**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: Vishvash C Batch ID: 23012024**

**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)

****

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

df = pd.read\_csv(r"C:\Users\Lenovo\Downloads\Cars.csv")

df.describe()

# Checking for normal distribution

normality\_tests = df.describe().loc[['mean', 'std']].T

normality\_tests['skew'] = df.skew()

normality\_tests['kurtosis'] = df.kurtosis()

normality\_tests['range'] = df.max() - df.min()

normality\_tests['var'] = df.var()

normality\_tests['mode'] = df.mode().T #*since the values are of type continuous, the mode is of no use*

pd.set\_option('display.max\_columns', 7)

print("Normality Tests:")

print(normality\_tests)

# Visualizing distribution using histograms

plt.figure(figsize=(10, 6))

for column in df.columns:

plt.subplot(2, 2, list(df.columns).index(column) + 1)

sns.histplot(df[column], kde=True)

plt.title(column)

plt.tight\_layout()

plt.show()

# Detecting outliers using boxplots

plt.figure(figsize=(10, 6))

for column in df.columns:

plt.subplot(2, 2, list(df.columns).index(column) + 1)

sns.boxplot(y=df[column])

plt.title(column)

plt.tight\_layout()

plt.show()

**Output:**

Normality Tests:

mean std skew kurtosis range var \

SPEED 34.422076 9.131445 -0.177947 -0.611679 41.599418 83.383283

DIST 98.765432 22.301497 -0.590197 0.920229 110.000000 497.356790

SP 121.540272 14.181432 1.611450 2.977329 70.033606 201.113002

WT 31.677984 8.320090 -0.587317 0.156570 37.284894 69.223898

mode

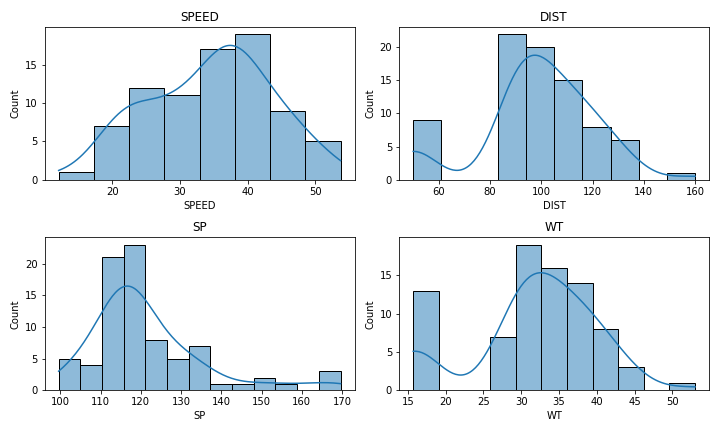
SPEED 29.629936

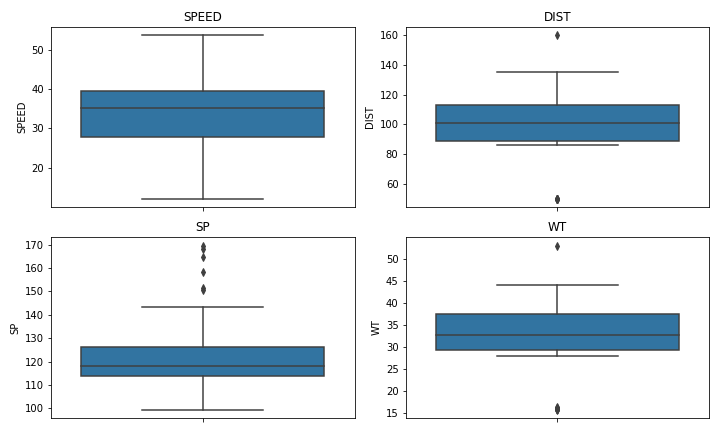
DIST 50.000000

SP 118.288996

WT 15.847758

#*since the values are of type continuous, the mode is of no use*





From the values of skew, kurtosis, outliers in graphical representation, we infer that SP is positively skewed with high number of outliers and other column variables or slightly negatively skewed. Whereas the column SPEED has negative kurtosis and other variables have positive kurtosis.

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: 25.529411764705884

Standard Deviation: 5.05266382858645

1. What can we say about the student marks?

The mean, median, and mode are all close together, indicating a relatively symmetric distribution of scores around the center

1. What can you say about the Excepted value for the student score?

The expected value for the student score is the mean value is 41.

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

P(Two heads and one tail)= Number of favorable outcomes/ Total number of outcomes = 3/8

​

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

1. Equal to 1 : Ans: 0
2. Less than or equal to 4 Ans: 6/36 = 1/6
3. Sum is divisible by 2 and 3 Ans: 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?

5/7 \* 4/6 = 10/21

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 |

To calculate the expected number of candies for a randomly selected child, we can use the formula for expected value:

E(X) = ∑ x P ( x ) =>

Expected number of candies = (1×0.015)+(4×0.20)+(3×0.65)+(5×0.005)+(6×0.01)+(2×0.12)=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.

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.

