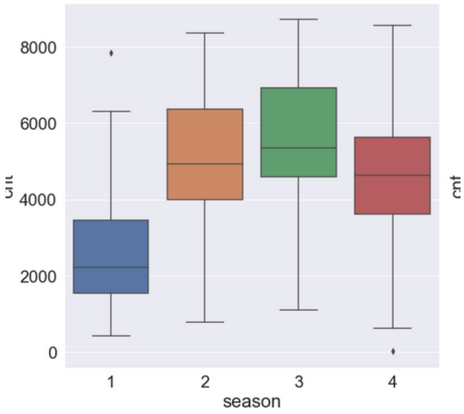
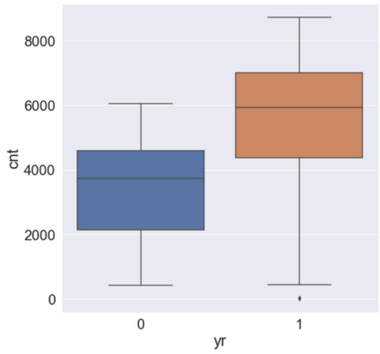
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

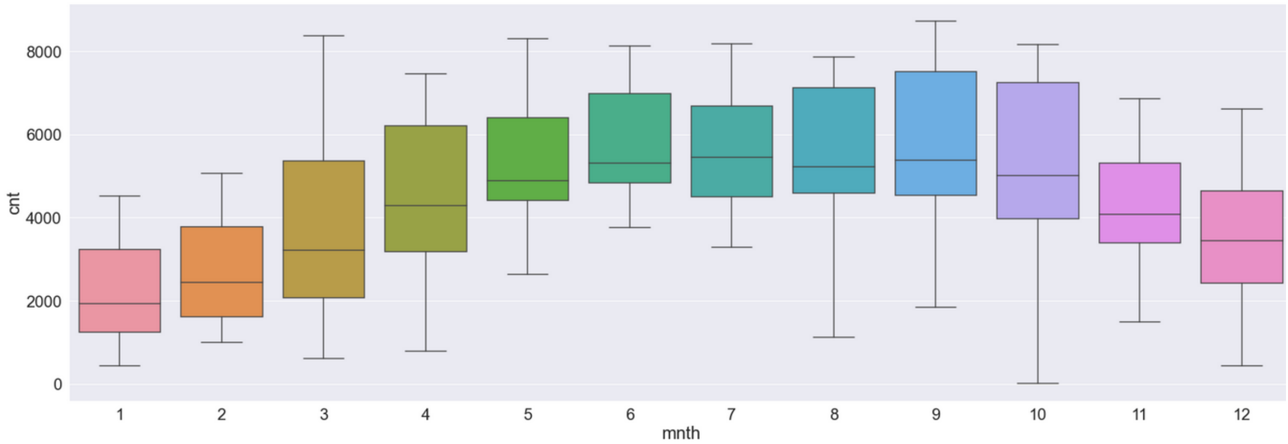
Answer:
1. season: when season is 3, i.e. fall, more counts of bikes.



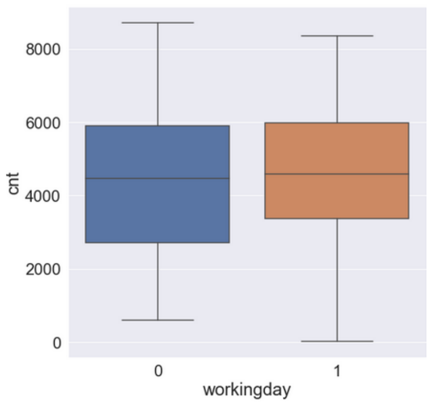
2. year: when year is 1, i.e. 2019, more counts of bikes.



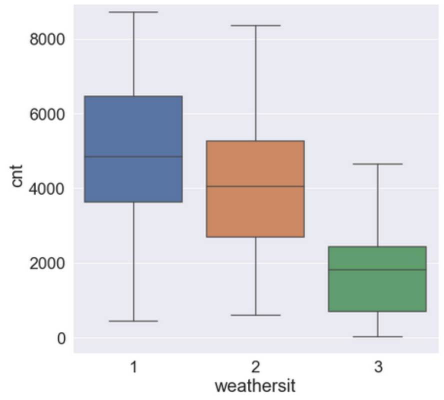
3. month: when month is from 5 to 10, i.e. may to October, more counts of bikes.



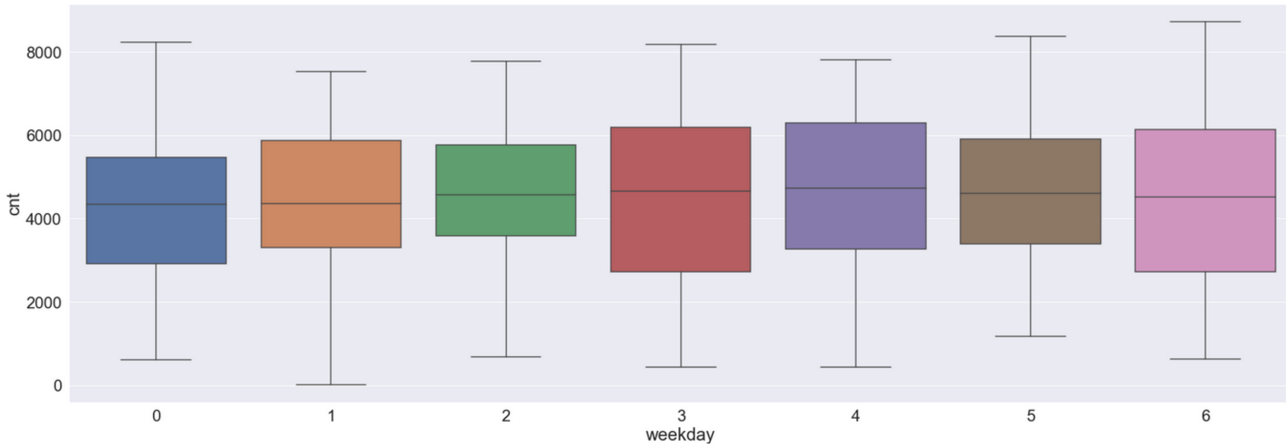
4. working day: no such effects on bike counts.



