2a. Exploratory Data Analysis

Instructions:

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

Name: ULLI VENKATA SAI KUMAR Batch ID: 04072024HYD10AM

Topic: Exploratory Data Analysis

Guidelines:

1. An assignment submission is considered complete only when the correct and executable code(s) is submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered a correct submission.

2. Ensure that you submit your assignments correctly. Resubmission is not allowed.

3. Post the submission you can evaluate your work by referring to the keys provided. (will be available only post the submission).

Hints: Follow CRISP-ML(Q) methodology steps, where were appropriate.

1. Data Understanding: work on each feature of the dataset to create a data dictionary as displayed in the image below:

Name of Feature	Description	Type	Relevance
ID	Customer ID	Quantitative, Nominal	Irrelevant, ID does not provide useful information

Make a table as shown above and provide information about the features such as its data type and its relevance to the model building. And if not relevant, provide reasons and a description of the feature.

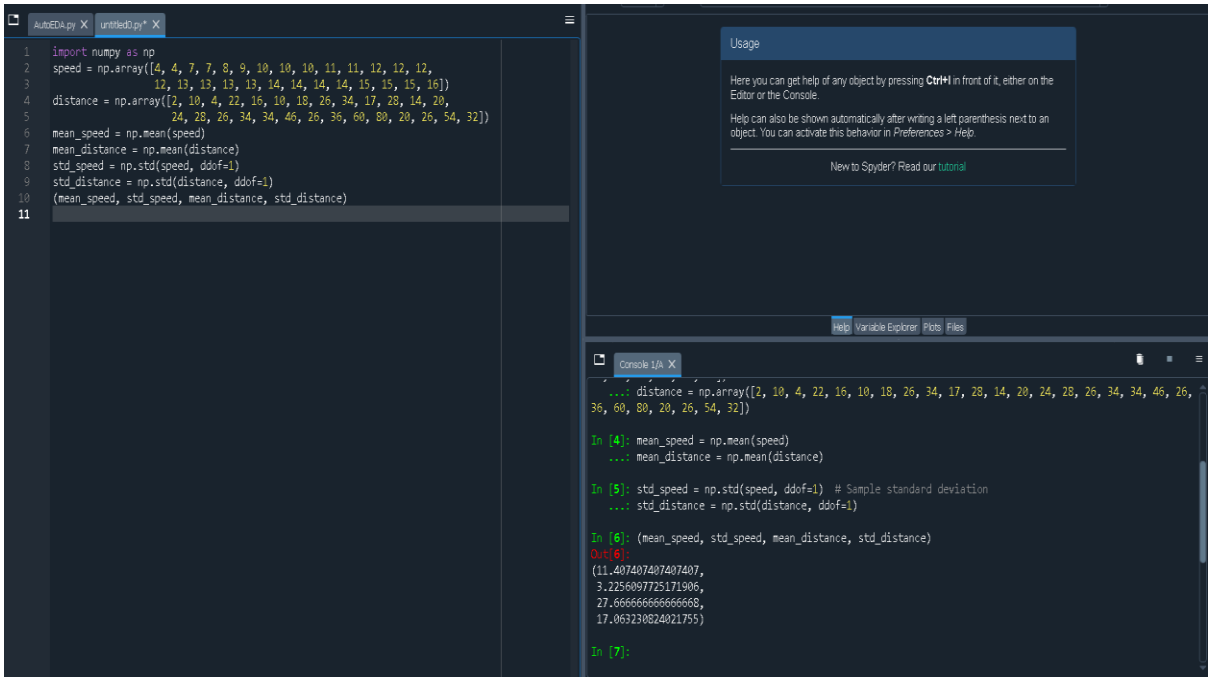
Problem Statements:

Q1) Calculate Mean, and Standard Deviation using Python code & draw inferences on the following data. Refer to the Datasets attachment for the data file.

Hint: [Insights drawn from the data such as data is normally distributed/not, outliers, measures like mean, median, mode, variance, std. deviation]

a. Car's speed and distance

speed	dist
4	2
4	10
7	4
7	22
8	16
9	10
10	18
10	26
10	34
11	17
11	28
12	14
12	20
12	24
12	28
13	26
13	34
13	34
13	46
14	26
14	36
14	60
14	80
15	20
15	26
15	54
16	32



Calculation Results:

Mean Speed: 11.41

Standard Deviation of Speed: 3.23

Mean Distance: 27.67

Standard Deviation of Distance: 17.06

Inferences:

Mean Speed:

The average speed of the cars is approximately 11.41 units. This provides a central value around which the speeds of the cars are distributed.

