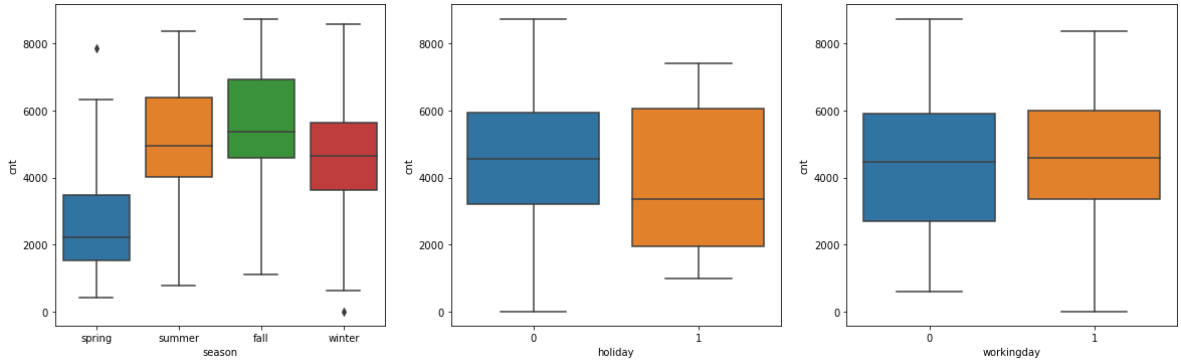
Assignment-based Subjective Questions:

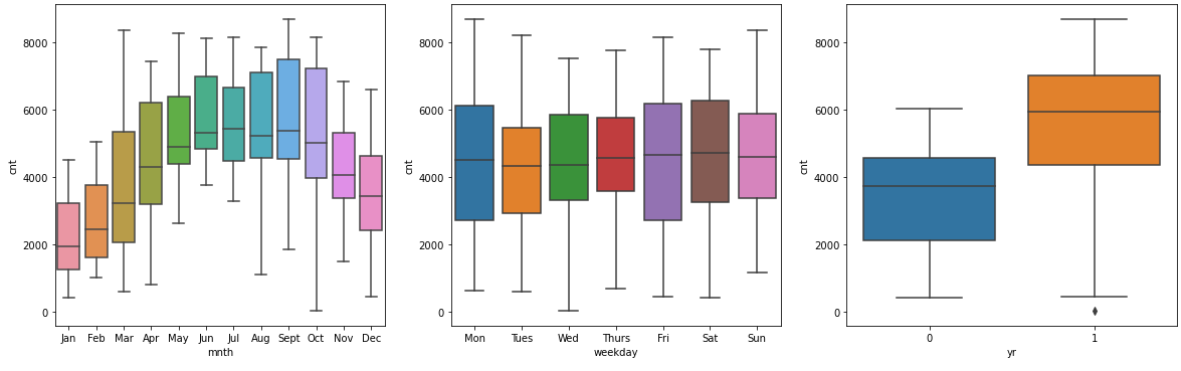
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

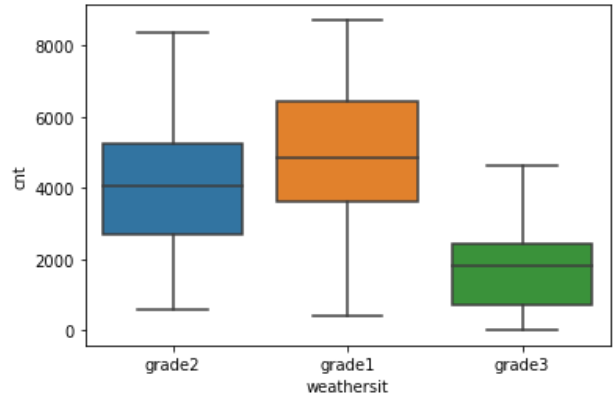
Answer: Following are the categorical variables in the dataset:

1. ‘season': Season appears to be strongly affecting the demand for shared bikes
2. ,’holiday’, : On holidays the median demand (count) is lower than non-holidays
3. ‘workingday’, (Neither a holiday nor a weekend): Has no significantly notable effect on d emand of shared bikes
4. 'mnth', : Month has strong & clearly visible impact on demand. Generally warmer mont hs cause more demand.
5. 'weekday', : Weekday does not have significant notable impact on demand.
6. ‘yr’: Year has strong impact on demand. yr value 0 = 2018 has lesser demand, however, yr value 1 = 2019 has shown higher demand.

7. 'weathersit’: Weather situation is provided in 4 graded numbers from good to bad. Wea ther situation is appearing to impact demand significantly. Good weather causing more demand (boxplots on next page)







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

Answer: If we have n variables, then while converting to dummies, they can be reduced by one count (n-1). In binary value assignment (1 or 0), all zeros can be considered as the 'first column' which is being dropped. Advantages are:

a. It reduces number of features, enabling brevity & makes the model leaner
b. Many times, it eliminates the need of 'unknown' category.

