Assignment 1
Machine Learning

Mohammad Hossein Basouli - 401222020

Theoretical Questions

Exercise 1

First, we assume that we are using z and t - tests for estimating the population mean.

Assumptions

- The z-test assumes that:
  - The sample size is large ($n > 30$).
  - Samples are chosen independently.
  - The population standard deviation is known.
  - The sample mean follows a normal distribution.
- The t-test assumes that:
  - The sample size does not need to be large.
  - Samples are chosen independently.
  - The population variance is unknown.
  - The statistic $\frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}}$ follows a t-distribution.

Population Standard Deviation

We discussed this earlier.

Use Cases

- z-test:
  - Used to check whether the sample mean differs significantly from the population mean when the population variance is known.
  - Used to check whether two populations differ significantly in their means when the variances of both populations are known.
- t-test:
  - Used to check whether the sample mean differs significantly from the population mean when the population variance is unknown.
  - Used to check whether two populations differ significantly in their means when the variances of both populations are unknown.

Exercise 2

As n goes beyond 30, from the Central Limit Theorem (CLT), we can say:

$$P(4.5 \leq \bar{X}) = P\left(\frac{4.5 - 4.2}{\frac{0.5}{\sqrt{30}}} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= P\left(\frac{0.3 \times \sqrt{30}}{0.5} \leq Z\right) = 0.00051904$$

Exercise 3

Scenario a)

We apply Pearson's correlation test to determine whether there is a linear relationship between Age and Annual Income. Pearson's correlation is the most suitable test for detecting linear relationships:

$$\text{correlation} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{(n-1)\sigma_x \sigma_y}$$

$$r \approx 0.98$$

$$\text{Correlation test: } t = r\sqrt{\frac{n-2}{1-r^2}}$$

$$p\text{-value} \approx 0.0035 < 0.05$$

Scenario b)

for monotonic relationships but not necessarily linear, Spearman's correlation test is the most appropriate:

$$\text{Spearman correlation} = \frac{\text{Cov}(R[x], R[y])}{\sigma_{R[x]} \cdot \sigma_{R[y]}}$$

$$r = -0.9$$

$$\text{Spearman correlation test}$$

$$p\text{-value} = 0.037 < 0.05$$

Exercise 4

- Key difference: The Mann-Whitney U test is used for comparing two independent samples, whereas the Wilcoxon signed-rank test is used for paired samples.
- Conclusion: We use the Mann-Whitney U test because the strategies being compared are independent.

Exercise 5

a We need to answer two questions:

(a) Are the residuals within each group normally distributed? In this case, they are, since the Shapiro-Wilk test yields a p-value of $0.43 > 0.05$.

(b) Is the variance equal across groups? According to Levene's test, the p-value is $0.86 > 0.05$, indicating homogeneity of variance.

Conclusion: Since both assumptions for ANOVA are met, we can perform the test. The ANOVA test gives a p-value of 1.4×10^{-9}, suggesting a significant difference between groups in terms of "Feature X".

b The results from part (a) confirm a significant difference between groups based on the p-value.

c The MANOVA test could be used to examine whether there are significant differences between groups across multiple dependent variables.

d If at least one feature contributes significantly, we can decide between ANOVA and the Kruskal-Wallis test. Each variable must be analyzed separately to determine if it significantly differentiates the groups. If it does, it can be considered as a feature.

Practical Questions

Sleep Health

1. **Exploratory Data Analysis (EDA)**:

   - **Number of rows (data samples) and columns (features) are in the dataset**: (374, 13)
   - **Values for each categorical feature**:
     - **Gender**:

       ```
       ['Male' 'Female']
       ```

     - **Occupation**:

       ```
       ['Software Engineer' 'Doctor' 'Sales Representative' 'Teacher' 'Nurse'
       'Engineer' 'Accountant' 'Scientist' 'Lawyer' 'Salesperson' 'Manager']
       ```

     - **BMI Category**:

       ```
       ['Overweight' 'Normal' 'Obese' 'Normal Weight']
       ```

     - **Sleep Disorder**:

       ```
       [NaN 'Sleep Apnea' 'Insomnia']
       ```

- **Description of numerical features**:

	Person ID	Age	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	Heart Rate	Daily Steps
count	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000
mean	187.500000	42.184492	7.132086	7.312834	59.171123	5.385027	70.165775	6816.844920
std	108.108742	8.673133	0.795657	1.196956	20.830804	1.774526	4.135676	1617.915679
min	1.000000	27.000000	5.800000	4.000000	30.000000	3.000000	65.000000	3000.000000
25%	94.250000	35.250000	6.400000	6.000000	45.000000	4.000000	68.000000	5600.000000
50%	187.500000	43.000000	7.200000	7.000000	60.000000	5.000000	70.000000	7000.000000
75%	280.750000	50.000000	7.800000	8.000000	75.000000	7.000000	72.000000	8000.000000
max	374.000000	59.000000	8.500000	9.000000	90.000000	8.000000	86.000000	10000.000000

Figure 1: Description of numerical features

- **Datatype of each feature**:

```
#   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Person ID                374 non-null    int64
 1   Gender                   374 non-null    object
 2   Age                      374 non-null    int64
 3   Occupation               374 non-null    object
 4   Sleep Duration           374 non-null    float64
 5   Quality of Sleep         374 non-null    int64
 6   Physical Activity Level  374 non-null    int64
 7   Stress Level             374 non-null    int64
 8   BMI Category             374 non-null    object
 9   Blood Pressure           374 non-null    object
 10  Heart Rate               374 non-null    int64
 11  Daily Steps              374 non-null    int64
 12  Sleep Disorder           155 non-null    object
dtypes: float64(1), int64(7), object(5)
```

- **Data pre-processing**:
  - We begin Data pre-processing, by first looking at what columns have NaN entries in them and how many there are: *From the last bullet, we can see that 374 - 155 = 219 entries in the column, 'Sleep Disorder' are filled with NaN, and the are the only NaN entries in our dataset.*
  - Next, we divide the dataset into three different groups, according to the range of values that features 'Sleep Disorder' can take on, including NaN values. If we plot the distribution of all of the features for each of the groups, we would end up with something like Figure 2:

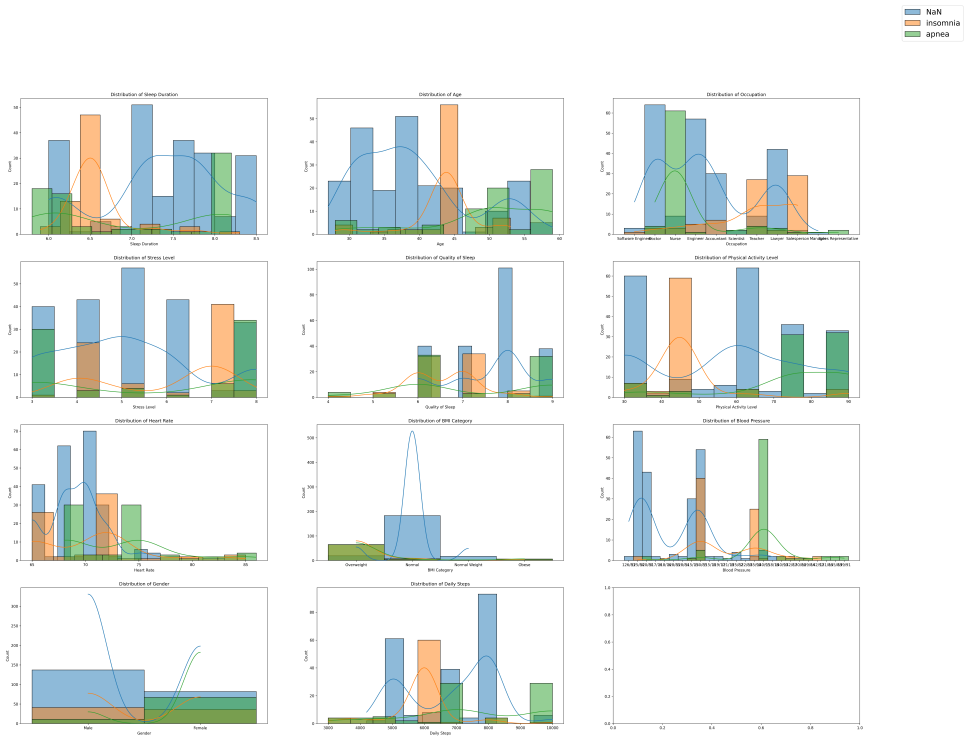


Figure 2: Plot of features for each of the groups (zoom-in)

