**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: \_\_\_Seema Iyengar\_\_\_\_\_\_\_\_ Batch ID:** \_\_\_\_\_\_EDA\_06042023\_10AM\_\_\_\_\_

**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. (Keys 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. Cars speed and distance

****

**Answers:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name of the Feature** | **Description** | **Type** | **Relevance** | **Explanation** |
| speed | Speed of the Car | Ratio,Quantitative,Structured | Speed is a Relevant Column | Ratio Type because all statistical analysis can be performed Quantitative because it is a numerical data Structured because it can be represented in tabular format |
| dist | Distance travelled by the car | Ratio,Quantitative,Structured | Distance is a Relevant Column | Ratio Type because all statistical analysis can be performed Quantitative because it is a numerical data Structured because it can be represented in tabular format |

**First Moment Decision -> Measures of Central tendency**

**Below is the mean and Standard Deviation of the columns of the cars data set.**

1. **Mean -> cars.speed.mean()**

**Out[33]: 11.407407407407407**

**cars.dist.mean()**

**Out[34]: 27.666666666666668**

1. **Median -> cars.speed.median()**

**Out[36]: 12.0**

**cars.dist.median()**

**Out[37]: 26.0**

1. **Mode -> cars.speed.mode()**

**Out[38]:**

**0 12**

**1 13**

**2 14**

**Name: speed, dtype: int64**

**Second Moment Decision -> Measures of Dispersion**

1. **Variance -> cars.speed.var()**

**Out[40]: 10.404558404558403**

**cars.dist.var()**

**Out[41]: 291.15384615384613**

1. **Standard Deviation -> cars.speed.std()**

**Out[42]: 3.2256097725171906**

**cars.dist.std()**

**Out[43]: 17.063230824021755**

1. **Range -> range\_speed = max(cars.speed) - min(cars.speed)**

**range\_speed**

**Out[45]: 12**

**range\_dist = max(cars.dist) - min(cars.dist)**

**range\_dist**

**Out[47]: 78**

**Third Moment Business Decision -> Skewness**

1. **cars.speed.skew() # skew is negative(<0) hence negatively skewed**

**Out[48]: -0.895424912156393**

**b) cars.dist.skew() # skew is positive(>0) hence positively skewed**

**Out[49]: 1.29076266205913**

**Fourth Moment Business Decision -> Kurtosis**

**a) cars.speed.kurt() # kurt is less than 3 hence negative kurtosis**

**Out[50]: 0.24956097515622666**

**b) cars.dist.kurt() # kurt is less than 3 hence negative kurtosis**

**Out[51]: 2.464545503089613**

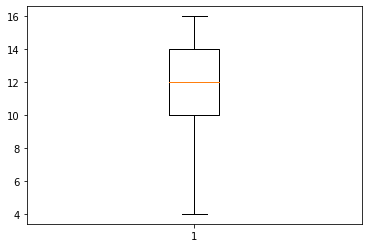
**Outliers ->**

**# To detect outliers using boxplot**

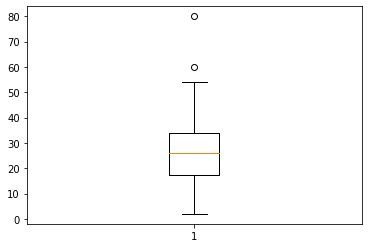
**import matplotlib.pyplot as plt**

**plt.figure()**

**plt.boxplot(cars.speed) # no outliers as per the boxplot visualization**



**plt.boxplot(cars.dist) # there are outliers visible as per boxplot visualization**



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

****

**Answers:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name of the Feature** | **Description** | **Type** | **Relevance** | **Explanation** |
| SP | Speed of the Car | Ratio,Quantitative,Structured | Speed is a Relevant Column | Ratio Type because all statistical analysis can be performed Quantitative because it is a numerical data Structured because it can be represented in tabular format |
| WT | Weight of the Car | Ratio,Quantitative,Structured | Distance is a Relevant Column | Ratio Type because all statistical analysis can be performed Quantitative because it is a numerical data Structured because it can be represented in tabular format |

