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?

Solution:

From the Plots drawn using the categorical columns we can infer the following:

1. In fall season the booking is maximum and the booking for each season has increased drastically in 2019
2. the booking increases from the start of the year and peaks from June to September and starts to decrease
3. When its holiday, the booking seems to be less.
4. Thu, Fir, Sat and Sun have a greater number of bookings as compared to the start of the week.
5. Booking seemed to be almost equal either on working day or non-working day.
6. The bookings are more when there is a clear weather which is as expected.
7. Overall, in general, we can see similar pattern of booking 2018 and 2019, but the booking count has increased in 2019, which is a god sign for business.

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

Solution:

drop_first=True helps in avoiding the redundancy in the dataframe, for any categorical variable with n type of values, it only needs n-1 dummy variables to depict all type of values, when all the dummy variables are zero it implies that the given row belongs to the dropped variable.

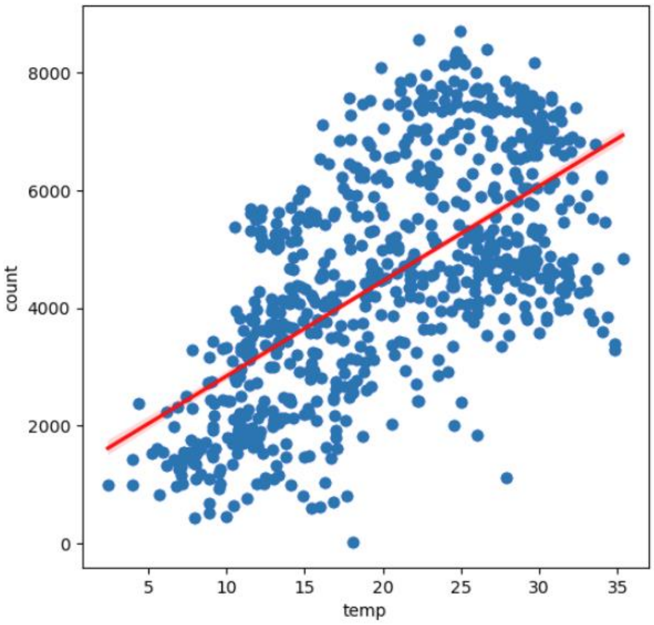
For example: if we have a variable with 3 types of value namely A , B, C and let's say we create n-1 dummy variables i.e. A and B. Now, when A is '1' and 'B' is '0', then the row has type 'A' and vice versa, but if both 'A' and 'B' are '0' then it will be 'C'. Thus, we need only 2 dummy variables to depict 3 types of values in the variable.

Hence **drop_first=True** plays an important role in maintaining an efficient dataset, and in reducing multicollinearity among the dummy variables.

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

Solution:

'temp' variable has the highest correlation with the target variable 'count'.

