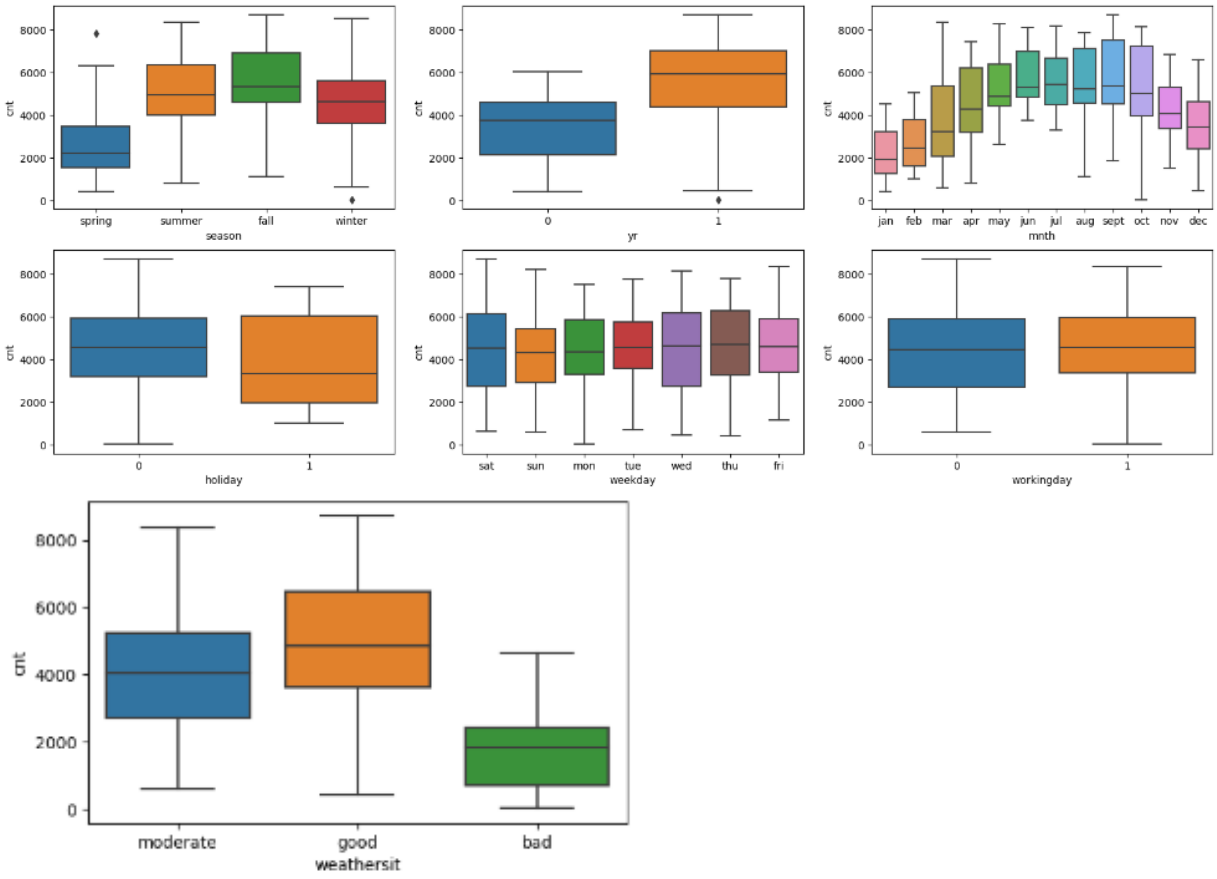
Assignment-Based Subjective Questions

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

Answer: The categorical variables named season, yr, mnth, holiday, weekday, workingday and weathersit in the dataset have major effect on the dependent variable 'cnt'. The below fig shows the correlation among the categorical variables and the dependent variable.



Inference from the above graphs:

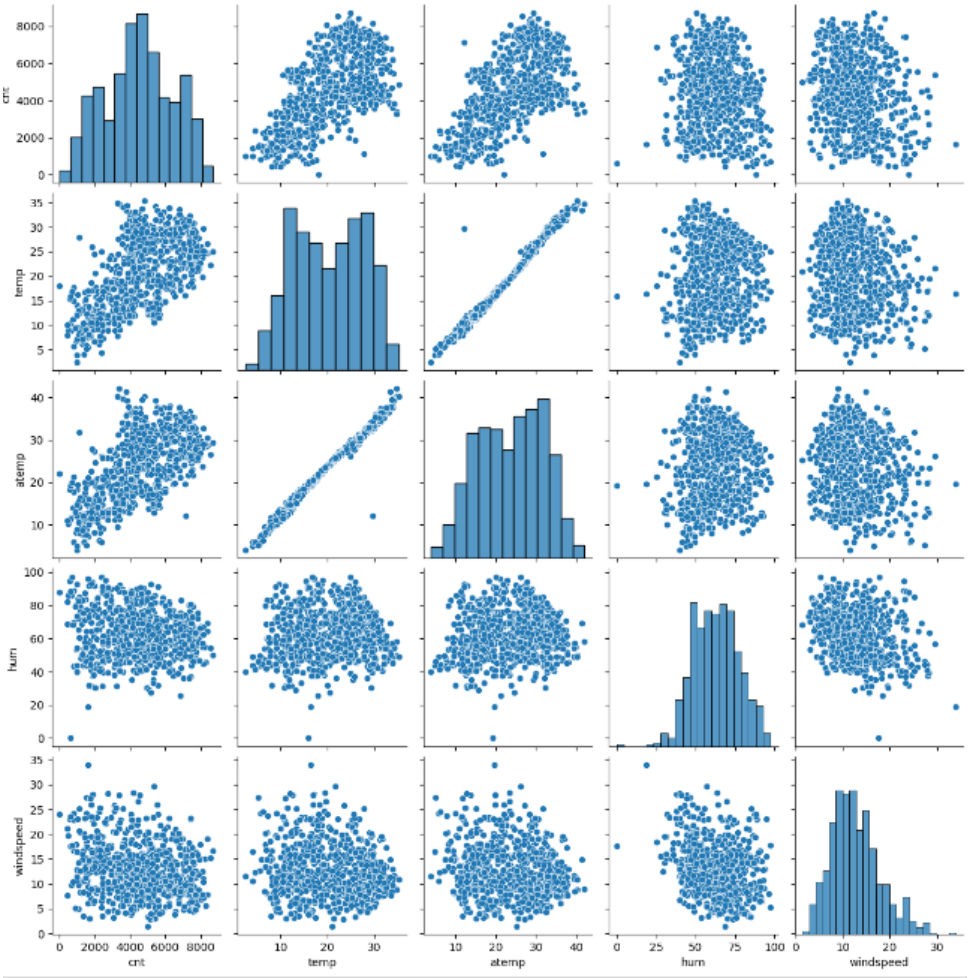
- fall Season has highest demand for rental bikes followed by summer and winter. In Spring season the bike demand is low.
- There is an increase in bike demand in next year 2019 compare to last year 2018
- September month has highest demand for bikes while November December January and February has the least demand, can be due to extreme weather conditions.
- Holiday have affected the mean demand negatively.
- No proper inference about bike demand can be drawn from Weekday.
- In working-days the demand is slightly more as compared to the non-working days.
- People have higher demand of bike during clear weather situations and lowest during bad weathers.

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

Answer: In general, when we use dummy variables to define categorical variables of level n, we create n-1 variables indicating the levels. We will be able to explain the categorical variable with n-1 levels.
Hence drop_first=True is used so that the resultant can match up n-1 levels. This is often used to avoid multi-collinearity issues and to maintain model interpretability.
Eg: If there are 3 levels, the drop_first will drop the first column.

