QMP 7.1 D/F

Channabasaveshwara Institute of Technology
(Affiliated to VTU, Belagavi & Approved by AICTE, New Delhi)
(NAAC Accredited & ISO 9001:2015 Certified Institution)
NH 206 (B.H. Road), Gubbi, Tumkur – 572 216. Karnataka.

Department of Artificial Intelligence
and Data Science

Statistical Machine Learning for Data Science
BAD702

(CBCS SCHEME)

B.E - VII Semester

Lab Manual 2025-26

Name: ____

USN: ____

Batch:_________________________Section: ___________________________

QMP 7.1 D/F

Channabasaveshwara Institute of Technology
(Affiliated to VTU, Belagavi & Approved by AICTE, New Delhi)
(**NAAC Accredited & ISO 9001:2015 Certified Institution)**
NH 206 (B.H. Road), Gubbi, Tumkur – 572 216. Karnataka.

Department of Artificial Intelligence and Data Science

Statistical Machine Learning for Data Science BAD702

(PRACTICAL COMPONENT OF IPCC)

Prepared by:

Mrs. Tejaswini S
Assistant Professor
AD Department
CIT. Gubbi

SYLLABUS

Statistical Machine Learning for Data Science

PRACTICAL COMPONENT OF IPCC

[As per Choice Based Credit System (CBCS) scheme]

SEMESTER – VII (AD)

Subject Code: BAD702 **CIE Marks: 20**

Hours/ Week : 02 (02 Hours Laboratory) **Test Hours: 03**

Using suitable simulation software, demonstrate the operation of the following programs:

Sl.No	Experiments
1.	A dataset contains the prices of houses in a city. Find the 25th and 75th percentiles and calculate the Inter quartile range (IQR). How does the IQR help in understanding the price variability?
2.	You are given a dataset with categorical variables about customer satisfaction levels (Low, Medium, High) and whether customers made repeat purchases (Yes/No). Create visualizations such as bar plots or stacked bar charts to explore the relationship between satisfaction level and repeat purchases. What can you infer from the data?
3.	A dataset contains information about car models, including the engine size (in Liters), fuel efficiency (miles per gallon), and car price. Use a pair plot or correlation matrix to explore the relationships between these variables. Which variables seem to have the strongest relationships, and what might be the practical significance of these findings?
4.	You want to estimate the mean salary of software engineers in a country. You take 10 different random samples, each containing 50 engineers, and calculate the sample mean for each. Plot the distribution of these sample means. How does the Central Limit Theorem explain the shape of this sampling distribution, even if the underlying salary distribution is skewed?
5.	A researcher conducts an experiment with a sample of 20 participants to determine if a new drug affects heart rate. The sample has a mean heart rate increase of 8 beats per minute and a standard deviation of 2beats per minute. Perform a hypothesis test using the t-distribution to determine if the mean heart rate increase is significantly different from zero at the 5% significance level.
6.	A company is testing two versions of a webpage (A and B) to determine which version leads to more sales. Version A was shown to 1,000 users and resulted in 120 sales. Version B was shown to 1,200 users and resulted in 150 sales. Perform an A/B test to determine if there is a statistically significant difference in the conversion rates between the two versions. Use a 5% significance level.
7.	You are comparing the average daily sales between two stores. Store A has a mean daily sales value of$1,000 with a standard deviation of $100 over 30 days, and Store B has a mean daily sales value of $950with a standard deviation of $120 over 30 days. Conduct a two-sample t-test to determine if there is a significant difference between the average sales of the two stores at the 5% significance level.

8.	A company collects data on employees' salaries and records their education level as a categorical variable with three levels: "High School", "Bachelor's", and "Master's". Fit a multiple linear regression model to predict salary using education level (as a factor variable) and years of experience. Interpret the coefficients for the education levels in the regression model.
9.	You have data on housing prices and square footage and notice that the relationship between square footage and price is nonlinear. Fit a spline regression model to allow the relationship between square footage and price to change at 2,000 square feet. Explain how spline regression can capture different behaviours of the relationship before and after 2,000 square feet.
10.	A hospital is using a Poisson regression model (a type of GLM) to predict the number of emergency room visits per week based on patient age and medical history. The model is given by: $\text{Log}(\lambda) =2.5-0.03*\text{Age}+0.5*\text{condition}$ where λ is the expected number of visits per week, Age is the patient's age, and condition is a binary variable (1 if the patient has a chronic condition, 0 otherwise). Interpret the coefficients of Age and condition. What is the expected number of visits per week for a 60-year-old patient with a chronic condition? How would the expected number of visits change if the patient did not have a chronic condition?
11.	A bakery claims that its new cookie recipe is lower in calories compared to the old recipe, which had a mean calorie count of 200. You sample 40 new cookies and find a mean of 190 calories with a standard deviation of 15 calories. Perform a one-tailed t-test to determine if the new recipe has significantly fewer calories at a 5% significance level.

General Instructions to Students

1. Students should come with thorough preparation for the experiment to be conducted.
2. Students should take prior permission from the concerned faculty before availing the leave.
3. Students should come with formals and to be present on time in the laboratory.
4. Students will not be permitted to attend the laboratory unless they bring the practical record fully completed in all respects pertaining to the experiments conducted in the previous session.
5. Students will be permitted to attend the laboratory unless they bring the observation book fully completed in all respects pertaining to the experiments conducted in the present session.
6. They should obtain the signature of the staff-in –charge in the observation book after completing each experiment.
7. Practical record should be neatly maintained.
8. Ask lab Instructor for assistance for any problem.
9. Do not download or install software without the assistance of laboratory Instructor.
10. Do not alter the configuration of system.
11. Turn off the systems after use.

Program 1

A dataset contains the prices of houses in a city. Find the 25th and 75th percentiles and calculate the Inter quartile range (IQR). How does the IQR help in understanding the price variability?

```
import pandas as pd

# Load the dataset (assuming it's in a CSV file named 'house_prices.csv' with a 'price' column)
df = pd.read_csv('Housing.csv')

# Calculate the 25th percentile (Q1)
q1 = df['price'].quantile(0.25)

# Calculate the 75th percentile (Q3)
q3 = df['price'].quantile(0.75)

# Calculate the Interquartile Range (IQR)
iqr = q3 - q1

print(f"25th Percentile (Q1): {q1:.2f}")
print(f"75th Percentile (Q3): {q3:.2f}")
print(f"Interquartile Range (IQR): {iqr:.2f}")
```

Output:

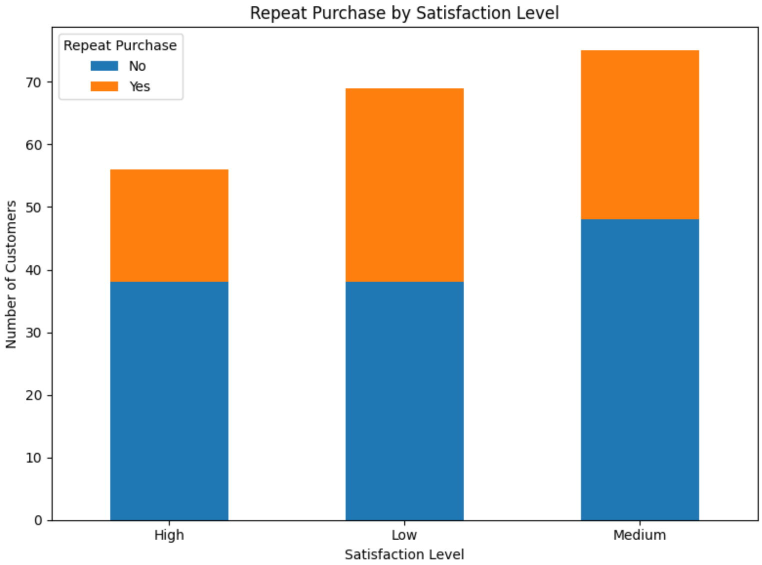
```
25th Percentile (Q1): 3430000.00
75th Percentile (Q3): 5740000.00
Interquartile Range (IQR): 2310000.00
```

