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)

Answer :

1.1. Categorical variable such as Season has got 4 values, out of which we see the value Summer & Winter is having linear relationship with the dependent variable ant which is a target variable.

The coefficients for season summer with cat is 0.1050 whereas for winter it is 0.1467.

1.2. Categorical variable mnth which is month of a year is also important in the analysis. We see that

The coefficients of rental of bike is 0.0548 for Aug month and 0.1186 for Sept month

1.3. Categorical variable yr which shows two years 2018 & 2019 shows that rentals of bikes went up in year 2019

And having coefficients of 0.2285 with year value 2019

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

Answer: For any categorical variable having n possible values, it is advisable to create n-1 dummy variables.

The pd.get_dummies(categorical_column_name) will by default create n dummy variables, hence drop_first=True will create n-1 dummy variables.

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

with the target variable? (1 mark)

Answer : temp

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer : Plotting the residuals(which is y_train - y_train_predicted) in a histogram. As per the assumption of linear regression model, the error terms of a linear regression model are normally distributed with mean value of 0

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

Answer: temp, winter & yr 2019 are the top 3 contributors having positive coefficients to the demand of shared bikes

General Subjective Questions

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

Answer: Linear Regression algorithm provides a linear relationship between set of independent and dependent variables.

Using the linear regression algorithm we can predict the outcome of future events, it gives us a mathematical relationship between independent variables and dependent variables and helps us determine continuous variable which is a dependent target variable. In the assignment we can predict that

- Demand for sharing of bikes goes up during winter and summer
- Demand for sharing of bikes is higher in the month of August and September.
- Demand for sharing of bikes goes low when the weather is cold and windy
- More people rent the bikes during the summer season.

Using above information, the management of the bike sharing company can made decisions to increase the demand for rentals of bike(dependent target variable) by increasing their spend on marketing during summer season.

2. Explain the Anscombe's quartet in detail. (3 marks)
Answer : Anscombe's quartet is a group of dataset that has same mean, standard deviation and same regression line when plotted graphically but the actual dataset is different quantitatively. Anscombe's quartet is used to illustrate the importance of looking at a dataset graphically.

3. What is Pearson's R? (3 marks)
Answer: Pearson's R is a correlation coefficient between two datasets that measures the linear correlation between two sets of data. It is the ratio of covariance of the two variables and product of their standard deviation.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling
and standardized scaling? (3 marks)
Answer: Goal of scaling is to bring all the parameters of a dataset into same scale before building the data model. The importance of this is to reduce the time required for gradient decent method which goes into the background of model building. If we have a dataset of salary of person which is in thousand units and years of employment of employee which is between number 1-50. If the goal is to find out the whether the individual will default the loan by using these two parameters then bringing them into the same scale is important. Scaling changes the coefficient of the parameters in a regression model. The difference between normalised scaling and standardised is that normalised scaling brings the values of a variable between range 0 and 1 whereas standardised scales the data to have mean value of 0 and standard deviation of 1

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Answer: In a scenario where two independent variables are having perfect correlation with each other, VIF of those variables comes to be infinite.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Answer: A quantile-quantile plot, is a graphical tool which helps us decide if the a set of data comes from same distribution such as normal distribution or uniform distribution.
It helps asses if two data sets
- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes