Submitted by: Pieyush C Joy
Linear Regression- UpGrad

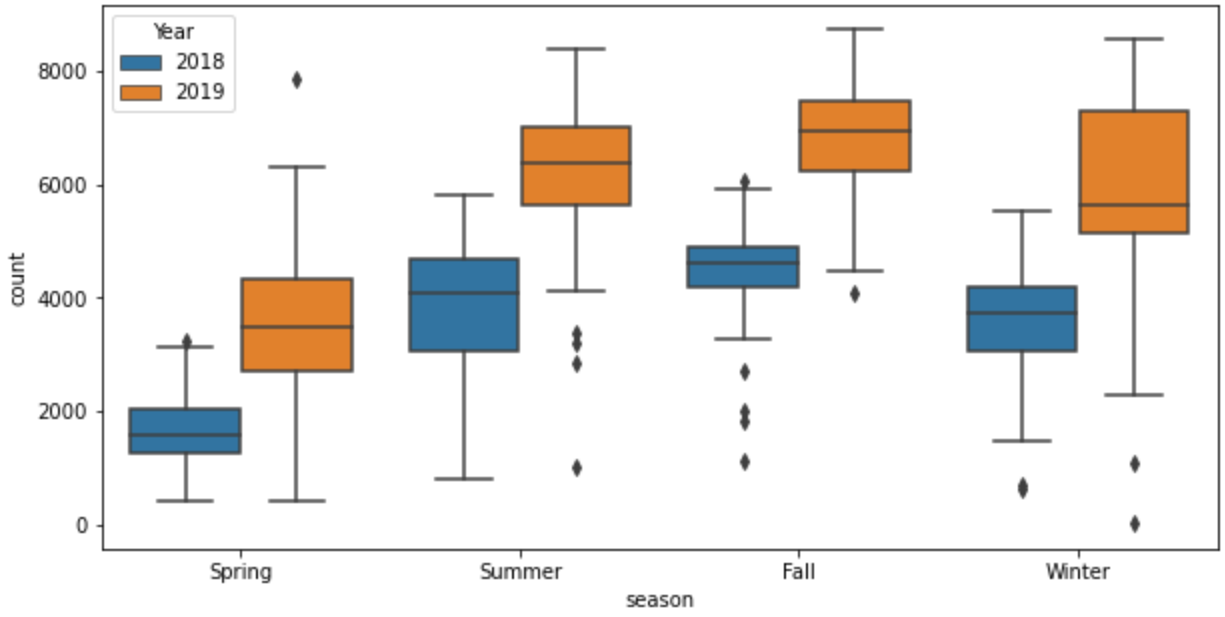
Assignment-based Subjective Questions- Boom Bikes

Q1. *From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?*

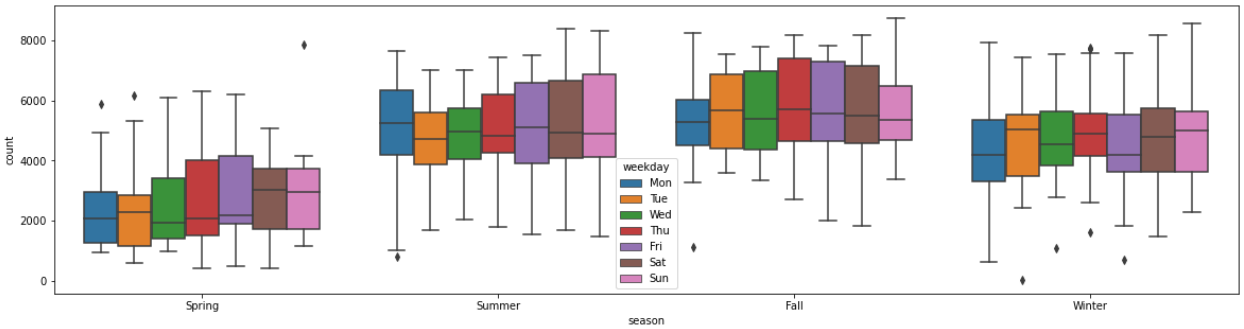
Answer:

From the dataset we can infer there are seven categorical variables in the dataset. The analysis of the effect on the dependent variable can be summarized as:

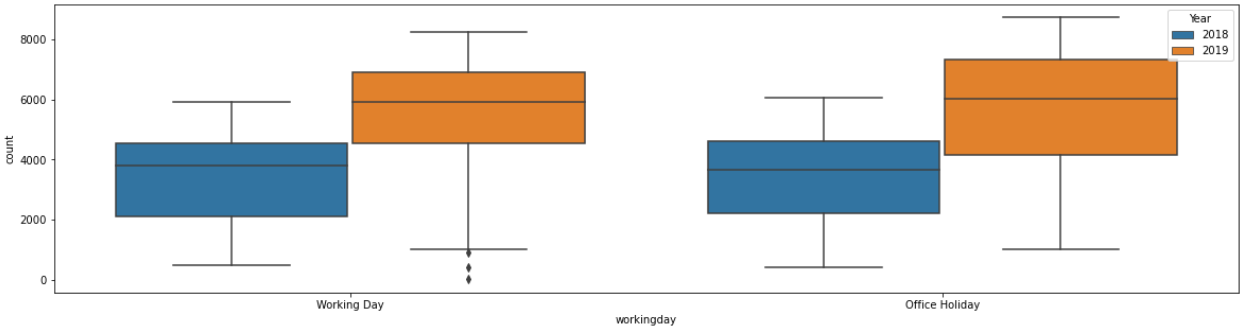
1. *From the boxplot, it can be inferred that in Summer and Fall, count increases in both years (2018 as well as 2019). Count has been increased from 2018 to 2019 .*



2. *Summer and Fall count is high and that count has been pretty high in Friday and Saturday*



3. *Count is high for clear weathersit in the year 2019*



4. *Count is considerably high during the month of September for both 2018 and 2019*
5. *Count is considerably high during the month of September for both 2018 and 2019*

Q2. Why is it important to use drop_first=True during dummy variable creation?
Answer:
Using drop_first=True is more common in statistics and often referred to as "dummy encoding" while using drop_first=False gives you "one hot-encoding" which is more common in ML. Using dummy encoding on a binary variable does not mean that a 0 has no relevance. If gender_male has high importance that does not generally say anything about the importance of gender_male==0 vs gender_male==1. It is variable importance and accordingly calculated per variable. Moreover, if the gender variable is binary, gender_male==1 is equivalent to gender_female==0. Therefore from a high variable importance of gender_male one cannot infer that being female (or not) is not relevant.
In this case gender_male==0 AND gender_female==0 means Transgender is true. For algorithmic approaches in ML there is no statistical disadvantage using one hot-encoding.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

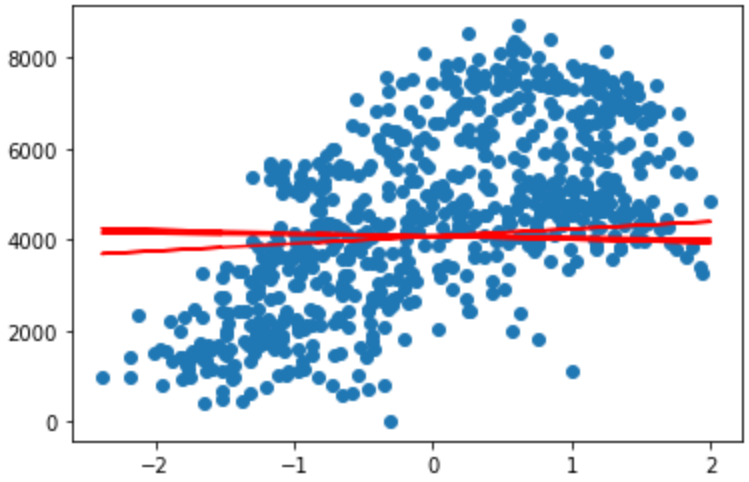
Answer:
"atemp" has the highest correlation with the target variable.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

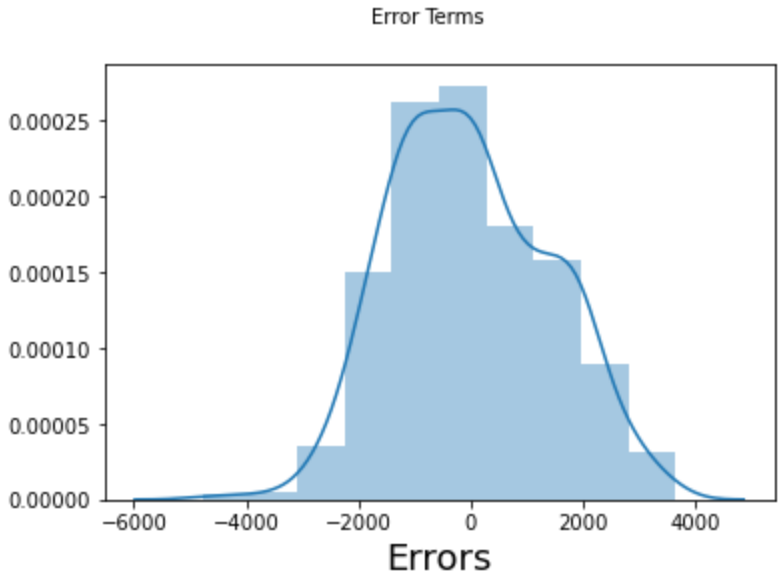
Answer:

The assumptions of Linear Regression after building the model can be validated as:

a. *Fitted regression line is linear.*



b. *Error terms are normally distributed.*



Q5. *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 are season, month and windspeed.