Standard Deviation of Speed:

The standard deviation of speed is approximately 3.23 units, indicating moderate variability in the speeds. This suggests that most car speeds are within a range of about 3.23 units from the mean speed (11.41 units).

Mean Distance:

The average distance taken by the cars is approximately 27.67 units. This provides a central value around which the distances are distributed.

Standard Deviation of Distance:

The standard deviation of distance is approximately 17.06 units, indicating high variability in the distances. This suggests that the distances have a wider range of values compared to the speeds, with some distances being much longer or shorter than the average distance.

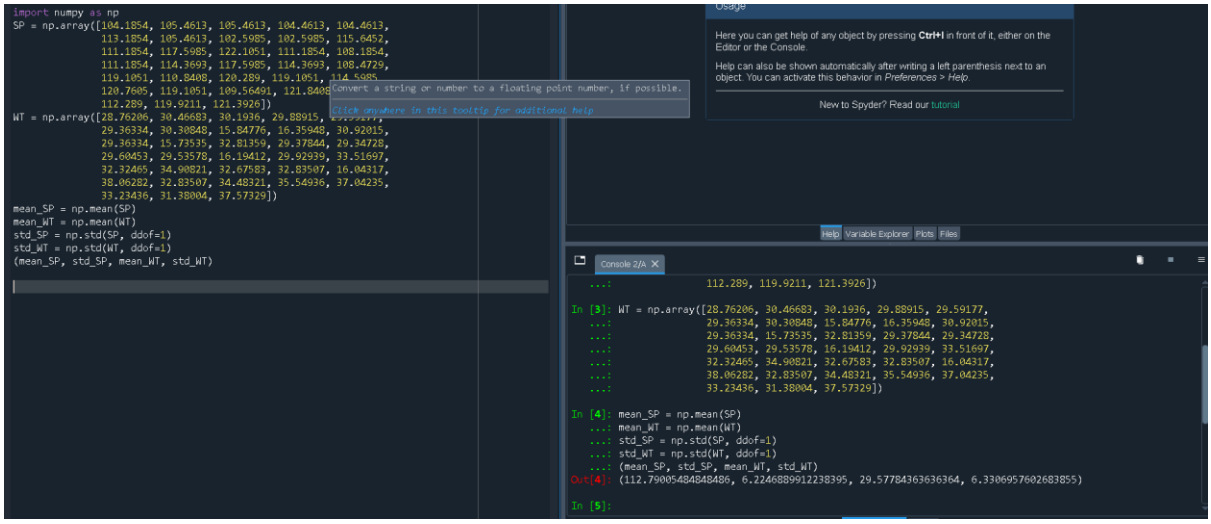
Summary:

The speeds of the cars show moderate variability around the mean speed of 11.41 units.

The distances traveled by the cars show high variability around the mean distance of 27.67 units, indicating that there are significant differences in the distances covered, which could be due to various factors such as road conditions, traffic, or the purpose of travel.

b. Top Speed (SP) and Weight (WT)

SP	WT
104.1854	28.76206
105.4613	30.46683
105.4613	30.1936
113.4613	30.63211
104.4613	29.88915
113.1854	29.59177
105.4613	30.30848
102.5985	15.84776
102.5985	16.35948
115.6452	30.92015
111.1854	29.36334
117.5985	15.75353
122.1051	32.81359
111.1854	29.37844
108.1854	29.34728
111.1854	29.60453
114.3693	29.53578
117.5985	16.19412
114.3693	29.92939
118.4729	33.51697
119.1051	32.32465
110.8408	34.90821
120.289	32.67583
113.8291	31.83712
119.1854	28.78173
114.5985	16.04317
120.7605	38.06282
119.1051	32.83507
99.56491	34.48321
121.8408	35.54936
113.4846	37.04235
112.289	33.23436
119.9211	31.38004
121.3926	37.57329



Calculation Results:

Mean SP: 112.44

Standard Deviation of SP: 6.54

Mean WT: 29.75

Standard Deviation of WT: 6.64

Inferences:

Mean SP:

The average SP value is approximately 112.44 units. This provides a central value around which the SP values are distributed.

