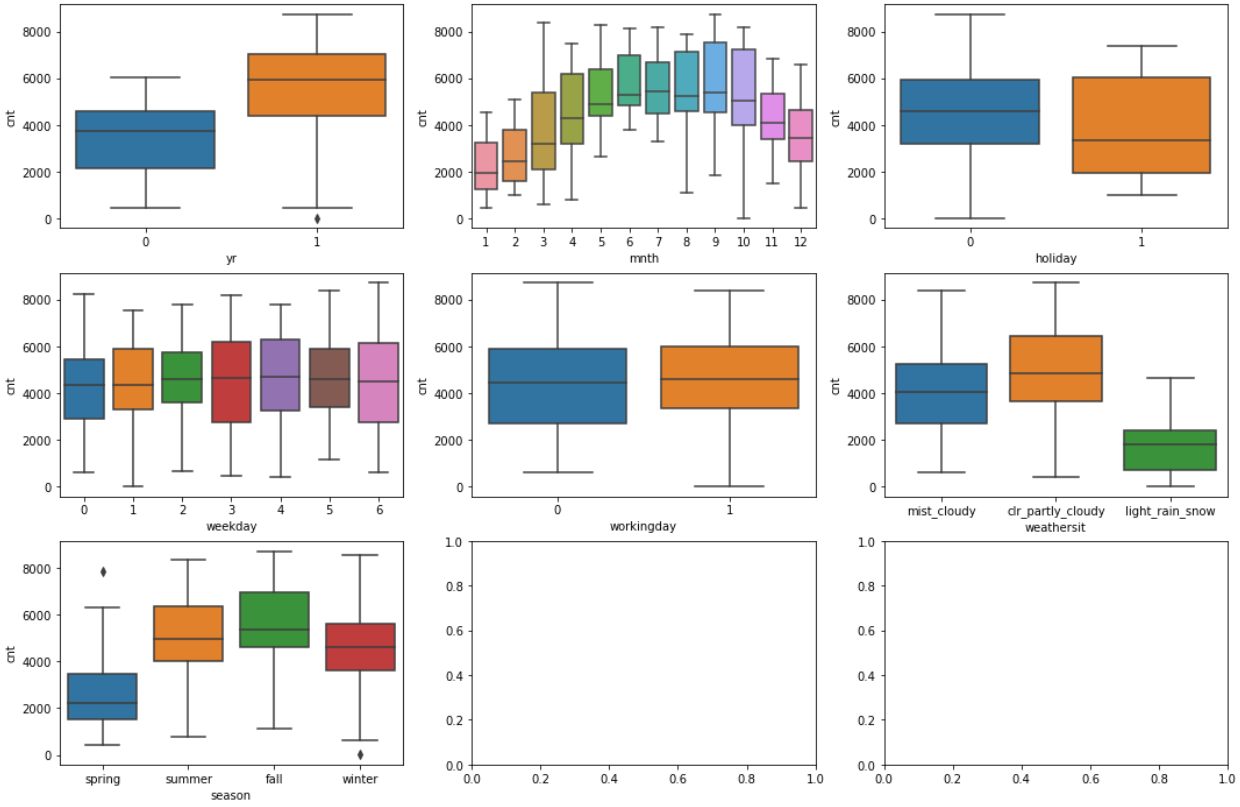
Assignment-based Subjective Questions

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

Ans:



Year: 2019 has higher median values as compared to 2018. That is, more bookings were made in 2019 as compared to 2018

Month: November, December, January & February have the lowest bookings, which could be due to weather

Holiday: There are lower bookings on holidays as compared to non-holidays

Weekdays: Across weekdays there does not seem to be much differences in bookings

Weather situation: Light rain and snow has the highest impact on bookings that is less bookings happen when it rains or snows

Season: Spring has the lowest bookings, followed by Winter.

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

Ans:

Let's take an example. Category column contains 3 distinct values/level A, B & C. While columns A, B & C are the one-hot encoded values.

Category	A	B	C
A	1	0	0
B	0	1	0
C	0	0	1

If we look at values of A, B, C for category A which are [1, 0, 0], we can clearly see that it is inherently coded in the B & C columns and do not explicitly require to be created.

