REPORT

DATA VISUALIZATION

Submitted By

Roya Dehghani

Contents

Q1) Bar Crawl: Detecting Heavy Drinking Data Set

Description:

For 13 individuals, we have acceleration along the x, y, and z-axis. We also have the TAC readings for each of them, which show the alcohol level throughout the study. Based on the acceleration data, this task determines whether the subject is sober or drunk. TAC readings would be utilized to categorize the data as sober or drunk. TAC reading is continuous; nevertheless, changing the problem to categorization is more convenient. As a result, a blood alcohol level of more than 0.08 g/dl is deemed drunk, whereas a blood alcohol level of less than 0.08 g/dl is considered sober. (It should be noted that the timestamp in the acceleration dataset is calculated by millisecond. However, the timestamp in TAC is based on second. The other note is about the duration time; for acceleration, we have data for 24 hours, but for TAC reading, we have data for 8 hours).

Several statistical numbers/information:

Before processing the dataset, it is necessary to gain insight into the data by statistical numbers, including mean, median, standard deviation, etc.

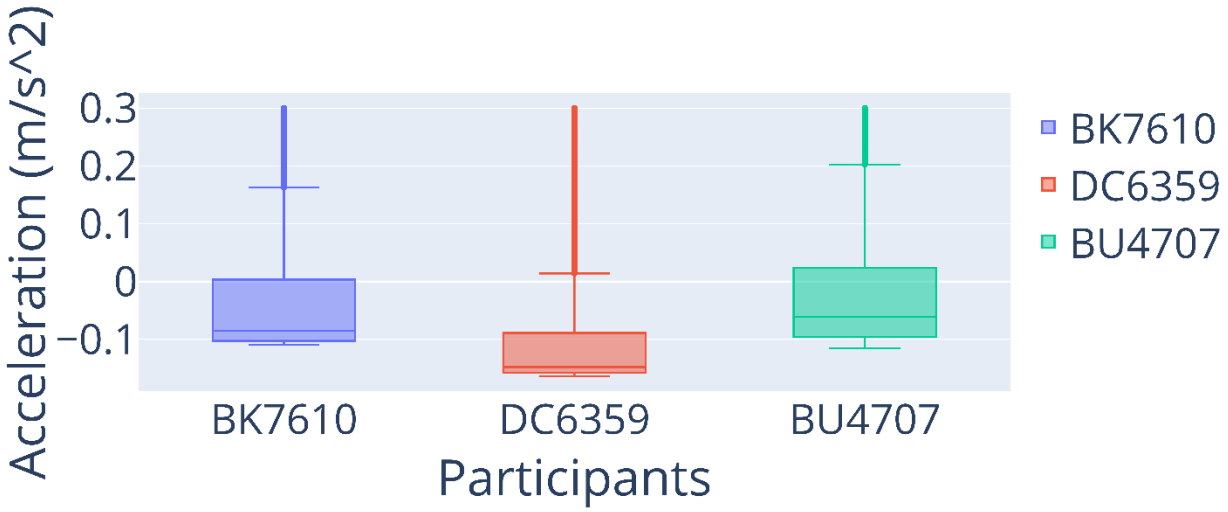


Figure 1: Distribution of normalized acceleration data

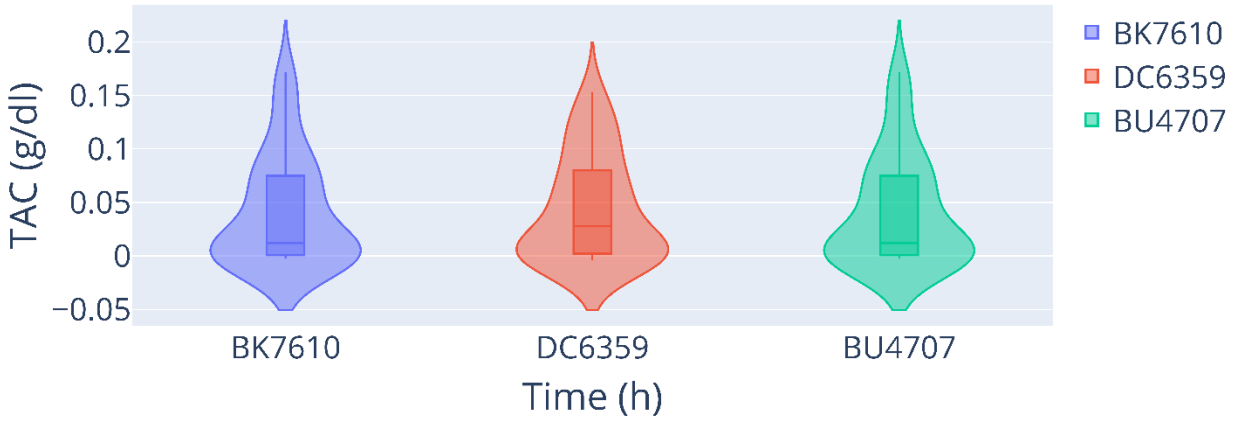


Figure 2:Distribution of TAC

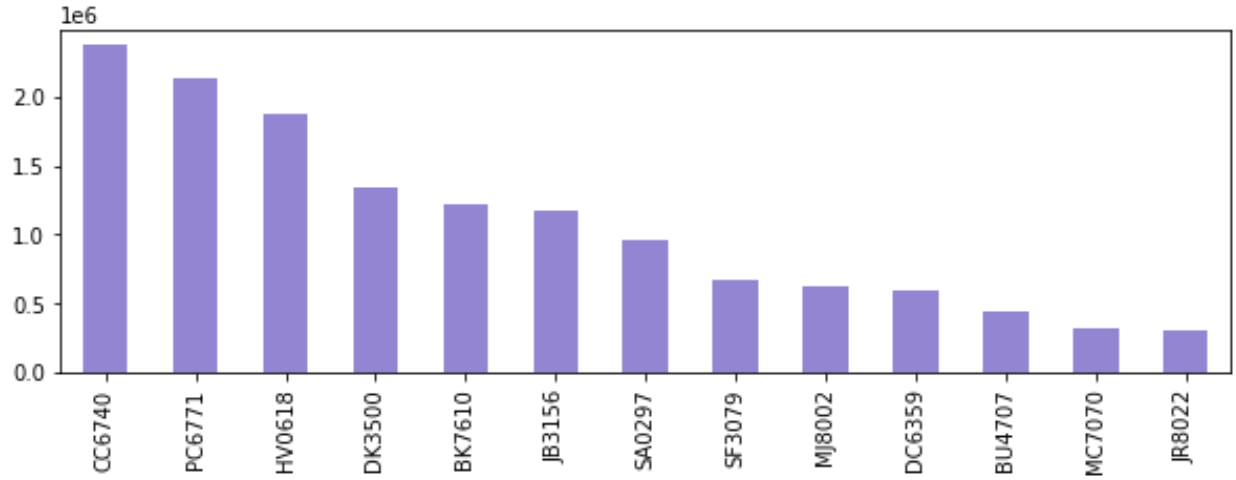


Figure 3:Number of data samples collected for all participants

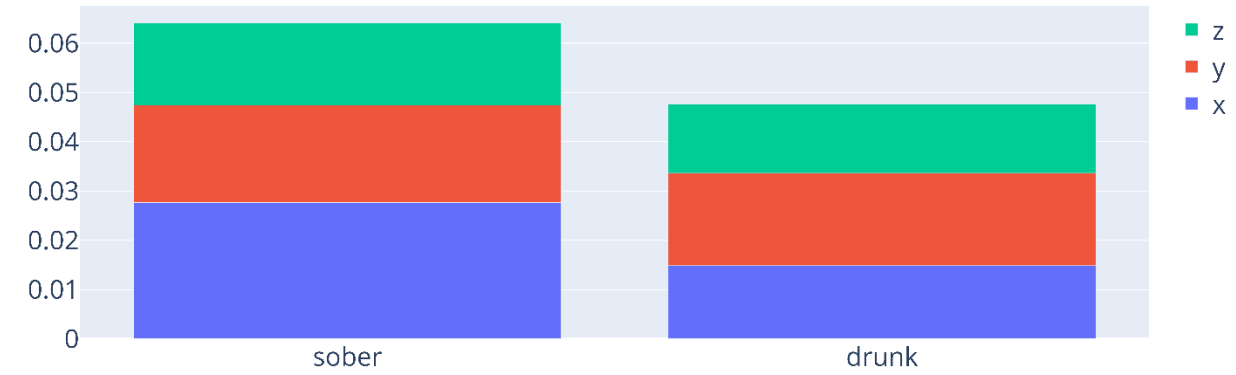


Figure 4:Stacked bar chart to show energy between sober and drunk mode

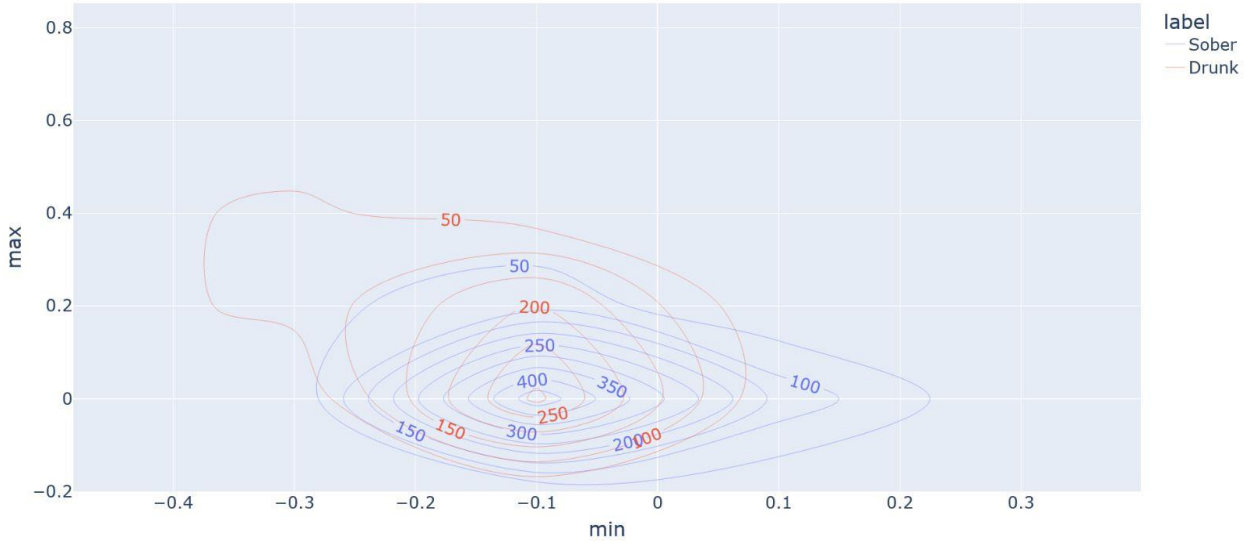


Figure 5:2D Histogram to show the distribution of min and max normalized acceleration

Visualization method(s):

We used a **Box and Whisker plot**, **violin**, and **histogram** to show the distribution of normalized acceleration, TAC level, and minimum and maximum normalized acceleration, respectively. We used a **bar chart** to compare the number of data samples for each user. Furthermore, we used a **stacked bar chart** to compare the amount of energy in sober and drunk mode on the x-axis, y-axis, and z-axis.

Hidden pattern(s), relationships, and/or associations:

Figure 1 indicates the mean, median, standard deviation, minimum, upper fence, q1, and q3 normalized acceleration for three users/participants. We normalize the acceleration values in the direction of the x-axis, y-axis, and z-axis by the following formula (1). Figure 1 shows that while the order of magnitude is the same, each participant's acceleration has a slightly distinct distribution. In other words, the acceleration range is approximately similar for all users/participants. However, for user DC6359, it is a little lower.

$$b(t) = \sqrt{x_t^2 + y_t^2 + z_t^2} - \sum_{t=1}^{n} \frac{\sqrt{x_t^2 + y_t^2 + z_t^2}}{n} \quad (1)$$

Figure 2: we first read the CSV file for each user, then converted the timestamp unit from millisecond to second. Next, we set start time zero, put it as origin, and calculate other times based on that. The graph indicates that most of the time in the targeted duration, the participants are sober because the density of the graph is around zero (the chart is wider around zero). The Median (Middle line in box plot) for user DC6359 is higher than the other two users. Also, most values are concentrated on the Q1 value (Wider around Q1).

Figure 3: It shows the number of data samples for each participant. As it is shown, more data samples were collected for users CC6740.

Figure 4: The goal is to find features by which we can classify new participants into sober and drunk. Energy is the first feature to be extracted. We calculate the energy of the normalized acceleration in each direction for each 4(s) intervals. We now show the Energy Stacked bar chart for drunk and sober samples. The graph shows that the user has more energy when sober than in drunk mode. However, this feature cannot be a sufficient criterion to classify because the bars do not have a bit different.

Figure 5: More features are required to classify accurately. We calculated the minimum and maximum normalized acceleration every four-interval time during 1 hour of the user's sobriety and 1 hour of being drunk. The 2D histogram shows the distribution of min and max features. It indicates that these two features cannot be suitable for classification because they overlap. However, the min and max of being drunk are a little higher.

Conclusion:

This report used a box plot and violin to show the data distribution. After gaining an insight into what the data is and how they are distributed, we tried

to find some features to see based on which we can predict users into sober and drunk or not. We extracted the energy feature to compare users at every level when they are heavy/drunk and intoxicated/sober by the stacked bar chart. However, we found that this feature cannot be sufficient to decide on users' sober or drunk mode. Then, we extracted other features, including minimum and maximum energy in four-time intervals, in a 2D histogram. These features overlap and are not good enough for the classification task. Finally, extracting characteristics from input data and evaluating their distribution simultaneously can aid classification. This report's graphic shows that by extracting enough information, we can readily forecast a person's state based on his acceleration data.

Q2) Wisconsin Breast Cancer Dataset

Description:

This breast cancer dataset from the University of Wisconsin Hospitals, Madison, from Dr. William H. Wolberg. Features are computed from a digitized image of a breast mass's fine needle aspirate (FNA). They describe the characteristics of the cell nuclei present in the picture. There are 30 features and 569 samples.

Attribute information:

- ID number
- Diagnosis (M = malignant, B = benign)
- 3-32: Ten real-valued features are computed for each cell nucleus:
    - a) radius (mean of distances from the center to points on the perimeter)
    - b) texture (standard deviation of gray-scale values)
    - c) perimeter
    - d) area
    - e) smoothness (local variation in radius lengths)
    - f) compactness (perimeter^2 / area - 1.0)
    - g) concavity (severity of concave portions of the contour)
    - h) concave points (number of concave portions of the contour)
    - i) symmetry
    - j) fractal dimension ("coastline approximation" -1)

- The mean, standard error, and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, and field 23 is Worst Radius.

Several statistical numbers/information:

First of all, we have a look at the data that we have in hand. In this regard, we used a violin plot to see how features are distributed.

Figure 6 is a **violin plot** to indicate feature value distribution by target/diagnosis. It shows that the median of the malignant and benign populations appears to be separated in the texture mean feature (the second feature from the left side of the graph), making it useful for classification. However, the malignant and benign median does not appear to be separated in the fractal dimension representing the feature, rendering it useless for classification.

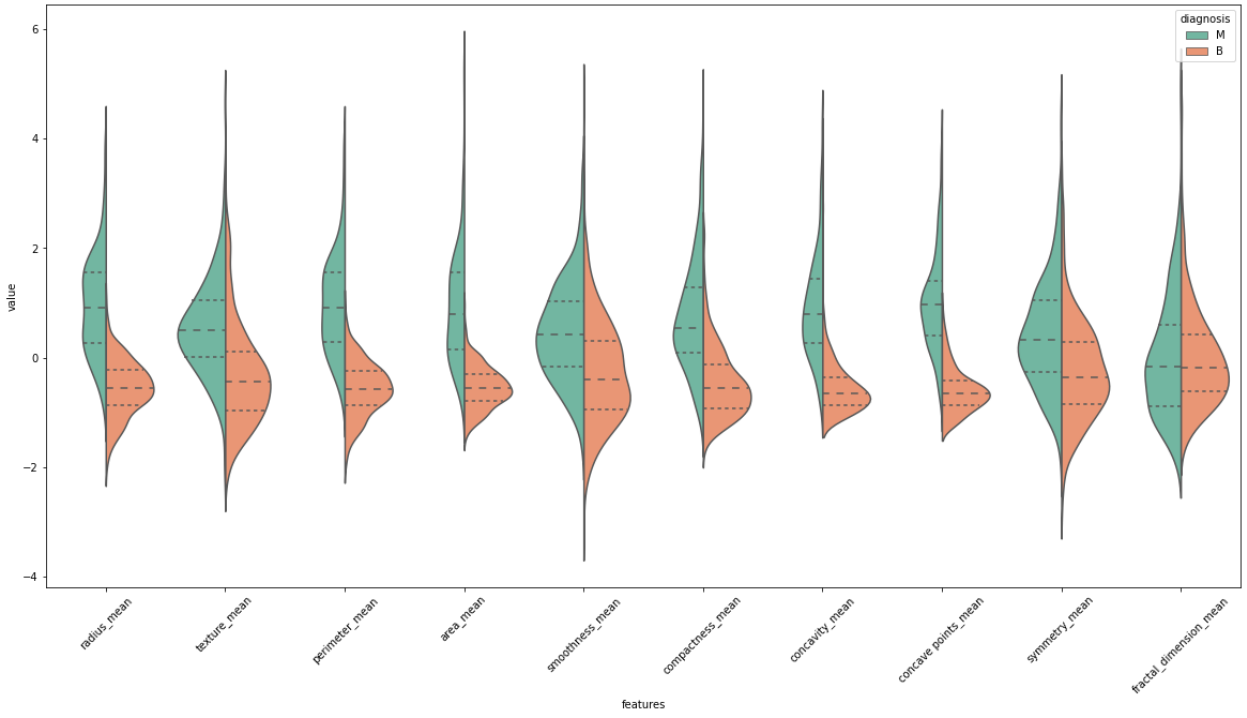


Figure 6:Mean features distribution by diagnosis

Figure 7 and Figure 8 show distributions of standard error and “worst” features. Also, the violin can show colinear relationships. We can use a heatmap to determine whether Concavity worst and Concave points worst

are connected or unrelated. If they are related, the best strategy is to eliminate the redundancy by removing one of the columns.