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

Answer:

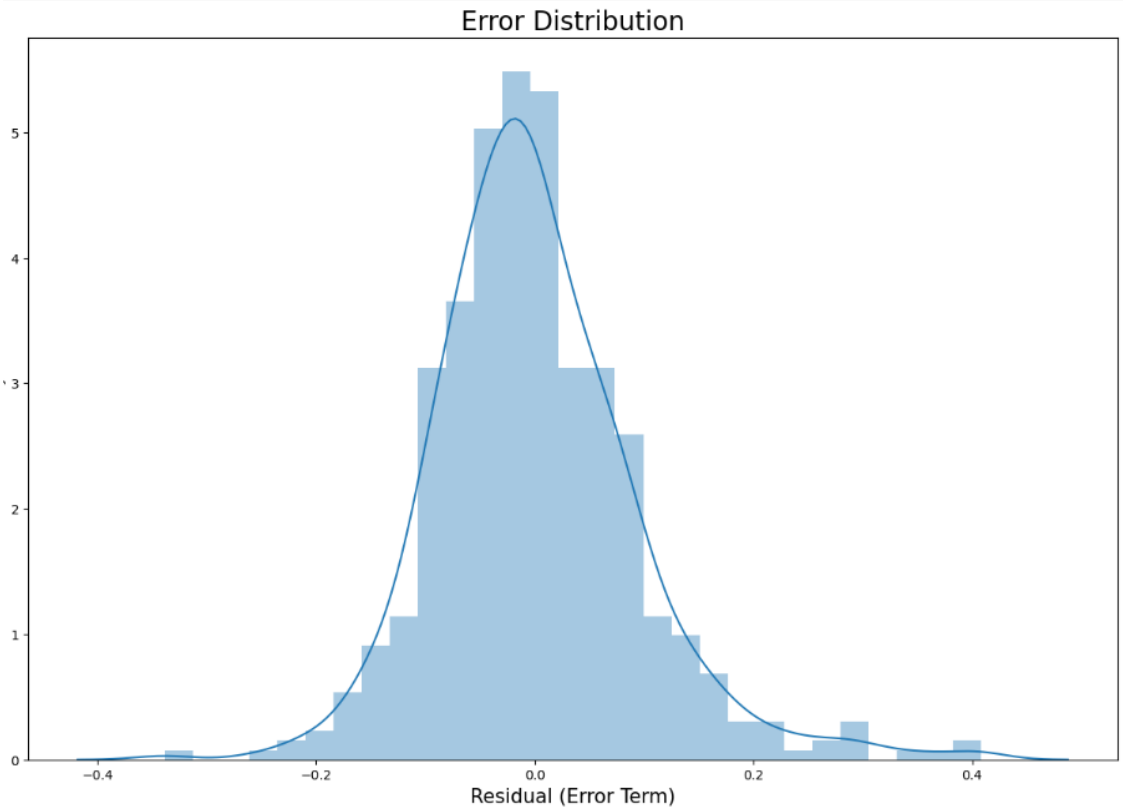


Analyzing the pair plot the variables 'temp' and 'atemp' has the highest correlation with target variable 'cnt'.

