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 :

Following inferences from the data visualization of categorical variables in the data set

a. On holiday number of bikes rented are less compared to working days.
b. **Median** of bike rented on all working days is almost same
c. **Median** of bikes rented is high from May to September. Demand is gradually increasing January and peaks in July. Demand start drop from October onwards.
d. Demand for rented bikes is high in
   "weathersit" 1: Clear, Few clouds, Partly cloudy, Partly cloudy
   followed buy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist.
   No bikes rented during 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

e. Highest bikes are rented during fall and summer season.
f. Working day yes or no is not impacting much in bikes rented

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

Ans :

Setting "drop_first=True" helps in reducing number of dummy variables created by 1 for each dummy variable. This reducing computing power needed for very big data base. Also helps to reduce multicollinearity.

If dummy variable has m level then it needs m-1 dummy variables.

If there are 'n' number of dummy variables and Ni is number of level for each dummy variable

Then total number of dummy variables = Sum of "Ni -1 " where is from 1 to n

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans :

atemp has highest correlation with cnt. Temp has next highest correction.

Both looks almost have same amount of correlation with cnt.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks).

Ans :

Validated the model with following thing.

a) Error terms distribution : Plotted the predicted y with "distplot". From the graph. It shows that error teams are normally distributed. This one of most important assumption of Liner regression.

b) Verified by predicting y on test data using the model. Compared predicted versus actual y. Plotted y-test versus y-pred to check the accuracy of model
c) Calculated r-squared score by using "r2_score function from sklearn library" and also calculated adjusted r-squared score. Compared these values with model r-squared and adjusted r-squared
d) Took the conclusion on factors affecting bike renting based on coefficient from the model.

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

Ans :

1. Temp – Temperature
2. Weathersit_3 – Weather situation 3
3. Yr – Year

Are the top features contributing towards shared bike demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans :

Linear Regression : Linear regression analysis is used to predict the value of a variable called dependent variable based on the value of another variable called predictor variable.

There are two types of linear regression models:

- Simple linear regression

  In simple liner regression, single dependent variable will be analysed again one predictor variable at a time

- Multiple linear regression

  In MLR : A single dependent variable will be analysed against multiple predictors variables at a same time

2. Explain the Anscombe's quartet in detail.

Ans :

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics.

It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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

3. What is Pearson's R?

Ans :

The Pearson correlation coefficient (r) : is most commonly used measurement of liner correlation between two variables

It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (*r*)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and -1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the

			lower the air pressure.

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

Ans :

Scaling : Transforming the given data to fit into a specific scale is called scaling. Scaling does not impact the model.

Scaling factor : Is a number which multiplies ("scales") a quantity.

Why is scaling performed to fit the data into given scale. So that it is easy to interpret the output of model

difference between normalized scaling and standardized scaling :

A normalized dataset will always have values that range between 0 and 1.

A standardized dataset will have a mean of 0 and standard deviation of 1, but there is no specific upper or lower bound for the maximum and minimum values.

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. The greater the VIF, the higher the degree of multicollinearity. That means when multicollinearity is perfect (i.e., the regressor is equal to a linear combination of other regressors), the VIF tends to infinity.

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

Ans :

Q-Q Plot : Aare an exploratory tool used to assess the similarity between the distribution of one numeric variable and a normal distribution, or between the distributions of two numeric variables.

The purpose of the quantile-quantile (QQ) plot is to show if two data sets come from the same distribution. Plotting the first data set's quantiles along the x-axis and plotting the second data set's quantiles along the y-axis is how the plot is constructed.

Uses :

In Linear Regression it used to to determine. If two populations are of the same distribution.