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?

The categorical variable in the dataset are season, weather situation, year, weekdays, holidays, month, workingdays.

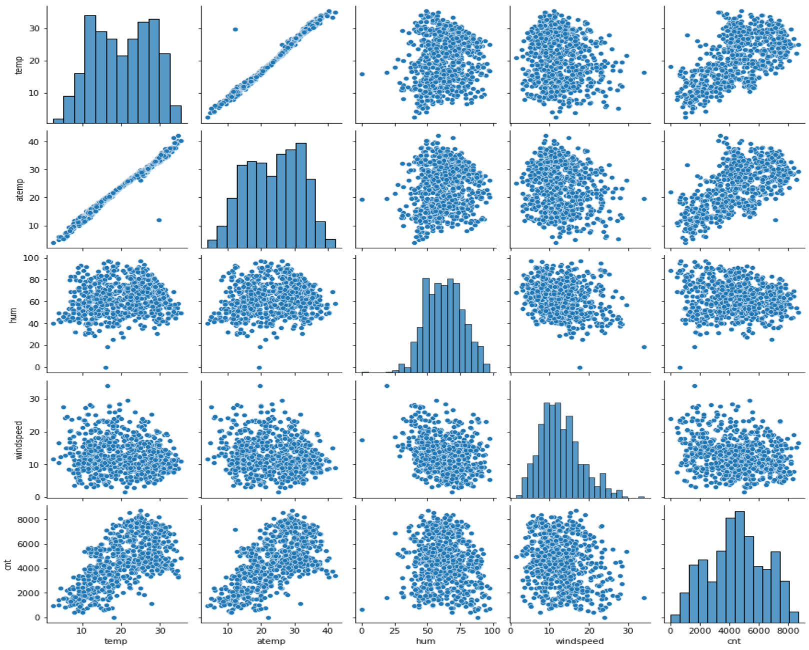
1. Season 3 - fall has highest demand for rental bikes.
2. More bikes are rent during 2019 when compared with 2018.
3. Demand of bike is similar throughout the weekdays.
4. Demand has decreased for holiday.
5. No significant change in working and non-working days.
6. Weathersit 1 (Few clouds) has highest demand.
7. During September, bike renting is more. During the year end and beginning, it is less.

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

It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. In iterative model there is trouble in converging if we have more number of variables.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

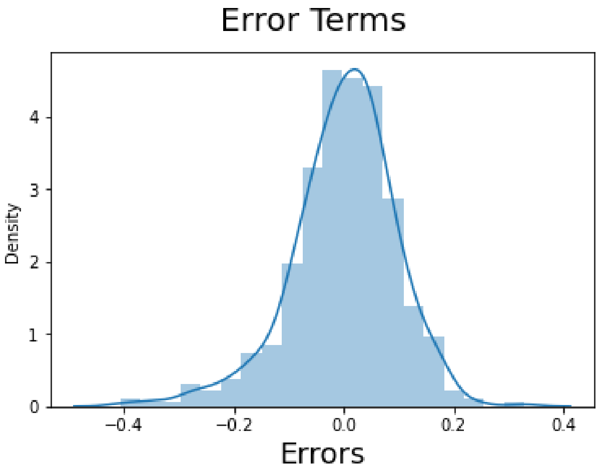
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?



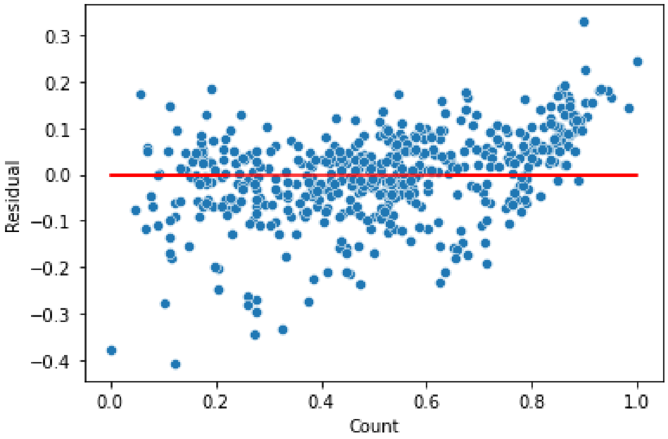
From above graphs we can say that temp and atemp have highest correlation with target variable count

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

1. The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity. This found using VIF value.
2. The error terms must be normally distributed.



3. The error terms must have constant variance. This phenomenon is known as homoscedasticity.



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 share bikes are:

1. Temp(Positive correlation).
2. Heavy Rain (weathersit3) (Negative correlation).
3. Year(Positive correlation).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Linear regression is a part of regression analysis. Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Linear regression is used to predict a quantitative response Y from the predictor variable X.

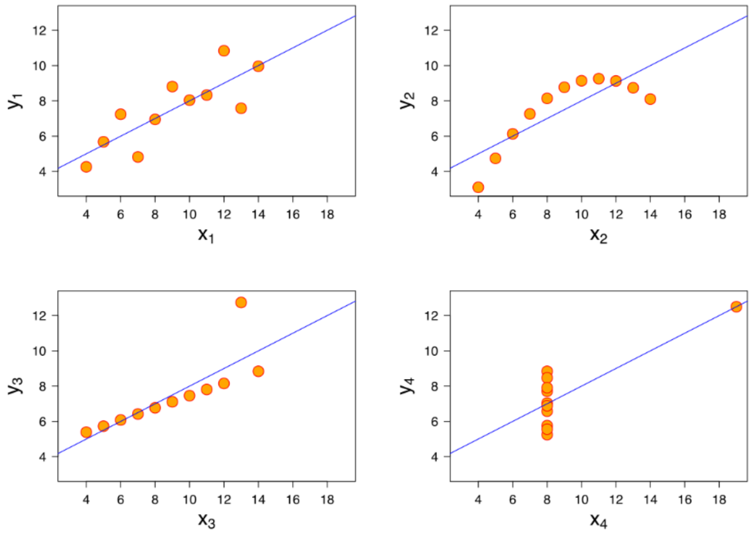
Mathematically, we can write a linear regression equation as:

$$y = mx + c$$

Here, x (Independent) and y (Dependent) are two variables on the regression line. m is Slope of the line. c is y-intercept of the line.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.



3. What is Pearson's R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

Pearson’s correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

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

Feature scaling is a method used to normalize the range of independent variables or features of data. In machine learning, feature scaling refers to putting the feature values into the same range.

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 must do scaling to bring all the variables to the same level of magnitude. It also helps in speeding up the calculations in an algorithm.

<u>Min-Max scaling</u>

In normalization, we map the minimum feature value to 0 and the maximum to 1. Hence, the feature values are mapped into the [0, 1] range:

$z = x - \min(x)/\max(x) - \min(x)$

<u>Standardization</u>

we don’t enforce the data into a definite range. Instead, we transform to have a mean of 0 and a standard deviation of 1. It not only helps with scaling but also centralizes the data.

$z = x\text{-mu}/\text{sigma}$

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.