ASSIGNMENT SUBJECTIVE QUESTIONS

Q1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

The categorical variables have an effect on the dependent variables:

- Season: The season possibly explains the variation in the number of bikes booked.
- Weather: The weather conditions, specifically overlapped with particular seasons, produce similar effects, and is therefore a determinant. However, we might see a collinearity between seasonality, month and weather. Therefore, the usage of all three variables might not be necessary
- Holiday: Holiday exhibits extreme variations and is not a good predictor.
- Working day and weekday: We let the model determine the usefulness of this variable.

Q2) Why is it important to use drop_first = True during the dummy variable creation?

During the dummy variable creation, the drop_first attribute helps us drop the redundant column. The extra column created during the dummy variable creation to explain the levels.

If there are n-levels for a ategorical variable, we need only (n-1) columns to represent these levels.

Q3)L ooking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The variable 'atemp' has the highest correlation (0.63) with the target variable

---

Q4) How did you validate the assumptions of Linear Regressions after building the model on the training set?

1. Check for the distribution of the residuals. The distribution must be normal. Heteroskedasticy is not acceptable.
2. The VIF, after every model, explains the multi-collinearity persistent within the independent variables.

Q5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Season_4, atemp, and yr.

GENERAL SUBJECTIVE QUESTIONS

Q1) Explain the Linear Regression Algorithm in detail

1. **REGRESSION ALGORITHIM (in detail)**

STEP 1: Exploratory DATA analysis

Understand the different attributes and how they contribute to the distributions. Visualizing the data can help you relate to the variables that are highly correlated and give you a sense of the data.

STEP 2: DATA PREPARATION

Once you are done exploring the data, check the nature of the variables: Numerical, Dummy, or Categorical. Depending upon the type of the variable, decide the columns that you need. Check what makes sense.

1. Convert the dummy variables (Encode) into numerical variables
2. Use scaling for the numerical variables (Standard scaler, or MinMax scaler)

STEP 3: TRAINING THE MODEL and CHECKING FOR MULTICOLLINEARITY

After data preparation, we move onto training the model using Scikitlearn or Statsmodels. After training the model, we check the VIF of the variables;

Based on the VIF, we eliminate those variables that exhibit a high multicollinearity. Dropping the variables that are unnecessary gives us the next step to building a robust model. Based on the VIF, and R-square, we eliminate and arrive at the correct set of '**Features'.**

STEP 4: RESIDUAL ANALYSIS

The residual analysis allowed us to ascertain the distribution of the error terms in question.

STEP 5: PREDICTING THE TARGET VARIABLE

Predicting the value of the target variable and evaluating it against the value of the trained y will help us understand the difference in the predicted vs the actual y values.

Q2) Explain the Anscombe's quartet in detail

Anscombe's quartet gives us statistical properties for graphing data. There are 4 data sets, with simple statistics that are nearly identical, which means that they have similar means, medians, and mode. Although summary statistics can give you quick overview of the data set, you can never reply only on it. This is where **Data visualization** plays a key role.

Four data sets with similar averages for x and y, similar variances, similar correlations can still produce different scatters

---

Q3) What is Pearson's R?

Pearson's R is basically the correlation between two variables (bivariate). The relationship between two variables is determined by the product of the two variables from its center; the idea is that it is standardized. The standardization ensures that the 'r' value ranges from -1to 1.

The Correlation coefficient is used to describe the 'strength' of the relationship between variables. A greater value of 'r' implies a stronger correlation between the variables.

Q4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is the process of bringing some semblance of sanity to the data. Imagine having one variable that runs into millions, while another variable has a range from 1 to 100. The data is then bound to a particular range and it helps us create useful interpretations because everything then becomes comparable

The output data and the coefficients would be comparable post scaling. Bringing the data set into the same magnitude, units, and range is essential. Scaling has to be done to bring everything to the same level (range)

Scaling affects coefficients only, and we are looking at the comparability of coefficients when we speak about scaling.

STANDARDIZED SCALING

Standardized scaling is the effect of using the standard deviation and the mean of the data to scale.

$X = \{x - mean(x)\} / std(x)$

NORMALIZED SCALING

MinMax scaling (or Normalized scaling) uses the range of the data, and plots all the points within that given range and scales it from 0 to 1

$X = \{x - min(x)\} / \{max(x) - min(x)\}$

Q5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The value of VIF is infinite if there is a perfect correlation among the variables. This means that the Coefficient of determination is 1. This does not make sense as no variable can completely explain the dependent variable. An infinite VIF indicates 'perfect multicollinearity'- multiple variables being able to explain another independent variable.

The problem is solved by dropping one of the variables, and then recalculating the VIF.

Q6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

The Quantile-Quantile plot is a graph describing the relationship between the quantiles of 2 data sets. This helps us establish how closely one graph follows another. The points should fall along the same cluster. When we sample from the same distribution, the data points should follow almost the same distribution.

Sample distribution vs Theoretical distribution : The relationship between these two is exactly what the Q-Q plot measures.

The Q-Q plot can tell us a characteristic distribution. Eg- In a normal distribution, the cluster of data points should be around the median.

The distribution of the error terms can be ascertained using the Error plot in a linear regression. The quantile data helps us realize whether the scatter of errors is correct.