Maryam Soleimani 401222075

# Theoretical Questions:

1 )

| | T-Test | Z-Test |
|---|---|---|
| Assumptions | Normally distributed data, approximate normality | Normally distributed data, known population standard deviation |
| Population Standard Deviation | Unknown | Known |
| Use Case | Small sample analysis, comparing means between groups | Large sample analysis, population mean comparisons |

Z-tests are used when the population variance is known and the sample size is large while t-tests are used when the population variance is unknown and the sample size is small.

2)

CLT tells us that if we take a large enough sample from any population even if the individual ratings are not distributed normal the distribution of the sample mean will tend to be approximately normal. This lets us to find confidence intervals and perform hypothesis tests using the normal distribution  even when the data isn't normal.

$Z = (X - \mu) / \sigma$

$(4.5 - 4.2) / 0.5 = 0.6$

$P(X>4.5)=P(Z>0.6)$

$P(Z>0.6)=1-0.7257=0.2743$ --------> This means there is approximately a 27.43% chance that a random selected user will rate the song more than 4.5.

3)

a) because both Age and Annual Income are continuous variables and we want a linear relationship Pearson's Correlation is appropriate.

Pearson's correlation coefficient (r) measures the strength and direction of a linear relationship between two continuous variables.

( the rest are in paper)

b) Customer Satisfaction Score is ordinal and Age is continuous. When we want to know if the relationship is monotonic Spearman's Rank Correlation is appropriate.

Spearman's correlation measures how well the relationship between two variables can be described by a monotonic function using ranks. ( the rest are in paper)

3a)

Mean age: $\bar{x} = \frac{25+32+40+50+60}{5} = 41.4$

Mean income: $\bar{y} = \frac{10^3(40+55+65+70+85)}{5} = 10^3 \times 63$

For each customer we calculate $(x_i - \bar{x})(y_i - \bar{y})$ & then sum them:

C1: $-16.4 \times -23 \times 10^3 = 377200$

C2: $-9.4 \times -8000 = 75200$

C3: $-1.4 \times 2000 = -2800$

C4: $8.6 \times 7000 = 60200$

C5: $18.6 \times 22000 = 409.200$

$+ = 919000$ (sum)

now we calculate the denominator:

Age: $(25-41.4)^2 + (32-41.4)^2 + (90-41.4)^2 + (50-41.4)^2 + (60-41.4)^2 = 779.20$

$\sqrt{779.20} = 27.91$



Income. $10^6 \left( (40-63)^2 + (55-63)^2 + (65-63)^2 + (70-63)^2 + (85-63)^2 \right) = 1130 \times 10^6$

$\sqrt{1130 \times 10^6} = 33638$

denominator $= 27.91 \times 33638 = 938836$

$r = \frac{919000}{938836} = 0.979$ → since it's close to 1

it shows a strong + linier rel between age & Income

Test statistic for peterson's correlation:

$t = r \times \sqrt{\frac{n-2}{1-r^2}} = 0.979 \times \sqrt{\frac{3}{0.042}} = 8.27$

3 degrees freedom → the t value shows a significant linear rel → p value very low

Interpretation: There is a very strong + linear relation between age & Annual income among the customers → As age increases & Income increases too.

Eg 3b)

Assign rank for age:

C1: R1

⋮

C6: R5

Assign rank for satisfaction score:

C5: R1

C4: R2

C2: R3

C3: R4

C1: R5

$\sum di^2 = 16+1+1+9+16 = 38$

$rs = 1 - \frac{6\sum di^2}{n(n^2-1)} = 1 - \frac{6\times38}{5(25-1)} = 1-1.9 =$ (-0.9)



This value shows a strong - monotric rel.

Interpretation: There's a strong negative monotric rel between age & customer satisfaction score. In this dataset as age increases the cus sat decreases.

4) The Wilcoxon Test is used to compare two dependent samples and the Mann-Whitney U Test is for two independent samples. Both tests are non-parametric and are used when the data has no normal distribution.

If we have two independent groups like strategy A on one group of customers and strategy B on another separate group we must use the Mann-Whitney U test.

If we have paired data like the same group of customers exposed to both strategies at different times we must use the Wilcoxon Signed-Rank test.

So the choice depends on if the data comes from the same things measured twice (Wilcoxon) or different groups of things (Mann-Whitney).

5)

a) Before choosing a test we need to find out if the data for feature X in each group is normally distributed (and if the variances are similar). we can do this by running a normality test such as the Shapiro-Wilk test on feature X for each group

If the normality assumption holds (and variances are roughly equal)then ANOVA is appropriate.

If normality is violated the Kruskal-Wallis testa non-parametric alternative should be used.

Applying the Test (Assuming Normality):
i assumed that we've checked normality for feature X and the data appears normal. Then we use ANOVA.

Our data for feature X are:

- Group 1: 10.2, 10.8, 11.0
- Group 2: 18.5, 17.9, 18.2
- Group 3: 30.1, 29.8, 30.3

( computations in paper)

Because this F-value is high the associated p-value will be to much less than 0.05.
We reject the null hypothesis that the groups have the same mean for feature X concluding that the groups are different.

b)assuming that the normality assumption holds ANOVA is the appropriate test. Because ANOVA is designed to compare the means of three or more groups with the assumption of normality. As computed above the ANOVA shows a highly significant F-statistic (2618), showing that there is a statistically difference in feature X among the groups ($p < 0.05$).

c) Multivariate Analysis of Variance (MANOVA Because it allows at-the-same-time testing of multiple dependent variables to see if the mean vectors are different between groups.

d) Running ANOVA (or Kruskal-Wallis if normality is violated) on each feature separately to see if there are statistically significant differences among the groups.

5a)

calculate group means & overall means:

Mean (G1) = 10.67

" (G2) = 18.20

" (G3) = 30.97

overall mean = (10.2 + 10.8 + 11 + 18.5 + 17.9 + 18.2 + 30.7 + 29.8 + 30.3) / 9 = 19.64

Compute SSB for each group:

$n \times (\text{Group mean} - \text{overall mean})^2$

G1: $3 \times 80.6 = 241.8$

G2: $3 \times 2.99 = 6.3$

G3: $3 \times 108.6 = 325.8$

Total SSB = 241.8 + 6.3 + 325.8 = 573.9





Comput SSW for each group:

G1: 0.35
G2: 0.18
G3: 0.13

+ = 0.66 = total SSW

Calculate mean squares & F-statistic:

degree of freedom:

between = k - 1 = 3 - 1 = 2

within = N k = 9 - 3 = 6

MSB = 573.9/2 = 286.95

MSW = 0.66/6 = 0.11

F statistic = F = MSB/MSW = 2618

Features with significant differences ($p < 0.05$) are likely good candidates as predictors because their distributions vary meaningfully across groups.

we can rank the features based on the strength of their significance and effect size as part of a feature selection process.

# Practical Questions:

## Sleep Health and Lifestyle Dataset

1)EDA:

Dataset Structure:

- The dataset contains 374 records and 13 columns.
- The data types are (int64, float64) and (object) features.

Feature:

- Numerical:
    - Age, Sleep Duration, Quality of Sleep, Physical Activity Level, Stress Level, Heart Rate, and Daily Steps.
- Categorical:
    - Gender, Occupation, BMI Category, Blood Pressure, and Sleep Disorder.

Objectives: the objective of this dataset is to find the relation ship between the life style of a person with some associated feature the kinds of sleep disorders that that person experiences.

Summery of the most important information of each col:

the information includes the count, mean, std, min, max , 75% and 50% of numerical features and count, unique, top and frequency of the categorical features.

(374, 13)

| | Person ID | Age | Sleep Duration | Quality of Sleep | \ |
|---|---|---|---|---|---|
| count | 374.000000 | 374.000000 | 374.000000 | 374.000000 | |
| mean | 187.500000 | 42.184492 | 7.132086 | 7.312834 | |
| std | 108.108742 | 8.673133 | 0.795657 | 1.196956 | |
| min | 1.000000 | 27.000000 | 5.800000 | 4.000000 | |
| 25% | 94.250000 | 35.250000 | 6.400000 | 6.000000 | |
| 50% | 187.500000 | 43.000000 | 7.200000 | 7.000000 | |
| 75% | 280.750000 | 50.000000 | 7.800000 | 8.000000 | |
| max | 374.000000 | 59.000000 | 8.500000 | 9.000000 | |

| | Physical Activity Level | Stress Level | Heart Rate | Daily Steps |
|---|---|---|---|---|
| count | 374.000000 | 374.000000 | 374.000000 | 374.000000 |
| mean | 59.171123 | 5.385027 | 70.165775 | 6816.844920 |
| std | 20.830804 | 1.774526 | 4.135676 | 1617.915679 |
| min | 30.000000 | 3.000000 | 65.000000 | 3000.000000 |
| 25% | 45.000000 | 4.000000 | 68.000000 | 5600.000000 |
| 50% | 60.000000 | 5.000000 | 70.000000 | 7000.000000 |
| 75% | 75.000000 | 7.000000 | 72.000000 | 8000.000000 |
| max | 90.000000 | 8.000000 | 86.000000 | 10000.000000 |

| | Gender | Occupation | BMI Category | Blood Pressure | Sleep Disorder |
|---|---|---|---|---|---|
| count | 374 | 374 | 374 | 374 | 155 |
| unique | 2 | 11 | 4 | 25 | 2 |
| top | Male | Nurse | Normal | 130/85 | Sleep Apnea |
| freq | 189 | 73 | 195 | 99 | 78 |

Figure 1 : summery info about the features.

## Age:

Mean (42.18) / Median (43): the dataset represents a slightly older adult population. The close proximity of mean and median shows a symmetric age distribution. Quarter (25%, 50%, 75%): it shows that the dataset has a wide range of people of different age.

Sleep Duration:

Mean (7.13) / Median (7.2): on average, individuals sleep around 7 hours per day. Again, the mean and median are close, so we have a symmetrical distribution. Range (5.8 - 8.5): Shows some variability in sleep duration, but most individuals are in a relatively narrow band.

Quality of Sleep:

Mean (6.31) / Median (7): With a 1-10 scale the average sleep quality rating is above the midpoint. However, the median being higher than the mean says that the distribution might be slightly skewed towards lower scores.

Physical Activity Level:

Mean (59.17) / Median (60): On average individuals take part in about an hour of physical activity per day.