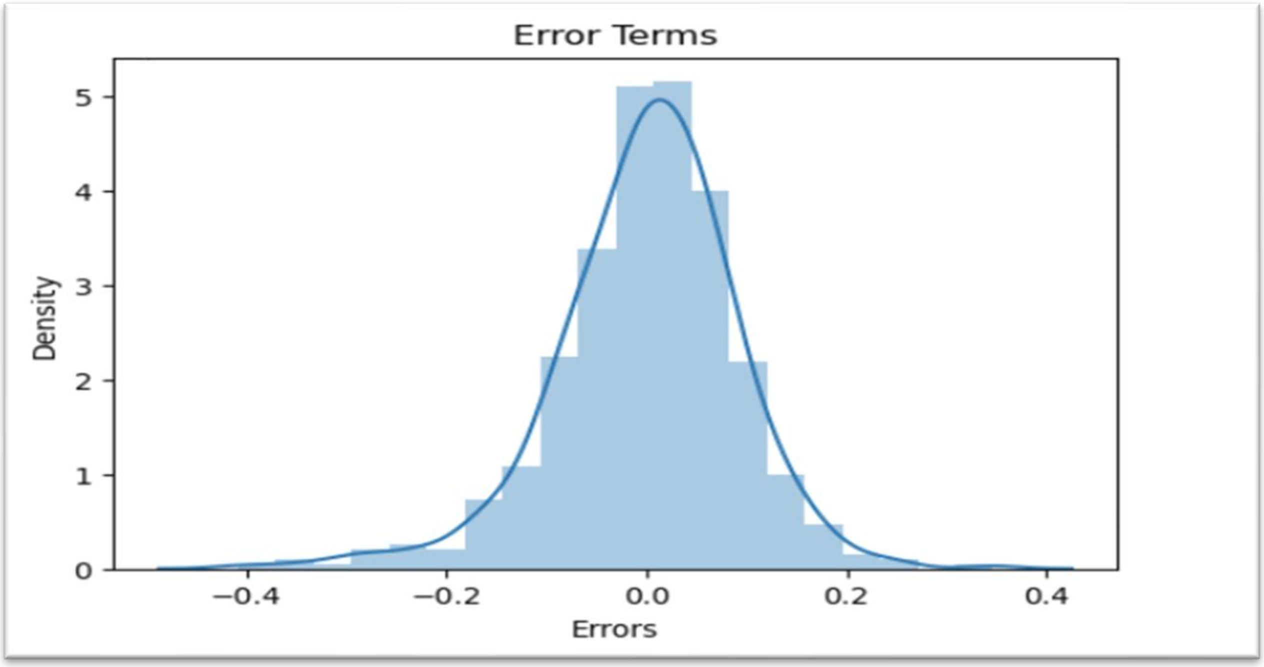


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

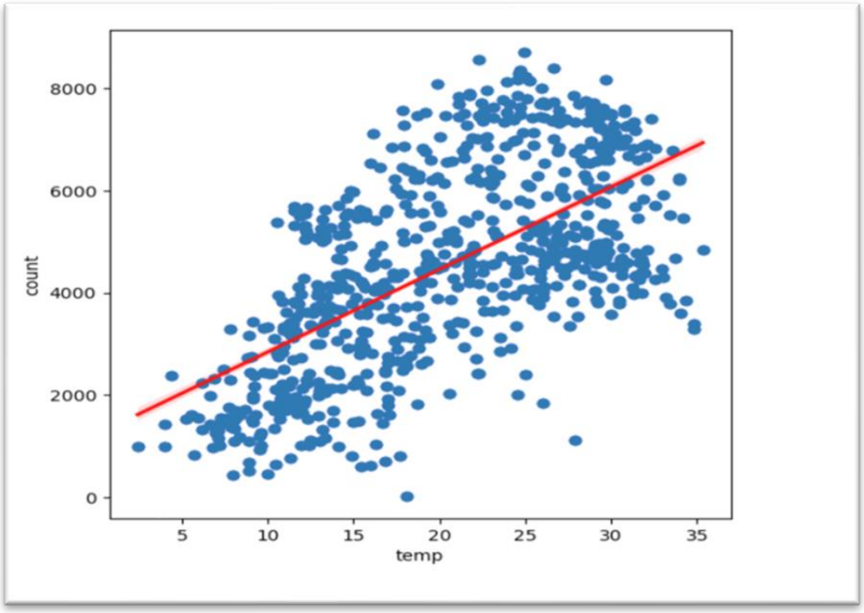
Solution:

The assumptions of Linear Regression were validated by plotting the charts and performing the calculation over the given model and resulting data. And they are as follows,

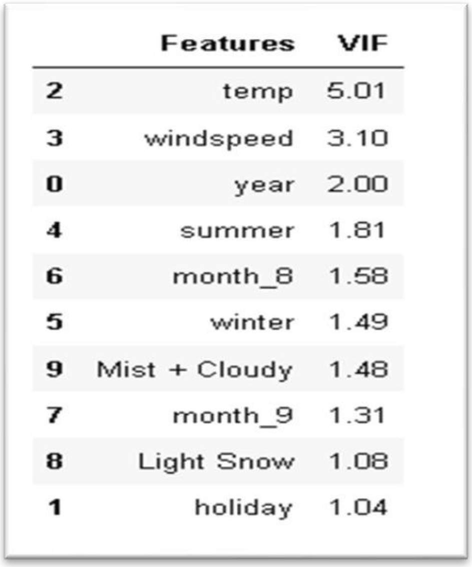
1) Normality of error terms – The error terms should be normally distributed with mean over 0.



2) Linearity among the variables – input and target variable display linearity among them.

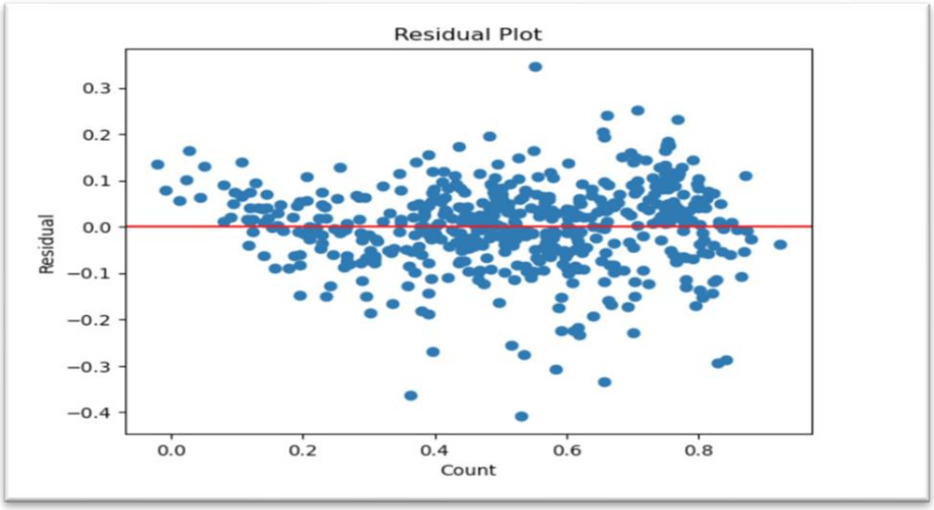


3) Multicollinearity – independent variables should not be highly correlated among themselves to avoid redundancy of one of the variables.