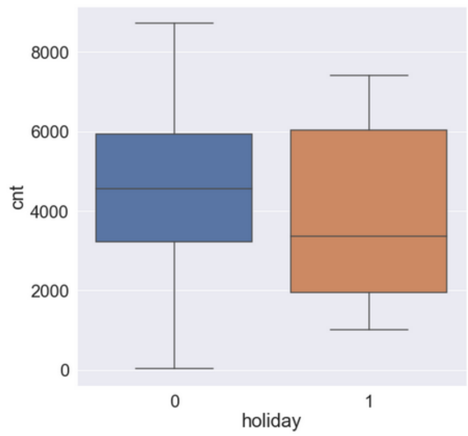
4. weather situation: more bike count at 1, i.e. clear weather.



5. weekday: no such effect on weekday on bike counts.



6. holiday: more counts when there is not a holiday.



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

Answer:
During manual one-hot encoding on a single column, we usually create dummy variables out of all the available discrete categorical values in the columns. During these process, dropping the first column signifies that when all the other dummy columns are 0 in a row, the dropped value is the value in that specific row.
It is important to drop first column to reduce the number of column as wel complexity for our model to learn.

When dummy variables are created, the values given are wither 0 or 1, where 0 represents 'negative' and 1 represents 'positive'.
Thus if a column have values like A,B,C, and D. New dummy columns created are A|B|C|D.
So for the row where column values was A, in the new dummy columns dataframe, values for A,B,C,D will be 1,0,0,0 respectively.
Accordingly for B; C; ; and D, values will be 0,1,0,0 ; 0,0,1,0 : and 0,0,0,1 respectively.
e.g.
season column – fall, winter, spring, summer

	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
171	21-06-2018	fall	0	6	0	2	1	2	27.914153	31.88230	77.0417	11.458675	774	4061	4835
79	21-03-2018	summer	0	3	0	1	1	2	17.647835	20.48675	73.7391	19.348461	401	1676	2077
0	01-01-2018	spring	0	1	0	6	0	2	14.110847	18.18125	80.5833	10.749882	331	654	985
265	23-09-2018	winter	0	9	0	5	1	2	24.975847	26.10625	97.2500	5.250569	258	2137	2395

creating dummy columns: 0 & 1 assigned to each column.

	fall	spring	summer	winter
171	1	0	0	0
0	0	1	0	0
79	0	0	1	0
265	0	0	0	1

dropping first column. When 1 is present at each specific column, that value is true for the row.

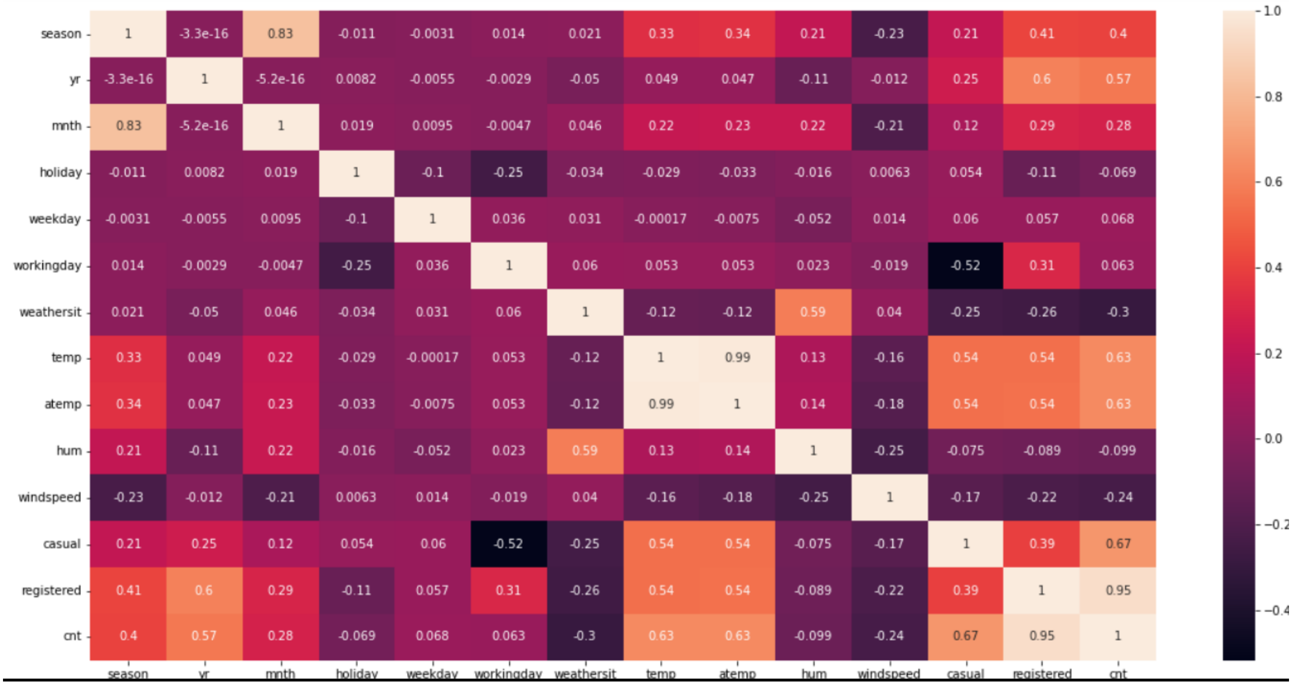
	spring	summer	winter
0	1	0	0
79	0	1	0
265	0	0	1

