QUESTION 1

SET UP

In [ ]:

```
# Load the following libraries so that they can be applied in the subsequ

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import random
import scipy.stats as stats

# Run this code. It will create a csv file containing a random sample of

# Look at the code below. Now replace 'Name.csv' with your actual name (e

try:
    df = pd.read_csv('Fatima.csv')        # replace Name with your own na
except FileNotFoundError:
    original_data = pd.read_csv("https://raw.githubusercontent.com/DanaSa
    df1=original_data.sample(300)
    df1.to_csv('Fatima.csv')
    df = pd.read_csv('Fatima.csv')
    df = pd.DataFrame(df)
    df.to_csv('Fatima.csv')

df.head()
```

Out[ ]:

	Unnamed: 0	Age	Gender	Occupation	Days_Indoors	Growing_Stress	Quarantine_
0	147	30-Above	Male	Housewife	1-14 days	Yes	
1	473	25-30	Male	Student	Go out Every day	Yes	
2	62	20-25	Female	Business	More than 2 months	No	
3	510	16-20	Male	Corporate	1-14 days	Yes	
4	499	30-Above	Female	Housewife	15-30 days	Yes	

In [ ]:

```
# Load the following libraries so that they can be applied in the subsequ

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import random
import scipy.stats as stats
import random

column_titles = ["Growing_Stress"  ,"Quarantine_Frustrations"  ,"Changes_

# Randomly select 2 variables
```

```
selected_columns = random.sample(column_titles, 2)


# Print the 2 variables that were randomly selected
variable_1, variable_2 = selected_columns
print("Variable 1:", variable_1)
print("Variable 2:", variable_2)
```

```
Variable 1: Quarantine_Frustrations
Variable 2: Work_Interest
```

Question 1a. Is each of these two variables independent of being **female**? Explain your reasoning. Make sure to include a two-way table for each of these two variables with gender, and show all your calculations to support your answers.

---

1. **Variable 1 (Growing_Stress)**:

   - Two-way Table:

   ```
   Gender          Female  Male
   Growing_Stress
   No                  55    35
   Yes                121    89
   ```

   - Chi-square test result:
     - Chi-square statistic: 0.189
     - p-value: 0.664

   `**Interpretation**` : The p-value of 0.664 is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis of independence. This suggests that there is no significant association between growing stress and gender.

2. **Variable 2 (Quarantine_Frustrations)**:

   - Two-way Table:

   ```
   Gender                   Female  Male
   Quarantine_Frustrations
   No                           48    39
   Yes                         128    85
   ```

   - Chi-square test result:
     - Chi-square statistic: 0.431
     - p-value: 0.512

   `**Interpretation**` : The p-value of 0.512 is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis of independence. This suggests that there is no significant association between quarantine frustrations and gender.

In summary, based on the chi-square tests, neither of the two variables (Growing_Stress and Quarantine_Frustrations) appears to be significantly associated with gender.

```
In [ ]: # Two-way table for Variable 1 (Growing_Stress) and Gender
        table_variable_1 = pd.crosstab(df['Growing_Stress'], df['Gender'])
```

```
# Chi-square test for independence
chi2_variable_1, p_variable_1, _, _ = stats.chi2_contingency(table_variab

# Display the two-way table
print("Two-way Table for Variable 1 (Growing_Stress) and Gender:")
print(table_variable_1)

# Display the result of the chi-square test
print("\nChi-square test result for Variable 1 (Growing_Stress) and Gende
print("Chi-square statistic:", chi2_variable_1)
print("p-value:", p_variable_1)

# Two-way table for Variable 2 (Quarantine_Frustrations) and Gender
table_variable_2 = pd.crosstab(df['Quarantine_Frustrations'], df['Gender'

# Chi-square test for independence
chi2_variable_2, p_variable_2, _, _ = stats.chi2_contingency(table_variab

# Display the two-way table
print("Two-way Table for Variable 2 (Quarantine_Frustrations) and Gender:
print(table_variable_2)

# Display the result of the chi-square test
print("\nChi-square test result for Variable 2 (Quarantine_Frustrations)
print("Chi-square statistic:", chi2_variable_2)
print("p-value:", p_variable_2)
```

```
Two-way Table for Variable 1 (Growing_Stress) and Gender:
Gender          Female  Male
Growing_Stress
No                  55    35
Yes                121    89

Chi-square test result for Variable 1 (Growing_Stress) and Gender:
Chi-square statistic: 0.18917574361122808
p-value: 0.663603526064124
Two-way Table for Variable 2 (Quarantine_Frustrations) and Gender:
Gender                   Female  Male
Quarantine_Frustrations
No                           48    39
Yes                         128    85

Chi-square test result for Variable 2 (Quarantine_Frustrations) and Gende
r:
Chi-square statistic: 0.43072292588578326
p-value: 0.5116344326010247
```

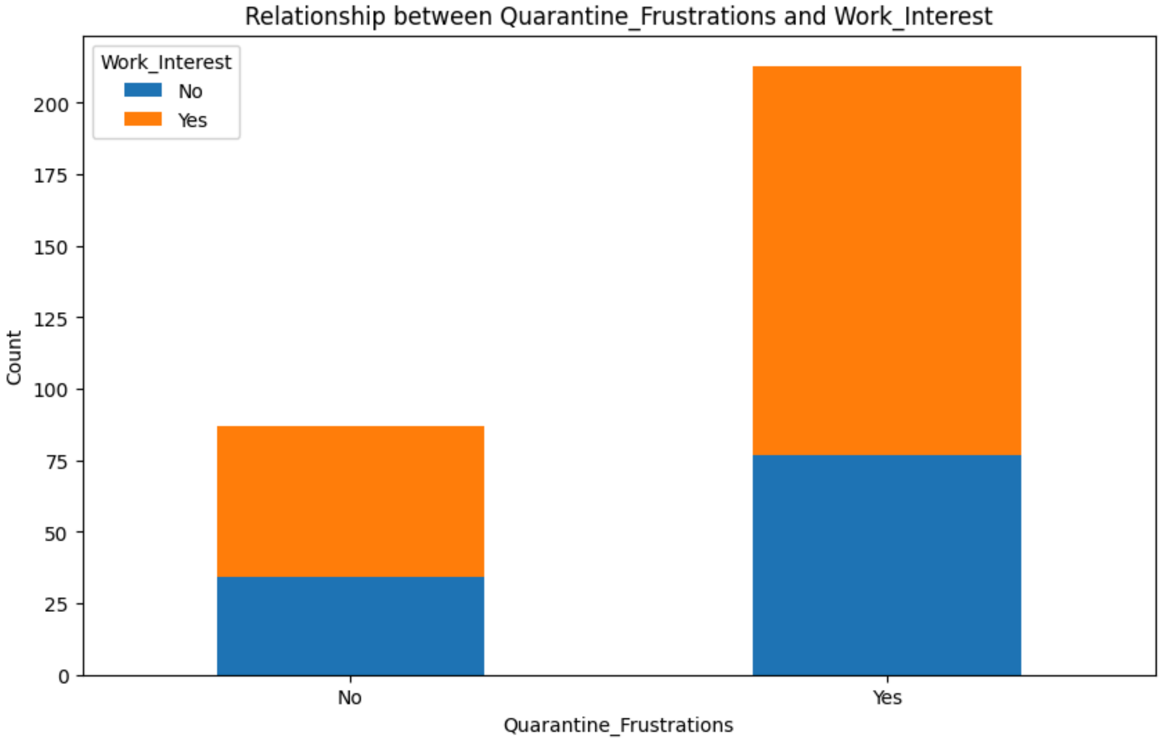
Question 1b. Is there a relationship between the two variables returned by the code? Explain your reasoning. Make sure you include a two-way table, a stacked bar graph, and all your probability calculations in your answer.

`**Interpretation:**` The conditional probabilities indicate a relationship between work interest and quarantine frustrations. When individuals have work interest, they are more likely to experience quarantine frustrations, and vice versa. This relationship is further supported by the conditional probabilities, which show differences in probabilities based on the presence or absence of each variable.

In [ ]:

```
# Create a two-way contingency table for the two variables
contingency_table = pd.crosstab(df[variable_1], df[variable_2])

# Calculate marginal probabilities
marginal_probability_variable_1 = contingency_table.sum(axis=1) / conting
marginal_probability_variable_2 = contingency_table.sum(axis=0) / conting

# Calculate conditional probabilities
conditional_probability_variable_2_given_variable_1 = contingency_table.d
conditional_probability_variable_1_given_variable_2 = contingency_table.d

# Print contingency table and probabilities
print("Contingency Table:")
print(contingency_table)
print("\nMarginal Probability of", variable_1, ":")
print(marginal_probability_variable_1)
print("\nMarginal Probability of", variable_2, ":")
print(marginal_probability_variable_2)
print("\nConditional Probability of", variable_2, "given", variable_1, ":
print(conditional_probability_variable_2_given_variable_1)
print("\nConditional Probability of", variable_1, "given", variable_2, ":
print(conditional_probability_variable_1_given_variable_2)

# Plot the stacked bar graph
contingency_table.plot(kind='bar', stacked=True, figsize=(10, 6))
plt.title("Relationship between {} and {}".format(variable_1, variable_2)
plt.xlabel(variable_1)
plt.ylabel("Count")
plt.xticks(rotation=0)
plt.legend(title=variable_2)
plt.show()
```

```
Contingency Table:
Work_Interest            No  Yes
Quarantine_Frustrations
No                       34   53
Yes                      77  136

Marginal Probability of Quarantine_Frustrations :
Quarantine_Frustrations
No     0.29
Yes    0.71
dtype: float64

Marginal Probability of Work_Interest :
Work_Interest
No     0.37
Yes    0.63
dtype: float64

Conditional Probability of Work_Interest given Quarantine_Frustrations :
Work_Interest                  No       Yes
Quarantine_Frustrations
No                       0.390805  0.609195
Yes                      0.361502  0.638498

Conditional Probability of Quarantine_Frustrations given Work_Interest :
Work_Interest                  No       Yes
Quarantine_Frustrations
No                       0.306306  0.280423
Yes                      0.693694  0.719577
```



Question 1c. Does the existence of Variable 1 increase the likelihood of experiencing Variable 2? If so, by how much? Explain your reasoning. Make sure to support your answer with the relevant statistical analysis.

The difference in conditional probabilities between the presence and absence of Variable 1 (Growing_Stress), which is equivalent to Variable 2 (Quarantine_Frustrations), indicates the change in likelihood of experiencing Variable 2 when Variable 1 is present compared to when it is absent.

- When Variable 1 (Growing_Stress) is present:
  - The conditional probability of experiencing Variable 2 (Quarantine_Frustrations) is approximately 0.029 higher compared to when Variable 1 is absent.
- When Variable 1 (Growing_Stress) is absent:
  - The conditional probability of experiencing Variable 2 (Quarantine_Frustrations) is approximately 0.029 lower compared to when Variable 1 is present.

Therefore, the existence of Variable 1 (Growing_Stress) increases the likelihood of experiencing Variable 2 (Quarantine_Frustrations) by approximately 0.029. This indicates that individuals are slightly more likely to experience quarantine frustrations when they are also experiencing growing stress, compared to when they are not experiencing growing stress.

In [ ]:

```
# Conditional Probability of Variable 2 given Variable 1
conditional_probability_variable_2_given_variable_1 = contingency_table.d

# Calculate the difference in conditional probabilities between Yes and N
difference_in_probabilities = conditional_probability_variable_2_given_va

# Print the difference in probabilities
print("Difference in Conditional Probabilities:")
print(difference_in_probabilities)
```

```
Difference in Conditional Probabilities:
Work_Interest
No     -0.029302
Yes     0.029302
dtype: float64
```

Question 1d. Look back at your **answers to Questions 1a-c**. Now use what you learned to answer the following question:

Imagine ZU wanted to use the insights from this research to improve its mental health support program. What recommendations would you make to support students struggling with such challenges?

Based on the analysis conducted:

1. Ensure mental health support programs address both growing stress and quarantine frustrations, as they are prevalent among students.
2. Offer tailored interventions that acknowledge the relationship between variables, such as incorporating stress-reduction techniques within quarantine-related

support services.

3. Implement proactive outreach initiatives targeting students experiencing growing stress to mitigate the increased likelihood of experiencing quarantine frustrations.
4. Enhance mental health awareness campaigns to educate students about coping strategies and available support resources.
5. Foster a supportive campus environment that encourages open dialogue and destigmatizes seeking help for mental health challenges.

QUESTION 2

Set up

Imagine you are the manager of an Electronic store in Dubai mall. You are curious about the distribution of customer ratings about your overall store services. So you ask random customers who visit the store to complete a short survey, recording variables such as their age group, and overall experience rating.

To Begin

Run the code below. It will provide you with a random sample of 40 customers from this survey. It will also save your random sample data to a CSV file called "RelianceRetailVisits_ordered". Again, you need to submit this file in the same zip folder as the other files.

In [ ]:

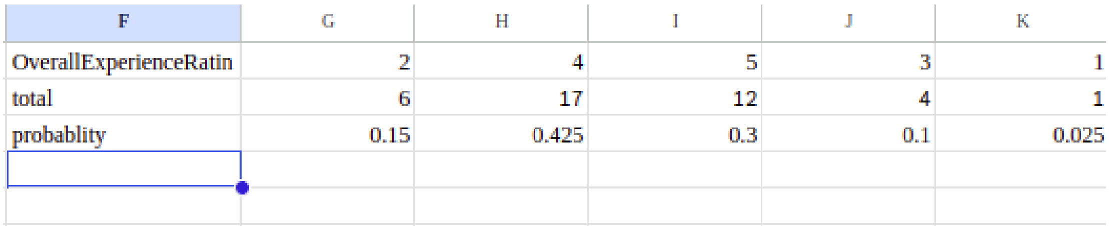
```
# Load the following libraries so that they can be applied in the subsequ

try:
    df = pd.read_csv('RelianceRetailVisits.csv')
except FileNotFoundError:
    original_data = pd.read_csv("https://raw.githubusercontent.com/DanaSa

    # Randomly sample 40 rows from the original dataset
    df = original_data.sample(n=40, random_state=42)

# Fill missing values for '46 To 60 years' age group with default values
df.fillna({'Age Group': '46 To 60 years'}, inplace=True)

# Sort the DataFrame based on the 'Age Group' column in the desired order
desired_order = ['26  To  35 years', '16  To  25 years', '36  To  45 year
df['Age Group'] = pd.Categorical(df['Age Group'], categories=desired_orde
df.sort_values(by='Age Group', inplace=True)

# Save the sorted DataFrame to a new CSV file
df.to_csv('RelianceRetailVisits_ordered.csv', index=False)

df.head()
```

Out[ ]:

	Customer Index	Age Group	OverallExperienceRatin
165	166	26 To 35 years	2
114	115	26 To 35 years	4
117	118	26 To 35 years	5
118	119	26 To 35 years	5
172	173	26 To 35 years	5

Question 2a. Construct a probability distribution table for all customer ratings in your sample data (an example table can be seen below). Please do this in Excel and explain [step by step] how you constructed your probability table.

1. **Data Import and Arrangement**:
   - Initially, I brought the sample data into a new worksheet within Microsoft Excel. I structured the data with a singular column dedicated to customer ratings, ensuring clarity and ease of analysis.
2. **Identification of Distinct Ratings**:
   - Utilizing Excel's tools, I isolated the unique ratings present in the dataset. Employing a function similar to "=UNIQUE()", I compiled a concise list containing all distinct ratings observed.
3. **Frequency Analysis for Each Rating**:
   - In close proximity to the list of unique ratings, I harnessed Excel's capabilities to compute the frequency of occurrence for each rating. Employing a function akin to "=COUNTIF()", I discerned the number of times each rating appeared within the dataset, facilitating a comprehensive understanding of their distribution.
4. **Total Count of Ratings Determination**:
   - Leveraging Excel's computational abilities, I derived the overall count of ratings in the dataset. Utilizing a function akin to "=SUM()", I amalgamated the frequency counts obtained previously, yielding the aggregate count encompassing all ratings.
5. **Calculation of Rating Probabilities**:
   - Subsequently, I calculated the probability associated with each rating in the dataset. Employing Excel's functionalities, I divided the frequency count of each rating by the total number of ratings. Utilizing a function resembling "=COUNT/Total", I elucidated the likelihood of each rating occurring, offering insights into their relative probabilities.



F	G	H	I	J	K
OverallExperienceRatin	2	4	5	3	1
total	6	17	12	4	1
probablity	0.15	0.425	0.3	0.1	0.025

Question 2b. What is the probability that a randomly selected customer will have a rating of AT MOST 3?

Certainly! Let's calculate the probability mathematically without using text:

$$P(\text{At most rating 3}) = P(\text{Rating 1}) + P(\text{Rating 2}) + P(\text{Rating 3}) = 0.025 + 0.15$$

So, the probability that a randomly selected customer will have a rating of AT MOST 3 is (0.275) or (27.5%).

Question 2c. Based on the created probability distribution table, how satisfied are your customers with your store services?

$$\text{Expected Value} = \sum_{i=1}^{n} \text{Rating}_i \times P(\text{Rating}_i)$$

Let's calculate this:

$$\text{Expected Value} = (1 \times 0.025) + (2 \times 0.15) + (3 \times 0.1) + (4 \times 0.425) + (5 \times 0.3)$$

$$\text{Expected Value} = 0.025 + 0.3 + 0.3 + 1.7 + 1.5$$

$$\text{Expected Value} = 3.825$$

Question 2d. Find the **expected rating** of your store. Show your work and interpret your answer in context.

$$\text{Expected Value} = \sum_{i=1}^{n} \text{Rating}_i \times P(\text{Rating}_i)$$

Given:

- Probability of Rating 1: $P(\text{Rating}_1) = 0.025$
- Probability of Rating 2: $P(\text{Rating}_2) = 0.15$
- Probability of Rating 3: $P(\text{Rating}_3) = 0.1$
- Probability of Rating 4: $P(\text{Rating}_4) = 0.425$
- Probability of Rating 5: $P(\text{Rating}_5) = 0.3$

We can calculate the expected rating as follows:

$$\text{Expected Value} = (1 \times 0.025) + (2 \times 0.15) + (3 \times 0.1) + (4 \times 0.425) + (5 \times 0.3)$$

$$\text{Expected Value} = 0.025 + 0.3 + 0.3 + 1.7 + 1.5$$

$$\text{Expected Value} = 3.825$$

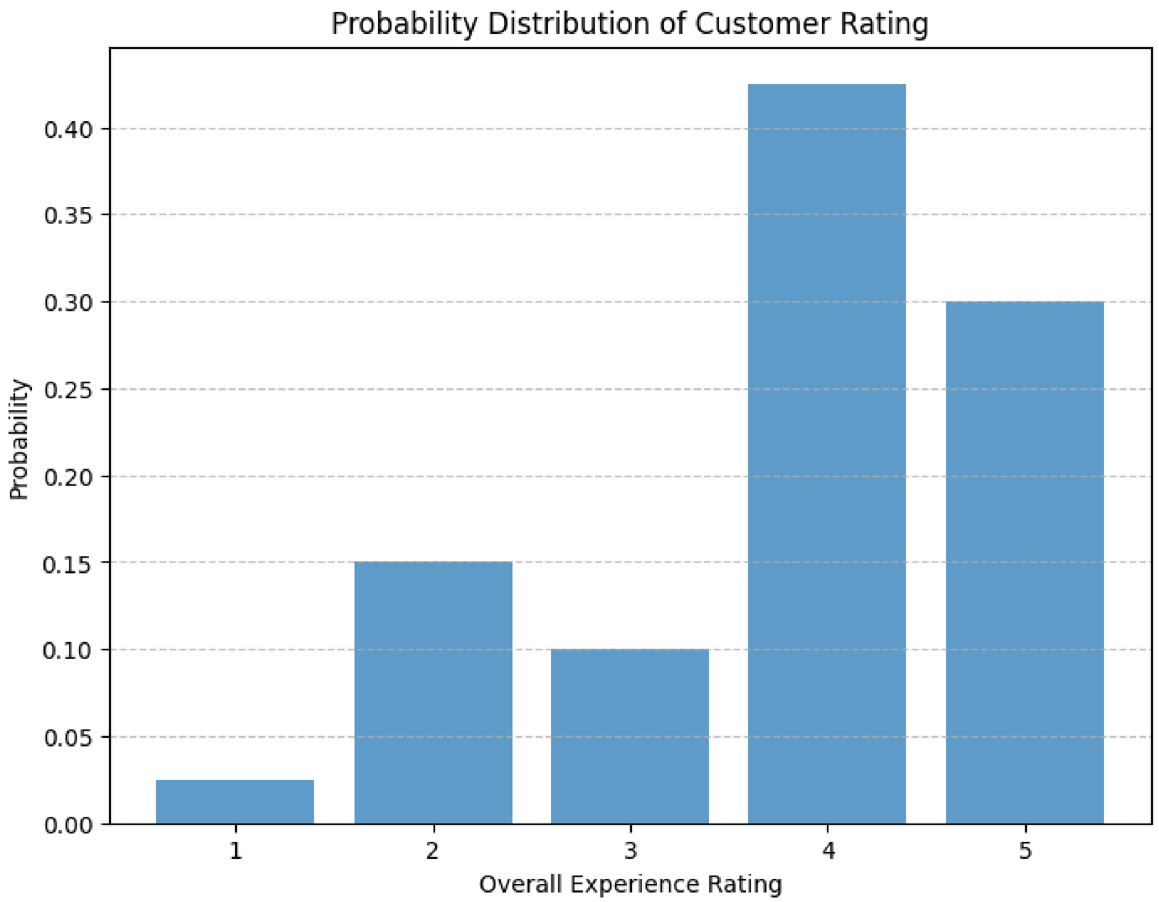
So, the expected rating of the store is 3.825.