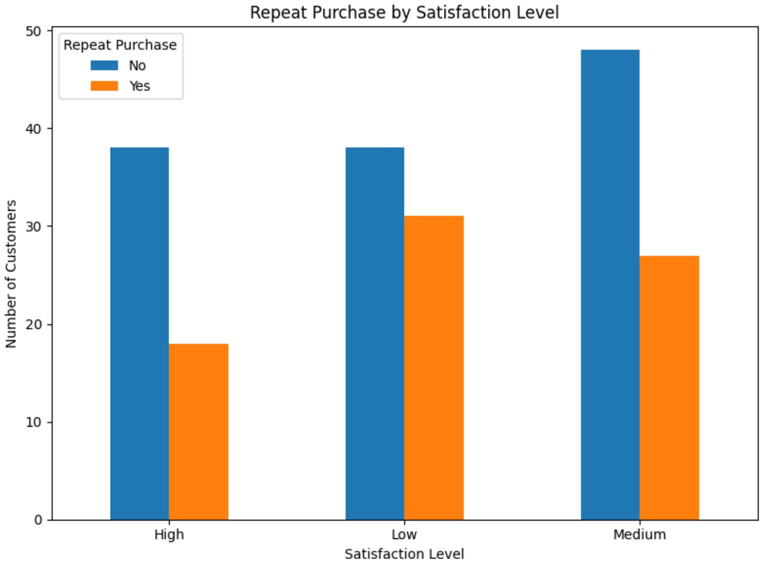
Program 2

You are given a dataset with categorical variables about customer satisfaction levels (Low, Medium, High) and whether customers made repeat purchases (Yes/No). Create visualizations such as bar plots or stacked bar charts to explore the relationship between satisfaction level and repeat purchases. What can you infer from the data?

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

# Create a dummy dataset for demonstration
data = {
    'Satisfaction Level': np.random.choice(['Low', 'Medium', 'High'], size=200),
    'Repeat Purchase': np.random.choice(['Yes', 'No'], size=200, p=[0.4, 0.6]) # Assuming 40% repeat
purchases
}
df_customer = pd.DataFrame(data)

# Create a cross-tabulation of the two variables
cross_tab = pd.crosstab(df_customer['Satisfaction Level'], df_customer['Repeat Purchase'])

# Create a stacked bar chart
cross_tab.plot(kind='bar', stacked=True, figsize=(8, 6))
plt.title('Repeat Purchase by Satisfaction Level')
plt.xlabel('Satisfaction Level')
plt.ylabel('Number of Customers')
plt.xticks(rotation=0)
plt.legend(title='Repeat Purchase')
plt.tight_layout()
plt.show()

# Create a grouped bar chart
cross_tab.plot(kind='bar', stacked=False, figsize=(8, 6))
plt.title('Repeat Purchase by Satisfaction Level')
plt.xlabel('Satisfaction Level')
plt.ylabel('Number of Customers')
plt.xticks(rotation=0)
plt.legend(title='Repeat Purchase')
plt.tight_layout()
plt.show()
```

Output:





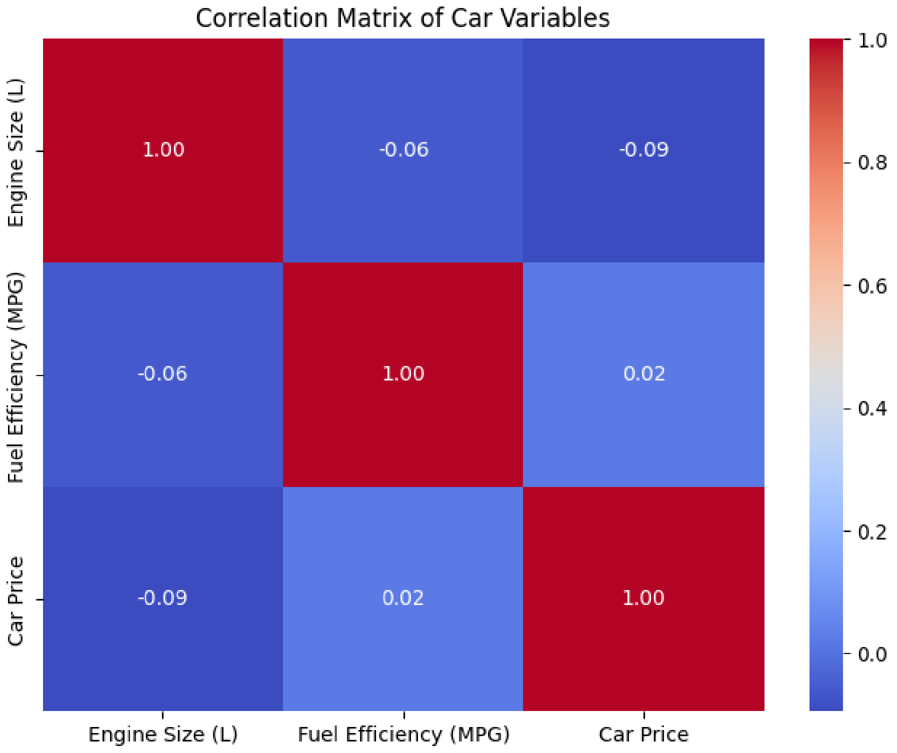
```
Repeat Purchase     No   Yes
Satisfaction Level
        High        38    18
        Low         38    31
       Medium       48    27
```

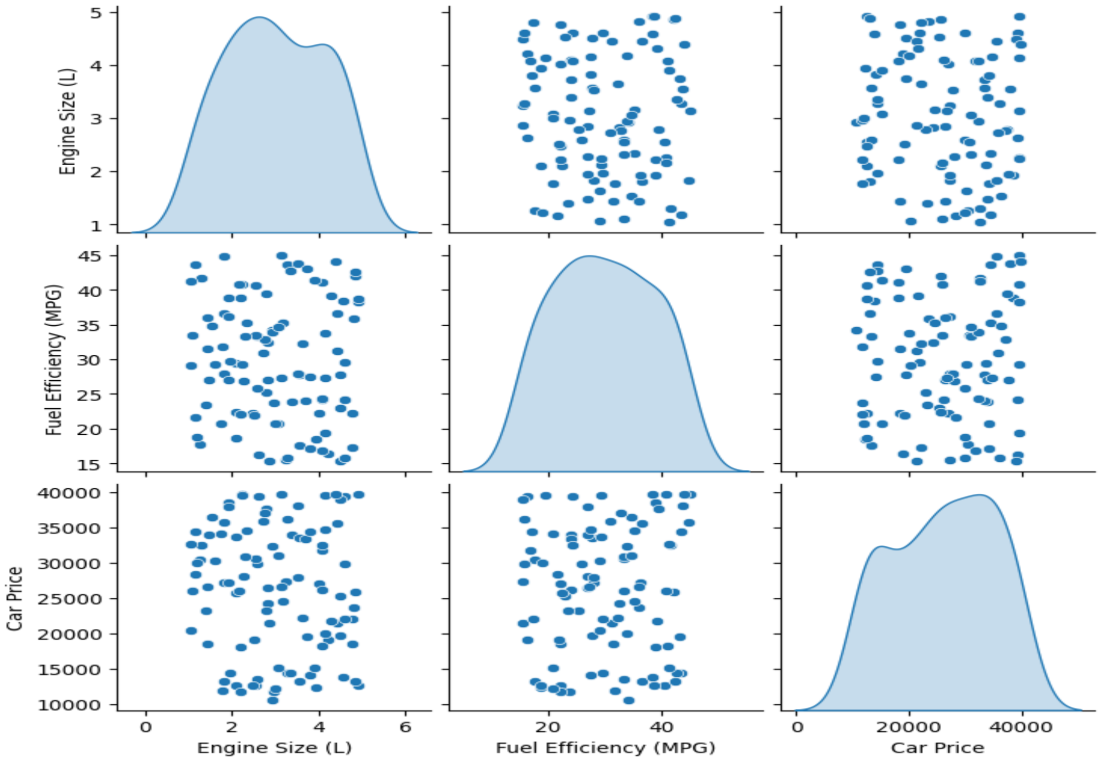
Program 3

A dataset contains information about car models, including the engine size (in Liters), fuel efficiency (miles per gallon), and car price. Use a pair plot or correlation matrix to explore the relationships between these variables. Which variables seem to have the strongest relationships, and what might be the practical significance of these findings?

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

# Create a dummy dataset for demonstration
data = {
    'Engine Size (L)': np.random.rand(100) * 4 + 1,  # Engine size between 1 and 5 Liters
    'Fuel Efficiency (MPG)': np.random.rand(100) * 30 + 15, # Fuel efficiency between 15 and 45 MPG
    'Car Price': np.random.rand(100) * 30000 + 10000 # Car price between $10,000 and $40,000
}
df_cars = pd.DataFrame(data)

# Calculate the correlation matrix
correlation_matrix = df_cars[['Engine Size (L)', 'Fuel Efficiency (MPG)', 'Car Price']].corr()

# Visualize the correlation matrix using a heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix of Car Variables')
plt.show()

sns.pairplot(df_cars[['Engine Size (L)', 'Fuel Efficiency (MPG)', 'Car Price']], diag_kind='kde')
plt.show()
```