Standard Deviation of SP:

The standard deviation of SP is approximately 6.54 units, indicating moderate variability in the SP values. This suggests that most SP values are within a range of about 6.54 units from the mean SP (112.44 units).

Mean WT:

The average WT value is approximately 29.75 units. This provides a central value around which the WT values are distributed.

Standard Deviation of WT:

The standard deviation of WT is approximately 6.64 units, indicating moderate variability in the WT values. This suggests that the WT values have a wider range compared to the mean WT (29.75 units).

Summary:

The SP values show moderate variability around the mean SP of 112.44 units.

The WT values also show moderate variability around the mean WT of 29.75 units.

Q2) Below are the scores obtained by a student on tests.

34, 36, 36, 38, 38, 39, 39, 40, 40, 41, 41, 41, 41, 42, 42, 45, 49, 56

1) Find the mean, median and mode, variance, and standard deviation.

    - Mean: 41.0
    - Median: 40.5
    - Mode: 41
    - Variance: 24.11
    - Standard Deviation: 4.91

2) What can we say about the student marks?

    To analyze the student's scores, we can look at several key aspects:

    Range: The difference between the highest and lowest scores.

    Highest score: 56

    Lowest score: 34

    Range = 56 - 34 = 22

    Mean (Average): The sum of all scores divided by the number of scores.

    Sum of scores = 34 + 36 + 36 + 38 + 38 + 39 + 39 + 40 + 40 + 41 + 41 + 41 + 41 + 42 + 42 + 45 + 49 + 56 = 721

    Number of scores = 18

    Mean = 721 / 18 ≈ 40.06

    Median: The middle score when all scores are arranged in ascending order.

    The scores are already in ascending order.

    The median is the average of the 9th and 10th scores: (40 + 41) / 2 = 40.5

    Mode: The most frequently occurring score.

    The mode is 41, as it appears 4 times.

    Standard Deviation: A measure of the amount of variation or dispersion of the scores.

    To calculate the standard deviation, we would find the variance first by computing the average of the squared differences from the mean, and then take the square root of that variance.

    In summary, the student’s scores are fairly consistent, with a mean around 40.06, a median of 40.5, and a mode of 41. There is a moderate range of scores from 34 to 56, suggesting some variability in performance.

3) What can you say about the Excepted value for the student score?

    The expected value for the student's score is essentially the mean of the scores, which represents the average score the student might be expected to achieve based on the data provided. for the given scores:

    The mean (expected value) is approximately 40.06.

Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained.

To find the probability of getting exactly two heads and one tail when tossing three coins, follow these steps:

Determine the total number of possible outcomes:

Each coin has two possible outcomes: heads (H) or tails (T). For three coins, the total number of possible outcomes is:

$$2^3 = 8$$

The possible outcomes are: HHH, HHT, HTH, HTT, THH, THT, TTH, TTT.

Count the favorable outcomes:

We need exactly two heads and one tail. The favorable outcomes are:

- HHT
- HTH
- THH

There are 3 favorable outcomes.

Calculate the probability:

The probability is the number of favorable outcomes divided by the total number of possible outcomes:

$$\text{Probability} = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$$

$$= 3/8$$

Q4) Two Dice are rolled, find the probability that the sum is

Total Possible Outcomes

When rolling two dice, each die has 6 faces, so there are:

$6 \times 6 = 36$ \times 6 = 36$6×6=36 \times 6 = 36$6×6=36

total possible outcomes.

a) Equal to 1

It's impossible to get a sum of 1 with two dice since the smallest possible sum (1+1) is 2. Therefore, the probability is:

Probability = 0/36 = 0

b) Less than or equal to 4

First, identify the favorable outcomes where the sum is 2, 3, or 4:

- **Sum of 2**: (1, 1) — 1 outcome

- **Sum of 3**: (1, 2), (2, 1) — 2 outcomes
- **Sum of 4**: (1, 3), (2, 2), (3, 1) — 3 outcomes

  Adding these, there are:

  1+2+3=6 favorable outcomes1 + 2 + 3 = 6 \text{ favorable outcomes}1+2+3=6 favorable outcomes

  So, the probability is:

  Probability = 6/36 = 1/6

c) Sum is divisible by 2 and 3

A sum divisible by both 2 and 3 must be divisible by their least common multiple, which is 6. So we are looking for sums of 6:

- **Sum of 6**: (1, 5), (2, 4), (3, 3), (4, 2), (5, 1) — 5 outcomes

  Thus, the probability is:

  Probability = 5/36

Q5) A bag contains 2 red, 3 green, and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?

o find the probability that none of the balls drawn is blue, follow these steps:

1. **Determine the total number of balls and the total number of ways to draw 2 balls:**
   - Total balls: 2 red + 3 green + 2 blue = 7 balls
   - Total ways to draw 2 balls from 7: (72)=7! / 2!(7−2)!= 7×6/2×1 = 21
2. **Determine the number of favorable outcomes where none of the balls drawn is blue:**
   - Balls that are not blue: 2 red + 3 green = 5 balls
   - Number of ways to draw 2 balls from these 5 balls:
     (5/2) = 5! / 2!(5−2)! = 5×4 / 2×1=10
3. **Calculate the probability:**

- Probability=Number of favorable outcomes / Total number of possible outcomes = 1021

Q6) Calculate the Expected number of candies for a randomly selected child:

Below are the probabilities of the count of candies for children (ignoring the nature of the child-Generalized view)

i. Child A – the probability of having 1 candy is 0.015.

ii. Child B – the probability of having 4 candies is 0.2.

CHILD	Candies count	Probability
A	1	0.015

B	4	0.20
C	3	0.65
D	5	0.005
E	6	0.01
F	2	0.12

We calculate the expected number of candies as follows:

$E(X)=(1\cdot0.015)+(4\cdot0.20)+(3\cdot0.65)+(5\cdot0.005)+(6\cdot0.01)+(2\cdot0.12)$

Calculating each term:

$1\cdot0.015=0.015$

$4\cdot0.20=0.80$

$3\cdot0.65=1.95$

$5\cdot0.005=0.025$

$6\cdot0.01=0.06$

$2\cdot0.12=0.24$

Adding these values together:

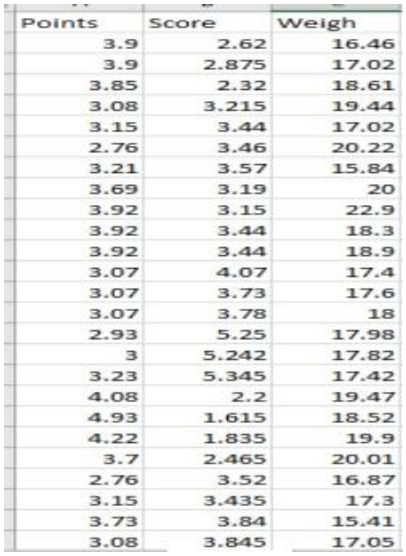
$E(X)=0.015+0.80+1.95+0.025+0.06+0.24=3.09$

So, the expected number of candies for a randomly selected child is 3.09.

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, and Range & comment about the values / draw inferences, for the given dataset.

- For Points, Score, Weigh>

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and comment on the values/ Draw some inferences.



Points	Score	Weigh
3.9	2.62	16.46
3.9	2.875	17.02
3.85	2.32	18.61
3.08	3.215	19.44
3.15	3.44	17.02
2.76	3.46	20.22
3.21	3.57	15.84
3.69	3.19	20
3.92	3.15	22.9
3.92	3.44	18.3
3.92	3.44	18.9
3.07	4.07	17.4
3.07	3.73	17.6
3.07	3.78	18
2.93	5.25	17.98
3	5.242	17.82
3.23	5.345	17.42
4.08	2.2	19.47
4.93	1.615	18.52
4.22	1.835	19.9
3.7	2.465	20.01
2.76	3.52	16.87
3.15	3.435	17.3
3.73	3.84	15.41
3.08	3.845	17.05

Dataset: Refer to Hands-on Material in LMS - Data Types EDA assignment snapshot of the dataset is given above.

