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 – [a]As compared to year 2018 in 2019 more booking took place.

[b]In year 2008 and 2019 from month may to sept the highest number of booking took place.

[c] Highest number of booking occur when there is Clear weathersit followed by Mist+Few clouds, Light snow.

[d] Highest number of booking occur when there is Fall season followed by summer, winter and spring.

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

Ans – The drop_first=True is important to use, because it helps in reducing the additional column created during dummy variable creation. Hence it reduces the correlations created with dummy variables. Also we did not need dummies to explain categorical variable. In addition to it we will also n-1 columns to represent dummy.

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

Ans - According to analysis temp and atemp have a highest correlation with the target varaiable.

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

Ans – [a] After building a model we will validate the model with the help of scatter plot we can get idea there is linear relationship between X and Y are not.

[b] With the help of distplot we can get are the errors terms are normally distributed or not.

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

Ans - As per model following results has been seen

1. Year (yr):- A coefficient value of yr indicated that a year wise the rental numbers are increasing.
2. Temperature (temp):- A coefficient value of temp indicates that temperature has significant impact on bike bookings and it is an imp parameter for building a model.
3. Light snow :- A coefficient value of Light snow indicates there will we the decrease in bike bookings if there is any rain or snowfall or due to bad weather condition.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans – [a] Linear regression is used to predict upcoming trend based on based data.

[b] The linear regression are basically divided into two types Simple linear regression and Multiple linear regression.

[c] Model With only one independent variable is called as Simple linear regression while model with more than 1 one independent variable is called as Multiple linear regression.

[d] Equation for best fit regression for Simple linear regression is given by

$$Y = B0 + B1*x$$

[e] Equation for best fit regression for Simple linear regression is given by

Y = B0 + B1*x1 + B2*X2 + B3*X3….. +Bn*Xn

[f] Assumption in linear regression algorithm are:

1. There is Linear relationship between X & Y.
2. Error terms should not have constant variance (Homoscedasticity)
3. Error terms are independent to each other.
4. Error terms are normally distributed with mean zero (not x, y).

Q2. Explain the Anscombe's quartet in detail.

Ans - Anscombe's Quartet are often defined as a group of 4 data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities within the dataset that fools the regression model if built. They need very different distributions and appear differently when plotted on scatter plots.

Q3. What is Pearson's R?

Ans – The Pearson's R correlation coefficient it helps us to find out relationship between two variable. It basically gives us measure of strength which are associated with two variable. The value of Pearson's R lies in range of -1 to 1, whereas 0 means having no correlation and 1 means having highly correlated.

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

Ans – [a] Scaling in linear regression is defined as method of normalising the range of independent variable.

[b] The requirement of Scaling occurs when we are dealing with Gradient decent and distance based algorithm as because these are having range of data points which are very sensitive.

[c] In normalised scaling generally values are rescales in range of 0 to 1, meanwhile in standardized scaling values are rescale to have a standard deviation as 1 and mean as 0.

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

Ans – If value of VIF is infinity it indicate that the it has perfect correlation between two independent variable. When value of R2 is equal to 1 then we will get the value of VIF as infinity by the formula (1/(1-R2)). To overcome this problem we need to drop one of variable from data set which gives ideal multi correlation.

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

Ans- [a] Q-Q plot is defined as graphical representation for determining if two data sets come from population with a common distribution

[b] The use and importance of Q-Q Plot is that a quantile may be the fraction where certain values fall below quantile.

[c] The purpose of Q-Q plots is to seek out out if two sets of knowledge come from an equivalent distribution.

[d] A 45 deg angle is plotted on the Q-Q plot; if the 2 data sets come from a standard distribution.