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?

Season:

- We can see that most bike sharing happens in Fall Season around 32%
- In summer bike sharing happens around 27% and 25% in winter and 14% in Srping.

Month:

- we can see that bike sharing slowly increases from jan to may and from may it kept increasing at a steady rate at 10% till october and then a slight decrease happens

holiday:

- 97% of total booking happens in non-holiday days

Workingday:

- about 69% of the booking happens on working days

weathersit:

- 68% of total sharing happens during clear sky
- 30% of total sharing happens during mist
- 1.1% of total sharing happens during light rains

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

Because it becomes redundant i.e; we only require (n-1) number of variables to explain n different levels.

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

The Variables temp and atemp are the variables that have 0.643517
and 0.646475 respectively.

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

By doing residual analysis and checking the normal distribution of error terms.

Also plotting the test_values vs test_pred_values to see if the model predicts the test dataset with the same accuracy as train dataset or not and confirming for no overfitting.

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

Temp, light_rain and year in that order.

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Linear regression explains the linear relationship between a **dependent variable (target)** and one or more **independent features (predictors)**. It explains the impact of each independent variable on the dependent variable.

basic equation for linear regression looks like –

y= c+ m1x1+ m2x2+ m3x3+….

Where x1,x2,x3,… re independent vars and y is dependent var. c- constant

M1, m2, …. Are coefficents.

The goal is to find the best-fitting linear equation that minimizes the difference between the predicted values and the actual target values. The algorithm estimates the coefficients by minimizing the cost function like Mean Squared Error.

Assumptions of linear regression:

- Linearity: Assumes a linear relationship between variables.
- Independence: Assumes independence of errors.
- Homoscedasticity: Assumes constant variance of errors.
- Normality: Assumes normally distributed errors.

2.Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a fascinating set of four datasets that highlight the importance of data visualization and the limitations of relying solely on summary statistics**.** These datasets have identical descriptive statistical properties in terms of means, variances, R-squared values, correlations, and linear regression lines. but when we plot them on a graph, they exhibit distinctive patterns and relationships. It teaches us that basic statistics alone are inadequate for describing realistic datasets. Always visualize data before making conclusions.

3.What is Pearson's R?

It is the most widely used correlation coefficient. It explains strength and direction of the linear relationship between two quantitative variables**.** Ranges between -1 to 1 and 0 means no correlation. >0.5 means strong positive correlation and <-0.5 means strong negative correlation. It also reveals the slope of the line +ve or –ve.

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

It is performed to ensure that all features are on a similar scale. Algorithms that compute distances between features are biased towards numerically larger values if the data is not scaled, hence it is necessary.

Normalization transforms variables to be within a specific range [0,1]. Subtracting min and dividing with max-min

Standardization transforms features by subtracting the mean and dividing by the standard deviation.

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

it indicates a perfect correlation between two independent variables. Some variables in your dataset might create perfect multiple regressions on other variables, leading to infinite VIF values.

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

A Q-Q plot, short for quantile-quantile plot, is a type of plot that we can use to determine whether or not the residuals of a model follow a normal distribution. If the points form a straight line roughly then we can assume it is normally distributed. Q-Q plots are used to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.