Output :


Correlation Matrix of Car Variables
Engine Size (L)
Fuel Efficiency (MPG)
Car Price
1.00
-0.06
-0.09
-0.06
1.00
0.02
-0.09
0.02
1.00
1.0
0.8
0.6
0.4
0.2
0.0


Engine Size (L)
Fuel Efficiency (MPG)
Car Price
5
4
3
2
1
45
40
35
30
25
20
15
40000
35000
30000
25000
20000
15000
10000
0
2
4
6
20
40
0
20000
40000

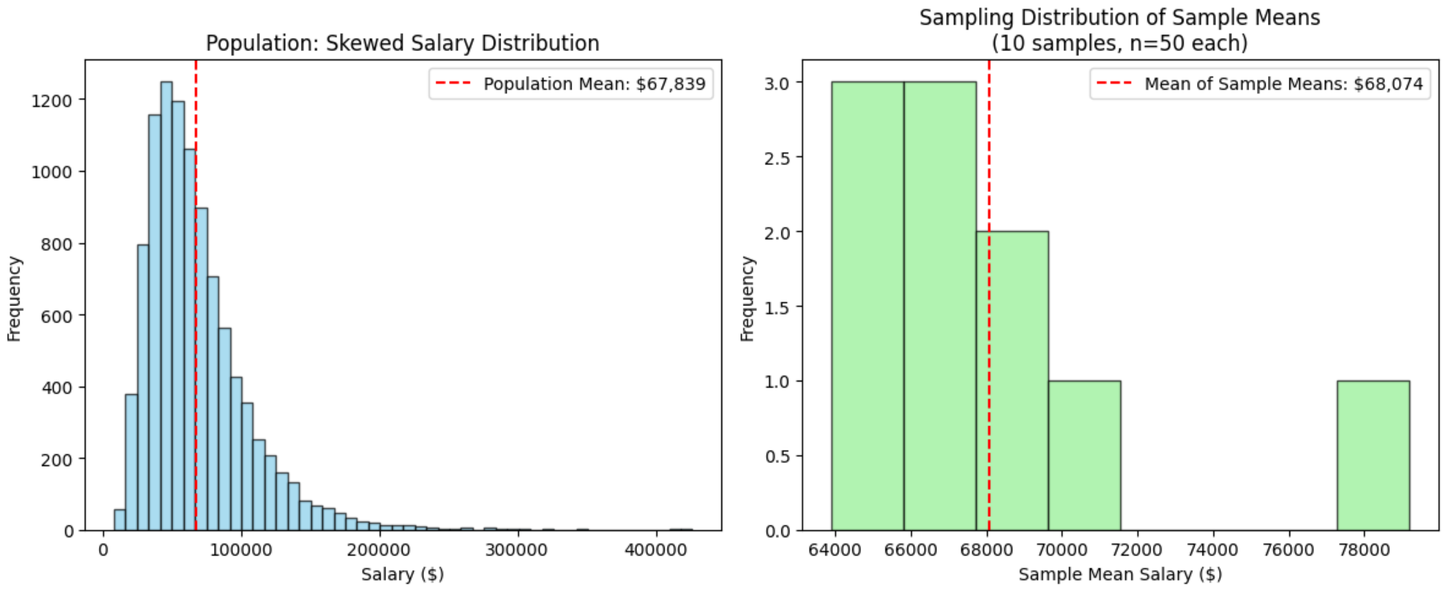
Program 4

You want to estimate the mean salary of software engineers in a country. You take 10 different random samples, each containing 50 engineers, and calculate the sample mean for each. Plot the distribution of these sample means. How does the Central Limit Theorem explain the shape of this sampling distribution, even if the underlying salary distribution is skewed?

```
# Set random seed for reproducibility
np.random.seed(42)

# Create a skewed population (log-normal distribution for salaries)
population_size = 10000
population = np.random.lognormal(mean=11, sigma=0.5, size=population_size)

# Take 10 samples of 50 engineers each and calculate sample means
n_samples = 10
sample_size = 50
sample_means = []

for i in range(n_samples):
    sample = np.random.choice(population, size=sample_size, replace=False)
    sample_means.append(np.mean(sample))

# Create plots
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 5))

# Plot 1: Original skewed population
ax1.hist(population, bins=50, alpha=0.7, color='skyblue', edgecolor='black')
ax1.set_title('Population: Skewed Salary Distribution')
ax1.set_xlabel('Salary ($)')
ax1.set_ylabel('Frequency')
ax1.axvline(np.mean(population), color='red', linestyle='--',
          label=f'Population Mean: ${np.mean(population):,.0f}')
ax1.legend()

# Plot 2: Distribution of sample means
ax2.hist(sample_means, bins=8, alpha=0.7, color='lightgreen', edgecolor='black')
ax2.set_title('Sampling Distribution of Sample Means\n(10 samples, n=50 each)')
ax2.set_xlabel('Sample Mean Salary ($)')
ax2.set_ylabel('Frequency')
ax2.axvline(np.mean(sample_means), color='red', linestyle='--',
          label=f'Mean of Sample Means: ${np.mean(sample_means):,.0f}')
ax2.legend()

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

```
# Print results
print(f"Population mean: ${np.mean(population):,.2f}")
print(f"Population std: ${np.std(population):,.2f}")
print(f"Sample means: {[f'${x:,.0f}' for x in sample_means]}")
print(f"Mean of sample means: ${np.mean(sample_means):,.2f}")
print(f"Std of sample means: ${np.std(sample_means):,.2f}")

print("\nCentral Limit Theorem Explanation:")
print("Even though the population is right-skewed, the sample means")
print("tend to be normally distributed around the population mean.")
print("With larger sample sizes, this becomes even more pronounced.")
```

Output:



```
Population mean: $67,839.02
Population std: $36,379.18
Sample means: ['$69,907', '$63,901', '$66,804', '$66,606', '$68,435', '$64,951',
'$65,673', '$66,894', '$79,199', '$68,369']
Mean of sample means: $68,073.80
Std of sample means: $4,067.48

Central Limit Theorem Explanation:
Even though the population is right-skewed, the sample means
tend to be normally distributed around the population mean.
With larger sample sizes, this becomes even more pronounced.
```

Program 5

A researcher conducts an experiment with a sample of 20 participants to determine if a new drug affects heart rate. The sample has a mean heart rate increase of 8 beats per minute and a standard deviation of 2beats per minute. Perform a hypothesis test using the t-distribution to determine if the mean heart rate increase is significantly different from zero at the 5% significance level.

