Assignment-based Subjective Questions

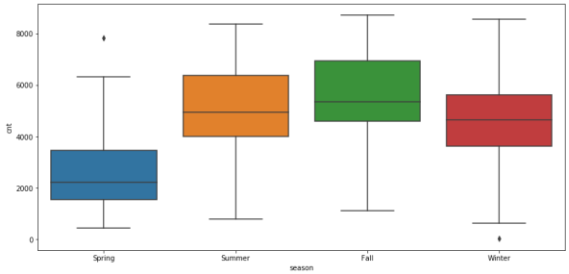
Question 1:

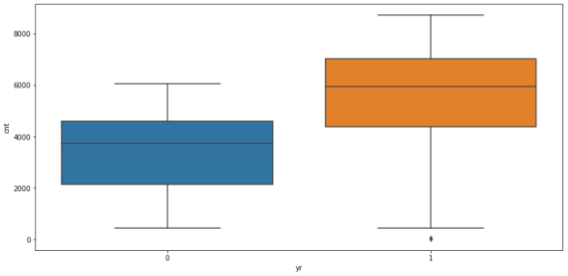
From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

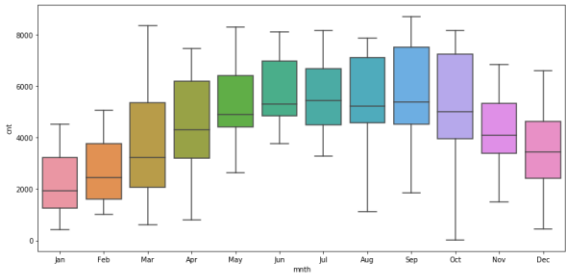
Answer:

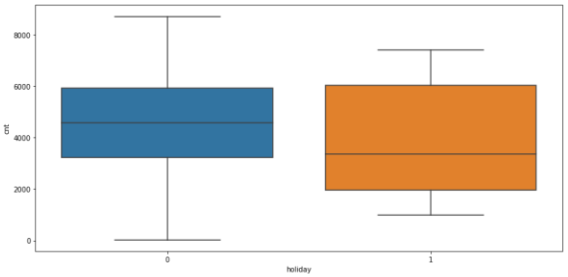
There are **7** categorical variables in dataset, these are visualized through boxplot against '**cnt**' and below is the insight from the boxplot:

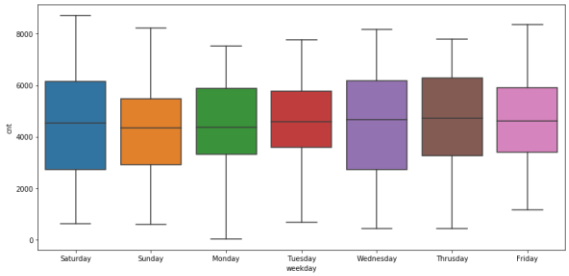
- **Season** : From the boxplot we can see 'spring' has lowest numbers whereas 'fall' has highest rentals.
- **yr** : 2019 witnessed more rentals compare to 2018.
- **mnth** : September has highest number of rentals
- **holiday** : Rentals reduced during holiday.
- **weekday** : Weekday is not really impacting rentals.
- **workingday** : Workingday is not having significant impact on rentals.
- **weathersit** : Clear weather witness more number of rentals.

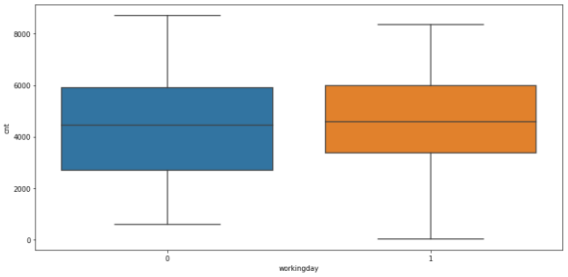


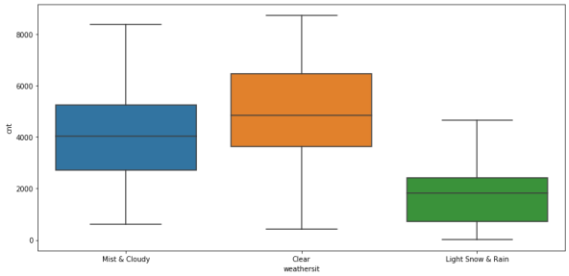












Question 2.

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

Answer:

'drop_first=True' used during dummy creation for categorical variables. It drops one of the feature as it helps getting K-1 dummy variable, with K-1 variable we can imply the meaning of the data and the dropping helps reducing complexity and extra feature from the dataset.

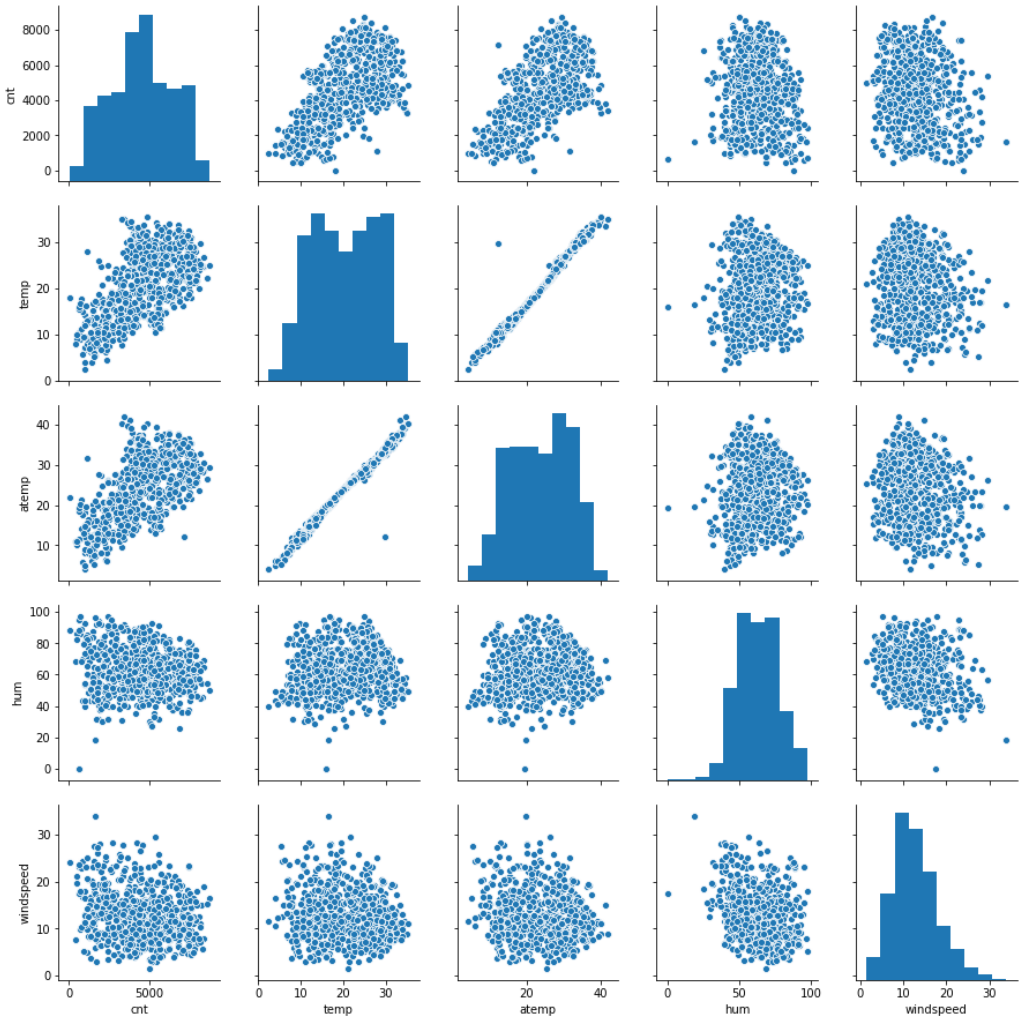
EX.: We have three variables: **Furnished, Semi-furnished** and **un-furnished**. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don't need unfurnished as we know 0-0 will indicate un-furnished, so we can remove it. It is also used to reduce the collinearity between dummy variables.

Question 3.

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

Answer:

'temp' and 'atemp' has highest correlation with target variable 'cnt'.