**First Moment Decision -> Measures of Central tendency**

**Below is the mean and Standard Deviation of the columns of the cars data set.**

1. **Mean -> cars1.SP.mean()**

**Out[33]: 113.08768264705884**

**cars1.WT.mean()**

**Out[34]: 29.563105000000004**

1. **Median -> cars1.SP.median()**

**Out[36]: 113.65684999999999**

**cars1.WT.median()**

**Out[37]: 30.387655000000002**

1. **Mode -> cars1.SP.mode()**

**Out[38]:**

**0 105.4613**

**1 111.1854**

**Name: SP, dtype: float64**

**cars1.WT.mode()**

**Out[10]:**

**0 15.75353**

**1 15.84776**

**2 16.04917**

**3 16.19412**

**4 16.35948**

**5 28.76206**

**6 28.78173**

**7 29.34728**

**8 29.36334**

**9 29.37844**

**10 29.53578**

**11 29.59177**

**12 29.60453**

**13 29.88915**

**14 29.92939**

**15 30.19360**

**16 30.30848**

**17 30.46683**

**18 30.63211**

**19 30.92015**

**20 31.38004**

**21 31.83712**

**22 32.32465**

**23 32.67583**

**24 32.81359**

**25 32.83507**

**26 33.23436**

**27 33.51697**

**28 34.48321**

**29 34.90821**

**30 35.54936**

**31 37.04235**

**32 37.57329**

**33 38.06282**

**Name: WT, dtype: float64**

**Second Moment Decision -> Measures of Dispersion**

1. **Variance -> cars1.SP.var()**

**Out[40]: 40.85535668708672**

**cars1.WT.var()**

**Out[41]: 38.72751522120758**

1. **Standard Deviation -> cars1.SP.std()**

**Out[42]: 6.391819513024966**

**cars1.WT.std()**

**Out[43]: 6.223143516038143**

1. **Range -> range\_SP = max(cars1.SP) - min(cars1.SP)**

**range\_SP**

**Out[19]: 22.540189999999996**

**range\_WT = max(cars1.WT) - min(cars1.WT)**

**range\_WT**

**Out[21]: 22.309290000000004**

**Third Moment Business Decision -> Skewness**

1. **cars1.SP.skew() # skew is negative(<0) hence negatively skewed**

**Out[48]: -0.4267574826349179**

**b) cars1.WT.skew() # skew is negative(<0) hence negatively skewed**

**Out[49]: -1.3474621607317778**

**Fourth Moment Business Decision -> Kurtosis**

**a) cars1.SP.kurt() # kurt is less than 3 hence negative kurtosis**

**Out[50]: -0.8637279023486579**

**b) cars1.WT.kurt() # kurt is less than 3 hence negative kurtosis**

**Out[51]: 1.1528173104724373**

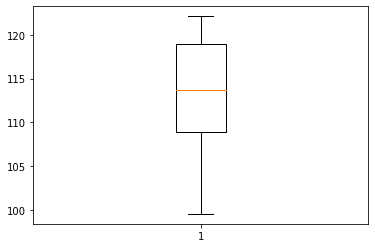
**Outliers ->**

**# To detect outliers using boxplot**

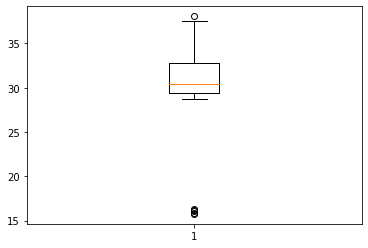
**import matplotlib.pyplot as plt**

**plt.figure()**

**plt.boxplot(cars.speed) # no outliers as per the boxplot visualization**



**plt.boxplot(cars1.WT) # there are outliers visible as per boxplot visualization**



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?

