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?

Answer:

1. These things negatively impact the sales:
    a. Windspeed
    b. Spring
    c. December
    d. November
    e. Wednesday
    f. Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    g. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
2. Below things Positively impact the sales:
    a. Workingday
    b. March and September may be because weather is good in this month
    c. Monday has positive impact on sales because this is the start of the week we can consider
3. The sale increased in 2019

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

Answer:

Because the first variable is redundant and the same information can be conveyed with remaining variables

If the number of columns are greater than 1 then we can remove the first.

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

Answer:

Registered has the highest corelation with cnt which is 0.95(95%)

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

Answer:

a. There is Linear relationship between X and Y
b. Error terms are normally distributed
c. Error terms are independent of each other
d. Error terms have constant variance(homoscedasticity)

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    a. Yr which is year where bike rental increased in 2019 compared to 2018

b. Workingday : On work day it has been observed that business increases
c. Windspeed: Windspeed negatively impact on rental, as the wind increases bike usage decreases

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

a. Linear regression is to identify the relationship between dependant and independent variable
b. Y=mx+c is the line equation which is used in this
c. We can use python stats model or sklearn to find out the relationship between multiple variables with independent variable
d. Steps:
    i. Clean the data
    ii. Create dummy variables if necessary
    iii. Scale the variables
    iv. Split the data into training and testing purpose
    v. Add constant
    vi. Create the model by using OLS
    vii. Check the summary of model
    viii. If the P value is greater than 0.05 remove the variable and rerun the steps

2. Explain the Anscombe's quarter in detail:

Answer:

a. It says that dataset have close to identical mean, variance, correlation and linear regression result but it appear very different when graphed. Hence visualization is important to identify patterns.

3. What is Pearson's R ?

Answer:

Person's R also known as Pearson correlation coefficient, measures the linear correlation between two variables. It ranges from -1 to 1. Where -1 Means negative linear relationship, 0 means no linear relationship and 1 means positive linear relationship.

R= Summession(xi-x)2 Summesion (yi-y)2 Summesion (xi-x)(yi-y)

4. What is scaling ? Why is scaling performed ?

   Answer: Scaling is adjusting the range of features in data. It is performed to make sure that each feature contributes equally to model.

   Normalized Scaling: Adjusts the data to a range of [0,1] using Min-Max scaling.
   Standardized Scaling: Adjusts the data to have a mean of 0 and a standard deviation of 1 using z score normalization

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

   Answer: VIF become inifinite when there is perfect multicollinearity, one predictor variable is a perfect linear comination of one or more other predictors. This results in a denominator of zero so it becomes infinite.

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

Answer: A Quanitile-Quantile plot is a graphical tool to assess if a dataset follows a specified distribution, typically the normal distribution. It plots the quantiles of the data against the quantiles of the theoretical distribution. In linear regression. Q-Q plots are used to check the normality of residuals, which is an important assumption for the validity of statistical tests and confidence intervals.