is rotating is all possible (Symmetry electronics, 2022). An example of this would be a watch with both sensors and a participant dribbling a basketball. When running up and down the court the accelerometer would record the speed of going from end to end of the court, the gyroscope sensor would be able to track the palm movement, for example if the palm is facing down dribbling or if it is holding the ball and facing the rim. This could be taken a step further and have a shot accuracy prediction based on the sensor readings and where the participant is currently on the court.

Participant with the ID 1629 had roughly double the amount of data points as all the other participants, this can be seen in Appendix H, I, J and K. This at first was thought to be an outlier with the sensor being faulty and possibly taking double readings when looking at just appendix H. The chances that all four of the participants sensors being faulty was unlikely so hence why this participant was not removed. The participant due to having a larger score could possibly skew the data set, normalisation should have been done on the dataset to avoid this.

Appendix L, M, N and O show each data set visualised in a rotatable 3D format, a few trends were discovered here.

For the watch gyro sensor data, dribbling was what made the gyroscope sensor have the most variance in the graph which was as expected. Appendix L looks like an L, from this smaller colour groupings could be found along the L shape, this shows the potential for different categorical analysis. A lot of activities do also overlap, and this is across Appendix L-O, this shows that some categories cannot be identified in their own right clearly due to having similar readings which may be mis classified, as such fore a data processing solution the number of activities should be reduced. Activities involving food have a low y-axis score due to not jumaring up and down but a high z-axis due to the rotation of the wrist from facing down to either left or right

For the phone gyro sensor data, jogging was what made the gyroscope sensor have the most variance in the graph which was as expected. Appendix M looks like an L, from this smaller colour groupings could be found along the L shape, this shows the potential for different categorical analysis but has even more overlap. Activities involving food have a low y-axis score due to not jumping up and down but a high z-axis due to the rotation of the hip to either left or right.

For the watch accel sensor data, jogging, dribbling and catch was what made the accelerometer sensor have the most variance in the graph which was as expected as these were the activities with the fastest movements. From this smaller colour groupings could be found along the graph, this shows the potential for different categorical analysis but has even more overlap. The lowest ranking accelerometer sensor readings are those where a participant is not accelerating much such as writing, standing, pasta and soup.

Next up, the phone accel sensor data, jogging, dribbling and catch was what made the accelerometer sensor have the most variance in the graph which was as expected as these were the activities with the fastest movements, primarily jogging due to the phone being in a pocket at the hip. From this smaller colour groupings could be found along the graph, this shows the potential for different categorical analysis but has overlap. The lowest ranking accelerometer sensor readings are those where a participant is not accelerating much such as writing, standing, pasta and soup.

PCA could have been attempted for the four data sets which could have given some insights into how to group the activities to be able to identify as many as possible with a high accuracy. This is as the number of categories overlaps a lot and does not allow for a clear analysis.

The final four graphs from Appendix P – S, was creating 3D scatter plots again, but each activity just had one plot on the 3D scatter plot, based off the standard deviation. This was to identify the frequency variance across activities and narrow down the categories. When looking at all four graphs a trend seems to emerge, both phone sensors appear to have four clusters sport, jogging, small movements, no movement. While the watch has six, jogging, hand sports, hand activities such as typing/writing, eating and hand swing movements such as walking, kicking and stairs.

Data Processing Solution

When thinking about what the data could be used to process data with, two ideas come to mind.

The first is that the data is gathered from different activities, as a few activities such as standing and running, pose such different results, then an artificial intelligence (machine learning algorithm) model could be created to help identify different movements. As these are different activities, they can be classified into different categories then a categorical machine learning algorithm could be used such as a decision tree or random forest algorithm. Before this a few activities seemed to have a similar variance. These activities are likely not going to be predicted accurately due to overlaps in variance and mean.

The second use would need further work in terms of collecting data, this would be as heart rates would need to be included. This would only be for users with watches and would work if the owner of the device was sat still for an hour presumably watching TV, of a certain age and their heart rate suddenly shot up with no wrist movement detected, then the device could automatically contact the healthcare services. This would also fall into a categorical Machine Learning Model as another category of Heart Attack would need to be added.

Further Work

Appendix T shows what Seb (2011) created, they used a sensor on their foot, instead of on a phone by their hip or a watch. What was created is called gait tracking and in the image the subject was walking up a spiral staircase onto the landing, as seen on the landing there are three steps as well. This could have been investigated further and an implementation attempted and then a deep dive could have been created to see if the gait tracking could be a form of labelled data as it would provide more accurate results. Gait tracking can also be great to identify disease such as Parkinson's and has been proven to accurately predict if a subject has it (Su and et al, 2021).