**Answers ->**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name of the Feature** | **Description** | **Type** | **Relevance** | **Explanation** |
| scores | Scores of the students | Ratio,Quantitative,Structured | scores is a Relevant Column | Ratio Type because all statistical analysis can be performed Quantitative because it is a numerical data Structured because it can be represented in tabular format |

**# Measures of Central Tendency / First moment business decision**

1. **stu\_scores.scores.mean()**

**Out[38]: 41.0**

1. **stu\_scores.scores.median()**

**Out[39]: 40.5**

1. **stu\_scores.scores.mode()**

**0 41**

**Name: scores, dtype: int64**

**# Measures of Dispersion / Second moment business decision**

1. **stu\_scores.scores.var()**

**Out[41]: 25.529411764705884**

1. **stu\_scores.scores.std()**

**Out[42]: 5.05266382858645**

**# Third Moment Business Decision / Skewness**

**stu\_scores.scores.skew() # skew is positive(>0) hence positively skewed**

**Out[44]: 1.686841191854795**

**# Fourth Moment Business Decision / Kurtosis**

**stu\_scores.scores.kurt() # kurt is greater than 3 hence positive kurtosis**

**Out[45]: 3.9532788974070376**

**# plotting the density/probability distribution plot to find insights on student marks**

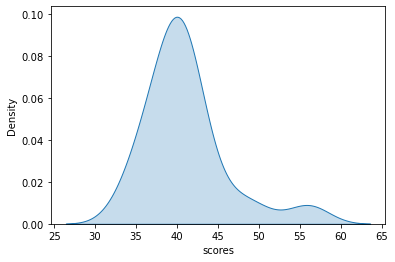
**import seaborn as sns**

**sns.kdeplot(stu\_scores.scores, bw = 0.5 , fill = True)**

**As per the below density plot, it can be inferred that most of the students have scored the marks between 35 and 45.**

**# Insights on Expected Value for the Student Score**

**As per the density plot the expected value is the mean of the student score which is 41.0.**



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

**Answers ->**

**3 coins -> Total Number of Probabilities -> {H,T,H,T,H,T} -> 6 -> total number of events**

**Probability that 2 heads and 1 tail -> 2H + 1T -> 3 - > number of interested events**

**As per probability formula -> Probability = number of interested events / total number of events**

**Therefore , 3/6 = ½ or 0.5**

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

**Answers ->**

**As per probability formula -> Probability = number of interested events / total number of events**

1. **Total Number of interested events = 12 (2 dice)**

**Number of interested events = equal to one = 1**

**Therefore, 1/12.(answer)**

1. **Total Number of interested events = 12 (2 dice)**

**Number of interested events = less than or equal to 4 = {4,3,2,1} = 4**

**Therefore, 4/12 = 1/3. (answer)**

1. **Total Number of interested events = 12 (2 dice)**

**Number of interested events = sum is divisible by 2 & 3 = {12,6} = 2**

**Therefore, 2/12 = 1/6.(answer)**

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?

**Answers ->**

**2R+2G+2B = Total 6 balls**

**Total Number of interested events = {RR,GG,BB,RG,RB,GB} = 6**

**Number of interested events = 2 Random Balls , none of them are blue = {RR,GG,RG} = 3**

**Therefore, 3/6 = ½ or 0.5 (answer)**

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 picture containing shape, arrow

Description automatically generateds**

**Answers ->**

**Expected number of candies - > sum of product of candies count with probability.**

**= 1\*0.025 + 4\*0.20+3\*0.65+5\*0.005+6\*0.01+2\*0.12**

**= 3.1000000000000005.**

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.