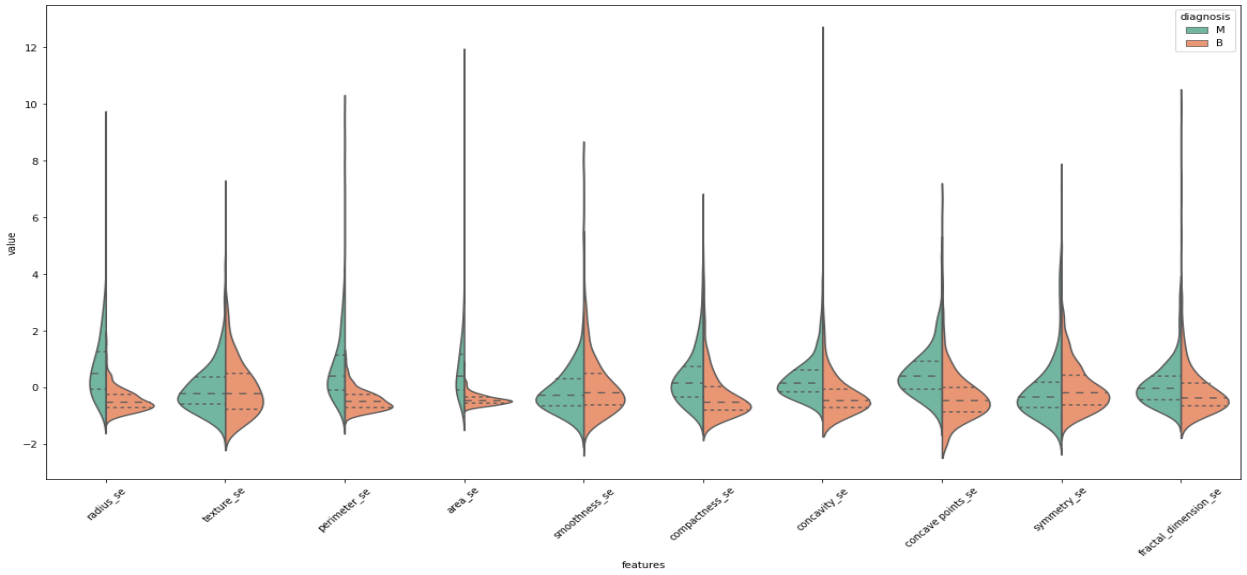


Figure 7:Standard error features distribution by diagnosis.

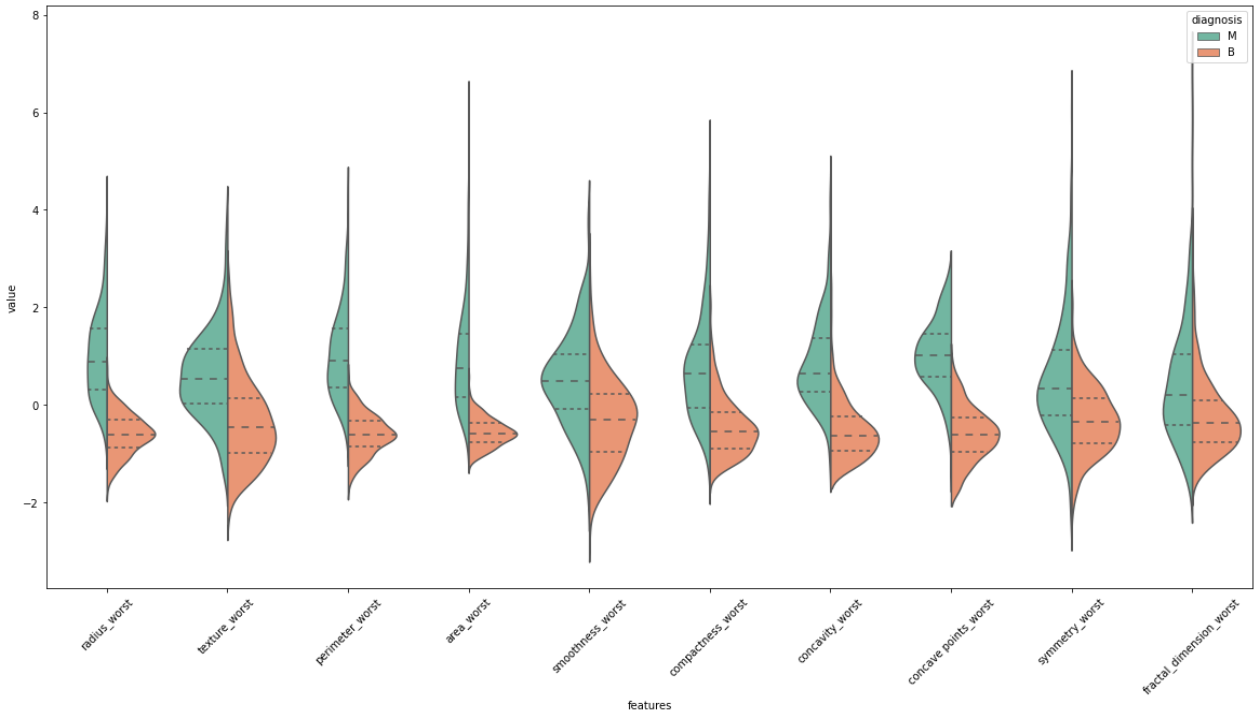


Figure 8: "Worst" or the largest features distribution by diagnosis

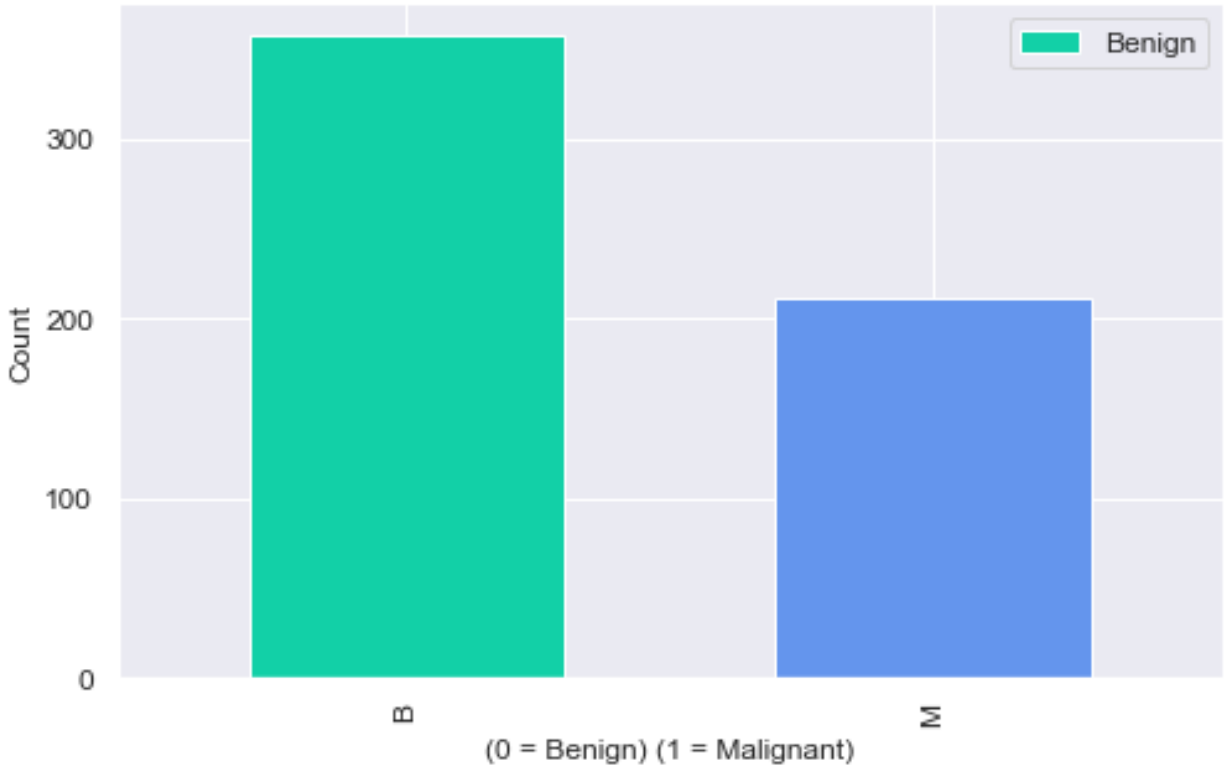


Figure 9:Number of Benign and Malignant (target)

Visualization method(s):

We used a **heatmap** to show a correlation between features and targets and co-linear relationships. The value of each cell indicates the amount of correlation. Moreover, Boxplot is used to display data distribution, see overlap between benign and malignant values in one feature, and see how good the feature is in classification. Also, we used a **pair plot** to see the distribution of each feature based on the target. Furthermore, we used a **scatter plot** to see how good when we use two features for classification.

Hidden pattern(s), relationships, and/or associations:

Figure 9: This graph shows whether our dataset is balanced or not. We have 357 malignant and 212 benign cases, so our dataset is imbalanced.

(the machine learning Classifier tends to be more biased towards the majority class, causing bad classification of the minority class.)

Figure 11: **Heatmap** is an effective tool to visualize correlation among the features. A positive correlation means that both variables move in the same direction, while a negative correlation means that when one variable's value rises, the other variable's value falls. The correlation might also be zero, indicating that the variables are unconnected. Light color and deep color show positive and negative co-linear relationships between features. Good features for classification are those with middle color.

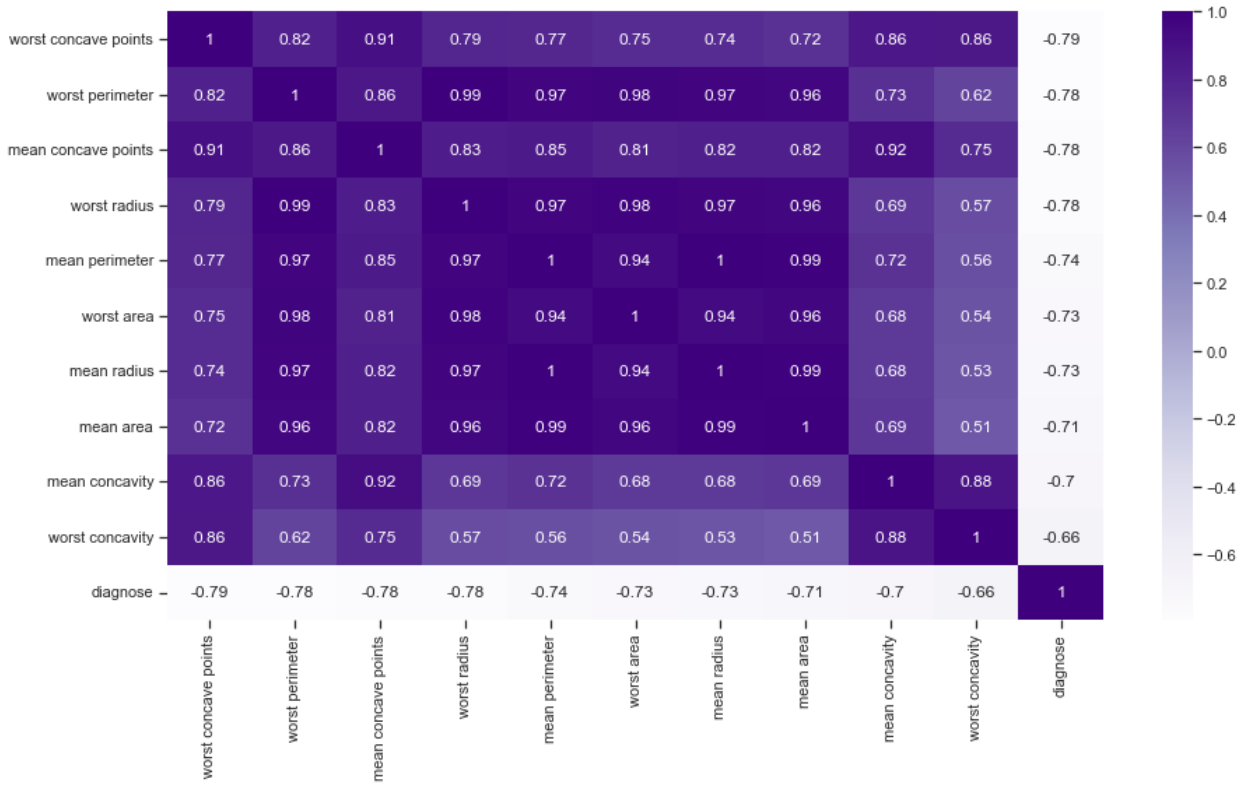


Figure 10:Only highly correlated features

Figure 10: It represents only highly correlated features. We may readily determine multicollinearity by looking at this graph. As a result, certain collinear features have been removed to have a more accurate classification task.

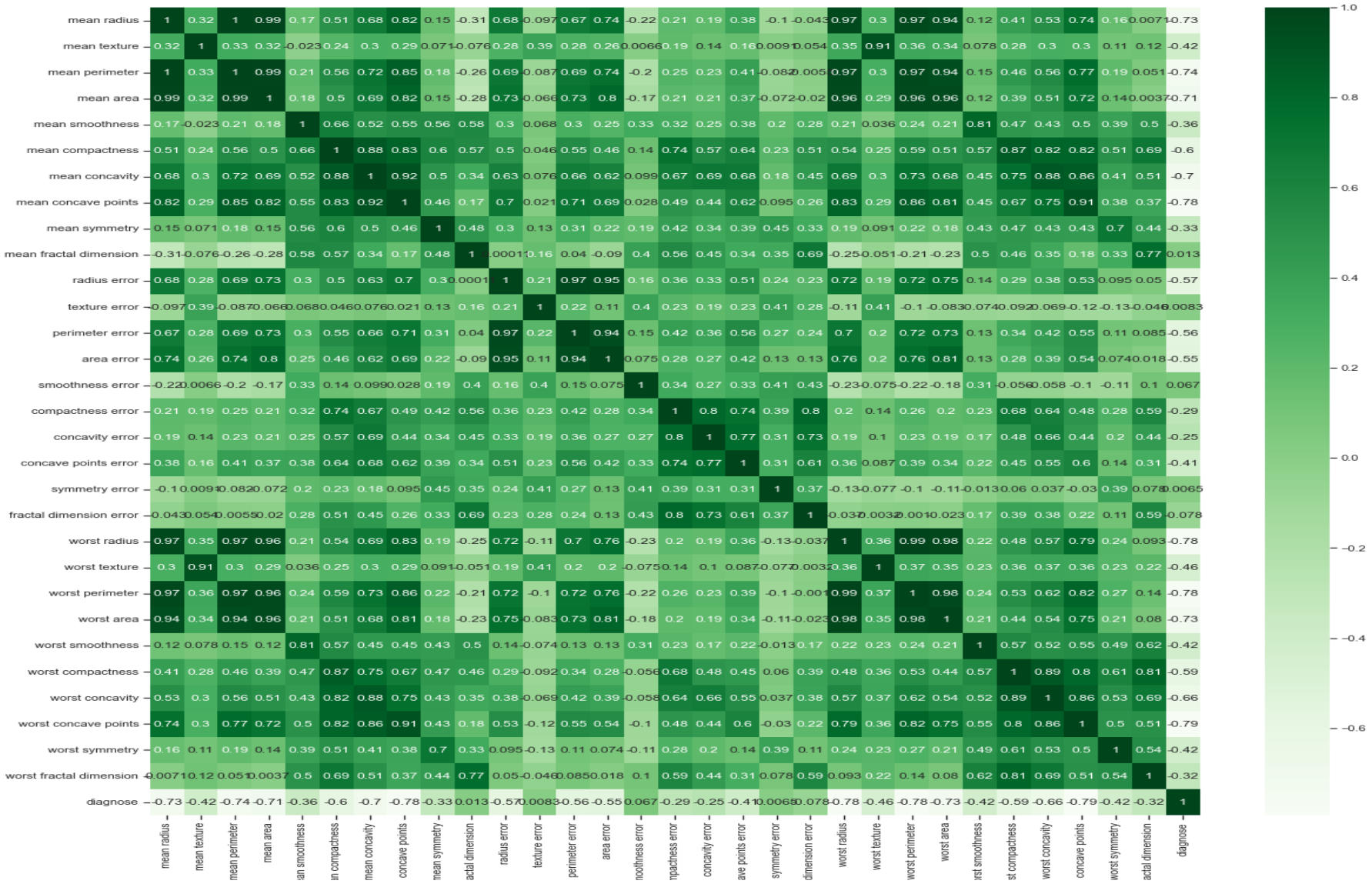


Figure 11:Correleration between target values and features

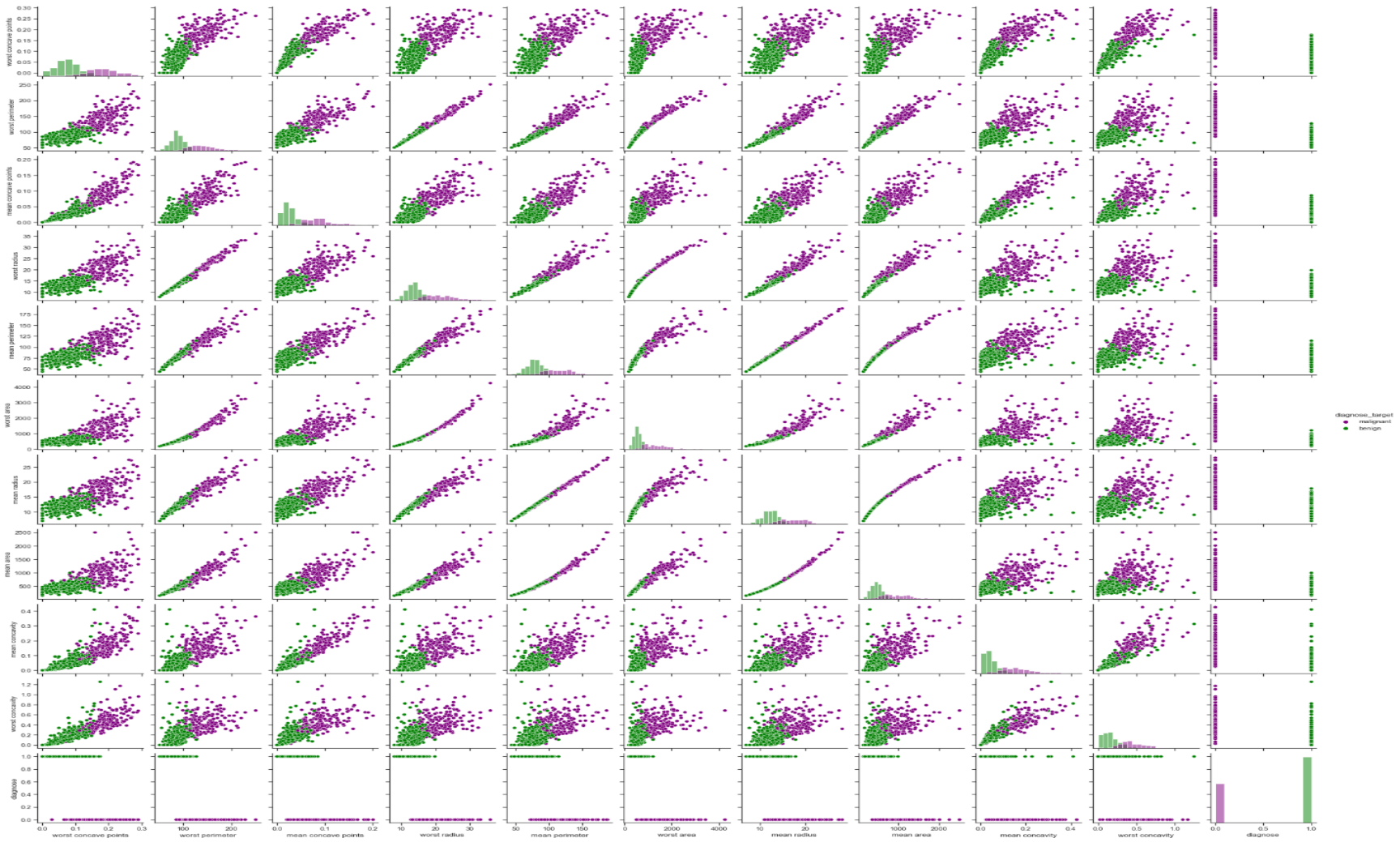


Figure 12:Distribution of features' values by targets

Figure12: **Paitplot** shows relationship between features. Moreover, It represents Lower levels of features that indicate a normal/benign cell, while greater values indicate a malignant tumor.