That is,

For category B, BC will contain [1, 0]

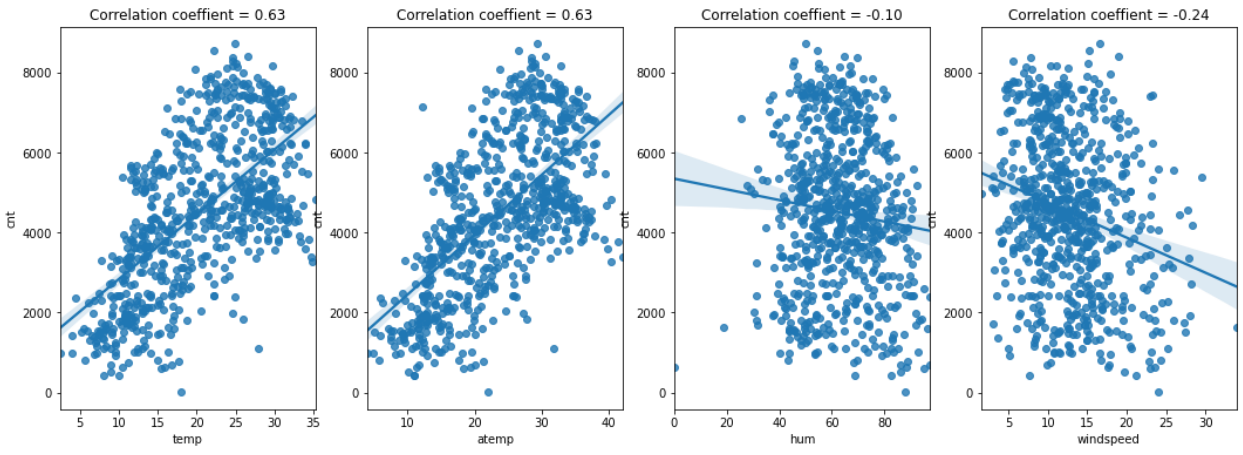
For category C, BC will contain [0, 1]

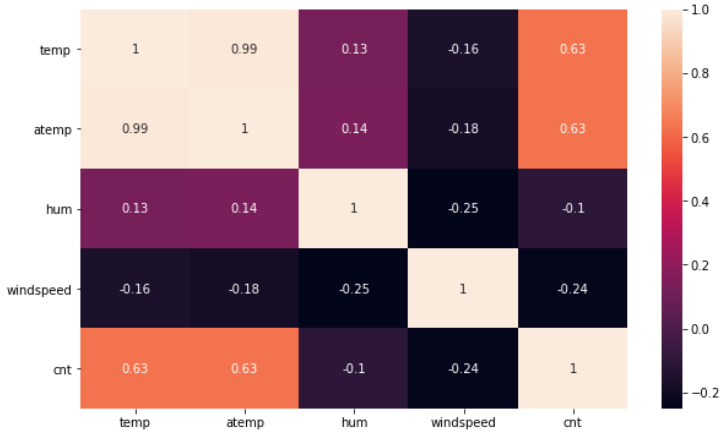
But, it will be [0, 0] when it is neither B or C, that it for A. So, A gets encoded implicitly and that is why we need 1 less column to represent.

So, if category contains 'N' levels, we need only 'N-1' columns to represent all the values.

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

Ans:



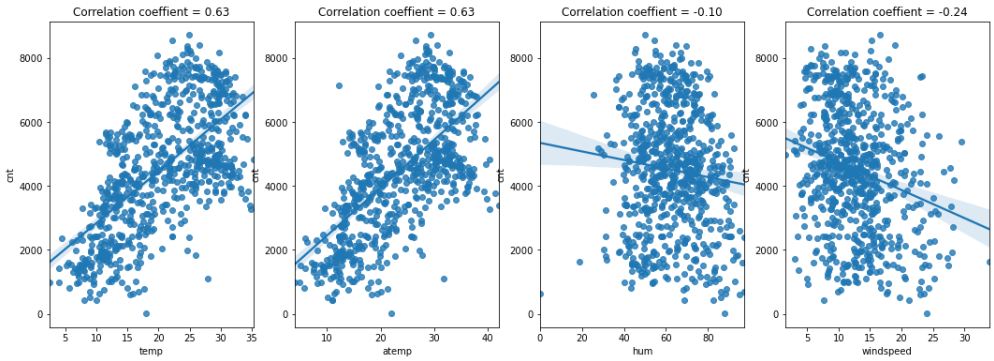


Temp and Atemp both are equally correlated with target variable (cnt). Each has correlation coefficient of 0.63

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

Ans:

a. **Linear Relationship:** there should be linear relationship between dependent and independent variables.

