**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: \_\_\_\_\_\_\_\_\_\_\_\_\_ Batch ID:** \_\_\_\_\_\_\_\_\_\_\_

**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

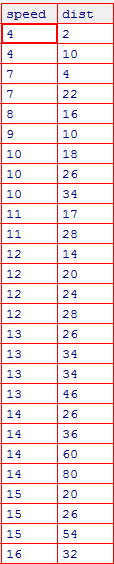
**Make a table as shown above and provide information about the features such as its data type and its relevance to the model building. And if not relevant, provide reasons and a description of the feature.**

**Problem Statements:**

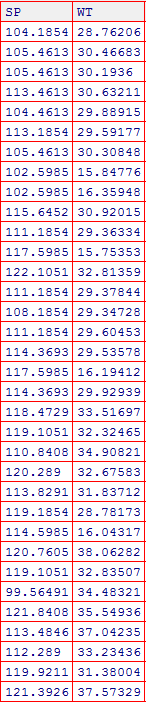
Q1) Calculate Mean, and Standard Deviation using Python code & draw inferences on the following data. Refer to the Datasets attachment for the data file.

**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

****

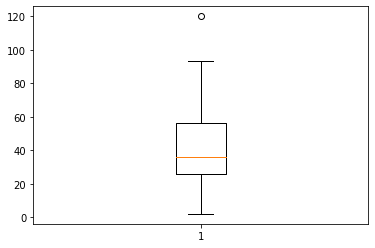
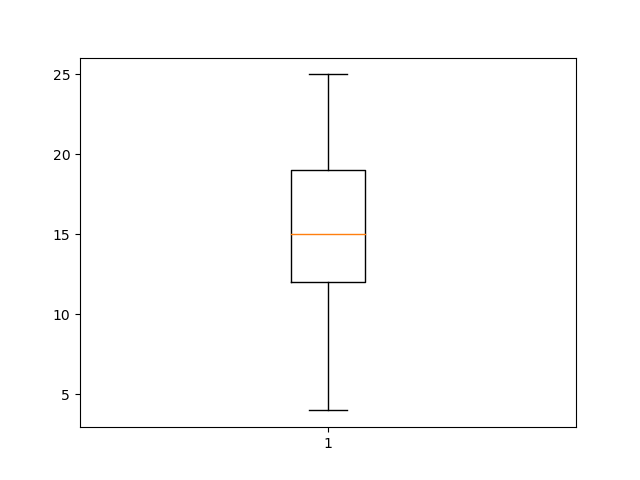
b. Top Speed (SP) and Weight (WT)

****

1a)

*Table 1 Speed and Distance Inferences*

|  | Speed | Distance |
| --- | --- | --- |
| Mean | 15.4 | 42.98 |
| Median | 15 | 36 |
| Mode | 20 | 26 |
| Variance | 27.96 | 664.06 |
| Standard Deviation | 5.29 | 25.77 |
| Range | 21 | 118 |
| Skewness | -0.1175 | -0.8068 |
| Distribution / Skewness | Close to symmetric/  slightly negative | Moderately Negative |
| Kurtosis | -0.5089 | 0.405 |
| Kurtosis(title) | Platykurtic | Mesokurtic |

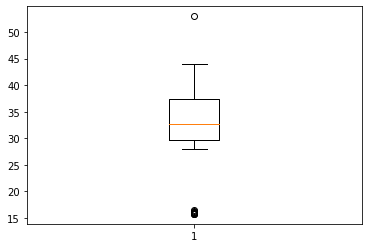
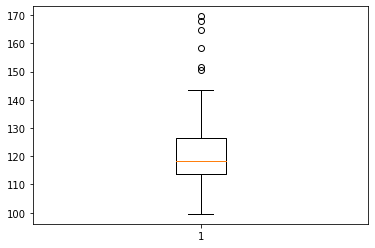


*Figure 1 Box Plots (a) Speed (b) Distance*

1b)

*Table 2 Top Speed and Weight Inferences*

|  | Top Speed | Weight |
| --- | --- | --- |
| Mean | 121.5 | 32.41 |
| Median | 118.208 | 32.734 |
| Mode | 118.289 | NA |
| Variance | 201.113 | 56.14 |
| Standard Deviation | 14.18 | 7.49 |
| Range | 70.03 | 37.28 |
| Skewness | 1.611 | -0.614 |
| Distribution / Skewness | Positive | Slightly Negative |
| Kurtosis | 2.977 | 0.95 |
| Kurtosis (title) | Leptokurtic | Platykurtic |



*Figure 2 Box Plots (a) Top Speed (b) Weight*

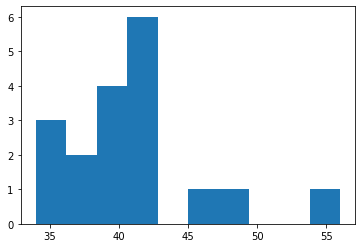
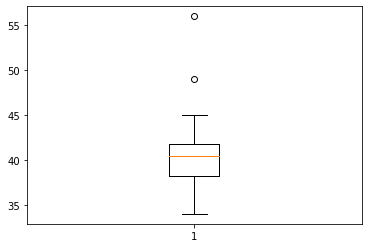
Q2) Below are the scores obtained by a student on tests.

