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?

Ans:

- Demand of rental bike is highest during 'fall' season followed by 'summer', then 'winter' and least during 'spring'
- There is an increase in demand, on year on year basis(from 2018 to 2019)
- People tend to rent bike in a good weather situation
- Wednesdays and Saturdays has a significant demand

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

Ans:

To avoid creation of duplicate/redundant column during dummy variable column. drop_first= True will drop the original variable and create the resultant dummy variables

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

Ans:

temp has the highest correlation with cnt. Also, atemp has the similar correlation with target variable but since both temp and atemp has correlation = 0.99, we have dropped atemp

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

Ans:

By performing residual analysis:
1.Assumption Validation: if error terms are normally distributed

2.Assumption Validation: if error terms are independent of each other

3.Assumption Validation: Model is Homoscedastic or not

Observations:

1. Normal Distribution: The error distribution is normal (i.e concentrated around 0) which is another assumption of linear regression.

2. Independent from each other : There is not pattern detected so it can be concluded that error terms are independent of each other

3. Error terms has constant Variance: Variance is constant. Hence we can say heteroscedasticity is not observed in the error terms. So model is truly homoscedastic

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

Ans:

temp : 0.5489

yr:0.2385

season_winter:0.1165

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

Linear Regression is a machine learning algorithm based on **supervised learning**

It is mostly used for finding out the r**elationship between variables and forecasting.**

Different regression models differ based on – the **kind of relationship** between dependent and independent variables they are considering, and t**he number of independent variables** getting used.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)). Hence, the name is Linear Regression.

Dependent variable are also called as an outcome variable, criterion variable, endogenous variable, or regressand.

Independent variable are also called as an exogenous variables, predictor variables, or regressors.

Example: in finance, linear regression might be used to understand the relationship between a company's stock price and its earnings, or to predict the future value of a currency based on its past performance.

2. Explain the Anscombe's quartet in detail.

Ans:

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyse it and build your model.

It is a set of four datasets that have nearly identical summary statistics, but very different visual patterns.

It tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

3. What is Pearson's R?

Ans:

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.

The Pearson correlation coefficient is a descriptive statistic, which summarises 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, which can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

Another way to think of the Pearson correlation coefficient (*r*) is as a measure of how close the observations are to a line of best fit.

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.

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

Ans:

What?

Scaling is a data Pre-Processing step which is applied to independent variables to normalise the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why?

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 results in 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.*

Normalisation/Min-Max Scaling brings all of the data in the range of 0 and 1.
sklearn.preprocessing.MinMaxScaler helps to implement normalisation in python.

Standardisation Scaling 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).
sklearn.preprocessing.scale helps to implement standardisation in python.

One disadvantage of normalisation over standardisation is that it **loses** some information in the data, especially about **outliers**.

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

Ans:

If there is perfect correlation, then VIF = infinity.
If R2 is 1 then it will lead to infinity value as VIF = 1/1-R2
A large value of VIF indicates that there is a correlation between the variables.
A general rule of thumb is that if VIF > 10 then there is multicollinearity

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

Ans:

Q-Q plots are also known as Quantile-Quantile plots.
It plots the quantiles of a sample distribution against quantiles of a theoretical distribution.
Doing this helps to determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.