The ARFF files were not analysed due to the size and lack of understanding of each line. This could have been combatted by attempting a PCA on the data set. The RAW .txt files were chosen due to ease of understanding, interpretation and visualisation. When it comes to big data this is ideal to allow all stakeholders to easily comprehend what was done and why. With big data a lot could be analysed but without quick, meaningful impact to non-technical stakeholders which is not ideal. A 2D visualisation was not done for this reason, a lot of graphs would come up but without quick and interpretable impact, the 3D visualisation allowed for each of the four datasets to be summarised into one meaningful graph.

Bibliography

BIBI, K., 2022. *Seaborn HeatMap Colors* [online]. Available from: Seaborn HeatMap Colors (linuxhint.com) [Accessed 12 December 2022].

CLARK, A., 2013. How to convert epoch time with nanoseconds to human-readable?. *Stackoverflow.* [Blog online]. 27 March. Available from: https://stackoverflow.com/questions/15649942/how-to-convert-epoch-time-with-nanoseconds-to-human-readable [Accessed 12 December 2022].

DATACAMP., 2021. *Python Seaborn Cheat Sheet* [online]. Available from: Python Seaborn Cheat Sheet | DataCamp [Accessed 12 December 2022].

EPOCHCONVERTER., N/A. *Epoch & Unix Timestamp Conversion Tools* [online]. Available from: Epoch Converter - Unix Timestamp Converter [Accessed 11 December 2022].

GEEKSFORGEEKS., 2020. *Convert Text File to CSV using Python Pandas* [online]. Available from: https://pandas.pydata.org/docs/reference/api/pandas.crosstab.html [Accessed 6 December 2022].

GEEKSFORGEEKS., 2020. *3D Scatter Plotting in Python using Matplotlib* [online]. Available from: 3D Scatter Plotting in Python using Matplotlib - GeeksforGeeks [Accessed 11 December 2021].

KASHYAP, A., 2017. *Top 5 tricks to make plots look better* [online]. Available from: Top 5 tricks to make plots look better. | by Anirudh Kashyap | Medium [Accessed 14 December 2022].

KOOLAC., 2022. *3D Line Chart Plotting in Python using Matplotlib [online video].* Available from: https://www.youtube.com/watch?v=86sJTZTLXG0 [Accessed 9 December 2022].

LOULY, A., 2019., *Exploratory Data Analysis on WISDM* [online]. Available from: Exploratory Data Analysis on WISDM | Kaggle [Accessed 8 December 2022].

LOULY., 2019. *Jogging line graph* [online image]. Available from: Exploratory Data Analysis on WISDM | Kaggle [Accessed 8 December 2022].

LOULY., 2019. *Sitting line graph* [online image]. Available from: Exploratory Data Analysis on WISDM | Kaggle [Accessed 8 December 2022].

LOULY., 2019. *Walking heatmap* [online image]. Available from: Exploratory Data Analysis on WISDM | Kaggle [Accessed 8 December 2022].

PANDAS., N/A. *pandas.crosstab* [online]. Available from: https://pandas.pydata.org/docs/reference/api/pandas.crosstab.html [Accessed 9 December 2022].

SEBMADGWICKRESEARCH., 2011. *3D Tracking with IMU* [online video]. Available from: (692) 3D Tracking with IMU - YouTube [Accessed 15 December 2022].

STATISTICSCANADA., 2021. *4.5.3 Calculating the Variance and Standard Deviation* [online]. Available from: 4.5.3 Calculating the variance and standard deviation (statcan.gc.ca) [Accessed 10 December 2022].

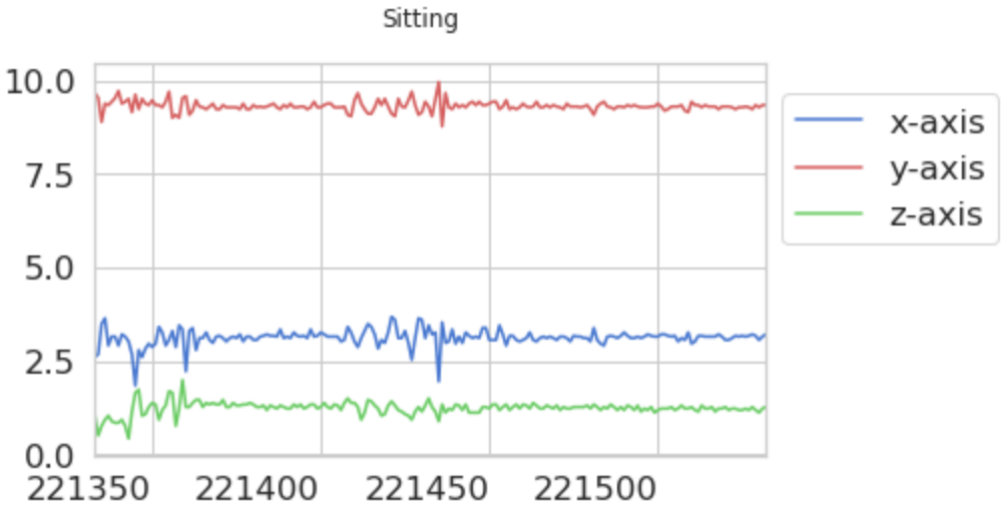
SU, D et al., 2021. *Simple Smartphone-Based Assessment of Gait Characteristics in Parkinson Disease: Validation Study* [online]. Available from: Simple Smartphone-Based Assessment of Gait Characteristics in Parkinson Disease: Validation Study - PMC (nih.gov) [Accessed 15 December 2022].

