STATS 202: DATA MINING AND ANALYSIS FINAL

Instructor: Linh Tran

Final Project

Due Date: August 2, 2023

Stanford University

Adam Kainikara

Kaggle Reference

```
Team Name: Adam Kainikara.
Work done by Adam Kainikara.
```

Part 1. Treatment Effect

Using Python I loaded the data from all available CSV sets. Created a function called find_patients which took the input of the loaded data and returned every unique patient id. After collecting every patient id, I created a blank dictionary. The dictionary would have the patient id as the key. For values, I stored all the associated values of each patient in an array. For patients that had one or more visit days, the information would also be stored. The following is an example of how I stored the data.

dtype=[(’study’, ’<U10’), (country, ’<U10’), (’txgroup’, ’<U10’), (’assesmentid’, ’<f8’), (’patientid’, ’<f8’), (’visitday’, ’<i4’), (’xvalues’, ’<f8’, (30,)), (’panss’, ’<f8’), (’leadstatus’, ’<U10’)]), 30951.0: array([(’"C"’, ’"China"’, ’"Control"’, 304958., 30951., 0, [4., 3., 2., 1., 3., 4., 2., 4., 5., 4., 5., 4., 5., 3., 1., 2., 1., 1., 4., 1., 4., 4., 3., 3., 4., 6., 3., 1., 5., 2.], 94., ’"Assign to’), (’"C"’, ’"China"’, ’"Control"’, 301327., 30951., 7, [4., 3., 2., 1., 2., 4., 2., 4., 5., 4., 5., 4., 5., 3., 1., 1., 1., 1., 4., 1., 4., 4., 3., 3., 4., 6., 3., 1., 4., 2.], 91., ’"Assign to’), (’"C"’, ’"China"’, ’"Control"’, 303725., 30951., 14, [4., 3., 2., 1., 1., 4., 2., 4., 5., 4., 5., 4., 5., 3., 1., 1., 1., 1., 4., 2., 4., 4., 3., 3., 4., 6., 3., 1., 4., 2.], 91., ’"Passed"’), (’"C"’, ’"China"’, ’"Control"’, 304954., 30951., 42, [4., 3., 4., 1., 1., 4., 2., 4., 5., 5., 5., 4., 5., 3., 1., 2., 1., 1., 4., 3., 4., 4., 3., 3., 4., 6., 3., 2., 5., 2.], 98., ’"Passed"’), (’"C"’, ’"China"’, ’"Control"’, 307645., 30951., 70, [4., 2., 2., 3., 1., 4., 3., 5., 5., 6., 6., 5., 5., 4., 2., 3., 1., 1., 4., 3., 5., 5., 3., 3., 4., 6., 3., 3., 5., 2.], 108., ’"Passed"’)

After this was done, patients were further filtered into a control group and a treatment group. In addition, patients with fewer than one visit day were removed. Below is a graph of treatment and control groups. The X axis is visit day and Y axis is PANSS score. To reduce clutter on the graph, only Study E is shown.

After, I calculated the difference between each patient’s final PANNS score and their initial score. The distribution of the difference in scores can be seen bellow.