**Answers ->**

**Measures of Central Tendency / First moment business decision**

1. **sports.Points.mean()**

**Out[5]: 3.4928000000000003**

1. **sports.Score.mean()**

**Out[6]: 3.39568**

1. **sports.Weight.mean()**

**Out[7]: 18.2222**

1. **sports.Points.median()**

**Out[8]: 3.23**

1. **sports.Score.median()**

**Out[9]: 3.44**

1. **sports.Weight.median()**

**Out[10]: 17.98**

1. **sports.Points.mode()**

**Out[11]:**

**0 3.07**

**1 3.92**

**Name: Points, dtype: float64**

1. **sports.Score.mode()**

**Out[12]:**

**0 3.44**

**Name: Score, dtype: float64**

1. **sports.Weight.mode()**

**Out[13]:**

**0 17.02**

**Name: Weight, dtype: float64**

**Measures of Dispersion / Second moment business decision**

1. **sports.Points.var()**

**Out[14]: 0.29077933333333333**

1. **sports.Score.var()**

**Out[15]: 0.9119040599999998**

1. **sports.Weight.var()**

**Out[16]: 2.6803189999999986**

1. **sports.Points.std()**

**Out[17]: 0.5392395880620536**

1. **sports.Score.std()**

**Out[18]: 0.9549366785290005**

1. **sports.Weight.std()**

**Out[19]: 1.6371679816072626**

1. **range\_Points = max(sports.Points) - min(sports.Points)**

**range\_Points**

**Out[21]: 2.17**

1. **range\_Score= max(sports.Score) - min(sports.Score)**

**range\_Score**

**Out[23]: 3.7299999999999995**

1. **range\_Weight = max(sports.Weight) - min(sports.Weight)**

**range\_Weight**

**Out[25]: 7.489999999999998**

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?

**Answers ->**

**Expected value is said to be the mean of the given values.**

**Mean = Sum of all values / total number of values**

**That is mean of all weights(X) of patients -> (108+110+123+134+135+145+167+187+199) /9**

**Mean = 145.33333333333334**

**Therefore, expected values of one of the patients chosen at random is 145.33333333333334.**

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]

**A picture containing shape, arrow

Description automatically generated**

|  |  |
| --- | --- |
| **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% |
|  |  |

**Answers ->**

**import pandas as pd**

**import numpy as np**

**import matplotlib.pyplot as plt**

**company = pd.read\_csv(r"C:\Users\seema\OneDrive\Desktop\COURSE\360DigiTMG Course\EDA\_06042023\_10AM\DataSets\company.csv")**

**company.info()**

**# as per the above information Measure column is an object type hence converting the**

**# same to float in order to calculate mean**

**company.Measure = company.Measure.str.replace('%', '').astype(float)**

**# Measures of Central Tendency / First moment business decision**

**company.Measure.mean()**

**Out[9]: 33.27133333333333**

**# Measures of Dispersion / Second moment business decision**

**company.Measure.var()**

**Out[10]: 287.1466123809524**

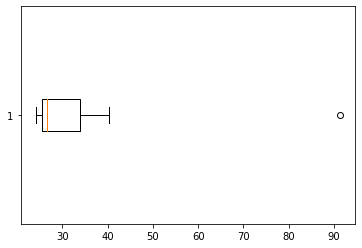
**company.Measure.std()**

**Out[11]: 16.945400921222028**

**# Box plot to show outliers,distribution of numerical data and skewness**

**plt.figure()**

**plt.boxplot(company.Measure,vert=False)**



**#Insights ->**

**As per the box plot analysis, There are outliers in the Measure\_X data of the company dataset, The skewness is positive and hence data is positively distributed.**

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 **A picture containing shape, arrow

Description automatically generated**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]

**Answers ->**

**Probability -> Number of interested events / Total Number of interested events**

**Number of interested events = 1 (one in five attempted calls reaches wrong number)**

**Total Number of interested events = 5 attempted telephone calls**

**P = 1/5 (answer).**

**Therefore, probability of at least one in five attempted telephone calls reaches the wrong number is 1/5 or 0.2.**

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:

2.1. Be thorough with the concepts of Probability, Central Limit Theorem and Perform the

calculation stepwise

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

2.3. R & Python code for Univariate Analysis (histogram, box plot, bar plots, etc.) the data

distribution 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