# Statistics Basics Assignment

## Name – Vishal Jadhav

# Statistics Assignment Answers

## Question 1: What is the difference between descriptive statistics and inferential statistics? Explain with examples?

**Descriptive Statistics** involves methods of organizing, summarizing, and presenting data in an informative way. It aims to describe the main features of a collection of data quantitatively. This type of statistics is used to describe the characteristics of a sample or population.

- **Examples:**
  - Calculating the average height of students in a classroom.
  - Determining the most frequent score on a test.
  - Creating a bar chart to show the distribution of different car colors in a parking lot.
  - Reporting the range of ages of participants in a survey.

**Inferential Statistics** involves methods that use data from a sample to make predictions or inferences about a larger population. It aims to draw conclusions and make generalizations beyond the immediate data. This type of statistics is used to test hypotheses and make predictions.

- **Examples:**
  - Using a sample of voters to predict the outcome of an election.
  - Testing a new drug on a sample of patients to infer its effectiveness on the entire population with the disease.
  - Estimating the average income of all households in a city based on a survey of a few hundred households.
  - Determining if there is a significant difference in test scores between two groups of students who received different teaching methods.

**Key Differences:**

| Feature | Descriptive Statistics | Inferential Statistics |
|---|---|---|
| **Purpose** | Describe characteristics of a dataset | Make inferences and predictions about a population |
| **Data Source** | Entire population or a sample (described as is) | Sample data used to generalize to a population |
| **Methods** | Measures of central tendency (mean, median, mode), measures of dispersion (range, variance, standard deviation), frequency distributions, graphs | Hypothesis testing, confidence intervals, regression analysis, ANOVA |
| **Outcome** | Summaries, charts, tables | Probabilities, predictions, conclusions, generalizations |
| **Scope** | Limited to the observed data | Extends beyond the observed data |

## Question 2: What is sampling in statistics? Explain the differences between random and stratified sampling?

**Sampling in statistics** is a process of selecting a subset (a sample) of individuals or items from a larger group (a population) to gather information and make inferences about the entire population. It is often impractical or impossible to study an entire population, so sampling allows researchers to collect data from a smaller, manageable group while still aiming for results that are representative of the whole.

**Key reasons for sampling:** * **Cost-effectiveness:** Studying a sample is generally less expensive than studying an entire population. * **Time-efficiency:** Data collection from a sample is much faster. * **Feasibility:** In some cases, studying the entire population might be impossible (e.g., destructive testing). * **Accuracy:** A well-chosen sample can provide accurate insights into the population.

## Differences between Random and Stratified Sampling

### 1. Random Sampling (Simple Random Sampling)

**Definition:** Simple random sampling is a method where every individual or item in the population has an equal chance of being selected for the sample. Each selection is independent of the others.

**How it works:** * Assign a unique number to each member of the population. * Use a random number generator or a lottery method to select the required number of individuals.

**Characteristics:** * **Unbiased:** It minimizes bias and ensures that the sample is representative of the population on average. * **Simplicity:** It is conceptually simple and easy to implement for small populations. * **Generalizability:** Results can be generalized to the entire population with a known margin of error.

**Limitations:** * May not be truly representative if the population has distinct subgroups that are not proportionally represented by chance. * Can be impractical for very large populations as it requires a complete list of all members.

**Example:** If you want to survey 100 students from a school of 1000 students, you could put all 1000 names into a hat and draw out 100 names randomly.

### 2. Stratified Sampling

**Definition:** Stratified sampling is a method that involves dividing the population into homogeneous subgroups called \'strata\' based on shared characteristics (e.g., age, gender, income, education level). Then, a simple random sample is drawn from each stratum, proportional to the stratum\'s size in the population.

**How it works:** * Divide the population into mutually exclusive and collectively exhaustive strata. * Determine the proportion of each stratum in the population. * Select a random sample from each stratum, ensuring that the sample size from each stratum is proportional to its representation in the population.

**Characteristics:** * **Increased Representativeness:** Ensures that specific subgroups are adequately represented in the sample, which is crucial when these subgroups are important for the study. * **Reduced Sampling Error:** Can lead to more precise estimates for the population parameters, especially if the strata are truly homogeneous. * **Comparisons:** Allows for comparisons between different subgroups.