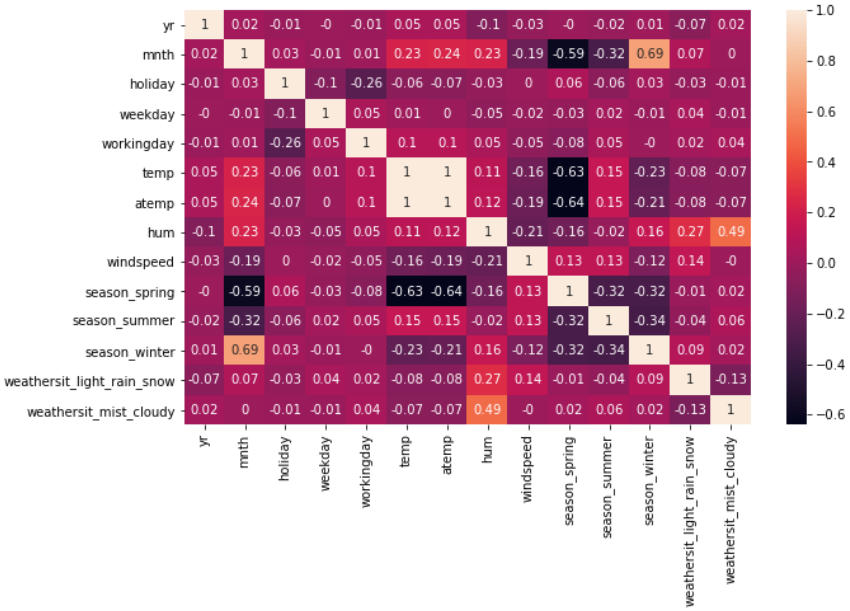


It can be clearly visualized that “temp” & “atemp” has high positive linear correlation. While humidity and windspeed and negative but slight correlation.

b. **Multicollinearity:** The independent variables should not be correlated with each other.

For the same checked the Pearson’s R correlation between all the independent

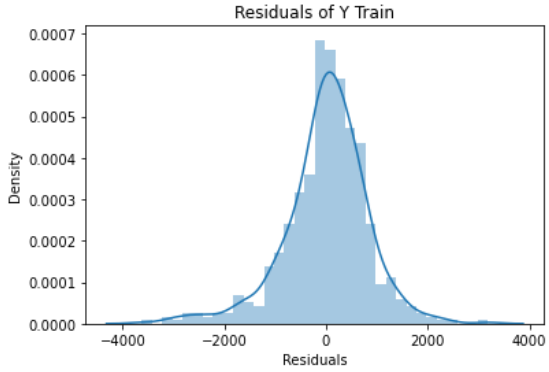
variables.



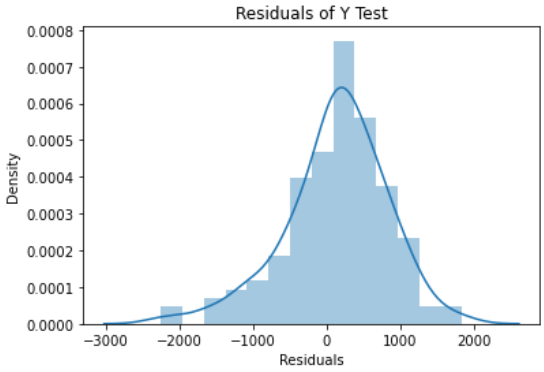
- Remove “atemp” which was highly correlated with “temp”
- Month and Winter season are also correlated but the coefficient value is small ~0.7
- Spring season and temperature are also correlated but have small coefficient of -0.64

c. **Errors/Residuals should be normally distributed:** Residuals that is the difference between Y Actual & Y Predicted

- For training data

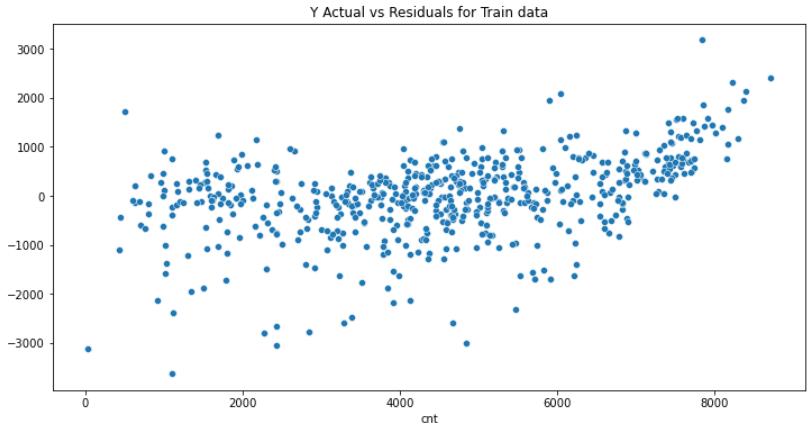


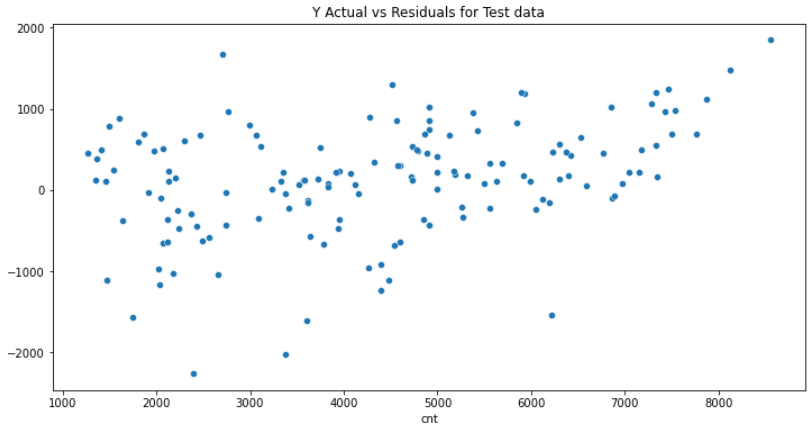
- For testing data



Since, the residuals are almost normally distributed, we can conclude that this assumption holds.

d. **Homoscedasticity: Constant variance.** Residuals should be randomly scattered when Residuals are plotted against Actual, Predicted or any index.