***General Subjective* Questions**

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

***Answer*:**

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of the data based on some variables. In the case of linear regression as the name suggests linear it meAnswer the two variables which are on the x-axis and y-axis should be linearly correlated. Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.

Regression analysis is used for three types of applications:

- *Finding out the effect of Input variables on Target variable.*
- *Finding out the change in Target variable with respect to one or more input variables.*
- *To find out upcoming trends.*

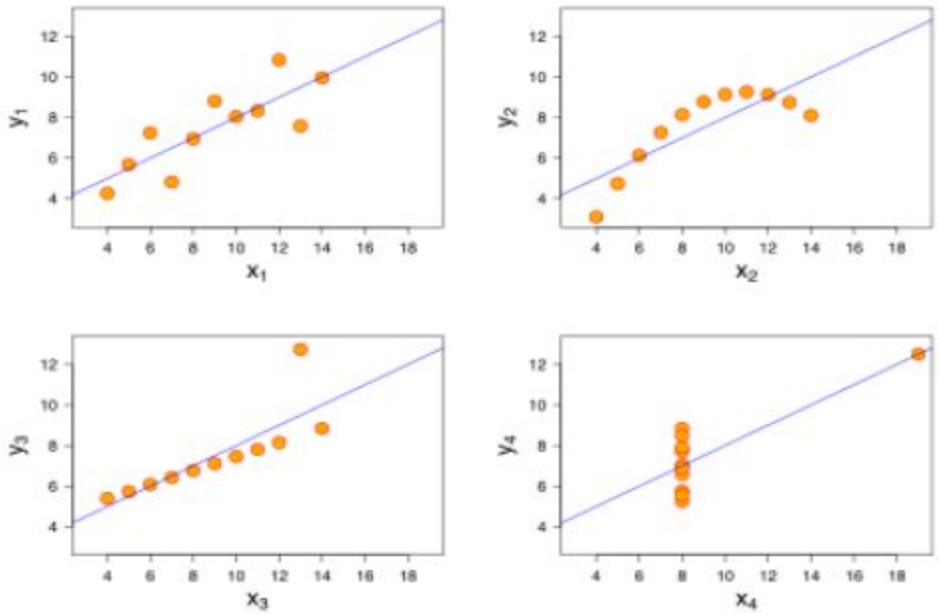
Q2. *Explain Anscombe's quartet in detail.*

***Answer*:**

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

(Please find image below)

Submitted by: Pieyush C Joy
Linear Regression- UpGrad



For all four datasets following parameters are calculated:

- *Mean and Sample variance of x and y*
- *Correlation between x and y*
- *Linear regression line*
- *Coefficient of determination of the linear regression: R2*

The four plots signifies:

a. The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

b. The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant.

c. The third graph (bottom left), the distribution is linear, but should have a different regression line . The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

d. The fourth graph shows when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

Submitted by: Pieyush C Joy
Linear Regression- UpGrad

Q2. *What is Pearson's R?*
***Answer*:**
In statistics, the Pearson correlation coefficient referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.[

The absolute values of both the sample and population Pearson correlation coefficients are on or between 0 and 1. Correlations equal to +1 or −1 correspond to data points lying exactly on a line (in the case of the sample correlation), or to a bivariate distribution entirely supported on a line (in the case of the population Correlation.

A key mathematical property of the Pearson correlation coefficient is that it is invariant under separate changes in location and scale in the two variables. That is, we may trAnswerform X to a + bX and trAnswerform Y to c + dY, where a, b, c, and d are
constants with b, d > 0, without changing the correlation coefficient

Q3. *What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?*
***Answer*:**
Scaling (also called min-max scaling), can trAnswerform the data such that the features are within a specific range e.g. [0, 1].
Scaling is important in the algorithms such as support vector machines (SVM) and k-nearest neighbors (KNN) where distance between the data points is important. Normalization is good to use when the distribution of the data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks. Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, Unlike normalization, standardization does not have a bounding range. So, even if there are outliers in the data, they will not be affected by standardization.

Submitted by: Pieyush C Joy
Linear Regression- UpGrad

Q4. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

***Answer*:**

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

When the the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which leads to 1/(1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Q5. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

***Answer*:**

Quantile-Quantile (Q-Q) plot, is a graphical tool to help assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.In a scenario of linear regression when there is training and test data set received separately and then it can confirmed using Q-Q plot that both the data sets are from populations with same distributions. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

***********Thank You*********