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?

Answer:

- **Season**: Almost 35% of the bikes were booked in season3 with a median of over 5000in 2 years. Similarly in season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.
- **mnth**: Almost 10% of the bike booking were happening in the months 5 to & 9 with a median of over 4500 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
- **weathersit**: Around 5000 bike booking were happening during 'weathersit1. This was followed by weathersit2 with 4000 median booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
- **holiday**: Maximum bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.
- **weekday**: weekday variable shows very close trend having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor.
- **workingday**: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

2. Why is it important to use drop_first=True during dummy variable creation?
Answer: drop_first=True helps in eliminating the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
Assuming we have 3 types of values in Categorical column as Furnished, Semi Furnished and Unfurnished and we want to create dummy variable for that column. If the variable is not Semi - furnished and Unfurnished, then it is obvious Furnished. So we do not need 3rd variable to identify the Furnished

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Answer: Target Variable cnt has the highest correlation with atemp.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Answer: The best way to determine if the assumption is met or not is by creating a scatter plot x VS y. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds true.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer: The top features contributing significantly towards explaining the demand of the shared bikes are Temperature, year, season_4 and mnth_9

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a supervised machine learning algorithm used for predicting a dependent variable based on one or more independent variables. It assumes a linear relationship between the input variables and the output.

The goal of linear regression is to find the values of $\beta 0$, $\beta 1$, ...,βn that minimize the sum of squared differences between the observed and predicted values:

Minimize: $\sum i=1m(Yi-(\beta 0+\beta 1X1i+\beta 2X2i+...+\beta nXni))2$

simple linear regression with one independent variable, the equation is:

$$y=mx+b$$

Linear regression is widely used for predicting numerical outcomes and understanding the relationships between variables in various fields, including economics, finance, biology, and social sciences.

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. What is Pearson's R?

Answer: The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

The Pearson's R 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. Specifically, we can test whether there is a significant relationship between two variables.

The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, r is negative. When the slope is positive, r is positive. When r is 1 or –1, all the points fall exactly on the line of best fit:

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: Scaling is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. Standardization centers data around a mean of zero and a standard deviation of one, while normalization scales data to a set range, often [0, 1], by using the minimum and maximum values.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer: If there is perfect correlation, then VIF is infinity. A large value of VIF indicates that there is a correlation between the variables.

So we can say that if the given variables has the highest correlation then the value of VIF can range up to infinity, AS there is no upper bound for the values of VIF.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer: Q-Q (Quantile-Quantile) plot is a graphical tool that compares the quantiles of a data distribution to the quantiles of a theoretical distribution. The purpose of a Q-Q plot is to determine if a dataset follows a particular type of probability distribution. In linear regression, the intercept and slope of a linear regression between the quantiles gives a measure of the relative location and relative scale of the samples.Q-Q Plot helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential, etc.