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)

I have done analysis on categorical variables using boxplots. Below are the points we can infer:

- fall season had more bookings
- bookings increased in subsequent year
- more number of bookings have been done during the months of May, June, July, Aug, Sep, Oct
- when it is holiday, bookings seem to be less in number
- fri, sat, sun seems to have more bookings
- bookings seem to be equal on working and non workings days
- more number of bookings in clear weather which seems obvious¶

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)

drop_first=True is important to use as it reduces the extra column created during dummy variable creation. Hence it reduces the correlations in the dummy variables.

Let's say we have 3 values in a categorical column (A,B,C) and we want to create dummy variables of that column. The column can be represented by two dummy variables only because if the variable is not A and B, then obviously it is C.

Therefore drop_first=True is used to create k -1 dummies out of k categorical levels.

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)

'Temp' variable has highest correlation with the target variable.

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)

I have validated the assumptions of Linear Regression on below assumptions:

- Normality of error terms –
  Error terms are normally distributed with mean zero.
- Multicollinearity check –
  There should be insignificant multicollinearity among variables (VIF less than 5)
- Linear Relationship validation –
  Linearity should be visible among variables.
- Homoscedasticity -
  The variance of error terms should be constant

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 are temperature, year, winter.

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 supervised machine learning algorithm that computes a linear relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data.

Mathematically the relationship can be represented by below equation-
Y=mX + c
Y is the dependent variable we are trying to predict.
X is the independent variable we are using to predict
m is the slope of the regression line which represents the effect X has on Y
c is the constant term or intercept

Linear Regression is two types-
1) Simple Linear Regression – It explains the relationship between a dependent variable and only

one independent variable using a straight line.
2) Multiple Linear Regression - It shows the relationship between one dependent variable and several independent variables.

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, where each produces the same summary statistics (mean, standard deviation, and correlation), which could lead one to believe the datasets are quite similar. However, after visualizing (plotting) the data, it becomes clear that the datasets are very different.
Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

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>

Pearson correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by r. It can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association, that is, as the value of one variable increases, the value of the other variable decreases.

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>

Scaling is a step of data pre-processing which is applied to independent variables to adjust the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of

the times, collected data set contains features highly varying in magnitudes and units. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling:

- It brings all the data in the range of 0 and 1.
- Scikit-Learn provides a transformer called MinMaxScaler for Normalization.
- Normalization is useful when there are no outliers as it cannot cope up with them

Standardization Scaling:

- It brings all the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
- Scikit-Learn provides a transformer called StandardScaler for standardization.
- Standardization does not get affected by outliers because there is no predefined range of transformed features.

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>

If there is perfect correlation, then VIF = infinity. It indicates that one or more independent variables in a model can be perfectly predicted by the other variables leading to the perfect multicollinearity. This happens when two or more independent variables in a model are perfectly linearly dependent.
In case of perfect correlation we get R- squared (R2)= 1 ,which leads to 1/1-R2 as infinity. In this case we need to drop variables from the dataset which are causing perfect collinearity.

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>

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.
Use of Q-Q plot in Linear Regression:
Q-Q plot can be used on the residuals of a simple linear regression to check if they are normally distributed. You can also create a Q-Q plot to check the distribution of the variables before you create a linear regression in the first place.