**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: N.Harsha vardhan Batch ID: nharsha569103528**

**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:**

Table

Description automatically generated

**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

****

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

****

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.
2. What can we say about the student marks?
3. What can you say about the Excepted value for the student score?

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

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

1. Equal to 1
2. Less than or equal to 4
3. Sum is divisible by 2 and 3

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?

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 |

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.

A picture containing table

Description automatically generated

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

Q8) Calculate the Expected Value for the problem below.

1. 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?

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

**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% |

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]

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 |

1. What is the most likely monetary outcome of the business venture?

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

1. Is the venture likely to be successful? Explain.

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

1. 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]

1. 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]

**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:

* 1. Be thorough with the concepts of Probability, Probability Distributions, Business Moments, and Univariate & Bivariate visualizations.
  2. For True/False Questions, or short answer type questions explanation is a must.
  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

ANSWERS

'''

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

import pandas as pd

import numpy as np

car\_s\_d = {'speed':[4,4,7,7,8,9,10,10,10,11,11,12,12,12,12,13,13,13,13,14,14,14,14,15,15,15,16],

'distance':[2,10,4,22,16,10,18,26,34,17,28,14,20,24,28,26,34,34,46,26,36,60,80,20,26,54,32]}

car\_s\_d\_df = pd.DataFrame(car\_s\_d)

car\_s\_d\_df

#mean

car\_s\_d\_mean = car\_s\_d\_df.mean()

#standard deviation

car\_s\_d\_std = car\_s\_d\_df.std()

# Dropping the duplicates fetching the mean and standard deviation

car\_s\_d\_dup = car\_s\_d\_df.drop\_duplicates(subset=['speed'])

car\_s\_d\_dup

#Finding the mean of duplicate data

car\_s\_d\_dup.mean()

car\_s\_d\_dup.std()

#b) Top speed(SP) and weight(WT)

car\_sp\_wt = pd.read\_excel(r'F:\Liser Time\360digitmg\2.EDA\Data Sets\Book1.xls')

# mean

car\_sp\_wt.mean()

# standard deviation

car\_sp\_wt.std()

'''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.

2) What can we say about the student marks?

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

'''

scores = pd.DataFrame(columns=['scores'])

scores['scores'] = np.array([34, 36, 36, 38, 38, 39, 39, 40, 40, 41, 41, 41, 41, 42, 42, 45, 49, 56])

scores

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

#mean

scores.mean()

# median

scores.median()

# mode

scores.mode()

#variance

scores.var()

# standard deviation

scores.std()

# 2) What can we say about the student marks?

The mean is close to the median so it is the distribution is roughly symmetric

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

The expected value for the student score is dependent on the mean value

'''

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

'''

When tossing three coins, the possible outcomes can be represented as a set of eight outcomes

But we need the probability of of 2 heads and 1 tail so final probalibility is 3/8

'''

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

a) Equal to 1

b) Less than or equal to 4

c) Sum is divisible by 2 and 3

'''

#a) Equal to 1

There is only one combination where the sum is 1, which is (1, 1).

Probability P(Equal to 1) = 1/36

# b) Less than or equal to 4

The combinations for sums less than or equal to 4 are (1, 1), (1, 2), (2, 1), (1, 3), (2, 2), and (3, 1).

Probability P(Sum<=4) = 6 / 36 = 1/6

# c) Sum is divisible by 2 and 3

The sums divisible by 2 and 3 are 2, 4, 6, 8, 10, and 12. The corresponding combinations are

(1, 1), (1, 2), (2, 1), (1, 5), (2, 4), (3, 3), (4, 2), and (5, 1),(6,6)

Probability P(Sum Divisible by 2and 3) = 9/36 = ¼

'''

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?

'''

The number of ways to draw 2 balls without getting any blue balls is given by the combination of the non-blue balls:

Ways to draw 2 non-blue balls

is 5C2 = 10

Total number ways we can draw 2 balls

Is 7C2

import math

# Total number of balls

total\_balls = 2 + 3 + 2

# Total ways to draw 2 balls

total\_ways = math.comb(total\_balls, 2)

# Ways to draw 2 non-blue balls

non\_blue\_ways = math.comb(2 + 3, 2)

# Probability

probability\_none\_blue = non\_blue\_ways / total\_ways

'''

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

'''

A, B, C, D, E, F = 'A', 'B', 'C', 'D', 'E', 'F'

exp\_candies = pd.DataFrame(columns=['CHILD','Candies count','Probability'])

exp\_candies['CHILD'] ,exp\_candies['Candies count'], exp\_candies['Probability']= [A,B,C,D,E,F],[1,4,3,5,6,2],[0.015,0.20,0.65,0.005,0.01,0.12]

exp\_candies.mean()

child\_values = [1, 4, 3, 5, 6, 2]

probabilities = [0.015, 0.20, 0.65, 0.005, 0.01, 0.12]