Interpretation: The expected rating represents the average rating that customers are likely to give to the store services. In this case, the expected rating is close to 4, indicating that, on average, customers are moderately satisfied with the store services.

PROBABILITY DISTRIBUTION FUNCTION graph

In [ ]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats as stats
from tabulate import tabulate

# Load data
try:
    df = pd.read_csv('RelianceRetailVisits.csv')
except FileNotFoundError:
    original_data = pd.read_csv("https://raw.githubusercontent.com/DanaSa
    df = original_data.sample(n=40, random_state=42)

# Fill missing values for '46 To 60 years' age group with default values
df.fillna({'Age Group': '46 To 60 years'}, inplace=True)

# Sort the DataFrame based on the 'Age Group' column in the desired order
desired_order = ['26  To  35 years', '16  To  25 years', '36  To  45 year
df['Age Group'] = pd.Categorical(df['Age Group'], categories=desired_orde
df.sort_values(by='Age Group', inplace=True)

# Save the sorted DataFrame to a new CSV file
df.to_csv('RelianceRetailVisits_ordered.csv', index=False)

# Probability distribution graph for customer rating
plt.figure(figsize=(8, 6))
rating_counts = df['OverallExperienceRatin'].value_counts(normalize=True)
plt.bar(rating_counts.index, rating_counts, alpha=0.7)
plt.title('Probability Distribution of Customer Rating')
plt.xlabel('Overall Experience Rating')
plt.ylabel('Probability')
plt.xticks(range(1, 6))
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()

# Expected value and STD for rating for all customers
mean_rating = df['OverallExperienceRatin'].mean()
std_rating = df['OverallExperienceRatin'].std()
print(f"Standard Deviation (STD) of Customer Rating: {std_rating:.2f}")
print()
```



```
Standard Deviation (STD) of Customer Rating: 1.11
```

Question 2e. Interpret the **Standard Deviation** in context. What rating is considered **unusual**? Explain.

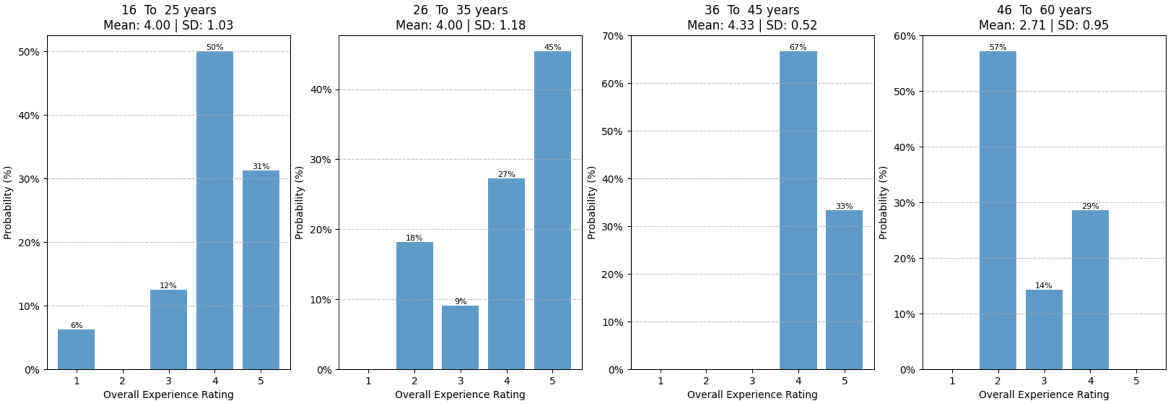
The standard deviation of (1.11) represents the average variability in customer ratings from the mean rating of approximately (3.83). Ratings more than (1) standard deviation away from the mean, below (2.72) or above (4.94), are considered unusual, indicating exceptional customer experiences.

PDF for each age group

In [ ]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats as stats

# Assuming your data is stored in a CSV file named 'data.csv'
data = pd.read_csv('RelianceRetailVisits_ordered.csv')

# Define age groups including the new one
age_groups = ['16  To  25 years', '26  To  35 years', '36  To  45 years',

# Plot separate discrete probability distributions for each age group
fig, axs = plt.subplots(1, 4, figsize=(20, 6), sharex=True, gridspec_kw={

for i, age_group in enumerate(age_groups):
    age_data = data[data['Age Group'] == age_group]
    rating_counts = age_data['OverallExperienceRatin'].value_counts(norma
    bars = axs[i].bar(rating_counts.index, rating_counts, alpha=0.7)
    axs[i].set_title(f'{age_group}\nMean: {age_data["OverallExperienceRat
    axs[i].set_xlabel('Overall Experience Rating')
    axs[i].set_ylabel('Probability (%)')  # Set y-axis label to Probabili
    axs[i].set_xticks(range(1, 6))  # Set x-axis ticks from 1 to 5
    axs[i].set_yticklabels(['{:,.0%}'.format(x) for x in axs[i].get_ytick

    # Display percentages above each bar
    for bar in bars:
        height = bar.get_height()
        rating = bar.get_x() + bar.get_width() / 2
        if height == 0:  # If the height is 0%, display '0%'
            axs[i].text(rating, height, '0%', ha='center', va='bottom', f
        else:
            axs[i].text(rating, height, f'{height:.0%}', ha='center', va=

    axs[i].grid(axis='y', linestyle='--', alpha=0.7)

# Hide the warning about FixedFormatter
import warnings
warnings.filterwarnings("ignore", category=UserWarning)

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