SYMMETRY ELECTRONICS., 2022. *Accelerometer vs Gyroscope – What's the Difference Between These Popular Sensors?* [online video]. Available from: https://www.youtube.com/watch?v=vFUlaRmuEHk [Accessed 08 December 2022].

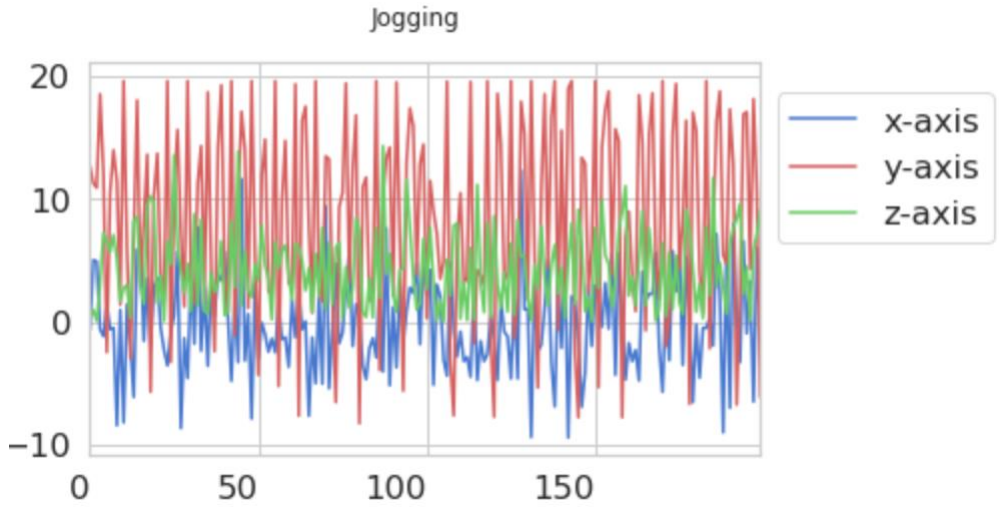
WU, W., et al., 2012. *Classification Accuracies of Physical Activities Using Smartphone Motion Sensors* [online]. Available from: Journal of Medical Internet Research - Classification Accuracies of Physical Activities Using Smartphone Motion Sensors (jmir.org) [Accessed 9 December 2022].

