Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Based on the Box Plot,

- Maximum booking happens in the **season3** with the median of 5000
- Most of the bookings happen in the **month of 5,6,7,8 & 9** with a median of over 4000 booking per month
- Maximum booking in the **weather set 1** with the median of 5000
- More Bookings are in the **working day** with the median of 5000. So it is the good predictor with the dependent variable

---

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

It is important to achieve k-1 dummy variables as it can be used to **delete extra column** while creating dummy variables.

For Example: We have three variables: Furnished, Semi-furnished and un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don't need unfurnished as we know 0-0 will indicate un-furnished. So we can remove it

It is also used to reduce the collinearity between dummy variables .

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

atemp and **temp** both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

To validate the assumptions of linear regression after building a model on a training set we used the below mentioned models

Correlation: Check for correlation between the independent variables

Scatter plot: Plot the data to check for a linear relationship between the independent and dependent variables

Residual plots: Use a histogram of residuals or a normal probability plot of residuals to check if the residuals are normally distributed

---

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)
The Top 3 features contributing significantly towards the demands of share bikes are:
Weather_fall (negative correlation).
yr_2019(Positive correlation).
temp(Positive correlation).

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>
Linear regression is a type of machine-learning algorithm more specifically a supervised machine-learning algorithm that learns from the labelled datasets and maps the data points to the most optimized linear functions, which can be used for prediction on new datasets.

Linear regression finds the line that best fits a set of data points. It models the relationship between a dependent variable and an independent variable as a linear equation
Linear regression can be used to predict future expenses based on past income and expenses.
There are different types of linear regression, including simple linear regression, multiple linear regression, and errors-in-variables models

---

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>
Anscombe's quartet is a set of four datasets that have similar descriptive statistics but look very different when graphed.
It is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analyzing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics (mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>
The Pearson correlation coefficient (r) is the most widely used correlation coefficient

- The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the

linear relationship between two quantitative variables.

- The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses

---

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- sklearn.preprocessing.scale helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

- The variance inflation factor (VIF) can be infinite when there is perfect correlation between variables. This happens when the R^2 value is 1 while calculating the VIF for one independent variable using all the other independent variables.
- The VIF is a statistical index that measures how much the variance of an estimated regression coefficient increases due to collinearity. A large VIF indicates a high degree of multicollinearity, and the regression coefficients may be poorly estimated.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

A Q-Q plot, or quantile-quantile plot, is a graphical tool that compares two probability distributions by plotting their quantiles against each other. Q-Q plots are useful for:

- Comparing distributions: Q-Q plots can show if two data sets come from the same distribution, or if a given sample fits a specified probability distribution.
- Detecting distributional aspects: Q-Q plots can show shifts in location, scale, symmetry, and outliers.
- Building machine learning models: Q-Q plots can help ensure that a machine learning model is based on the right distribution