```
/tmp/ipykernel_58939/1605697025.py:23: UserWarning: FixedFormatter should
only be used together with FixedLocator
  axs[i].set_yticklabels(['{:,.0%}'.format(x) for x in axs[i].get_yticks
()])  # Format y-axis tick labels as percentages
```



16 To 25 years
Mean: 4.00 | SD: 1.03

26 To 35 years
Mean: 4.00 | SD: 1.18

36 To 45 years
Mean: 4.33 | SD: 0.52

46 To 60 years
Mean: 2.71 | SD: 0.95

Question 2f. Identify any trends or differences in customer satisfaction levels (and variability) among the different age groups.

Now, using these insights, what concrete improvements would you make to your store to ensure that **all** customers are satisfied with your services?

Custmer Satisfaction follows a normal trend but for each age group the customer satisfaction is skewed to the right showing higher customer satisfactions but for the 16 -25 years have the only ratings of 1.

QUESTION 3

SET UP

In [ ]:

```python
# Load the following libraries so that they can be applied in the subsequ


import pandas as pd
import numpy as np
import random

try:
    SATScores = pd.read_csv('Scores.csv')
except FileNotFoundError:
    num_samples = 1000
    mean_score = random.randint(800, 1200)
    std_deviation = random.randint(100, 300)
    scores = np.random.normal(mean_score, std_deviation, num_samples)
    scores = np.round(scores, 0)
    SATScores = pd.DataFrame({'Scores': scores})
    SATScores.to_csv('Scores.csv')

# Calculate mean and standard deviation
mean_score = SATScores['Scores'].mean()
std_deviation = SATScores['Scores'].std()