Appendices

Appendix A

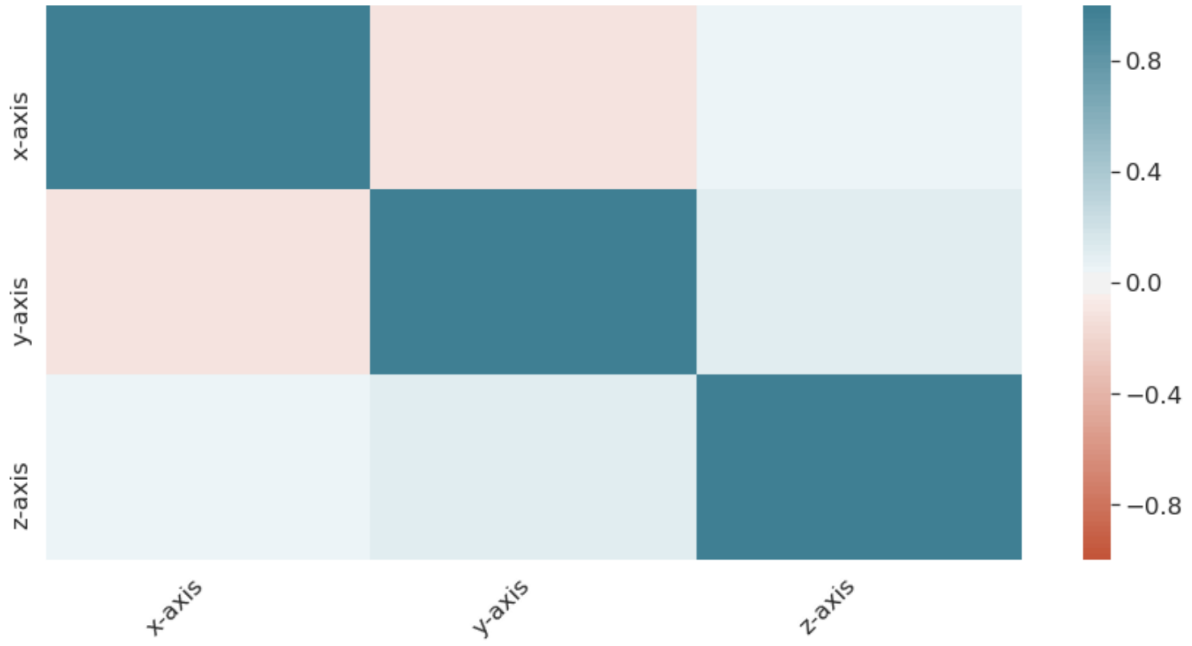


Appendix B



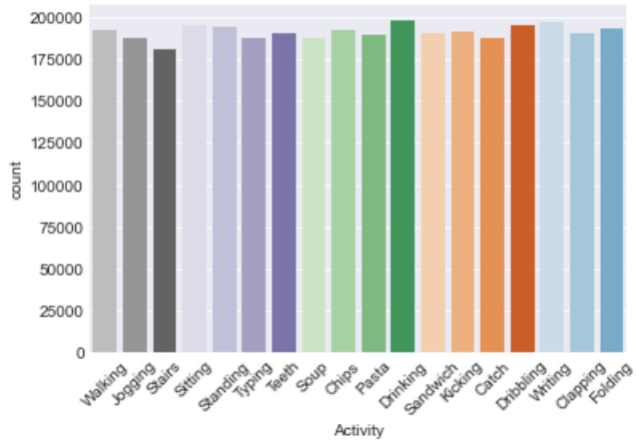
Appendix C

```
plot_corr("Walking",df)
```



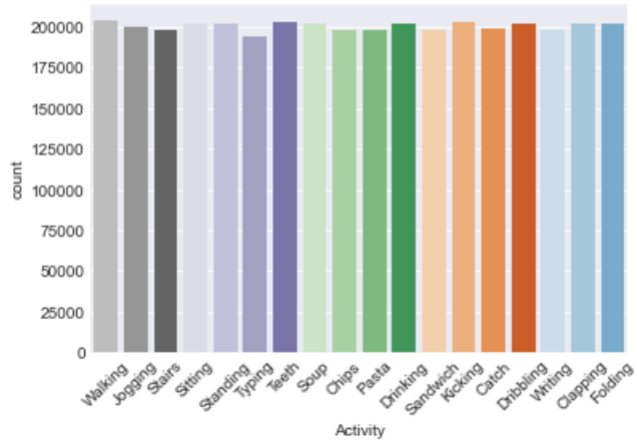
Appendix D

```
Watch Gyro Activity Count:
 AxesSubplot(0.125,0.125;0.775x0.755)
```



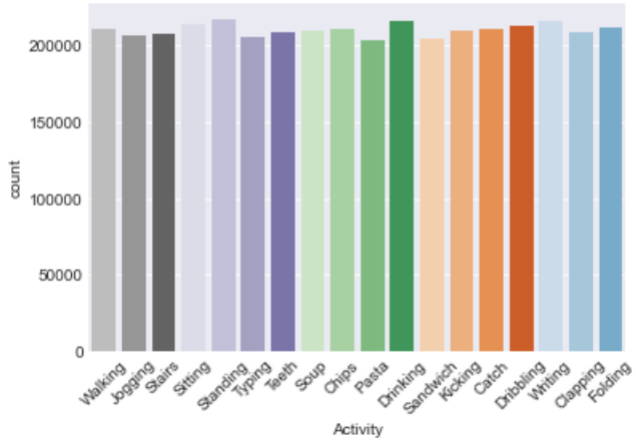
Appendix E

```
Phone Gyro Activity Count:
 AxesSubplot(0.125,0.125;0.775x0.755)
```