When all columns are 0, dropped value is true for the row.

	spring	summer	winter
171	0	0	0

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

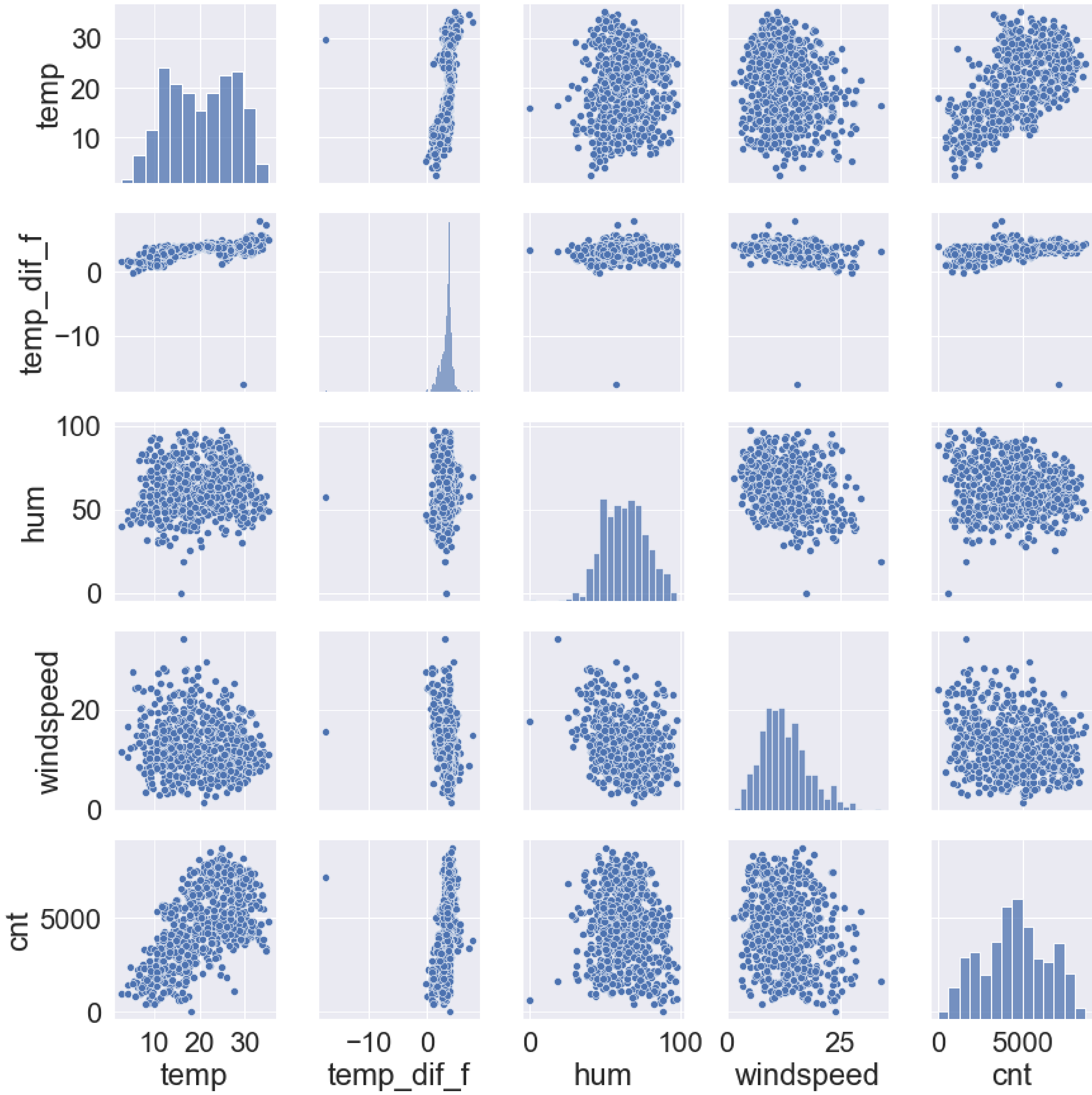
Answer:
column 'atemp' have highest correlation with the target variable 'cnt' with value of (0.6306853489531029).