Stress Level:

Mean (5.38) / Median (5): The average stress level is above the midpoint of the 1-10 scale. Similar to sleep quality the lower mean compared to the median shows skew towards higher stress levels for some individuals.

Heart Rate:

Mean (70.17) / Median (70): The average resting heart rate is around 70 beats per minute which is in the normal range for adults.

Daily Steps:

Mean (6816.84) / Median (7000): the avg of steps daily is 7000 for people. Quarter/ Range (3000 - 10000): There's a lot variation in daily steps showing different levels of activity among people .

Here we can see the cols and their data types:

| # | 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)

unique values or the categories of categorical features:

Gender

['Male' 'Female']

Occupation

['Software Engineer' 'Doctor' 'Sales Representative' 'Teacher' 'Nurse'

'Engineer' 'Accountant' 'Scientist' 'Lawyer' 'Salesperson' 'Manager']

BMI Category

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

Blood Pressure

['126/83' '125/80' '140/90' '120/80' '132/87' '130/86' '117/76' '118/76'
 '128/85' '131/86' '128/84' '115/75' '135/88' '129/84' '130/85' '115/78'
 '119/77' '121/79' '125/82' '135/90' '122/80' '142/92' '140/95' '139/91'
 '118/75']

Sleep Disorder

[nan 'Sleep Apnea' 'Insomnia']

Potential Data Issues:

1. Missing Values:
    - There are 219 missing values in the dataset. Only in Sleep Disorder col.
    - How to handle?
        - Since the creator of the dataset probably wanted to say that the entries with missed values in this col don't have any disorder, we can replace the None (missed) values with " no disorder".
2. In BMI the "Normal" and "Normal Weight" have the same meaning:
    - How to handle?
        - renaming the "Normal Weight" category to "Normal". ( so instead of 4 categories we will have 3 categories)

2) visualization:

In this section we have some plots that are derived from the basic information of the dataset like the distribution of each feature:

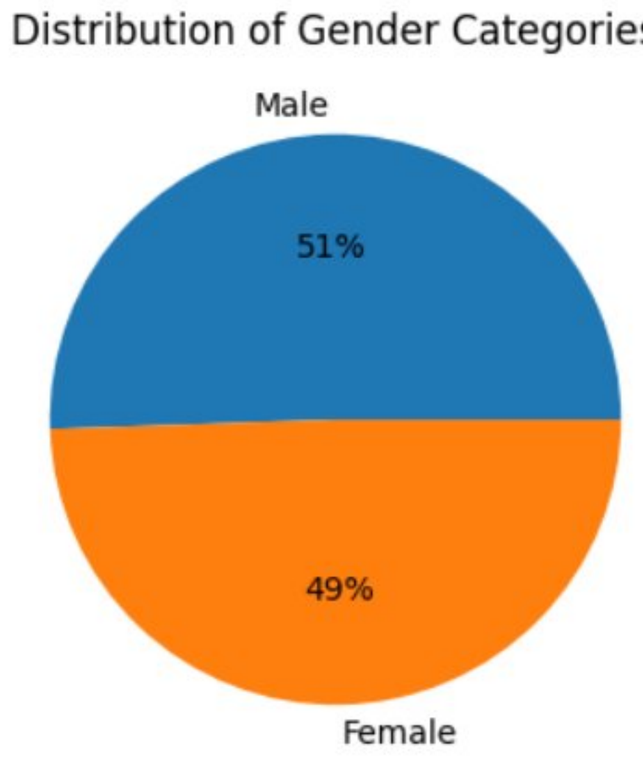



Figure 2: the distribution of gender categories. We can see that the genders are distributed almost evenly.

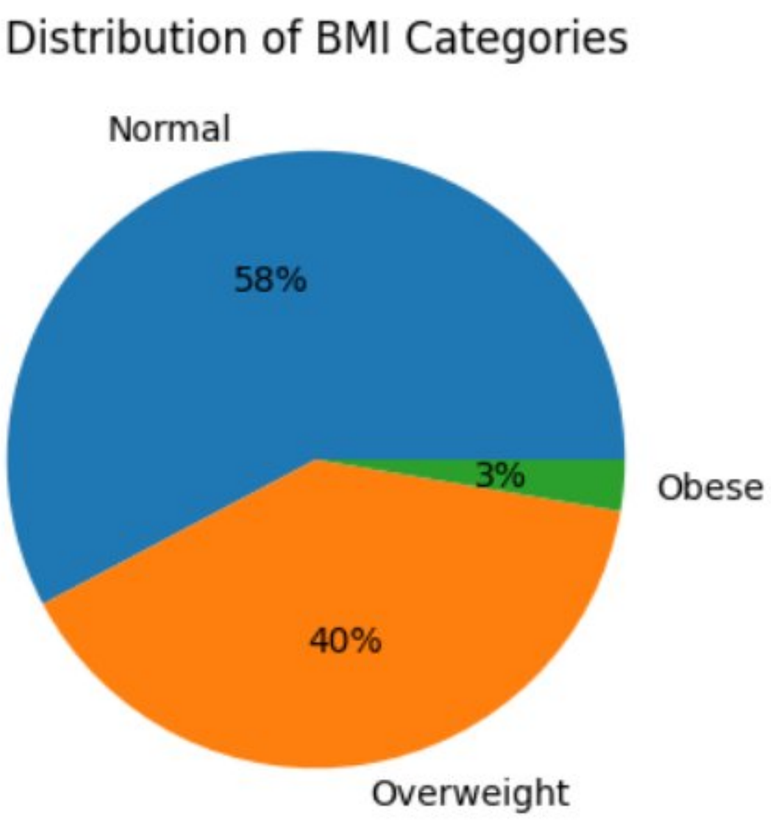



Figure 3: Distribution of BMI Categories:

Most of the entries have normal bmi

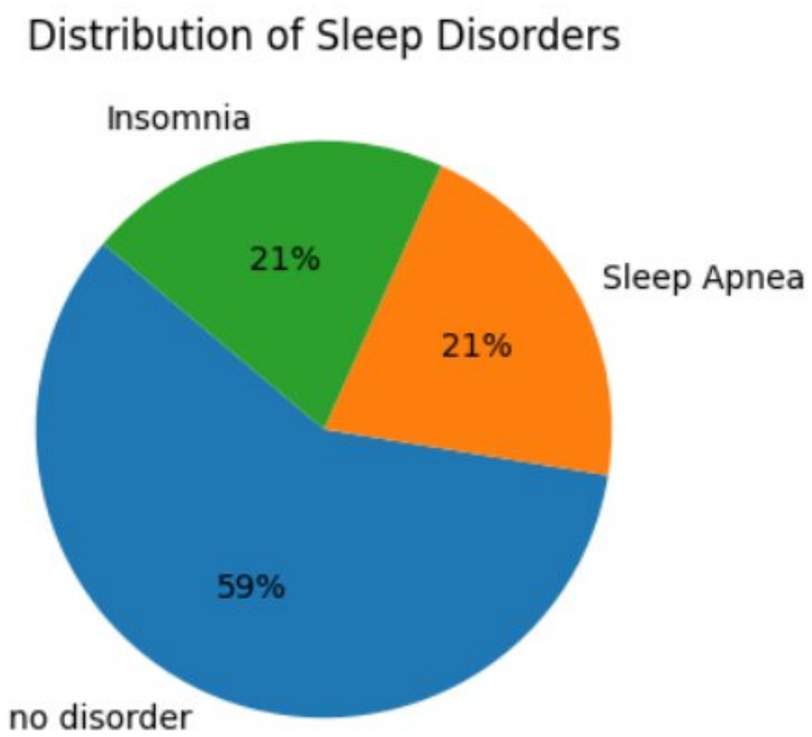
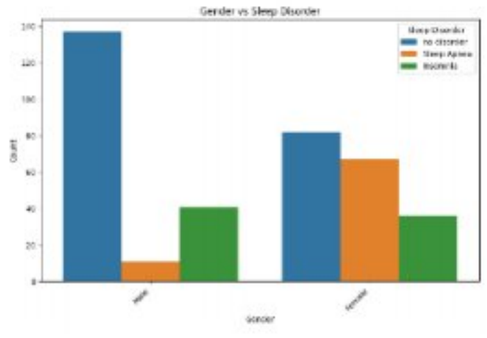
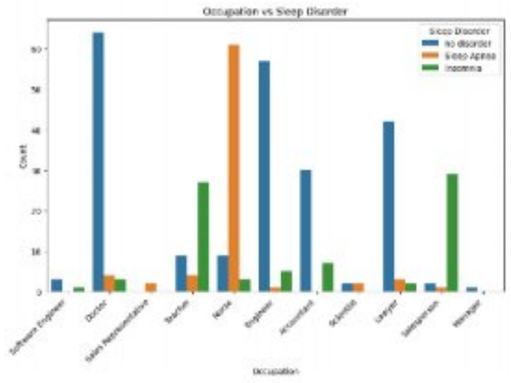


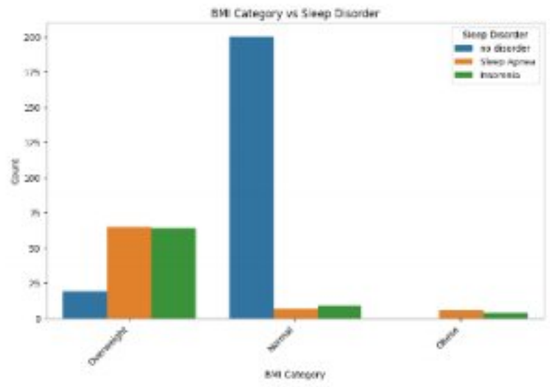

Figure 4: Distribution of Sleep Disorders

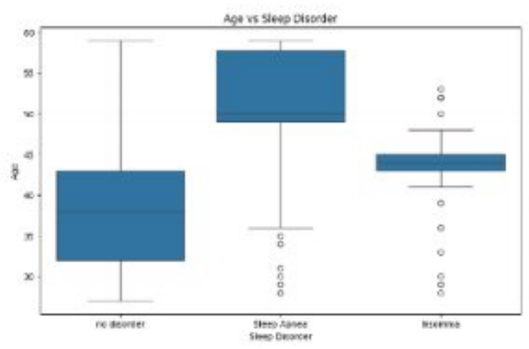
Most of the entries have no disorder

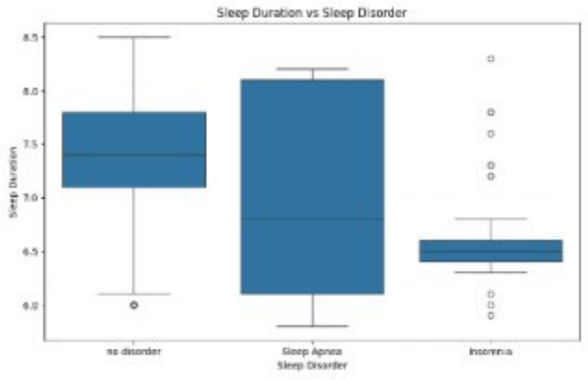
Figure 5: the distribution of features v.s sleeping disorders:

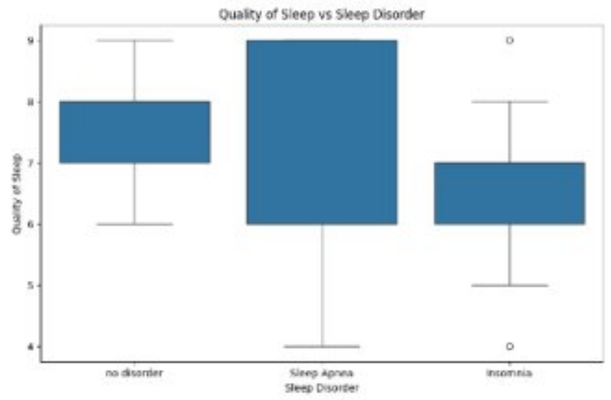


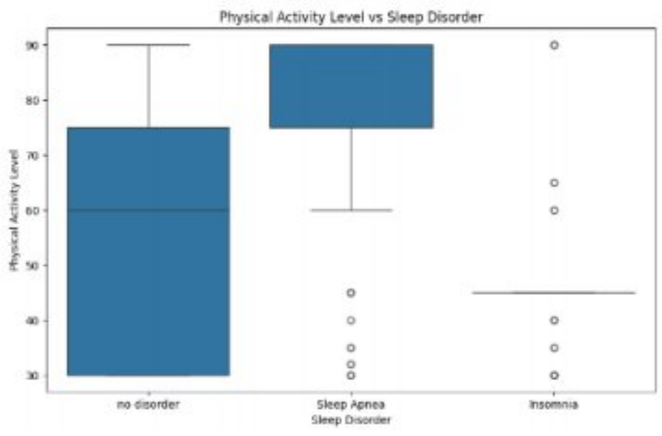



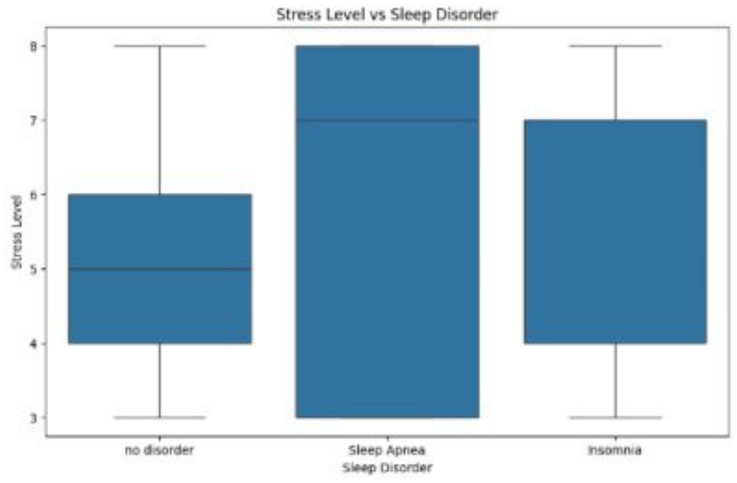



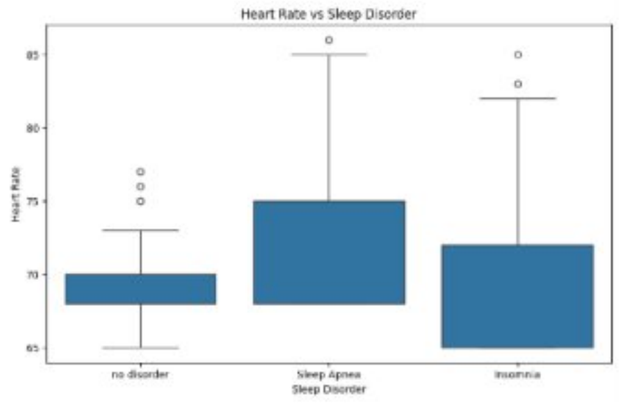



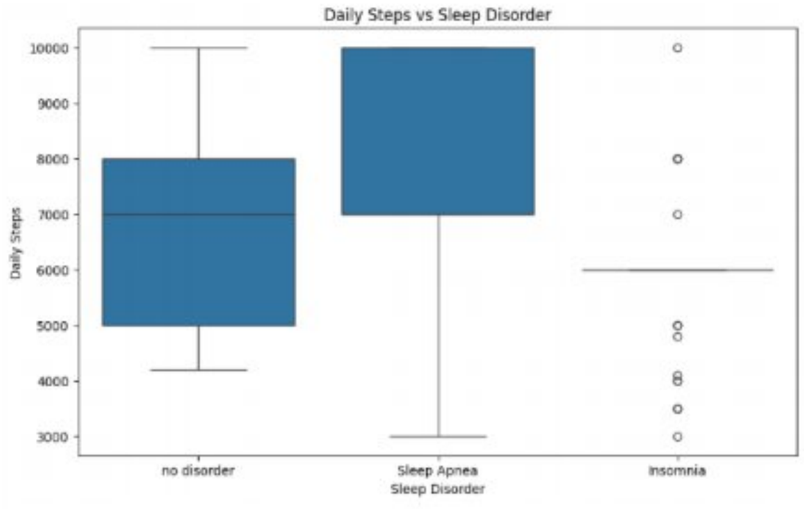



These plots help with the capability of new pattern identification as well as ease the analysis process due to more definite information related to each feature individually

## Hypothesis testing:

For each of the following hypothesis a specific test is applied and the reason for each in mentioned.

In all of the following tests if the p value is less than alpha=0.05 h0 will be rejected.

A: Does women’s sleep duration follow a normal distribution?

Null Hypothesis : Women's sleep duration follows a normal distribution.

Alternative Hypothesis : Women's sleep duration does not follow a normal distribution.

Test: Shapiro-Wilk normality test. Why? Because we want to check the normality of Women's sleep duration .

Result:

Shapiro-Wilk Test Statistic:

0.8985763558917751

p-value:

6.360571286948505e-10

Reject the H0.

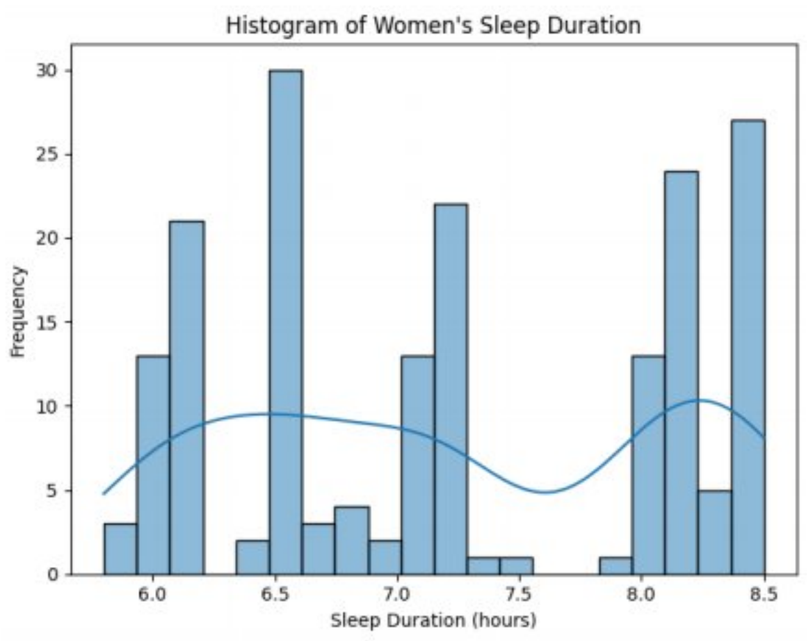



Figure 6 which shows it doesn’t follow a normal distribution.

B: Is having higher daily steps a contributing factor into better sleep? Check

the corresponding correlation of Daily Steps and Quality of Sleep.

Null Hypothesis ($H_o$): There is no significant correlation between Daily Steps and Quality of Sleep.

Alternative Hypothesis ($H_1$): There is a significant correlation between Daily Steps and Quality of Sleep.

Test: Pearson correlation test. Why? to measure the strength and direction of the linear relationship between two continuous variables.( Daily Steps and Quality of Sleep) and also the main question wants us to check for these two factors collorations.

Result:

Pearson Correlation Coefficient:

0.016791414924716 78

p-value:

0.7461906652961119

Fail to reject the H0

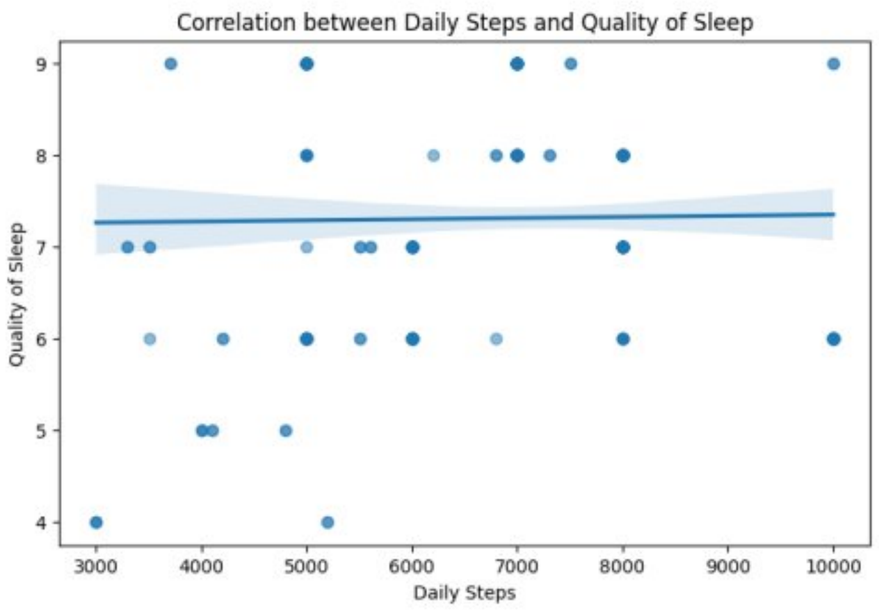



Figure 7: We can’t see significant correlation.

