**2a. Data Visualization**

**Name: Joo Zhin Mei**\_\_\_\_\_\_\_\_\_\_\_ **Batch ID: 050823**\_\_\_\_\_\_\_\_\_\_\_

1. **Data Understanding:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Name of Feature** | **Description** | **Type** | **Relevance** |
| 1 | Index | Index of the entry | Quantitative, Nominal | Irrelevant. Index is for entry organization & positioning, not for providing useful |
|  |  |  |  | information involving in model building. |
| 2 | speed | Car speed | Quantitative, Ratio | Relevant |
| 3 | dist | Distance travelled by car | Quantitative, Ratio | Relevant |
| 4 | SP | Top speed | Quantitative, Ratio | Relevant |
| 5 | WT | Weight | Quantitative, Ratio | Relevant |

**Q1) a. Cars speed and distance –** please refer to Python code file.

Skewness of cars speed= -0.11750986144663393

Kurtosis of cars speed= -0.5089944204057617

Inferences:

1. Skewness of an approximately -0.12 (which is between -0.5 & 0.5) means cars speed data is nearly normal distributed with slightly skewness to the left. Cars speeds are centered around the mean, with slightly more data points are smaller than the mean.
2. Kurtosis of approximately -0.51 (which is <3.0) means a negative Kurtosis (Platykurtic) which has light tails (fewer outliers). This suggest cars speeds data has fewer extreme values, and is spreading out more than a normal distribution.

Skewness of cars distance= 0.8068949601674215

Kurtosis of cars distance= 0.4050525816795765

Inferences:

1. Skewness of an approximately 0.81 (which is between 0.5 & 1.0) means cars distance data is moderately skewed to the right. Majority of the cars distance data points are more towards the lower range readings, with few larger cars distance reading.
2. Kurtosis of an approximately 0.41 (which is <3.0) means a Platykurtic distribution which has light tails (fewer outliers). This suggest cars distance data has fewer extreme values, and is spreading out more than a normal distribution. However, there are more extreme values in cars distance (potentially more outliers), plus the data is more concentrated around the mean, indicating heavier tails and a more peaked distribution as it has larger kurtosis.

**Q1) b. Top Speed (SP) and Weight (WT)-** please refer to Python code file.

Skewness of Top Speed (SP)= 1.6114501961773586

Kurtosis of Top Speed (SP)= 2.9773289437871835

Inferences:

1. Skewness of an approximately 1.61 (which is >1.0) means top speed data is heavily skewed to the right. Top speeds data points are heavily concentrated at the lower range readings, with few larger top speed readings at the higher range.
2. Kurtosis of an approximately 2.98 (which is very close to 3.0) means a slightly Platykurtic distribution in top speeds, which is close to the normal distribution, but with slightly lighter tails and is slightly less peaked than a normal distribution.

Skewness of Weight (WT)= -0.6147533255357768

Kurtosis of Weight (WT)= 0.9502914910300326

Inferences:

1. Skewness of an approximately -0.61 (which between -1.0 and -0.5) means weight data is moderately skewed to the left. Weight data points are more concentrated at the higher range readings, with few smaller weight readings at the lower end range.
2. Kurtosis of an approximately 0.95 (which is <3.0) Implies a Platykurtic distribution in weight, which is more spread out, less peaked and lighter tails than normal distribution. However, as weight kurtosis is smaller than top speed kurtosis, weight values are less extreme than top speed values. Additionally, weight values are less concentrated around their mean compared to top speeds.

**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]

Histogram of ChickWeight$weight:

1. Chick weights are positively skewed (right skewed) heavily. Most of the chicks are light in weight, with only a few of heavier chicks.
2. Majority of chicks are in the weight of 50-100, which represent the mode.
3. The average weight of the chicks is heavier than the majority weight of the chicks.

Boxplot:

1. There are 8 extremely large reading outliers in the dataset, which falls outside of the maximum boxplot range.
2. The dataset is positively skewed (right skewed) heavily. The middle 50% of the data points are concentrated at the smaller reading range side (left side), with some recorded larger data points, and 8 extremely large data points, the outliers.
3. The mean is much higher than the median, as it is pulled to the higher value range by both large and extreme large data points in the dataset.

