**2a. Exploratory Data Analysis (Answers Documentation)**

**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 |
| 6 | Name of company | Company names | Qualitative, Nominal | Relevant |
| 7 | Measure X | The percentage of X | Quantitative, Ratio | Relevant |

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

Mean of cars speed= 15.4

Standard deviation of cars speed= 5.2876444352347844

Inferences:

1. Average cars speed is around 15.4
2. Cars speed varies by around 5.2876444352347844 on average from the mean speed of 15.4.
3. Cars speed values are generally away from the mean of 15.4, as standard deviation is relatively large. (Data points are spreading out over a relatively large range from the mean.)

Mean of cars distance= 42.98

Standard deviation of cars distance= 25.769377492025892

Inferences:

1. Average cars distance is around 42.98
2. Cars distance varies by around 25.769377492025892 on average from the mean distance of 42.98.
3. Cars distance values are generally quite far away from the mean of 42.98, as standard deviation is quite large. (Data points are spreading out widely from the mean.)
4. Comparing to cars speed, cars distance has much higher variability in this dataset.

**Q2) b. Top Speed (SP) and Weight (WT)-** please refer to Python code file.

Mean of Top Speed (SP)= 121.54027218037035

Standard deviation Top Speed (SP)= 14.18143157452861

Inferences:

1. Average top speed is around 121.54027218037035
2. Top speed varies by around 14.18143157452861 on average from the mean top speed of 121.54027218037035.
3. Top speed values in dataset are away from mean 121.54027218037035, indicate by large standard deviation. (Data points are spreading out from the mean.)

Mean of Weight (WT)= 32.412576910246905

Standard deviation Weight (WT)= 7.492812997393198

Inferences:

1. Average weight is around 32.412576910246905
2. Weight varies by around 7.492812997393198 on average from the mean weight 32.412576910246905.
3. Weight values in dataset are away from mean 32.412576910246905, indicate by large standard deviation. (Data points are spreading out from the mean.)
4. Comparing to top speeds however, weights have much lower variability in this dataset.

**Q2)**

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

**2) What can we say about the student marks?**

The student marks have a mean of 41.0, median of 40.5, mode of 41, standard deviation of 5.05266382858645. As mean, median and mode are almost the same which is 41.0, hence can be conclude that the student marks are in normal distribution, majority of the marks scored by the student are around 41.0, with few scores vary by about 5.05 marks from 41.0.

**3) What can you say about the Expected value for the student score?**

Expected value for the student score is 41.0, which means on average, the student is expected to score 41.0 marks. However, as there is variability as indicated by the standard deviation, actual scores can fluctuate around 41.0 both higher and lower, follow the spreading of approximately 5.05 marks.

**Q3), Q4), Q5), Q6), Q7), Q8) & Q9) –** Please refer to Python code file.

**Q10)**

**(i)** Please refer to Python code file.

**(ii) Is the venture likely to be successful? Explain.**

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

* The most likely monetary outcome of the business venture is $800.0. When rounded to the nearest $1,000, it is $1,000.
* From probability distribution data, after running this business for some time, there are 20% of chances for it to give an average return of $1,000.
* In real life, for a business to be considered successful, it should at least have >50% of chances in giving a relatively stable and positive returns (X >0).
* From Python calculation, PX\_positive = 0.6, which means there is 60% of chances this business will give a positive return.
* Plus, as mentioned earlier there are 20% of chances for it to give an average return of $1,000 over some time, so yes, the venture is likely to be successful with an expected average return of $1,000.

**(iii) 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 $1,000, which is equals to the most likely monetary outcome of the business venture (the expected return).
* This is because expected return represents the expected average return generated by the business after it has run some time.

**(iv) 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 good measure of the risk involved in a venture of this kind is the standard deviation, σ. In this business, standard deviation measure how widely differs each actual returns (or losses) from the expected return. The greater the variability of returns (or loss) from the expected return of $1,000, the riskier the venture of this kind.

**Compute this measure:** Please refer to Python code file.

From Python calculation, standard deviation (the measure of risk) is 1469.6938456699068, which is considered high. As higher standard deviation implies higher variability from expected return, hence, we can conclude that a venture of this kind is a high-risk venture.