# Calculate the expected value

expected\_value = sum(value \* prob for value, prob in zip(child\_values, probabilities))

'''

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 = [3.9,3.9,3.85,3.08,3.15,2.76,3.21,3.69,3.92,3.92,3.92,3.07,3.07,3.07,2.93,3,3.23,4.08,4.93,4.22,3.7,2.76,3.15,3.73,3.08]

score = [2.62,2.875,2.32,3.215,3.44,3.46,3.57,3.19,3.15,3.44,3.44,4.07,3.73,3.78,5.25,5.242,5.345,2.2,1.615,1.835,2.465,3.52,3.435,3.84,3.845]

weigh = [16.46,17.02,18.61,19.44,17.02,20.22,15.84,20,22.9,18.3,18.9,17.4,17.6,18,17.98,17.82,17.42,19.47,18.52,19.9,20.01,16.87,17.3,15.41,17.05]

p\_s\_w = {'points':points,'score':score,'weigh':weigh}

psw = pd.DataFrame(p\_s\_w)

psw.mean()

psw.median()

psw.mode()

psw.var()

psw.std()

range\_p = max(psw.points) - min(psw.points)

range\_s = max(psw.score) - min(psw.score)

range\_w = max(psw.weigh) - min(psw.weigh)

psw.points.skew()

psw.score.skew()

psw.weigh.skew()

'''

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?

'''

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

# Calculate the expected value

expected\_value = sum(weight / len(weights) for weight in weights)

# Display the result

print("Expected value of the weight of a randomly selected patient:", expected\_value)

'''

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%

'''

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Data

data = {

'Name of Company': ['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]

}

df = pd.DataFrame(data)

# Plot the distribution

plt.figure(figsize=(12, 6))

sns.boxplot(x='Measure X', data=df)

plt.title('Distribution of Measure X with Outliers')

plt.show()

# Identify and print outliers

Q1 = df['Measure X'].quantile(0.25)

Q3 = df['Measure X'].quantile(0.75)

IQR = Q3 - Q1

outliers = df[(df['Measure X'] < (Q1 - 1.5 \* IQR)) | (df['Measure X'] > (Q3 + 1.5 \* IQR))]

print("Outliers:")

print(outliers)

# Calculate mean, standard deviation, and variance

mean\_value = np.mean(df['Measure X'])

std\_deviation = np.std(df['Measure X'])

variance = np.var(df['Measure X'])

print("\nMeasures of Centrality and Spread:")

print("Mean (μ):", mean\_value)

print("Standard Deviation (σ):", std\_deviation)

print("Variance (σ^2):", variance)

'''

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.

'''

import numpy as np

# Probability of a misdirected call

probability\_misdirected = 1 / 200

# Number of simulations

num\_simulations = 10000

# Simulate misdirected calls

misdirected\_calls = np.random.choice([0, 1], size=num\_simulations, p=[1 - probability\_misdirected, probability\_misdirected])

# Calculate the simulated probability of a misdirected call

simulated\_probability = np.mean(misdirected\_calls)

print(f"Simulated Probability of a Misdirected Call: {simulated\_probability:.4f}")

'''

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]

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

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

(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]

(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]

'''

import numpy as np

# Given data

outcomes = [-2000, -1000, 0, 1000, 2000, 3000]

probabilities = [0.1, 0.1, 0.2, 0.2, 0.3, 0.1]

df\_out\_pro = pd.DataFrame(columns=['outcomes','probabilities'])

df\_out\_pro['outcomes'] = outcomes

df\_out\_pro['probabilities'] = probabilities+

df\_out\_pro.var()

df\_out\_pro.std()

# (i) Expected return (most likely monetary outcome)

expected\_return = np.sum(np.array(outcomes) \* np.array(probabilities))

expected\_return.mean()

print(f"(i) Most Likely Monetary Outcome (Expected Return): ${expected\_return:,.0f}")

# (ii) Likelihood of success

success\_probability = np.sum(np.array(probabilities)[np.array(outcomes) > 0])

print(f"(ii) Likelihood of Success: {success\_probability:.2%}")

# (iii) Long-term average earnings

long\_term\_average\_earnings = expected\_return

print(f"(iii) Long-Term Average Earnings: ${long\_term\_average\_earnings:,.0f}")

# (iv) Measure of risk (standard deviation)

variance = np.sum((np.array(outcomes) - expected\_return)\*\*2 \* np.array(probabilities))

standard\_deviation = np.sqrt(variance)

print(f"(iv) Measure of Risk (Standard Deviation): ${standard\_deviation:,.0f}")