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

df = pd.read\_excel(r"C:\Users\Lenovo\Downloads\Cars.xlsx", sheet\_name = 'assignment')

df.describe()

# Checking for normal distribution

normality\_tests = df.describe().loc[['mean', 'std']].T

normality\_tests['skew'] = df.skew()

normality\_tests['kurtosis'] = df.kurtosis()

normality\_tests['range'] = df.max() - df.min()

normality\_tests['var'] = df.var()

#since the values are of type continuous, the mode is of no use

pd.set\_option('display.max\_columns', 7)

print("Normality Tests:")

print(normality\_tests)

# Visualizing distribution using histograms

plt.figure(figsize=(10, 6))

for column in df.columns:

plt.subplot(2, 2, list(df.columns).index(column) + 1)

sns.histplot(df[column], kde=True)

plt.title(column)

plt.tight\_layout()

plt.show()

# Detecting outliers using boxplots

plt.figure(figsize=(10, 6))

for column in df.columns:

plt.subplot(2, 2, list(df.columns).index(column) + 1)

sns.boxplot(y=df[column])

plt.title(column)

plt.tight\_layout()

plt.show()

**Output:**

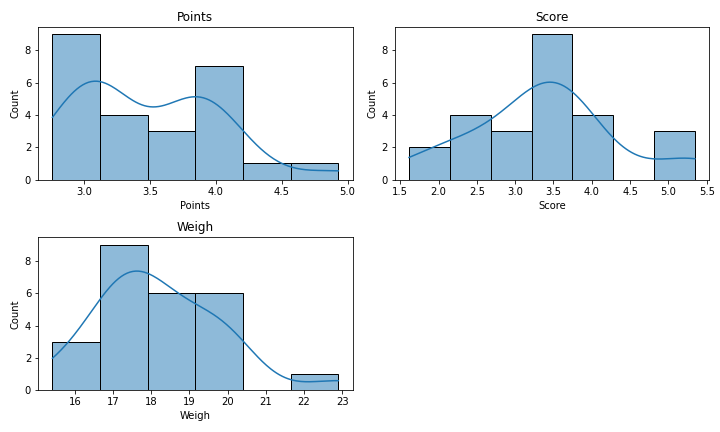
Normality Tests:

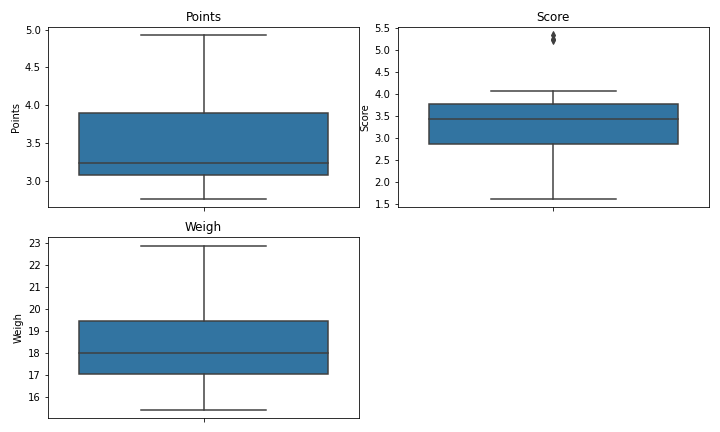
mean std skew kurtosis range var

Points 3.49280 0.539240 0.700816 0.238499 2.17 0.290779

Score 3.39568 0.954937 0.354650 0.367522 3.73 0.911904

Weigh 18.21840 1.638060 0.827334 1.376526 7.49 2.683239





"Weight" shows the highest variability among the variables, as indicated by its higher standard deviation, skewness, kurtosis, and variance.

All variables exhibit positive skewness, indicating a right-skewed distribution, with more data points towards higher values.

“Score” variable has outliers which is visible in the boxplot.

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?

The expected value of the patient weights is total sum of all patients divided by the number of patients(number of patients is 9).

Sum the weights of all patients: 108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199 = 1308.

E(X) = 1308 / 9 = 145.33.

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% |

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

df = pd.read\_excel(r"C:\Users\Lenovo\Downloads\Cars.xlsx", sheet\_name = 2)

df.describe()

# Convert percentages to numerical values

df["Measure X"] = df["Measure X"]\*100

# Plot the data

plt.figure(figsize=(10, 6))

sns.boxplot(y="Measure X", data=df)

plt.title("Boxplot of Measure X")

plt.ylabel("Measure X")

plt.show()

# Visualizing distribution using histograms

plt.figure(figsize=(10, 6))

sns.histplot(df["Measure X"], kde=True)

plt.title("Measure X")

plt.tight\_layout()

plt.show()

# Identify outliers using IQR method

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

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

IQR = Q3 - Q1

lower\_bound = Q1 - 1.5 \* IQR

upper\_bound = Q3 + 1.5 \* IQR

outliers = df[(df["Measure X"] < lower\_bound) | (df["Measure X"] > upper\_bound)]

print("Outliers:")

print(outliers)