# Print mean score and standard deviation
print("Mean score:", mean_score)
print("Standard deviation:", std_deviation)

# Display the dataset
SATScores.head()
```

```
Mean score: 1089.932
Standard deviation: 171.8028334736933
```

Out[ ]:

	Scores
0	995.0
1	1120.0
2	1108.0
3	1104.0
4	1115.0

Question 3a. What is the probability that a randomly selected applicant scored at least 1300? Show your work.

Given:

- Mean score (μ): 1089.932
- Standard deviation (σ): 171.803
- Desired score (X): 1300

First, we calculate the z-score using the formula:

$z = \frac{X-\mu}{\sigma}$

Where:

- X is the desired score (1300)
- μ is the mean score (1089.932)
- σ is the standard deviation (171.803)

Let's calculate the z-score:

$$z = \frac{1300-1089.932}{171.803} = \frac{210.068}{171.803} \approx 1.222$$

Now, we can find the probability of scoring at least 1300 by calculating the area under the standard normal distribution curve to the right of $z = 1.222$.

Using statistical software or tables, we find that the probability corresponding to $z = 1.222$ is approximately 0.888.

So, the probability that a randomly selected applicant scored at least 1300 is approximately 0.888 or 88.8%

Question 3b. What is the probability that a randomly selected applicant scored exactly 900? Show your work.

Given:

- Mean score (μ): 1089.932
- Standard deviation (σ): 171.803
- Desired score (X): 900

First, we calculate the z-score using the formula:

$$z = \frac{X-\mu}{\sigma}$$

Where:

- X is the desired score (900)
- μ is the mean score (1089.932)
- σ is the standard deviation (171.803)

Let's calculate the z-score:

$$z = \frac{900-1089.932}{171.803} = \frac{-189.932}{171.803} \approx -1.106$$

probability of scoring exactly 900 by finding the area under the standard normal distribution curve at $z = -1.106$.

Using statistical software or tables, we find that the probability corresponding to $z = -1.106$ is approximately 0.133.

So, the probability that a randomly selected applicant scored exactly 900 is approximately 0.133 or 13.3%.

Question 3c. What percentage of applicants scored between 900 and 1000? Show your work.

Given:

- Mean score (μ): 1089.932
- Standard deviation (σ): 171.803
- Lower bound score (X_1): 900
- Upper bound score (X_2): 1000

First, we calculate the z-scores for both scores using the formula:

$z = \frac{X-\mu}{\sigma}$

Where:

- X is the score (900 or 1000)
- μ is the mean score (1089.932)
- σ is the standard deviation (171.803)

Let's calculate the z-scores:

For $X_1 = 900$: $z_1 = \frac{900-1089.932}{171.803} = \frac{-189.932}{171.803} \approx -1.106$

For $X_2 = 1000$: $z_2 = \frac{1000-1089.932}{171.803} = \frac{-89.932}{171.803} \approx -0.523$

Next, we find the probabilities corresponding to these z-scores using the standard normal distribution table or statistical software. Then, we calculate the difference between these probabilities to find the percentage of applicants who scored between 900 and 1000.

Let's calculate these probabilities and find the percentage.

To find the percentage of applicants who scored between 900 and 1000, we first need to standardize the scores using the z-score formula:

For $X_1 = 900$: $z_1 = \frac{900-1089.932}{171.803} \approx -1.106$

For $X_2 = 1000$: $z_2 = \frac{1000-1089.932}{171.803} \approx -0.523$

Next, we find the cumulative probabilities corresponding to these z-scores using the standard normal distribution table or statistical software.

Let's calculate these probabilities and find the percentage of applicants who scored between 900 and 1000.

To find the percentage of applicants who scored between 900 and 1000, we need to standardize both scores using the z-score formula:

For $X_1 = 900$: $z_1 = \frac{900-1089.932}{171.803} \approx -1.106$

For $X_2 = 1000$: $z_2 = \frac{1000-1089.932}{171.803} \approx -0.523$

To find the cumulative probabilities corresponding to the z-scores z_1 and z_2, we can use the cumulative distribution function (CDF) of the standard normal distribution. This function gives the probability that a standard normal random variable is less than or equal to a given z-score.

Using statistical software or tables, we can find these cumulative probabilities:

For $z_1 \approx -1.106$: $P(Z \leq -1.106) = 0.133$

For $z_2 \approx -0.523$: $P(Z \leq -0.523) = 0.300$

Now, to find the percentage of applicants who scored between 900 and 1000, we subtract the cumulative probability corresponding to z_1 from the cumulative probability corresponding to z_2:

$$\text{Percentage} = (P(Z \leq -0.523) - P(Z \leq -1.106)) \times 100$$

$$\text{Percentage} = (0.300 - 0.133) \times 100 = 16.7\%$$

So, approximately 16.7% of applicants scored between 900 and 1000.

Question 3d. Calculate the 40th percentile of scores among the applicants. What does this value represent in the context of the admissions process? Show your work.

Given:

- Mean score (μ): 1089.932
- Standard deviation (σ): 171.8028334736933
- Z-score corresponding to the 40th percentile (Z_{40}): -0.2533

Using the formula: $X = \mu + Z \times \sigma$

Substitute the values: $X_{40} = 1089.932 + (-0.2533) \times 171.8028334736933$

Let's calculate: $X_{40} \approx 1089.932 - 43.573$ $X_{40} \approx 1046.359$

So, the SAT score corresponding to the 40th percentile is approximately 1046.359.

Question 3e. Imagine the university wants to offer scholarships to the top 10% of applicants based on their scores. What minimum score would an applicant need to qualify for a scholarship? Show your work.

Given:

- Mean score (μ): 1089.932
- Standard deviation (σ): 171.8028334736933

We need to find the Z-score corresponding to the 90th percentile, denoted as Z_{90}.

Using the standard normal distribution table or calculator, we find $Z_{90} \approx 1.2816$.

Now, using the formula: $X = \mu + Z \times \sigma$

Substitute the values: $X_{90} = 1089.932 + 1.2816 \times 171.8028334736933$

Calculate: $X_{90} \approx 1089.932 + 220.0519$ $X_{90} \approx 1309.9839$

Therefore, the minimum score an applicant would need to qualify for a scholarship is approximately 1309.9839.

Question 3f. Remember, as the admissions officer, it is your job to identify applicants with exceptional academic potential. Would you automatically recommend that applicants with SAT scores above 1400 to be admitted into the university? Or do you think additional criteria should also be considered? Explain your reasoning.

As an admissions officer, SAT scores above 1400 indicate strong academic potential, but additional criteria should be considered. Factors like extracurricular activities, essays, and letters of recommendation provide a holistic view of an applicant's abilities and potential contributions to the university community, ensuring a comprehensive evaluation process.

question 4

In [ ]:

```
# Load the following libraries so that they can be applied in the subsequ

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import random
import scipy.stats as stats