```
import numpy as np
from scipy import stats

# Sample data
n = 20
mean_sample = 8
std_dev = 2
mu0 = 0  # hypothesized mean

# Calculate t statistic
t_stat = (mean_sample - mu0) / (std_dev / np.sqrt(n))

# Degrees of freedom
df = n - 1

# Two-tailed p-value
p_value = 2 * (1 - stats.t.cdf(abs(t_stat), df))

# Critical value for 95% confidence (two-tailed)
t_critical = stats.t.ppf(1 - 0.025, df)

print("t statistic:", round(t_stat, 3))
print("Degrees of freedom:", df)
print("Critical value (±):", round(t_critical, 3))
print("p-value:", p_value)

if abs(t_stat) > t_critical:
    print("Result: Reject H0 (significant effect).")
else:
    print("Result: Fail to reject H0 (no significant effect).")
```

Output:

```
t statistic: 17.889
Degrees of freedom: 19
Critical value (±): 2.093
p-value: 2.395861287141088e-13
Result: Reject H0 (significant effect).
```

Calculation:

We are testing whether the mean increase in heart rate is significantly different from zero.

Step 1: State the hypotheses

Null hypothesis (H_0): μ = 0 (no increase in heart rate)
Alternative hypothesis (H_1): μ ≠ 0 (there is an increase or decrease)

This is a two-tailed t-test.

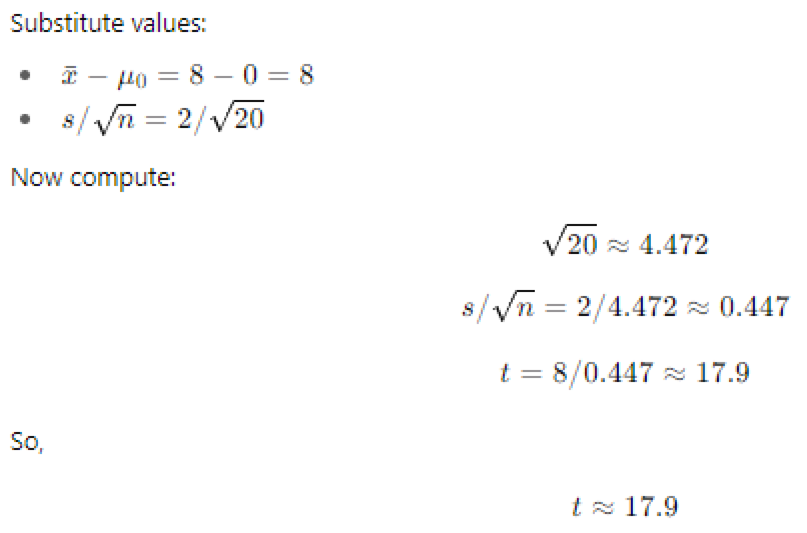
Step 2: Gather sample statistics

Sample size: n = 20
Sample mean: $\bar{x}$ = 8
Standard deviation: s = 2
Hypothesized mean: μ_0 = 0

Step 3: Compute the test statistic

The t-statistic formula is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$



Substitute values:

- $\bar{x} - \mu_0 = 8 - 0 = 8$
- $s/\sqrt{n} = 2/\sqrt{20}$

Now compute:

$$\sqrt{20} \approx 4.472$$

$$s/\sqrt{n} = 2/4.472 \approx 0.447$$

$$t = 8/0.447 \approx 17.9$$

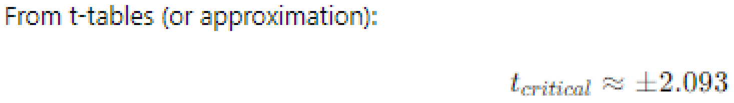
So,

$$t \approx 17.9$$

Step 4: Degrees of freedom

$$df = n - 1 = 20 - 1 = 19$$

Step 5: Critical t-value (α = 0.05, two-tailed, df = 19)



From t-tables (or approximation):

$$t_{critical} \approx \pm 2.093$$

Step 6: Decision rule

If |t| > 2.093 , reject H_0.
Our t = 17.9, which is far greater.

Conclusion

Since t = 17.9 >> 2.093$, we reject the null hypothesis.

At the 5% significance level, there is strong evidence that the new drug causes a significant change (increase) in heart rate.

Program 6

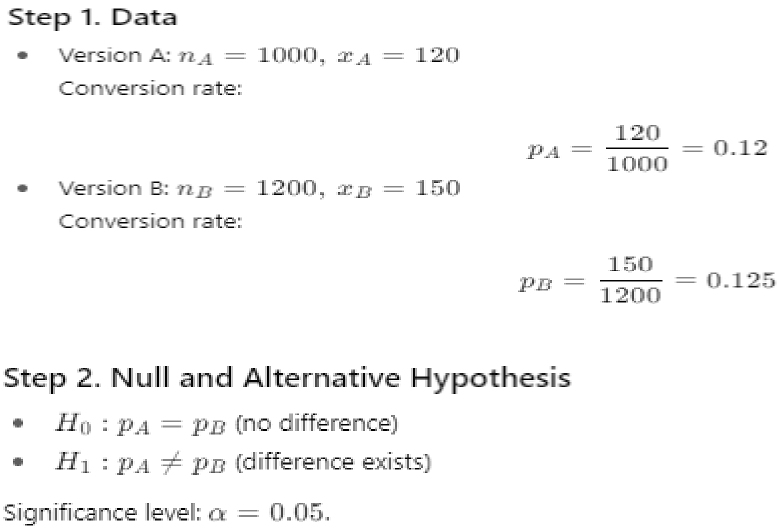
A company is testing two versions of a webpage (A and B) to determine which version leads to more sales. Version A was shown to 1,000 users and resulted in 120 sales. Version B was shown to 1,200 users and resulted in 150 sales. Perform an A/B test to determine if there is a statistically significant difference in the conversion rates between the two versions. Use a 5% significance level.

```python
import numpy as np
from statsmodels.stats.proportion import proportions_ztest

# Data
n_A, x_A = 1000, 120
n_B, x_B = 1200, 150

# Conversion rates
p_A = x_A / n_A
p_B = x_B / n_B
print("Conversion Rate A:", p_A)
print("Conversion Rate B:", p_B)

# Perform two-proportion z-test
count = np.array([x_A, x_B])      # successes
nobs = np.array([n_A, n_B])       # trials

z_stat, p_value = proportions_ztest(count, nobs, alternative='two-sided')

print("Z-statistic:", z_stat)
print("P-value:", p_value)

alpha = 0.05
if p_value < alpha:
    print("Reject H0: Significant difference in conversion rates.")
else:
    print("Fail to reject H0: No significant difference in conversion rates.")
```

Output:

```
Conversion Rate A: 0.12
Conversion Rate B: 0.125
Z-statistic: -0.3558864321126734
P-value: 0.7219256170797248

Fail to reject H0: No significant difference in conversion rates.
```

Calculation:



Step 1. Data

- Version A: $n_A = 1000,\ x_A = 120$
  Conversion rate:

$$p_A = \frac{120}{1000} = 0.12$$

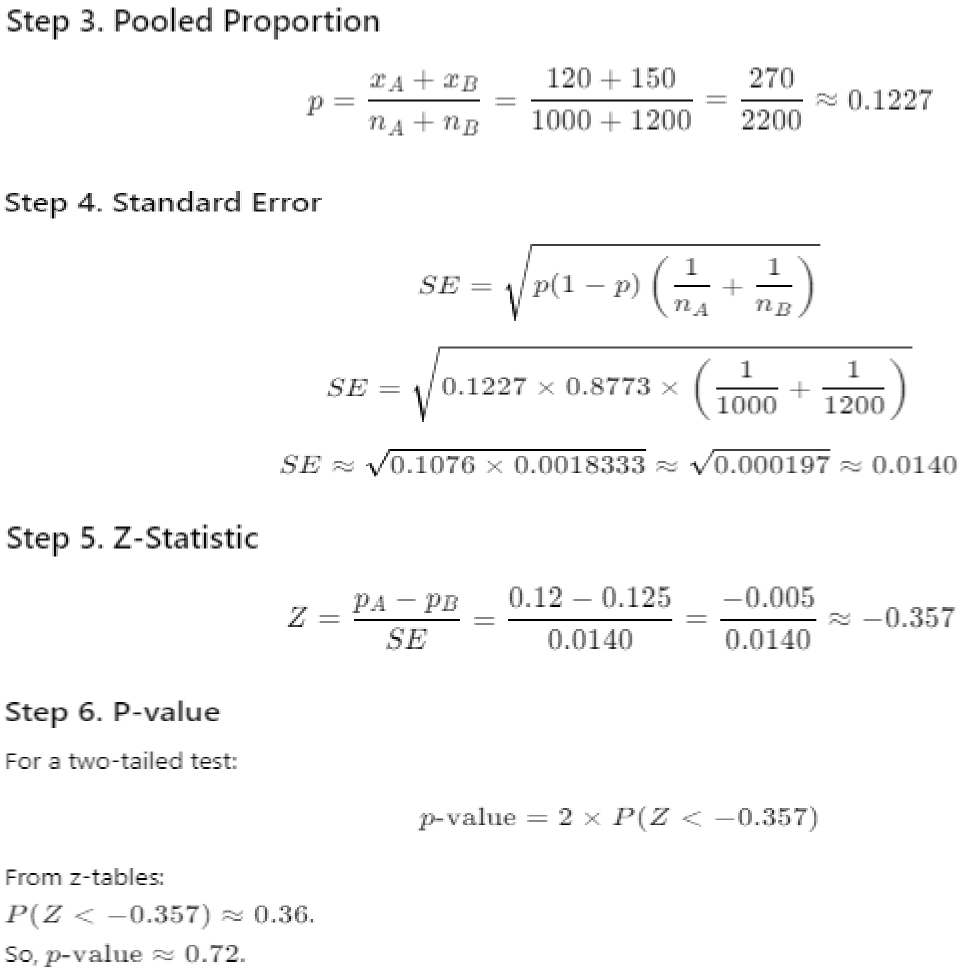
- Version B: $n_B = 1200,\ x_B = 150$
  Conversion rate:

$$p_B = \frac{150}{1200} = 0.125$$

Step 2. Null and Alternative Hypothesis

- $H_0 : p_A = p_B$ (no difference)
- $H_1 : p_A \neq p_B$ (difference exists)

Significance level: $\alpha = 0.05$.



Step 3. Pooled Proportion

$$p = \frac{x_A + x_B}{n_A + n_B} = \frac{120 + 150}{1000 + 1200} = \frac{270}{2200} \approx 0.1227$$

Step 4. Standard Error

$$SE = \sqrt{p(1-p)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

$$SE = \sqrt{0.1227 \times 0.8773 \times \left(\frac{1}{1000} + \frac{1}{1200}\right)}$$

$$SE \approx \sqrt{0.1076 \times 0.0018333} \approx \sqrt{0.000197} \approx 0.0140$$

Step 5. Z-Statistic

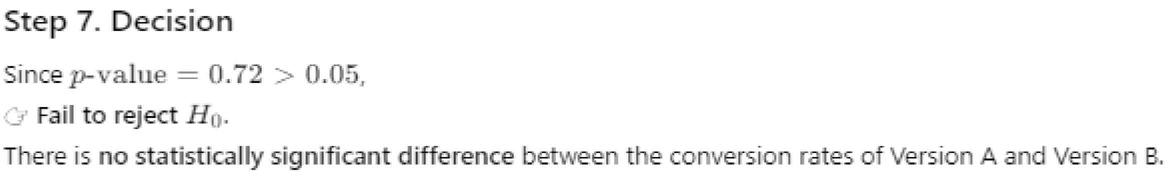
$$Z = \frac{p_A - p_B}{SE} = \frac{0.12 - 0.125}{0.0140} = \frac{-0.005}{0.0140} \approx -0.357$$

Step 6. P-value

For a two-tailed test:

$$p\text{-value} = 2 \times P(Z < -0.357)$$

From z-tables:
$P(Z < -0.357) \approx 0.36.$
So, $p\text{-value} \approx 0.72.$



Step 7. Decision

Since $p\text{-value} = 0.72 > 0.05$,
☞ Fail to reject H_0.
There is **no statistically significant difference** between the conversion rates of Version A and Version B.

Program 7

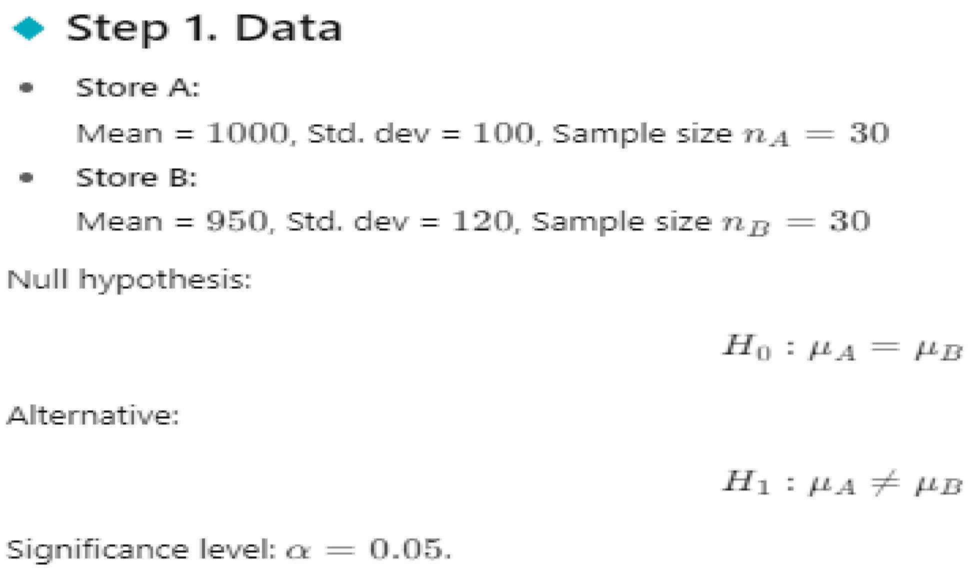
You are comparing the average daily sales between two stores. Store A has a mean daily sales value of $1,000 with a standard deviation of $100 over 30 days, and Store B has a mean daily sales value of $950 with a standard deviation of $120 over 30 days. Conduct a two-sample t-test to determine if there is a significant difference between the average sales of the two stores at the 5% significance level.

```
import scipy.stats as stats

# Given data for Store A
mean_A = 1000
std_A = 100
n_A = 30

# Given data for Store B
mean_B = 950
std_B = 120
n_B = 30

# Significance level
alpha = 0.05

# Perform independent two-sample t-test (assuming equal variances, or check for unequal variances)
# For simplicity, we'll assume equal variances here. If variances are significantly different,
# use the Welch's t-test by setting equal_var=False in stats.ttest_ind
t_statistic, p_value = stats.ttest_ind_from_stats(mean1=mean_A, std1=std_A, nobs1=n_A,
                                                  mean2=mean_B, std2=std_B, nobs2=n_B,
                                                  equal_var=True) # Set equal_var=False for Welch's t-test

print(f"Mean daily sales for Store A: ${mean_A}")
print(f"Standard deviation for Store A: ${std_A}")
print(f"Sample size for Store A: {n_A}")
print(f"Mean daily sales for Store B: ${mean_B}")
print(f"Standard deviation for Store B: ${std_B}")
print(f"Sample size for Store B: {n_B}")

print(f"\nT-statistic: {t_statistic:.4f}")
print(f"P-value: {p_value:.4f}")

# Compare p-value with the significance level
if p_value < alpha:
    print("\nResult: Reject the null hypothesis. There is a statistically significant difference in average daily sales
between the two stores.")
else:
    print("\nResult: Fail to reject the null hypothesis. There is no statistically significant difference in average
daily sales between the two stores.")
```

Output:

```
Mean daily sales for Store A: $1000
Standard deviation for Store A: $100
Sample size for Store A: 30
Mean daily sales for Store B: $950
Standard deviation for Store B: $120
Sample size for Store B: 30