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.

h0: There is no statistically significant difference in stress levels among different occupations.

h1:Stress levels differ significantly among occupations.

Test: Kruskal. Why? to determine whether there are statistically significant differences between the medians of three or more independent groups (different occupations here)

Result:

H-statistic

136.39693171205218

p-value:

2.304155098706687e-24

Reject the H0.

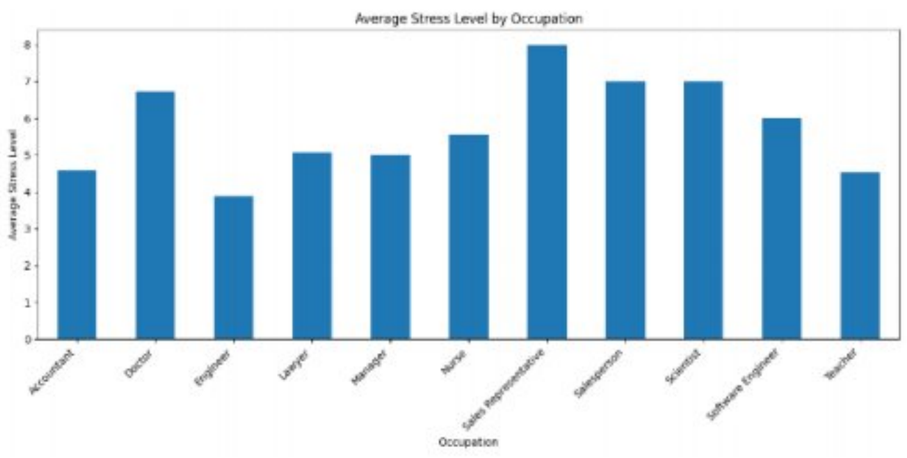



Figure 8: here we can see the avg stress level for each occupation which are not the same.

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.)

Test:  kruskal ( explained in the previous one) we want to check if BMI categories differ for different blood pressure.

h0: No significant difference in blood pressure among BMI groups

h1:Blood pressure differs significantly across BMI groups.

We divide the blood pressure to 2 groups and perform the test for each.

Result:

Systolic nH-statistic

215.78340584354038

p-value:

1.390685024254713e-47

Diastolic nH-statistic

220.8054883352936

p-value:

1.1290090443421815e-48

Reject the H0.

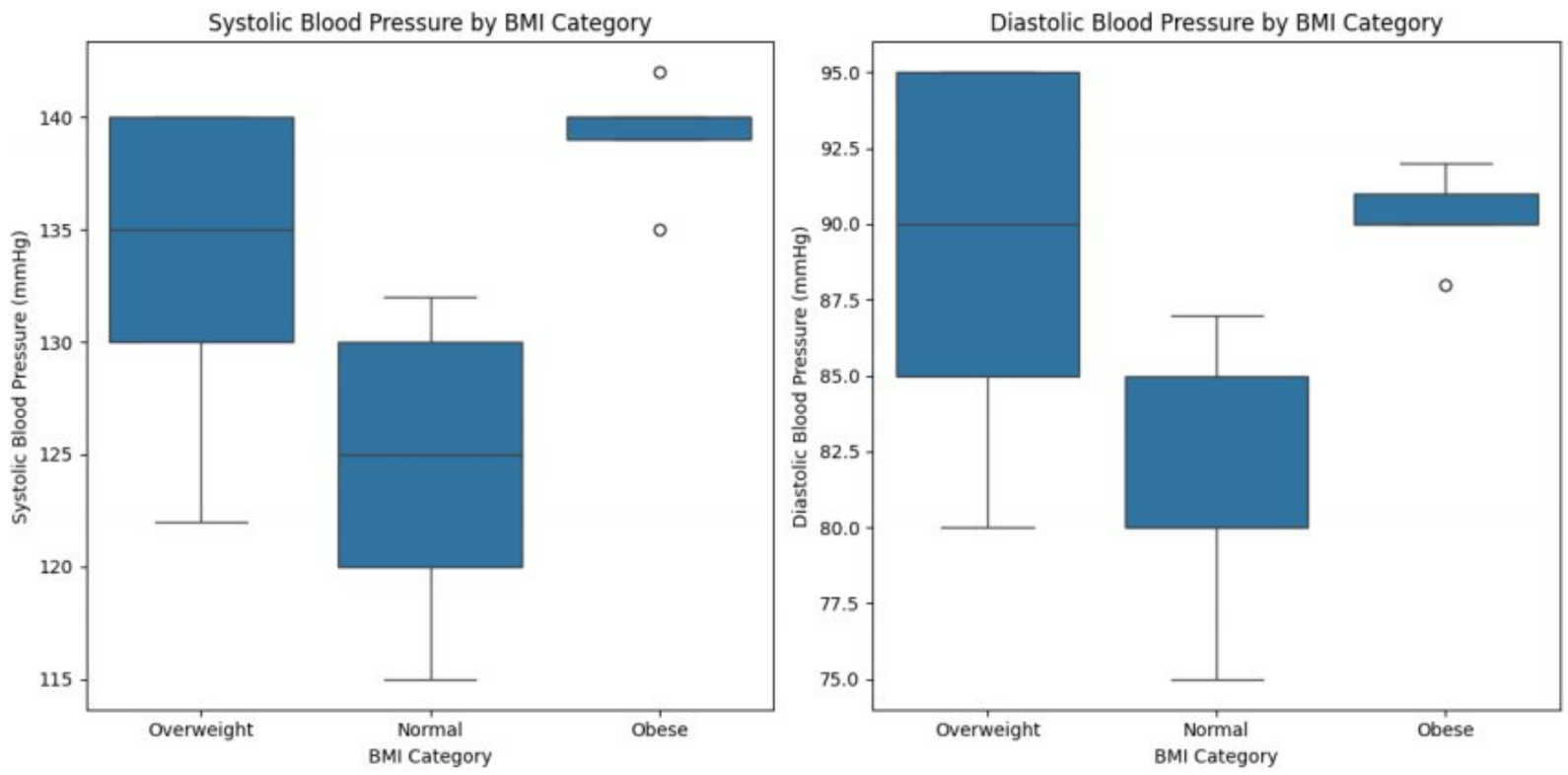



Figure 9: we can see that Blood pressure differs significantly across BMI groups according to both groups of blood pressure.

E: Do people with sleep disorders have higher heart rates than those without

any sleep disorder?

Null Hypothesis ($H_0$):There is no difference in median heart rates between people with and without sleep disorders.

Alternative Hypothesis ($H_1$): People with sleep disorders have higher median heart rates than those without.

Test: Mann-Whitney U  to compare the distributions of two independent groups. We want to compare people with sleep disorders and people without sleep disorders over heart rate.

Result:

statistic

22399.0

p-value:

3.9580425954857 81e-08

Reject the H0.

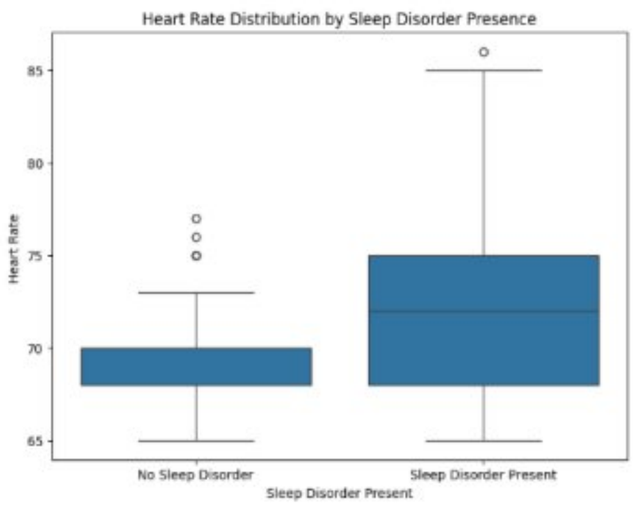



Figure 10: we see that the hear rates differ from one group to another.

Bonus:

1: Null Hypothesis ($H_0$):There is no significant difference between the means of Sleep Duration and Quality of Sleep

Alternative Hypothesis ($H_1$):There is a significant difference between the means of Sleep Duration and Quality of Sleep.

Test: z test to determine whether there is a significant difference between the mean of a sample and a known population mean (or between the means of two samples) when the population standard deviation is known. Here between Sleep Duration and Quality of Sleep.

Result:

z-statistic

-2.432033681094019

p-value:

0.015014311061261587

Reject $H_o$

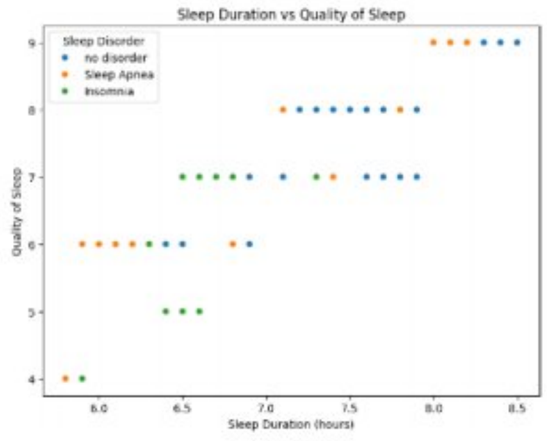



Figure 11

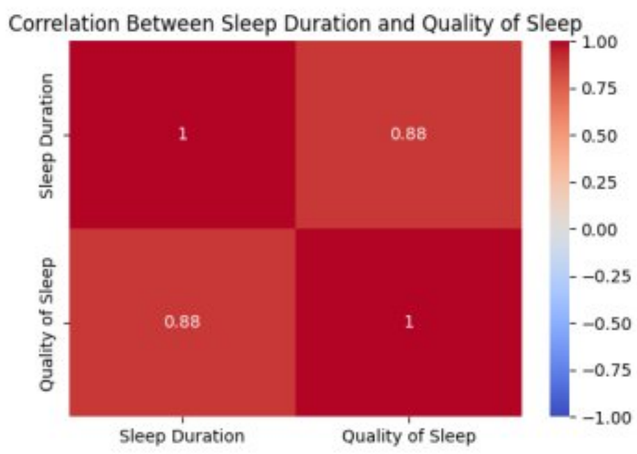



Figure 12

INCE Pvalue = 0.015< 0.05,

we reject the null hypothesis that says there is no difference between the average sleep duration and Quality Of sleep.