**Limitations:** * Requires prior knowledge of the population characteristics to create strata. * More complex to implement than simple random sampling. * If strata are not well-defined or overlap, it can introduce bias.

**Example:** If you want to survey 100 students from a school of 1000 students, and you know that a0% are female and 40% are male, you would select a0 female students and 40 male students randomly from their respective groups to ensure proportional representation.

**Key Differences Summarized:**

| Feature | Random Sampling | Stratified Sampling |
|---|---|---|
| **Population Division** | No division; treats population as a single unit | Divides population into homogeneous subgroups (strata) |
| **Selection** | Individuals selected randomly from the entire population | Individuals selected randomly from each stratum |
| **Representativeness** | Relies on chance for representativeness; may not guarantee representation of subgroups | Ensures representation of specific subgroups |
| **Complexity** | Simpler to implement | More complex; requires prior knowledge of population |
| **Precision** | Generally less precise for heterogeneous populations | More precise for heterogeneous populations |
| **Use Case** | Homogeneous populations or when subgroup analysis is not critical | Heterogeneous populations where subgroup analysis is important or specific subgroups need to be represented |

# Question 3: Define mean, median, and mode. Explain why these measures of central tendency are important?

**Measures of central tendency** are single values that attempt to describe a set of data by identifying the central position within that set of data. They are often called averages. The three main measures of central tendency are the mean, median, and mode.

## 1. Mean

**Definition:** The mean (or arithmetic mean) is the most commonly used measure of central tendency. It is calculated by summing all the values in a dataset and then dividing by the number of values in that dataset.

**Formula:** For a set of $n$ values $x_1, x_2, \ldots, x_n$, the mean (denoted as $\bar{x}$) is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

**Importance:** * **Balances all values:** The mean takes into account every value in the dataset, making it a good representation of the overall data. * **Mathematical properties:** It has desirable mathematical properties that make it suitable for further statistical analysis (e.g., in regression, ANOVA). * **Widely understood:** It is a familiar and easily understood concept.

**Limitations:** * **Sensitive to outliers:** The mean can be heavily influenced by extreme values (outliers), which can pull the average away from the typical value. * **Not suitable for skewed data:** For skewed distributions, the mean may not accurately represent the \'center\' of the data.

**Example:** For the dataset [1, 2, 3, 4, 5], the mean is (1+2+3+4+5)/5 = 15/5 = 3.

## 2. Median

**Definition:** The median is the middle value in a dataset when the values are arranged in ascending or descending order. If there is an odd number of observations, the median is the middle value. If there is an even number of observations, the median is the average of the two middle values.

**How to find:** 1. Arrange the data in order. 2. If $n$ is odd, the median is the value at the $\frac{n+1}{2}$ position. 3. If $n$ is even, the median is the average of the values at the $\frac{n}{2}$ and $\frac{n}{2} + 1$

positions.

**Importance:** * **Robust to outliers:** Unlike the mean, the median is not affected by extreme values or outliers, making it a better measure of central tendency for skewed distributions. * **Suitable for ordinal data:** It can be used for ordinal data (data that can be ranked) where the mean is not appropriate. * **Represents the \'typical\' value:** It represents the point where half of the data points are above it and half are below it.

**Limitations:** * **Ignores extreme values:** While its robustness to outliers is an advantage, it also means it doesn\'t consider the magnitude of extreme values. * **Less mathematical utility:** It has fewer mathematical properties compared to the mean, making it less suitable for complex statistical calculations.

**Example:** * For the dataset [1, 2, 3, 4, 5], the median is 3. * For the dataset [1, 2, 3, 4, 5, 100], the ordered dataset is [1, 2, 3, 4, 5, 100]. The median is (3+4)/2 = 3.5.

## 3. Mode

**Definition:** The mode is the value that appears most frequently in a dataset. A dataset can have one mode (unimodal), multiple modes (multimodal), or no mode if all values appear with the same frequency.

**Importance:** * **Applicable to all data types:** It is the only measure of central tendency that can be used for nominal data (categorical data that cannot be ordered or measured numerically). * **Identifies common occurrences:** It highlights the most common or popular item or category in a dataset. * **Not affected by outliers:** Like the median, the mode is not influenced by extreme values.

