PROJECT 2 REPORT

PART A:

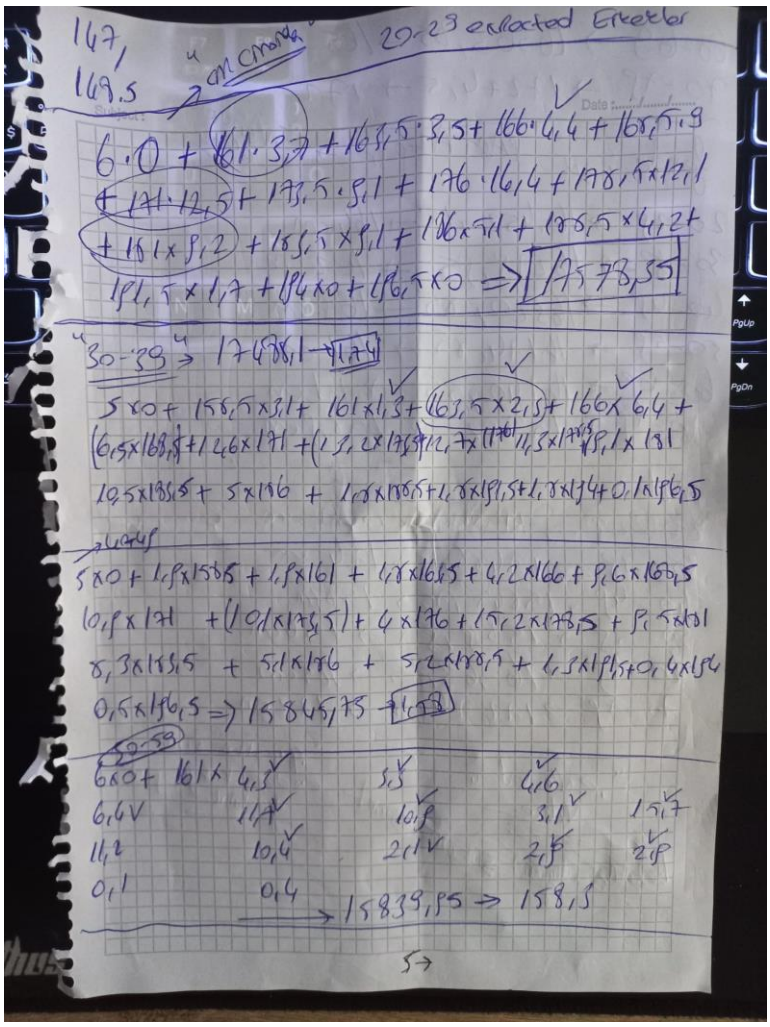
STEP 7)