2:

Null Hypothesis: there is no difference in sleep quality of males and females.

Alternative Hypothesis: there is a difference in sleep quality of males and females.

Test: t test to determine whether there is a significant difference between the mean of a sample and a known population mean (or between the means of two samples) when the population standard deviation is unknown. Here between sleep quality of males and females.

Result:

t-statistic

-5.874547760454642

p-value:

9.416446532689304e-09

Reject $H_o$.

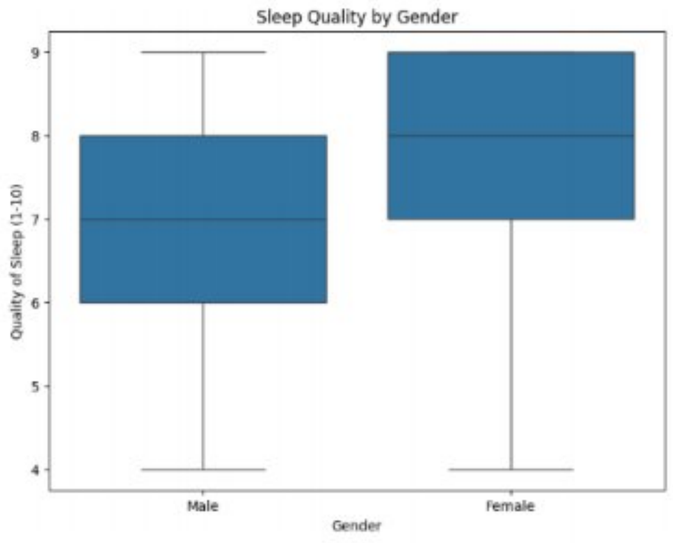



Figure 13: we can obviously see that the sleep quality differs for genders .

3:

Null Hypothesis ($H_o$): There is no relationship between physical activity level and sleep quality.

Alternative Hypothesis ($H_1$): There is a relationship between physical activity level and sleep quality.

Test: Pearson correlation ( it is already explained why to use this)

Result:

Coefficient

0.19289645493975321

p-value:

0.00017454895247838838

Reject $H_o$

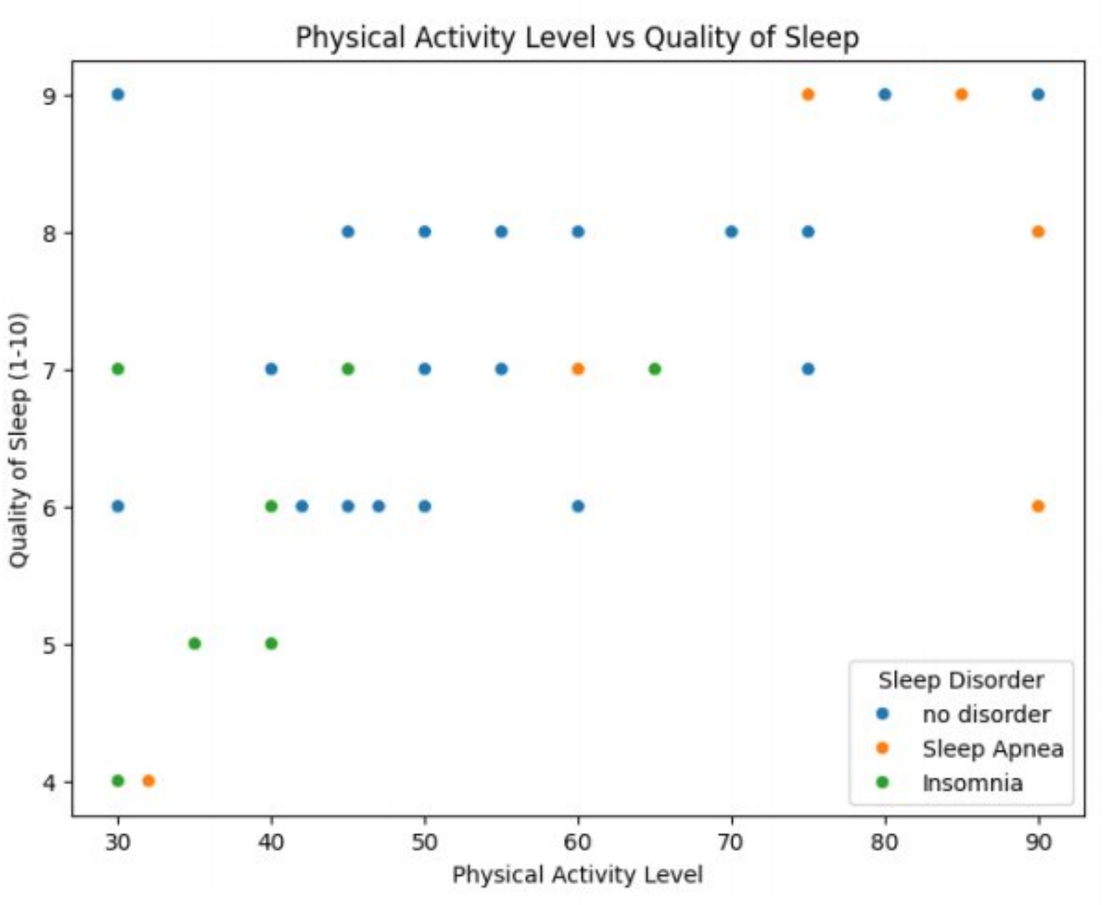



Figure14

## Student Performance Factors Dataset

1)EDA:

Dataset Structure:

- The dataset contains 6607 records and 20 columns.
- The data types are (int64, float64) and (object) features.

Feature:

- Numerical:
  - Hours_Studied, Attendance, Sleep_Hours, Previous_Scores, Tutoring_Sessions, Physical_Activity, Exam_Score

- Categorical:
  - Parental_Involvement, Access_to_Resources, Extracurricular_Activities, Motivation_Level, Internet_Access, Family_Income, Teacher_Quality, School_Type, Peer_Influence, Learning_Disabilities, Parental_Education_Level, Distance_from_Home, Gender

Objectives: the objective of this dataset is to find the relation ship between the factors related to students and how well they do on their exams

Summery of the most important information of each col:

the information includes the count, mean, std, min, max , 75% and 50% of numerical features and count, unique, top and frequency of the categorical features.

| | Hours_Studied | Attendance | Sleep_Hours | Previous_Scores |
|---|---|---|---|---|
| count | 6607.000000 | 6607.000000 | 6607.00000 | 6607.000000 |
| mean | 19.975329 | 79.977448 | 7.02906 | 75.070531 |
| std | 5.990594 | 11.547475 | 1.46812 | 14.399784 |
| min | 1.000000 | 60.000000 | 4.00000 | 50.000000 |
| 25% | 16.000000 | 70.000000 | 6.00000 | 63.000000 |
| 50% | 20.000000 | 80.000000 | 7.00000 | 75.000000 |
| 75% | 24.000000 | 90.000000 | 8.00000 | 88.000000 |
| max | 44.000000 | 100.000000 | 10.00000 | 100.000000 |

| | Tutoring_Sessions | Physical_Activity | Exam_Score |
|---|---|---|---|
| count | 6607.000000 | 6607.000000 | 6607.000000 |
| mean | 1.493719 | 2.967610 | 67.235659 |
| std | 1.230570 | 1.031231 | 3.890456 |
| min | 0.000000 | 0.000000 | 55.000000 |
| 25% | 1.000000 | 2.000000 | 65.000000 |
| 50% | 1.000000 | 3.000000 | 67.000000 |
| 75% | 2.000000 | 4.000000 | 69.000000 |
| max | 8.000000 | 6.000000 | 101.000000 |

| | Parental_Involvement | Access_to_Resources | Extracurricular_Activities |
|---|---|---|---|
| count | 6607 | 6607 | 6607 |
| unique | 3 | 3 | 2 |
| top | Medium | Medium | Yes |
| freq | 3362 | 3319 | 3938 |

| | Motivation_Level | Internet_Access | Family_Income | Teacher_Quality |
|---|---|---|---|---|
| count | 6607 | 6607 | 6607 | 6529 |
| unique | 3 | 2 | 3 | 3 |
| top | Medium | Yes | Low | Medium |
| freq | 3351 | 6108 | 2672 | 3925 |

| | School_Type | Peer_Influence | Learning_Disabilities |
|---|---|---|---|
| count | 6607 | 6607 | 6607 |
| unique | 2 | 3 | 2 |
| top | Public | Positive | No |
| freq | 4598 | 2638 | 5912 |

| | Parental_Education_Level | Distance_from_Home | Gender |
|---|---|---|---|
| count | 6517 | 6540 | 6607 |
| unique | 3 | 3 | 2 |
| top | High School | Near | Male |
| freq | 3223 | 3884 | 3814 |

Here we can see the cols and their data types:

| # | 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)

unique values or the categories of categorical features:

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']

Peer_Influence

['Positive' 'Negative' 'Neutral']

Learning_Disabilities

['No' 'Yes']

Parental_Education_Level

['High School' 'College' 'Postgraduate' nan]

Distance_from_Home

['Near' 'Moderate' 'Far' nan]

Gender

['Male' 'Female']

Potencial issue:

Missing value: the followings are the number of missing values for each col:

Hours_Studied 0

Attendance 0

Parental_Involvement 0

Access_to_Resources 0

Extracurricular_Activities 0

Sleep_Hours 0

Previous_Scores 0

Motivation_Level 0

| | |
|---|---|
| Internet_Access | 0 |
| Tutoring_Sessions | 0 |
| Family_Income | 0 |
| Teacher_Quality | 78 |
| School_Type | 0 |
| Peer_Influence | 0 |
| Physical_Activity | 0 |
| Learning_Disabilities | 0 |
| Parental_Education_Level | 90 |
| Distance_from_Home | 67 |
| Gender | 0 |
| Exam_Score | 0 |