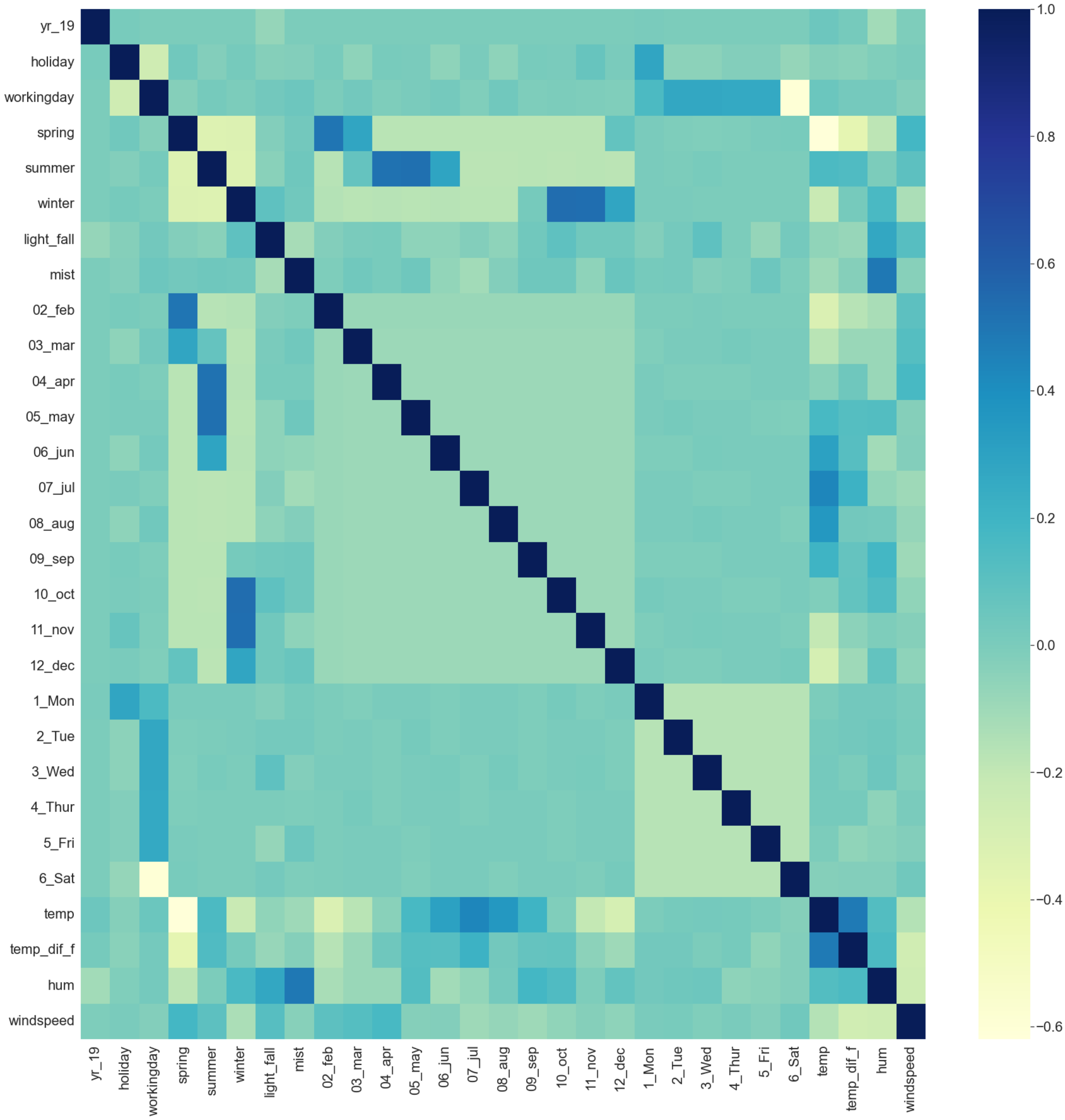
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:

1. Linear relationship assumption: linear relationship present between target and feature variables(continuous).

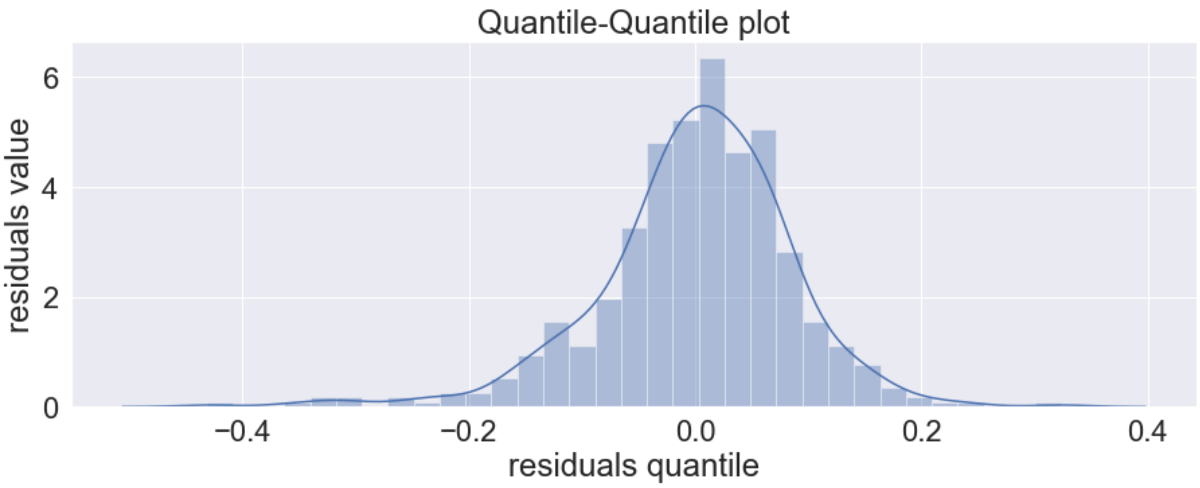


2. no correlation assumption: no collinearity observed between independent feature variables.