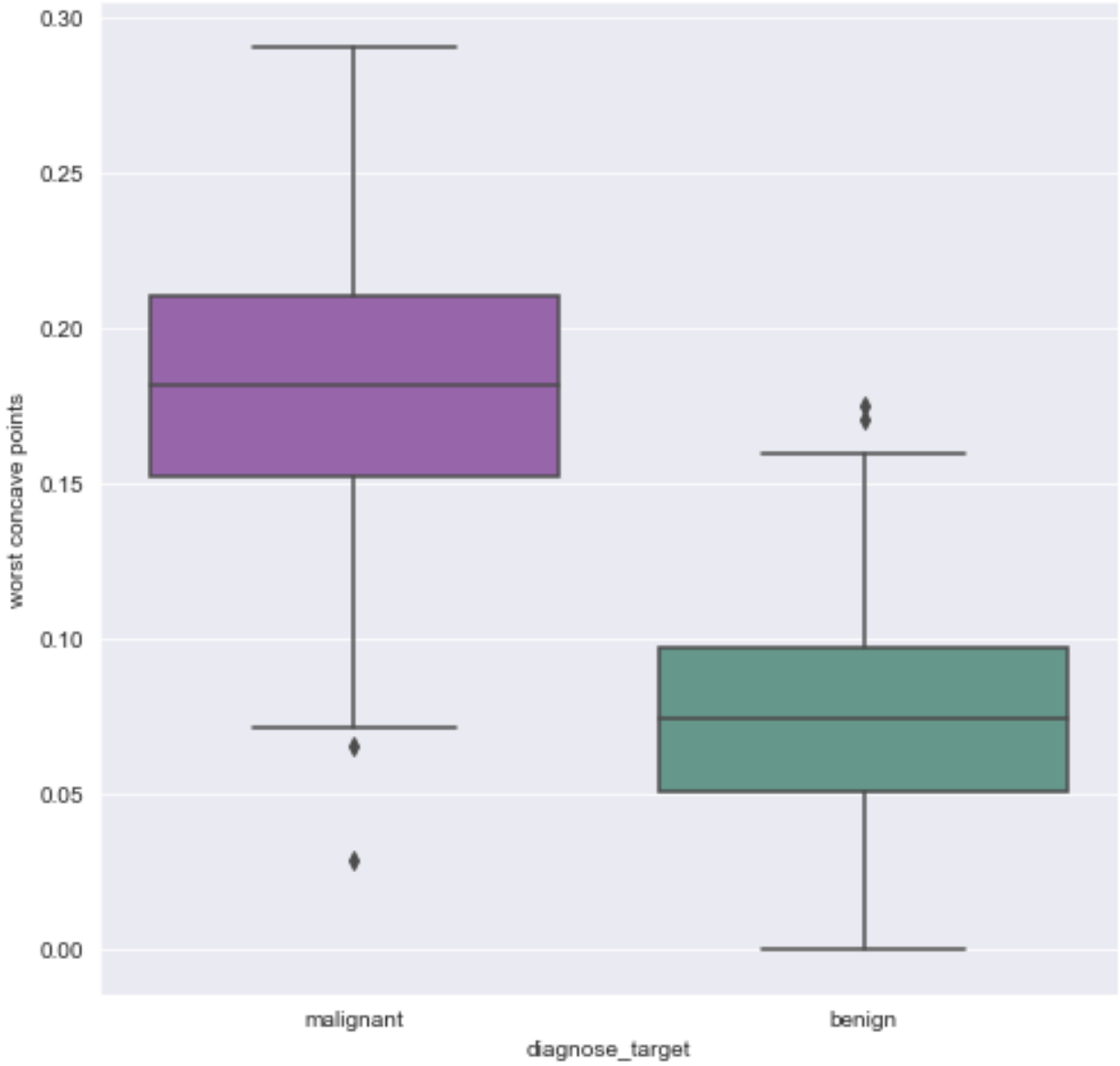


Figure 13: Distribution of Worst Concave Points feature

Figure 13: We used a **boxplot** to show the distribution of values in one feature. If the values representing benign do not overlap those indicating malignant, we can predict the new cell based on this feature. In other words, this feature can be a good feature to classify the new cell into benign or malignant. However, It is insufficient, and we should use other features for more accurate classification.

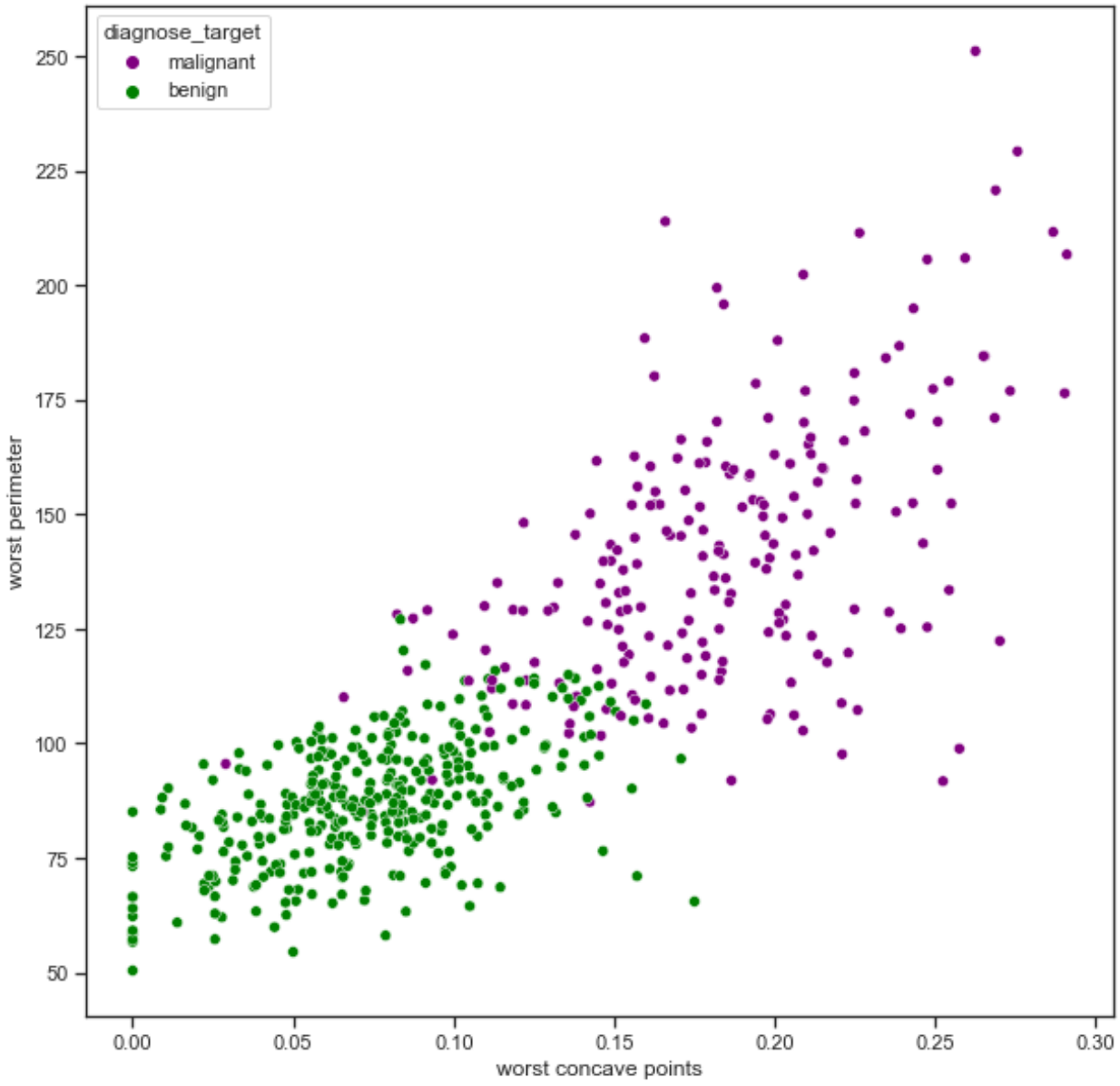


Figure 14 Relationship between two features

Figure 14: The goal is to find features by which we can predict whether a cell is benign or malignant. For this purpose, we used a **scatter plot** representing the relationship between Worst Concave and Worst Perimeter. Looking at this graph, we can find that these two features can help us predict a cell into benign or malignant. This is because the boundary between benign and malignant is about clear.

Conclusion:

The goal is to find sufficient features to classify new values into benign or malignant. First, we used a violin plot to show our data distribution. Those features with the similar violin are likely to be correlated. Then, we employ a heatmap to decide on the correlation between features. It shows both positive and negative relationships between features. Those highly correlated features are demonstrated in deeper and lighter colors, so we used middle correlated features for classification. Next, thanks to the pair plot, we found that lower values in features can represent benign cells, but the difference between values is not the same in all features. Therefore, we need some features by which we can predict more accurately. In this regard, we examine the overlap between values distribution in benign and malignant in one feature. We can decide on this good feature to predict new value if there is no overlap. However, there is overlap (shown in figure14), so we need more features to predict more accurately. We used a scatter plot to show the relationship between two features, and the combination of these can classify benign and malignant to a great extent (Figure15). In conclusion, this dataset is not imbalanced and has highly correlated features that should be removed.

Q3) Human Activity Recognition Using Smartphones Dataset

Description:

The dataset consists of an experiment that has been carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, they captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The experiments have been video-recorded to label the data manually. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.

We first investigate the labels file which shows the corresponding activity for each data instance. We show in the bar chart below the number of occurrences for each activity in the dataset to confirm whether the dataset is balanced or not so we can know if the given dataset is good enough to train a model to differentiate between these activities without being biased.

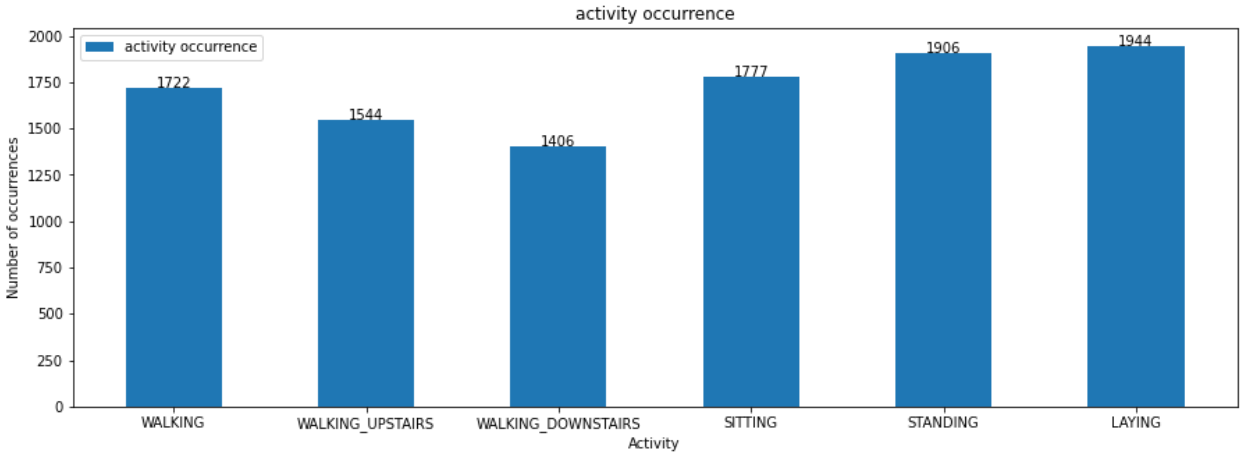


Figure 3.1: Activity occurrence bar chart

The above graph shows that the activities are balanced fairly and the dataset can be indeed used to train a model to differentiate between them.

The next step in evaluating the dataset was to see the correlation between the features and the labels in the dataset. The main features that were gathered in the dataset were the body acceleration and gyroscope measurements in the three coordinates and merging them to get the total acceleration of the body in one instance. In the line graph below we show the change of total acceleration, body acceleration, and gyro measurements in the three axes with time along with the corresponding activity. All the measurements belong to the first volunteer in the dataset, the dataset contains the measurements of 30 volunteers and the change in the line graph is negligible when we try to use different volunteers like all of them belong to the same age bracket. From the line graph, we can find that regions with high change in the acceleration values correspond to activities that require some movement, and regions with low acceleration values correspond to activities that require minimal movement or no movement at all.

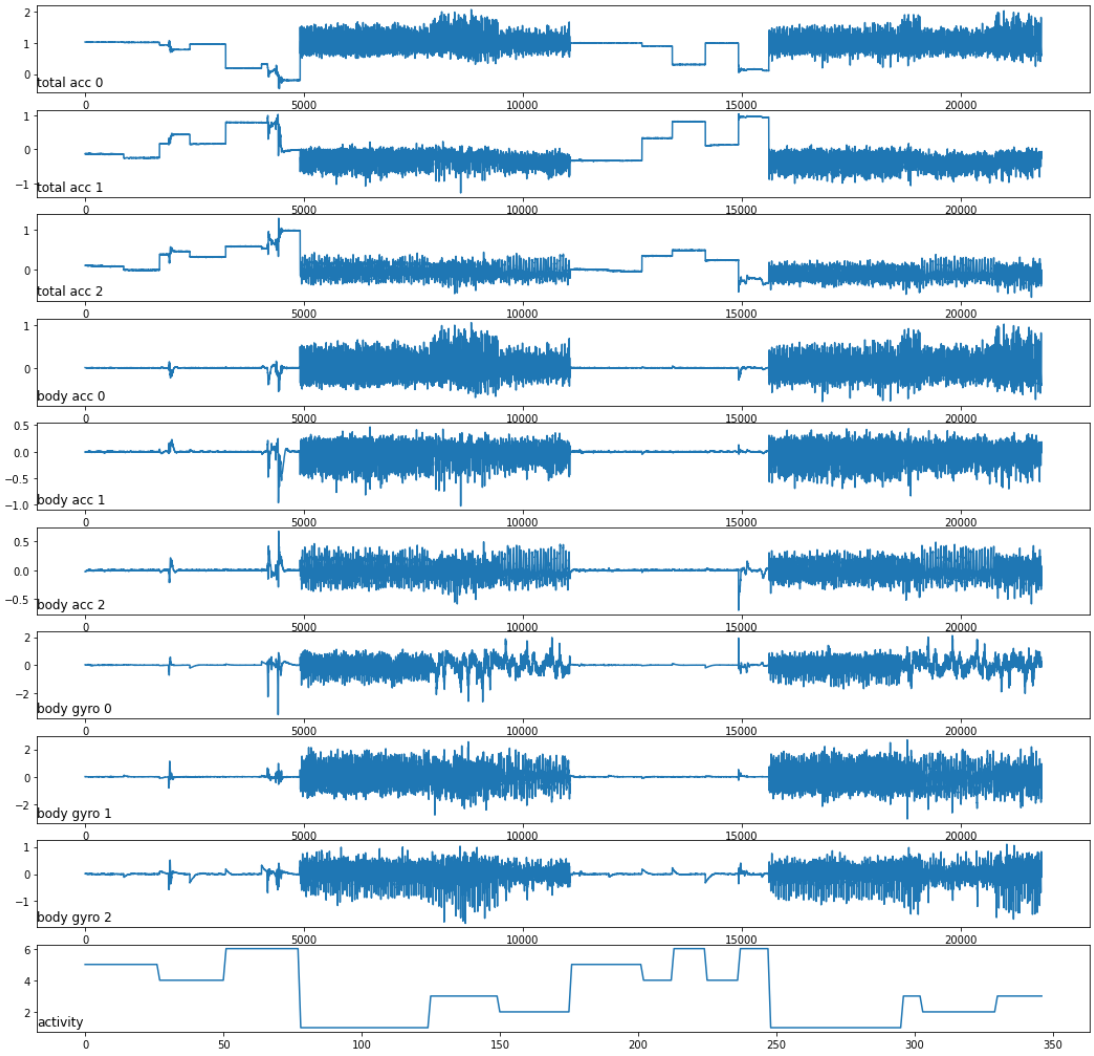


Figure 3.2: Line graph of the acceleration and the corresponding activity for a given test subject

The histogram of the total acceleration of five volunteers is shown in the graph below to test whether the volunteers have similar acceleration distribution or not. We expect the subjects to have similar acceleration distribution as all of them are doing the same activities and this will help us generalize this acceleration distribution to find the activity of new subjects after training.

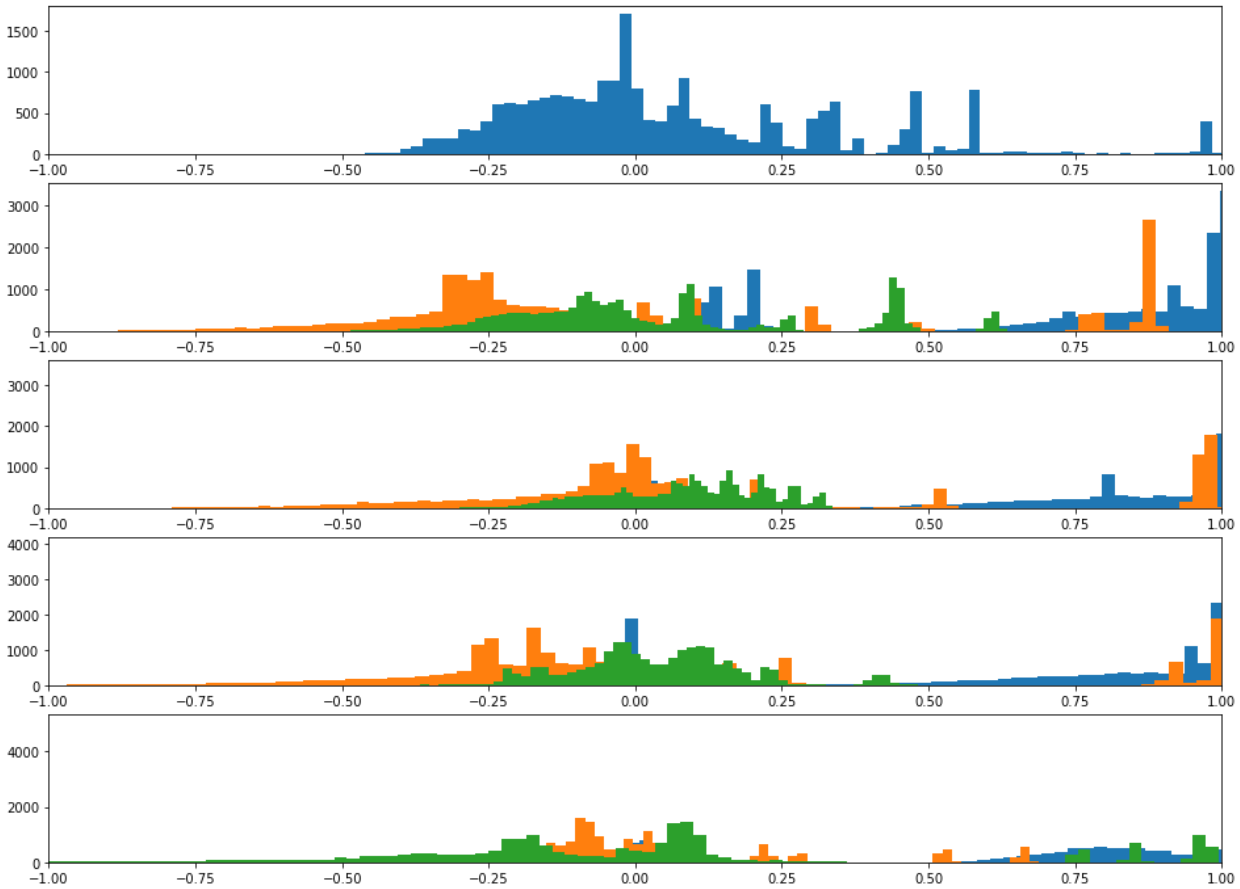


Figure 3.3: Histogram of the total acceleration for five subjects

The histogram of the total acceleration in X, Y, and Z directions are shown in blue, orange, and green respectively. Results show that the volunteers have similar acceleration distribution.

Finally, we test whether this data truly differentiates between the different activities with the given features. The main 6 attributes that were collected from volunteers (the body acceleration and the gyro measurements in the 3 coordinates) were used to make 561 features that were used to train the model to predict the activity of a given subject. We used the tsne algorithm to perform dimensionality reduction to the 561 features into 2 features only in order to be able to visualize them. Keep in mind that the tsne algorithm preserves the clustering property of the data points while reducing the dimensionality by preserving the similarity matrix between the original data

domain and the transformed domain. The results of the tsne model were visualized using a scatter plot as shown below.

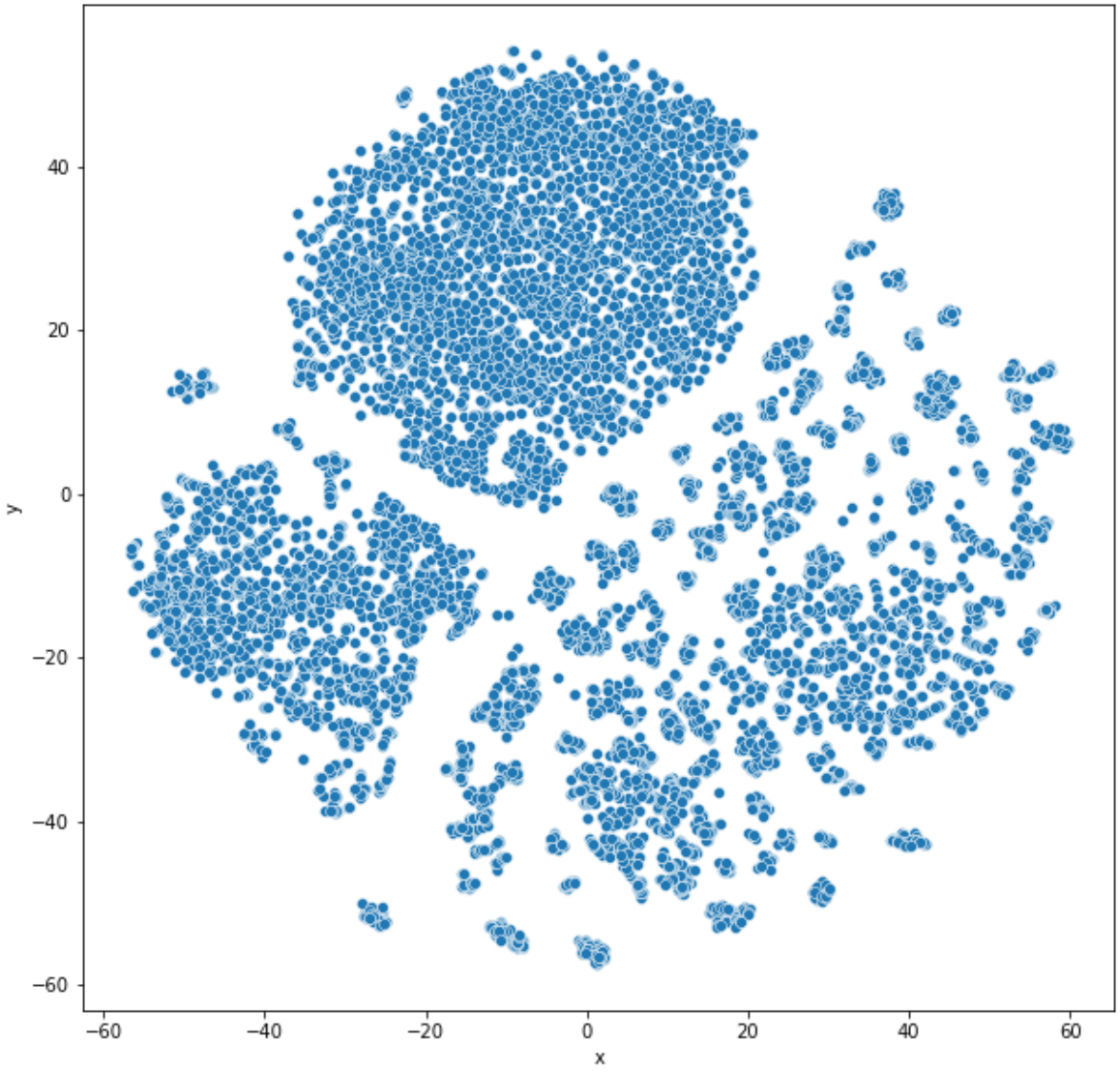


Figure 3.4: Tsne model output of the data features

The results show that the features form 5 main clusters although we have 6 labels in the dataset adding the labels of each instance as the hue of the point gives us the graph below

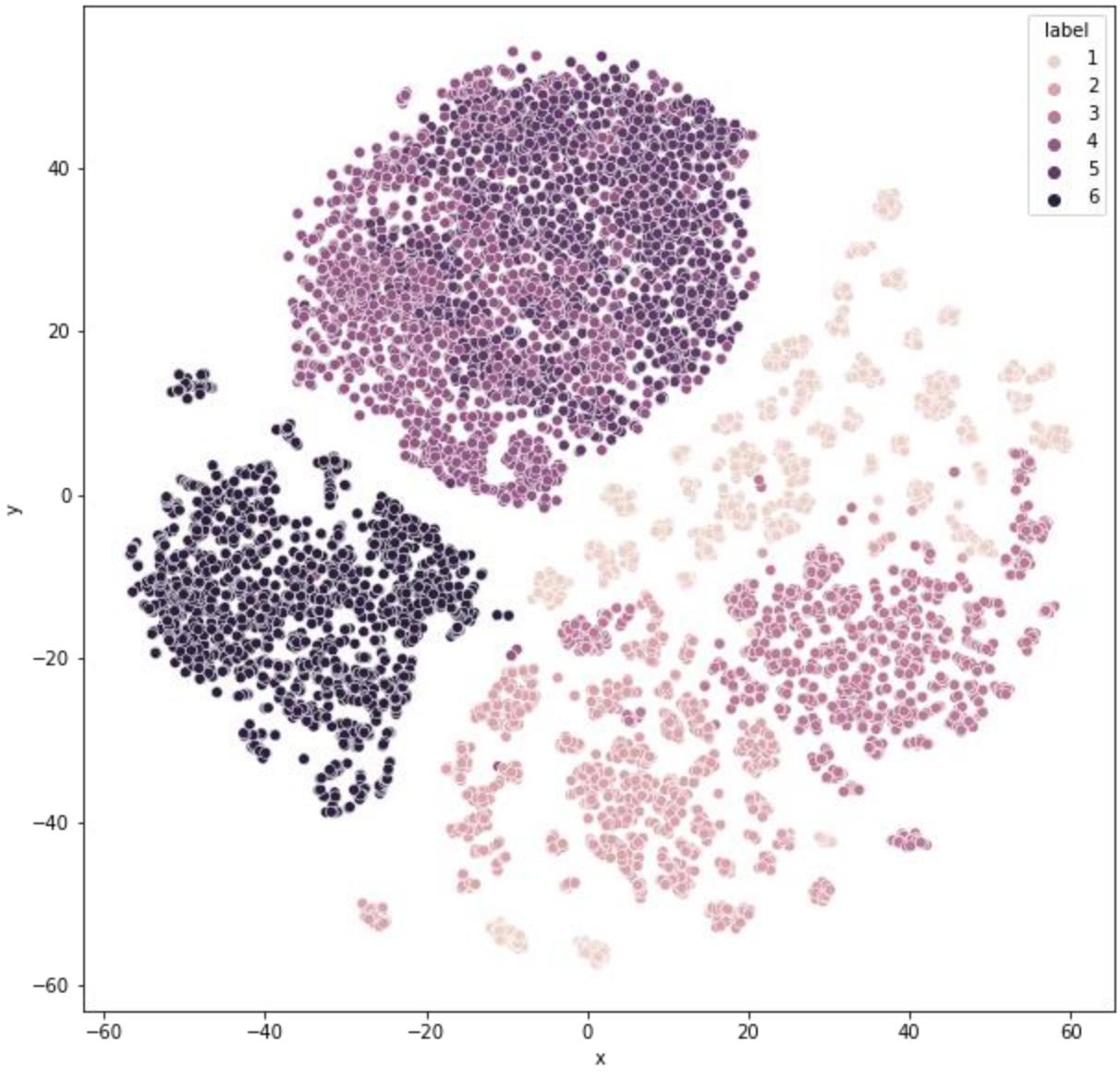


Figure 3.5: Tsne model output of the data features with labels

Conclusion:

Now we can observe that class 4 and 5 seems to be interconnected with each other but the other classes seem to be well separated by the given features. The connection between classes 4 and 5 can be explained by knowing that these 2 classes represent the activities of Sitting and Standing respectively and because both of them require almost no movement at all, measuring the body acceleration will not be very helpful in differentiating between these 2 activities. This finding allows us to conclude that this dataset

can be indeed used to train a machine learning model to separate different activities.

Q4) A study of Asian Religious and Biblical Texts Data Set

Description:

The dataset consists of texts from Asian and Christianity Religious books. Most of the sacred texts in this dataset were collected from Project Gutenberg.

For this dataset we try to find the similarities and differences between some Asian religions and Christianity. We first try to find the common words in all the mentioned religion books by forming the word cloud of the dataset. The figure below shows the word cloud for the entire dataset where we can see that the word “shall” is the most common word in the dataset followed by the word “man” and “thy”. It’s also observed that some religious words like “god”,” lord”,” spiritual”, and “soul” are commonly used in all the religious books.

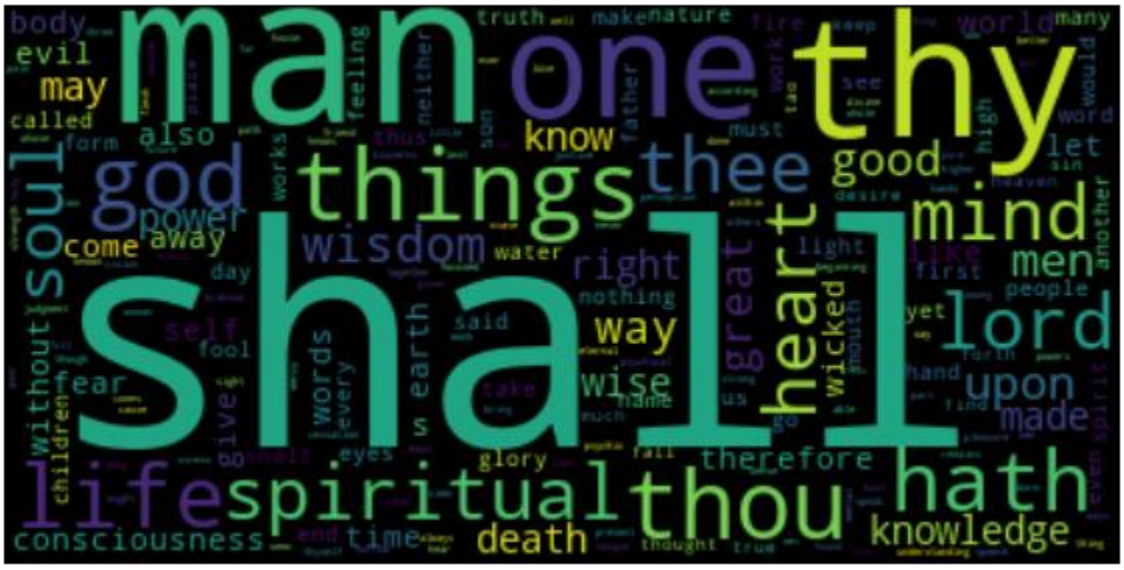
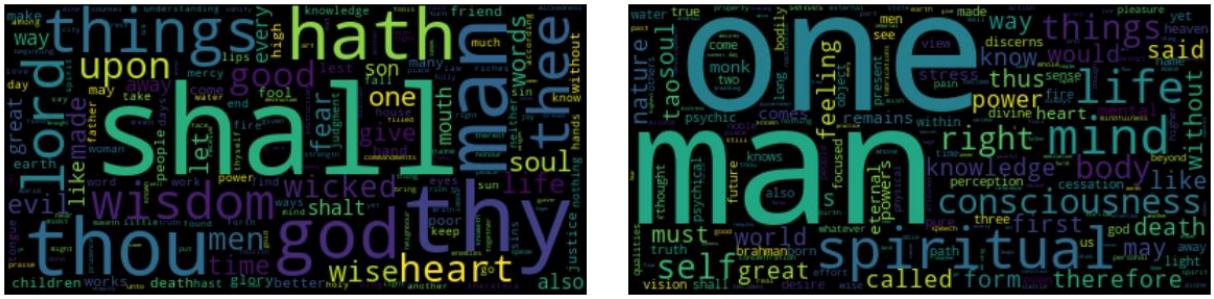


Figure 4.1: Word cloud for the whole dataset

In the next step, we show the word cloud for the books related to Asian religions and the word cloud for all the Christianity to find the differences between the vocabulary used in each of them.

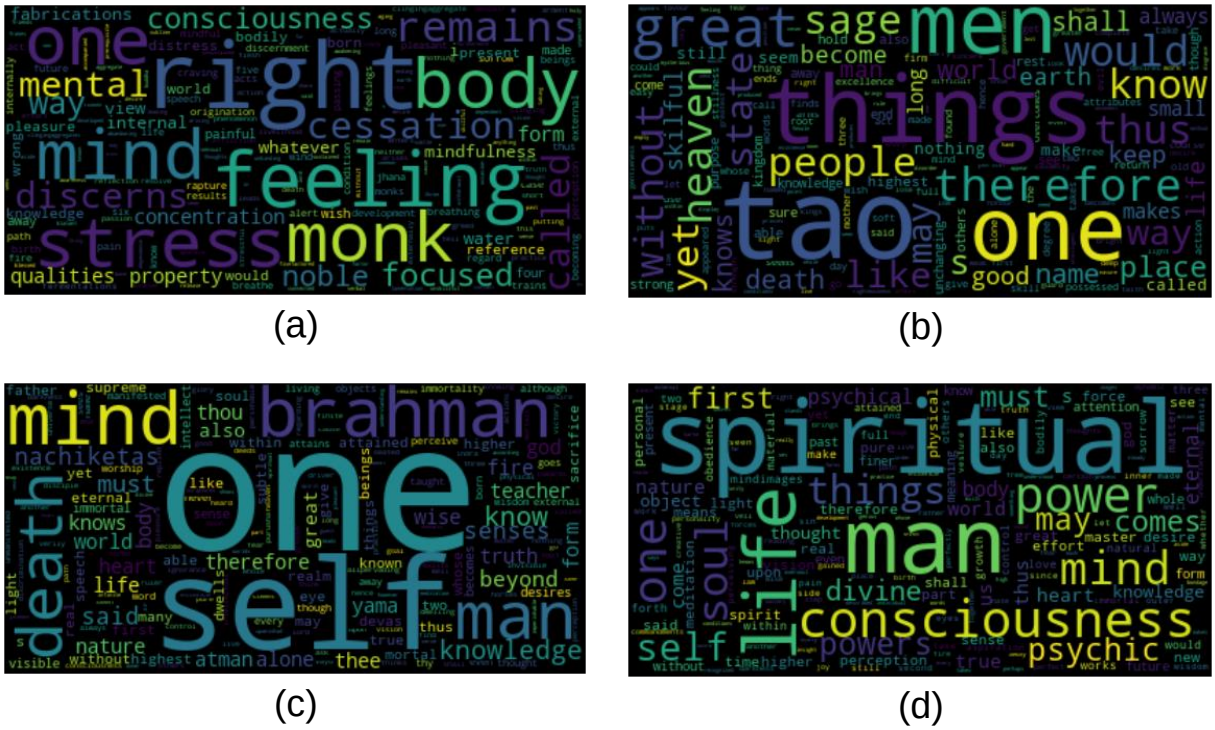


(a) (b)

Figure 4.2: (a) Word cloud for the christian books
(b) Word cloud for the asian religious books

The figure above shows the difference between the words commonly used in the Christian books and the Asian religious books. We can observe that Christian books talk about god explicitly in words like “god” and “lord” Asian religions uses more words that focus on the soul and the inner power within a man like “spiritual”, “power”, and “self”. We can also observe some similarities between the vocabulary used in both classes like the words “man” and “life”. Also, we can observe some words with similar meanings used in both classes of books like “soul-self”, and “wisdom-knowledge”.