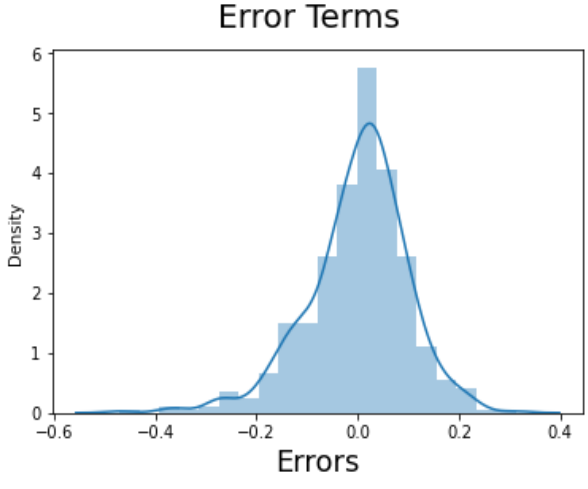
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: Registered users, (feature name: 'registered') & then casual users (feature name: 'casual') have top two highest correlations with target variable (cnt). However, as it is adding up with 'casual' feature to make cnt, this correlation is obvious. Other than 'registered' & 'casual', 'temp' & 'atemp' have next strongest correlation with cnt. Note that temp & atemp are also directly related to each other & thus are redundant up to certain extent.

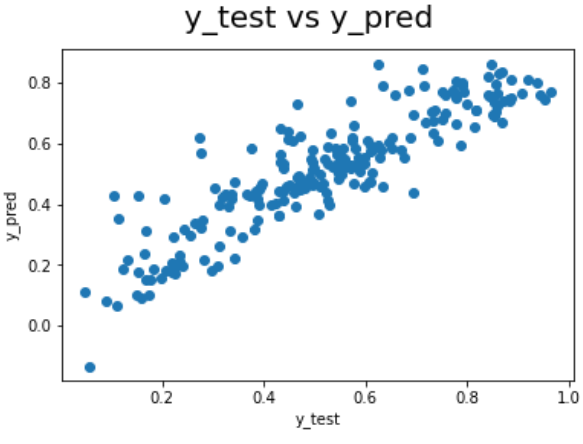
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer: This can be checked by:

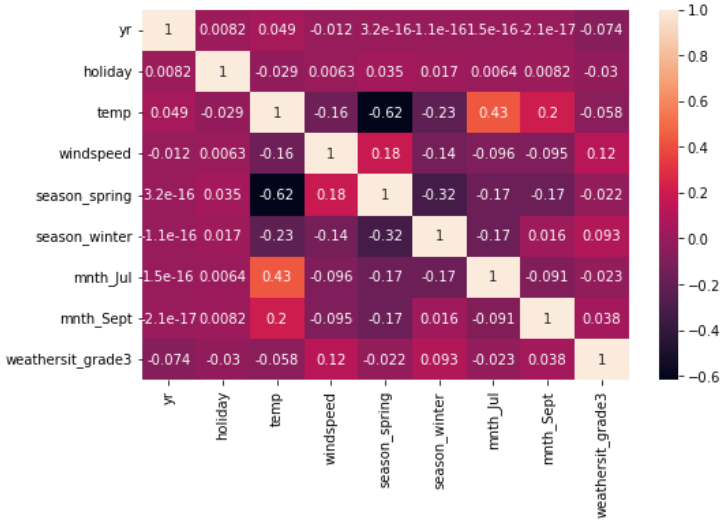
1. Checking histogram of error terms. Seeing that they have fairly normal distribution.



2.Checking a scatter plot of y (test) vs y (predicted)



3. Seeing a correlation matrix for finally selected independent variables to confirm that there is no multicollinearity within these independent variables.



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:
Following three are the top factors affecting demand of the shared bikes:

1. Temperature
2. Year
3. Season

General Subjective Questions

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

Answer: Linear Regression algorithm is one of the basic methodologies used in machine learning. It is a supervised learning involving an intention to predict a dependent variable (also known as target variable) based on available independent variable or variables.

This algorithm finds best fit linear relationship between target variable & independent variable. If there are multiple independent variables, it finds the best fit linear relationship between target variable & each independent variable, as well as their resultant effect on target variable by calculating the coefficients for each independent variable.

