**2a. Data Visualization**

**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: E PAVAN KUMAR Batch ID:** 04/08/2023

**Topic: Data Visualization**

**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 Skewness, and Kurtosis 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

****

**ANSWER;**

#first moment decision

df.speed.mean()

df.speed.median()

df.speed.mode()

# second moment decision

df.speed.std()

df.speed.var()

range = (max(df.speed)-min(df.speed))

#3rd moment decision

df.speed.kurt()

df.speed.skew()

|  |  |  |
| --- | --- | --- |
|  | 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 |

my insights from the provided data for both Speed and Distance:

Speed:

- The mean speed is approximately 15.4, indicating a moderate average speed.

- The median speed (15) is close to the mean, suggesting a relatively symmetric distribution.

- The mode (20) indicates a peak at that speed.

- The variance for speed (27.96) is moderate, indicating some spread of data points from the mean.

- The standard deviation for speed (5.29) quantifies the dispersion of data around the mean.

- The range for speed (21) shows the spread between the minimum and maximum values.

- The skewness for speed is slightly negative (-0.1175), indicating a slightly left-skewed distribution.

- The kurtosis for speed (-0.5089) suggests a platykurtic distribution, which means it has lighter tails compared to a normal distribution.

Distance:

- The mean distance is approximately 42.98, indicating a moderate average distance.

- The median distance (36) is slightly lower than the mean, suggesting a slightly left-skewed distribution.

- The mode (26) indicates a peak at that distance.

- The variance for distance (664.06) is relatively high, indicating a significant spread of data points from the mean.

- The standard deviation for distance (25.77) quantifies the dispersion of data around the mean.

- The range for distance (118) is wide, showing a substantial spread between the minimum and maximum values.

- The skewness for distance is moderately negative (-0.8068), indicating a moderately left-skewed distribution.

- The kurtosis for distance (0.405) suggests a mesokurtic distribution, which means it is relatively close to a normal distribution in terms of tail behavior.

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

****

|  |  |  |
| --- | --- | --- |
|  | 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 |

MY insights from this data;

Top Speed:

- The mean top speed is approximately 121.5, indicating a moderate average speed.

- The median (118.208) is slightly lower than the mean, suggesting a slightly right-skewed distribution.

Weight:

- The mean weight is approximately 32.41, indicating a moderate average weight.

- The median (32.734) is slightly higher than the mean, suggesting a slightly left-skewed distribution.

Variance and Standard Deviation:

- The variance for top speed (201.113) is relatively high, indicating a significant spread of data points from the mean.

- The variance for weight (56.14) is lower, indicating less spread from the mean.

- The standard deviation for top speed (14.18) and weight (7.49) quantifies the dispersion of data around their respective means.

Range:

- The range for top speed (70.03) is relatively wide, suggesting a wide spread between the minimum and maximum values.

- The range for weight (37.28) is also notable but somewhat narrower than the top speed.

Skewness:

- The skewness for top speed is positive (1.611), indicating a right-skewed distribution with a longer tail on the right.

- The skewness for weight is slightly negative (-0.614), suggesting a slightly left-skewed distribution with a longer tail on the left.

Kurtosis:

- The kurtosis for top speed (2.977) indicates a leptokurtic distribution, which means it has heavier tails and more outliers compared to a normal distribution.

- The kurtosis for weight (0.95) suggests a platykurtic distribution, which means it has lighter tails and fewer outliers compared to a normal distribution.

**Q2) Draw inferences about the following boxplot & histogram.**

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





My insights for the above data;

Histogram:

- The histogram displays the distribution of chick weight data with bars representing frequency intervals.

- The data appears to be right-skewed because there are more data points with lower chick weights (left side of the histogram) and fewer data points with higher chick weights (right side).

- The histogram bars have the following frequencies for each interval: 80, 200, 120, 100, 50, 30, 20, and 10.

- The highest frequency interval is 200, which suggests that many chicks fall within that weight range.

- There is a noticeable drop in frequency as chick weight increases, indicating that fewer chicks have higher weights.

Boxplot:

- The boxplot provides a summary of the chick weight data's central tendency, spread, and presence of outliers.

- The boxplot shows that the data has a right-skewed distribution, consistent with the histogram.

- The boxplot's median line is likely closer to the lower end of the box, indicating that the median chick weight is lower.

- The box's length represents the interquartile range (IQR), which is relatively small, suggesting that most data points are concentrated within a narrow weight range.

- There are potential outliers on the right side of the boxplot, represented by individual data points above the upper whisker. These outliers indicate some chicks with significantly higher weights compared to the majority.

Inferences:

- The chick weight data is not normally distributed; it is right-skewed, with a tail on the right side.

- The majority of chicks have lower weights, as indicated by the high frequency in the lower weight intervals.

- The presence of outliers suggests that there are some chicks with unusually high weights compared to the rest of the population.

- The median chick weight is likely lower due to the skewness of the data.

