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?

   Categorical variables such as season and weather situation have a high influence on the dependent variable. If the season or weather situation changes it has impact on the bike booking count

   There are few other categorical variables such as month and weekday. These variables were least significant and did not influence on the dependent variable.

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

   Drop_first=True helps to reduce multicollinearity. When you create dummies for k levels it's better to drop one of the columns and just have k-1 dummies(columns) to represent k levels. Separate column will not be needed since one of the combinations will be uniquely representing this redundant column.

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

   Temp variable has the highest correlation with the target variable

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

   Assumptions in the linear Regression are

   - Linearity Assumption: There exists a linear relationship between the independent variable and the dependent variable. You can use scatterplot to check the relationship between independent and dependent variable.

   - Normality assumption: It is assumed that the error terms, $\varepsilon^{(i)}$, are normally distributed.
   - Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.

Use a distplot to plot the residuals, it should show a normal distribution centered at zero

- Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, σ^2 . This assumption is also known as the assumption of homogeneity or homoscedasticity.
- Independent error assumption: It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero
- The independent variables are measured without error.
- The independent variables are linearly independent of each other, i.e. there is no multicollinearity in the data

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

   Temp, year, and weather situation are the top 3 features

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Linear regression is a machine learning algorithm which estimates how a model is following a linear relationship between target variable and one or more predictor variables

Steps Involved in Linear Regression

1.Importing required Libraries

2.Reading and Understanding the Data : Cleaning and Manipulating the data, handling null values, removing redundant columns,rows,changing data types

3, Perform EDA, understand variables and correlation between the variables, visualize numerical and categorical data using the different plots available

4.Data Preparation: Create dummy variables for categorical variables with varying degrees of levels

5. Split the data into train and test set (70:30 or 80:20)

6. Perform scaling: Normalize range of numerical variables

7. Create linear model (Forward/Backward/RFE methods)

8. Check the various linear regression assumptions, residual analysis

9.Report the final model, make predictions using the final, evaluate the model using test set

10. Calculate r2_score for test and train set. Calculate MSE (mean square error) RMSE (Root Mean square error)

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

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots

It is used to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties

It also suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc

Consider 4 data sets with approximately similar statistical information. When the model is plotted all four data sets give different kind of plot

Dataset1 : Fit Regression model

Dataset 2 : Data is non linear so cannot fit regression model

Dataset 3 : Outliers in the dataset cannot be handled by linear regression model

Dataset 4 : Outliers in the dataset which cannot be handled by linear regression model

So , *all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model*

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

The Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

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

Scaling: Means that you're transforming your data so that it fits within a specific scale like 0-1. By scaling your variables, you can help compare different variables on equal footing. I*t is a step of data Pre-Processing which is applied to independent variables. It also helps in speeding up the calculations in an algorithm*.

Why Scaling : Collected data set might contain features highly varying in magnitudes, units and range. 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.

NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

   If there is perfect correlation between 2 independent variables (perfectly predicted by other variable), then VIF will be equal to infinity. This happens when R^2 approaches 1.

6. 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 graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set

Use and Importance

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behavior.