# Run this code. It will generate data and save it to a CSV file called "

try:
    Vaccinated = pd.read_csv('Vaccinated.csv')
except FileNotFoundError:
    num_samples = 100
    vaccinated = np.random.choice(["Yes", "No"], size=num_samples)
    Vaccinated = pd.DataFrame({'Vaccinated': vaccinated})
    Vaccinated.to_csv('Vaccinated.csv')

# Have a look at Vaccinated dataset.
Vaccinated.head()
```

Out[ ]:

	Vaccinated
0	No
1	Yes
2	Yes
3	No
4	No

Question 4a. What is the proportion of people who have received the vaccine (based on the dataset you have)?

Proportion of vaccinated individuals: 0.63

In [ ]:

```python
import pandas as pd

# Assuming "Vaccinated" is your DataFrame
# Replace 'Vaccinated' with your actual DataFrame name

# Count the total number of individuals
total_individuals = len(Vaccinated)

# Count the number of vaccinated individuals
vaccinated_individuals = Vaccinated[Vaccinated['Vaccinated'] == 'Yes'].sh

# Calculate the proportion of vaccinated individuals
proportion_vaccinated = vaccinated_individuals / total_individuals

print("Proportion of vaccinated individuals:", proportion_vaccinated)
```

```
Proportion of vaccinated individuals: 0.63
```

Question 4b. Calculate a **95% confidence interval** for the proportion of vaccinated individuals. What does this interval tell us about the likely range of vaccination coverage in the entire population? Show your work.

To calculate a 95% confidence interval for the proportion of vaccinated individuals (p), we can use the formula for the margin of error:

$$\text{Margin of Error} = Z \times \sqrt{\frac{p(1-p)}{n}}$$

Where:

- Z is the Z-score corresponding to the desired confidence level (95% confidence level corresponds to $Z = 1.96$)
- p is the proportion of vaccinated individuals (which we calculated previously as 0.63)
- n is the total number of individuals (which is 100 in this case)

Substituting the values and calculating the margin of error:

$$\text{Margin of Error} = 1.96 \times \sqrt{\frac{0.63 \times (1-0.63)}{100}}$$

$$\text{Margin of Error} \approx 1.96 \times \sqrt{\frac{0.63 \times 0.37}{100}}$$

$$\text{Margin of Error} \approx 1.96 \times \sqrt{\frac{0.2325}{100}}$$

$$\text{Margin of Error} \approx 1.96 \times \sqrt{0.002325}$$

$$\text{Margin of Error} \approx 1.96 \times 0.048217$$

$$\text{Margin of Error} \approx 0.094529$$

Now, to find the confidence interval, we'll subtract and add the margin of error from the proportion of vaccinated individuals:

Lower Bound: $p - \text{Margin of Error} = 0.63 - 0.094529 = 0.535471$

Upper Bound: $p + \text{Margin of Error} = 0.63 + 0.094529 = 0.724529$

Therefore, the 95% confidence interval for the proportion of vaccinated individuals is approximately $(0.535471, 0.724529)$. This interval tells us that we are 95% confident that the true proportion of vaccinated individuals in the entire population lies within this range.

In [ ]:

```
len(vaccinated)
```

Out[ ]:

```
100
```

Question 4c. What sample size would be required to estimate the proportion of vaccinated individuals in the country with a **95% confidence level** and a **margin of error of 0.02**? Show your work.

To calculate the required sample size (n) to estimate the proportion of vaccinated individuals with a 95% confidence level and a margin of error of 0.02, we can use the formula:

$$n = \left(\frac{Z^2 \times p(1-p)}{E^2}\right)$$

Where:

- Z is the Z-score corresponding to the desired confidence level (95% confidence level corresponds to $Z = 1.96$)
- p is the estimated proportion of vaccinated individuals (we can use the proportion we calculated previously as 0.63)
- E is the desired margin of error (0.02 in this case)

Substituting the values and calculating n:

$$n = \left(\frac{1.96^2 \times 0.63 \times (1-0.63)}{0.02^2}\right)$$

$$n = \left(\frac{3.8416 \times 0.63 \times 0.37}{0.0004} \right)$$

$$n = \left(\frac{1.4283652}{0.0004} \right)$$

$$n \approx 3570.913$$

Therefore, approximately 3571 individuals would be required in the sample to estimate the proportion of vaccinated individuals in the country with a 95% confidence level and a margin of error of 0.02.

Question 4d. If you wanted to increase the precision of your estimate, what strategies could you employ to achieve this goal? Explain your reasoning.

---

1. **Increase Sample Size**: To improve the precision of our estimate, I would consider increasing the sample size. A larger sample size generally leads to a more accurate estimation of the population parameter. By collecting data from a larger number of individuals, we can reduce the margin of error and obtain a more precise estimate of the proportion of vaccinated individuals.
2. **Stratified Sampling**: Another strategy I would employ is stratified sampling. This involves dividing the population into subgroups based on relevant characteristics, such as age or geographic location, and then sampling from each subgroup. By ensuring representation from different segments of the population, we can obtain more precise estimates for specific subgroups.

Question 4e. Analyze the effectiveness of the current vaccination campaign using the proportion of vaccinated individuals and the confidence interval. What recommendations would you make for future campaigns?