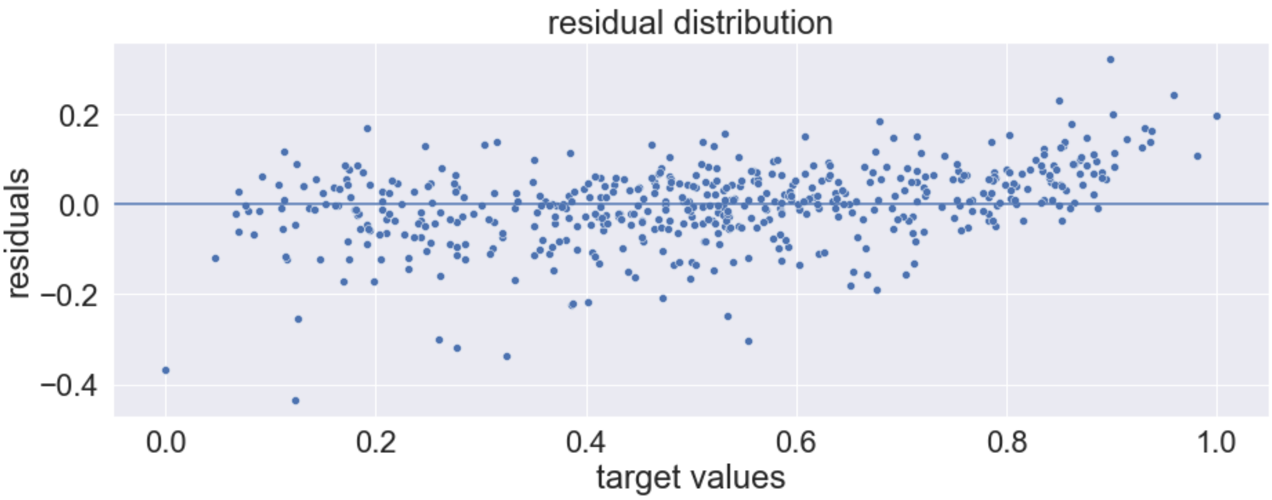


3. error terms

a. normal distribution of error terms and zero mean of errors



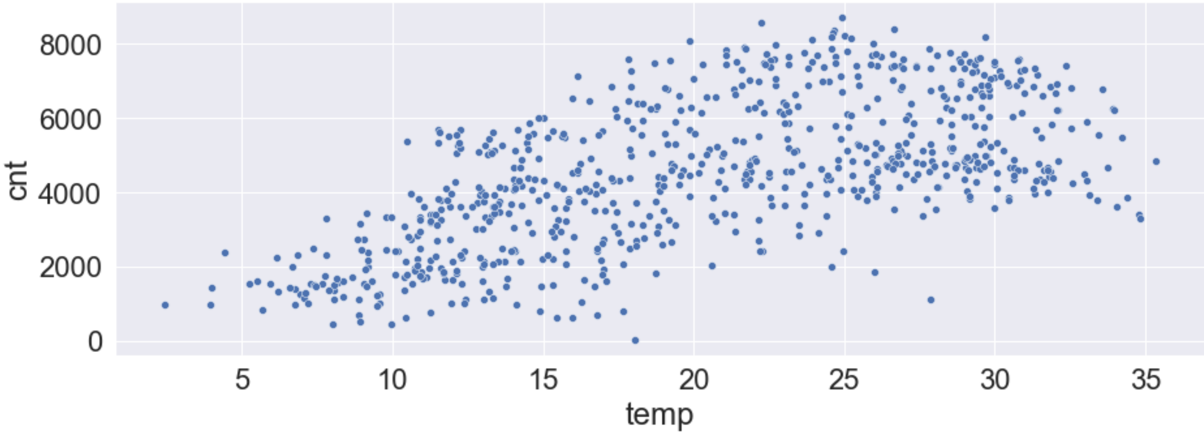
b. residuals are homoscedasticity in nature, i.e. have constant variance.



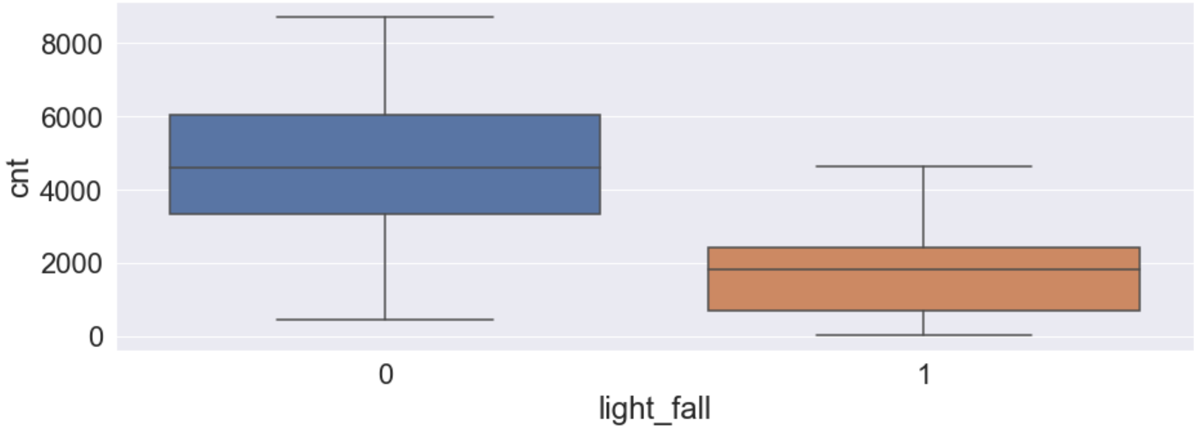
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer:
features having very high coefficients absolute value are more contributing to the shared bikes demand.

