Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans:

1> We see the demand is highest in fall and least in spring.
2> We see a significant increase in demand from year 2018 to 2019, this suggest that this business has the potential to grow overtime.
3> We see a increase in the median for demand from jan to jun then its kinda constant for jul, aug, sep and then it declines.
4> We also observe the demand is highest in month sept.
5> We observe on sunday the demand is little less compared to other days but its not clear.
6> Nothing substantial can be inferred from working day.
7> We see when the weather is good there is increase in the demand. But its significantly low when the weather is bad.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Ans:

n-1 variables are able to predict the value of all n variables, hence one variable can be dropped and if we don't do this as these variables are themselves correlated which is known as multicollinearity, and it leads to dummy variable trap.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans:

temp and atemp

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans:

It can be validated by plotting a scatter plot between the features and the target.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans:

Year, temperature and weather situation

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear regression is a type of supervised ML algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Simple Linear regression, and in the case of more than one feature, it is known as Multiple Linear regression. The goal of the algorithm is to find the best fit linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable.

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans: Anscombe's quartet comprises of four sets that are identical when examined using simple summary statistics, but vary considerably when graphed.

3. What is Pearson's R? (3 marks)

Ans: It is a correlation coefficient that measures linear correlation between two sets of data. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. In normalized data is fixed between 0 and 1 using min and max whereas in standardization we subtract mean and then divide by standard deviation. One more difference standardization is not affected by outliers.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: When there is perfect correlation.

(3 marks)

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans: The QQ plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a normal or exponential.
(3 marks)