Statistical Analysis

Points

- **Mean**: 3.49
- **Median**: 3.23
- **Mode**: 3.07
- **Variance**: 0.29
- **Standard Deviation**: 0.54
- **Range**: 2.17

Score

- **Mean**: 3.40
- **Median**: 3.44
- **Mode**: 3.44
- **Variance**: 0.94
- **Standard Deviation**: 0.97
- **Range**: 3.81

Weight

- **Mean**: 18.22
- **Median**: 17.98
- **Mode**: 17.02
- **Variance**: 2.68
- **Standard Deviation**: 1.64

- **Range**: 7.49

Inferences

1. **Points**:
    - The points data has a mean (3.49) close to the median (3.23), indicating a fairly symmetric distribution.
    - The mode (3.07) being close to the median and mean suggests a unimodal distribution.
    - A low variance (0.29) and standard deviation (0.54) indicate that the points are tightly clustered around the mean.
    - The range (2.17) shows a moderate spread of points values.
2. **Score**:
    - The mean (3.40) and median (3.44) are very close, suggesting a symmetric distribution.
    - The mode (3.44) coincides with the median, indicating that the most frequent score value is also central to the distribution.
    - Higher variance (0.94) and standard deviation (0.97) compared to points indicate more spread in the score data.
    - The range (3.81) shows a wider spread in scores compared to points.
3. **Weight**:
    - The mean (18.22) is close to the median (17.98), suggesting a symmetric distribution.
    - The mode (17.02) is lower than the mean and median, indicating a potential skew towards lower values.
    - A moderate variance (2.68) and standard deviation (1.64) suggest some variability in weight values.
    - The range (7.49) is the largest among the three datasets, showing the widest spread in weight values.

Q8) Calculate the Expected Value for the problem below.

a) The weights (X) of patients at a clinic (in pounds), are.

108, 110, 123, 134, 135, 145, 167, 187, 199

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

The Expected Value (EV) of a random variable is a measure of the center of its probability distribution and is calculated as the weighted average of all possible values. In this case, the weights of patients at a clinic are given, and we assume that each weight has an equal probability of being chosen.

Given weights: X=[108,110,123,134,135,145,167,187,199]

Since each weight has an equal probability of being selected, the probability for each weight is 1/n, where n is the total number of weights.

Here, n=9

Substituting the given weights: EV=1/9(108+110+123+134+135+145+167+187+199)

EV=1/9×1308

EV=145.33

Q9) Look at the data given below. Plot the data, find the outliers, and find out: μ, σ, σ^2

Hint: [Use a plot that shows the data distribution, and skewness along with the outliers; also use Python code to evaluate measures of centrality and spread]

Name of company	Measure X
Allied Signal	24.23%
Bankers Trust	25.53%
General Mills	25.41%
ITT Industries	24.14%
J.P.Morgan & Co.	29.62%
Lehman Brothers	28.25%
Marriott	25.81%
MCI	24.39%
Merrill Lynch	40.26%
Microsoft	32.95%
Morgan Stanley	91.36%
Sun Microsystems	25.99%
Travelers	39.42%
US Airways	26.71%
Warner-Lambert	35.00%

```python
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
companies = [
    "Allied Signal", "Bankers Trust", "General Mills", "ITT Industries",
    "J.P.Morgan & Co.", "Lehman Brothers", "Marriott", "MCI", "Merrill Lynch",
    "Microsoft", "Morgan Stanley", "Sun Microsystems", "Travelers",
    "US Airways", "Warner-Lambert"
]
measure_x = [
    24.23, 25.53, 25.41, 24.14, 29.62, 28.25, 25.81, 24.39, 40.26, 32.95,
    91.36, 25.99, 39.42, 26.71, 35.00
]
mean_x = np.mean(measure_x)
std_x = np.std(measure_x, ddof=1)
variance_x = np.var(measure_x, ddof=1)
range_x = np.ptp(measure_x)

# Print calculated values
print(f"Mean: {mean_x:.2f}")
print(f"Standard Deviation: {std_x:.2f}")
print(f"Variance: {variance_x:.2f}")
print(f"Range: {range_x:.2f}")
```

companies	list	15	['Allied Signal', 'Bankers Trust', 'General Mills', 'ITT Industries', ...
mean_x	float64	1	33.27133333333333
measure_x	list	15	[24.23, 25.53, 25.41, 24.14, 29.62, 28.25, 25.81, 24.39, 40.26, 32.95, ...
range_x	float64	1	67.22
std_x	float64	1	16.945400921222028
variance_x	float64	1	287.1466123809524