# Calculate measures of centrality and spread

mean\_X = df["Measure X"].mean()

std\_X = df["Measure X"].std()

var\_X = df["Measure X"].var()

skew\_X = df["Measure X"].skew()

kurt\_X = df["Measure X"].kurt()

print("\nMean:", mean\_X)

print("Standard Deviation:", std\_X)

print("Variance:", var\_X)

print("Skewness:", skew\_X)

print("Kurtosis", kurt\_X)

**Output:**

Outliers:

Name of company Measure X

10 Morgan Stanley 91.36

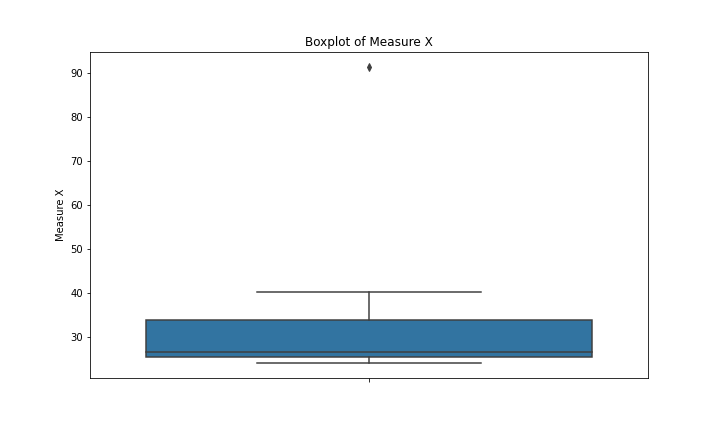
Mean: 33.27133333333333

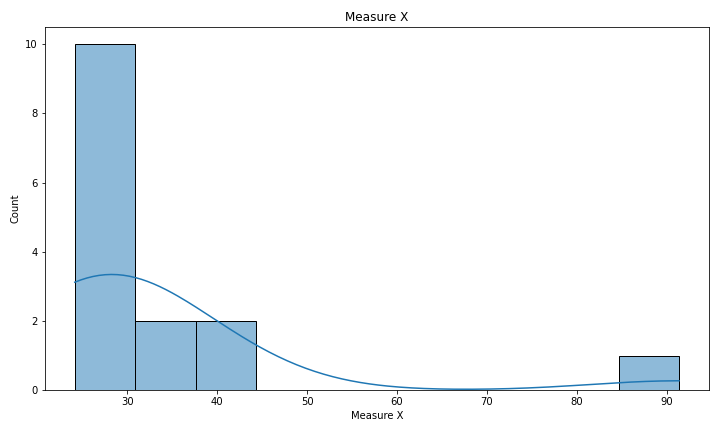
Standard Deviation: 16.945400921222028

Variance: 287.1466123809524

Skewness: 3.2551132228850337

Kurtosis 11.458222957026209





**Inference:**

The data of "Measure X" exhibits a highly skewed and leptokurtic distribution, with a large standard deviation suggesting considerable variability around the mean. The high kurtosis value indicates that extreme values are more likely to occur compared to a normal distribution

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]

# Probability of a single call being correct

p\_correct = 199 / 200

# Probability of all five calls being correct

p\_all\_correct = p\_correct \*\* 5

# Probability of at least one wrong call out of five

p\_at\_least\_one\_wrong = 1 - p\_all\_correct

print("Probability of at least one wrong call out of five:", p\_at\_least\_one\_wrong)

Output: Probability of at least one wrong call out of five: 0.025

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]

The outcome is most likely the expected returns of the venture is 2000 as it has the highest probability of 0.3

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

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

Yes the venture is likely to be successful as the probability of the venture to get a profit is 60 percent and to get loss is only 20 percent and remaining 20 percent the venture doesn’t gain or lose.

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]

The long-term average earning of business ventures of this kind is equal to the expected value, which is $800. This represents the average return the business can expect over the long term.

E(X)=(−2000×0.1)+(−1000×0.1)+(0×0.2)+(1000×0.2)+(2000×0.3)+(3000×0.1)=800

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]

A common measure of risk in financial analysis is the standard deviation. The standard deviation of the returns in this venture is 1469.69. This means that there is a significant spread of possible outcomes, and the actual return could be much higher or lower than the expected return.

# Given probability distribution

data = {

"X": [-2000, -1000, 0, 1000, 2000, 3000],

"P(x)": [0.1, 0.1, 0.2, 0.2, 0.3, 0.1]

}

# Create DataFrame

df = pd.DataFrame(data)

# (i) Most likely monetary outcome

most\_likely\_outcome = df.loc[df["P(x)"].idxmax(), "X"]

# (ii) Likelihood of success

success\_probability = df[df["X"] > 0]["P(x)"].sum()

# (iii) Long-term average earning

long\_term\_average\_earning = (df["X"] \* df["P(x)"]).sum()

# (iv) Measure of risk (standard deviation)

mean\_return = long\_term\_average\_earning

risk = np.sqrt(((df["X"] - mean\_return) \*\* 2 \* df["P(x)"]).sum())

**Output:**

(i) Most likely monetary outcome: 2000

(ii) Likelihood of success: 0.6

(iii) Long-term average earning: 800.0

(iv) Measure of risk (standard deviation): 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:

* 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