Bytewise Fellowship Program

DATA SCIENCE

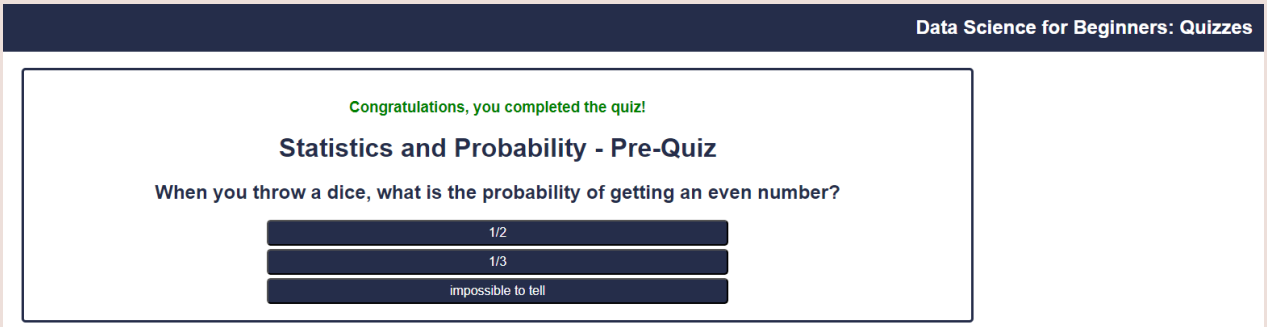
Task #3

BWT- Data Science (Group1)

Submitted to: Mahrukh Khan
Submitted by: Usama Malik

Task: A Brief Introduction to Statistics and Probability

Pre Quiz



Probability and Random Variables

Probability:

- **Def:** The chance that something will happen.
- **Example:** The probability of it raining today is 40%, so you might want to carry an umbrella.

Random Variables:

- **Def:** A number that can change based on chance.
- **Example:** The amount of time people spend on social media each day is a random variable.

Mean, Variance, and Standard Deviation

Mean:

- **Def:** The average of a set of numbers.
- **Example:** The mean score of the test was 75 out of 100, which indicates that most students performed well.
- **Formula:** Add set of number and / divide by total number = mean

Variance:

- **Def:** A measure of how much the numbers in a set differ from the average.
- **Example:** The variance in the temperatures this month shows that the weather has been quite unpredictable.

Standard Deviation:

- **Def:** A measure of how spread out the numbers are in a set.
- **Example:** The sales data indicates that our monthly sales have been consistently.
- **Formula:** firstly calculate mean, calculate each deviation from mean and square it, then find average and take the square root of variance.

Mode, Median, and Quartiles

Mode:

- **Simple Definition:** The number that appears most often in a set.
- **Usage Example:** The mode of the survey responses was 'satisfied,' showing that most customers are happy.
- **Formula:** The most same number repeat in data set is Mode

Median:

- **Simple Definition:** The middle number in a set when the numbers are arranged in order.
- **Usage Example:** "The median income in the area is $50,000, which gives us an idea of the typical earnings."
- **Formula:** The point lies between greater than zero or less than highest value = Median

Quartiles:

- **Def:** The Values that divide a set of numbers into four equal parts.

Covariance and Correlation

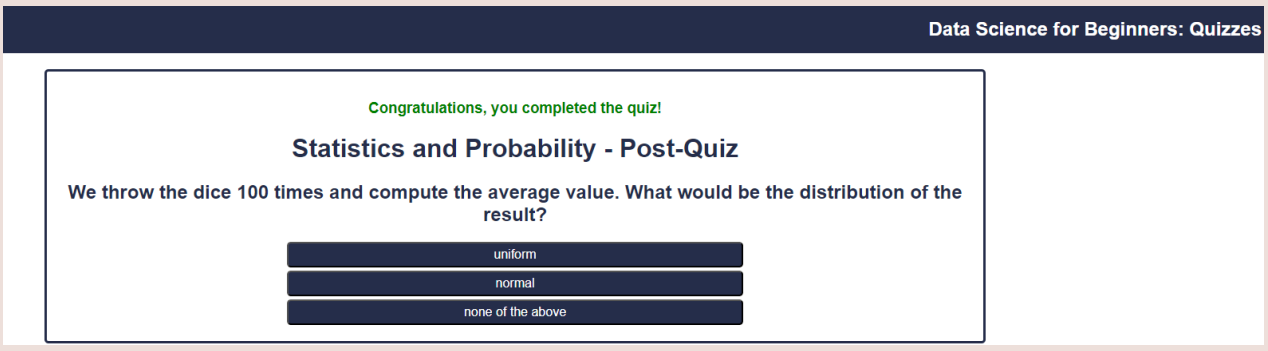
Covariance:

- **Def:** A measure of how two sets of numbers change together.
- **Example:** The covariance between hours studied and exam scores is positive, suggesting that more studying leads to higher scores.

Correlation:

- **Def:** A measure of how strongly two sets of numbers are related.
- **Example:** The correlation between exercise and health is strong, indicating that regular exercise is linked to better health outcomes.

Post quiz



Assignment:

Compute mean values and variance for all values

Plot boxplots for BMI, BP and Y depending on gender

What is the the distribution of Age, Sex, BMI and Y variables?

Test the correlation between different variables and disease progression (Y)

Test the hypothesis that the degree of diabetes progression is different between men and women.

Code:

```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import ttest_ind
from scipy.stats import pearsonr


# Define the dataset
data = {
    'AGE': [59, 48, 72],
    'SEX': [2, 1, 2],
    'BMI': [32.1, 21.6, 30.5],
    'BP': [101.0, 87.0, 93.0],
    'S1': [157, 183, 156],
    'S2': [93.2, 103.2, 93.6],
    'S3': [38.0, 70.0, 41.0],
    'S4': [4.0, 3.0, 4.0],
    'S5': [4.8598, 3.8918, 4.0],
    'S6': [87, 69, 85],
```

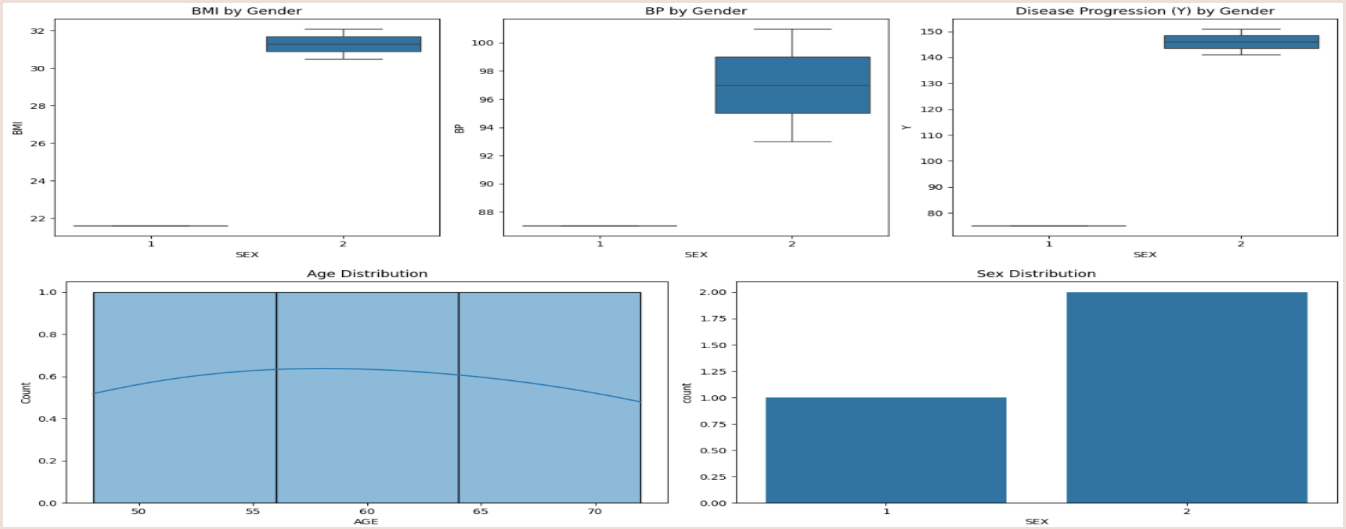
```
    'Y': [151, 75, 141]
}

# Convert the dictionary into a DataFrame
df = pd.DataFrame(data)
# Display the first few rows of the dataset
print(df.head())

# Compute mean values and variance for all values
mean_values = df.mean()
variance_values = df.var()

print("Mean values:\n", mean_values)
print("\nVariance values:\n", variance_values)

# Boxplots for BMI, BP and Y depending on gender
plt.figure(figsize=(15, 5))

plt.subplot(1, 3, 1)
sns.boxplot(x='SEX', y='BMI', data=df)
plt.title('BMI by Gender')

plt.subplot(1, 3, 2)
sns.boxplot(x='SEX', y='BP', data=df)
plt.title('BP by Gender')
plt.subplot(1, 3, 3)
```

