Subjective Questions

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?

Working days have almost equal impact on the count
Holidays have more impact on the count. The interquartile range is more on Holidays which indicate more users use on holidays.
For month wise, It gradually increases and during Apr-Oct the usage is high and then it decreases. It seems obvious that during winter the usage/count reduces
Summer and Fall have more usage compared to winter and spring. Spring also has an outlier.
The weekday distribution is almost same with Fridays having a more interquartile range.
Average count is more in 2019 compared to 2018. It gives a slight hint of a yearly increase in usage.
When the weather situation is clear without clouds or few clouds people uses bike more.
When there is snow the usage/count is low.

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

When creating dummy variables for a categorical feature with n levels, n dummy variables are created. If all n dummy variables are included in the model, they will be linearly dependent. For example, if you have a categorical variable with three levels (A, B, and C), creating three dummy variables will result in a situation where the third variable is perfectly predicted by the first two (A + B + C = 1). This causes multicollinearity.

Dropping one dummy variable (using drop_first=True) reduces redundancy. For a categorical feature with n levels, only n-1 dummy variables are needed to uniquely represent the feature.

Without drop_first=True

Color_Red	Color_Blue	Color_Green	
1	0	0	(Red)
0	1	0	(Blue)
0	0	1	(Green)

With drop_first=True

Color_Blue	Color_Green	
0	0	(Red - reference category)
1	0	(Blue)
0	1	(Green)

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

When we exclude casual and registered, the highest correlation with the target variable is for temp with 0.63.
The sum of casual and registered is the target variable. If we consider these two variables the the highest correlation is for the registered variable with 0.95

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

The p value is less than 0.05 and VIF is less that 5 for all our independent variables.
The error terms should be normally distributed with the mean as zero. Used a distplot to verify it.
There is a linear relationship between the independent variables and the dependent variable.
The independent variables are not highly correlated which were verified using VIF.
The residuals have constant variance at every level of the independent variables (Homoscedasticity) using a scatter plot..

.

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

Windspeed
Weather situation with light snow, light rain
Year 2019

General Subjective Questions

1. Explain the linear regression algorithm in detail?

Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). It is used for prediction and forecasting in various fields.
We use linear regression to find the best-fitting linear relationship between the dependent variable y and the independent variable(s) X.
The equation is
Y = B0 + B1X1 + B2X2 + ….+BnXn + e

Linearity: The relationship between the independent and dependent variables is linear.
Independence: Observations are independent of each other.
Homoscedasticity: The residuals (errors) have constant variance at every level of X
Normality: The residuals of the model are normally distributed.
No multicollinearity: The independent variables are not highly correlated with each other.

Steps in Linear Regression

1. Data Preparation
2. EDA
3. Model Building
4. Model Evaluation
5. Residual analysis
6. Prediction using the final model.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but appear very different when graphed.

This set of data was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and the effect of outliers and the distribution of data on statistical properties. The quartet illustrates that relying solely on summary statistics can be misleading without a visual inspection of the data.

We need to always visualize our data. Graphs can reveal patterns, trends, and anomalies that are not obvious from summary statistics alone. It can detect outliers, see trends that helps to identify the patterns so that we can choose the right model.

x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68

x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74

x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73

x: 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 19
y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 5.56, 7.91, 6.89, 12.50

3. What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient or simply the correlation coefficient, is a measure of the strength and direction of the linear relationship between two continuous variables. It is denoted by r and ranges from -1 to 1.

+1 - perfect positive linear relationship
-1 - perfect negative linear relationship
0 - No linear relationship

When it is positive, it indicates that when one value increases the other value also increases. When it is negative, it indicates that when one value increases the other value also decreases.

When it is close to zero, it suggests weak relationship, when it is close to +-1 it suggests strong relationship.

Pearson’s R is calculated as the covariance of the two variables divided by the product of their standard deviations.

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

Scaling refers to the process of adjusting the range of feature values in a dataset so that they are on a comparable scale. It is a crucial preprocessing step in machine learning and data analysis to ensure that features contribute equally to the model and to improve the performance and convergence speed of algorithms.

Two major methods are employed to scale the variables: standardisation and MinMax scaling. Standardisation brings all the data into a standard normal distribution with mean 0 and standard deviation 1. MinMax scaling, on the other hand, brings all the data in the range of 0-1. The formulae used in the background for each of these methods are as given below:

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

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

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

Variance Inflation Factor (VIF) is a measure used to detect the presence of multicollinearity in a multiple regression model. Multicollinearity occurs when two or more predictor variables in the model are highly correlated, meaning they provide redundant information.

VIF = 1/(1 - Ri sqaure)
where Ri square is the R-squared value obtained by regressing the i-th predictor on all other predictors

When Ri square become 1, VIF will be infine.
This happens due to high multicollinearity. Adding the same variable by mistake.
Adding temperature in Celsius and temperature in Fahrenheit as predictors in the same model. Creating dummy variables without dropping one.

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a given theoretical distribution. In the context of linear regression, it is commonly used to check if the residuals (errors) of the model are normally distributed, which is one of the key assumptions of linear regression.

Quantiles of the sample data are plotted against the quantiles of a specified theoretical distribution (usually the normal distribution).

The sample quantiles are plotted on the y-axis, and the theoretical quantiles are plotted on the x-axis.

If the data follows the theoretical distribution, the points on the Q-Q plot will approximately lie on a straight line (the 45-degree line).

sm.qqplot(residuals, line ='45') is the command for Q-Q plot.

A Q-Q plot can help in diagnosing problems with the model. If residuals are not normally distributed, it may indicate that the model is not capturing some aspect of the data, such as non-linearity, or that there are outliers influencing the model.