**Q3)**

**1)** Please refer to Python code file.

**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].**

The student marks have a mean of 41.0, median of 40.5, standard deviation of 5.05266382858645, and variance of 25.529411764705884. As mean and median are almost the same which is 41.0, and standard deviation is relatively low around 5.05 which implies low degree of marks variability from the average mark of 41.0, the student marks are potentially normally distributed (still need to be confirmed by graph plotting). To determine the presence of outlier however, the interquartile range (IQR) method is needed.

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

It is potentially normally distributed (zero of skewness). The data does not skew to both left and right, all data points are equally distributed across left and right range, resulting a normal bell-shaped curve.

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

It is right skewed (positively skewed). Most of the data points are concentrated at the left (the range with smaller readings), while few larger or extremely large values at the right, resulting a long right tail. The mean is pulled to the right tail (> median) by large and extremely large values in the dataset.

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

It is left skewed (negatively skewed). Most of the data points are concentrated at the right (the range with larger readings), while fewer, but smaller or extremely small values at the left, resulting a long left tail. The mean is pulled to the left tail (< median) by small and extremely small values in the dataset.

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

1. The data has more extremely values, potentially more outliers (heavier tails).
2. The data is more concentrated around the mean, giving a sharp peak (leptokurtic).
3. The data is more concentrated around the mean, and more peaked than a normal distribution.

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

1. The data has lesser extremely values, potentially lesser outliers (lighter tails).
2. The data is less concentrated around the mean, giving a wide peak (platykurtic).
3. The data is less concentrated around the mean, and less peaked than a normal distribution.

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

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

1. It has a relatively large median (Q2), which means the middle value of the data, when sorted ascendingly, is relatively large (between 15 and 16).
2. The middle 50% of the data (Interquartile Range, IQR) are distributed at the larger side of the scale, which means the central portion of the data is lying approximately in between 10 and 18.
3. There is no any outlier observed, neither below the minimum range nor above the maximum range.

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

It is left skewed (negatively skewed). Most of the data points are concentrated at the larger scale end, while fewer but smaller data points are distributed at smaller scale end, giving a long tail at the left.

**What will be the IQR of the data (approximately)? -** please refer to Python code file.

**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.]**

1. Both Boxplot 1 & Boxplot 2 are normally distributed, as both boxplots are having an IQR (Interquartile Range) which are distributed right at the middle of their whiskers, with median (Q2) at the center of their IQR.
2. Both Boxplot 1 & Boxplot 2 have no outliers observed, neither below their minimum range nor above their maximum range.
3. Both Boxplot 1 & Boxplot 2 are in zero skewness (no skew), which means data are symmetrically distributed at their lower and upper range respectively.

**Q12)**

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

Please refer to Python code file.

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

The range between the third quartile (Q3) and the first quartile (Q1) in the dataset is 7.5 units (IQR = Q3 - Q1).

1. **What can we say about the skewness of this dataset?**

It is moderately right skewed (positively skewed). The IQR of this data set is slightly concentrated towards the lower range of the reading scale, with median(Q2) shifted towards the lower range reading scale as well. This means the data set is distributed towards the lower value range, giving a right tail.

1. **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.)**

There will be no more outlier observed above the maximum range, nor below the minimum range, as the 2.5 is now distributed in between the range of minimum value and the first quartile.

**Q13)**

1. **Where would the mode of this dataset lie? Hint: [In terms of values On the Y-axis]**

Since mode refers to the data category which is having the highest counts or frequency, the mode of this dataset lies in between the Value Y of 5 and 7.5.

1. **Comment on the skewness of the dataset**

It is right skewed (positively skewed). From the histogram, majority of the data points are concentrated at the lower range value, with fewer but larger data points and extremely large data points (outliers) at the higher right range, resulting a long right tail.

1. **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]**
2. Boxplot gives a clearer insight for outliers, including position and number of counts of outliers, whereas histogram is providing a general idea about outlier existence.
3. Histogram is capable in demonstrating central tendency and data skewness, whereas boxplot is providing better understanding in data distribution.