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?

Ans. In the Bike Sharing Dataset, categorical variables are as follows:

1. In season feature, Fall turns out to be the season with highest bike rentals count and sprint with lowest bike rentals count.

2. In Year feature, 2019 saw the highest bike rentals count.

3. In Month feature, For Jan, Feb count was low. It started increasing from March, was stable around June to August and was highest in the month of September.

4. In weekday feature, there is not much difference in the 50% value of bike rentals count for weekdays.

5. In Weather feature, Clear weather was preferred by the customer.

6. In working Day Feature, there is no much difference in count as the 50% values are almost same for working day as well as holiday.

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

Ans- In dummy variable creation, it is the basic rule that if there are n categorical variables, there should be (n-1) Dummy variables. So, if a categorical feature has 4 possible values, then for getting a stable model 4-1 = 3 dummy variables will be sufficient to get those 4 values.

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

Ans. Among all the numerical variables, 'registered' has the highest correlation with 'cnt'. According to my model, Temperature has the highest correlation.

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

Ans. 1. A Linear Relationship between the dependent variable and their predictors can be justified using Pair Plots.

2. If error terms form a normal distribution, it validates the training model.

3. VIF (Variance Inflation Factor) value for all the feature should not be greater than 5.

4. P-values for all the features should be less than 0.05(assumed) significance level.

5. Error terms should be independent of each other and for that Durbin – Watson (DW) statistic can be used.

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

Ans. After removing highly collinear and unwanted features these are the 3 features which are significantly contributing to the model are:

1. Temperature

2. Year

3. Winter (Season)

For all these 3 features value of coefficient is positive and has VIF values less than 5.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans- Regression helps us to determine the strength of the Relationship between one dependent Variable and one or more independent variables. 'Linear' in Linear Regression here doesn't mean a perfect straight line but it means linearity in the parameters. Dependent Variable is that Variable on which modal has to be build. Dependent variable is also known as the target variable. Independent variable is known as the Predictor variable.

Some important things to note:

1. Target variable that has to be predicted must be a numerical/continues one.

2. Linear Regression comes under Supervised learning method because labels are present.

Formula: $Y= a + bX$

Where, Y=Dependent variable,

X=Independent variable,

b=slope,

a= constant

Types of Linear Regression:

1. Simple Linear Regression: Here predictor variable is one.

$$y = B0 + B1*X + E$$

B0 – Intercept

B1 – Slope

E – Error Term

2. Multiple Linear Regression: Here predictor variable can be more than one.

$$Y = B0 + B1X1 + B2X2 + ... + BnXn + E$$

Assumptions in Linear Regression:

1. There must be Linear Relationship between the dependent variable and their predictors.

2. Error Terms are independent of each other's

3. Absence of Multicollinearity

4. Homoscedasticity: Error terms have constant Variance

5. Error terms are normally distributed with mean =0

2. Explain the Anscombe's quartet in detail.

Ans- Anscombe's quartet comprises of four datasets that have approximately identical statistics, yet have very different distributions and appear very different when plotted on graph. Anscombe's quartet was discovered by Francis John "Frank" Anscombe in 1973. He plotted 4 datasets with 11 Datapoints each on 4 different graphs and he observed that all the four datasets have same summary statistics but have different graphs. The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R?

Ans- Pearson's R is the correlation coefficient which is named after Karl Pearson. In simple words, Pearson's R is the covariance of 2 variables divided by product of their standard deviation. The value of the Pearson 's R lies between -1.0 to +1.0. Negative Pearson's R means if the value of one variable increases, the other variable value decreases proportionately i.e., Inversely proportional and vice versa for positive value.

Formula

$$r= N\Sigma xy-(\Sigma x)(\Sigma y)/root([N\Sigma x^2 - (\Sigma x)^2] [N\Sigma y^2 - (\Sigma y)^2])$$

Where, N = the number of pairs

Σxy = the sum of the products

Σx = the sum of x

Σy = the sum of y

Σx2 = the sum of squared x

Σy2 = the sum of squared y

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

Ans- Scaling is a pre - processing step which is taken to scale the value of the data within specified range to improve the accuracy of the model. The scaling is optional in case of simple linear regression while it must be compulsory for multiple linear regression modals.

The reason why we do scaling can be explained by an example: -

Let's suppose, if an algorithm is not using the feature scaling method, then it can consider the value 3000 meters to be greater than 5 km but that's actually not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to the same magnitudes and thus, tackle this issue.

Normalised scaling is also known as Min-Max scaling. Here scaling is done in such a way so that all the values of all the variables lie between 0 to 1. Minimum and maximum value of features are used for scaling in normalization.

$$(X-Xmin)/(Xmax-Xmin)$$

Standardized Scaling is the type where Mean and standard deviation is used for scaling. In this mean is 0 and sigma=1.

(X-Xmean)/std. deviation

Here, the difference between two the two techniques is first, Normalised scaling handles outliers while standardized one can't handle. Second, normalization is used when features are of different scales and Standardization is used when we want to ensure zero mean and unit standard deviation.

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

Ans- Variance Inflation Factor (VIF) is the method that shows the relationship of one independent variable with another independent variable. For a particular variable, if VIF is high then it means it has high association with other variable and vice versa.

When R=1, VIF = 1/(1-r^2) = Infinite, which means one independent variable is perfectly Correlated with another variable. This leads to multicollinearity.

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

Ans-Q-Q plot also known as Quantile-Quantile plot. Q-Q plot is the graphical technique which is used to check whether two datasets that are coming from the same Population have same distribution or not. The major advantage of the Q-Q plot is to validate the assumptions of the multiple linear regression which is Error Terms should be normally distributed. If the plot shows straight line, then it is normally distributed otherwise not. Q-Q plots are also used to find the skewness of the distribution. It can be Left Skewed or right Skewed.