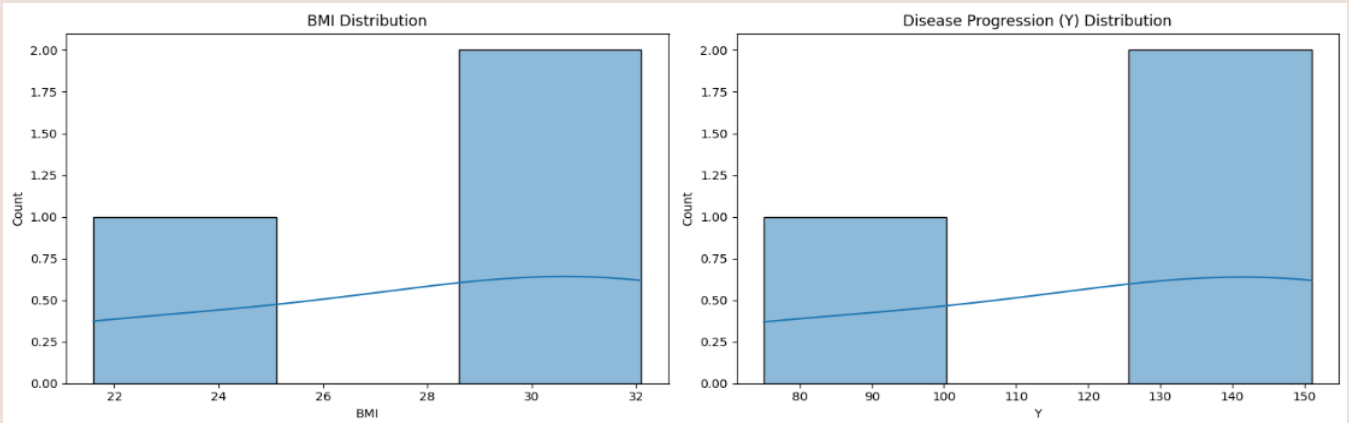
```
sns.boxplot(x='SEX', y='Y', data=df)

plt.title('Disease Progression (Y) by Gender')

plt.tight_layout()
plt.show()

# Distribution of Age, Sex, BMI, and Y variables
plt.figure(figsize=(15, 10))
plt.subplot(2, 2, 1)
sns.histplot(df['AGE'], kde=True)
plt.title('Age Distribution')

plt.subplot(2, 2, 2)
sns.countplot(x='SEX', data=df)
plt.title('Sex Distribution')

plt.subplot(2, 2, 3)
sns.histplot(df['BMI'], kde=True)
plt.title('BMI Distribution')

plt.subplot(2, 2, 4)
sns.histplot(df['Y'], kde=True)
plt.title('Disease Progression (Y) Distribution')

plt.tight_layout()
plt.show()
```

```
# Test the correlation between different variables and disease progression (Y)
correlation_matrix = df.corr()
print("Correlation Matrix:\n", correlation_matrix)

# Correlation values with Y
correlation_with_Y = correlation_matrix['Y'].sort_values(ascending=False)
print("\nCorrelation with Disease Progression (Y):\n", correlation_with_Y)

# Test the hypothesis that the degree of diabetes progression is different between men and women
men_Y = df[df['SEX'] == 1]['Y']
women_Y = df[df['SEX'] == 2]['Y']

t_stat, p_value = ttest_ind(men_Y, women_Y)
print(f"\nT-test results:\nT-statistic: {t_stat}, P-value: {p_value}")

if p_value < 0.05:
    print("Reject the null hypothesis: The degree of diabetes progression is different between men and women.")
else:
    print("Fail to reject the null hypothesis: The degree of diabetes progression is not significantly different between men and women.")
```

Output:

```
   AGE  SEX   BMI     BP   S1     S2    S3   S4      S5  S6    Y
0   59    2  32.1  101.0  157   93.2  38.0  4.0  4.8598  87  151
1   48    1  21.6   87.0  183  103.2  70.0  3.0  3.8918  69   75
2   72    2  30.5   93.0  156   93.6  41.0  4.0  4.0000  85  141
Mean values:
 AGE      59.666667
SEX        1.666667
BMI       28.066667
BP        93.666667
S1       165.333333
S2        96.666667
S3        49.666667
S4         3.666667
S5         4.250533
S6        80.333333
Y        122.333333
dtype: float64

Variance values:
 AGE      144.333333
SEX         0.333333
BMI        32.003333
BP         49.333333
S1        234.333333
S2         32.053333
S3        312.333333
S4          0.333333
S5          0.281331
S6         97.333333
Y        1705.333333
dtype: float64
```





```
Correlation Matrix:
          AGE       SEX       BMI        BP        S1        S2        S3  \
AGE  1.000000  0.840996  0.756034  0.383175 -0.858219 -0.821359 -0.792041
SEX  0.840996  1.000000  0.989951  0.821995 -0.999466 -0.999376 -0.996392
BMI  0.756034  0.989951  1.000000  0.894269 -0.984803 -0.994328 -0.998381
BP   0.383175  0.821995  0.894269  1.000000 -0.802955 -0.841600 -0.867365
S1  -0.858219 -0.999466 -0.984803 -0.802955  1.000000  0.997689  0.993088
S2  -0.821359 -0.999376 -0.994328 -0.841600  0.997689  1.000000  0.998768
S3  -0.792041 -0.996392 -0.998381 -0.867365  0.993088  0.998768  1.000000
S4   0.840996  1.000000  0.989951  0.821995 -0.999466 -0.999376 -0.996392
S5   0.054073  0.585725  0.694456  0.943044 -0.558938 -0.613991 -0.652403
S6   0.781825  0.994850  0.999186  0.875486 -0.991008 -0.997809 -0.999863
Y    0.769301  0.992643  0.999790  0.884901 -0.988159 -0.996301 -0.999338

          S4        S5        S6         Y
AGE  0.840996  0.054073  0.781825  0.769301
SEX  1.000000  0.585725  0.994850  0.992643
BMI  0.989951  0.694456  0.999186  0.999790
BP   0.821995  0.943044  0.875486  0.884901
S1  -0.999466 -0.558938 -0.991008 -0.988159
S2  -0.999376 -0.613991 -0.997809 -0.996301
S3  -0.996392 -0.652403 -0.999863 -0.999338
S4   1.000000  0.585725  0.994850  0.992643
S5   0.585725  1.000000  0.664862  0.679550
S6   0.994850  0.664862  1.000000  0.999803
Y    0.992643  0.679550  0.999803  1.000000

Correlation with Disease Progression (Y):
 Y      1.000000
S6     0.999803
BMI    0.999790
SEX    0.992643
S4     0.992643
BP     0.884901
AGE    0.769301
S5     0.679550
S1    -0.988159
S2    -0.996301
S3    -0.999338
Name: Y, dtype: float64

T-test results:
T-statistic: -8.198373822492686, P-value: 0.07727025676475614
Fail to reject the null hypothesis: The degree of diabetes progression is not significantly different between men and women.
```

//Comments for understanding

I am familiar with Python and am currently working on my final year project using the Python language. I studied Machine Learning in the previous semester, so I also have a basic understanding of it.