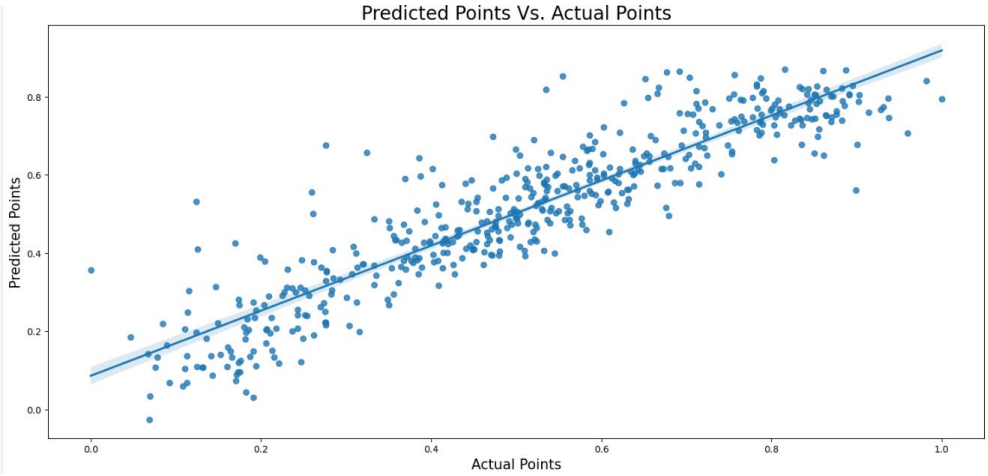
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Answer: Below are the linear regression assumptions validated after building the model on training set,

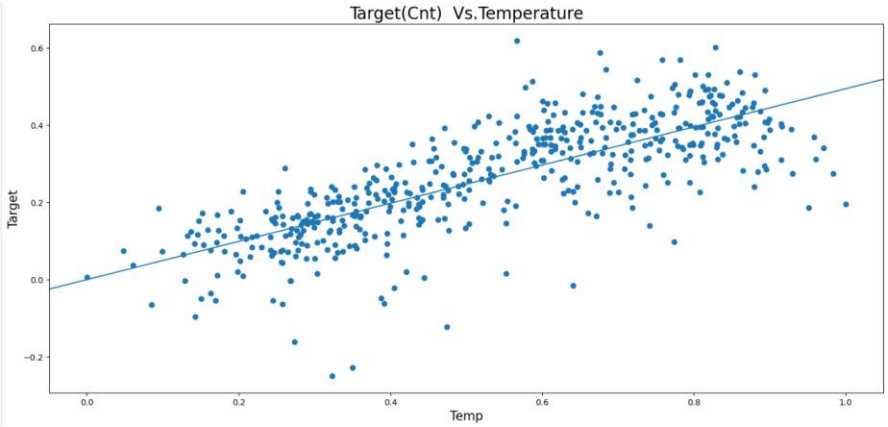
1. Error terms are normally distributed with mean zero – Through residual analysis and graph representation we have validated that the error terms are normally distributed with mean zero.

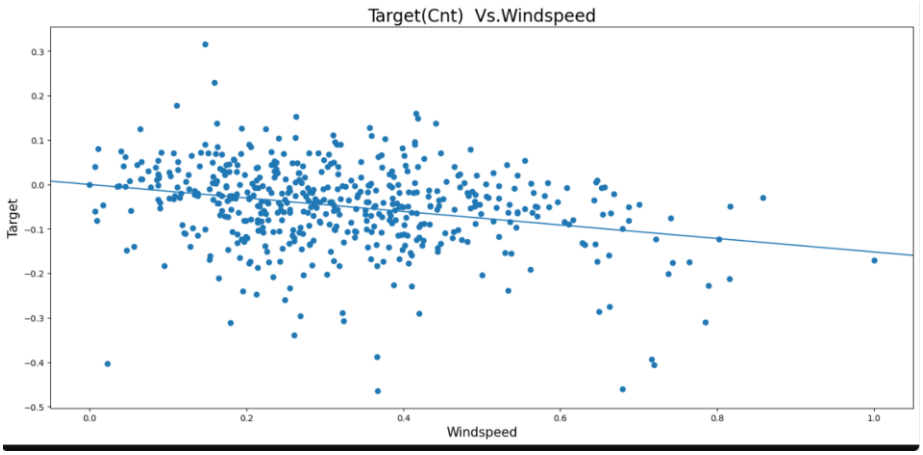


2. Error terms have constant variance (homoscedasticity) – From the below graph, we can say that residuals are equally distributed across predicted value. This means we see equal variance and we do not observe high concentration of data points in certain region & low concentration in certain regions. This proves Homoscedasticity of Error Terms.