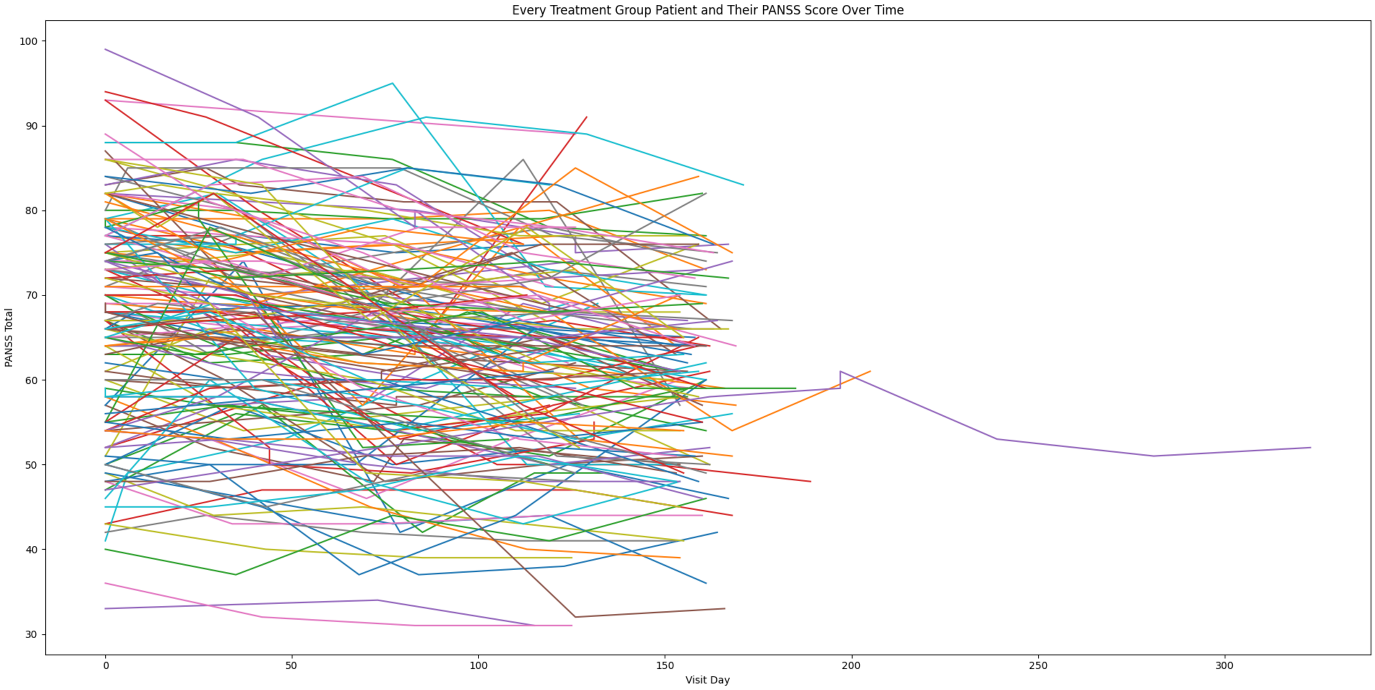

FIGURE 0.1. Treatment Group: Study E

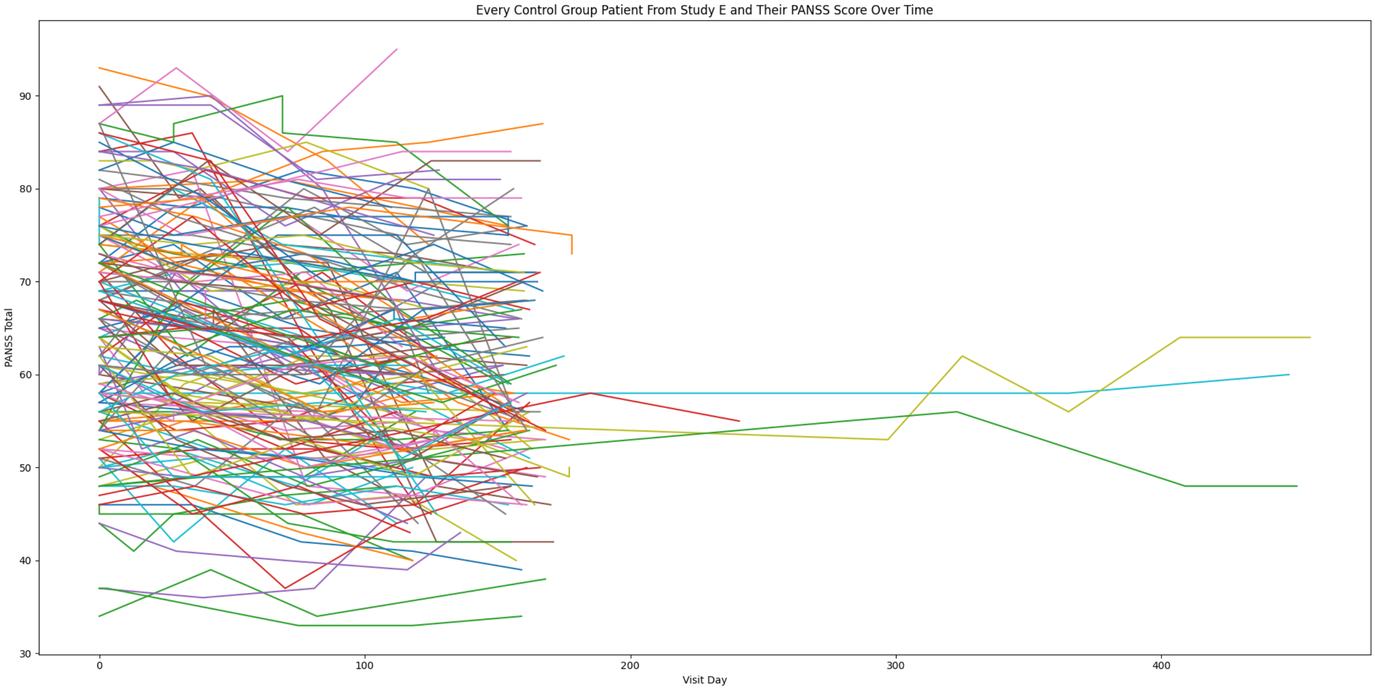

FIGURE 0.2. Control Group: Study E

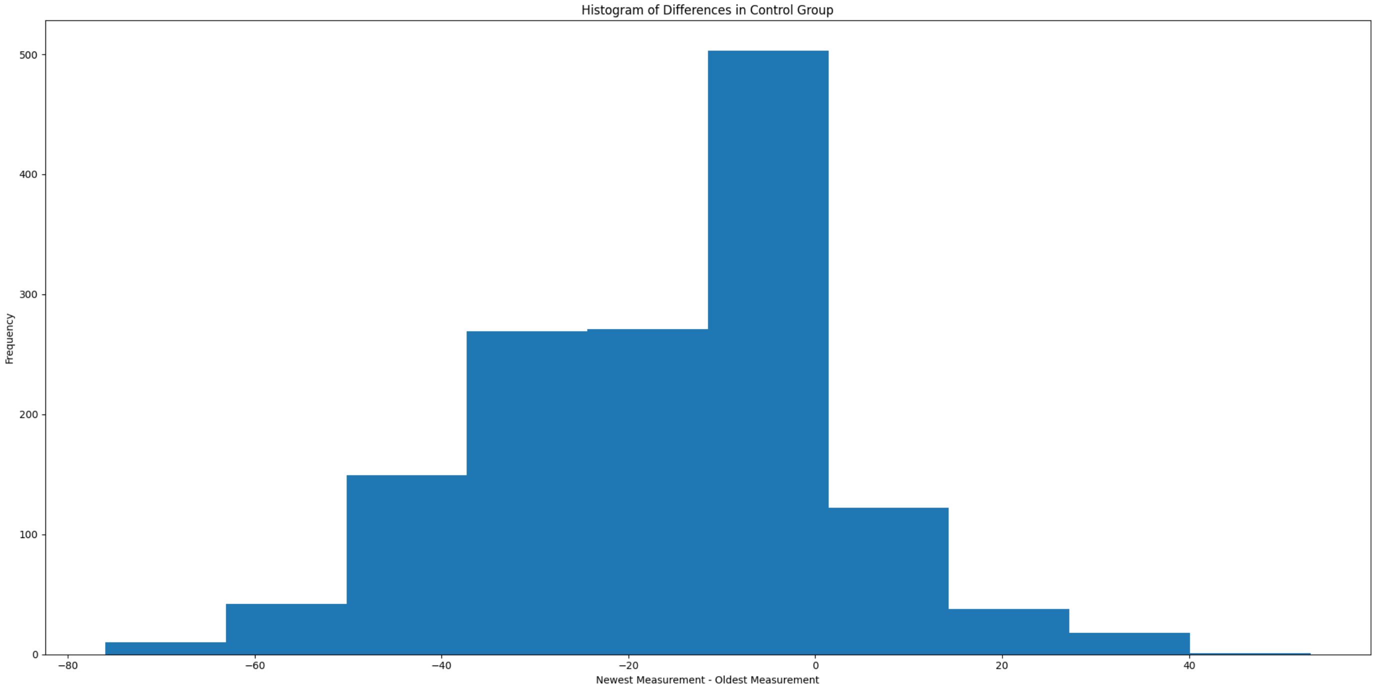

Figure 0.3. Differences in Control Group

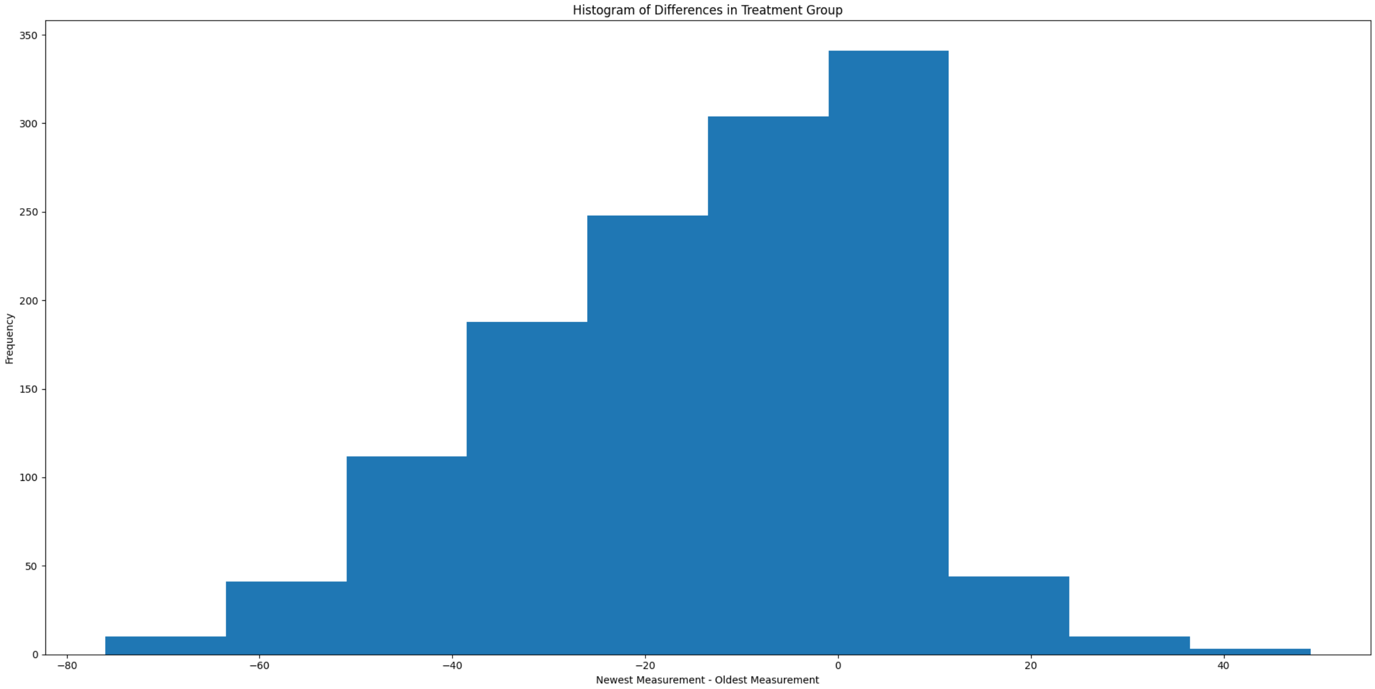

Figure 0.4. Differences in Treatment Group

The distribution of difference scores in the control group is somewhat normal where as the distribution of difference of scores in the treatment group is skewed.The mean difference of patients in the control group was -16.13 with a standard deviation of 19.40. That is on average, patients in the control group on average saw their PANSS score decrease by 16 at the end of the study. The mean difference of patients in the treatment group was -15.98 with a standard deviation of 19.24.