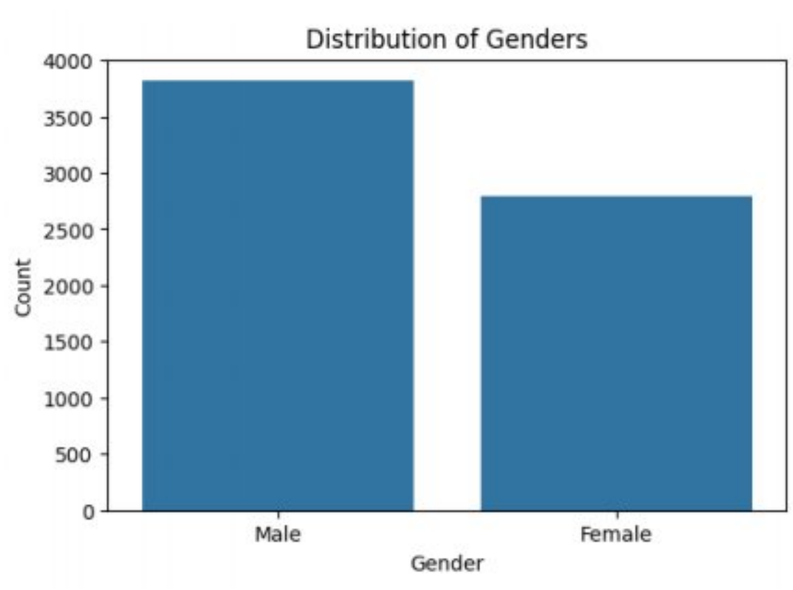
As it is obvious the number of missing values is so small in comparison to the whole records:

(90+67+78)/20*6607

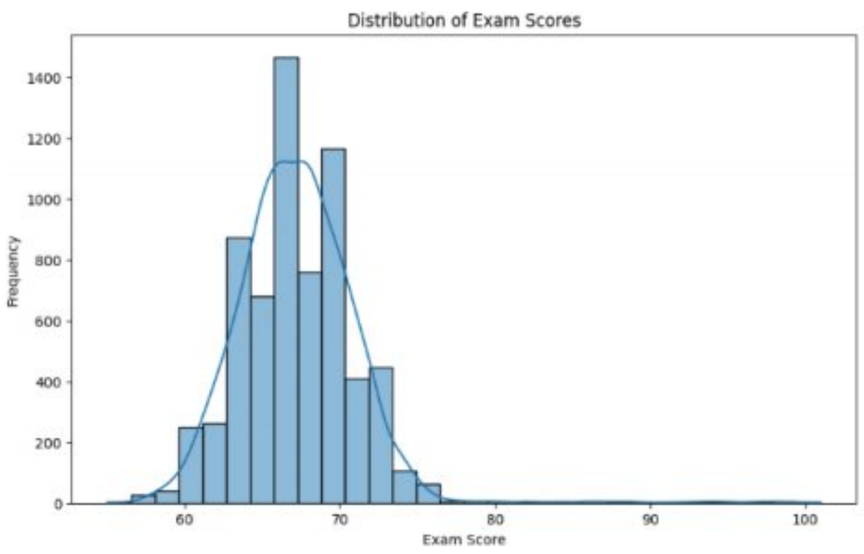
So the best strategy here is to drop those rows with missing values because they are too few that won't harm our research.

Visualization:

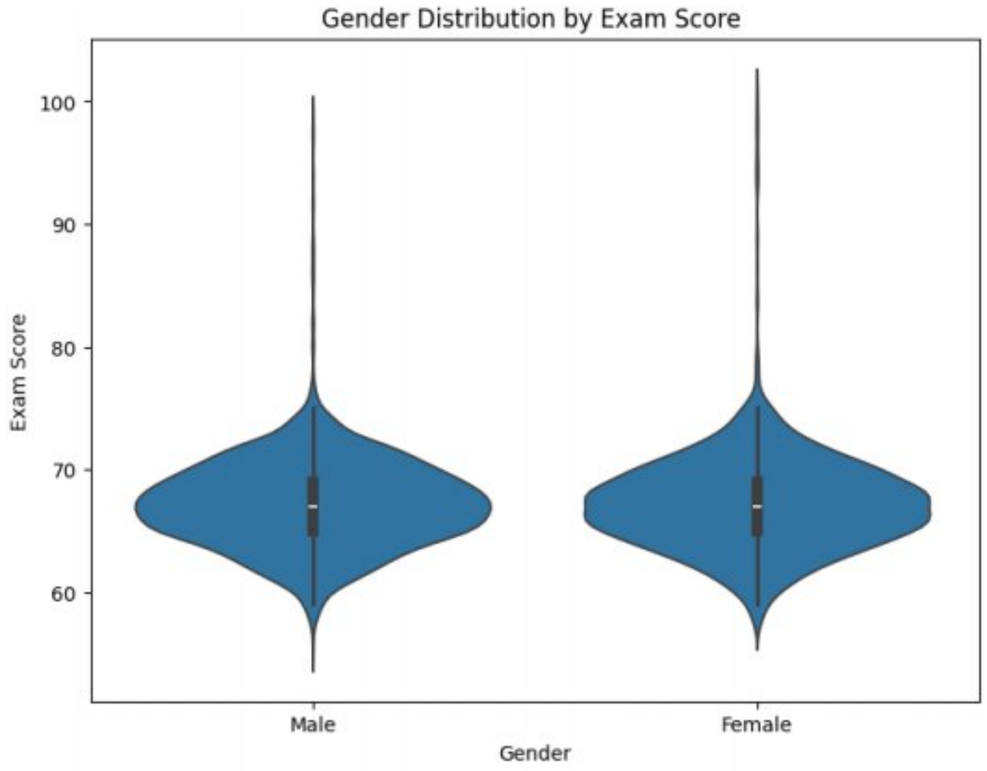
distribution of genders: Males are more than Femails



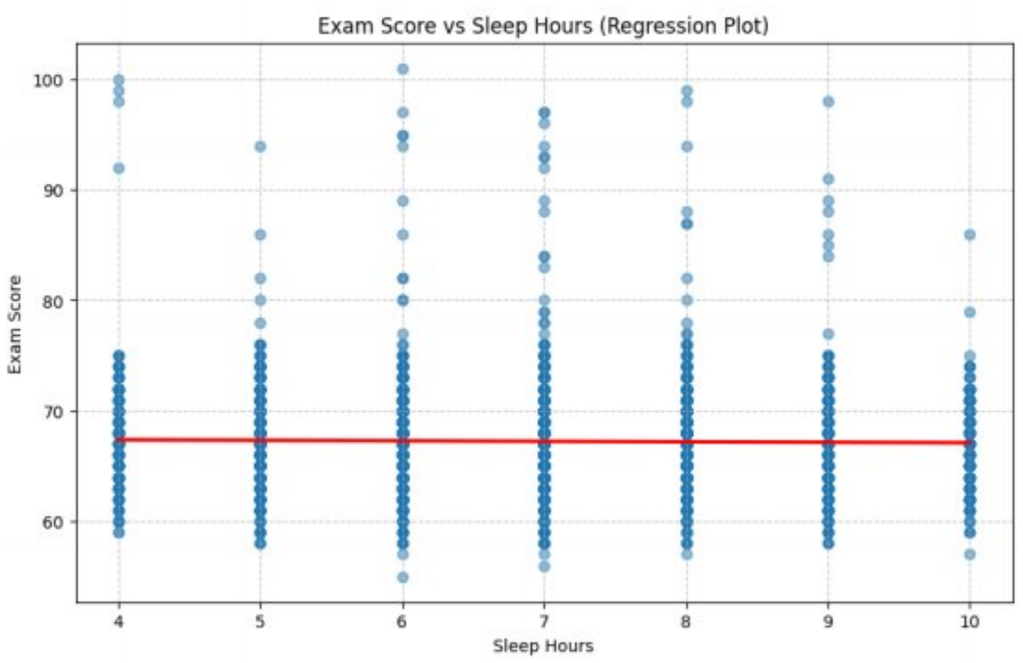

distribution of exam scores( it is not normal)