- Figure 2 roughly shows that, people with NaN value in the column 'Sleep Disorder' are people who don't have any Sleep Disorder and are healthy and OK.
- Some of the arguments that would lead us to this conclusion that people with NaN values don't have any Sleep Disorder are as follows:
  (a) They roughly have an even distribution of 'Sleep Duration' over the range between 7 and 8 hours. Which makes sense, because healthy people have a similar distribution of Sleep Duration as well.
  (b) Most of the them (80%), have Normal BMI.
  (c) They roughly have an even distribution of 'Heart Rate' over the range between 65 and 75. Which matches with what we would expect from healthy people.
  (d) They also have an even distribution of 'Stress Level', and their distribution is not skew to one direction. which makes sense for normal people.

2. **Hypothesis Testing**:

   (a) Does women's sleep duration follow a normal distribution?

   - Answer based on visualization (Figure 3): *No, it doesn't seem to be normally distributed*

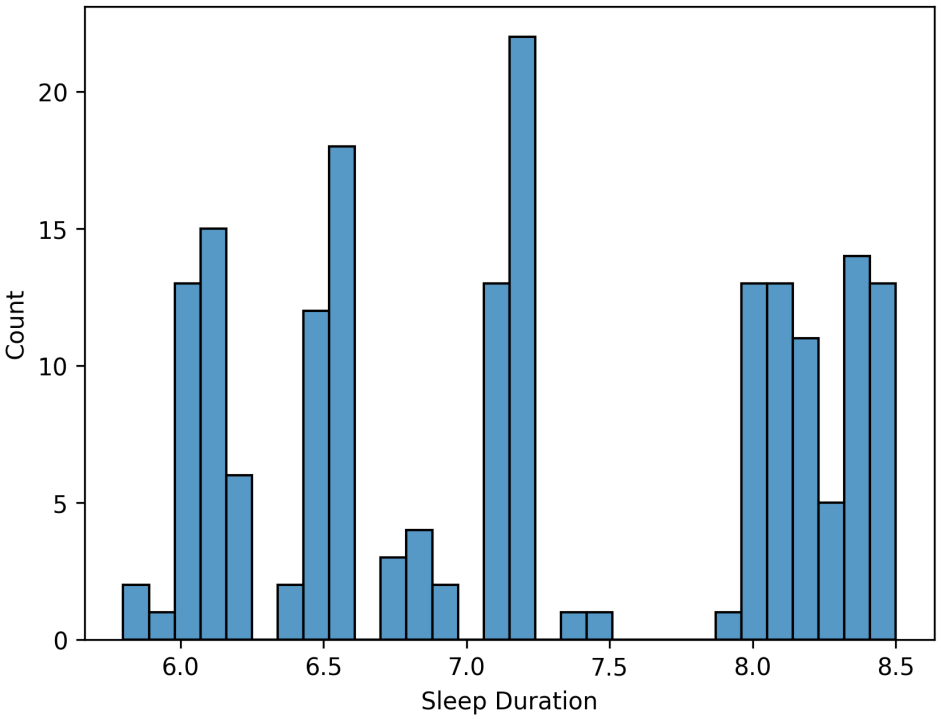


Figure 3: Disribution of **Sleep Duration** for Women

- Answer based on hypothesis test: *No, the p-value is so small, so we reject hypothesis that distribution is normal.*

(b) Is having higher daily steps a contributing factor into better sleep? Check the corresponding correlation of Daily Steps and Quality of Sleep.

- First, we have to check to see what correlation test have to use for these two variables.
  - If we look at the scatter plot of these two variables, we can see that the relationship between these two variables is not linear; So we head over to the Spearman Correlation test.

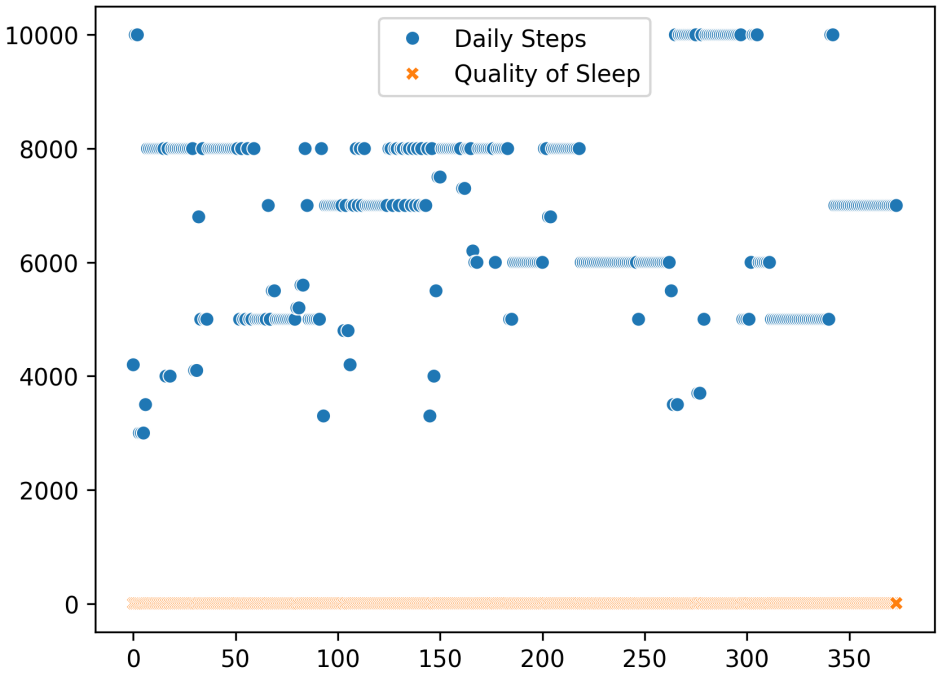


Figure 4: Scatter plot of two variables, **Daily Steps and Quality of Sleep**

- If we run Spearman Correlation test, we would get the following result: Spearman Correlation Coefficient: 0.022779213378418182, P-value: 0.6605808344543287. Which shows that, there is no significant correlation between these two features.

(c) Is stress level different among different occupations? First, check this hypothesis with a test, and then compute the average stress level among different occupations. Use a bar chart or any other desired visualization method to demonstrate the result.

- Answer based on visualization (Figure 5): *Yes, it really seems like that we have a significant difference in* ***stress level*** *in terms of different* ***Occupations****. e.g. we can easily see that stress level of salespersons is only distributed around 7 which is really high, but the stress level of most of the engineers is about 3 which is really low.*

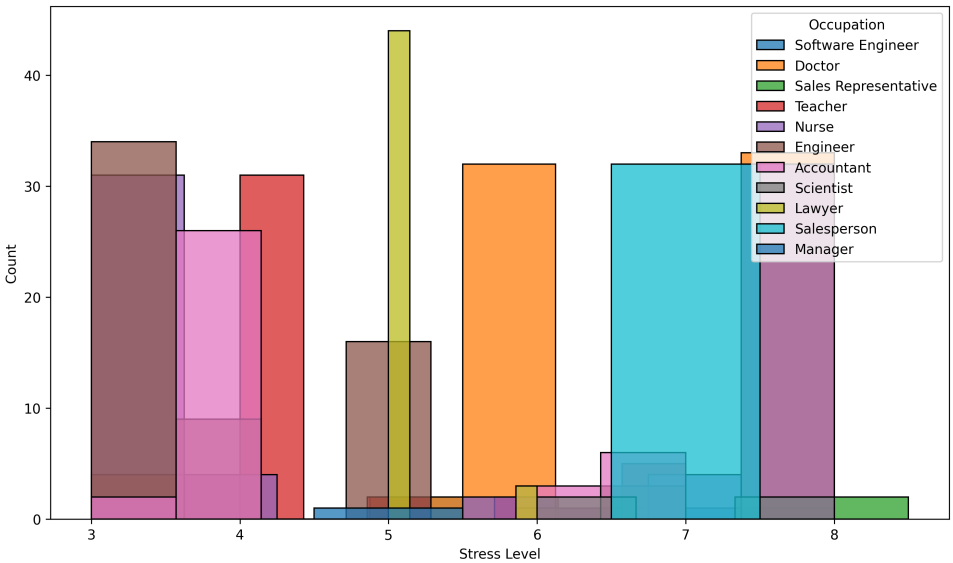


Figure 5: Disribution of **Stress Level** for different **Occupations**

- Answer based on hypothesis test: *If we run ANOVA one-way test, we get results as*

follows: F-statistic: 21.63598878521177 P-value: 1.3550912313042 78e-31. So we can say that, the different groups of occupation differ significantly in terms stress level.