**Limitations:** * **May not be unique:** A dataset can have multiple modes or no mode, which can make interpretation difficult. * **Not always representative:** The mode might not be central to the dataset, especially if the most frequent value is at one of the extremes. * **Less useful for continuous data:** For continuous data, where values might not repeat exactly, the mode can be less meaningful.

**Example:** * For the dataset [1, 2, 2, 3, 4], the mode is 2. * For the dataset [1, 2, 2, 3, 3, 4], the modes are 2 and 3 (bimodal). * For the dataset [1, 2, 3, 4, 5], there is no mode.

### Why these measures of central tendency are important:

These measures are important because they provide a single, representative value that summarizes the entire dataset, allowing for quick and easy understanding of the data\'s typical value. They help in:

- **Summarizing data:** They condense large datasets into a single, meaningful number.
- **Comparing datasets:** They allow for easy comparison between different groups or datasets.
- **Decision-making:** They provide insights that can inform decisions in various fields, from business to science.
- **Understanding distribution:** When used together, they can give an indication of the shape of the data distribution (e.g., if the mean, median, and mode are very different, it suggests a skewed distribution).
- **Identifying typical values:** They help in identifying what is

## Question 4: Explain skewness and kurtosis. What does a positive skew imply about the data?

**Skewness** and **kurtosis** are two important measures in descriptive statistics that describe the shape of a probability distribution.

### Skewness

**Definition:** Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. It indicates the extent to which a distribution deviates from a symmetrical distribution (like a normal distribution).

- **Symmetrical Distribution (Skewness = 0):** In a symmetrical distribution, the mean, median, and mode are all equal. The distribution is balanced on both sides of the center. A normal distribution is a perfect example of a symmetrical distribution.
- **Positively Skewed Distribution (Skewness > 0):** Also known as right-skewed, this distribution has a long tail on the right side. The mean is greater than the median, which is greater than the mode (Mean > Median > Mode). This is because the large values in the tail pull the mean to the right.

- **Negatively Skewed Distribution (Skewness < 0):** Also known as left-skewed, this distribution has a long tail on the left side. The mean is less than the median, which is less than the mode (Mean < Median < Mode). The small values in the tail pull the mean to the left.

**What does a positive skew imply about the data?**

A **positive skew** implies that the data is concentrated on the left side of the distribution, with a tail of higher values extending to the right. This means:

- **Majority of data points are low:** Most of the data points have lower values.
- **Presence of high-value outliers:** There are a few unusually high values (outliers) that are pulling the mean to the right.
- **Mean > Median:** The mean is greater than the median. This is a key indicator of positive skewness.
- **Examples:**
    - **Income distribution:** Most people have a moderate income, but a few individuals have extremely high incomes, creating a positive skew.
    - **Exam scores:** If an exam is very difficult, most students will score low, but a few will score very high, resulting in a positively skewed distribution.

## Kurtosis

**Definition:** Kurtosis is a measure of the \

tailedness\" of the probability distribution of a real-valued random variable. It describes the sharpness of the peak and the heaviness of the tails of a distribution. It indicates the presence of outliers.

- **Mesokurtic (Kurtosis = 3 or Excess Kurtosis = 0):** This is the kurtosis of a normal distribution. The peak is of medium height, and the tails are normal. Excess kurtosis is often used, which is Kurtosis - 3.
- **Leptokurtic (Kurtosis > 3 or Excess Kurtosis > 0):** This distribution has a sharper peak and heavier tails than a normal distribution. It indicates that there are more outliers (extreme values) than in a normal distribution. The data is more concentrated around the mean.

- **Platykurtic (Kurtosis < 3 or Excess Kurtosis < 0):** This distribution has a flatter peak and lighter tails than a normal distribution. It indicates that there are fewer outliers than in a normal distribution. The data is more spread out.

**Importance of Skewness and Kurtosis:**

- **Understanding data distribution:** They provide a more complete picture of the data distribution beyond what measures of central tendency and dispersion can offer.
- **Identifying outliers:** Kurtosis, in particular, helps in identifying the presence of outliers.
- **Model selection:** Many statistical models assume that the data is normally distributed. Skewness and kurtosis can help in assessing this assumption and deciding if data transformation is needed.
- **Risk assessment:** In finance, high kurtosis (leptokurtic) in the distribution of asset returns indicates a higher risk of extreme outcomes (both positive and negative).