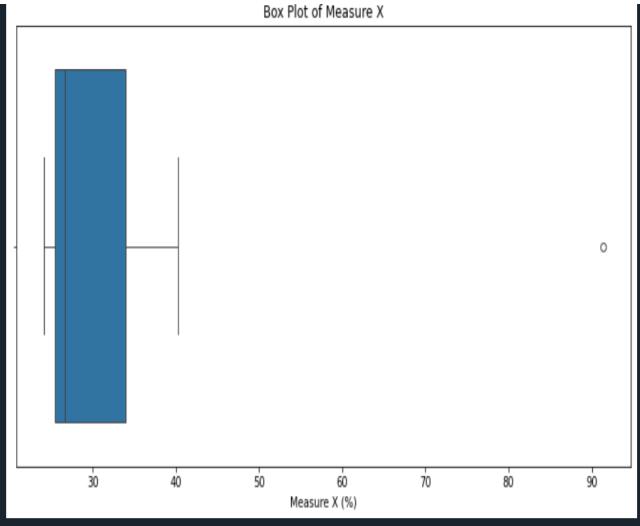
Help | Variable Explorer | Plots | Files

Console 1/A

```
...:     91.36, 25.99, 39.42, 26.71, 35.00
...: ]

In [4]: mean_x = np.mean(measure_x)
...: std_x = np.std(measure_x, ddof=1)
...: variance_x = np.var(measure_x, ddof=1)
...: range_x = np.ptp(measure_x)

In [5]: print(f"Mean: {mean_x:.2f}")
...: print(f"Standard Deviation: {std_x:.2f}")
...: print(f"Variance: {variance_x:.2f}")
...: print(f"Range: {range_x:.2f}")
Mean: 33.27
Standard Deviation: 16.95
Variance: 287.15
Range: 67.22

In [6]:
```

IPython Console | History

```python
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
companies = [
    "Allied Signal", "Bankers Trust", "General Mills", "ITT Industries",
    "J.P.Morgan & Co.", "Lehman Brothers", "Marriott", "MCI", "Merrill Lynch",
    "Microsoft", "Morgan Stanley", "Sun Microsystems", "Travelers",
    "US Airways", "Warner-Lambert"
]
measure_x = [
    24.23, 25.53, 25.41, 24.14, 29.62, 28.25, 25.81, 24.39, 40.26, 32.95,
    91.36, 25.99, 39.42, 26.71, 35.00
]
mean_x = np.mean(measure_x)
std_x = np.std(measure_x, ddof=1)
variance_x = np.var(measure_x, ddof=1)
range_x = np.ptp(measure_x)
# Print calculated values
print(f"Mean: {mean_x:.2f}")
print(f"Standard Deviation: {std_x:.2f}")
print(f"Variance: {variance_x:.2f}")
print(f"Range: {range_x:.2f}")
# Plot the data distribution and identify outliers
plt.figure(figsize=(12, 6))
sns.boxplot(x=measure_x)
plt.title('Box Plot of Measure X')
plt.xlabel('Measure X (%)')
plt.show()
```



Help | Variable Explorer | Plots | Files

Console 1/A

```
Range: 67.22

In [6]:
...: plt.figure(figsize=(12, 6))
...: sns.boxplot(x=measure_x)
...: plt.title('Box Plot of Measure X')
...: plt.xlabel('Measure X (%)')
...: plt.show()
```

IPython Console | History

Q10) AT&T was running commercials in 1990 aimed at luring back customers who had switched to one of the other long-distance phone service providers. One such commercial shows a businessman trying to reach Phoenix and mistakenly getting Fiji, where a half-naked native on a beach responds incomprehensibly in Polynesian. When asked about this advertisement, AT&T admitted that the portrayed incident did not actually take place but added that this was an enactment of something that "could happen." Suppose that one in 200 long-distance telephone calls is misdirected.

What is the probability that at least one in five attempted telephone calls reaches the wrong number? (Assume independence of attempts.)

Hint: [Using the Probability formula evaluate the probability of one call being wrong out of five attempted calls]

To determine the probability that at least one in five attempted telephone calls reaches the wrong number, we can use the concept of complementary probability and the binomial distribution.