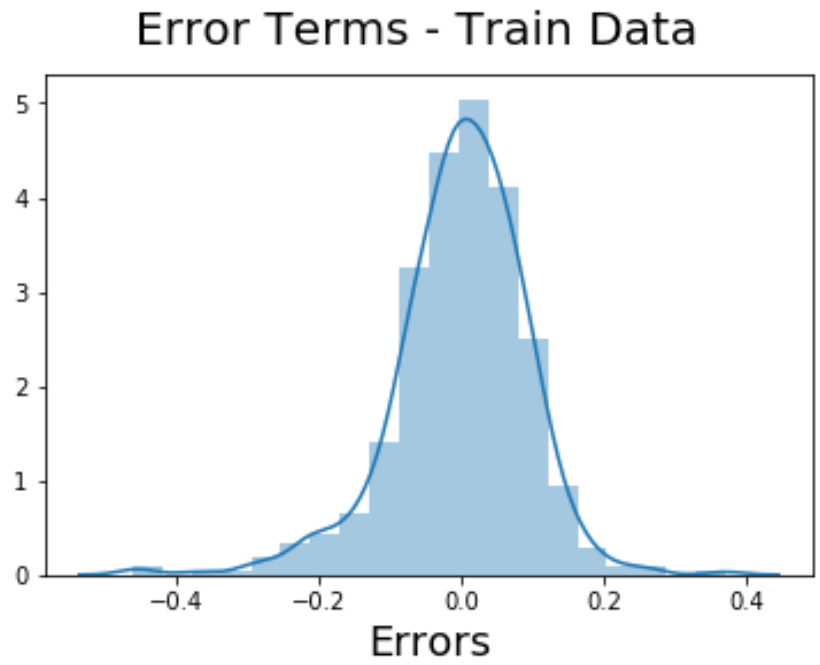


Question 4.

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

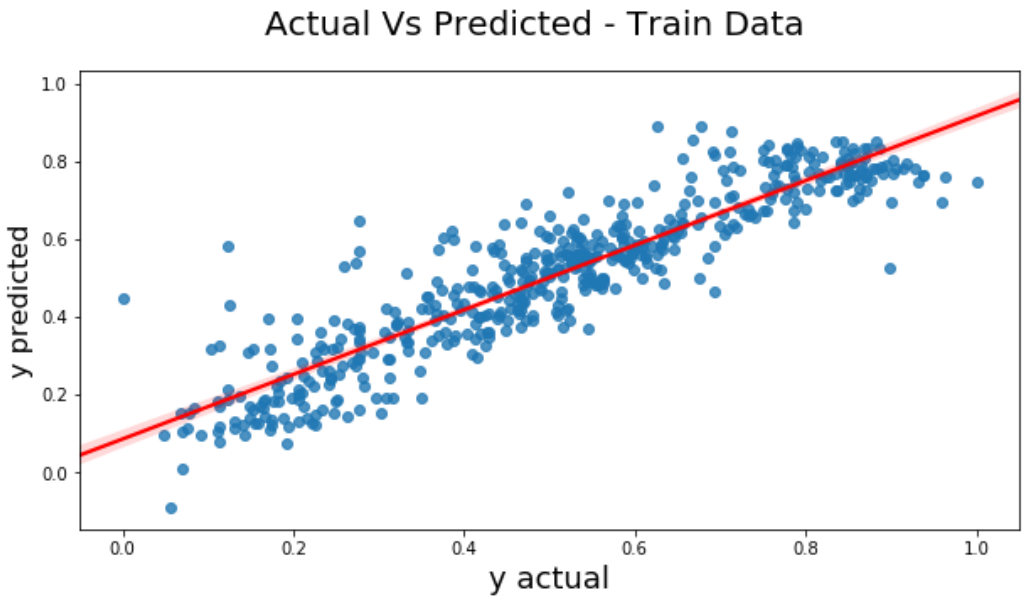
Answer:

Residual distribution should follow normal distribution and centred around mean value 0. Below plot shows the expected behaviour.



Error terms have constant variance (homoscedasticity):

- The variance should not increase (or decrease) as the error values change.
- Also, the variance should not follow any pattern as the error terms change.



Question 5.

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

Answer:

Below are the top 3 contributor to impact target variable.

1. **temp** 0.415195
2. **yr** 0.238922
3. **weathersit_Light Snow & Rain** -0.250532

General Subjective Questions

Question 1.

Explain the linear regression algorithm in detail.

Answer:

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of continuous values where independent variable can either be *continuous* or *categorical*. It is used for finding linear relationship between target and one or more predictors.

In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

1. Simple Linear Regression : In Simple linear regression we find the relationship between a dependent Y and independent variable X, the mathematical equation that approximates linear relationship between X and Y is.

Linear regression is based on the popular equation “y = mx + c”.

2. Multiple Linear Regression :MLR is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

y = m1x1 + m2x2 + ... +mnxn + c

Question 2.