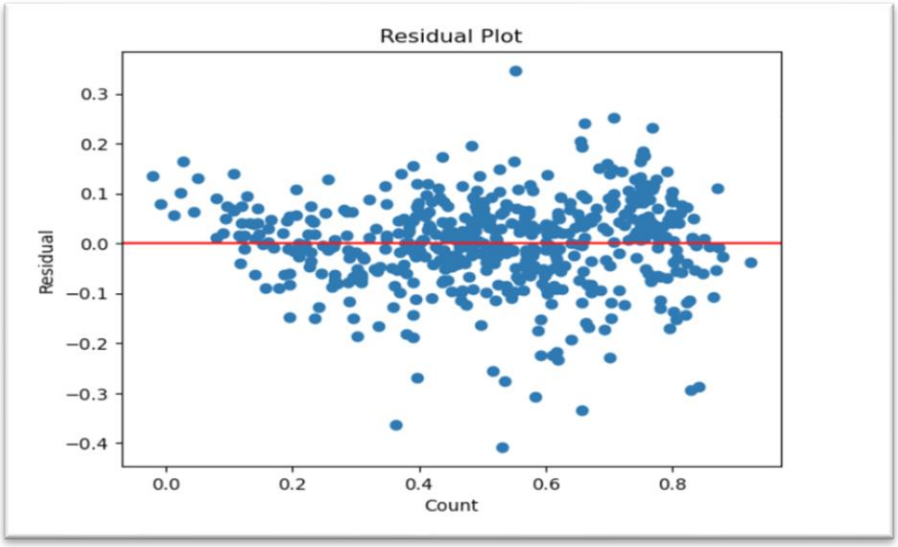


	Features	VIF
2	temp	5.01
3	windspeed	3.10
0	year	2.00
4	summer	1.81
6	month_8	1.58
5	winter	1.49
9	Mist + Cloudy	1.48
7	month_9	1.31
8	Light Snow	1.08
1	holiday	1.04

4) Homoscedasticity of the error terms – The variance should not increase or decrease for the error terms.



5) Error terms are independent of each other: The error terms should not be dependent on the previous error term.



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

Solution:

The top 3 features contributing significantly towards explaining the demand of the shared bikes are as follows:

1. Temp
2. Light Snow (weathersit – 3)
3. Year

General Subjective Questions:

1. **Explain the linear regression algorithm in detail.**

Solution:

Linear regression is a supervised machine learning algorithm where the past available data is used to identify the pattern between the input (one or more) and target variable. And such patterns are used in predicting the continuous target variable based on one or more input features.

There are two types of linear regression namely:
1. Simple Linear Regression: only one input variable
2. Multiple Linear Regression: two or more input variables

But the target variable in both the cases is always one.
In linear regression the goal is to find the best fitting line (or best fitting hyper plane in case of multiple linear regression) that minimises the sum of the squared difference between the actual value and the predicted value.

Linear regression involves the process of building the model and validating the model based on the assumption of linear regression,

1. Building the model involves the following steps:
    1. Data loading and understanding – loading and understanding each column of the data
    2. Pre-processing: Dropping the unnecessary columns and handling the missing values and potential outliers.
    3. Model selection: based on the number of input features (SLR or MLR)
    4. Training the model by selecting certain feature either automatically or manually upon which the model is modified multiple times based on the multicollinearity between the input variables and also significance of the model.
    5. The final model contains the variables with high significance (low p value) and low multicollinearity (VIF value) among themselves.
    6. This trained model is used to predict the values of the unseen data points.
    7. Model Evaluation: the model can evaluated using the R^2 , Mean Squared Error and Root Mean Squared Error methods.

2. Assumptions: There are certain assumption the are used to validate the Linear Regression Model, such as:

    - Normality of error terms – The error terms should be normally distributed with mean over 0.
    - Linearity among the variables – input and target variable display linearity among them.
    - Multicollinearity – independent variables should not be highly correlated among themselves to avoid redundancy of one of the variables.

- Homoscedasticity of the error terms – The variance should not increase or decrease for the error terms.
- Error terms are independent of each other: The error terms should not be dependent on the previous error term.

Using the above process and principle we will build a linear model which can be used to predict the values based on the given input features to a certain degree of accuracy.

2. Explain the Anscombe's quartet in detail.

Solution:

Anscombe's quartet is a set of four small dataset that have nearly identical statistical summary yet they appear different when they are visually represented. These datasets are often used to emphasise on the importance of data visualization and the limitation of the summary statistics in understanding the data.

The four datasets when visualised appear as explained below:

1) Dataset-1: This dataset contains a simple linear relationship between the independent and dependent variable

2) Dataset-2: There is a relation between the variables, but it is no more linear. It could be a quadratic relationship with some curvature

3) Dataset-3: This still has a linear relationship but the dataset contains an outlier which will influence the linear relationship and skews the summary statistics.

4) Dataset-4: In this dataset the data do not follow any pattern or relationship with dependent and independent variables.

The significance of the Anscombe's quartet is to show that the summary statistics often have the limitation when it comes to understanding the data and one should always make sure to visualize the data in order fully understand the before taking the decisions based on the data.

3. What is Pearson's R?

Solution:

Pearson's R or The Pearson correlation coefficient (r) is a measure of the strength and direction of the linear relationship between the two variables. Its value ranges from -1 to +1 and the meaning of the values is as below:

- Value +1 represent a strong positive correlation between the variables
- Value 0 indicates that there is no relationship between the variables
- Value -1 represent a strong negative correlation between the variables

- The values in between -1 and +1 indicate the relative strength of the relationship and the sign indicates their direction (positive or negative).

The formula for Pearson's r is:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where:

n is sample size,
x_i and y_i are the i^{th} sample point
$\bar{x}$ and $\bar{y}$ are sample mean.

Some of the key points of the Pearson's R:

- It only measures the linear relationships, if the relationship is non linear then Pearson's r may not capture it correctly
- It is sensitive to outliers, the outliers can either increase or decrease the correlation coefficient
- The r is symmetrical, i.e. coefficient between x and y is same as y and x
- The r value does not get affected due to scaling, hence the scale of the value can be changed as per convenience

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

Solution:

Scaling is a data preprocessing technique used in machine learning to bring the values of all the numerical continuous column in the same range. Scaling basically brings all the values to a closer range which is essential for some algorithms which are sensitive to the magnitude of data, such as gradient descent in linear regression.

Reasons for performing scaling:

- It is done to ensure that the data are in same range which results in the coefficients of the model also being in the same range.
- Scaling also helps in achieving the minimizing the cost function and training the model at a faster rate.
- Scaling also makes it easier to interpret the effect of different coefficients on the model, so that the business decision can be made easily.

Differences between normalizing and standardised scaling are as follows:

- The range in normalizing is typically between (0 and 1) where as in standardised scaling the values are centred around the mean '0' and standard deviation 1.

- Normalization scaling preserves the distribution pattern whereas in the standardization scaling the data is transformed into normal distribution
- Normalisation scaling are sensitive to outliers whereas standardisation scaling is not sensitive to outliers because it uses mean and standard deviation.

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

Solution:

VIF measures the collinearity of the regression model. Higher value represents the high collinearity of that feature with the rest of the feature.

As we all know that the formula for VIF,

$$VIF = 1/(1-R^2_j),$$

where R^2_j indicates the correlation of one feature with all the other features.

Now, when VIF becomes infinite, as per formula it means that the R^2_j is moving toward 1 or it is already 1.

In reality $R^2_j = 1$ means that there is a very high correlation between the given feature and all the other features. This indicates a very complex relationship among the features.

In this case the feature becomes relatively unimportant as the other feature are already accounting for its significance in the model and this particular feature will become a redundant feature for the model and should be dropped from the model.

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

Solution:

A Q–Q plot is a plot of the quantiles of two distributions against each other. It is use to check whether the dataset follows a particular theoretical distribution like normal distribution. It is a visual comparison of quantile of the observed data and quantiles of the expected theoretical values.

They are commonly used in hypothesis testing to make the important decision about the appropriateness of the chosen distribution.

Uses of Q-Q Plot in Linear Regression:

Assumption checking: Q-Q plot can be used to perform the residual analysis of the error terms and checked whether they are normally distributed or not with the mean of zero and homoscedasticity. As the deviation from the normality can affect the validity of the regression model

Importance of Q-Q Plot in linear regression:

1. If the Q-Q plot indicate that the residuals deviate from the normal distribution, then it indicates that the model can be improved.
2. If the points in the Q-Q plot deviate from the expected straight line, it may indicate that there could be skewness or outlier in the data
3. The normality of the residual is very important for the reliability and validity of the model. Q-Q plot will help in visualising the distribution of the residual upon which the model's reliability and the business decision depends on.