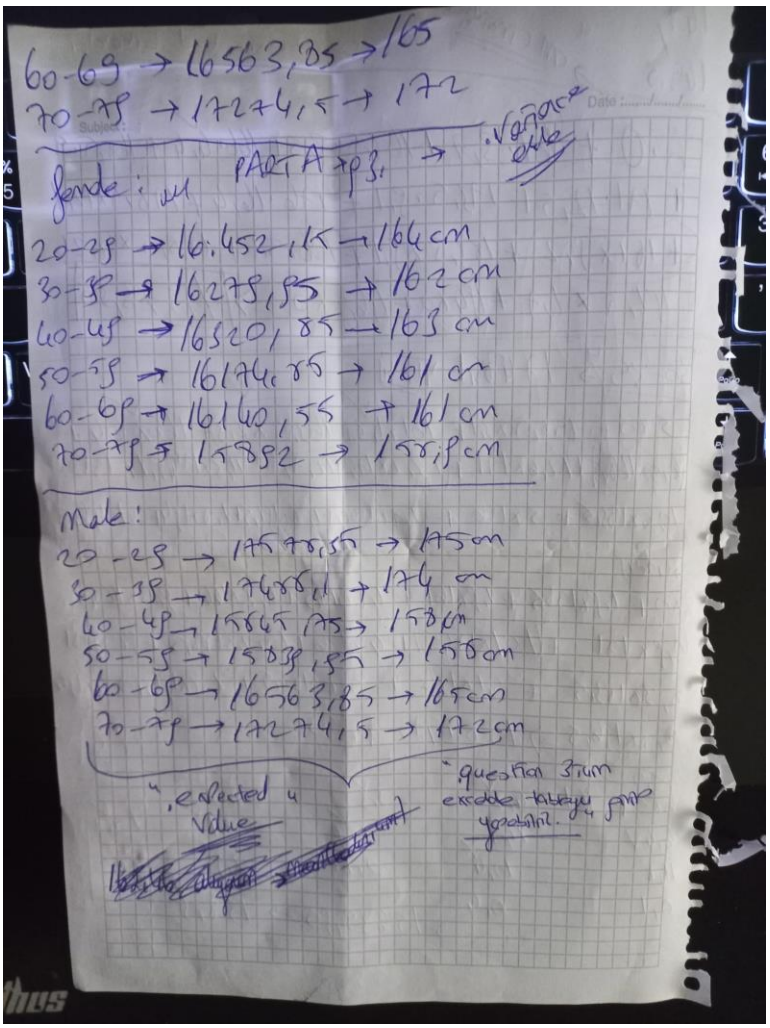
First of all, I will explain the project and what I have done in detail, and then I will explain Step 7. After step 7, I will explain step 6.

In Part A, when we look at the table given as a reference, the distribution of men and women in certain age ranges according to their height is given. Looking at this table, I first create the Probability Mass Functions.

- **Probability Mass Function (PMF):** The PMF of a probability distribution is expressed as a list of all possible values and the probabilities corresponding to each value. The PMF of a probability distribution can be used to directly calculate the probability corresponding to any value.
- In the first question, I created the Probability Mass function table, and for this, we first calculate its own probability percentages for each size range from the cumulative values of the values in the table. When we subtract the percentile in the next line from the percentile in the bottom line, we find the percentile between the heights it is attached to. Below I indicate the percentages that I calculated on the paper. Then I specified this table as a screenshot of the excel file in the last part of the data file.

Then I plot the graph of this table. For each age range, I take the average according to the specified height. I averaged each column but did not calculate the variance values for each column. I did these calculations myself first and checked my calculations using Python code. Below are my calculations for the averages:





- To calculate the average value of each column, I multiply the average of the heights it is attached to and the percentage of probability it has, and repeat this process until the end of the column and add up the results. After reaching the end of the column, we divide the total value by one hundred and calculate the average height value. I have shown the data in the data file.

- In step 3, I calculated the percentages corresponding to the heights. Now, in Step 4, to plot the height of men and women between the ages of 40-49, we need how many men and women in which heights, in which age range, that is, the percentage of men and women. For this, we have probability values in the whole age range by subtracting a lower value from an upper value in cumulative percentages in step 3. We calculate the p values by entering the calculated mean and standard deviation values, together with the N values, into the GraphPad T test calculator.

- In step 5, I draw a graph by looking at the percentages of men between the ages of 40-49 and 70-79 in their height ranges. At the same time, I calculate the P values by entering the mean and standard deviation values that I have created, together with the N values, into the system. I repeat the same process for women. I show this process on the last page of the data file.

Beginning Of Step 7:

In Steps 2 through 6 I gained experience working with real-life data on how to calculate the standard deviation and mean values. Below are the P values I found in Steps 4 and 5.

In fourth step:

- N=10 → P = 0.45; N=100 → P = 0.01; N=1000 → P = <0.0001

In fifth step:

- for MALE: N=10 → P = 0.03; N = 100 → P = <0.0001; N = 1000 → P = <0.0001
- for FEMALE: N=10 → P = 0.17; N = 100 → P = <0.0001; N = 1000 → P = <0.0001

- As we can clearly see above, when the sample size values, that is, the N values, increase, the P values, the realism of the result, increase. Therefore, it will be more realistic for us to achieve the test result we want to achieve, and we will not get it right by chance. The realism and precision of the result will increase. At the same time, we are more likely to reach a certain test result with real data in the operations we performed for men in the 5th step, compared to the operations in the 4th step.