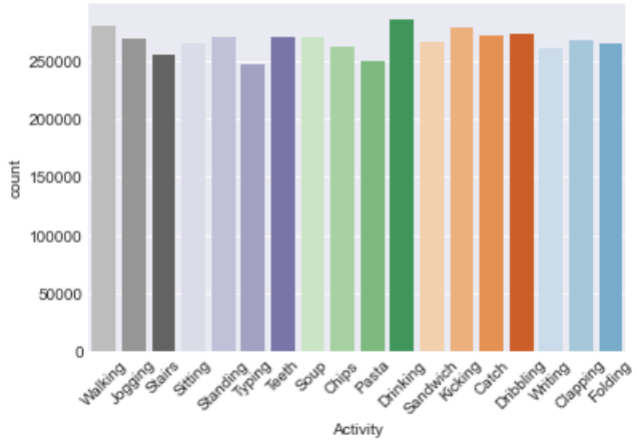
Appendix F

```
Watch Accel Activity Count:
 AxesSubplot(0.125,0.125;0.775x0.755)
```



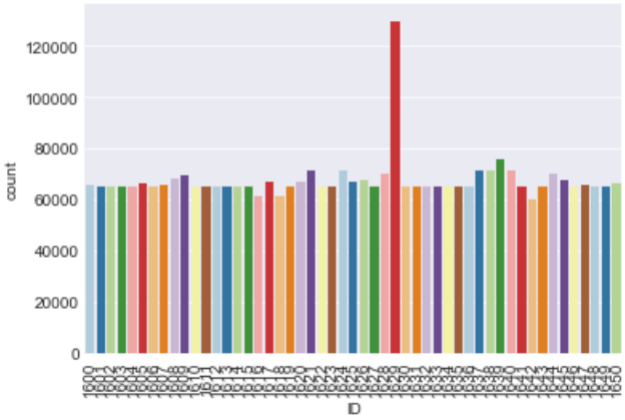
Appendix G

```
Phone Accel Activity Count:
 AxesSubplot(0.125,0.125;0.775x0.755)
```



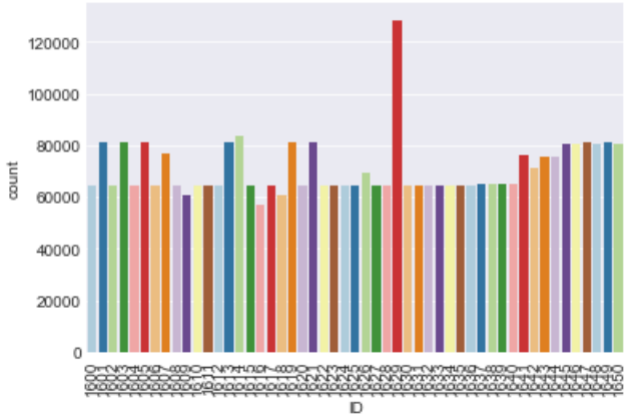
Appendix H

```
Watch Gyro Participant Count:
 AxesSubplot(0.125,0.125;0.775x0.755)
```



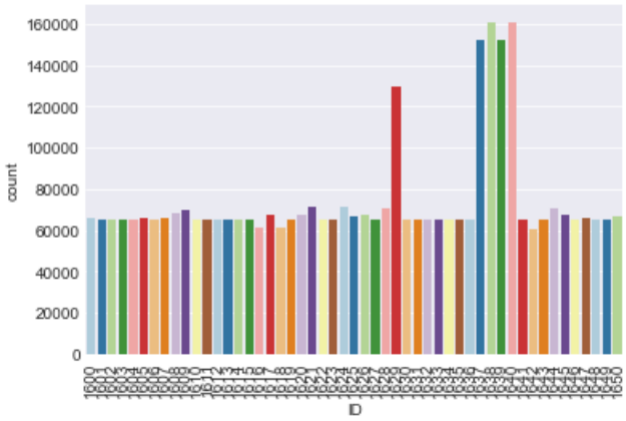
Appendix I

```
Phone Gyro Participant Count:
 AxesSubplot(0.125,0.125;0.775x0.755)
```



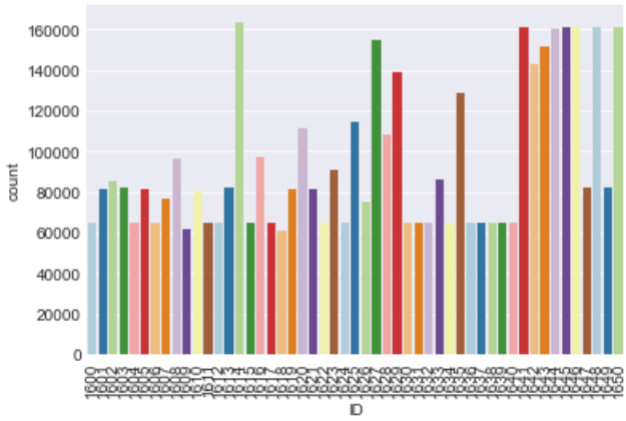
Appendix J

```
Watch Accel Participant Count:
 AxesSubplot(0.125,0.125;0.775x0.755)
```

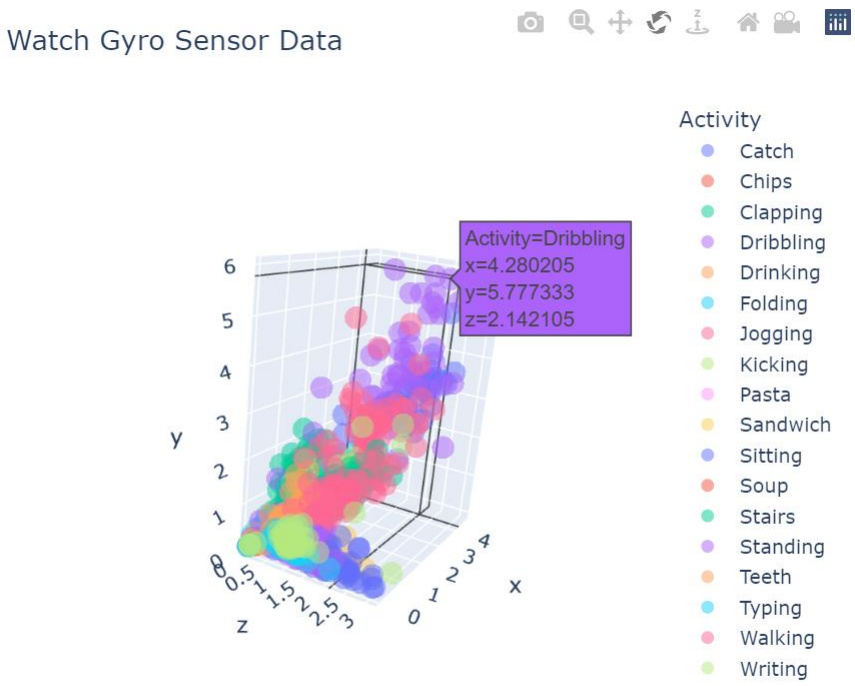


Appendix K

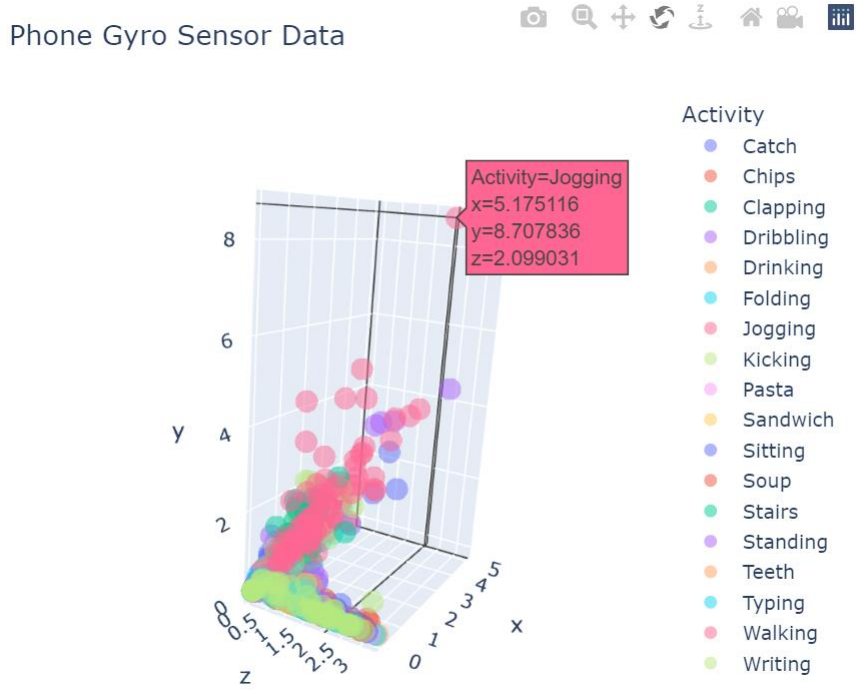
```
Phone Accel Participant Count:
 AxesSubplot(0.125,0.125;0.775x0.755)
```