T-statistic: 1.7532
P-value: 0.0848

Result: Fail to reject the null hypothesis. There is no statistically significant difference in average
daily sales between the two stores.
```

Calculation:



◆ Step 1. Data

- Store A:

  Mean = 1000, Std. dev = 100, Sample size $n_A = 30$
- Store B:

  Mean = 950, Std. dev = 120, Sample size $n_B = 30$

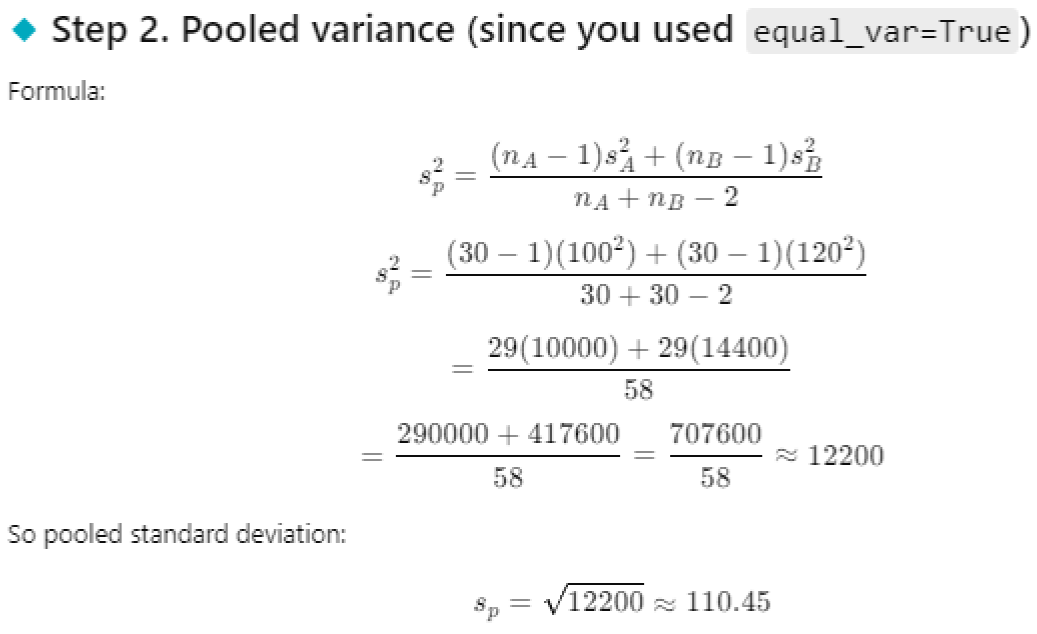
Null hypothesis:

$$H_0 : \mu_A = \mu_B$$

Alternative:

$$H_1 : \mu_A \neq \mu_B$$

Significance level: $\alpha = 0.05$.



◆ Step 2. Pooled variance (since you used `equal_var=True`)

Formula:

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$$

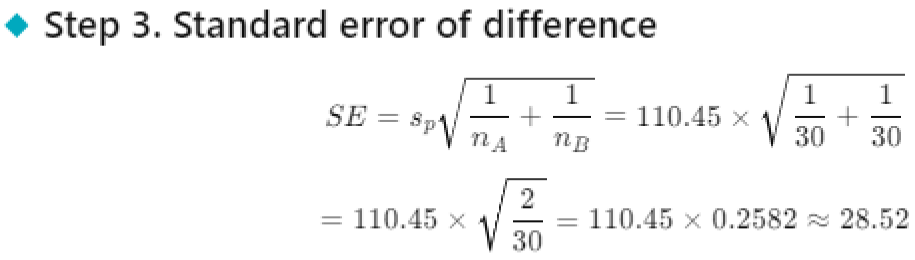
$$s_p^2 = \frac{(30-1)(100^2) + (30-1)(120^2)}{30 + 30 - 2}$$

$$= \frac{29(10000) + 29(14400)}{58}$$

$$= \frac{290000 + 417600}{58} = \frac{707600}{58} \approx 12200$$

So pooled standard deviation:

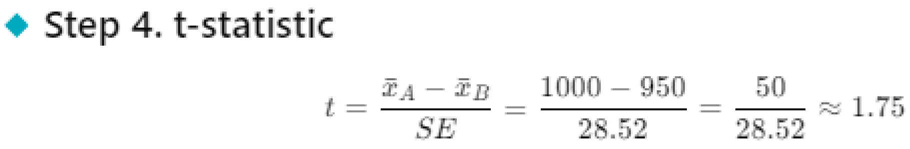
$$s_p = \sqrt{12200} \approx 110.45$$



◆ Step 3. Standard error of difference

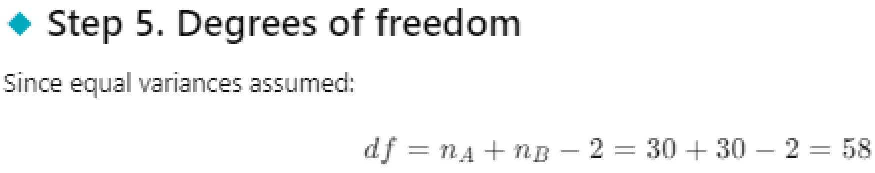
$$SE = s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = 110.45 \times \sqrt{\frac{1}{30} + \frac{1}{30}}$$

$$= 110.45 \times \sqrt{\frac{2}{30}} = 110.45 \times 0.2582 \approx 28.52$$



◆ Step 4. t-statistic

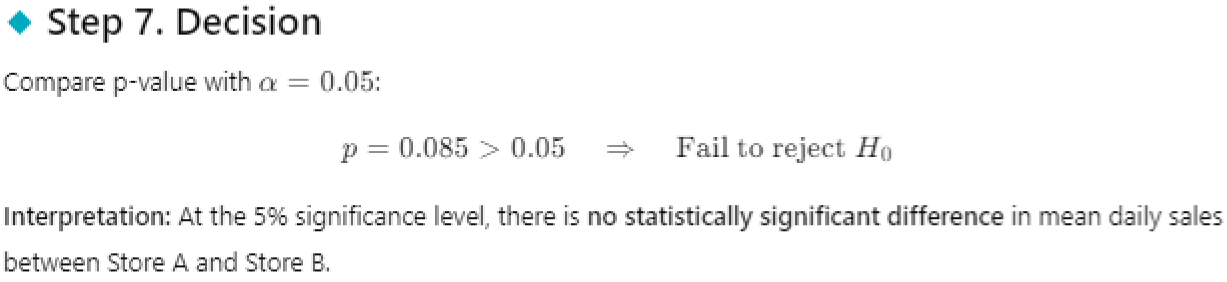
$$t = \frac{\bar{x}_A - \bar{x}_B}{SE} = \frac{1000 - 950}{28.52} = \frac{50}{28.52} \approx 1.75$$



◆ Step 5. Degrees of freedom

Since equal variances assumed:

$$df = n_A + n_B - 2 = 30 + 30 - 2 = 58$$



◆ Step 7. Decision

Compare p-value with $\alpha = 0.05$:

$$p = 0.085 > 0.05 \quad \Rightarrow \quad \text{Fail to reject } H_0$$

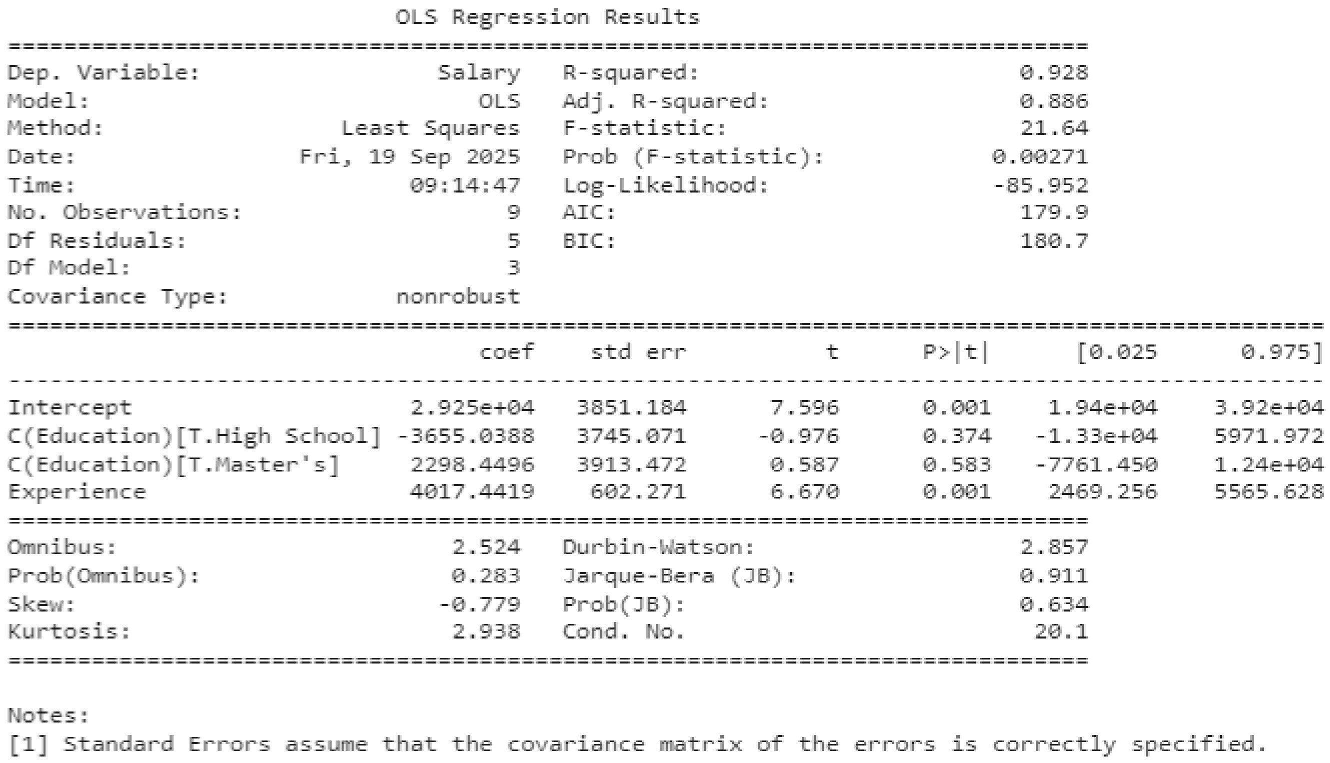
Interpretation: At the 5% significance level, there is **no statistically significant difference** in mean daily sales between Store A and Store B.

Program 8

A company collects data on employees' salaries and records their education level as a categorical variable with three levels: "High School", "Bachelor's", and "Master's". Fit a multiple linear regression model to predict salary using education level (as a factor variable) and years of experience. Interpret the coefficients for the education levels in the regression model.

