Assignment-based Subjective Questions

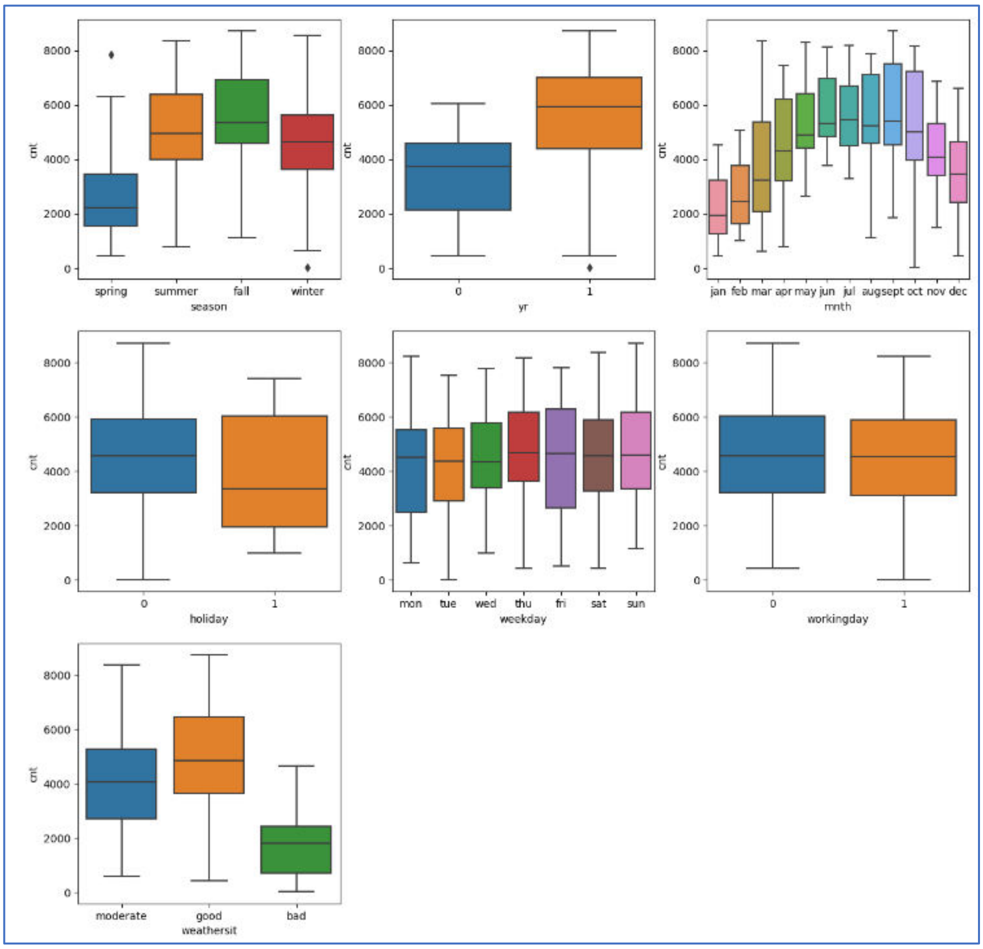
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:

To analyse effect of the categorical variables on the dependant variable we have plotted boxplot for each categorical variable against the target variable 'cnt'.

Inference:

1. Season Fall has maximum demand for shared bikes.
2. Year 2019 has more bike demands than 2018. It will boom each year it seems.
3. September month has the highest demand and the demand increases from Jan to Sep and then decreases.
4. When holiday is not there, shared bikes are requested more.
5. Good weather has maximum demand for the shared bikes. In moderate weather also demand is there compared to bad weather.
6. Whether it is working day or not, demand for the shared bikes are similar (not much difference).



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

Answer:

When you have a categorical variable with n levels(It has n type of values). The idea behind using the **drop_first=True** is to create n-1 dummies for that variable.

Let's take example of season. Season has 4 values (4 levels).

When creating dummies with using **drop_first=False :**

Season	Fall	Spring	Summer	Winter
Fall	1	0	0	0
Spring	0	1	0	0
Summer	0	0	1	0
Winter	0	0	0	1

When creating dummies with using **drop_first=True :**

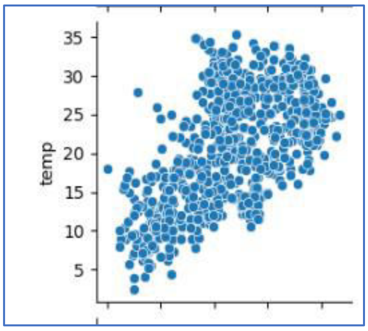
Season	Spring	Summer	Winter
Fall	0	0	0
Spring	1	0	0
Summer	0	1	0
Winter	0	0	1

Still Fall value can be predicted by '000' even if it is removed. This reduces the extra column and the correlations among 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:

Looking at the pair-plot among the numerical variables, 'temp' has the highest correlation with the target variable.