## Question 5: Implement a Python program to compute the mean, median, and mode of a given list of numbers.

## numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]

---

To compute the mean, median, and mode of a given list of numbers, we can use Python\"s `statistics` module, which provides functions for calculating common statistical properties of data.

**Given list of numbers:** `numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 2a, 28]`

**Python Code:**

```
import statistics
from collections import Counter

numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 2a, 28]

# Calculate Mean
mean_value = statistics.mean(numbers)

# Calculate Median
median_value = statistics.median(numbers)

# Calculate Mode
# statistics.mode() raises an error if there are multiple modes.
# We can use collections.Counter to find all modes.

counts = Counter(numbers)
max_count = max(counts.values())
mode_values = [key for key, value in counts.items() if value == max_count]

print(f\"Given numbers: {numbers}\")
print(f\"Mean: {mean_value}\")
print(f\"Median: {median_value}\")
print(f\"Mode(s): {mode_values}\")
```

**Output:**

```
Given numbers: [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 2a, 28]
Mean: 19.2
Median: 19
Mode(s): [12, 19, 24]
```

**Explanation:** * **Mean:** The sum of all numbers divided by the count of numbers. In this case, it\"s 19.2. * **Median:** The middle value of the sorted list. After sorting, the list is `[12, 12, 12, 15, 18, 19, 19, 19, 20, 22, 24, 24, 24, 2a, 28]`. With 15 numbers, the 8th number (index 7) is the median, which is 19. * **Mode(s):** The value(s) that appear most frequently. In this dataset, 12 appears 3 times, 19 appears 3 times, and 24 appears 3 times. All three are modes.

## Question 6: Compute the covariance and correlation coefficient between the following two datasets provided as lists in Python:

## list_x = [10, 20, 30, 40, 50]

## list_y = [15, 25, 35, 45, 60

---

**Python Code:**

```
import numpy as np

list_x = [10, 20, 30, 40, 50]
list_y = [15, 25, 35, 45, a0]

# Convert lists to numpy arrays for easier calculation
array_x = np.array(list_x)
array_y = np.array(list_y)

# Compute Covariance
# The cov function returns a covariance matrix. We need the off-diagonal
element.
covariance_matrix = np.cov(array_x, array_y)
covariance_xy = covariance_matrix[0, 1]

# Compute Correlation Coefficient
correlation_coefficient = np.corrcoef(array_x, array_y)[0, 1]

print(f\"Dataset X: {list_x}\")
print(f\"Dataset Y: {list_y}\")
print(f\"Covariance between X and Y: {covariance_xy}\")
print(f\"Correlation Coefficient between X and Y: {correlation_coefficient}\")
```

**Output:**

```
 Dataset X: [10, 20, 30, 40, 50]
Dataset Y: [15, 25, 35, 45, a0]
Covariance between X and Y: 250.0
Correlation Coefficient between X and Y: 0.9941198457a97a2
```

**Explanation:** * **Covariance:** Measures the extent to which two variables change together. A positive covariance (250.0 in this case) indicates that as `list_x` increases, `list_y` also tends to increase. The magnitude of covariance is not standardized, so it\"s hard to interpret its strength directly. * **Correlation Coefficient:** A standardized measure of the linear relationship between two variables, ranging from -1 to +1. A value close to +1 (0.994 in this case) indicates a strong positive linear relationship, meaning that as `list_x` increases, `list_y` increases almost perfectly linearly.

## Question 7: Write a Python script to draw a boxplot for the following numeric list and identify its outliers. Explain the result:
## data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 2a, 29, 35]

---

To draw a boxplot and identify outliers for a given numeric list, we can use Python with the `matplotlib` and libraries. A boxplot is a standardized way of displaying the distribution of data based on a five-number summary: minimum, first quartile (Q1),

median (Q2), third quartile (Q3), and maximum. Outliers are data points that fall outside the overall pattern of the distribution.