It also provides magnitude of each independent variable for it's effect on target variable (outcome). General formula is:

y = B0 + B1X (Single independent variable)

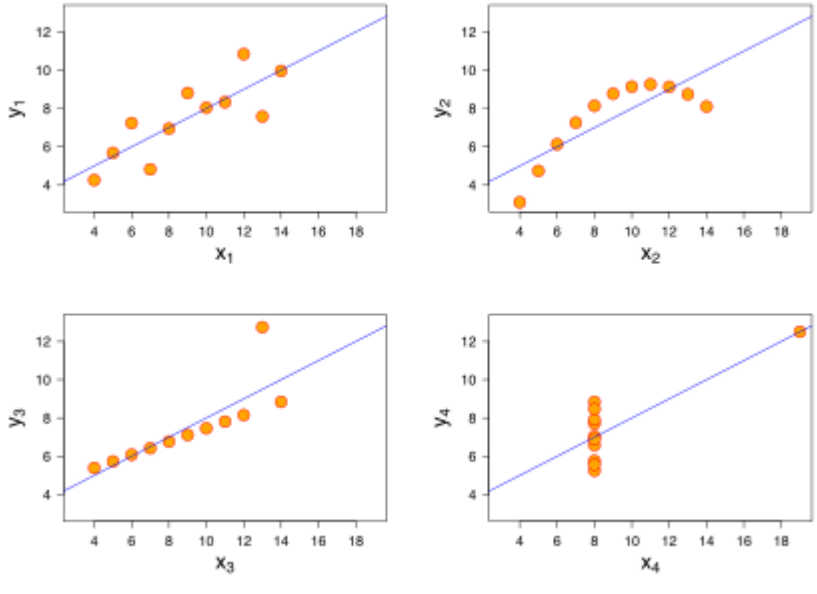
y = B0 + B1X1 + B2X2 +…+ BnXn (Multiple independent variables)
(y = target variable, X, X1, X2 = independent variable(s)

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

Answer: This term basically refers to a collection of 4 datasets of two variables each, which were created by statistician Frank Anscombe to highlight the importance of visualization of data. These four datasets created as an example have nearly same statistical values however, their distributions are entirely different from each other. The statistical parameters like mean, sample variance, correlation between two variables, linear regression line equation & R squared (coefficient of determination) were extremely similar to each other.

However, when they are plotted on chart, they show entirely different distributions. Each set has a unique distribution.

Diagram of charts provided below from Wikimedia for reference.



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

Answer: Pearson's R is a correlation coefficient which measures the magnitude & direction of relationship between two variables.

It is one of the most common ways of measuring linear correlation & thus it is sometimes referred to as just "correlation coefficient", which is understood as Pearson's correlation coefficient by default.

Formula to calculate Pearson's R is as follows.

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

However, software applications like MS Excel & language tools in Python, R etc have in-built methods to calculate the value of this coefficient.

The value is always between -1 & +1. The sign (+) or (-) indicates the direction of relationship (variables positively change with each other & negatively change with each other respectively.)

The value itself indicates the magnitude of relationship. -1 being strongest negative correlation & +1 being strongest positive correlation. Values near zero indicate no or poor correlation.

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

Answer: Input variables are likely to have different units & they are usually having different scales. This makes values of some variables much larger than values of other variables.

In some machine learning algorithms (such as linear regression) which use the difference between values of different variables, create bias when the value of one variable is in very higher range than the others. Higher weightage is assigned to higher values which may make the learning process less effective & less accurate. There are two ways of scaling the variables.

a. **Normalized scaling** (also known as Min-Max scaling) uses minimum & maximum values of variables in dataset. The following formula is used to bring multiple variables to similar scale.

   Var_Scaled = (Var - Var_min) / (Var_max - Var_min)

   This method scales down the values to a range between 0 to 1 in the same proportion as original values. However, this method is not desirable when there are too many large value outliers.

   It can be used for data like age, human height etc. But not suitable for income amounts, distances etc.

b. **Standardized scaling:** This method uses mean & standard deviation to squish the variable value range to a smaller range. However, the scaled variables are not bound between any fixed range.

   Var_Scaled = (Var - mean)/Std Dev

   Standardized scaling is basically subtracting the mean from value or variable and dividing by the standard deviation of a variable. This type of scaling is not affected largely by outliers.

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

Answer: VIF (variance inflation factor) is the measure of the amount of multicollinearity in a set of multiple regression variables. Higher the value of VIF, stronger the correlation. A perfect correlation between variables will cause a VIF value of infinity. An infinite VIF value shows that the referenced variable may be expressed exactly by a linear combination of other variables.

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

Answer: Q-Q plot a short name for quantile-quantile plot, is used for comparing two probability distributions. It is formed by plotting the quantiles of both probability distributions against each other. It helps you to understand if two datasets are from a population with a common distribution (e.g. normal distribution).

In linear regression, we can use Q-Q plots to check if the residuals follow a normal distribution. This is one of the assumptions of linear regression & it can be validated by Q-Q plot.

(3 marks)