Assignment-Based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans. 1.More bikes are rented in fall and summer seasons.

2. More bikes are rented from May to October months

3. More number of bikes were rented in 2019 than in 2018.

4. Sales is more in Saturday and Wednesday

5. As temp increases the bikes selling are more . as we saw ppl are taking more bikes in summer.

6. People are taking more bikes when it is not a holiday.

7. Most of the people are biking on working day.

8. If the weather is clear , then more bikes are rented. No bikes are rented in heavy weather.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans. It is important to use it because for example if we create a **dummy variable** for **'Gender'.** So it will be 1 for 'Male' and 0 for 'Female' , here get_dummies will create a 2 cols like 'Male' and 'Female' and it will have according to gender wherever it's a male it will be 1 and for female it will be 0 so here the model will understand wherever it is 0 it's a male , so now we don't need female columns . Just male is able to explain both the columns by 0 and 1. Now the model can understand that 0 is ' Female' and 1 is ' Male' . So basically to avoid complexity we use drop_first= True.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. We observed that temperature has the highest correlation with the target variable ie . count.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans. We validated the assumptions of Linear Regression by performing **' Residual Analysis',** where we observed that the if the error is distributed normally or not. It's the analysis done for checking the error distribution.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans. Yr ,temperature,hum are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

Q1 Explain the linear regression algorithm in detail**.**

Ans. Linear Regression is a machine learning algorithm based on supervised learning. Its having dependent and independent feature , where in dependent feature is our target variables on which we are making predictions . Target Variable is continuous in Linear regression. For eg. How the treatmenr cost of the patients varies based on BMI and Based on CGPA predicting the salary .It has 2 types :

1. **Simple Linear Regression:** Wherein we have only 1 independent and 1 dependent features. By which we get to know about the relationship between those 2 variables. .
   Equation : y=beta0+beta1.xi +e
   Where beta0,beta1 are the regression coefficients

2. **Multiple Linear Regression:** Here we have Multiple independent variables and 1 dependent variables . Here we make predictions and tells which variables are significant with the target variable.

Q2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet has four datasets that have nearly identical in descriptive statistics but have very different distributions.

Q3. What is Pearson's R?

Ans. It determines how much of the total variation in dependent variable is explained by the variation in independent variables. It is the measure of the strength of the relationship between the variable also among the variables.

1 implies a good relationship.

-1 indicates a negative correlation

0 implies no correlations

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. Scaling is a technique which is done for normalizing the data . It is done because some of the independent features may not have the same units like somewhere the amount will be in rupees , for other variable it could be pound and for some it could be in dollars so which may lead to a wrong interpretations . To avoid this conflict we scale so that all have the units. When data is not normally distribution we use the Normalization technique so as to normalise the data, wherein standardisation is used when data is normally distributed.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans.If the VIF is infinite this clearly shows that there is a absolute correlation between the independent variables among themselves. I got inf VIF for Holiday, weekday_Sunday, workingday. This shows that these variables are perfectly correlated with the independent variables.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans. It shows the distribution of the dataset. It shows that if the data follows a normal curve and tells about the distribution of the data