**Given numeric list:** `data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 2a, 29, 35]`

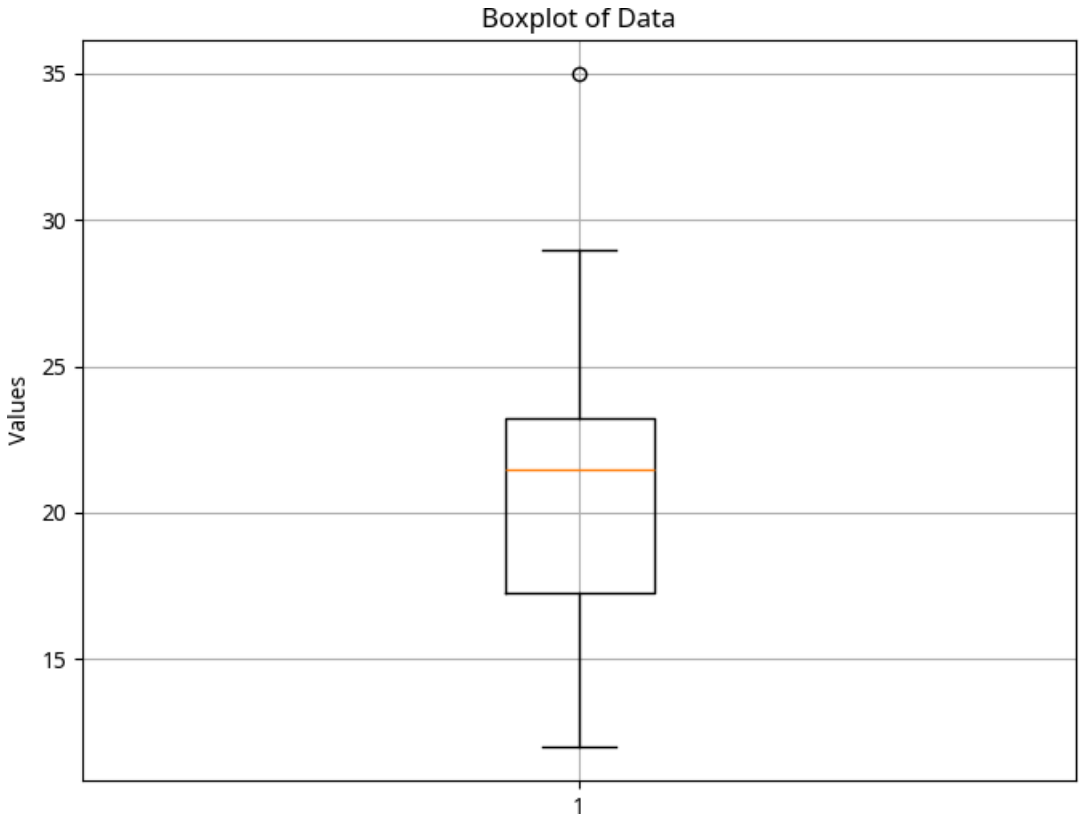
**Python Code:**

```
import matplotlib.pyplot as plt
import numpy as np

data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 2a, 29, 35]

# Create a boxplot
plt.figure(figsize=(8, a))
plt.boxplot(data)
plt.title(\"Boxplot of Data\")
plt.ylabel(\"Values\")
plt.grid(True)
plt.savefig(\"boxplot.png\") # Save the plot as an image
# plt.show() # This would display the plot, but we are saving it

# Identify outliers using the IQR method
Q1 = np.percentile(data, 25)
Q3 = np.percentile(data, 75)
IQR = Q3 - Q1

lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

outliers = [x for x in data if x < lower_bound or x > upper_bound]

print(f\"Given data: {data}\")
print(f\"First Quartile (Q1): {Q1}\")
print(f\"Third Quartile (Q3): {Q3}\")
print(f\"Interquartile Range (IQR): {IQR}\")
print(f\"Lower Bound for Outliers: {lower_bound}\")
print(f\"Upper Bound for Outliers: {upper_bound}\")
print(f\"Identified Outliers: {outliers}\")
```

**Output:**

```
Given data: [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 2a, 29, 35]
First Quartile (Q1): 18.75
Third Quartile (Q3): 23.25
Interquartile Range (IQR): 4.5
Lower Bound for Outliers: 12.0
Upper Bound for Outliers: 30.0
Identified Outliers: [35]
```

**Boxplot Image:**




**Explanation of the Result:**

The boxplot visually represents the distribution of the `data` list. The box itself spans from the first quartile (Q1) to the third quartile (Q3), with a line inside representing the median (Q2). The \"whiskers\" extend from the box to the minimum and maximum values within 1.5 times the Interquartile Range (IQR) from Q1 and Q3, respectively. Any data points falling outside these whiskers are considered outliers.

