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. temp: A coefficient value of 0.4777 indicated that a unit increase in temp variable, increases the bike hire numbers by 0.4722 units.
2. weathersit_rain: A coefficient value of -0.2908 indicated that a unit increase in Weathersit rain variable, decreases the bike hire numbers by 0.2908 units.
3. yr: A coefficient value of 0.2341 indicated that a unit increase in yr variable,increases the bike hire numbers by 0.2341 units
4. season_winter: A coefficient value of 0.0945 indicated that,a unit increase in season_winter variable increases the bike hire numbers by 0.0784 units.
5. windspeed: A coefficient value of -0.1481 indicated that,a unit increase in windspeed variable decreases the bike hire numbers by 0.1563 units.

Essentially we can conclude that with a positive coefficient value the target variable would increase and decrease with an increase in the value of the negative coefficient dependent variable.

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

Answer: Its important to use drop_first = true as it does not create an extra column while dummy variable creation. In a way this is important to reduce the correlations in the 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 has the highest correlation with the target variable cnt

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

Answer: The assumptions of the linear regression were validated based on the below 2 methods:

a. The R Squared value was 0.83 which essentially meant that the model can predict only 81% of the times.
b. Using residual analysis which proved that the errors are normally distributed with mean 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: Based on final model below are the three features that contribute towards explaining the demand of the shared bikes :

a. temp: A coefficient value of 0.4777 indicated that a unit increase in temp variable, increases the bike hire numbers by 0.4722 units.

b. weathersit_rain: A coefficient value of -0.2908 indicated that, a unit increase in Weathersit rain variable, decreases the bike hire numbers by 0.2908 units.

c. yr: A coefficient value of 0.2341 indicated that a unit increase in yr variable, increases the bike hire numbers by 0.2341 units.

General Subjective Questions

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

Answer: Linear regression algorithm is an algorithm which falls under supervised learning. In this technique the model is trained to predict the behavior based on some variables. There are 2 main entities- Dependent variables and target variable. The 2 variables need to be linearly correlated to predict the behavior. Linear regression is based on the below formula –

$$y = mx+C$$

here m is the slope and c is the intercept. Essentially we are trying to predict the quantitative behavior of y based on the predictor x. As an example, linear regression algorithm can be used to help predict the sales target of a company (y) based on some independent variables (x)

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer: Anscombe's quartet are a set of 4 data sets containing 11 data points (x,y). This quartet resembles the importance of data visualization. The main point in the quartet is that the data points share the same mean, variance, std dev etc. but appear differently when plotted on a graph. This proves the importance of visualizing the data before concluding.

3. What is Pearson's R? (3 marks)

Answer: Pearson's R is the covariance of 2 variables divided by the product of their standard deviation. This means that Pearson's R would always lie in between -1 and 1. The value depicts whether the data is perfectly linear with positive slope (1) or perfectly linear with negative slope (-1)

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

Answer: Scaling is a technique to handle the standardization of independent features present in the data. The whole idea of standardizing the data is to make sure that the machine learning algorithm can work with smaller set of values and thus predict more accurately. This is usually handled during the pre processing of data.

Standardized scaling used the mean and standard deviation for scaling and is not bounded to any range. However, normalized scaling uses minimum and maximum value of the feature and is scales in a bounded range. Another key difference between standardized and normalized scaling is that the standardized scaling is not affected by the outliers but normalized scaling is hugely affected by the outliers.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: VIF is Variance inflation factor. This essentially detects multicollinearity in linear regression which means it can detect the correlation between the predictors in the model. An infinite value of VIF indicates a perfect correlation amongst the predictors and one of the variables need to be dropped to handle this scenario.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: Q-Q plot is a quantile- quantile plot which determines the relation between 2 sample data sets by plotting the quantiles of both the data sets.

Q-Q plot is an important plot which can determine if 2 data sets come from the same population or not. It can also recognize if 2 samples have the same common location behavior or the same distribution set. To do a Q-Q plot the data must be sorted and a normal distribution curve needs to be drawn. At the end, the z value needs to be calculated for each segment and the data set values need to be plotted against the cut off points.