We then compare the word clouds for the books that belong to different Asian religions to see how they affect each other and how they use similar words.



(a) (b)

(c) (d)

Figure 4.3: (a) Word cloud for Buddhism book,
(b) Word cloud for TaoTeChing book,
(c) Word cloud for Upanishad book,
(d) Word cloud for YogaSutra book

The figure shows many similar words that are used in different Asian religious books like “One”,” man”,” consciousness”, and “mind”. We can also observe that most of these religions will focus on the inner power inside a man and talk more about the spirit than talking about higher power like a god, for example, this conclusion comes from observing the absence of words like “god”,” lord”,” angels” in all of these books.

For the next part, we show the word cloud for a different Christian books to see how similar they are and how they differ from Asian religious books.

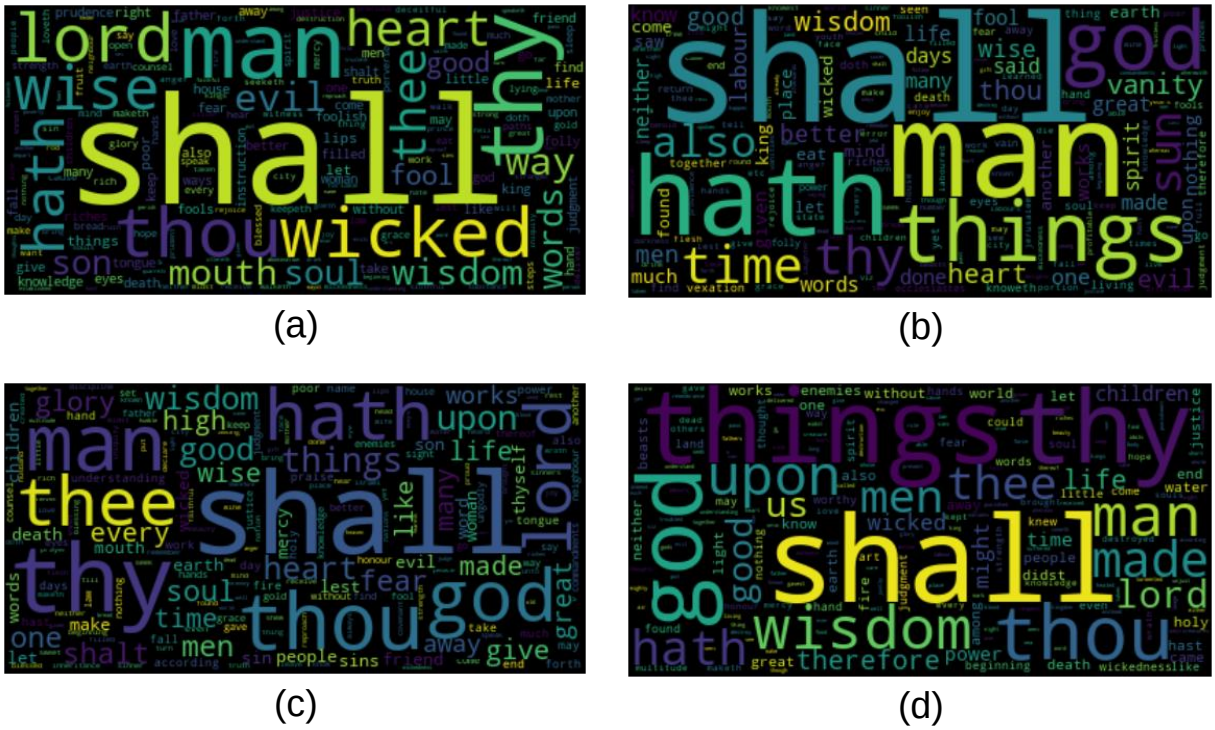


Figure 4.4: (a) Word cloud for Book Of Proverb,
(b) Word cloud for Book Of Ecclesiastes,
(c) Word cloud for Book Of Eccleasiasticus,
(d) Word cloud for Book Of Wisdom

The above figure shows how different Christian books use similar words like “shall”,” god”,” thou”,” thy”, and “lord”. This observation shows that different Christian religious books talk almost about the same topics and use similar vocabulary. Also, the appearance of words like “god”, and “lord” with high frequency in these books show that these books talk more about the existence of a higher power that guides humankind to some righteous path.

Finally, we add the scatter plot for the output of a tsne algorithm to show if there are clusters within the provided data.

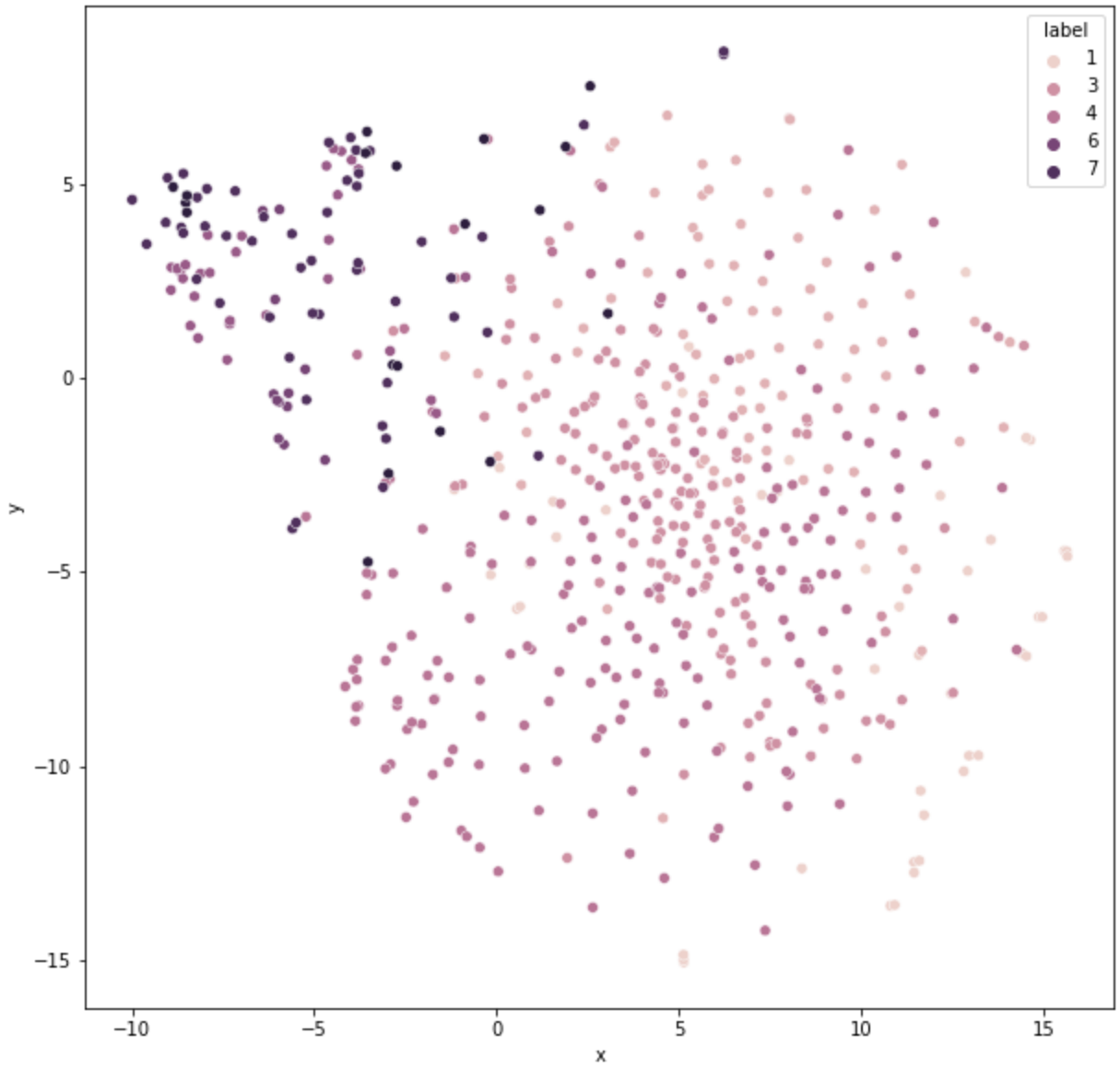


Figure 4.5: Tsne model output of the data features with labels

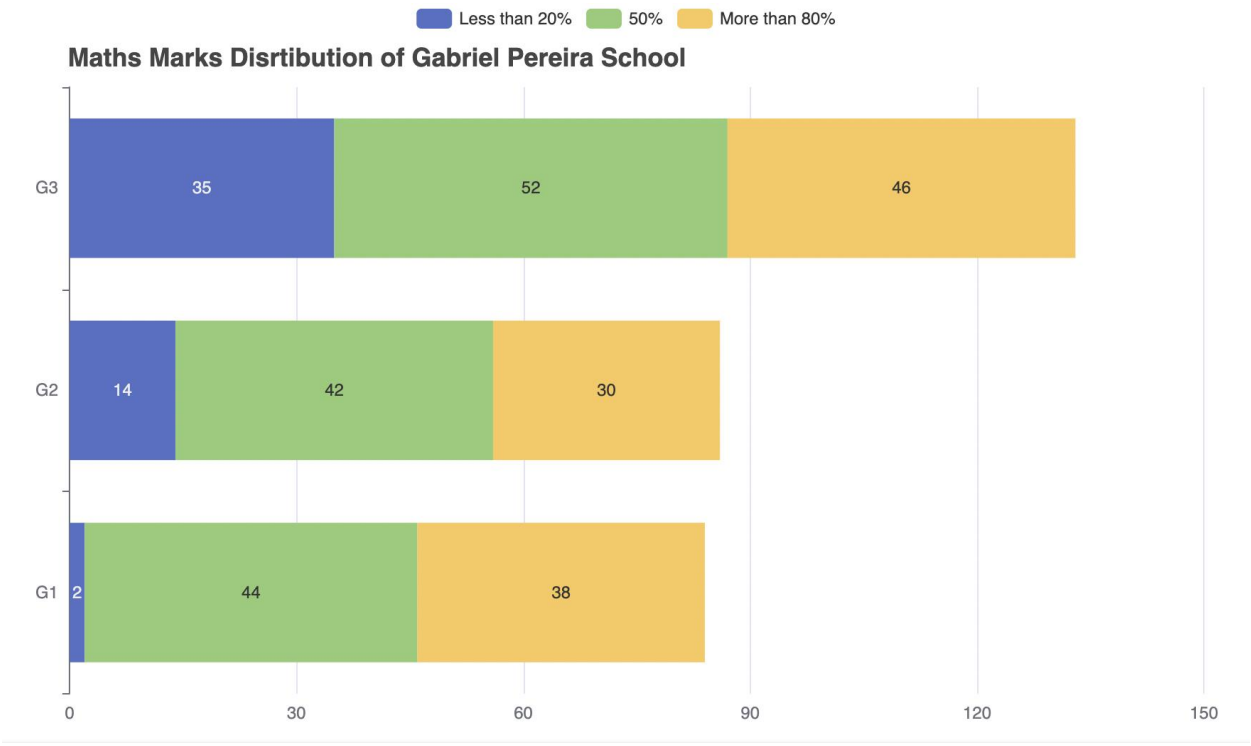
Conclusion:

As we can observe from the figure we don't see small clusters as seen in the previous dataset but rather we find one large cluster in the middle and a small cluster in the top left corner that contains different books related to Christianity. We can see that the given features show some similarities in the vocabulary used in different religious books and a strong relationship between words used in different Christian books.

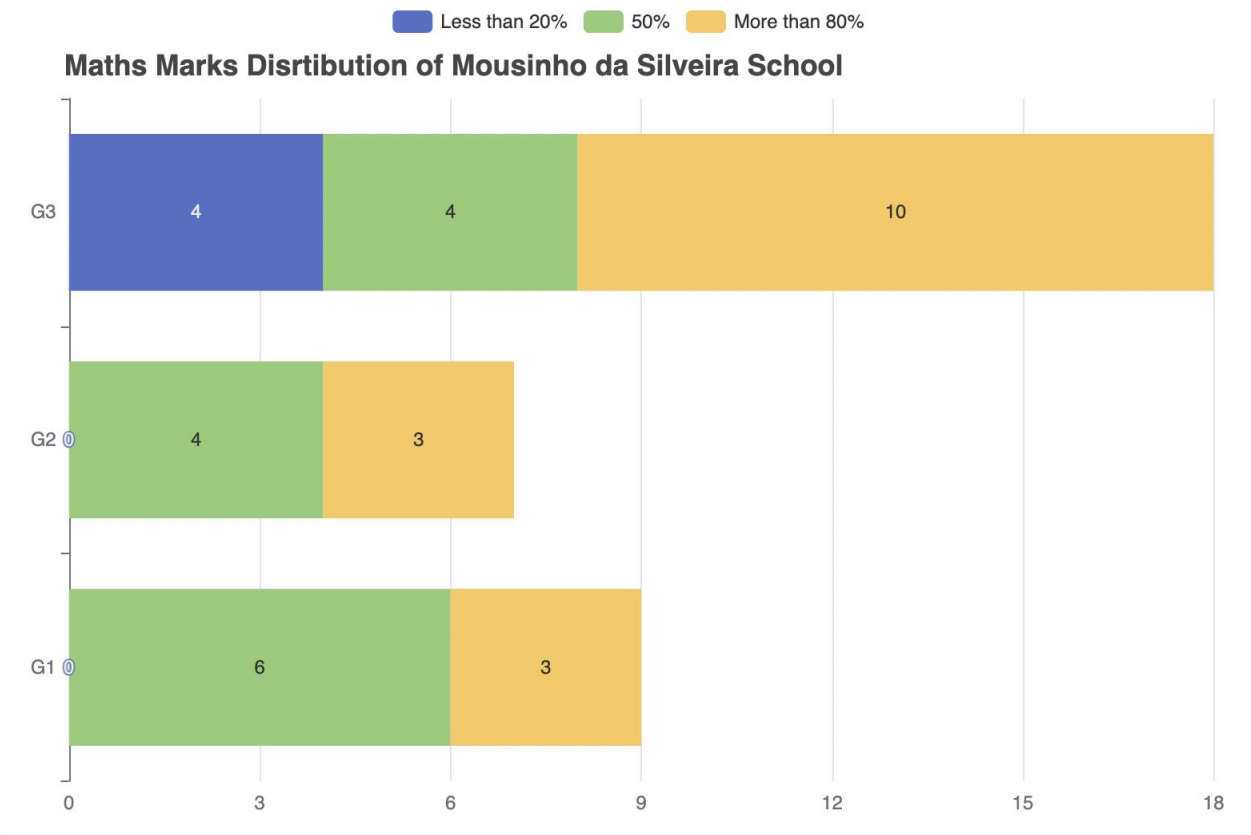
Q5) Student Performance Data Set

Description:

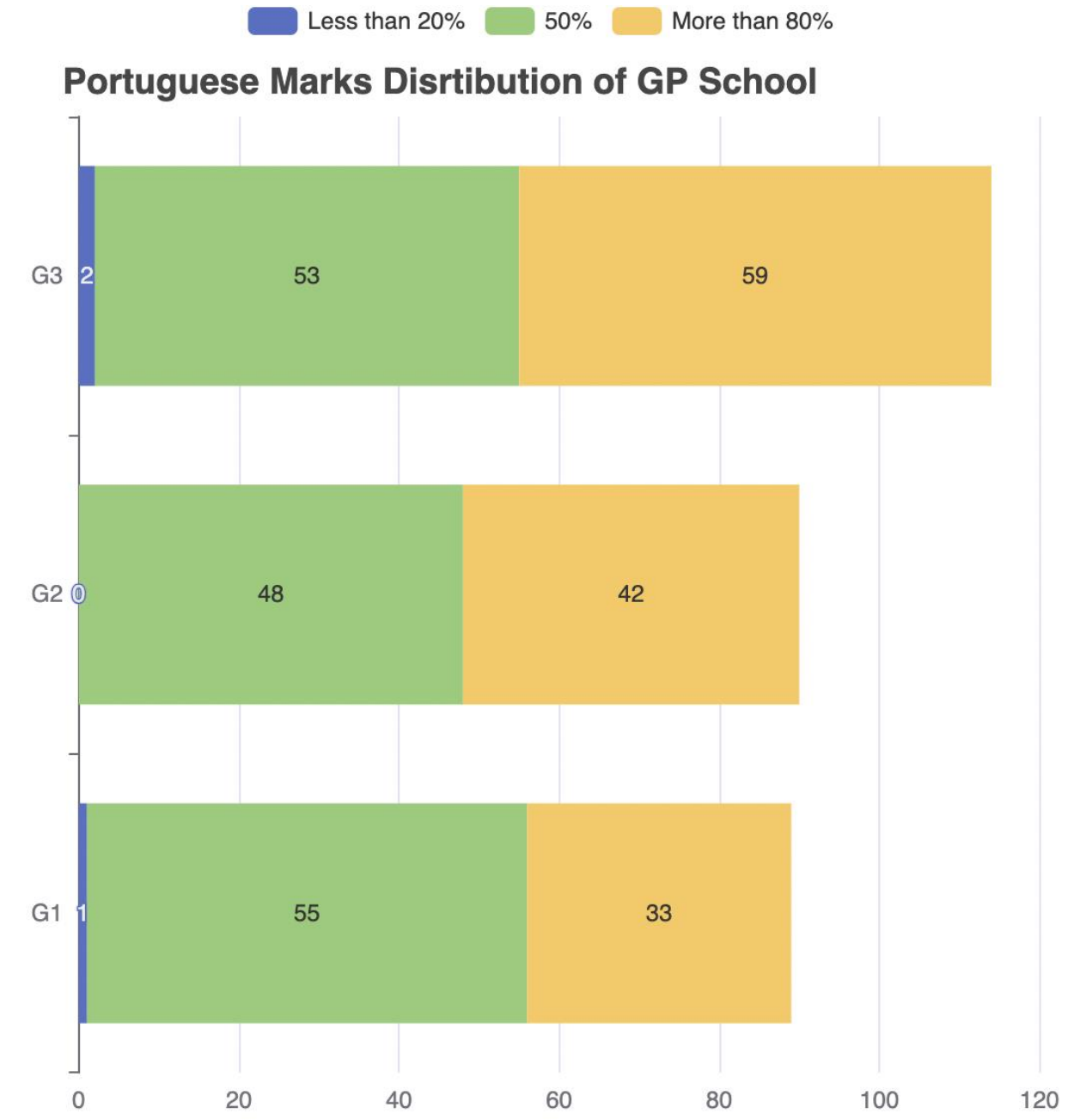
The dataset consists of students' performance in secondary education from 2 different Portuguese schools; **Gabriel Pereira** and **Mousinho da Silveira**.