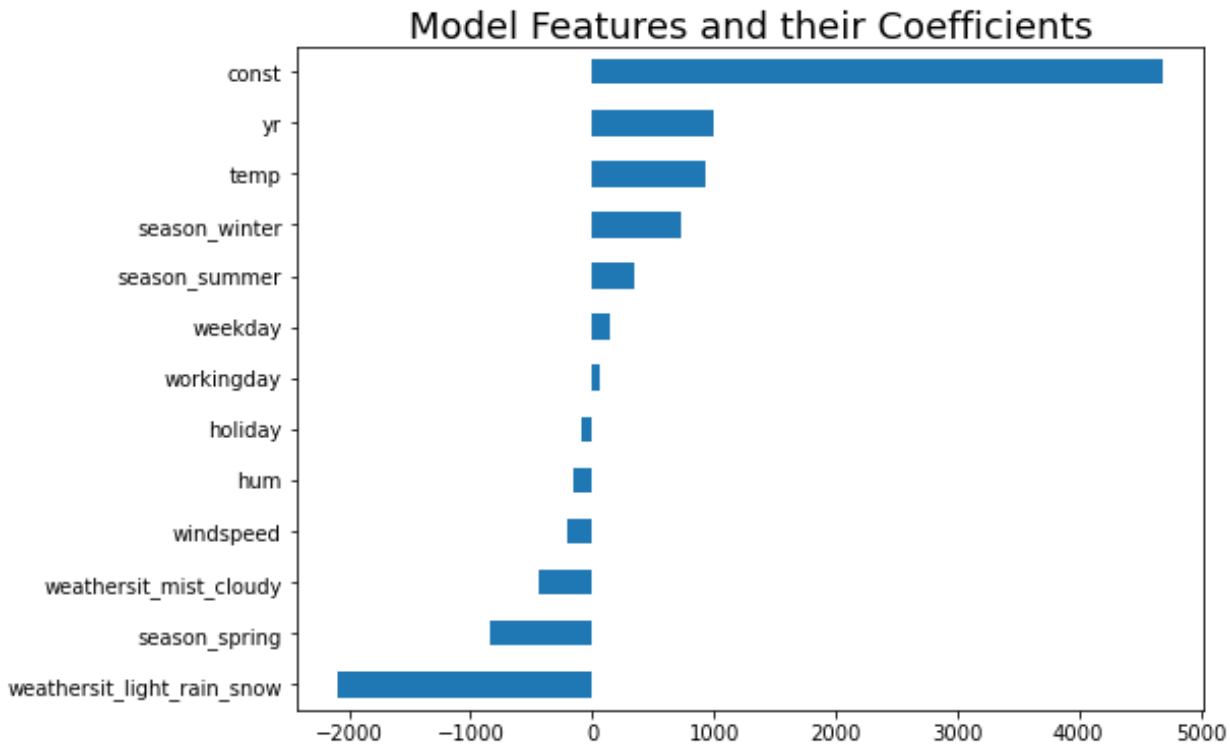




For Actual values greater than 5,000 the Residuals are larger. This indicates that model is under-estimating (predicting less) for the higher values of actual values.

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

Ans:



- **Light Rain & Snow category from variable Weather situation:** has the highest coefficient value of -2093.86, indicating that in rainy or snowing days there is less demand than normal
- **Year:** has the coefficient value of 992.4, indicating that for 2019 there is higher demand as compared to 2018
- **Temperature:** coefficient is 935.64, indicating that for higher temperatures there is more demand. It could be because, people prefer to take a ride rather than walking

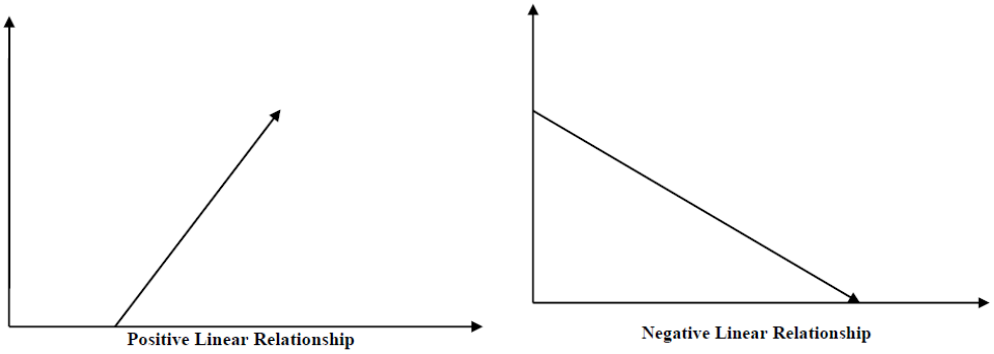
General Subjective Questions

1. Explain the linear regression algorithm in detail.

Overview:

Linear Regression is a statistical model which is used to estimate a variable called dependent variable (denoted with a vector y, where y is continuous) based on one or more variables called independent variables or features (denoted with X, since it is a matrix of numerical or ordinal values).

The LR statistical model analyzes the Linear relationship between the dependent and independent variables. The Linear relationship here means, the impact of change in features (X) on the dependent variable y.



The relationship is positive if on increasing the values of X, y also increases. While, if on increasing X, y decreases then the relationship is negative.

Simple Linear Regression:

It takes the following form:

$$y = mx + c$$

Where,

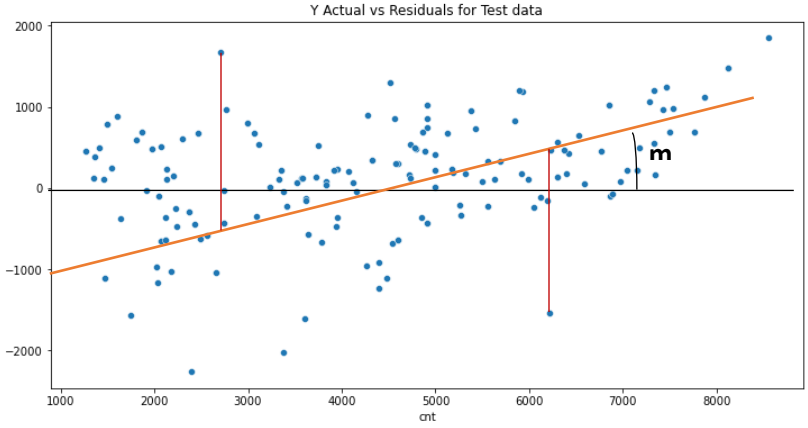
y – dependent variable

x – is the independent variable

m – is the slope that is the degree of the relationship between x and y

c – is a constant value

Let's take the following scatter plot as an example, where y is on y axis and x is on x axis.



We had a scatter plot of points where each point is at location (xi, yi).

Here, the **orange** line is the best fit line we have found which is inclined at slope **'m'** with x-axis and has an intercept of -1,000 that is the line cuts y-axis at -1,000 when **x** is 0. This indicates that value of y will be -1000 when x is 0.

Fitting the Line:

In Linear Regression we try to find this line of best fit and its parameters slope **'m'** & intercept **'c'.**

The line of best fit is the one which passes through the points such it is closest to all the points that it minimizes the error. The error here is defined as the mean residual sum of squared error that is the mean of squares of the residuals.

Residuals are the difference between actual values of y and predicted values of y also called y_hat. In the plot above the residuals are shown with **red** line.

The formula of the error is as follows:

$$MSE = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2$$

where,
$y_i - \hat{y}_i$ is the residual of the ith observation

$$MSE = \frac{1}{n}\sum_{i}^{n}(y_i - x_i \times m - c)^2$$

So, we keep trying different lines with different slopes and intercept, until we find a line which has minimum MSE loss that is the **parameters 'm' & 'c'** for which the MSE is the lowest.

In case of Simple Linear Regression, **parameters 'm' & 'c'** can be estimated using the following:

$$\mu_x = \frac{1}{n}\sum_{i}^{n} x_i$$

$$\mu_y = \frac{1}{n}\sum_{i}^{n} y_i$$

$$SS_{xy} = \sum_{i}^{n} y_i \times x_i - n \times \mu_y \times \mu_x$$

$$SS_{xx} = \sum_{i}^{n} x_i^2 - n \times \mu_x^2$$

$$m = \frac{SS_{xy}}{SS_{xx}}$$

$$c = \mu_y - m \times \mu_x$$

In case of Multiple Linear Regression we use Gradient Descent approach to estimate the parameters/coefficients associated with each of the independent variable.

Goodness of Fit:

To find out the overall quality of the model we use R^2 metric, which is defined as follows:

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$R^2 = 1 - \frac{\sum_i^n y_i - \hat{y}_i}{\sum_i^n y_i - \bar{y}_i}$$

Where,

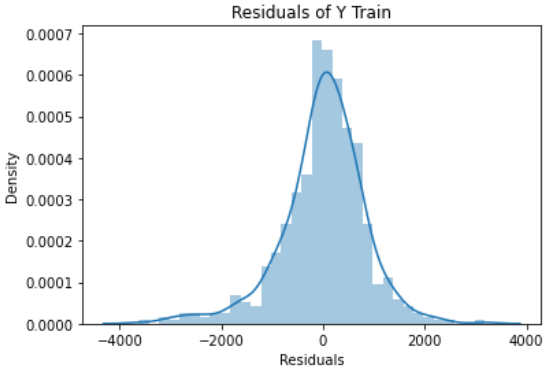
RSS is the residual sum of squares that is the difference between predicted values and actual values

TSS is the total sum of squares that is the difference between average of actual values and actual values.
R^2 has values between (-infinity, 1]

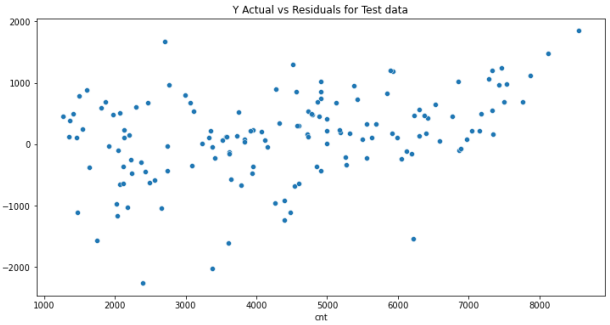
Assumptions:

There are certain assumptions of Linear Regression which must hold true:

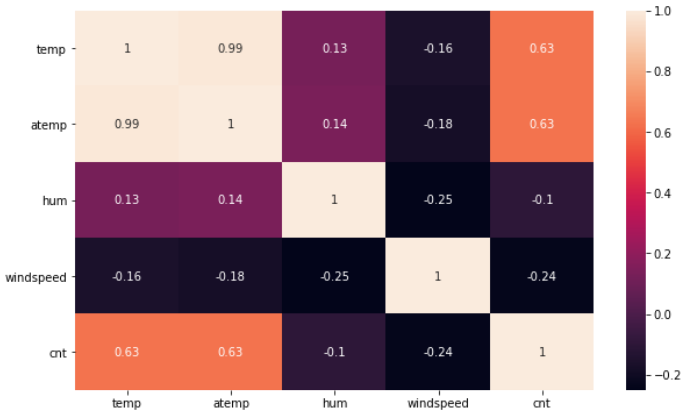
1. **Linear Relationship:** The dependent and independent variables must have linear relationship
2. **Normality:** Errors must be normally distributed, error here mean difference between y actual, and y predicted



3. **Constant Variance:** Errors must have constant variance; this assumption is called homoscedasticity



4. **Independence:** Errors must not be correlated i.e. should be independent of each other.
5. **Multicollinearity:** In case of multiple Linear Regression that is where number of independent variables are more than 1, there should not be any correlation between the independent variables.

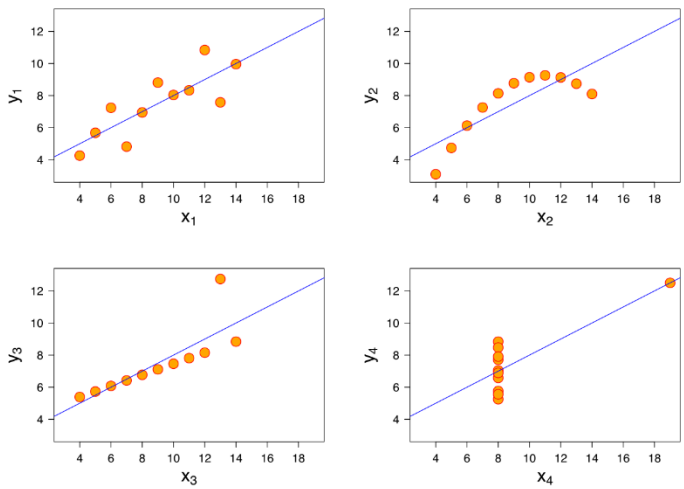


2. Explain the Anscombe's quartet in detail.

Overview:

- 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 Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.
- It was intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

Below are the 4 datasets proposed:



Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*: s^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*: s^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear Regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
R^2 off the linear regression:	0.67	to 2 decimal places

Observations:

- All the 4 datasets have the same Coefficient of Determination (R2) which is 0.67. But, only **X1** and **X3** plots have Linear Relationship between their dependent variables. **X2** does not have Linear but quadratic relation. While **X4** does not seem to have any relationship with its dependent variable at all.
- Similar observations can be found for other statistics like Mean and Variance

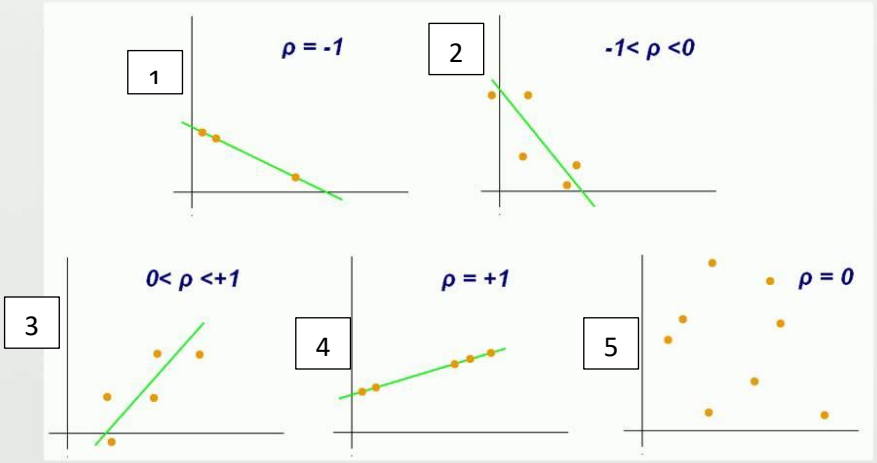
So, this proves that we cannot completely rely on Statistics, but we should always check visually the relationship between dependent and independent variables.

3. What is Pearson's R?

Correlation:
Correlation is a statistic that measures the relationship between two variables. It shows the strength of the relationship between the two variables as well as the direction of the relationship that is whether, they are positively correlated or negatively correlated.

Pearson's R:
One such statistical measure is **Pearson's R** correlation, which measures the **Linear Relationship** between two variables. Its coefficient R lies between -1 & 1, where value less than 0 is indicates negative relationship, while values greater than indicate positive relationship.

Pearson product-moment correlation coefficient



https://en.wikipedia.org/wiki/File:Correlation_coefficient.png

- When coefficient is negative then the slope of the line is > 90 degree as show in plot 1 & 2
- When coefficient is positive then the slope is 0< slope <90 as show in chart 3 & 4
- As we can observed in chart 5, the points are randomly scattered that is there is no linear relationship and hence the coefficient is 0

Calculation:
It is calculated using the following formula:

$$\rho = \frac{\sum_i^n (x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2} \times \sqrt{\sum_i^n (y_i - \bar{y})^2}}$$

where,

Numerator is the covariance between x & y

And Denominator is the standard deviation of x & y, which is to normalize the coefficient and due to this the values are bounded between -1 & +1

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

Scaling:
Scaling means to change the range of values without changing the distribution.

For ex: in below table X contains original values while X-scaled contains the scaled values of X. As we can observe, now the range of X was 1 to 3, while it has been changed to 0.33 to 1.

X	X_scaled
1	0.33
2	0.66
3	1.00

Scaling is performed for following reasons:

- Algorithms converge faster when the features are scaled. For ex: in Gradient Descent algorithm it has been proven that we reach the global minima much faster when the values are scaled.
- Helps in Interpretation of Coefficients of algorithms like Linear Regression. If the scales of different features in Multiple Linear regression will be different, then we won't be able to interpret directly which feature have the highest impact on the dependent variables. But, with all features scaled this interpretation becomes easy and Intuitive

Different Scaling Techniques:

- Normalization or Min-Max Scaling:
    - It subtracts the minimum value in the feature and then divides by the range. The range is the difference between the original maximum and original minimum.
    - It preserves the shape of the original distribution. It doesn't meaningfully change the information embedded in the original data.
    - Note this scaling technique doesn't reduce the importance of outliers.
    - The default range for the feature returned is 0 to 1.

      Formula is:

      $$x_i\_Scaled = \frac{x_i - \min(X)}{\max(X) - \min(X)}$$

      Ex:

X	X_scaled
1	0.0
2	0.5
3	1.0

    -
- Standardized Scaling:
    - It standardizes a feature by subtracting the mean and then scaling to unit variance. Unit variance means dividing all the values by the standard deviation.
    - It results in a distribution with a standard deviation equal to 1.
    - It makes the mean of the distribution 0. About 68% of the values will lie be between -1 and 1.

      Formula is:

      $$x_i\_Scaled = \frac{x_i - \mu_X}{\sigma_X}$$

Where,
μ_X: is the mean of the feature X
σ_X: is the standard deviation of the feature X

Ex:

X	X_scaled
1	-1.2
2	0.0
3	+1.2

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

Overview:

- In ordinary least square (OLS) regression analysis, multicollinearity exists when two or more of the independent variables demonstrate a linear relationship between them.
- With multicollinearity, the regression coefficients are still consistent but are no longer reliable since the standard errors are inflated. It means that the model's predictive power is not reduced, but the coefficients may not be statistically significant with a Type II error.
- VIF is commonly used tool to detect whether multicollinearity exists in a regression model. It measures how much the variance (or standard error) of the estimated regression coefficient is inflated due to collinearity.
- It can be calculated using the following:

$$VIF_i = \frac{1}{1 - R_i^2}$$

  Where R_i^2 represents the unadjusted coefficient of determination for regressing the ith independent variable on the remaining independent variables.
- VIF values less than 2 are good, > 2 and < 10 needs to be reviewed, but > 10 are highly correlated and must be excluded.