**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?

**Ans1)** Mean = 41.0  
Median = 40.5  
Variance = 24.11  
Standard Deviation = 4.91

**Ans2)** Histogram & Box Plot for scores

   
The students might be graded out of 60.  
Data is negatively skewed because the median is towards the top of the boxplot.   
Outliers are present.

**Ans3)** Expected value of weight for a randomly selected patient = 40.59

Probability(of any student’s score being selected) = 1 / 18 = 0.055

µ = ∑X \* p(X) = 1.87 + 1.98 + 1.98 + 2.09 + 2.09 + 2.145 + 2.145 + 2.2 + 2.2 + 2.255 + 2.255 + 2.255 + 2.255 + 2.31 + 2.31 + 2.475 + 2.695 + 3.08 = 40.59

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

**Ans3)** Total outcomes = 2^3 = 8

Outcomes where (2T&1H) are possible – TTH , THT, HTT

Probability = 3/8 = 0.375

**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

**Ans4)** Total number of ways where the outcomes of 2 dice rolled can be sum = 6^2 = 36

a) Outcomes when sum is 0 = 0

Probability is 0/36 = 0

b) Outcomes when sum is <=4, are : (1,1), (1,1) ,(1,2) , (2,1) , (1,3) , (3,1) , (2,2) , (2,2) = 8

Probability is 8/36 = 0.223

c) Outcomes when sum is divisible by 2 & 3 : (1,5),(5,1),(2,4),(4,2),(3,3),(3,3) = 6

Probability is 6/36 = 0.1667

**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?

**Ans5)** 7 Balls, the number of outcomes is given by multiplying number of balls initially and the number after the first ball is removed.   
Total number of of outcomes = 7 x 6 = 42

Outcome of finding 2 blue balls = (B1 , B2) & (B2 , B1) = 2

Probability of 2 blue balls = 2/42 = 0.047

Probability that both balls are not blue = 1 – P( 2 blue balls) = 0.953

**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 |

**Ans6)** Expected value of candies for a randomly selected child = 3.09

µ = ∑X \* p(X) = 0.015+0.8+1.95+0.025+0.06+0.24 = 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.

**Ans7)** Data types = Float type for all columns ( Points, Score, Weight )

Points & Weigh columns are bimodal.

|  | Points | Score | Weigh |
| --- | --- | --- | --- |
| Mean | 3.596 | 3.211 | 17.84 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | 3.07 & 3.92 | 3.44 | 17.02 & 18.9 |
| Variance | 0.285 | 0.93 | 3.19 |
| Standard Deviation | 0.53 | 0.96 | 1.786 |
| Range | 2.17 | 3.832 | 8.399 |

**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?

**Ans8)** Expected value of weight for a randomly selected patient = 145.33

Probability(any patient being selected) = 1 / 9 = 0.11111

µ = ∑X \* p(X) = 12 + 12.22 + 13.667 + 14.88 + 15 + 16.11 + 18.55 + 20.77 + 22.11 = 145.33

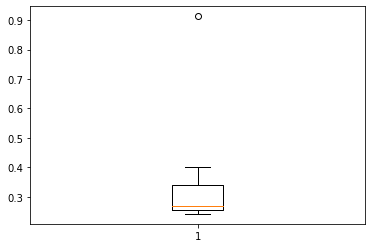
**Q9)** Look at the data given below. Plot the data, find the outliers, and find out:

**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% |

**Ans9)** ##q7['MeasureX'] = q7['MeasureX'].replace('%','')

Measure X is float Data type, Name of Company is nominal type.



*Figure 2 Box Plot for Measure X*

|  |  | Measure X |
| --- | --- | --- |
| Measures of Centrality | Mean, µ | 0.3327 |
| Median | 0.2671 |
| Mode | Multimodal data |
| Measures of Dispersion (spread) | Variance , σ2 | 0.028 |
| Standard Deviation, σ | 0.169 |
| Range | 0.672 |

**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]

**Ans10)** Probability (wrong call) = 1/200 = 0.005

Probability (wrong call in 5 attempted calls) = 5 \* Probability ( wrong call ) = 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]

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]

**Ans11.i)** Expected returns of the venture = 800$  
µ = ∑X \* p(X) = (-)200 + (-) 100 + 0 + 200 + 600 + 300 = 800

**Ans11.ii)** Probability ( Good returns) = 0.6   
Probability (Breakeven) = 0.2   
Probability (Influx of funds) = 0.2

This venture is likely to succeed , due to its low probability of influx in funds.

**Ans11.iii)** Long term average is the same as the expected returns. The probability value doesn’t change for more years .

**Ans11.iv)** Probability to Breakeven is a good measure of risk in this venture.

Influx of funds is the risk to run this venture , the negative returns require the venture to put money into the business when times are bad.

**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