```
import pandas as pd
import statsmodels.formula.api as smf

# Example dataset
data = pd.DataFrame({
   "Salary": [30000, 35000, 40000, 42000, 50000, 55000, 60000, 62000, 70000],
   "Education": ["High School", "High School", "Bachelor's", "Bachelor's", "Master's", "Master's", "High
School", "Bachelor's", "Master's"],
   "Experience": [1, 3, 2, 5, 4, 6, 8, 7, 10]
})

# Fit regression model
# C(Education) tells statsmodels to treat Education as a categorical (factor) variable
model = smf.ols("Salary ~ Experience + C(Education)", data=data).fit()

# Show results
print(model.summary())
```

Output:



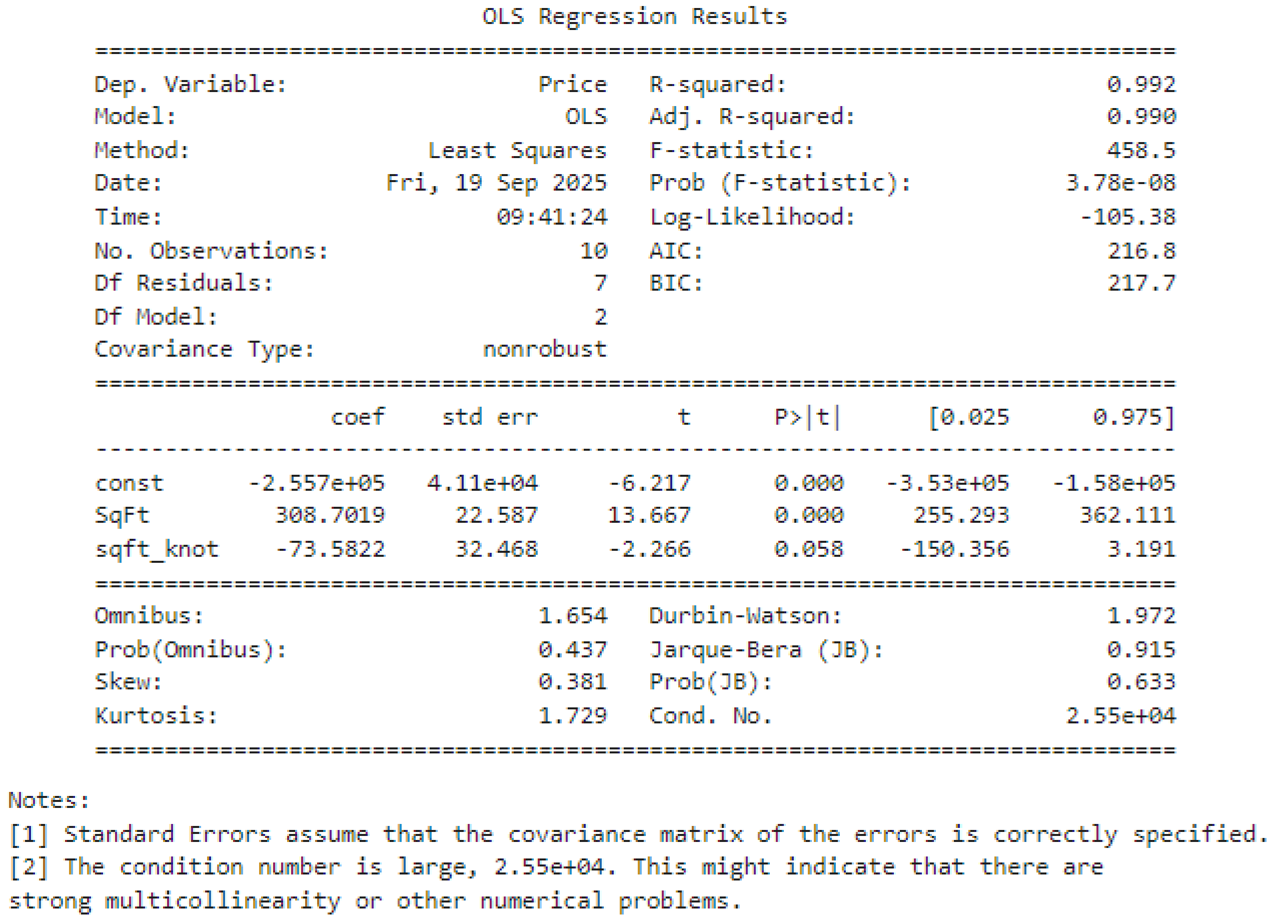
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Salary   R-squared:                       0.928
Model:                            OLS   Adj. R-squared:                  0.886
Method:                 Least Squares   F-statistic:                     21.64
Date:                Fri, 19 Sep 2025   Prob (F-statistic):            0.00271
Time:                        09:14:47   Log-Likelihood:                -85.952
No. Observations:                   9   AIC:                             179.9
Df Residuals:                       5   BIC:                             180.7
Df Model:                           3
Covariance Type:            nonrobust
=================================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
Intercept                    2.925e+04   3851.184      7.596      0.001    1.94e+04    3.92e+04
C(Education)[T.High School] -3655.0388   3745.071     -0.976      0.374   -1.33e+04    5971.972
C(Education)[T.Master's]     2298.4496   3913.472      0.587      0.583   -7761.450    1.24e+04
Experience                   4017.4419    602.271      6.670      0.001    2469.256    5565.628
==============================================================================
Omnibus:                        2.524   Durbin-Watson:                   2.857
Prob(Omnibus):                  0.283   Jarque-Bera (JB):                0.911
Skew:                          -0.779   Prob(JB):                        0.634
Kurtosis:                       2.938   Cond. No.                         20.1
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Program 9

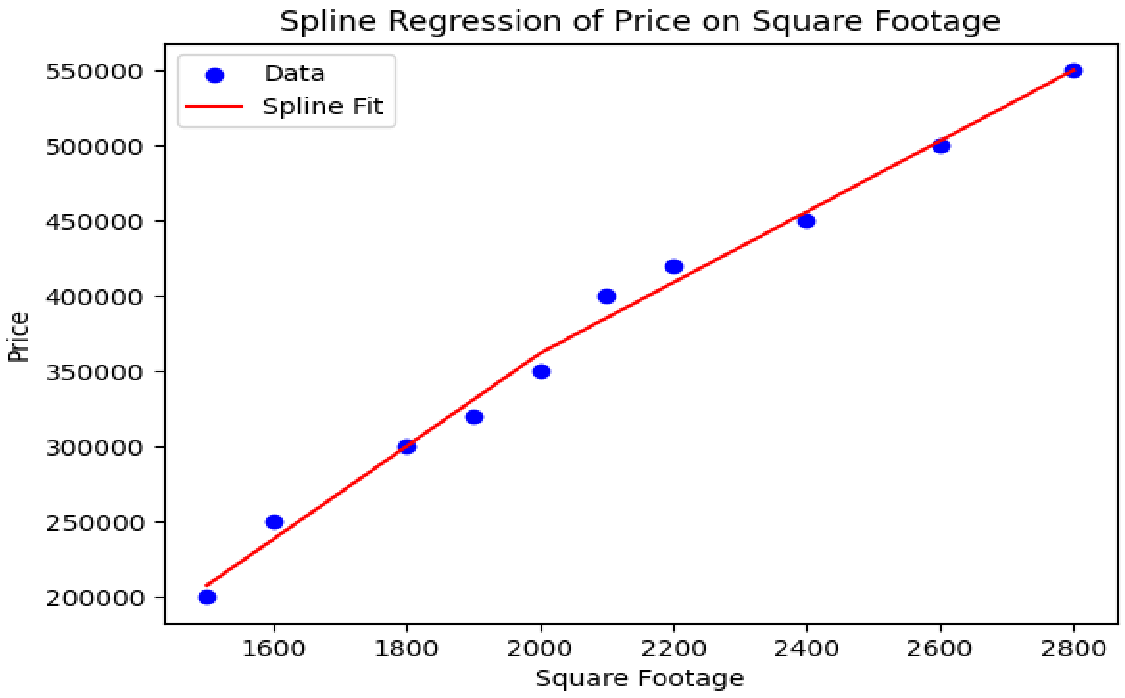
You have data on housing prices and square footage and notice that the relationship between square footage and price is nonlinear. Fit a spline regression model to allow the relationship between square footage and price to change at 2,000 square feet. Explain how spline regression can capture different behaviours of the relationship before and after 2,000 square feet.

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt

# dataset
data = {
    'Price': [200000, 250000, 300000, 320000, 350000, 400000, 420000, 450000, 500000, 550000],
    'SqFt': [1500, 1600, 1800, 1900, 2000, 2100, 2200, 2400, 2600, 2800]
}
df = pd.DataFrame(data)

# Define the spline term for SqFt with knot at 2000
df['sqft_knot'] = np.maximum(0, df['SqFt'] - 2000)

# Define independent variables (including intercept)
X = sm.add_constant(df[['SqFt', 'sqft_knot']])
y = df['Price']

# Fit linear regression with spline
model = sm.OLS(y, X).fit()

# Print summary
print(model.summary())

# Plot the fitted spline regression
plt.scatter(df['SqFt'], df['Price'], color='blue', label='Data')
plt.plot(df['SqFt'], model.predict(X), color='red', label='Spline Fit')
plt.xlabel('Square Footage')
plt.ylabel('Price')
plt.title('Spline Regression of Price on Square Footage')
plt.legend()
plt.show()
```