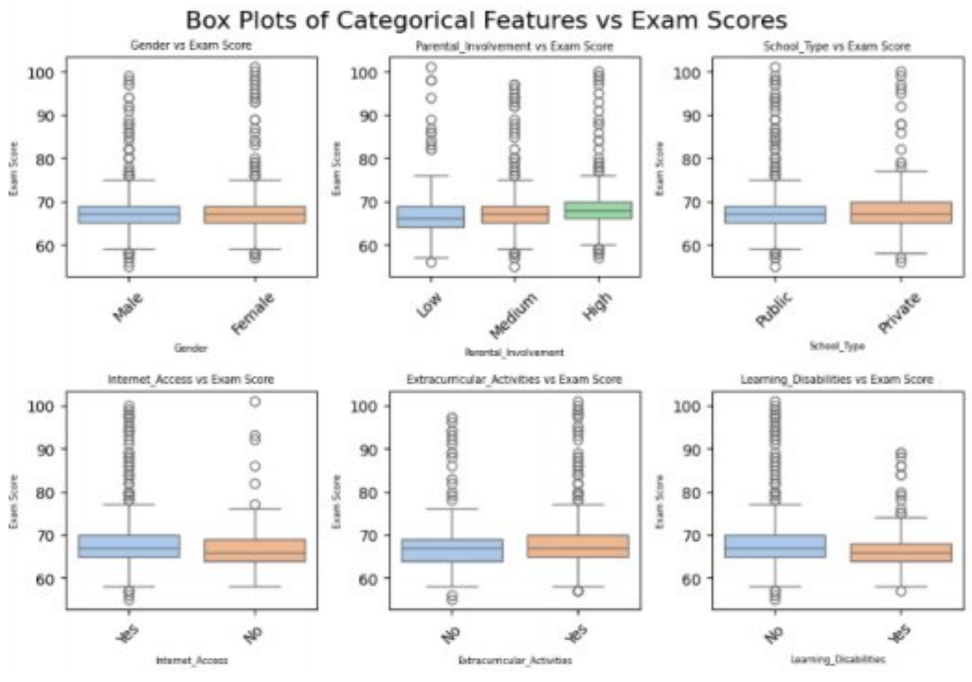

gender distribution by score: there is almost no difference between genders



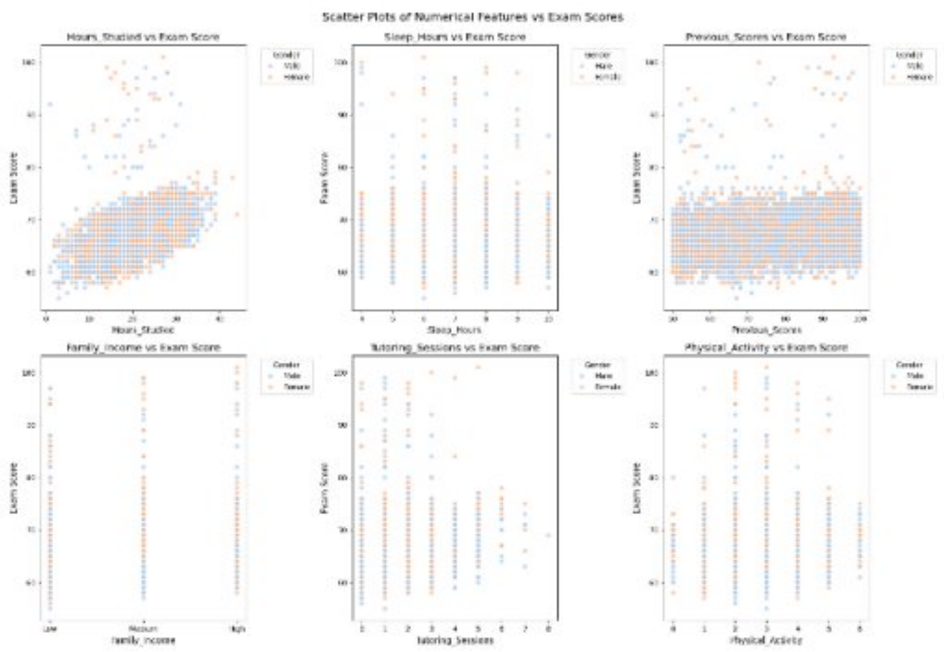

exam score vs sleep hours




all categorical features vs exam score in a single figure: These plots help with the capability of new pattern identification as well as ease the analysis process due to more definite information related to each feature individually




all numerical features vs exam score in a single figure




Hypothesis test:

For each of the following hypothesis a specific test is applied.

In all of the following tests if the p value is less than alpha=0.05 h0 will be rejected.

The reason of why each of these tests are used are already explained in the last section

1:

Null Hypothesis ($H_o$): There is no significant difference in exam scores between males and females.

Alternative Hypothesis ($H_1$): There is a significant difference in exam scores between males and females.

Test:  t-test: t test to determine whether there is a significant difference between the mean of a sample and a known population mean (or between the means of two samples) when the population standard deviation is unknown. Here between scores of males and females.

Result;

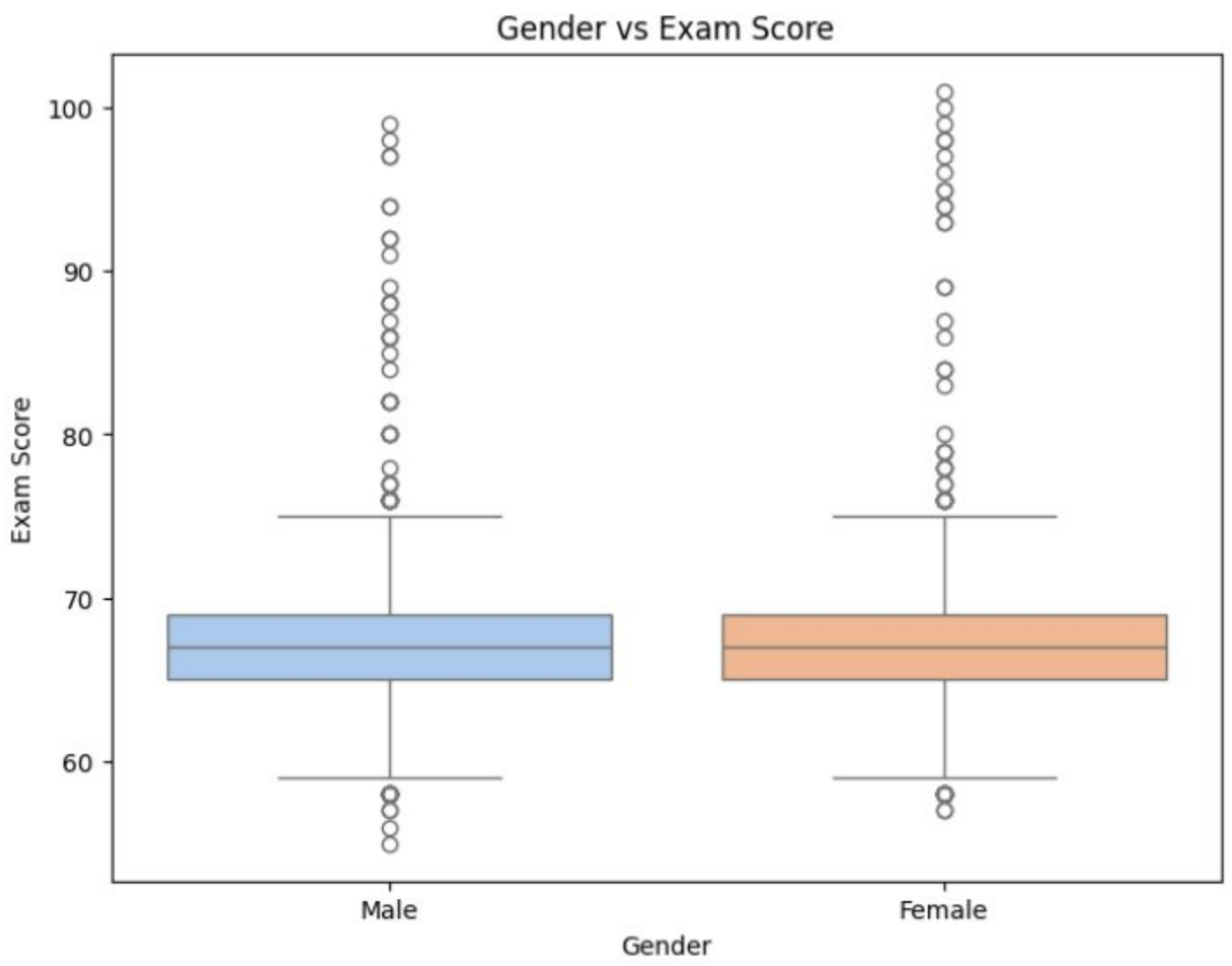
T-statistic

-0.16516987601406408

p-value:

0.8688153297340319

Fail to reject the null hypothesis




The exam score are almost the same across the genders.

2:

Null Hypothesis (H0): The exam scores are normally distributed.

Alternative Hypothesis (H1): The exam scores are not normally distributed.

Test; Shapiro-Wilk ( alredy explaibed): we use this test to examine normality.

Result:

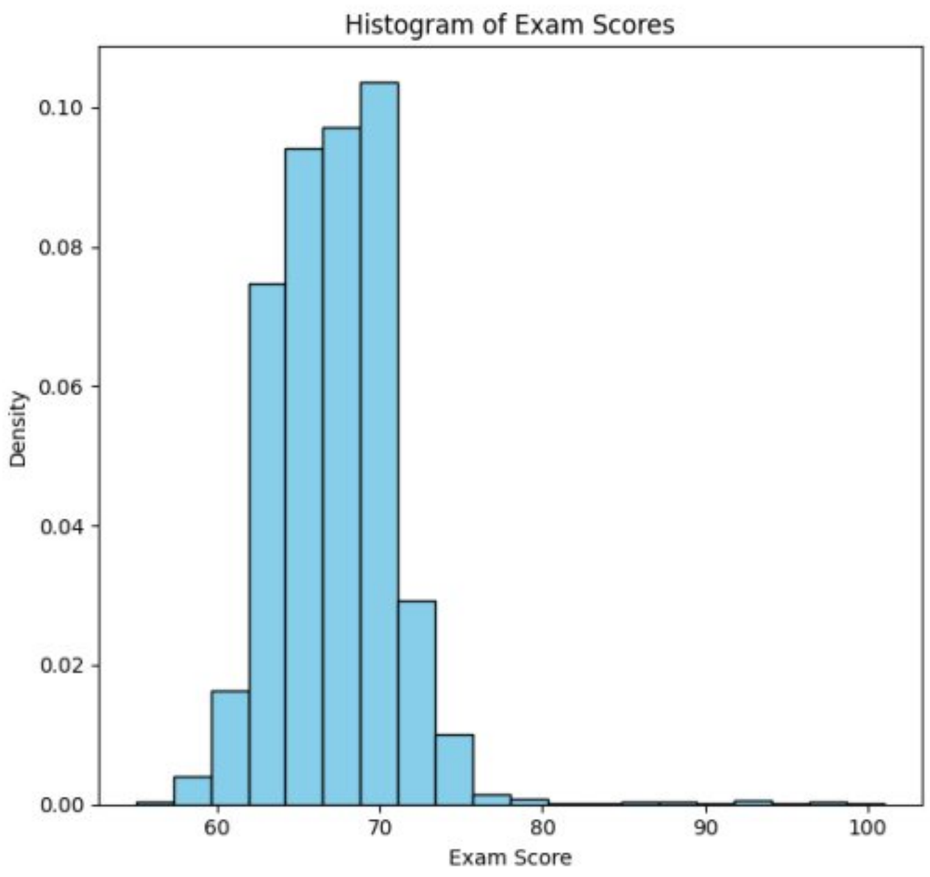
statistic

0.9029140953482367

p-value:

8.4304278588705e-54

Reject the null hypothesis




Its distribution is not normal

3:

Null Hypothesis (H0): There is no significant correlation between hours studied and exam scores.

Alternative Hypothesis (H1): There is a significant correlation between hours studied and exam scores.