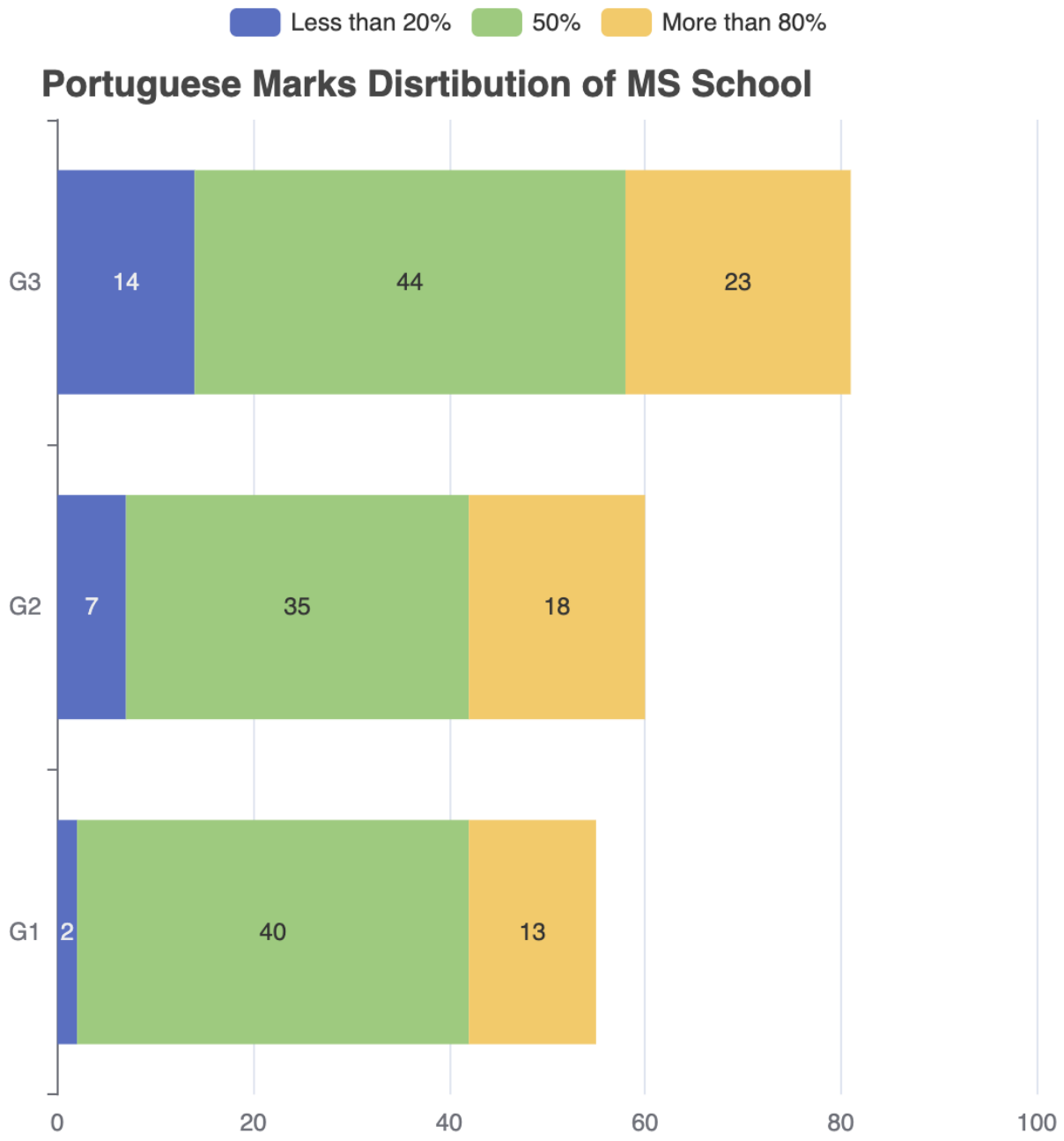
The above graph depicts how many students of **Gabriel Pereira School** gained less than **20%**, **50%**, and more than **80%** made in **Mathematics** 1st period exam (**G1**), 2nd-period exam(**G2**), and lastly, Final exam (**G3**)



The above graph depicts how many students of **Mousinho da Silveira School** gained less than **20%**, **50%**, and more than **80%** makes in **Mathematics** 1st period exam (**G1**), 2nd-period exam(**G2**), and lastly Final exam (**G3**)

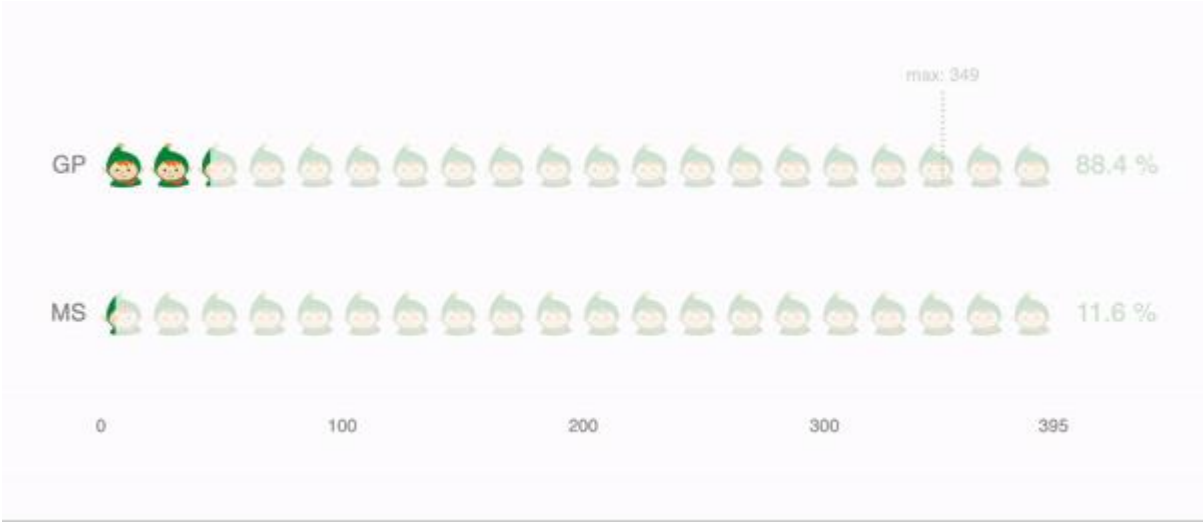


The above graph depicts how many students of **Gabriel Pereira School (GP)** gained **less than 20%, 50%**, and **more than 80%** makes in **Portuguese** 1st period exam (**G1**), 2nd-period exam(**G2**), and lastly Final exam (**G3**)



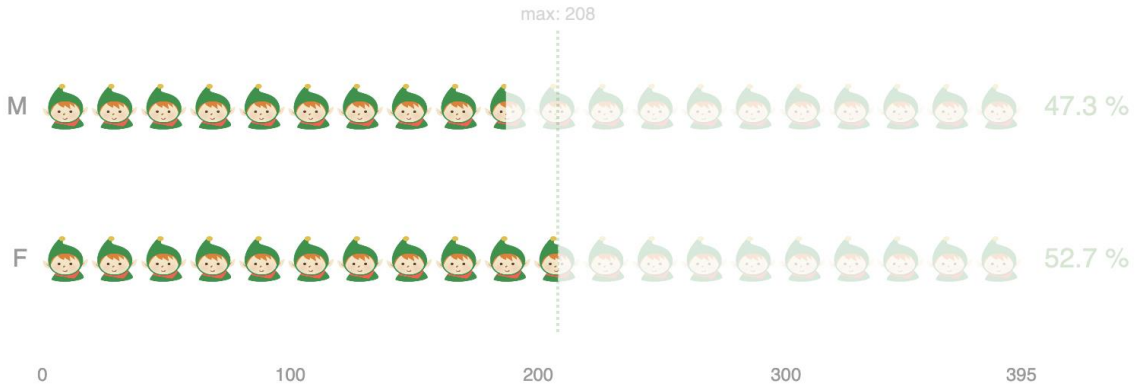
The above graph depicts how many students of **Mousinho da Silveira School (MS)** gained **less than 20%, 50%**, and **more than 80%** makes in **Portuguese** 1st period exam (**G1**), 2nd-period exam(**G2**), and lastly Final exam (**G3**)

This proves that both of the school's students worked hard and did better in their Final Exam of Math - G3. However, there is less number of students from Mousinho da Silveira School (MS) who achieved more than 80% in the Final Exam (G3) - Mathematics as compared to Gabriel Pereira School (GP). The reason is illustrated in the following chart.

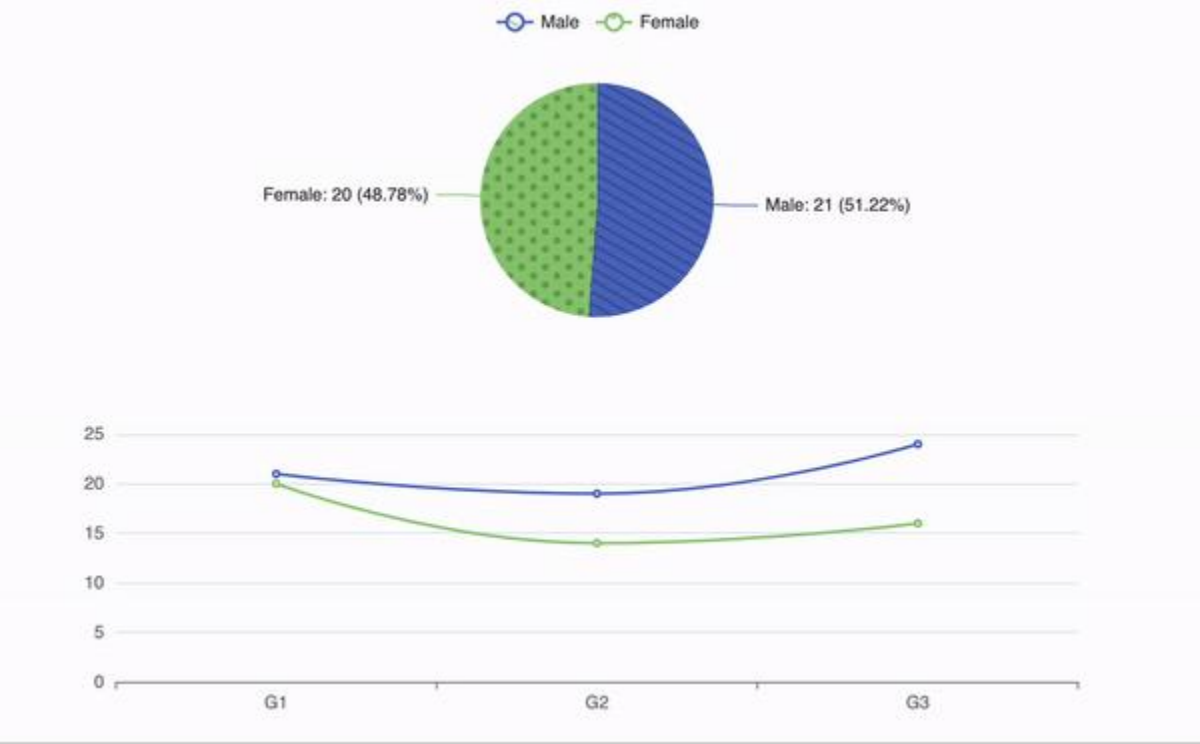


There are **395** students in the dataset total out of which **88.4%** are from **Gabriel Pereira School (GP)** on the other hand only **11.6%** are from **Mousinho da Silveira School (MS)**, this shows that the dataset is **not balanced** therefore, there are more students from Gabriel Pereira School who achieved more than 80% marks in their Final Exam (G3) - Mathematics

On the other hand, if we look at the Male-Female ratio it is slightly balanced, so it is fair to analyze which gender performed better in the Final math exam.



Male-Female ratio of more than 80% marks

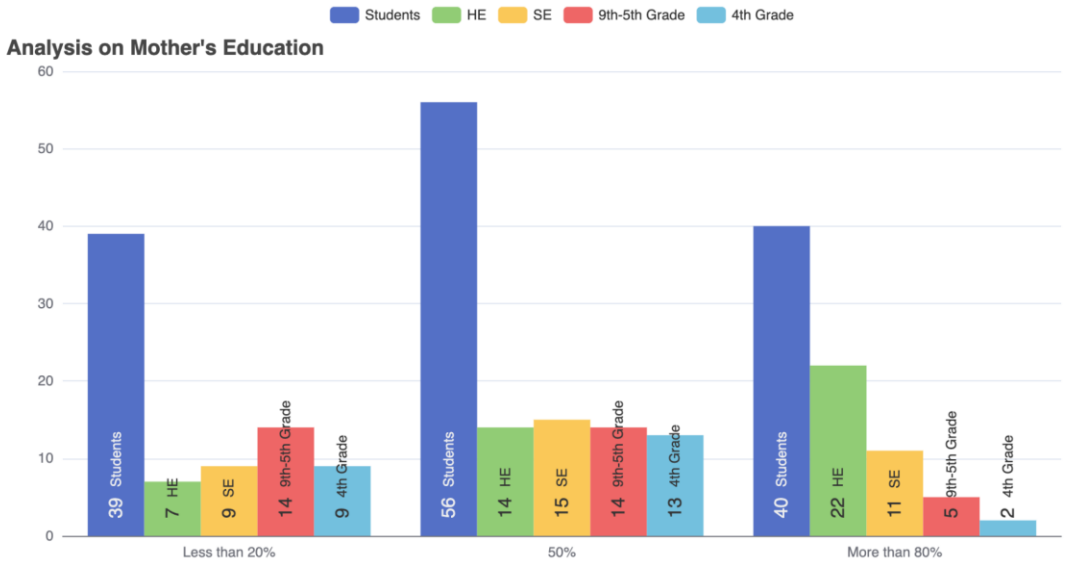


The above chart depicts that male students have performed better than female students in all 3 exams G1, G2, and G3. In G1 male and female students performed nearly the same but in G2 this difference grew more and lastly, in final exam-G3 **60% of the male students** got more than **80%** whereas **only 40% of the female students** made it to **above 80%**.

Let's investigate **why** these students were able to secure **more than 80%** in the final exam **G3**.

According to a study done at New York University, it is found that **a mother's education has a significant impact on the success of her child**. Therefore, to prove this study, we analyzed students' mother education and how it impacts their marks in exams.

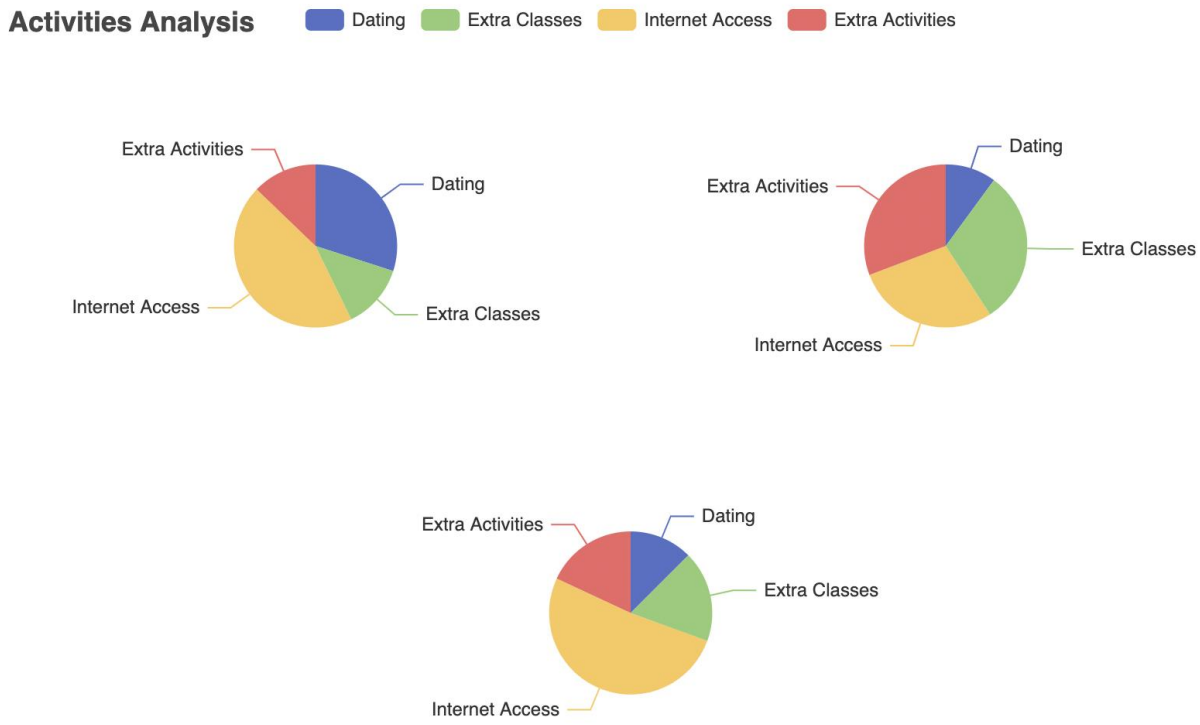
The following chart shows the total number of students for each category (Less than 20% marks, 50% marks, and more than 80%) for the G3 exam. **It is proved that the mothers who have completed higher or secondary education has helped their children in getting more than 80% marks whereas mothers who are simple school pass out, their children got less than 20% marks**.



HE = Higher Education
SE = Secondary Education

There are 40 students who got more than 80% marks in G3 and out of them, **22 students' mothers have completed higher education, and 11 students' mothers have completed secondary education**. On the other hand, if we look at those students who got less than 20% marks in the G3

exam, **only 7 students' mothers have completed higher education and most of them are 9th to 5th-grade graduates**.



For further investigation, we did an analysis of the activities of students who got less than 20 % marks, students who 50% marks, and students who got more than 80% marks in G3 exam.

From our analysis, we found out that, students who performed **poorly** were in **a romantic relationship with someone**, which might have caused distraction from their studies whereas students who got **more than 80%** marks were **mostly single and were not involved in any kind of romantic relationship** so they might have been able to focus more on their studies. Similarly, students who got **less than 20%** were **less involved in extracurricular activities** as compared to students who **got 50% or greater results**. If we talk about **internet access** then less than 20% marks students and more than 20% marks students had an almost **same ratio**. But it depends on an individual how are they utilizing that access. Looks like more than 80% marks students took **help for their exams** and studies from this accessibility on the internet, whereas less than 20% marks students might

have misused this accessibility on the internet and wasted their time chatting or internet surfing. Looking at the ratio of their extra classes, we can say that **50%** marks students **took more extra classes** as compared to the other **two** groups of students, the reason could be **this group is trying hard to do better** and achieve more marks through hard work and dedication.