- Statistical Significance value varies according to N, that is, sample size values. When we increase the sample size values, that is, increase the sample sizes, the P value gets closer to 0 and the result realism increases. Basically, the further the P value falls below 0, the higher the statistical significance value will be.

- As soon as the P value is more than 0.1 (for example, 0.1749, 0.4515), it is stated as “not statistically significant”. On the other hand, when the value is below 0.1 (for example, 0.0158,0.0330,0.029) it is specified as “statistically significant”. However, there is a measure of statistically significant values. When the P value falls below 0.0001, it is specified as “extremely significant”.

- At the same time, when the sample size values, that is, the N values, increase, the standard error of difference values get closer to 0 and the error rate is almost zero. As a result, the standard error parameter can never be reset, it just approaches 0.

- As a result, sample size values significantly affect statistical significance values, P values, and standard error of difference values.

STEP 6)

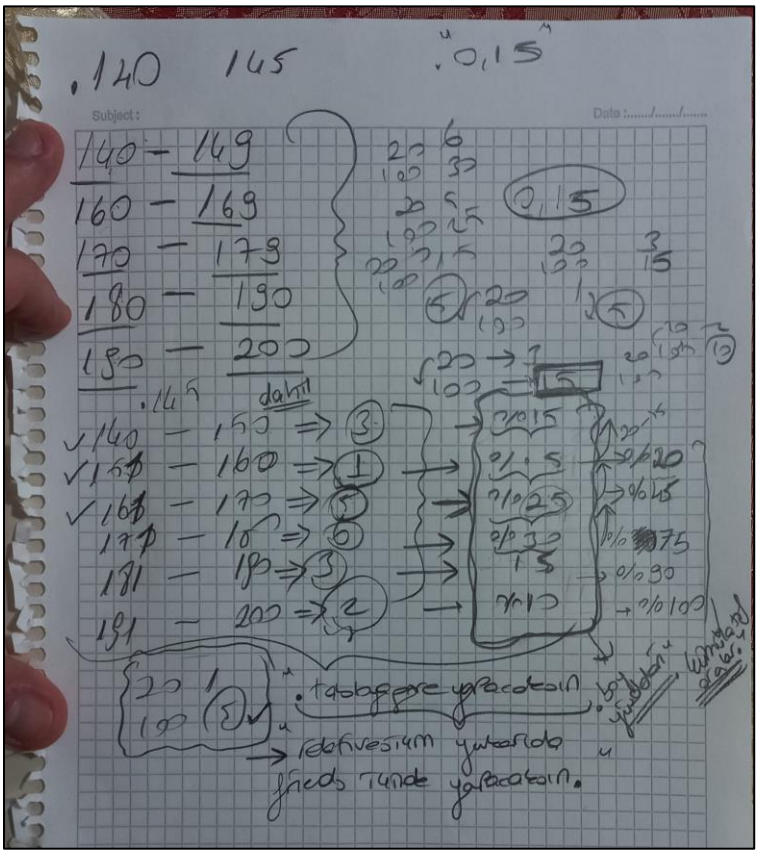
- In Part A, when we look at the questions asked in Steps 4 and 5, we can see that the T test is used to compare the results obtained in different groups with each other. When these different groups are compared, the mean values of the obtained data are compared with each other. Standard deviation and mean are two parameters used in comparing different groups with each other, that is, in the T test. As a result of T tests, the P value emerges. The P value used to obtain the result of a particular test statistic is an indicator of the realism of the test result. The lower the P-value, the more realistic and reliable the result.