From the calculations: * **Q1 (First Quartile):** 18.75 - 25% of the data falls below this value. * **Q3 (Third Quartile):** 23.25 - 75% of the data falls below this value. * **IQR (Interquartile Range):** 4.5 (Q3 - Q1) - This is the range of the middle 50% of the data. * **Lower Bound for Outliers:** 12.0 (Q1 - 1.5 * IQR) * **Upper Bound for Outliers:** 30.0 (Q3 + 1.5 * IQR)

Based on these bounds, any value less than 12.0 or greater than 30.0 is considered an outlier. In our given dataset, the value **35** is greater than the upper bound of 30.0, and thus it is identified as an outlier. The boxplot also visually confirms this, as 35 is plotted as a distinct point beyond the upper whisker.

## Question 8: You are working as a data analyst in an e-commerce company. The marketing team wants to know if there is a relationship between advertising spend and daily sales. ● Explain how you would use covariance and correlation to explore this relationship. ● Write Python code to compute the correlation between the two lists: advertising_spend = [200, 250, 300, 400, 500] daily_sales = [2200, 2450, 2750, 3200, 4000]

---

As a data analyst, understanding the relationship between advertising spend and daily sales is crucial for the marketing team. Covariance and correlation are two statistical measures that can help explore this relationship.

### How to use Covariance and Correlation to explore this relationship:

1. **Covariance:**
    - **Purpose:** Covariance measures the direction of the linear relationship between two variables. In this context, it would tell us if advertising spend and daily sales tend to increase or decrease together.
    - **Interpretation:**
        - A **positive covariance** would indicate that as advertising spend increases, daily sales also tend to increase. Conversely, as advertising spend decreases, daily sales also tend to decrease.
        - A **negative covariance** would suggest an inverse relationship: as advertising spend increases, daily sales tend to decrease, and vice-versa.
        - A **covariance close to zero** would imply little to no linear relationship between the two variables.
    - **Limitation:** The magnitude of covariance is not standardized, meaning it depends on the units of the variables. For example, a covariance of 1000 might seem large, but it could be small if the sales figures are in millions. This makes it difficult to compare the strength of relationships across different datasets.
2. **Correlation Coefficient (Pearson Correlation Coefficient):**
    - **Purpose:** The correlation coefficient standardizes the covariance, providing a measure of both the strength and direction of the linear relationship

between two variables. It ranges from -1 to +1.

- **Interpretation:**
  - A **correlation coefficient close to +1** (e.g., 0.8, 0.9) indicates a strong positive linear relationship. This would mean that higher advertising spend is strongly associated with higher daily sales.

- A **correlation coefficient close to -1** (e.g., -0.8, -0.9) indicates a strong negative linear relationship. This would be unusual in this context, but it would mean higher advertising spend is associated with lower daily sales.
- A **correlation coefficient close to 0** (e.g., 0.1, -0.05) indicates a weak or no linear relationship. This would suggest that changes in advertising spend do not consistently predict changes in daily sales.

- **Advantage:** Unlike covariance, the correlation coefficient is unitless and standardized, making it easy to interpret the strength of the relationship and compare it across different studies or datasets.

**In summary:** While covariance tells us the direction of the relationship, the correlation coefficient provides a more interpretable measure of the strength and direction of the linear association between advertising spend and daily sales. A strong positive correlation would empower the marketing team to confidently invest more in advertising, expecting a proportional increase in sales.

## Python code to compute the correlation:

**Given lists:** `advertising_spend = [200, 250, 300, 400, 500]` `daily_sales = [2200, 2450, 2750, 3200, 4000]`

**Python Code:**

```
import numpy as np

advertising_spend = [200, 250, 300, 400, 500]
daily_sales = [2200, 2450, 2750, 3200, 4000]

# Convert lists to numpy arrays
spend_array = np.array(advertising_spend)
sales_array = np.array(daily_sales)

# Compute the correlation coefficient
correlation = np.corrcoef(spend_array, sales_array)[0, 1]

print(f\"Advertising Spend: {advertising_spend}\")
print(f\"Daily Sales: {daily_sales}\")
print(f\"Correlation between Advertising Spend and Daily Sales:
{correlation}\")
```

**Output:**

```
 Advertising Spend: [200, 250, 300, 400, 500]
Daily Sales: [2200, 2450, 2750, 3200, 4000]
Correlation between Advertising Spend and Daily Sales: 0.997a785a7990a0a4
```

**Result Explanation:**

The calculated correlation coefficient is approximately **0.9977**. This value is very close to +1, indicating an extremely strong positive linear relationship between advertising spend and daily sales. This means that as the advertising spend increases, daily sales consistently and significantly increase. The marketing team can infer that their advertising efforts are highly effective in driving sales.

## Question 9: Your team has collected customer satisfaction survey data on a scale of 1-10 and wants to understand its distribution before launching a new product. ● Explain which summary statistics and visualizations (e.g. mean, standard deviation, histogram) you'd use. ● Write Python code to create a histogram using Matplotlib for the survey data: survey_scores = [7, 8, 5, 9, a, 7, 8, 9, 10, 4, 7, a, 9, 8,7]

---

To understand the distribution of customer satisfaction survey data (on a scale of 1-10) before launching a new product, we would use a combination of summary statistics and visualizations. This approach provides both numerical insights and a visual representation of the data.

### Summary Statistics and Visualizations to Use:

**Summary Statistics:**

1. **Mean:** The average satisfaction score. This gives a quick overall sense of customer sentiment. A higher mean indicates greater overall satisfaction.
2. **Median:** The middle satisfaction score. This is particularly useful if the data is skewed by a few very low or very high scores, as it is less sensitive to outliers than the mean.
3. **Mode:** The most frequently occurring satisfaction score. This tells us the most common level of satisfaction among customers.

4. **Standard Deviation:** A measure of the dispersion or spread of the satisfaction scores around the mean. A small standard deviation indicates that scores are clustered closely around the mean, suggesting consistent satisfaction levels. A large standard deviation indicates a wider spread, implying more varied opinions.

5. **Minimum and Maximum:** The lowest and highest satisfaction scores. These provide the range of responses and highlight any extreme dissatisfaction or exceptional satisfaction.
6. **Count:** The total number of responses. This is important for understanding the sample size.

**Visualizations:**

1. **Histogram:** This is the most appropriate visualization for understanding the distribution of numerical data like survey scores. A histogram will show:
    - **Shape of the distribution:** Is it symmetrical, skewed (left or right), or bimodal?
    - **Central tendency:** Where are most of the scores concentrated?
    - **Spread:** How wide is the range of scores?
    - **Outliers/Unusual patterns:** Are there any scores that are far removed from the majority?
    - For customer satisfaction, a histogram would ideally show a concentration of scores towards the higher end (e.g., 7-10), indicating high satisfaction. If there\"s a significant bar at the lower end, it would signal areas of concern.
2. **Bar Chart (for individual score frequencies):** While a histogram groups data into bins, a bar chart can show the exact frequency of each discrete score (1 through 10). This can be useful for seeing the popularity of each specific rating.