To see if the treatment had any affect, a t test will be conducted where H_0 is there is no significant difference in the mean change in score between the treatment and control groups and H_a is there is a significant difference in the mean change in score between the treatment and control groups. Patients were randomly assigned to either treatment or control. With $\alpha = 0.05$ a test statistic of 0.2105 and 2945 degress of freedom. The p value is 0.416. The result is not significant. The treatment affect does not make a significant impact on the decrease in PANSS Scores.

Part 2. **Patient Segmentation**

For this problem I used two forms of clustering. One being k means clustering, and the other being hierarchical clustering so I could see a dendogram. For K means clustering, I began by clustering on various components that made up the PANSS score. For example I tried P1 with N1, P1 with G1 etc. I choose 3 centroids. I also used cross validation to choose the best k. In the scatter plot below, it shows the locations of the centorids of the clusters when clustering using P1 and N1. The centroids are in orange. The blue dots are the xvalue combinations with P1 and N1.

I then wanted to know what a dendogram of this would look like, so one was constructed. There where two clusters. One thing I found quite interesting is that one of the clusters (green is N1) is larger than the other cluster. I thought that the clusters would be of similar size. Also the clusters don't meet for a while.

Part 3. **Forecasting**

Kaggle Reference. Team Name: Adam Kainikara.

Work done by Adam Kainikara.

There were many data wrangling steps involved in this part of the project. I used data from set E for this part. At first, I created a structured array which was organized by study, country, txgroup, assesment id, patient id, visit day, xvalues, panss score and lead status. Then I made a function that found every unique patient id. This was to figure out how many patients were in the study. After collecting every patient id, I created a blank dictionary. The dictionary would have the patient id as the key. For values, I stored all the associated values of each patient in an array. These associated values include: PANSS score, country, etc. For patients that had one or more visit days, the information would also be stored in arrays. It is stored in the same way as seen in section 1. Patients who did not have at least two visit days where removed from the study. In a spread sheet, I did some basic calculations. These calculations included figuring out the average % change for each patient from the initial reading in the study to the last reading in the study.