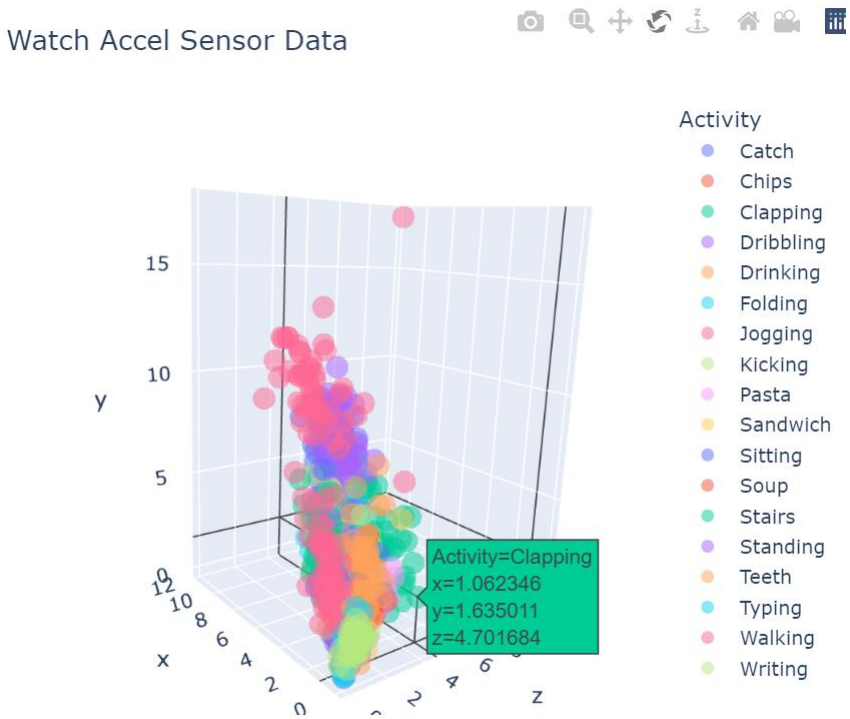
Appendix L



Appendix M

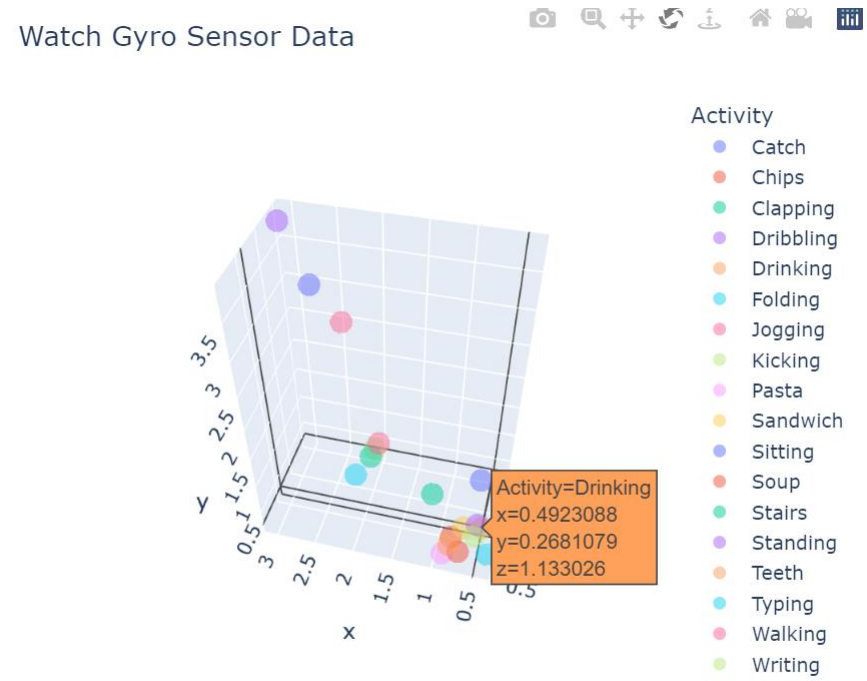


Appendix N

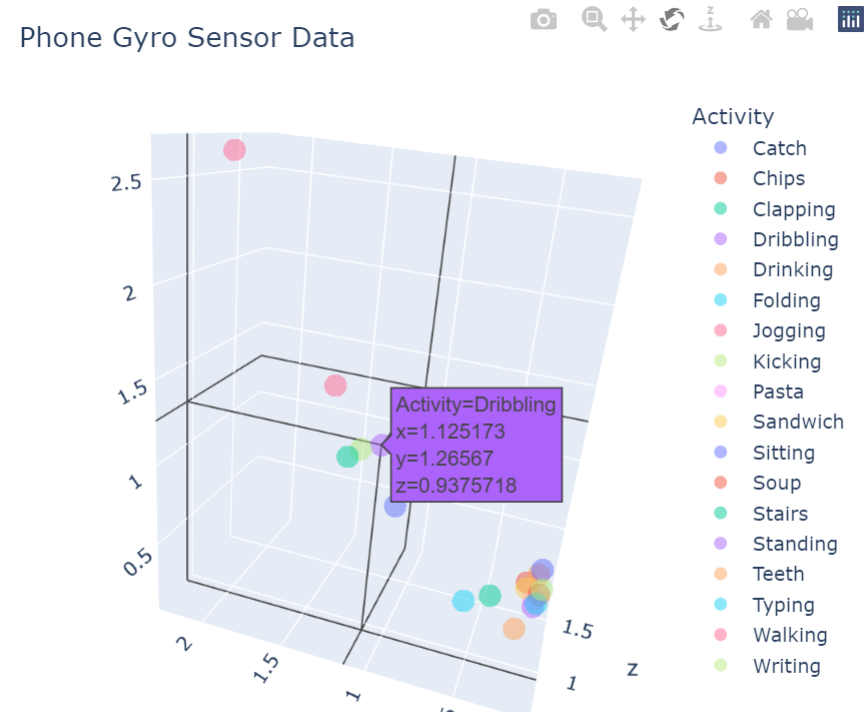


Appendix O

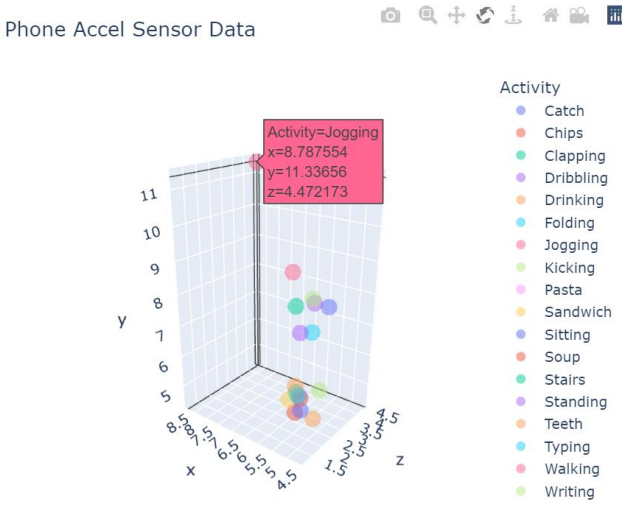
Appendix P



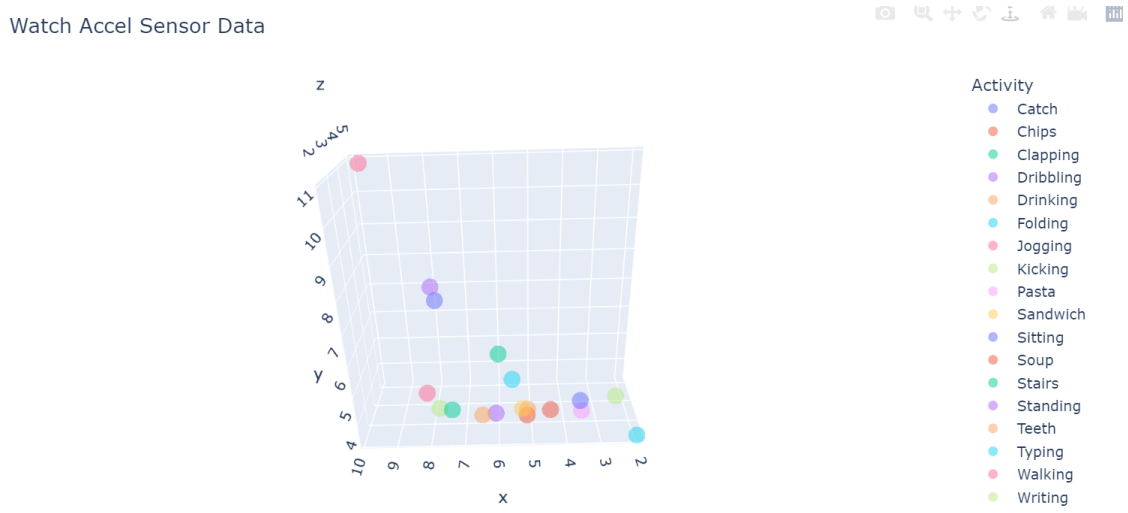
Appendix Q



Appendix R


Phone Accel Sensor Data
Activity=Jogging
x=8.787554
y=11.33656
z=4.472173
Activity
Catch
Chips
Clapping
Dribbling
Drinking
Folding
Jogging
Kicking
Pasta
Sandwich
Sitting
Soup
Stairs
Standing
Teeth
Typing
Walking
Writing

Appendix S


Watch Accel Sensor Data
Activity
Catch
Chips
Clapping
Dribbling
Drinking
Folding
Jogging
Kicking
Pasta
Sandwich
Sitting
Soup
Stairs
Standing
Teeth
Typing
Walking
Writing

Appendix T