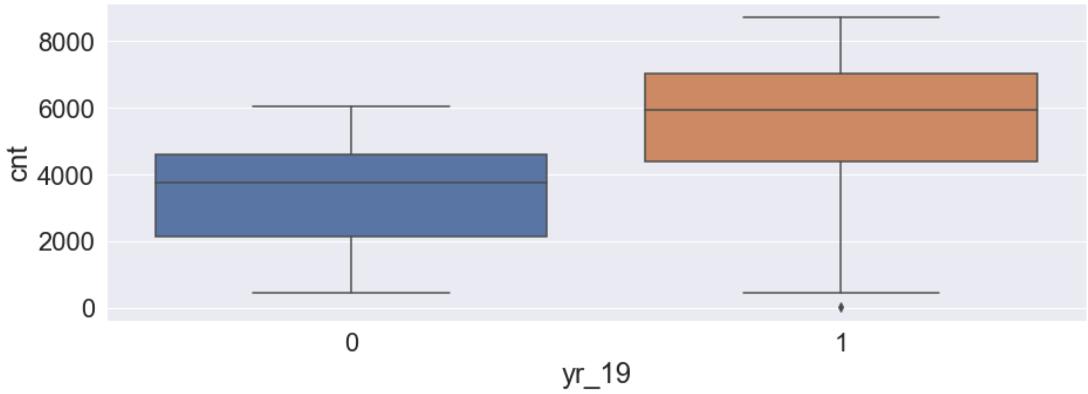
- temp: temperature: 0.434234
    - bike demand increases with temperature.



- light_fall: light fall weather situation: -0.254504
    - more bike demand when there is no light fall.



- yr_19: year 2019: 0.232232
    - more demand for year 2019



1. Explain the linear regression algorithm in detail.

Answer -->
Linear Regression is Supervised Machine Learning Algorithm which find the best linear fit relationship between dependent and independent variables.

The best fit line is created by minimizing sum of residual squares, i.e. sum of squares of difference between predicted and actual values.

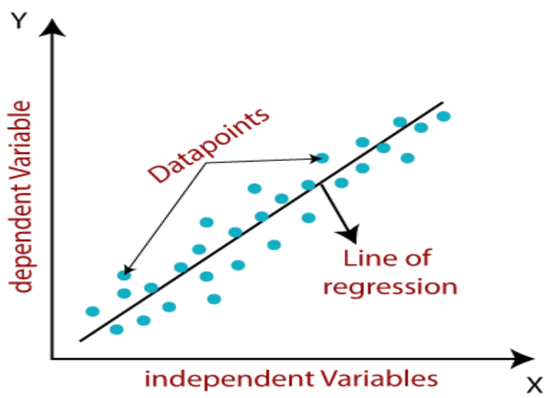
$$y=\beta_0X_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n+\ \Omega$$

β – weights/coefficients/feature_gradient

Ω - bias/intercept/default_value

- **weights**- gradient value that signifies, change in the target value with unit change in one of the feature variable when all other variables are held constant.
- **bias** – default value when all the features are not considered and are insignificant.

Basic Assumptions: before using this algorithm are:

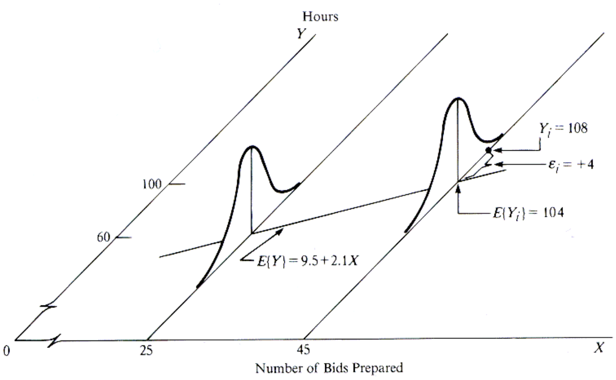


1. almost linear relationship between target and each individual feature variables, checked by two-dimensional pairwise scatter/line plot.

2. feature variables that are recorded without any error, are independent of each other, i.e. no multi-collinearity in our data.

3. error terms:

FIGURE 1.6 Illustration of Simple Linear Regression Model (1.1).



- normal distribution of each error terms.
- mean of error terms is 0, as we have both equal numbers of positive and negative errors.
- error terms are independent of each other, no correlation between the error terms, as our feature variables are independent.
- all error terms must have equal standard deviation.
- Error terms plot should not represent a pattern, random values for error terms.

$$Y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n$$

Where Y = Dependent Variable(DV)

X_1, x_2, x_n – Indepdent Variable(IV)

b_0 – intercept

b_1, b_2 – coefficients

n – No. of observations

Hypoothesis Testing: we assume a null hypothesis that the features are insignificant, thus coefficients of these features in the equation are 0.
Thus when we study on model, we infer about the significance of the features by:

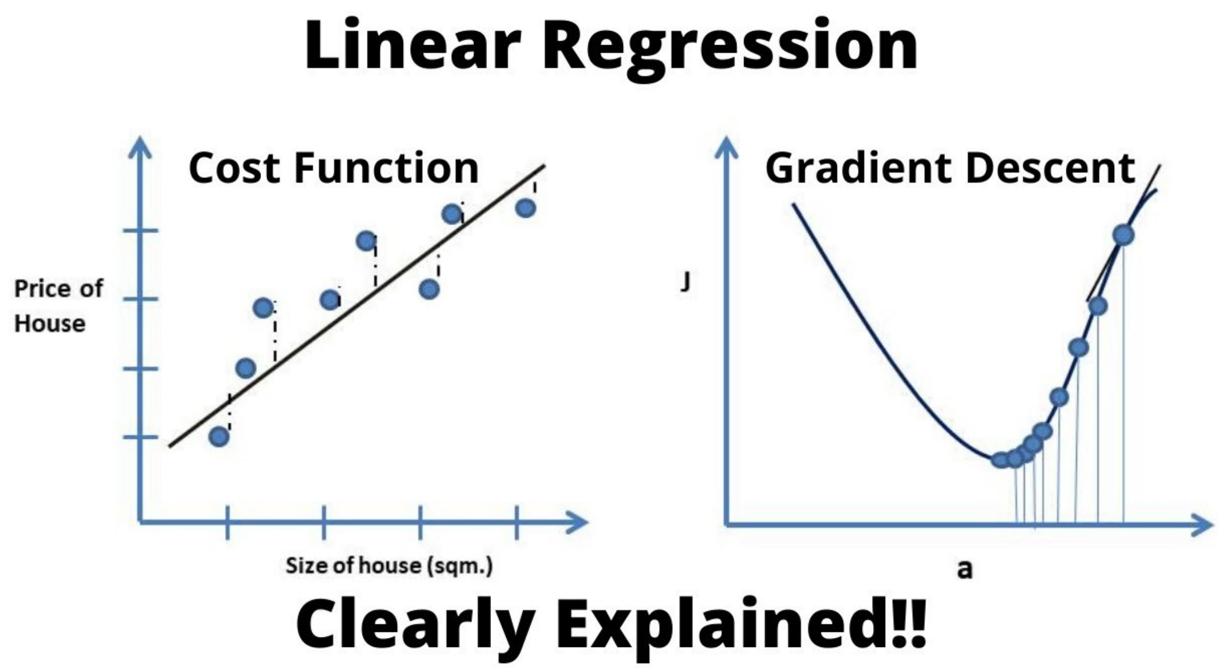
- p-value: if value is less than the significance level, i.e. 0.05 (default), the feature is significant, otherwise it's insignificant.
- Coefficient: if the coefficient value is 0, the variable is insignificant.
- VIF: if the variation inflation factor of the value is very high, we remove those features and check our all values again.
- Adjusted R^2: if the value is very less, we remove some variables which are not logically correct to use in our model.

Cost Function: we determine the values of the coefficients using this method. Cost function is basically mean square error of the function.

$$Cost\ Function(MSE) = \frac{1}{n}\sum_{i=0}^{n}(y_i - y_{i\ pred})^2$$

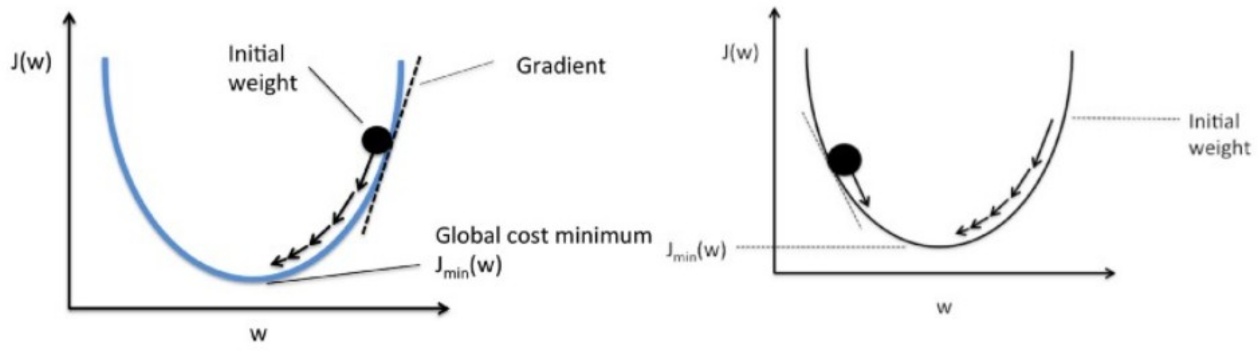
Replace $y_{i\ pred}$ with $mx_i + c$

$$Cost\ Function(MSE) = \frac{1}{n}\sum_{i=0}^{n}(y_i - (mx_i + c))^2$$



gradient descent algorithm: In linear regression model, we normally use the iterative algorithm to find th eour best possible values of coefficients and intercept.
Usually the best possible value is obtained at the global minima of the cost function curve which is also called convex curve.
We choose random values for the coefficients and intercepts, e.g. 0 & 0, and a very small learning rate (a step

size value for iterations) to move to the next values.



The equation for the next values of coefficients and intercept is the difference in the initial value and the learning rate times the gradient with respect to the specific weight/bias at the initial value.

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left((h_\theta(x_i) - y_i)x_i\right)$$

If our initial value is greater than the minima, the gradient will be positive and the new value will be less than initial value. Otherwise if the initial value is less than the minima, gradient will be negative and thus the sign will reversed in the equation and the new value will be more than the initial value.

Using this iterative approach, we reach towards the optimal values of the coefficients and intercepts where cost function is minimum globally.

Model Training:

1. Split our dataset into training, validation and test dataset.
2. Fit our model using training dataset, our models learns specific values like mean, standard deviation and distribution from the dataset, and calculates coefficients and intercepts for the best fit line using gradient descent method.
3. Use validation dataset to validate our model and predict values on the validation dataset and calculate r2 score.
4. Use the model to predict values on the test dataset, calculate r2 score for the predicted values and actual values in the test dataset.
5. Compare r2 score for validation and test dataset and check for under-fitting or over-fitting.
6. If some major discrepancy found, do tuning (coarse-fine) on the dataset using recursive feature elimination and manual testing.