(d) Are different BMI categories significantly different given their blood pres- sure? (Hint: Convert blood pressure into two columns and apply your test given these new two features.)

- Answer based on hypothesis test (I have divided the Blood Pressure column into two separate columns, Systolic Pressure and Diastolic Pressure.): *If we run MANOVA test, we get results represented in Figure 6. Which shows that we have a significant difference between BMI categories in terms of Blood Pressure.*

```
                  Multivariate linear model
================================================================

----------------------------------------------------------------
       Intercept         Value   Num DF  Den DF   F Value   Pr > F
----------------------------------------------------------------
          Wilks' lambda   0.0029 2.0000 369.0000 63410.2541 0.0000
         Pillai's trace   0.9971 2.0000 369.0000 63410.2541 0.0000
 Hotelling-Lawley trace 343.6870 2.0000 369.0000 63410.2541 0.0000
    Roy's greatest root 343.6870 2.0000 369.0000 63410.2541 0.0000
----------------------------------------------------------------

----------------------------------------------------------------
       C(BMI_Category)     Value  Num DF  Den DF  F Value  Pr > F
----------------------------------------------------------------
             Wilks' lambda 0.3412 6.0000 738.0000  87.5857 0.0000
            Pillai's trace 0.7556 6.0000 740.0000  74.8918 0.0000
    Hotelling-Lawley trace 1.6475 6.0000 490.2262 101.1869 0.0000
       Roy's greatest root 1.4522 3.0000 370.0000 179.1029 0.0000
================================================================
```

Figure 6: MANOVA test results for checking significancy of difference between different groups of BMI in terms of their Blood Pressure.

(e) Do people with sleep disorders have higher heart rates than those without any sleep disorder?

- Answer based on visualization (Figure 7): *Yes, based on the plot below we can say that, the people with Sleep Disorder have higher heart rates usually.*

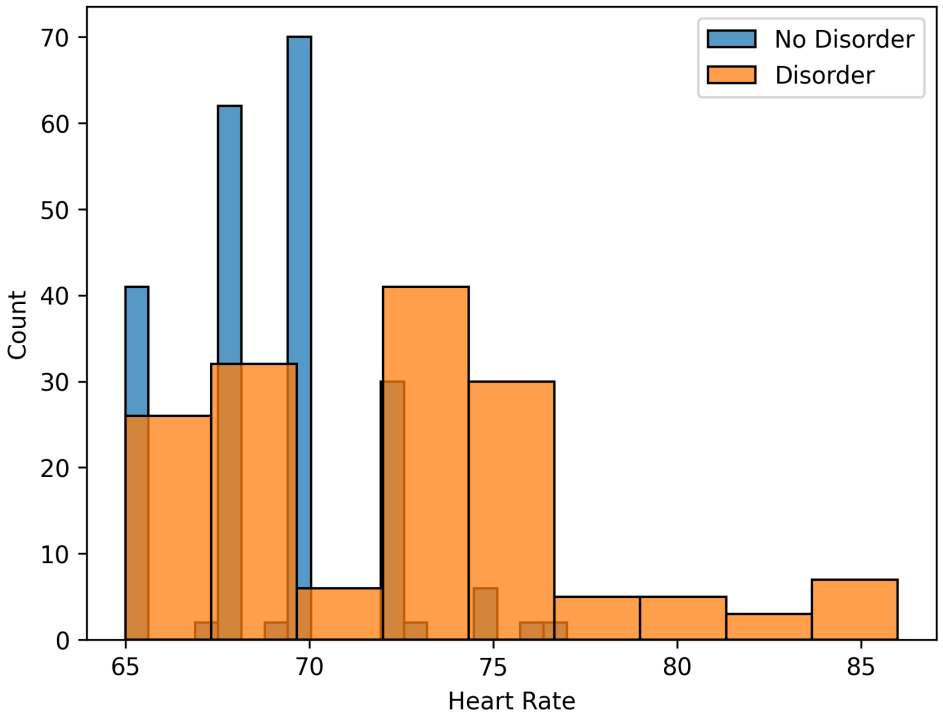


Figure 7: Distribution of Heart Rate for two groups, people with and without Sleep Disorder.