Conclusion:

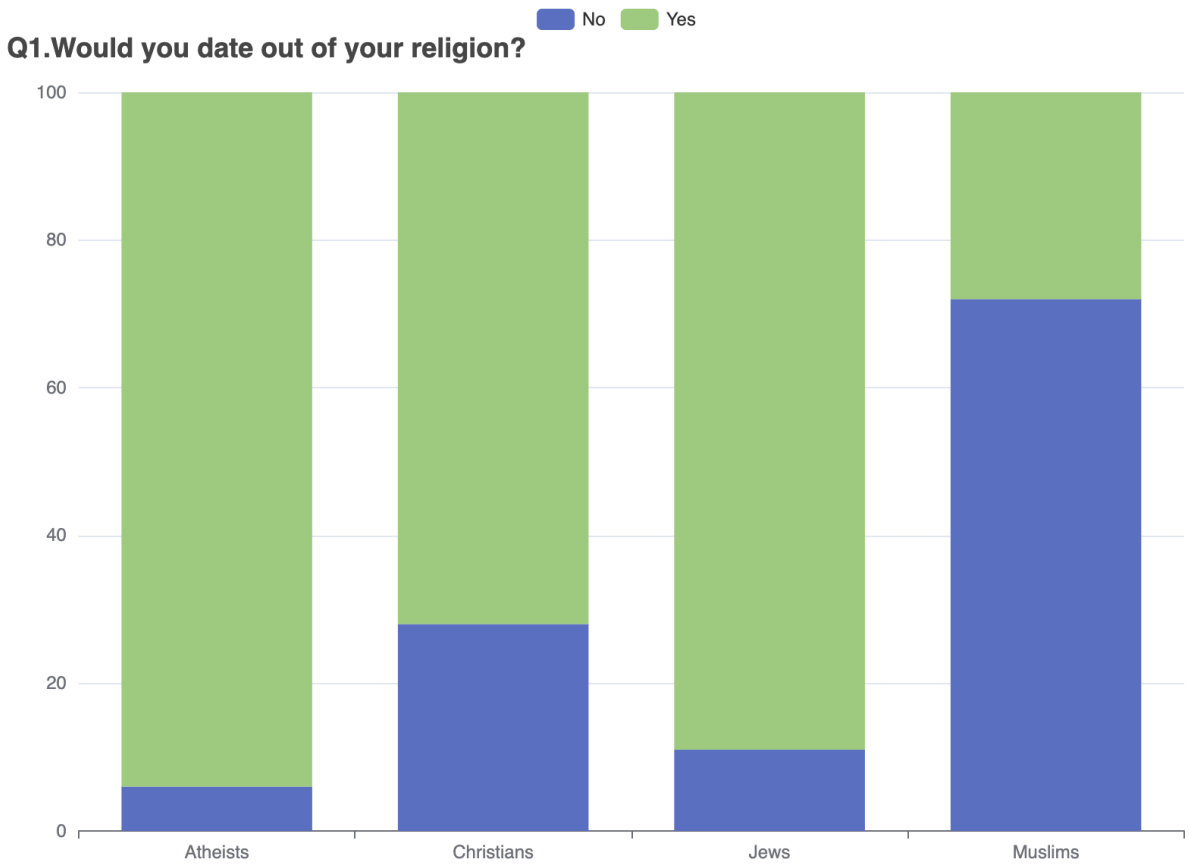
To conclude, we can say that students worked harder in G3 exams for both math and Portuguese as seen from the graphs and one of the major reasons of the success of these students was their mother's education. Apart from that whether they were involved in a romantic relationship or not played an important part too with the inclusion of other activities and extra classes.

Q6) Survey Questions

Description:

The dataset consists of a survey performed on 4 different groups of people; Atheists, Christians, Jews, and Muslims. In this survey, 15 different questions regarding beliefs, religion, marriage, and gender equality have been asked to these people.
The answer of the people are divided into 3 categories; **yes, no**, and **undecided**.

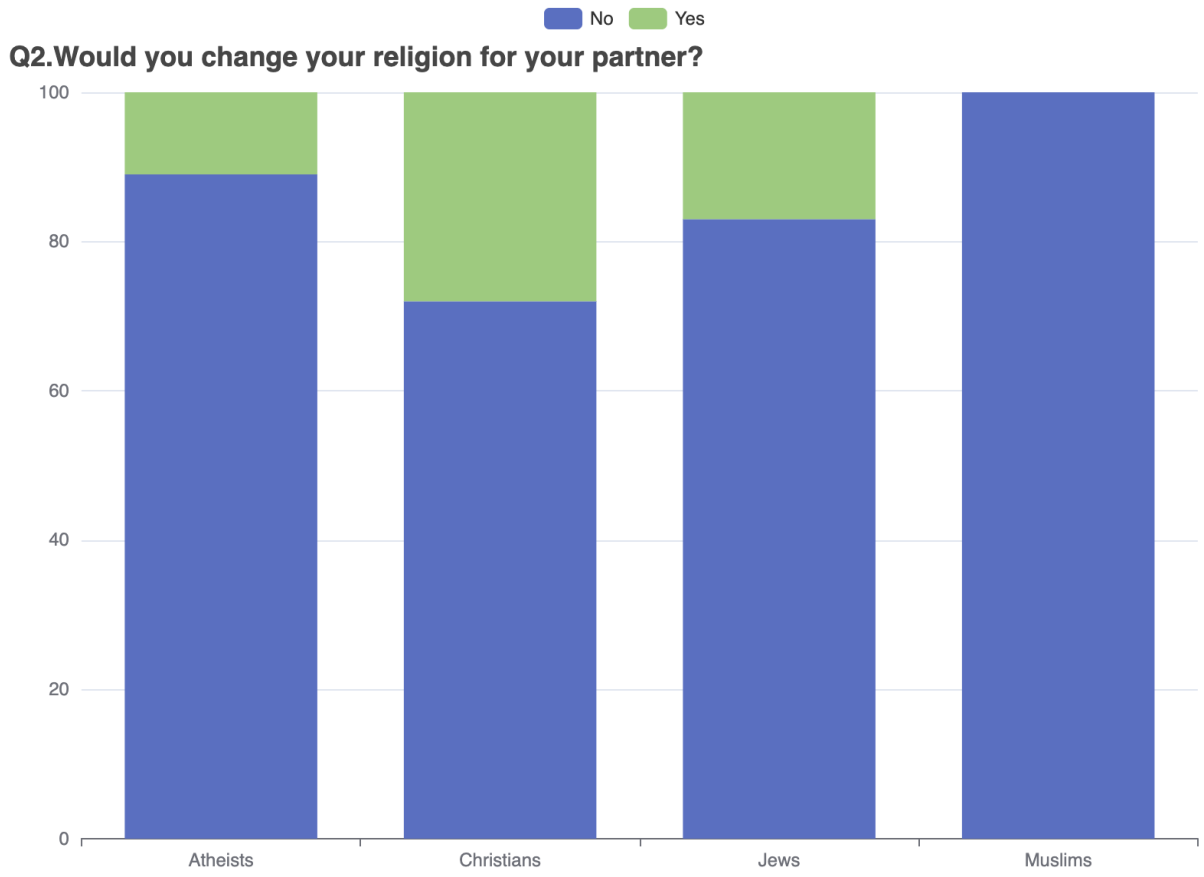
The following bar chart depicts the ratio of Yes and No for the question "Would you date out of your religion?"



From the chart, it can be concluded that most of the Muslims answered **No** to dating someone out of their religion, it makes sense as it is not allowed in their religion, however, not all Muslims practice their religion fully, therefore, 28 of them answered **Yes**.

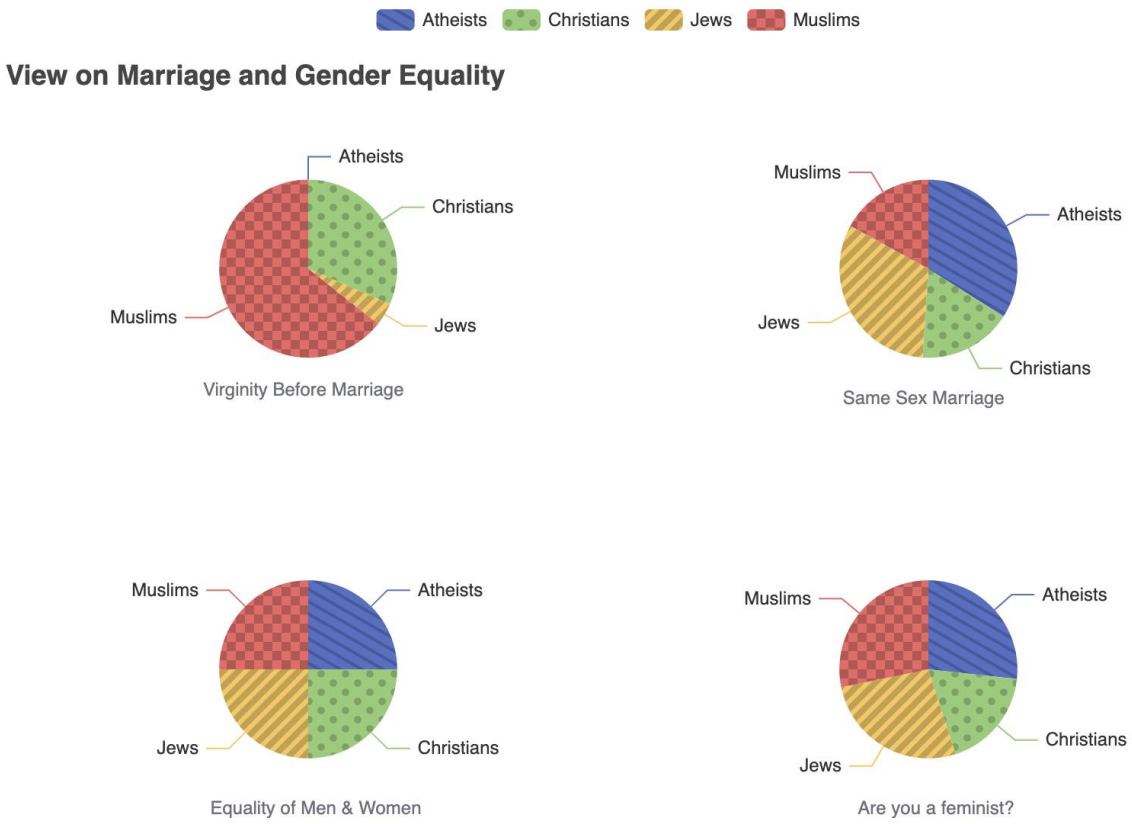
If we talk about other groups; Atheists, Christians, and Jews so the majority of them answered **Yes** to the question, the reason for their yes might be they are willing to convert their religion for their partner.

To find the reason for their Yes we did the analysis on question 2 which is "**Would you consider converting to a different religion for your partner?**"



To our surprise, we see that most of the people the majority yes in question 1 answered **No** here, with an exception of the Muslim community.

It makes sense for the Muslims as the majority said they will not date someone out of their religion, therefore, they are not willing to change their religion for their spouse too. But surprisingly Atheists, Christians, and Jews are also not willing to change their religion for their spouse. This shows that they will date someone out of their religion but they will never change their religion for them.



Next, we investigated their reviews on **Marriage and gender equality**. All four pie charts above show people who answered yes in each scenario.

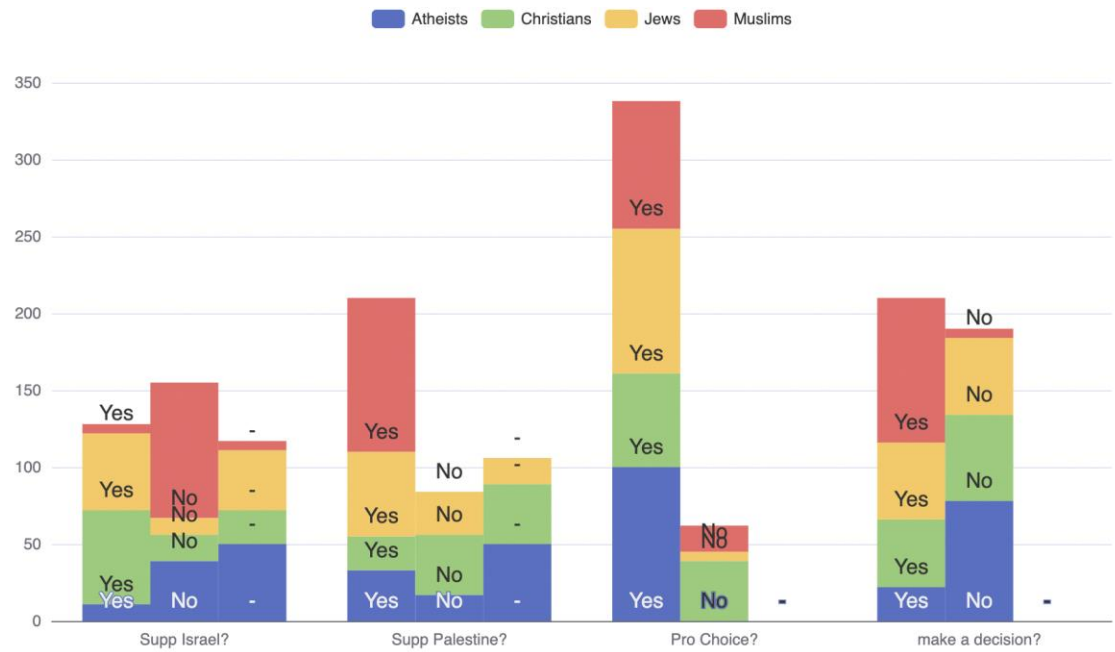
Q. Do you believe people should save their virginity for marriage?
From the pie chart, we can see that majority of the Muslims strongly believe that people should save their virginity before marriage, this statistic makes sense as it is what their religion says about it, and from the first two charts, we can see that they practice their religion strongly as compared to other 3 groups. On the contrary, there are **0 Atheists** who support this statement. Similarly, only **6 Jews** support this statement too whereas Christians are the second majority after Muslims who belives to save their virginity before the marriage.

Q. Do you support same-sex marriage?
Following the same trend in question 3, Jews and Atheists have a similar ratio of yes **94 vs 100**, and Muslims and Christians have the same ratio of **50 vs 50**.

Q. Do you think men and women deserve equal rights?
Surprisingly, all the four groups; Atheists, Muslims, Christians, and Jews have the same ratio for this question. All **100 people from 4 different groups agreed** that women and men should have equal rights.

Q. Are you a feminist?
Following the same trend in questions 3 and 14, Atheists and Jews answered the question in the same ratio of **89 vs 89**. On the other hand, **94 out of 100** Muslims said they are feminists, while **61** Christians said yes.

To get information on how their **general views coincide with each other,** we answered the 4 groups some **controversial questions** and the results are as follow:



Q. Do you support Israel?
From the bar chart, it can be seen that Jews and Christians are in majority who are supporting Israel in this survey and it makes sense as the majority of the Jewish lives in Israel, whereas on the other hand, Muslims are in

majority who are not supporting Israel in this survey, the reason could be their support for the Muslims in Palestine. However, the majority of the Atheists are confused about what should be their take on it.

Q. Do you support Palestine?

The reason for Muslims not supporting Israel is justified here as we can see that the majority of the Muslims are in favor of Palestine, therefore they are against Israel. But surprisingly, Jewish are the second majority supporting Palestine after Muslims despite the fact that most of the population of Israel is Jewish. Following the same trend in Question, last question majority of the Atheists remain undecided.

Q. Are you pro-choice?

None of the 4 groups were undecided on this question. All of the atheists answered yes and following the same trend Jews answered Yes being the second majority after Atheists. Whereas, 39 Christians and 17 Muslims are against it.

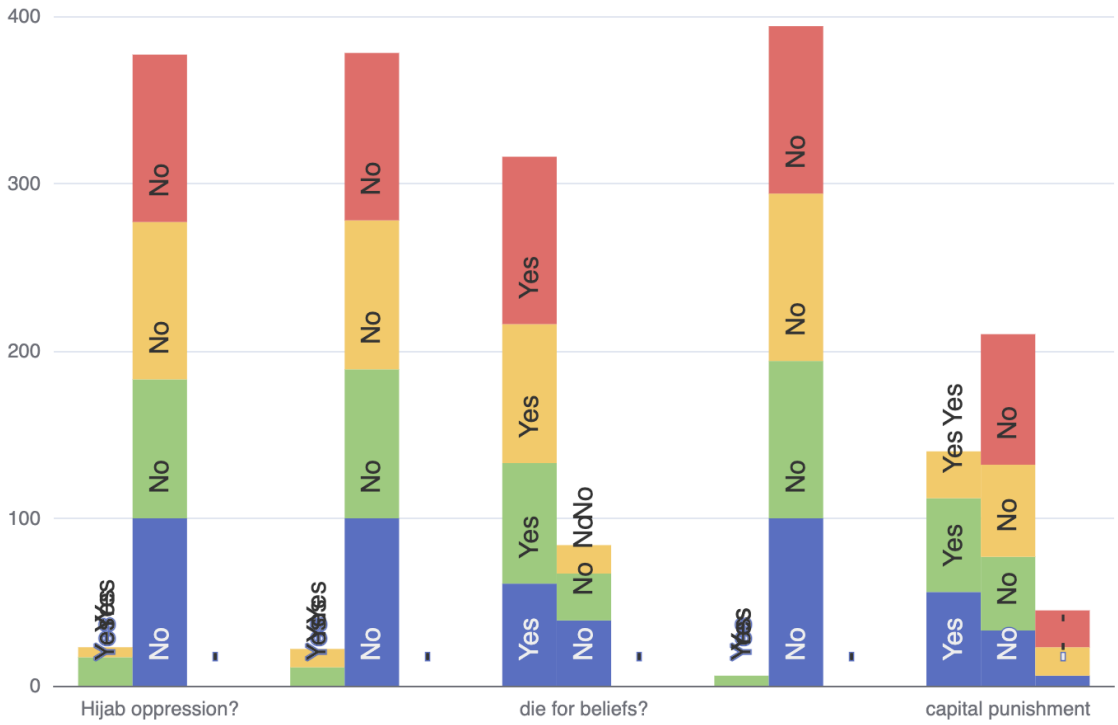
Q. Do you feel informed enough to make a decision?

For this question majority of the Atheists answered No, this shows they are doubtful about their decision-making ability, whereas Christians and Jews have an almost 50-50 ratio of yes and no. On the hand Muslim group is in majority who believes that they are informed enough to make a decision.

To analyse which group is considered as extremist we asked them the following questions;

Is Hijab symbol of oppression?
Would you kill for your beliefs?
Would you die for your beliefs?
Are non-religious people immoral?
Do you support capital punishment?

And here’s what we found out:



The above graph shows that for questions; **Is Hijab symbol of oppression, Would you kill for your beliefs** and **Are non-religious people immoral** the majority of the groups have answered **NO.** This proves that none of these 4 groups are considered as extremists. For questions regarding **Would you die for your beliefs,** the marjority of 4 groups answred **YES** which means that all the 4 groups belonging to different religions are firm and have the same level of committment to their own beliefs and values. For the last question regarding the support to **capital punishment,** the all have a mixed views with **Yes, no and some are even undecided**.

Conclusion:

To sum up, we can say that we found a similarity between the answers of Muslims & Christians, and Atheists and Jews. One of the main interesting finding was that all the four groups were together on the point of gender quality.

Q7) MOO_Solution_Candidate_Set

For this dataset we first need to formulate the problem in order to try and find hidden patterns within the dataset. By observing both the decision domain and the objective domain data we find that we are facing a multiobjective optimization problem. The histogram for four column of the objective domain are shown below to find the trend of values in the objective space.

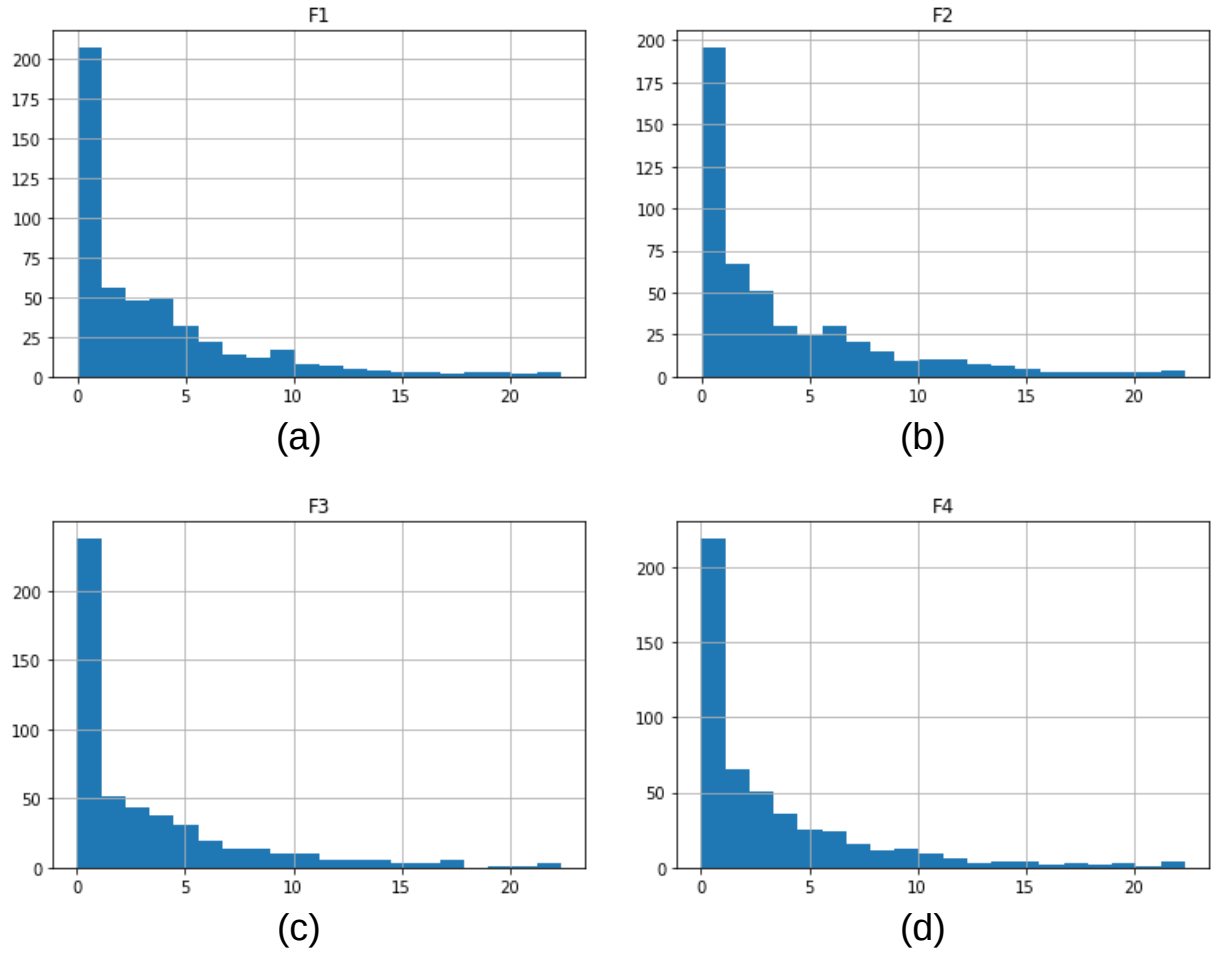


Figure 7.1: Histogram for (a) F1, (b) F2, (c) F3, (d) F4

The figures show that most of the values for any objective variable are low values which imply that the goal of the optimization problem is to minimize

the values for all variables in the objective space by controlling the values in decision space.

So we basically need to minimize all the values for the objective functions **F** by changing the values of **X.** The main goal here is to visualize the relation between decision space and objective space to find which values of **X** minimize most of the values for **F**. The first challenge is the high dimensionality nature of the problem as the feature domain **X** has 100 dimensions and the objective domain F has 15 dimensions, another challenge is showing the mapping between values in **X** domain and values in **F** domain. We begin tackling the first problem by using the tsne algorithm to perform dimensionality reduction on both decision domain variables and objective domain variables. Results are shown in the figure below.

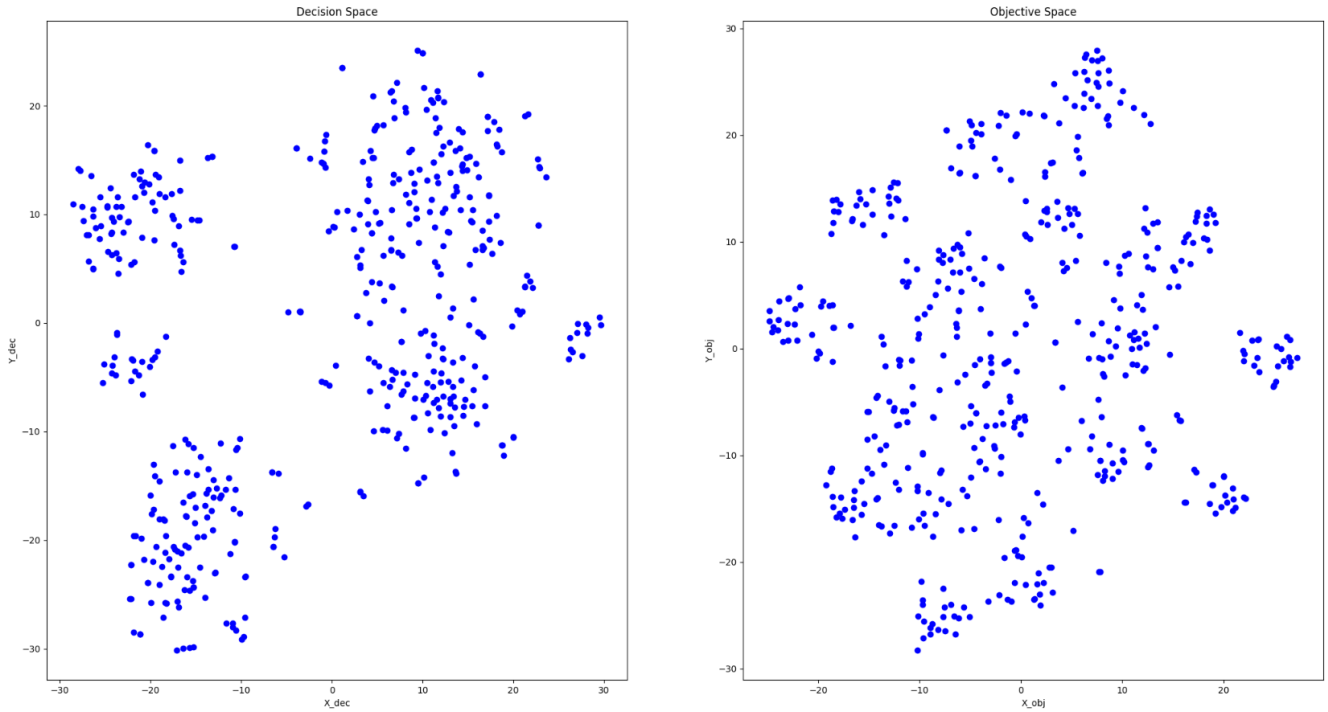


Figure 7.2: scatter plot for the output of t-sne algorithm for the decision domain (left) and objective domain (right)

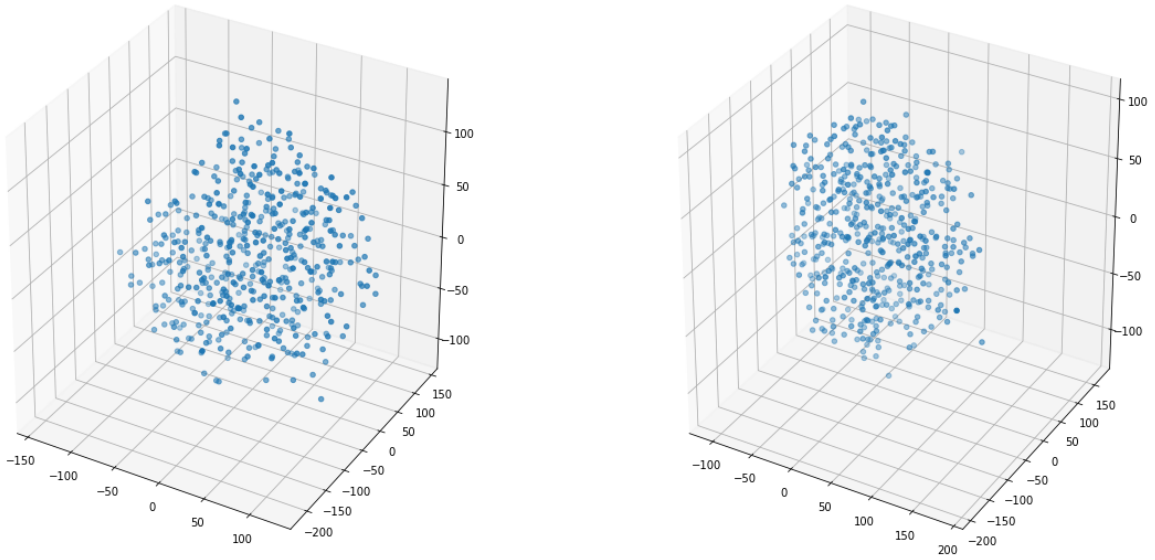


Figure 7.3: 3D scatter plot for the output of t-sne algorithm for the decision domain (left) and objective domain (right)

After this, we needed to find how to visualize the mapping from the decision domain to the objective domain. The first approach was to draw a connection line from every point in the decision space to the corresponding point in the objective space the results are shown in figure 7.4

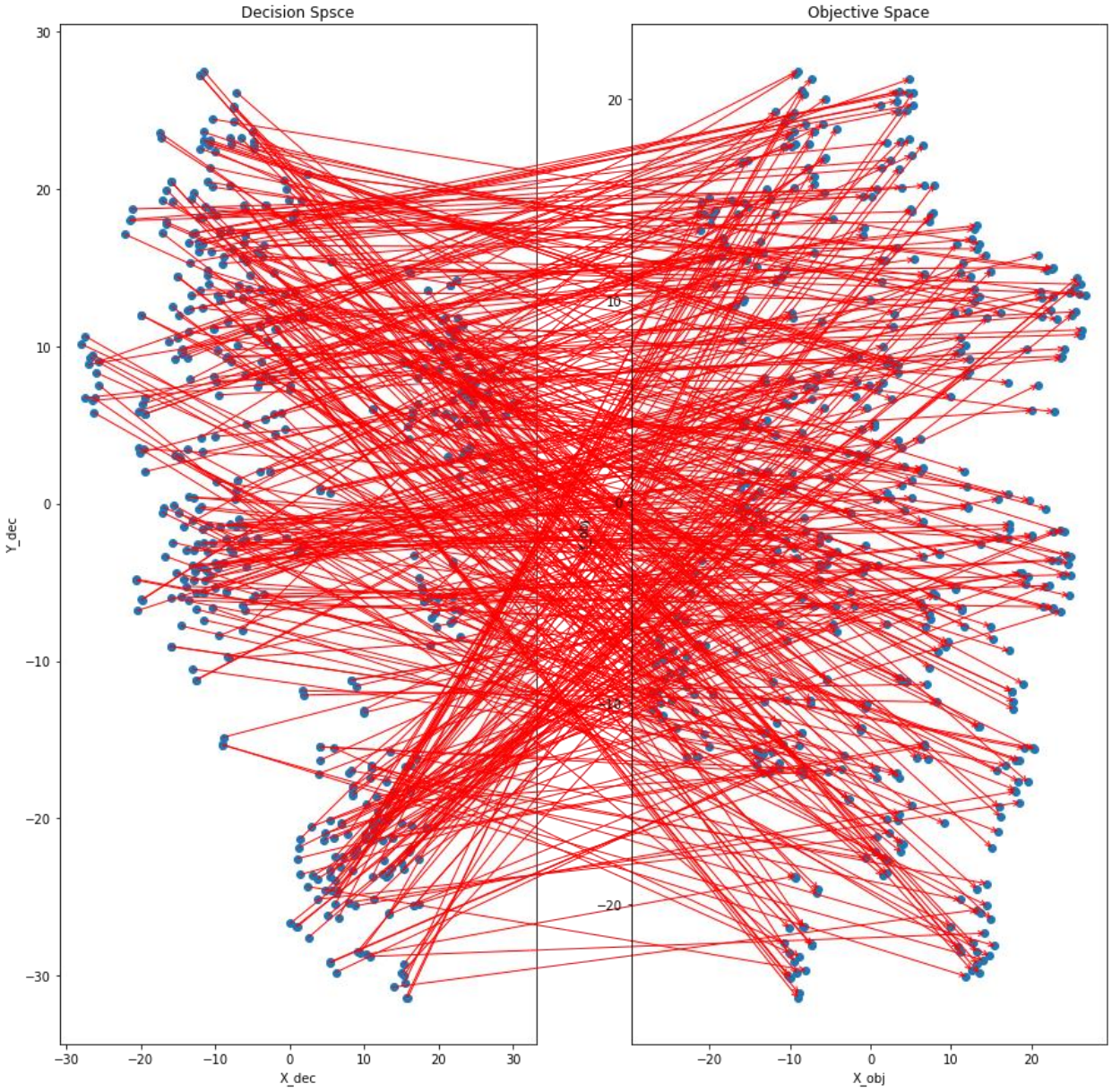


Figure 7.4: Connection lines from decision space to objective space

The figure shows that the huge number of connections makes the visualization of one-to-one connections between a point in the decision domain and the corresponding point in the objective domain very difficult. The user will not be able to gain any useful information from this graph. Wee

then try to visualize the mapping using parallel coordinate plot. The result is shown in figure 7.5

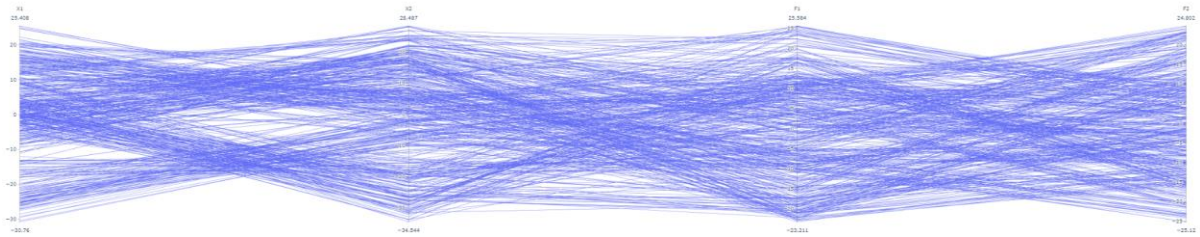


Figure 7.5: Parallel coordinate plot between decision space and objective space

We ended up facing the same problem which is huge number of connections that makes the graph un recognaizable.

Finally, we tried was to make this graph user interactive. In order to do this, a new python class was developed to handle the visualization of each point as an object and a method in this class was created to handle the event of a mouse click on one point in the decision domain or in the objective domain. When this event happens a connection line from the point in the decision space to the corresponding point in the objective space will be drawn if a connection line is already drowned between these two points and the user clicks on any of these two points the connection line will disappear this allows users to only show the mapping for points of interest for the user which is commonly the points that form the Pareto front in the objective space. Results of the proposed approach for the case of 3 connections are shown below.

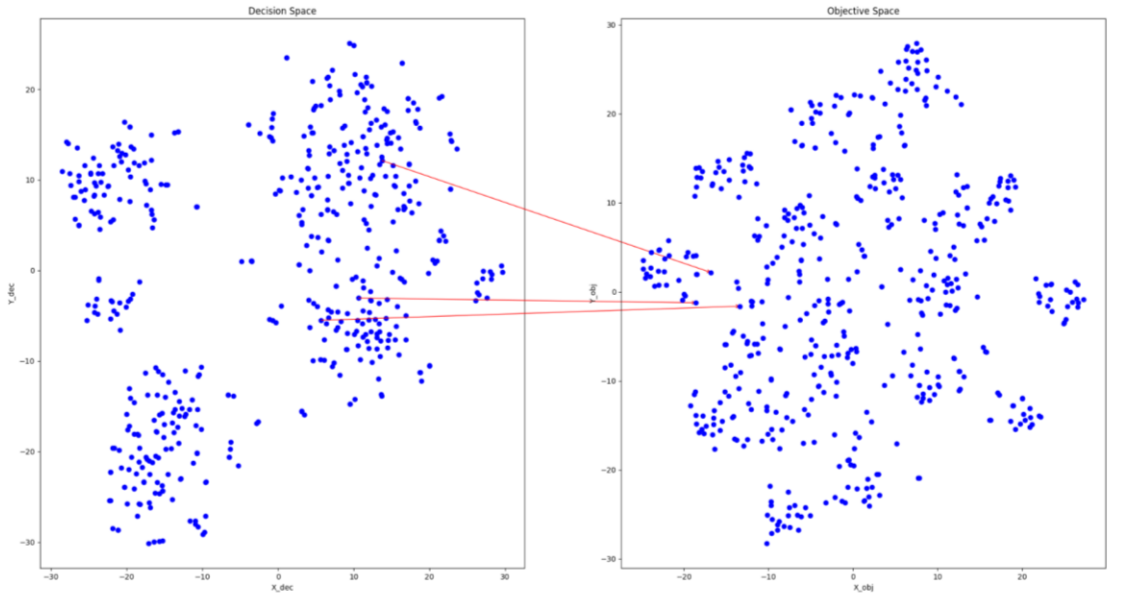


Figure 7.6: Proposed approach in case of using 3 connections

Conclusion:

Results show that the mapping approach proposed enabled the user to find the mapping between the decision space and objective space for points of interest to find which points in the decision space performed best i.e minimized most variables in the objective space.