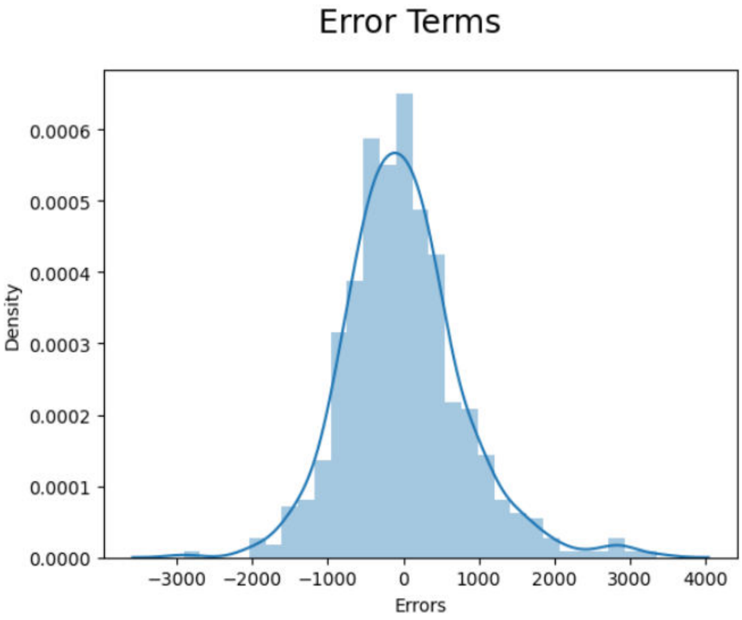


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

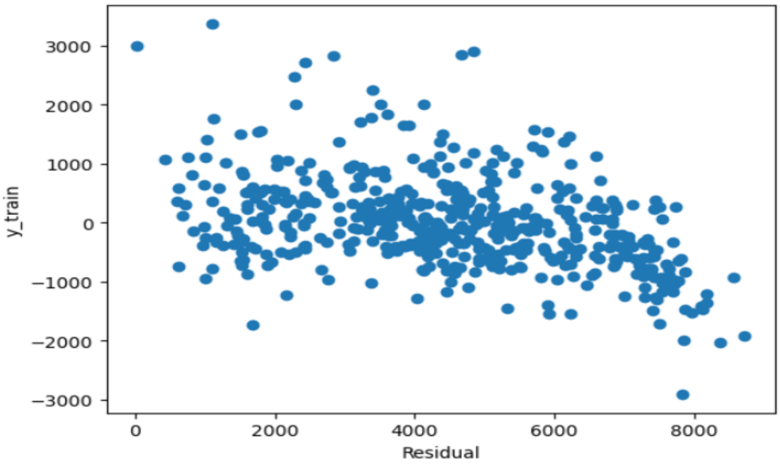
Answer:

We validated the below assumptions of Linear Regression after building the model on the training set.

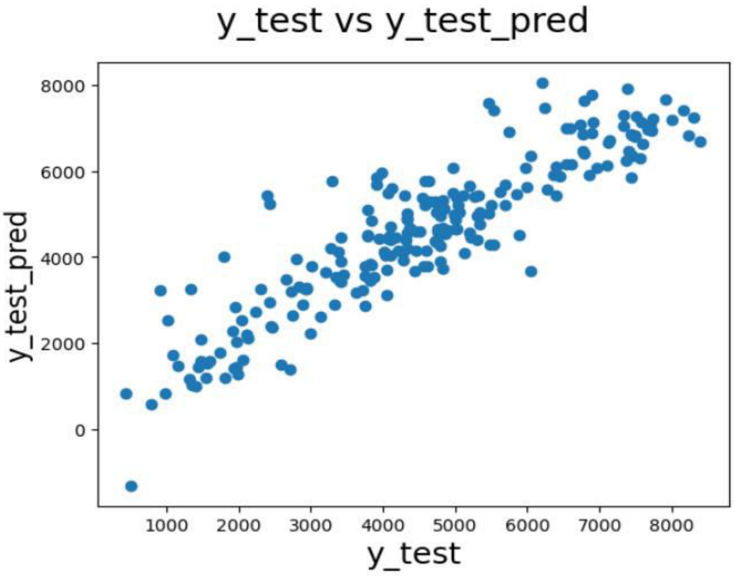
- Error terms are normally distributed with mean zero.



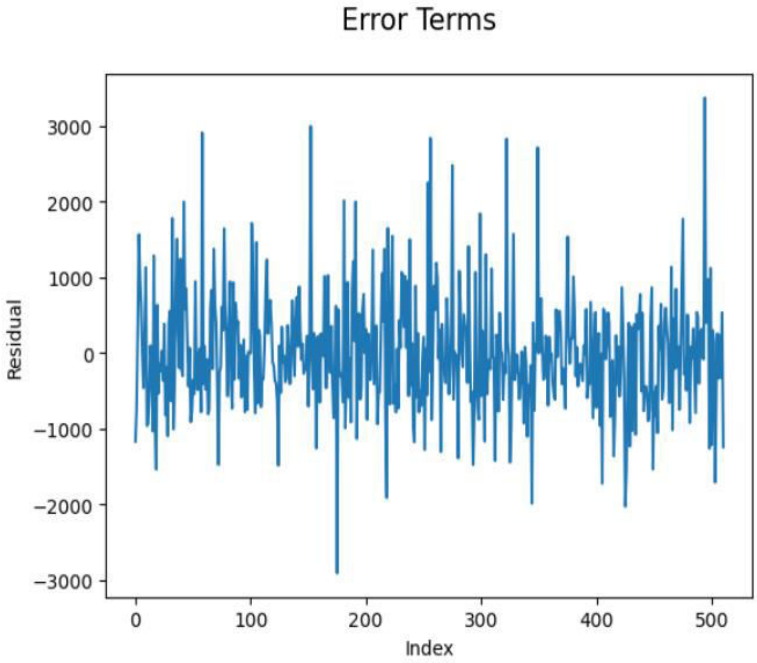
- Homoscedasticity (no visible pattern).