- Answer based on hypothesis test: *After performing ANOVA one-way test, we get the following results: F-statistic: 45.5400838475208, P-value: 5.749587721474387e-11, So we can say that, There are significant differences between the groups.*

Sleep Health

1. **Exploratory Data Analysis (EDA)**:

   - **Number of rows (data samples) and columns (features) are in the dataset**: (6607, 20)
   - **Values for each categorical feature**:
     - **Parental_Involvement**:

       ```
       ['Low' 'Medium' 'High']
       ```

     - **Access_to_Resources**:

       ```
       ['High' 'Medium' 'Low']
       ```

     - **Extracurricular_Activities**:

       ```
       ['No' 'Yes']
       ```

     - **Motivation_Level**:

       ```
       ['Low' 'Medium' 'High']
       ```

     - **Internet_Access**:

['Yes' 'No']

- **Family_Income**:
  ['Low' 'Medium' 'High']

- **Teacher_Quality**:
  ['Medium' 'High' 'Low' nan]

- **School_Type**:
  ['Public' 'Private']

- **Learning_Disabilities**:
  ['Positive' 'Negative' 'Neutral']

- **Parental_Education_Level**:
  ['High School' 'College' 'Postgraduate' nan]

- **Distance_from_Home**:
  ['Near' 'Moderate' 'Far' nan]

- **Gender**:
  ['Near' 'Moderate' 'Far' nan]

- **Description of numerical features**:

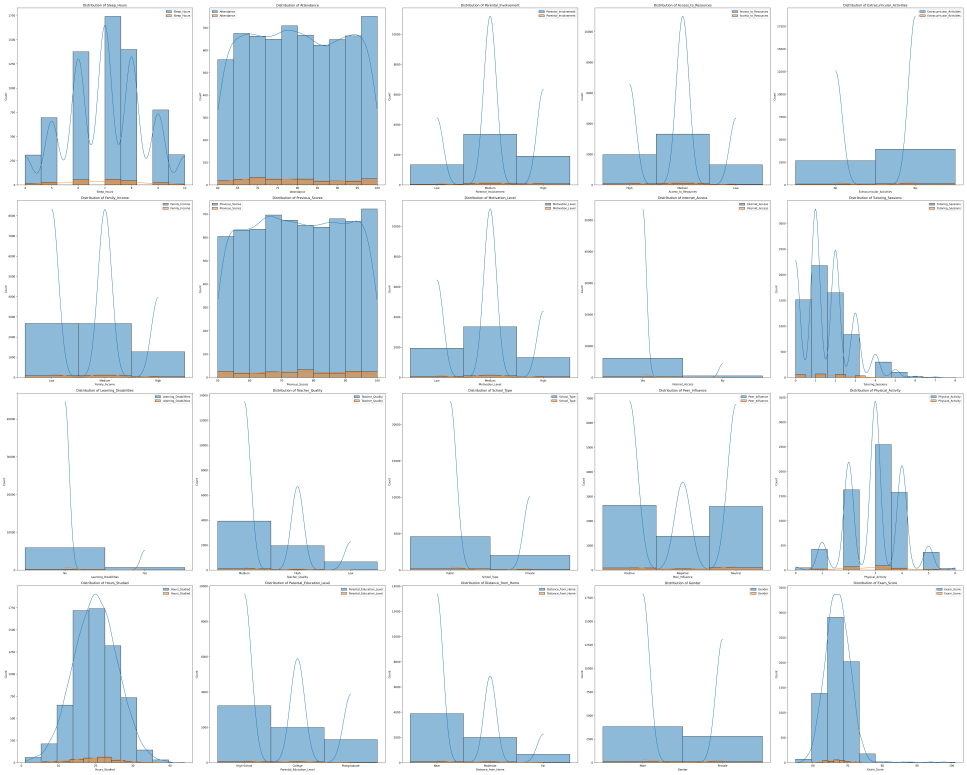


Figure 8: Description of numerical features

- **Datatype of each feature**:

```
#   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Hours_Studied               6607 non-null   int64
 1   Attendance                  6607 non-null   int64
 2   Parental_Involvement        6607 non-null   object
 3   Access_to_Resources         6607 non-null   object
 4   Extracurricular_Activities  6607 non-null   object
 5   Sleep_Hours                 6607 non-null   int64
 6   Previous_Scores             6607 non-null   int64
 7   Motivation_Level            6607 non-null   object
 8   Internet_Access             6607 non-null   object
 9   Tutoring_Sessions           6607 non-null   int64
 10  Family_Income               6607 non-null   object
 11  Teacher_Quality             6529 non-null   object
 12  School_Type                 6607 non-null   object
 13  Peer_Influence              6607 non-null   object
 14  Physical_Activity           6607 non-null   int64
 15  Learning_Disabilities       6607 non-null   object
 16  Parental_Education_Level    6517 non-null   object
 17  Distance_from_Home          6540 non-null   object
 18  Gender                      6607 non-null   object
 19  Exam_Score                  6607 non-null   int64
dtypes: int64(7), object(13)
```

- **Data pre-processing**:

– We begin Data pre-processing, by first looking at the distribution of columns for students who have some missing values and those who don't have any.: *Based on the distribution of each column, showing on Figure 9, we can see that, the students who don't have any missing values(the blue bars) and the ones who have some missing entries(the yellow ones) are approximately the same. Because the distribution of students who have some missing values is evenly distributed over the whole range of values for the original distribution (distribution of students who don't have any missing values).*

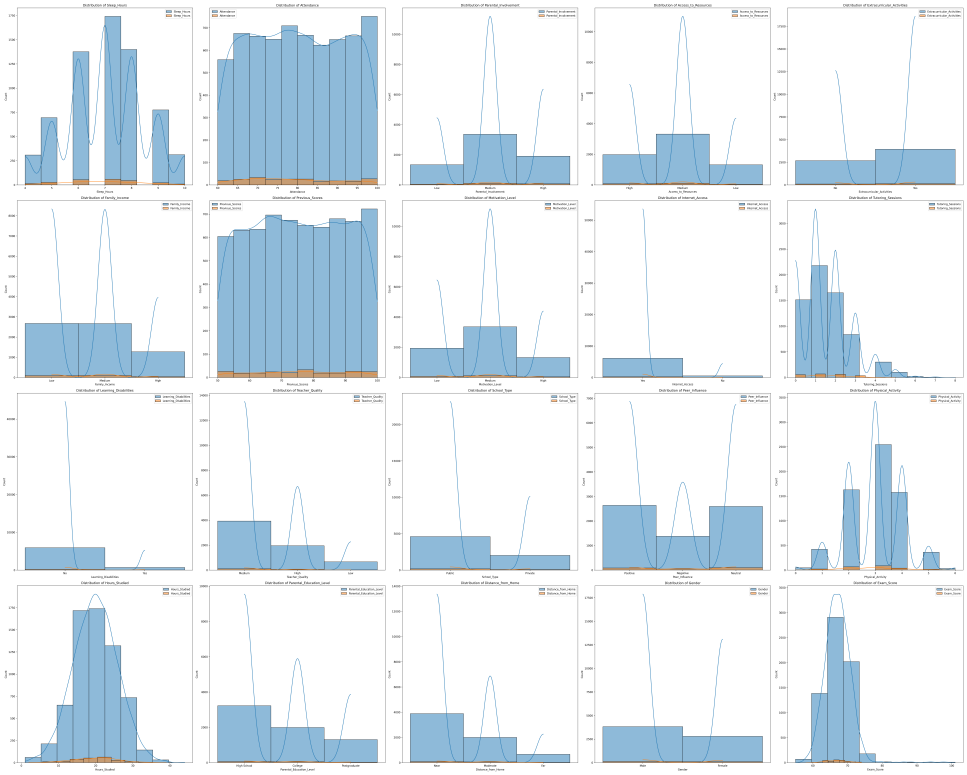


Figure 9: Plot of features for each of the groups (zoom-in)