Why VIF can be infinite?

- If R_i^2 for any i^{th} independent variable is equal to **1**, then it will make the denominator **0** which will lead the value of VIF as infinite as any constant divided by 0 is infinite.
- The R_i^2 can be 1 in scenario where we have built a perfect model and are able to predict the i^{th} variable from other independent variables or if there exists an independent variable which is perfectly correlated with i^{th} variable (that is person correlation coefficient = 1)

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

Overview:

In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

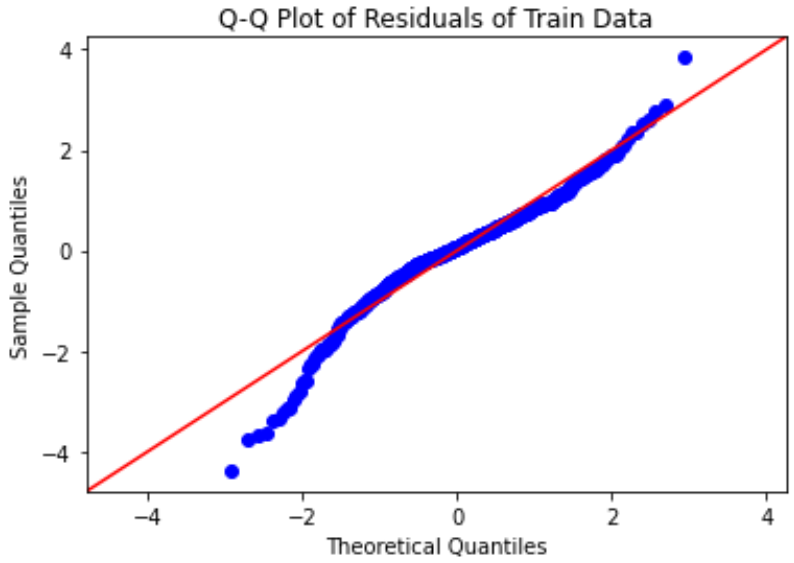
First, a set of variables and their quantiles is calculated. Then, a point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the

first distribution (x-coordinate). Thus, the line is a parametric curve with the parameter which is the number of the interval for the quantile.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x.

For example, below is a sample Q-Q plot and as we can observed most of the central points lie on the red line (y = x), but the points at the tails are not on the red line.

These off points show that they do not have different deviation in both the distribution. Like the bottom left point as quantile of around -3 in x distribution, while it is less than -4 on y distribution, which indicates that this point is farther away from mean in y-distribution as compared to x-distribution.

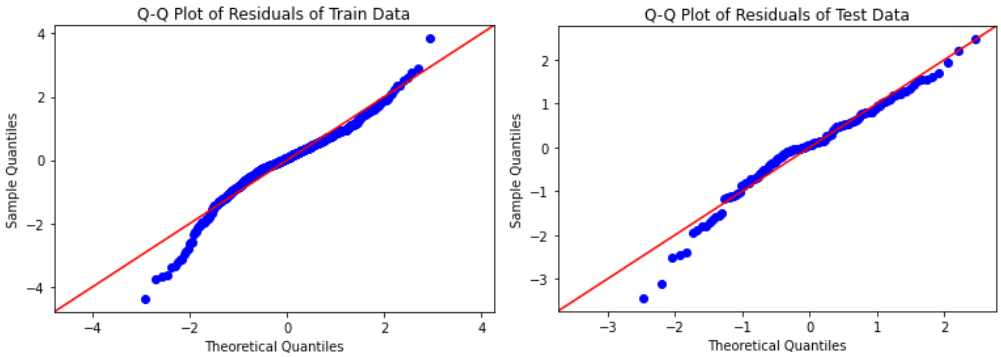


Importance of Q-Q Plot in Linear Regression:

One of the assumption of Linear regression is that Errors or Residuals should be normally distributed. The Q-Q plot helps use to validate that assumtion graphically.

- In case of Linear Regression we compare the distribution of Residuals against the Standard Normal distribution (mu=0 and std=1).
- The y-axis contains the Standardized Residuals, while x-axis contains the quantiles of Standard Normal distribution. This comparison helps us to validate if the residuals are normally distributed or not.
- If the residuals are normally distributed then all the points will lie on the red line, else the points will be scattered around.

Let’s take the example of our case study above.



These are the Q-Q plots are of Train and Test data residuals. As we can clearly observe, most of the points are on the red line (y=x), but there are points far of the red lines.

This deviation from the red line shows that the residuals/points are somewhat normally distributed but not completely.