- Correlation among independent variables must be insignificant(multicollinearity). This is achieved by maintaining VIF’s value less than 5.
- Linear relationship must be validated on test sets as well.



- Error terms are independent.



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 the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes are :

- Temp
- Year
- Season_winter

General Subjective Questions

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

   Answer:

   Linear regression algorithm falls under supervised machine learning methods. A linear regression algorithm tries to explain the linear relationship between dependant and independent variables using a straight line. Equation of straight line is given by:

   Y=c+mX

   Y: Dependent variable.

   X: independent variable.

   m: slope of the line.

   C: constant

   We can say that if the value of x increases by one unit the value of y increases by m.

   We are assuming that the x and y has linear relationship. x and y can have positive or negative relationship among them.

   Linear regression is of two types:

   - Simple Linear Regression
   - Multiple Linear Regression

   In other words, regression is to find the best fit line. To achieve this, we follow certain steps:

   - We plot scatter plot to check the relationship between X and Y.
   - We calculate R-square, adjusted R-square(in case of multiple linear regression) and residuals for any line passing through scatter plot.
   - We find the equation of the best fit line and optimal value for m and c.
   - The equation of best fit line can be found by minimizing the cost function.
   - Strength of the linear regression model is explained by R-square.

There some assumptions for linear regression:

- Multicollinearity: When the independent variables are dependent on each other. Independent variables correlation must be insignificant or small.
- Error terms are normally distributed with mean zero.
- X and Y must have linear relationship.
- Error terms must be independent.
- Error terms must not show any visible patterns.

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

Answer:

Anscombe's quartet was developed by Francis Anscombe. It illustrates us about the power of visualisation and why we must use it every day. Sometimes numerically all datasets seem to be same, but when they are plotted as graph tells entirely a different story.

Anscombe's quartet contains 4 datasets, each with eleven (x ,y) values. The interesting thing about it is that all below parameters are same for all four dataset:

- Mean of x and mean of y is same for all the four datasets.
- Variance of x and variance of y is same for all the four datasets.
- Correlation coefficient for x and y is same in all four datasets.

After plotting the eleven points in graph all four datasets showed different picture.

- Dataset-1: Perfect well fit linear model.
- Dataset-2: Showing normal distribution like curvature.
- Dataset-3: Perfect linear distribution with extreme outlier.
- Dataset-4: all points lying parallel to y-axis with one outlier.

Conclusion: We must not only rely on numerical data ,visualizing is also important.

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

Answer:

Pearson's R indicates the strength of association between two variables. Value of the Pearson's correlation coefficient ranges from -1 to 1.

- When increase in the value of one variable increases the value of other variable, it is called positive correlation. Value of R will be greater than 0.
- When any change in value of variable one does not impact the value of second variable it is called no correlation. Value of R will 0.
- When increase in the value of one variable decreases the value of other variable, it is Called negative correlation. Value of R will be less than 0.

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

Answer:

Scaling is a way to frame the independent variables in same or similar range. Feature scaling is done to prohibit the dominating nature of the features with large values.

For example we have salary and experience of the employee. As salary can have big values as compared to the experience, so to avoid dealing with such large values (interpretation becomes difficult) we prefer to standardize or normalize the variable in preprocessing data phase.

Normalized scaling	**Standardized scaling**
Min and max values of variables are used for scaling.	Mean and sigma values are used for scaling.
Affected by outliers.	Not affected by outliers.
It is range bound like between 0 and 1.	It is not range bound.
We use MinMaxScalar for normalized scaling.	We use StandardScalar for standardized scaling.

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

Answer:

Within independent variables multicollinearity is detected by VIF values. High (infinite) VIF value indicates the significant or perfect multicollinearity. To solve this issue we must drop the independent variable causing multicollinearity.

Usually when R-square =1, then VIF=infinity

VIF=1/(1-R-square)

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

Answer:

A Q-Q plot is called as quantile – quantile graph where the quantiles of one dataset is plotted against the quantiles of second dataset. For example 0.5 quantiles is a point below which 50% of data exist and 50% above it. By plotting quantiles we can infer whether the the two data set follow the same distribution(like normal, exponential) or not.

Importance:

- Q-Q plot is useful in detecting whether the two samples are from same Population or not.
- It is also used to study/compare the various distributions theoretically.