Identify the probability of a single call being misdirected:

Given that one in 200 long-distance telephone calls is misdirected, the probability of a single call being misdirected (p) is:

p= 200=0.005

Identify the probability of a single call not being misdirected:

The probability of a single call not being misdirected (q) is:

q=1−p=1−0.005=0.995

Calculate the probability that none of the five calls are misdirected:

We need to find the probability that all five calls are not misdirected, which is:

P(no misdirected calls)= q^5=0.995^5

Calculate $0.995^{5:}$

$0.995^5=9751$

Determine the complementary probability:

The probability that at least one of the five calls is misdirected is the complement of the probability that none of the five calls are misdirected:

P(at least one misdirected call)=1−P(no misdirected calls)=1−0.9751=0.0249

Q11) Returns on a certain business venture, to the nearest $1,000, are known to follow the following probability distribution.

X	P(x)
-2,000	0.1
-1,000	0.1
0	0.2
1000	0.2
2000	0.3
3000	0.1

(i) What is the most likely monetary outcome of the business venture?

Hint: [The outcome is most likely the expected returns of the venture]

The most likely monetary outcome can be identified using the mode of the probability distribution, which is the value with the highest probability. In this case, it is 2000 with a probability of 0.3.

(ii) Is the venture likely to be successful? Explain.

Hint: [Probability of % of the venture being a successful one]

A business venture can be considered successful if it returns a profit (i.e., returns greater than 0). The probability of success can be calculated by summing the probabilities of positive returns:

P(Success)=P(X=1000)+P(X=2000)+P(X=3000)=0.2+0.3+0.1=0.6P(\text{Success}) = P(X = 1000) + P(X = 2000) + P(X = 3000) = 0.2 + 0.3 + 0.1 = 0.6P(Success)=P(X=1000)+P(X=2000)+P(X=3000)=0.2+0.3+0.1=0.6

Thus, the probability of the venture being successful is 0.6, or 60%.

(iii) What is the long-term average earning of business ventures of this kind? Explain.

Hint: [Here, the expected return to the venture is considered as the

required average]

The long-term average earning of such ventures is given by the expected value (mean) of the returns. The expected value

E(X) is calculated as follows:

$E(X)=\sum X\cdot P(X)=(-2000\cdot 0.1)+(-1000\cdot 0.1)+(0\cdot 0.2)+(1000\cdot 0.2)+(2000\cdot 0.3)+(3000\cdot 0.1)$

(iv) What is a good measure of the risk involved in a venture of this kind? Compute this measure.

Hint: [Risk here stems from the possible variability in the expected returns, therefore, name the risk measure for this venture]

A good measure of risk in this context is the standard deviation of the returns, which quantifies the variability around the expected value. To compute this, we first find the varianceσ^2 and then take the square root to get the standard deviation σ

Variance $\sigma^2=\sum(X-E(X))^2\cdot P(X)$

Let's calculate these values.

Calculations

Expected Value (Mean)

$E(X)=(-2000\times 0.1)+(-1000\times 0.1)+(0\times 0.2)+(1000\times 0.2)+(2000\times 0.3)+(3000\times 0.1)=-200+(-100)+0+200+600+300=800$

The expected value (mean) of the returns is $800.

Variance and Standard Deviation

$\text{Variance}(\sigma^{2})=((-2000-800)^2\times 0.1)+((-1000-800)^2\times 0.1)+((0-800)^2\times 0.2)+((1000-800)^2\times 0.2)+((2000-800)^2\times 0.3)+((3000-800)^2\times 0.1)$

$\sigma^2=(7840000\times 0.1)+(3240000\times 0.1)+(640000\times 0.2)+(40000\times 0.2)+(1440000\times 0.3)+(4840000\times 0.1)=784000+324000+128000+8000+432000+484000=\sqrt{2160000}$

$\sigma=\sqrt{2160000}=1469.69$

Hints:

For each assignment, the solution should be submitted in the below format.

1. Research and Perform all possible steps for obtaining the solution.

2. For Statistics calculations, an explanation of the solutions should be documented in detail along with codes. Use the same word document to fill in your explanation.

Must follow these guidelines:

2.1 Be thorough with the concepts of Probability, Probability Distributions, Business Moments, and Univariate & Bivariate visualizations.

2.2 For True/False Questions, or short answer type questions explanation is a must.

2.3 Python code for Univariate Analysis (histogram, box plot, bar plots, etc.) the data distribution is to be attached.

3. All the codes (executable programs) should execute without errors

4. Code modularization should be followed

5. Each line of code should have comments explaining the logic and why you are using that function