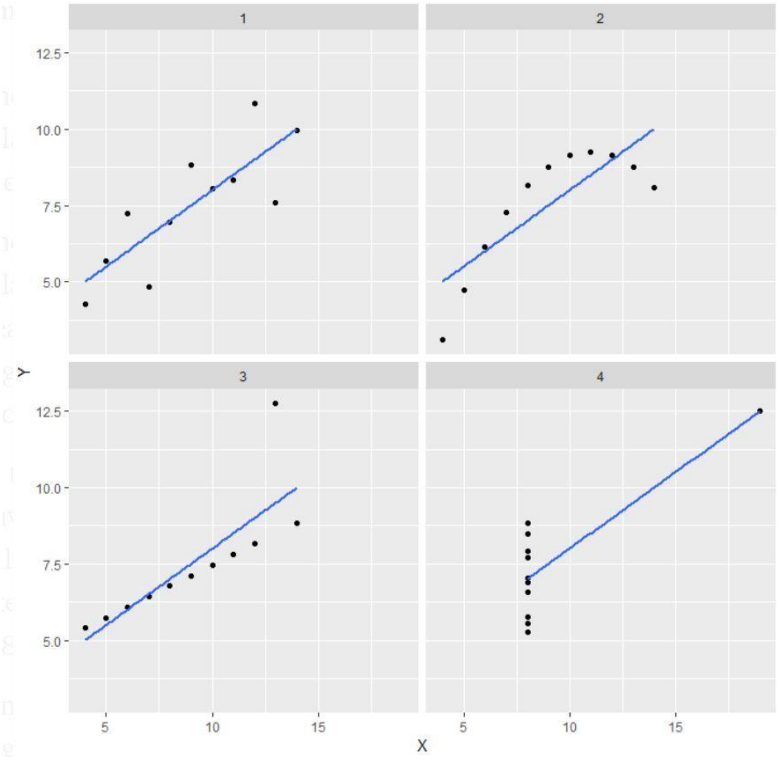
Explain the Anscombe's quartet in detail.

Answer:

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.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



Question 3.

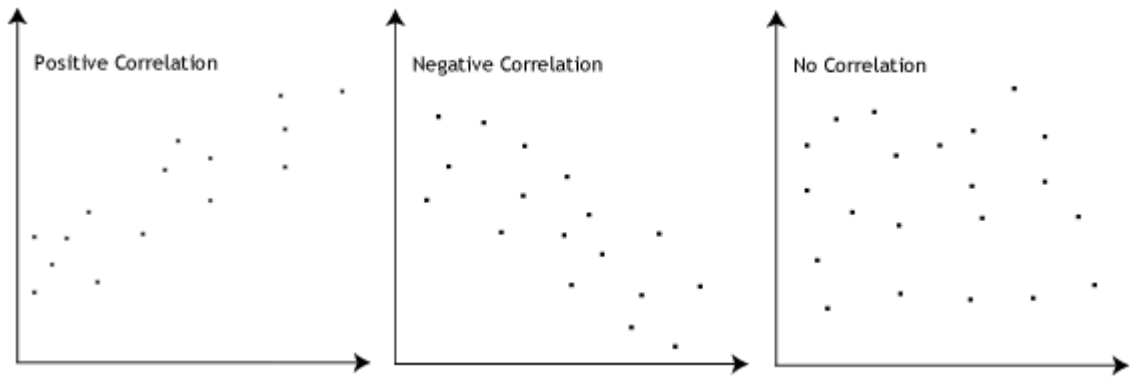
What is Pearson's R?

Answer:

The Pearson correlation method is the most common method to use for numerical variables; it assigns a value between − 1 and 1, where 0 is no correlation, 1 is total positive correlation, and − 1 is total negative correlation. This is interpreted as follows: a correlation value of 0.7 between two variables would indicate that a significant and positive relationship exists between the two. A positive correlation signifies that if variable A goes up, then B will also go up, whereas if the value of the correlation is negative, then if A increases, B decreases.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association



Question 4.

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

Answer:

Feature scaling is steps of data pre-processing which applied to normalize the data within range and helps to speed up the model. There are below two types of scaling used:

Normalisation(Scaling Normalization**)**:

- Minimum and maximum value of features are used for scaling.
- It is used when features are of different scales.
- Scales values between [0, 1] or [-1, 1 but really affected by outliers.
- Scikit-Learn provides a transformer called MinMaxScaler for Normalization.
- It is useful when we don't know about the distribution.

Standardisation(Z-Score Normalization**)**:

- Mean and standard deviation is used for scaling.
- It is used when we want to ensure zero mean and unit standard deviation.
- It is not bound to certain range and much less affected by outliers.

- Scikit-Learn provides a transformer called StandardScaler for standardization.
- It is useful when the feature distribution is Normal or Gaussian.

Question 5.

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

Answer:

The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. VIF is calculated based on R^2.

VIF = 1/(1-R^2)

When we get perfect correlation which is R~R^2~1. So, 1/(1-1) ~ 1/0 ~ Infinity.

Question 6.

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

Answer:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. The q-q plot is used to answer the following questions:

The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?