Output:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.992
Model:                            OLS   Adj. R-squared:                  0.990
Method:                 Least Squares   F-statistic:                     458.5
Date:                Fri, 19 Sep 2025   Prob (F-statistic):           3.78e-08
Time:                        09:41:24   Log-Likelihood:                -105.38
No. Observations:                  10   AIC:                             216.8
Df Residuals:                       7   BIC:                             217.7
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -2.557e+05   4.11e+04     -6.217      0.000   -3.53e+05   -1.58e+05
SqFt         308.7019     22.587     13.667      0.000     255.293     362.111
sqft_knot    -73.5822     32.468     -2.266      0.058    -150.356       3.191
==============================================================================
Omnibus:                        1.654   Durbin-Watson:                   1.972
Prob(Omnibus):                  0.437   Jarque-Bera (JB):                0.915
Skew:                           0.381   Prob(JB):                        0.633
Kurtosis:                       1.729   Cond. No.                     2.55e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.55e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```



Program 10

A hospital is using a Poisson regression model (a type of GLM) to predict the number of emergency room visits per week based on patient age and medical history. The model is given by:

$$\text{Log}(\lambda) =2.5-0.03*\text{Age}+0.5*\text{condition}$$

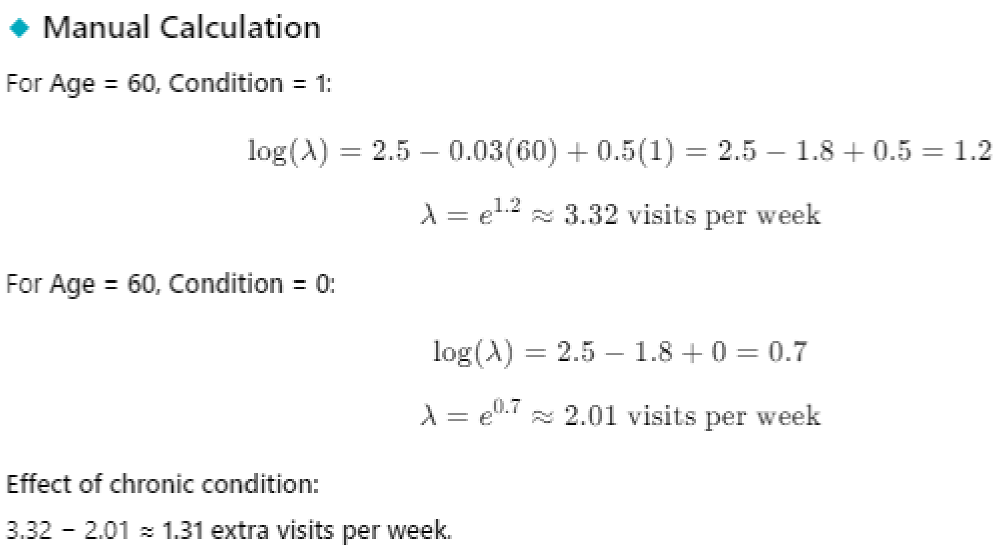
where λ is the expected number of visits per week, Age is the patient's age, and condition is a binary variable (1 if the patient has a chronic condition, 0 otherwise).

Interpret the coefficients of Age and condition. What is the expected number of visits per week for a 60-year-old patient with a chronic condition? How would the expected number of visits change if the patient did not have a chronic condition?

```python
import numpy as np

# Model coefficients
intercept = 2.5
coef_age = -0.03
coef_condition = 0.5

# Patient details
age = 60
condition_yes = 1
condition_no = 0

# Expected visits with chronic condition
log_lambda_yes = intercept + coef_age * age + coef_condition * condition_yes
lambda_yes = np.exp(log_lambda_yes)

# Expected visits without chronic condition
log_lambda_no = intercept + coef_age * age + coef_condition * condition_no
lambda_no = np.exp(log_lambda_no)

print(f"Expected visits (age 60, chronic condition): {lambda_yes:.2f} per week")
print(f"Expected visits (age 60, no chronic condition): {lambda_no:.2f} per week")
print(f"Change due to chronic condition: {lambda_yes - lambda_no:.2f} visits per week")
```

Output:

```
Expected visits (age 60, chronic condition): 3.32 per week
Expected visits (age 60, no chronic condition): 2.01 per week
Change due to chronic condition: 1.31 visits per week
```



◆ Manual Calculation

For Age = 60, Condition = 1:

$$\log(\lambda) = 2.5 - 0.03(60) + 0.5(1) = 2.5 - 1.8 + 0.5 = 1.2$$

$$\lambda = e^{1.2} \approx 3.32 \text{ visits per week}$$

For Age = 60, Condition = 0:

$$\log(\lambda) = 2.5 - 1.8 + 0 = 0.7$$

$$\lambda = e^{0.7} \approx 2.01 \text{ visits per week}$$

Effect of chronic condition:

3.32 − 2.01 ≈ **1.31 extra visits per week.**

Program 11

A bakery claims that its new cookie recipe is lower in calories compared to the old recipe, which had a mean calorie count of 200. You sample 40 new cookies and find a mean of 190 calories with a standard deviation of 15 calories. Perform a one-tailed t-test to determine if the new recipe has significantly fewer calories at a 5% significance level.

```
from scipy import stats
import math

# Given data
old_mean = 200        # mean calories of old recipe
sample_mean = 190     # mean calories of new recipe
sample_std = 15       # standard deviation of new recipe
n = 40                # sample size
alpha = 0.05          # significance level

# Calculate t-statistic
t_stat = (sample_mean - old_mean) / (sample_std / math.sqrt(n))

# Degrees of freedom
df = n - 1

# One-tailed p-value (testing if new mean < old mean)
p_value = stats.t.cdf(t_stat, df=df)

# Print results
print(f"T-statistic: {t_stat:.3f}")
print(f"P-value: {p_value:.5f}")

# Conclusion
if p_value < alpha:
    print("Reject the null hypothesis: The new recipe has significantly fewer calories.")
else:
    print("Fail to reject the null hypothesis: No significant difference in calories.")
```

Output:

```
T-statistic: -4.216
P-value: 0.00007
Reject the null hypothesis: The new recipe has significantly fewer calories.
```