Assignment-based Subjective Questions

1Q cFrom your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A) Below explained Analysis

A) Bike demand in the fall is the highest.
B) Bike demand in year 2019 is higher as compared to 2018.
C) The demand of bike is almost similar throughout the weekdays.
D) Bike demand doesn't change whether day is working day or not

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

A) it helps in reducing the extra column created during dummy variable creation

It useful bcoz it reduces the number columns, here is how, when all the other columns are zero that means the first columns is 1

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

A) 'temp' variable has the highest correlation with the target variable.

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

A) 1. Normality of Error Terms: This test checks whether the error terms (residuals) follow a normal distribution, indicating that most residuals should be clustered around zero.
2.Multicollinearity: This test checks for the presence of multicollinearity, which occurs when two or more predictor variables are highly correlated. The assumption is that there should be no significant multicollinearity among the features.
3. Linearity: This test checks if the relationship between the independent variables and the dependent variable is linear. Residuals should be randomly scattered around zero when plotted against the predicted values.

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

A) Top 3 features features contributing significantly towards explaining the demand of the shared bikes are
    1) Temperature
    2) Year
    3) Holiday

General Subjective Questions

1. Explain the linear regression algorithm in detail

A) Linear regression is also a type of machine-learning algorithm more specifically a supervised machine-learning algorithm that learns from the labelled datasets and maps the data points to the most optimized linear functions. which can be used for prediction on new datasets.

First of we should know what supervised machine learning algorithms is. It is a type of machine learning where the algorithm learns from labelled data. Labeled data means the dataset whose respective target value is already known. Supervised learning has two types:

- Classification: It predicts the class of the dataset based on the independent input variable. Class is the categorical or discrete values. like the image of an animal is a cat or dog?
- Regression: It predicts the continuous output variables based on the independent input variable. like the prediction of house prices based on different parameters like house age, distance from the main road, location, area, etc.

2) Explain the Anscombe's quartet in detail.

A) Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots.

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.

The Pearson correlation coefficient (*r*) is the most widely used correlation coefficient and is known by many names:

- Pearson's *r*
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

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

A) *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*
- *sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.*

  MinMax Scaling: x=x-min(x)/max(x)-min(x)

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).*

  Standardisation = x=x-mean(x)/sd(x)

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

A) This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

A) A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference. For example, you can use a Q-Q plot to check if the residuals of the model are normally distributed, which is an assumption for many parametric tests and confidence intervals.