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?

- It seems that demand of bikes is higher in fall followed by summer season.
- Also more bikes are rent during working days that on weekends or hoidays .
- More bikes are rent during 2019 as compared to 2018.

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

=> It is used to remove redundant columns for dummy variables. While creating dummy variables if we have n categories then we can create n-1 dummy variables for that feature. So it is necessary to drop redundant column if other dummy variables are representing it. For example we have four categories for season so we can create three dummy variables with below representation.

- 000 will correspond to fall

- 100 will correspond to spring

- 010 will correspond to summer

- 001 will correspond to winter

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

=> temp feature is highly correlated with target variables. Correlation for temperature is almost 0.63.

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

=> We can validate the assumptions of Linear Regression by calculating Variance Inflation Factor(VIF), Error distribution for residuals and R2 for test datset.

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

=> The top 3 features contributing significantly are temperature, year and holiday.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

=> Linear Regression is a machine learning algorithm which is used to perform predictive analysis for features continuos in nature. Equation for linear regression is

y = mx +c where m reprsents slope and c is intercept

There are two types of linear regression :

- Simple Linear Regression – Where only one feature considered
- Multiple Linear Regression – In this model more than one feature are considred

2. Explain the Anscombe’s quartet in detail.

=> 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.

Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

3. What is Pearson’s R?

=> 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.

- Between 0 and 1 - Positive correlation (When one variable changes, the other variable changes in the same direction.)
- 0 - No correlation (There is no relationship between the variables.)
- Between 0 and –1 - Negative correlation (When one variable changes, the other variable changes in the opposite direction.)

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

=> This 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 scaling.It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains 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.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

- Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1.
- Standardization Scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one(σ). z = (x - u) / s

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

=> The value of VIF is infinite when there is perfect correlation between the two independent variables. The R squared value is 1 that case and leads VIF to infinity. If you see this in model it represents there is a multicollinearity between the features and it may cause your model not useful so we need to drop one of the feature to define a working model.

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

=> Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

Normal distributions

We regularly make the assumption of normality in our distribution as we perform statistical analysis and build predictive models. Machine learning algorithms like linear regression and logistic regression perform better where numerical features and targets follow a Gaussian or a uniform distribution.