Overall, the chick weight data exhibits a right-skewed distribution with potential outliers on the higher end of weights. The median weight is lower than the mean weight due to the skewness of the data.

**Q3) 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, variance, and standard deviation.
2. What can we say about the student marks? [**Hint**: Looking at the various measures calculated above whether the data is normal/skewed or if outliers are present].

Test Scores IN ASCENDING ORDER: 34, 36, 36, 38, 38, 39, 39, 40, 40, 41, 41, 41, 41, 42, 42, 45, 49, 56

1) Mean (Average):

Mean = (Sum of all scores) / (Total number of scores)

Mean = (34 + 36 + 36 + 38 + 38 + 39 + 39 + 40 + 40 + 41 + 41 + 41 + 41 + 42 + 42 + 45 + 49 + 56) / 18

Mean ≈ 41.11

2) Median (Middle Value):

Since there are 18 scores, the median is the average of the 9th and 10th values (when arranged in ascending order).

Median = (40 + 41) / 2

Median = 40.5

3) Variance (Measure of Spread):

1. Step 1: Calculate the Mean (Average)

- First, calculate the mean (μ) of the dataset.

- Mean (μ) = (Sum of all data points) / (Total number of data points)

- For our dataset:

Mean = (34 + 36 + 36 + 38 + 38 + 39 + 39 + 40 + 40 + 41 + 41 + 41 + 41 + 42 + 42 + 45 + 49 + 56) / 18

Mean ≈ 41.11

2. Step 2: Find the Deviation from the Mean

- Subtract the mean (μ) from each data point (xi) to find the deviation from the mean.

- Deviation (xi - μ)

3. Step 3: Square the Deviations

- Square each deviation (xi - μ). This ensures that negative and positive deviations don't cancel each other out.

- Squared Deviation (xi - μ)²

4. Step 4: Calculate the Sum of Squared Deviations

- Sum all the squared deviations obtained in step 3.

- Σ(xi - μ)² (Sum of squared deviations)

5. Step 5: Divide by the Number of Data Points

- Divide the sum of squared deviations (step 4) by the total number of data points (N).

- Variance = [Σ(xi - μ)²] / N

Using this formula, we will obtain the variance. In the case of our dataset:

Variance ≈ [Σ(xi - 41.11)²] / 18

Variance ≈ [Σ(34-41.11)² + (36-41.11)² + ... + (56-41.11)²] / 18

After calculating this, WE will find that the variance is approximately 15.53.

The variance is a measure of how data points in the dataset are dispersed or spread out from the mean. A higher variance indicates more spread, while a lower variance indicates data points are closer to the mean.

In this dataset, a variance of approximately 15.53 suggests that the scores have some degree of spread from the mean score of approximately 41.11.

4) Standard Deviation (Measure of Dispersion):

Standard Deviation is the square root of the variance.

Standard Deviation ≈ √15.53 ≈ 3.95 (rounded to two decimal places)

**4B)ANSWER**

- The mean and median scores are close to each other, suggesting that the data is relatively symmetric or normally distributed. There is no strong skewness observed.

- The variance and standard deviation values are not extremely high, indicating that the scores do not have large variations.

- There are no extreme outliers in the data, as all the scores fall within a reasonable range given the context of test scores.

In summary, the student's marks appear to be reasonably normally distributed with no significant outliers. The scores exhibit moderate variability around the mean, but there is no strong evidence of skewness or extreme deviations from the central tendency.

**Q5) What is the nature of skewness when the mean and median of data are equal?**

ANS: When the mean and median of a dataset are equal, it suggests a nearly symmetric distribution.

- A perfectly symmetric distribution would have the mean and median exactly equal, but real-world data often exhibits some degree of deviation from perfect symmetry.

**Q6) What is the nature of skewness when mean > median?**

ANS : - The data has a positive skew or is right-skewed.

- The tail of the distribution extends more towards the higher values.

- It suggests the presence of outliers or extreme values on the right side of the distribution.

- The majority of the data points are concentrated on the left side of the distribution, where the median is located.

**Q7) What is the nature of skewness when median > mean?**

ANS: - The data has a negative skew or is left-skewed.

- The tail of the distribution extends more towards the lower values.

- It suggests the presence of outliers or extreme values on the left side of the distribution.

- The majority of the data points are concentrated on the right side of the distribution, where the median is located.

**Q8) What does a positive kurtosis value indicate for data?**

ANS: - A positive kurtosis value indicates that the data has heavy tails or is leptokurtic.

- It means that the data has more extreme values (outliers) than would be expected in a normal distribution.

- The peak of the distribution is sharper (more peaked) than that of a normal distribution.

- Positive kurtosis indicates that the data has a higher concentration of values around the mean and heavier tails with more extreme values in the distribution.

**Q9) What does a negative kurtosis value indicate for data?**

ANS: - A negative kurtosis value indicates that the data has light tails or is platykurtic.

- It means that the data has fewer extreme values (outliers) than would be expected in a normal distribution.

- The peak of the distribution is flatter than that of a normal distribution.

- Negative kurtosis indicates that the data has a more dispersed or flattened distribution with fewer extreme values compared to a normal distribution.

**Q10) Answer the below questions using the below boxplot visualization.**



1. **What can we say about the distribution of the data?**

ANS: The majority of the data points are concentrated in the lower part of the range, from 0 to 15, where Q1 and Q2 are located.

- There are some data points beyond the upper quartile (Q3), which are considered outliers or extreme values.

- The range of the data extends from 0 to 20, with the bulk of the values clustered in the lower half of this range.

In summary, the box plot suggests that the data is right-skewed, with the majority of values concentrated in the lower part of the range and some outliers present beyond the upper quartile.

1. **What is the nature of the skewness of the data?**

ANS: The data is positively skewed or right-skewed. This is because the median (Q2) is closer to the lower quartile (Q1) than to the upper quartile (Q3).

**C)What will be the IQR of the data (approximately)?**

ANS: The interquartile range (IQR), which represents the spread of the central 50% of the data, is relatively narrow, indicating that most of the data points are close to the median.

**Q11) Comment on the below Boxplot visualizations.**



**Draw an Inference from the distribution of data for Boxplot 1 with respect to Boxplot 2.**

**Hint: [On comparing both the plots and check if the data is normally distributed/not, outliers present, skewness, etc.]**

**COMMON INFERENCES**

From the information provided about the two boxplot visualizations (Boxplot 1 and Boxplot 2), we can draw some inferences regarding the distribution of data and the differences between the two plots:

1. Q2 (Median) : Both Boxplot 1 and Boxplot 2 have the same median (Q2), which is 262.5. This suggests that the central tendency of the data is similar in both cases.

2. Interquartile Range (IQR) : Boxplot 1 has a narrower interquartile range (IQR) compared to Boxplot 2. The IQR for Boxplot 1 is from 260 (Q1) to 278 (Q3), while the IQR for Boxplot 2 is from 225 (Q1) to 305 (Q3). This indicates that the data in Boxplot 1 is more concentrated around the median compared to Boxplot 2.

3. Total Range : Boxplot 1 has a smaller total range (from 240 to 280) compared to Boxplot 2, which has a wider total range (from 200 to 350). This suggests that the data in Boxplot 1 is less spread out compared to Boxplot 2.

4. Outliers : According to the information provided, neither Boxplot 1 nor Boxplot 2 has any outliers. This indicates that both datasets do not contain extreme values that significantly deviate from the overall distribution.

5. Skewness : To assess skewness (whether the data is normally distributed or not), we would need additional information or a visualization of the data, such as a histogram or a density plot. The presence of skewness would suggest that the data is not symmetrically distributed.

In summary, based on the provided information:

- Boxplot 1 has data that is more tightly clustered around the median with a narrower range.

- Boxplot 2 has data that is more spread out with a wider range.

- Both boxplots have the same median and no outliers.

Q12)



**Answer the following three questions based on the boxplot above.**

1. **What is inter-quartile range of this dataset? [Hint: IQR = Q3 – Q1]**

**In one line, explain what this value implies. (Hint: Based on IQR definition)**

1. **What can we say about the skewness of this dataset?**
2. **If it were found that the data point with the value 25 is 2.5, how would the new boxplot be affected?**

**(Hint: On changing the data point from 25 to 2.5 in the data, how is it different from the current one.)**

ANSWER;

1. Q1 = 5 , q3 = 12 and iqr = 7 # iql depicts the maximum range of data presence in this range
2. The data is right skewed. The outliers present in above 0 hence it is positively skewed data. The outlier values are high
3. New box plot will be effected and the range of data will be decreased the median did not get effected but the data will be norrow down

**Q13)**

****

**Answer the following three questions based on the histogram above.**

1. **Where would the mode of this dataset lie? Hint: [In terms of values On the Y-axis]**
2. **Comment on the skewness of the dataset**
3. **Suppose that the above histogram and the boxplot in question 2 are plotted for the same dataset. Explain how these graphs complement each other in providing information about any dataset. Hint: [Visualizing both the plots, draw the insights]**
4. The mode of this dataset lies at the bin with the highest frequency on the Y-axis, which is 20
5. The dataset exhibits a slight negative skewness because the mode is at a lower value on the Y-axis, and the frequencies decrease as we move away from the mode. However, the skewness is not very pronounced.
6. Histograms and boxplots complement each other in providing insights about a dataset. Histograms visually display the distribution of data in detail, showing frequency within specific intervals. They reveal the shape and spread of data. Boxplots, on the other hand, summarize key statistics like median, quartiles, and outliers. Together, these plots offer a comprehensive view of the dataset, allowing you to assess its overall pattern, central tendency, variability, and the presence of extreme values.