3. Validating Independence of residuals – the calculated Durbin-Watson value for Final Model is 2.01, which shows that the model is absence of autocorrelation of the residuals. The residuals are independent of each other.
4. Validating Linearity – the below graph between the variables temp and target variable cnt, variable Windspeed and target variable cnt have a linear relationship which is validating linearity.





I have followed the above steps to validate the Linear regression assumptions.

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

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

1. Temperature
2. Year
3. Month September

These features co-efficient significantly affecting the bike hire count.

General Subjective Questions

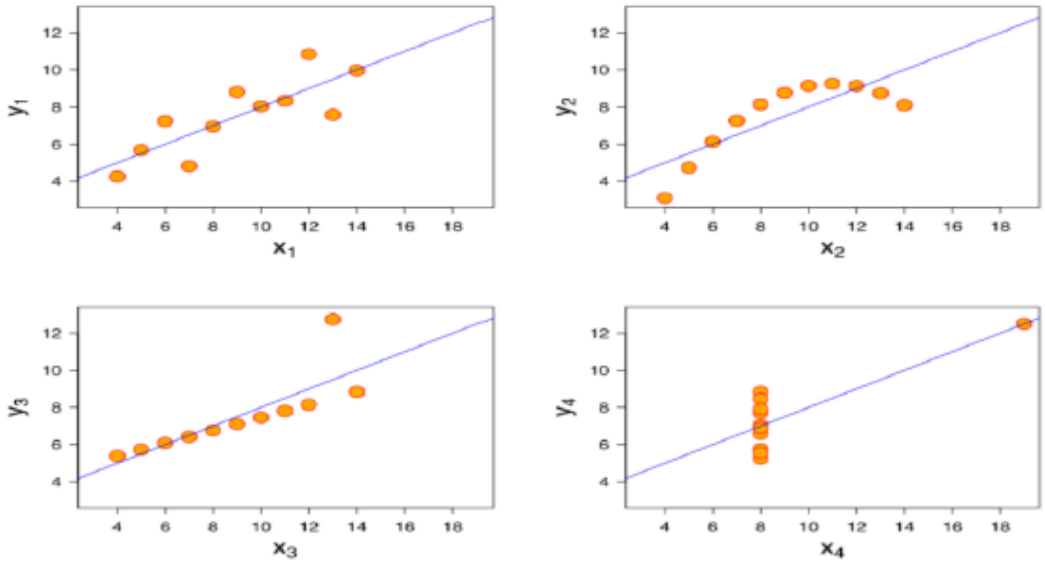
1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables. A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line and the best fit line should have the least error. In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points.

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It was constructed 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.



- 1st data set fits linear regression model as it seems to be linear relationship between X and y
- 2nd data set does not show a linear relationship between X and Y , which means it does not fit the linear regression model.
- 3rd data set shows some outliers present in the dataset which can't be handled by a linear regression model.
- 4th data set has a high leverage point means it produces a high correlation coeff.

Its conclusion is that regression algorithms can be fooled so, it's important to data visualization before build machine learning model.

3. What is Pearson's R?

Answer: In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.

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

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

Answer: Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modelling.

Difference between Normalizing Scaling and Standardize Scaling:

1. In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.
2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
5. Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.
6. Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization.

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

Answer: VIF(Variance Inflation Factor) basically helps explain the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below:

- ➔ A VIF value of greater than 10 is definitely high,
- ➔ a VIF of greater than 5 should also not be ignored and inspected appropriately.

A very high VIF value 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 multi-collinearity.

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

Answer: Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.
Q-Q plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can

best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.
Importance of Q-Q Plot in Linear Regression :
In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.
Advantages:

- It can be used with sample size also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot Q-Q plot use on two datasets to check.
- If both datasets came from population with common distribution
- If both datasets have common location and common scale
- If both datasets have similar type of distribution shape
- If both datasets have tail behaviour