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?

Bike demand is high during May, June, July, August, September, and October. Bike demand is high in 2019 than 2018.Bike demand is similar in weekdays and does not change if it is working or not working.

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

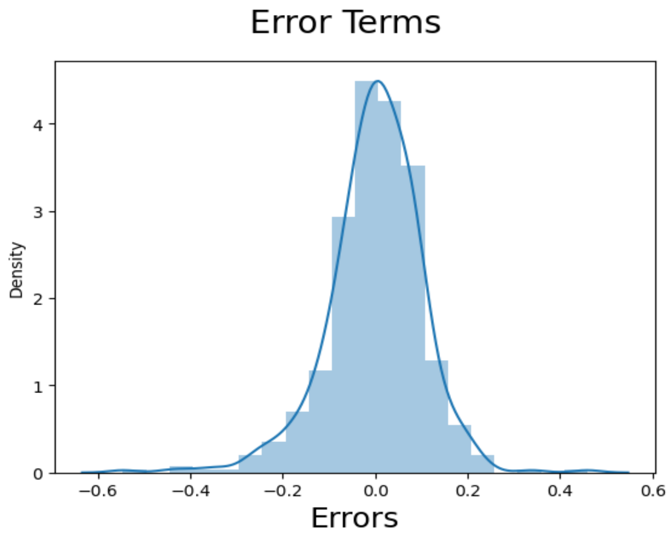
To reduce the collinearity between dummy variables and to achieve K-1 dummy variables which helps to delete the extra column while creating dummy variables.

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

atemp and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

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

We validate the assumptions of Linear Regression by using distplot of the residuals and also to validate if the normal distribution mean = 0.



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

The Top 3 features contributing significantly towards the demands of shared bikes are Weathersit_light_snow (Negative correlation), Yr_2019(Positive correlation, temp(positive correlation

GENERAL SUBJECTIVE QUESTIONS

1.Explain the linear regression algorithm in detail?

A Linear Algorithm explains the relationship between independent and dependent variables using a straight line. Applicable only to numerical variables.

>> The Dataset is divided into test and train dataset

>> Train Dataset is divided into feature(dependent variable) and target(independent variable)

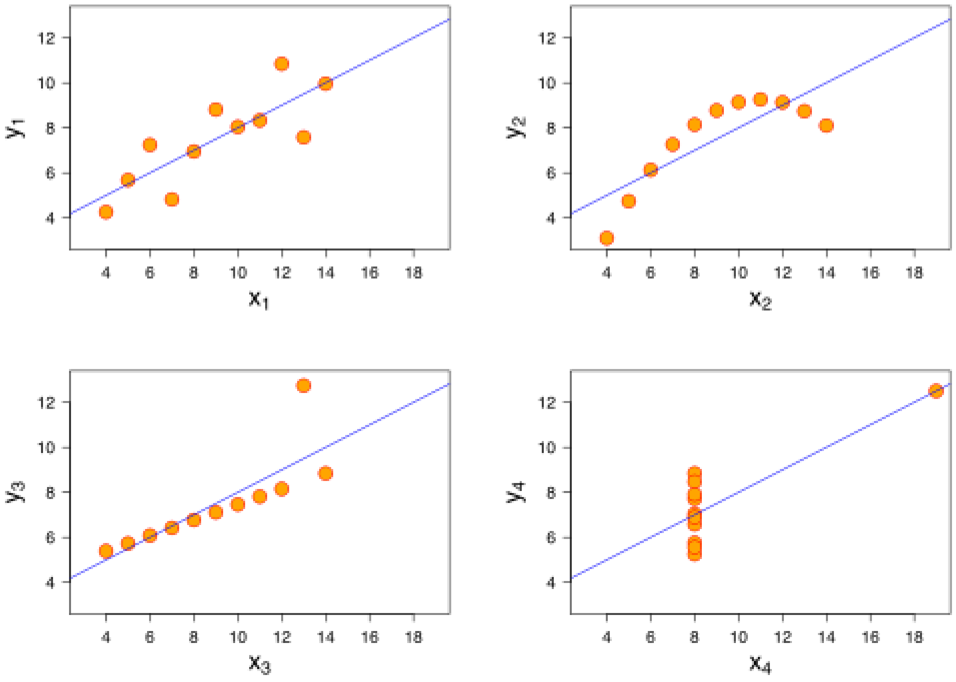
>> A Linear model is fitted using Training dataset. The gradient descent algorithm works by minimizing cost function. Example of cost function is residual sum of squares.

>> In case of multiple features, the predicted variable is a hyperplane instead of line. The predicted variable takes the following form

$Y = \beta 0+ \beta 1\chi 1+\beta 2\chi 2+\beta 3\chi 3+\ldots+\beta n\chi n$

Finally, the predicted variable is compared with the test data for final outcome.

2.Explain the Anscombe's quartet in detail?



Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.

There are these four data set plots which have nearly same statistical observations, which provides same statistical information that
involves variance, and mean of all x,y points in all four datasets.

3.What is Pearson's R?

Pearson's R measures the strength of association of two variables. It is the covariance of two variables divided by product of their Standard deviation . It has a value from +1 to -1.

1------→ Positive correlation,if one variable increases other will also increase

0-----→ no correlation

-1----→ Negative correlation , if one variable increases other will decrease

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

*Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.*

Normalized Scaling	Standardized Scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1 - {R_i}^2}$$

Where, ‘i’ refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.