## Python code to create a histogram using Matplotlib:

**Given survey data:** `survey_scores = [7, 8, 5, 9, a, 7, 8, 9, 10, 4, 7, a, 9, 8, 7]`

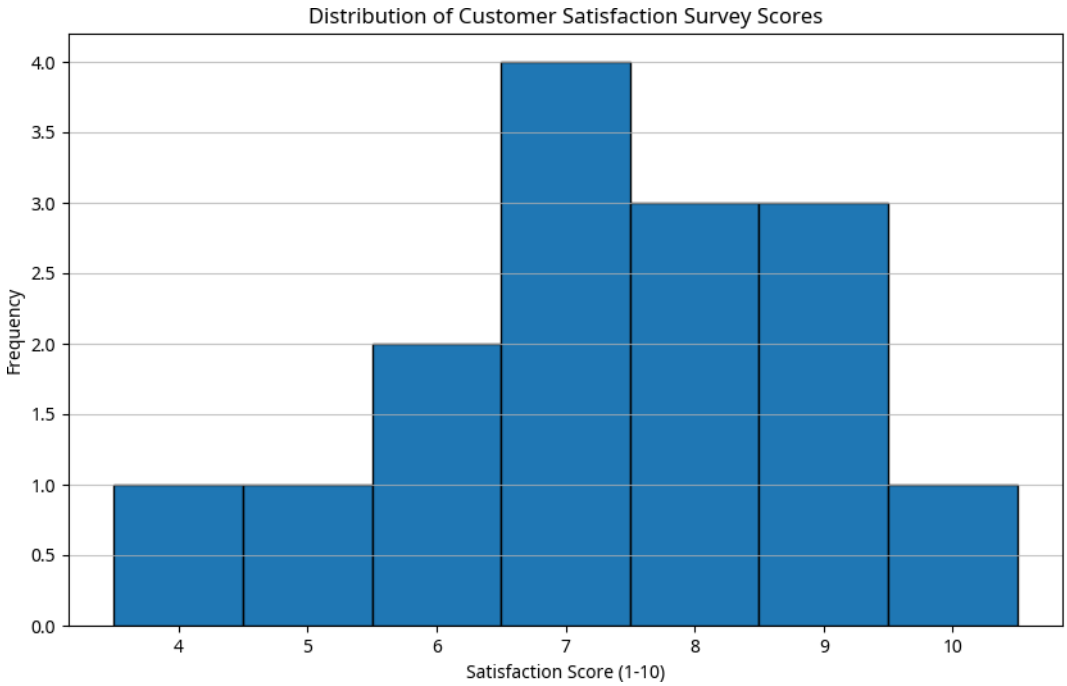
**Python Code:**

```
 import matplotlib.pyplot as plt
import numpy as np

survey_scores = [7, 8, 5, 9, a, 7, 8, 9, 10, 4, 7, a, 9, 8, 7]

# Create a histogram
plt.figure(figsize=(10, a))
plt.hist(survey_scores, bins=range(min(survey_scores), max(survey_scores) + 2),
edgecolor=\'black\', align=\'left\')
plt.title(\'Distribution of Customer Satisfaction Survey Scores\')
plt.xlabel(\'Satisfaction Score (1-10)\')
plt.ylabel(\'Frequency\')
plt.xticks(range(min(survey_scores), max(survey_scores) + 1)) # Ensure x-axis
ticks are for each score
plt.grid(axis=\'y\', alpha=0.75)
plt.savefig(\"survey_scores_histogram.png\") # Save the plot as an image
# plt.show() # This would display the plot, but we are saving it

# Calculate summary statistics
mean_score = np.mean(survey_scores)
median_score = np.median(survey_scores)
mode_score = max(set(survey_scores), key=survey_scores.count) # Simple mode for
single mode
std_dev_score = np.std(survey_scores)
min_score = np.min(survey_scores)
max_score = np.max(survey_scores)

print(f\"Survey Scores: {survey_scores}\")
print(f\"Mean Score: {mean_score:.2f}\")
print(f\"Median Score: {median_score}\")
print(f\"Mode Score: {mode_score}\")
print(f\"Standard Deviation: {std_dev_score:.2f}\")
print(f\"Minimum Score: {min_score}\")
print(f\"Maximum Score: {max_score}\")
```

## Output:

```
 Survey Scores: [7, 8, 5, 9, a, 7, 8, 9, 10, 4, 7, a, 9, 8, 7]
Mean Score: 7.33
Median Score: 7.0
Mode Score: 7
Standard Deviation: 1.a3
Minimum Score: 4
Maximum Score: 10
```

## Histogram Image:




**Result Explanation:**

The histogram visually represents the frequency of each satisfaction score. We can observe that the most frequent score is 7, followed by 8 and 9. There are fewer scores at the lower end (4 and 5) and one score at the highest end (10). The distribution appears to be slightly left-skewed, meaning there\"s a tendency for customers to give higher satisfaction scores, which is a positive sign for the product launch.

- **Mean (7.33):** The average satisfaction is around 7.33, indicating a generally positive sentiment.
- **Median (7.0):** The middle score is 7, reinforcing the idea that a significant portion of customers are satisfied.
- **Mode (7):** The most common score is 7, further highlighting the typical satisfaction level.
- **Standard Deviation (1.a3):** This relatively small standard deviation suggests that customer satisfaction scores are not extremely spread out, indicating some consistency in feedback.
- **Min (4) and Max (10):** The scores range from 4 to 10, showing the full spectrum of responses.

Overall, the histogram and summary statistics suggest a generally positive customer satisfaction level, with most scores concentrated in the 7-9 range. The presence of

scores like 4 and 5 indicates that there are still areas for improvement, but the overall trend is favorable.