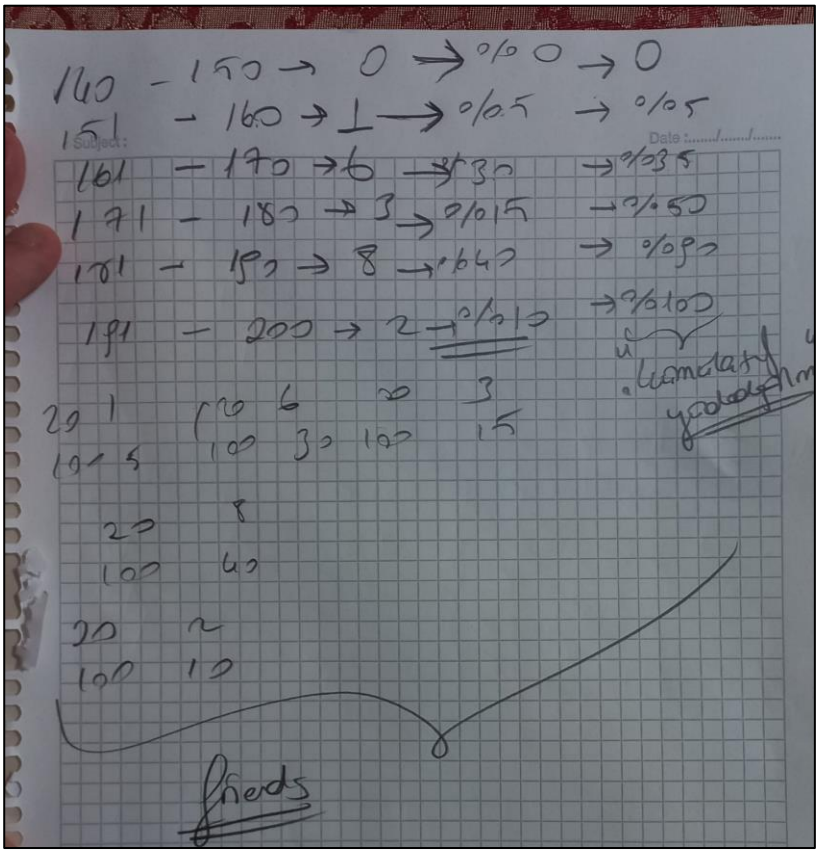
In Part A, I just did not calculate the variance values of all the columns in step 3. I think I have successfully completed all the remaining questions.

PART B:

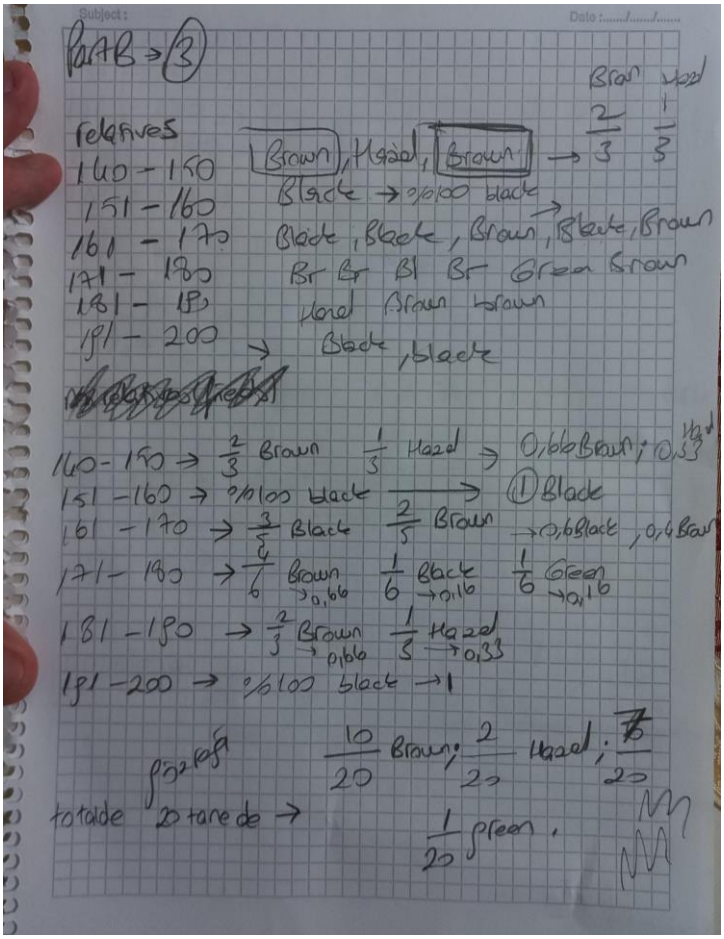
COMMENTS ON STEP 6:

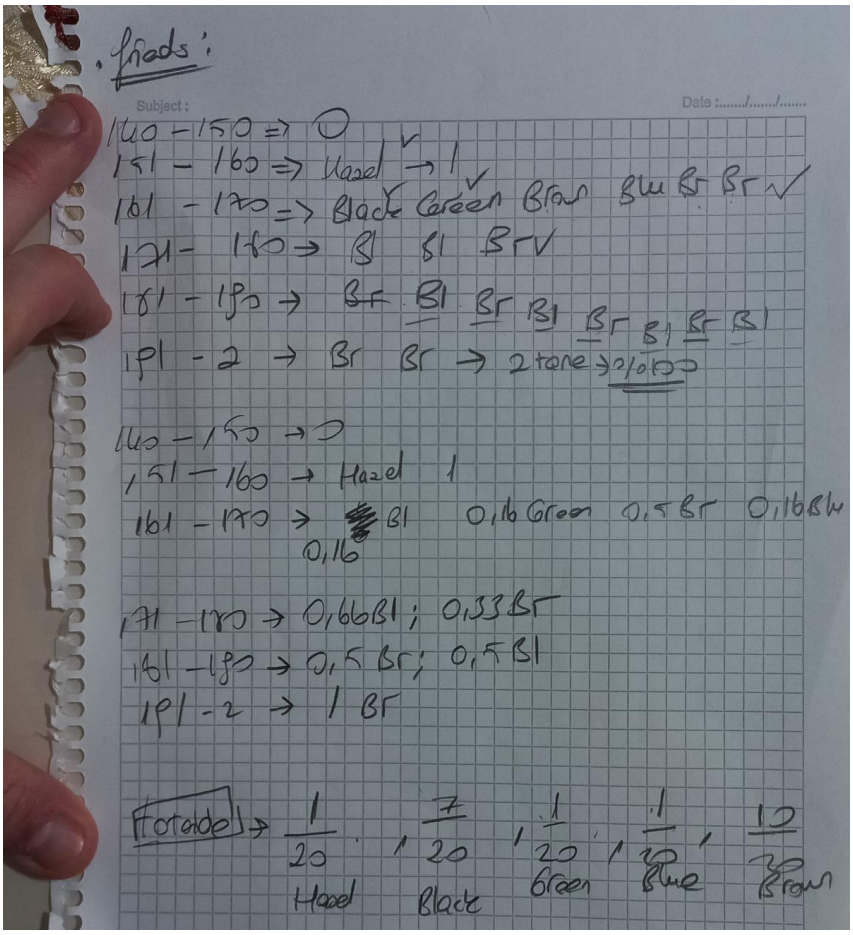
I specified the height ranges of relatives and friends and manually calculated the number of people in each height range, the probability percentage and the cumulative values. Then I specified these values in the data file as an excel snapshot. Relative values are shown in left side, friends values shown in right side.





After calculating the probability percentages in the height range, the eye color probabilities were calculated based on the height ranges. First of all, the percentages of eye colors of the relatives on the left and the percentages of eye colors of the friends on the right were calculated.





In the report file, I shared the calculations I made on paper and the operations I dealt with. These papers can be messy but I wanted to put my calculations on paper. Therefore, my studies about Project will be clearly shown. I shared all my calculations and operations in the data file.