**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? (3 marks)

Below are the inferences for categorical variables.

1. On a holiday, demand decreases.
2. Fall and spring seasons have the highest and lowest bike rental bookings.
3. Demand increases each month from Jan to June. The highest demand is in the month of September. Demand starts decreasing after September month.
4. Demand is more when whether is Clear. It can be seen from last chart.

Above findings can be verified from below chart.

A group of graphs with different colored squares

Description automatically generated with medium confidence

1. Why is it important to use **drop\_first=True** during dummy variable creation? (2 mark)

Dummy variables are created when we have categorical variables in our dataset. These dummy variables convert the categorical values into 0 and 1 values such that if there are n types of values in a single categorical column it will create n-1 variables. Drop\_first=True instructs runtime to delete the nth variable when creating dummy variable such that final output is n-1.

A very common example is to explain the flat type column. Suppose you have a housing dataset where you have flattype column with 3 values furnished, semifurnished and unfurnished. In this case when we create dummy variables it will create n-1 i.e. 3-1 = 2 variables with below values which would be like below

|  |  |
| --- | --- |
| Furnished | semifurnished |
| 1 | 0 |
| 0 | 1 |
| 0 | 0 |

When both columns have zero value, it means flat is unfurnished.

1. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
2. Variables having highest correlation with cnt variable are Temp and atemp
3. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

By using below chart. We can see that the error terms are centered around zero. Using VIF and p-values we verified assumptions of Linear Regression. P-value greater than 0.05 indicates that the test result is non-significant. Verifying that there is linear relationship between independent and dependent variables.

A graph of error terms

Description automatically generated

1. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

The top 3 features contributing significantly to explaining the demand of the shared bikes are season, weather, and temperature.

From the charts we can see that the Fall season has the highest demand for rental bikes.

From the charts we can see that the demand for rental bikes increases when the weather is clear.

As temperatures increase so does the demand for bike increases.

**General Subjective Questions**

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression algorithm explains a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

Independent variable is the cause. Its value is independent of other variables. Dependent variable is the effect. Its value depends on changes in the independent variable. When there is only one independent feature, it is known as Simple Linear Regression, and when there is more than one feature, it is known as Multiple Linear Regression.

Linear regression algorithm is a type of supervised machine learning algorithm because the algorithm learns through the training data. Training data is past data with labels is used for building the model.

E.g where we use Linear regression is as follows.

a) predicting house prices based on features like size, location, and number of bedrooms.

b) predicting the sales of the products.

1. Explain the Anscombe’s quartet in detail. (3 marks)

3. What is Pearson’s R? (3 marks)

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

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

(3 marks)

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

(3 marks)