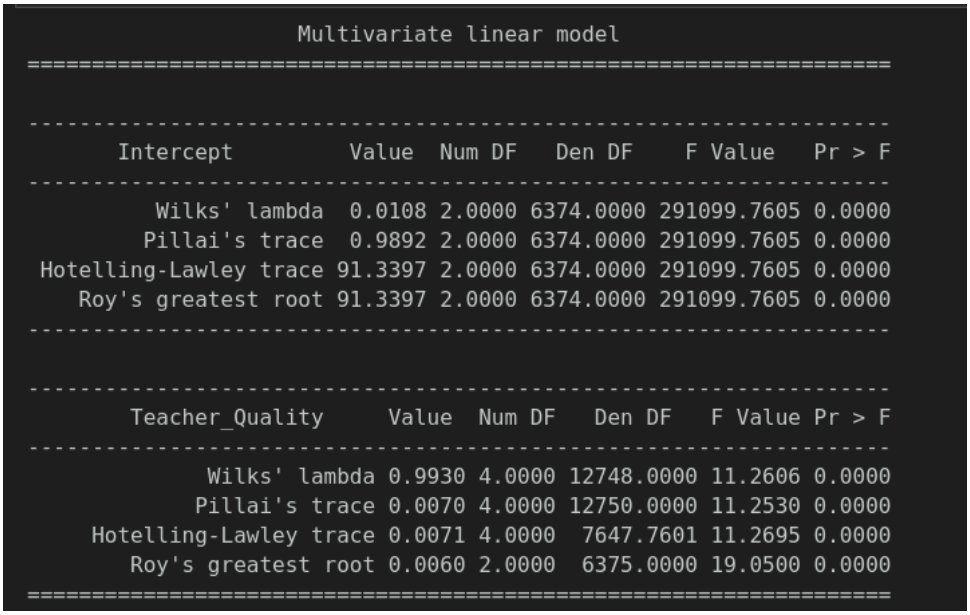
– Decission on the rows which have some missing values: *Based on the evidences given above, we can say that the students who have some missing values are approximately the same as the ones who don't have any missing entries. So we can simply drop them out of the dataset.*

2. Hypothesis Testing:

a Is Sleep_Hours correlated with Exam_Score? *If we perform Pearson Correlation Test, we get the following results: Pearson Coefficient: -0.0171714462163484 7 P-value is: 0.17031732819959972. So we can say that Sleep_Hours and Exam_Score are not possibly correlated*

b Does School_Type affect Exam_Score? *If we perform ANOVA one-way test, we get the following results: f_stat is: 0.7531963873423186, p_value is: 0.3854987810260675. So we can say that School_Type doesn't significantly determine Exam_Score*

c Does Teacher_Quality affect Previous_Scores or Exam_Score? *if we perform MANOVA test, we get the results shown in Figure 10, we can say there, there is a significant differnce in atleast one of the groups. So we perform two ANOVA one-way separate tests to see where is the difference.*



Multivariate linear model

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.0108	2.0000	6374.0000	291099.7605	0.0000
Pillai's trace	0.9892	2.0000	6374.0000	291099.7605	0.0000
Hotelling-Lawley trace	91.3397	2.0000	6374.0000	291099.7605	0.0000
Roy's greatest root	91.3397	2.0000	6374.0000	291099.7605	0.0000

Teacher_Quality	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.9930	4.0000	12748.0000	11.2606	0.0000
Pillai's trace	0.0070	4.0000	12750.0000	11.2530	0.0000
Hotelling-Lawley trace	0.0071	4.0000	7647.7601	11.2695	0.0000
Roy's greatest root	0.0060	2.0000	6375.0000	19.0500	0.0000

Figure 10: results of MANOVA test for different categories of Teacher_Quality

- ANOVA one-way test results for Exam_Score and Teacher_Quality: F=18.597, p=0.000. So Teacher_Quality significantly affect Exam_Score.
- ANOVA one-way test results for Previous_Score and Teacher_Quality: Previous Score ANOVA: F=3.494, p=0.030 . So Teacher_Quality significantly affect Previous_Score.

d Do Previous_Scores influence Exam_Score? *We perform the Pearson Correlation test and get the following results: Correlation: 0.19111637222616798, p-value is: 1.5821805765966484e-53. Previous_Scores and Exam_Score are possibly correlated*

e Is Sleep_Hours normally distributed? *We perform KS test to see whether the distribution of Sleep_Hours is normal or not. we get the following results: KS Test Statistic: 0.132, p-value: 0.000. So we can say that, Data does not follow a normal distribution*