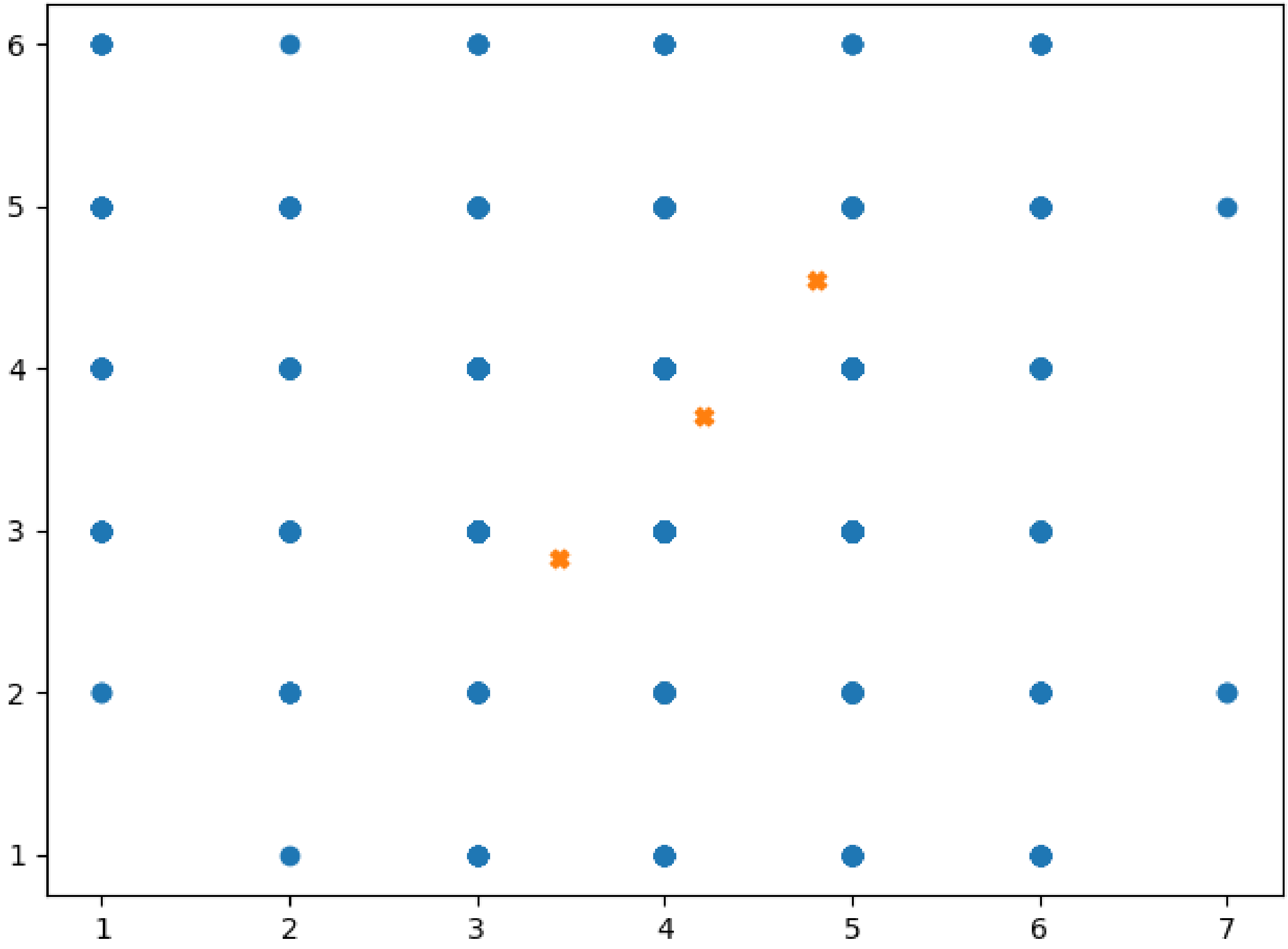

FIGURE 0.5. K Means Clustering

For my first prediction, I uploaded a CSV that contained the most recent visit day's PANSS score. I did this to see if the most recent visit PANSS score would be similar to the real 18th week score. This resulted in a score of 6.61736.

My next submission was the score decreased by the average decrease calculated in the spread sheet. To do this, I calculated each patients change in PANSS score from the first week to the last week. I found the average % drop across the data set. I then changed the initial reading (first day's reading) by the average % change.The average % change across the data set was about 10%. This did not change my score significantly.

For my final few submissions I created a function that did linear interpolation where:
$\alpha = \frac{y_2 - y_1}{x_2 - x_1}$ and
$\hat{y} = (1 - \alpha)y_1 + \alpha y_2$

The linear interpolation resulted in better results. My best score was 6.5099. In my created function, I tried several variations. The first variation I did was predict the average PANSS score if a patient did not get to day 129.5. I used day 129.5 as it is 18.5 weeks. In my spreadsheet I calculated that the average visit day turns into about 17.6 weeks. Although day

