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)

Following are my observations from categorical variables:

1. Demand goes down in spring and is the highest in fall
2. Demand is lowest at the start of the year and rises close to 5800 towards the end of the year.
3. The Demand is higher on a non-holiday than a holiday.
4. Demand is lowest on days when there is Light Rain + Thunderstorm.
5. Workingday vs Non Working day does not make much difference to the demand.
6. The number of rentals in 2019 were much higher than 2018

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)

So that we always keep n-1 number of dummy variables (n are the count of categories in the column). This ensures that there is no multicolinearity in the data as the sum of dummies now do not add up to 1.

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)

The temperature and feels like temperature column has the 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)

1. We validated the linear correlation assumption using scatter plots to check if there is a linear correlation between the independent and target variables.
2. We Checked the multicollinearity assumption using the heatmaps of correlations and the VIF scores.
3. We also made a plot of the distribution of the residuals to check if the residuals have a mean of 0 and are normally distributed.
4. We checked that the residuals do not follow a certain pattern.

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 three features that contribute to the demand of the bikes are –

1. Temperature
2. Year – 2018 or 2019
3. Weathersit category – light rain or thunderstorm Dummy Variab;e.

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 supervised technique that is used to predict numerical values. It does so by fitting a line through the data by a method called Least Square Method.
In other works, it helps us to get an estimate of the value of y given any value of x using the equation:
y = mx + c
where y is the target, x is the predictor, c is the y-intercept and m is the slope of the line.
Linear regression is all about calculating the best values for m & c such that the squared error between the actual and predicted values is minimized.

Linear regression can be performed with multiple variables as well with the below equation:
Y = m1x1 + m2x2 + m3x3 + m4x4 + …… + MnXn + C

Interpretation: The Change in Y with change in Xn given all other Xs are help constant is Mn.

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 collection of 4 data with similar mean, standard deviation and regression line. However, when we visualize the data, we see underlying differences in the data.
The Anscombe's quartet demonstrates the importance of visualizing the data before using the data.

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's R measures the linear relationship between two continuous variables.
The formula for Pearson's correlation coefficient is:
$r = Cov(X,Y) / \sigma X * \sigma Y$
The value of the pearson's R varies from -1 to 1 in which:
1 indicates a strong positive correlation between two continuous variables – One increases with the other and vice - versa
0 indicates a weak correlation between two continuous variables – One does not change with other and vice-versa
-1 indicates a strong negative correlation between the two variables – One increases as the other decreases.

However, it takes these 2 assumptions:
1. Variables should be approximately normally distributed.
2. The relationship between the two variables should be linear.

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 method to adjust the range or distribution of a particular variable in a dataset. This does not change the relationship of the variables with other variables in the dataset but just brings the column to a right range.

We use scaling because:
1. It adjusts the values of high magnitude variables which get undue importance in their weights.
2. It optimizes the performance of the optimization algorithm like gradient descent.

With normalized scaling, the ranges of the variables are fixed (eg [0,1] or [-1,1]) whereas with standard scalar, the scalar adjusts the mean of the data to 0 with a standard deviation of 1
Normalized scaling is more sensitive to outliers than standard scaling.

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>
VIF score gives an idea of how a particular variable is explained by the other variables in the data.
VIF is refined as 1/(1-R2). This becomes infinite when R2 = 1. The R2 value goes to 1 when we have perfect multicollinearity in the data. This means that the there are either **duplicate columns** in the data, or the data is able to perfectly explain a particular variable in the data.

Infinite values are mostly seen when there are two or more identical columns in the data.

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 indicates if our data follows a particular theoretical distributions. For example, it can help us asses and see the distribution of one of our columns in the data and compare if it matches with a theoretical distribution.

In a linear regression, we generally use a Q-Q plot to plot the residuals to check if the residuals are normally distributed. In this plot, deviations from the strainght line may suggest non-normality, which may indicate the presence of outliers in our data.