Test: Pearson: we use this to check for correlation

Result:

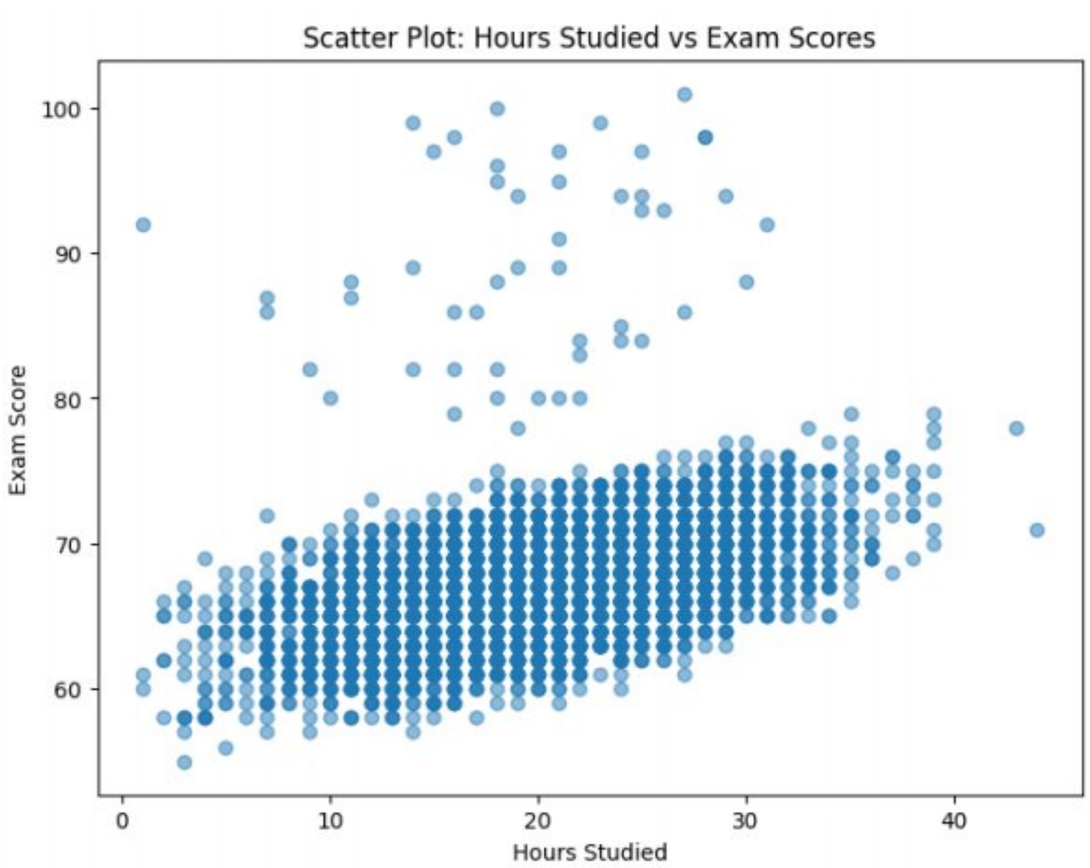
Coefficient

0.44545495407528185

p-value:

1.28635e-319

Reject the null hypothesis




As we can see there is a significant correlation between hours studied and exam scores and we can also draw a regression line with a high accuracy.

4:

Null Hypothesis (H0): There is no significant difference in the median exam scores across different levels of parental involvement.

Alternative Hypothesis (H1): There is a significant difference in the median exam scores across different levels of parental involvement.

Test: Kruskal: ? to determine whether there are statistically significant differences between the medians of three or more independent groups (different levels of parental involvement.

here)

Result:

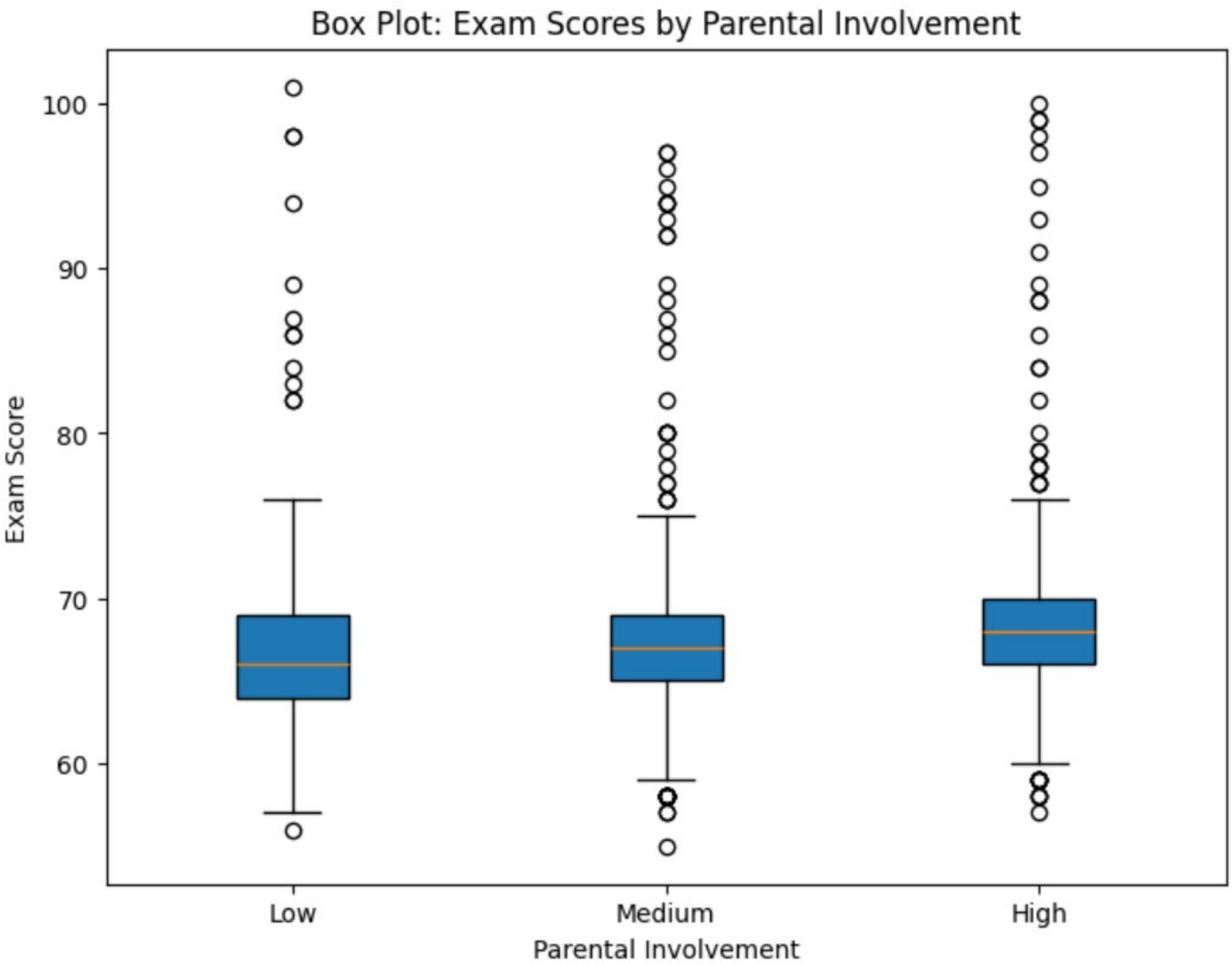
H-statistic

195.9244961378635

p-value:

2.8545406295009852e-43

Reject the null hypothesis




5:

Null Hypothesis (H0): There is no significant difference in the median exam scores between students from public and private schools.

Alternative Hypothesis (H1): There is a significant difference in the median exam scores between students from public and private schools

Test: Mann-Whitney U test to compare the distributions of two independent groups. We want to compare the scores of  students in public schools and the ones in private schools.

Result;

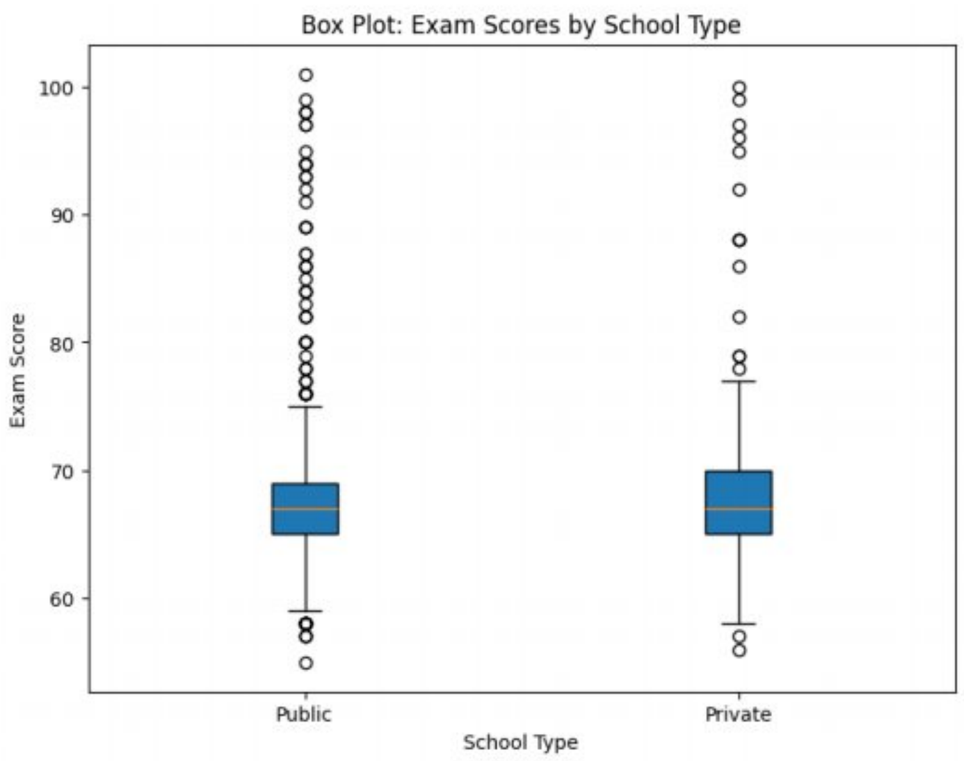
U-statistic

4552099.5

p-value:

0.3485872249726576

Fail to reject the null hypothesis




There is no significant difference in the median exam scores between students from public and private schools

Conclusion:

During this project we ended up performing EDA on data set, visualize it and perform different hypothesis tests according to the hypotheses. And also analyze the plots