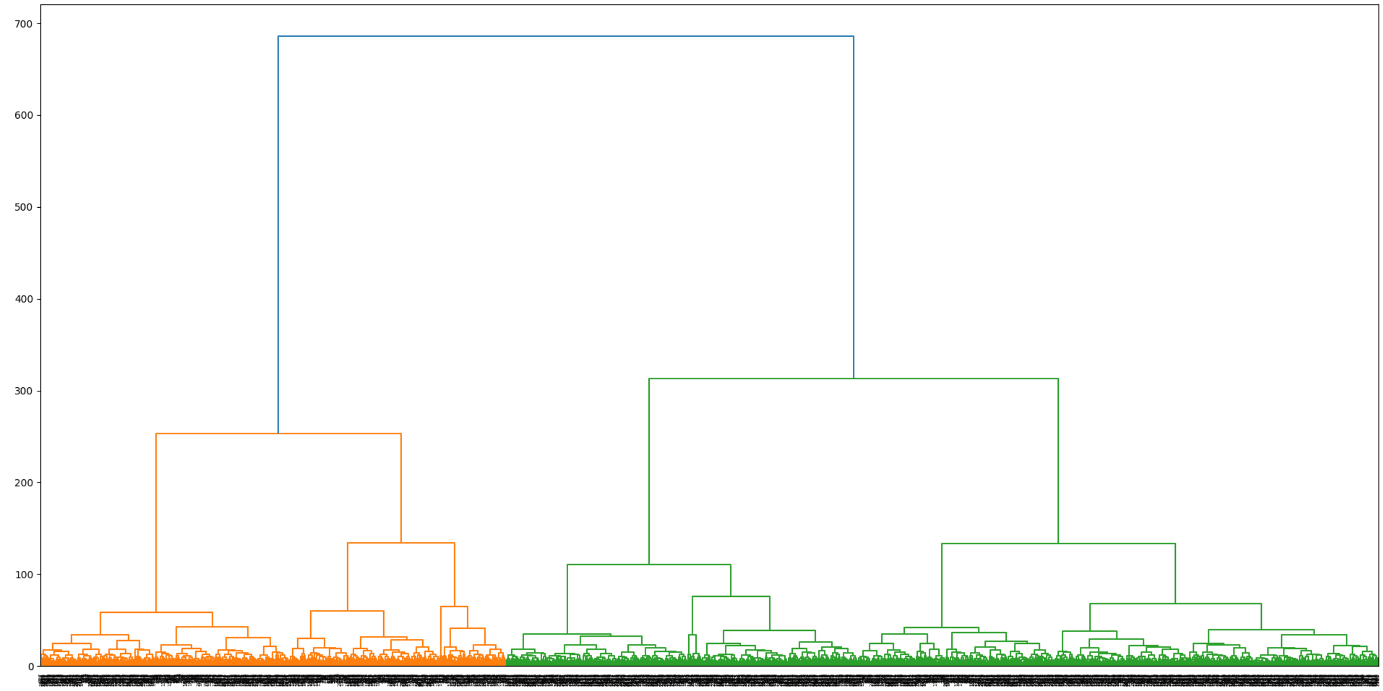

FIGURE 0.6. Dendogram

126 (18 weeks) does not necessarily mean that that was the patient's 18th week, I assumed that most patients would be around this time period. This assumption was made because the average patient stayed for 17.6 weeks.

Another variation I did if a patient did not make it to day 129.5 was to calculate the slope between the last two points, fit a line, and predict what the score for day 129.5 would have been.

The last attempt I did involved extrapolation. When I used extrapolation though, my score got worse (7.74936) so I did not use extrapolation. In the end using linear interpolation with the day as 129.5 days produced the best results for me. The code for this section is the function(s) desired_data in the interpol3.py file.

Part 4. 4 Binary Classification

Kaggle Reference

Team Name: Adam Kainikara.

Work done by Adam Kainikara.

There were many data wrangling steps involved in this part of the project. I used data sets A,B,C,D for training and E as the test set. At first, I created a structured array which was organized by study, country, txgroup, assesment id, patient id, visit day, xvalues, panss score and lead status. Then made a function that found every unique patient id. This was to figure out how many patients were in the study. After collecting every patient id, I created a blank dictionary. The dictionary would have the patient id as the key. For values, I stored all the associated values of each patient in an array. For patients that had one or more visit days, the information would also be stored. Patients who did not have at least two visit days where removed from the study.

For classification I used many classification methods. These include knn, naive bayes, decision trees, random forests and support vector machines. When using knn, I choose many different K values, however my accuracy was never very good so I decided not to use it. My score on kaggle was 16. The others (naive bayes, decision trees, random forests) produced poor training accuracy.

My best scores occurred while using a support vector machine. The support vector machine gave me a score of 0.858 with a training accuracy of 0.812.

To see if I could improve my score, I tried using different kernels such as linear or sigmoid however these did not improve my score. I did add a regularization parameter of 0.95. This changed my score to 0.854