2. Explain the Anscombe’s quartet in detail?

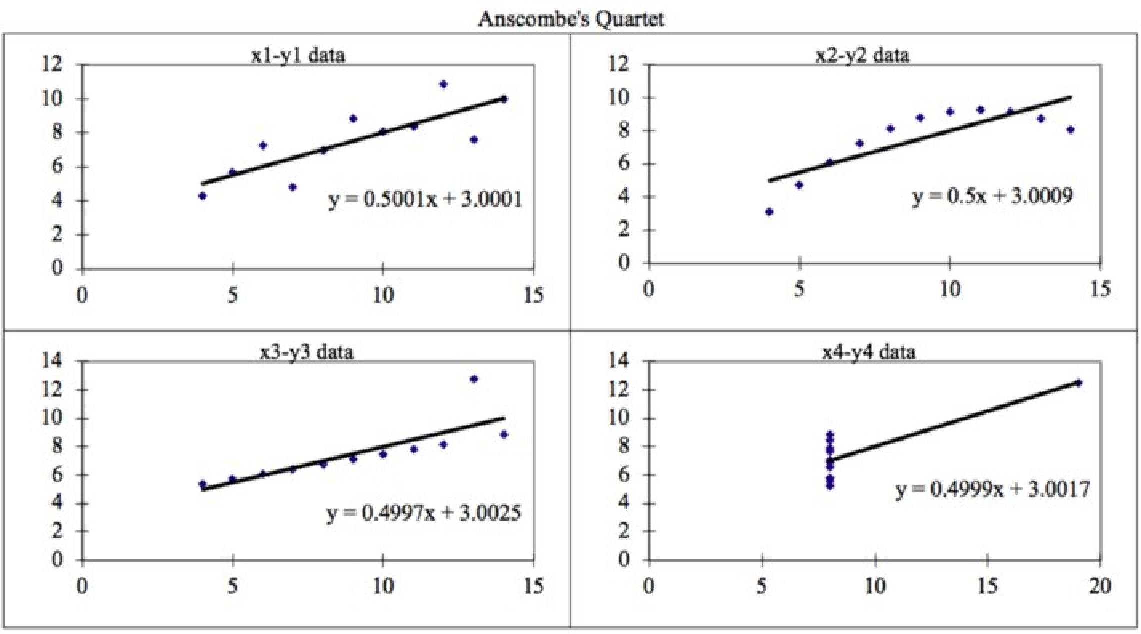
Answer -->

Anscombe’s quartet is a set of 4 graphs that advices to look on our data before building and running linear regression models.

Linear regression models have some features that may create problems:
1. sensitiveness to outliers.
2. build linear relationship only.
3. pre-acknowledges all assumptions are being held true.

Sometimes for different types of dataset, the statistic values of the best fit line provided by the linear regression is almost the same, irrespective of the nature of values.

This problem is well defined with Ansombe’s Quartet.



Explanations for graphs:
1. linear regression model is perfectly aligned with the dataset as the plot represents linear relationships.
2. though the relationship is non-linear, model still creates a linear relationship.
3. due to presence of some outliers, the best fit line is deviated from the best possible alignment, if these outliers were absent, our best fit line would be passing from all the data points.
4. few distance outliers and values concenterated over a single points may creates a best fit line which is totally false in nature.

3. What is Pearson’s R?

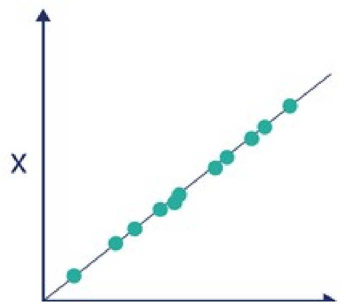
Answer-->

Pearson Correlation Coefficient determines linear relationship/correlation between two variables.

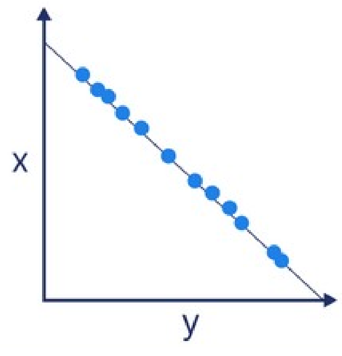
- It’s values ranges from -1 to 1.
- 0 to 1, signifies positive correlation, means with one variable, other variable also increases.
  - **e.g.** population and resource requirements.
- -1 to 0, signifies negative correlation, means with increase in one variable, other variable decreases.
  - **e.g.** life and price of a non-remodificable/resuable product.
- 0, signifies no correlation between variables.
  - **e.g.** new additional feature and supply chain of a product.
- -1 and 1, signifies perfect correlation, either negative or positive respectively.

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

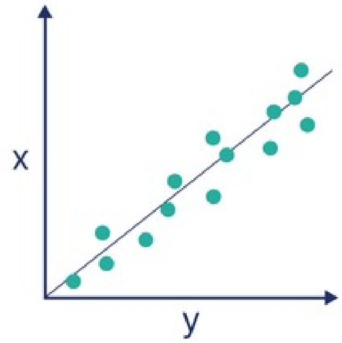
Perfect positive correlation



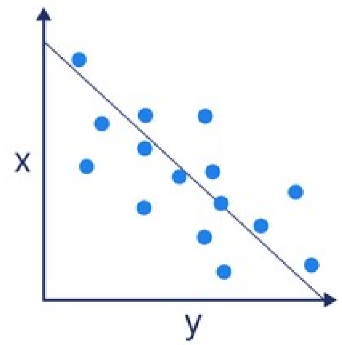
Perfect negative correlation



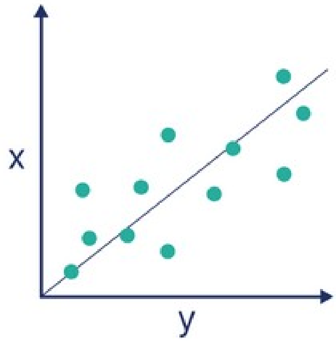
Strong positive correlation



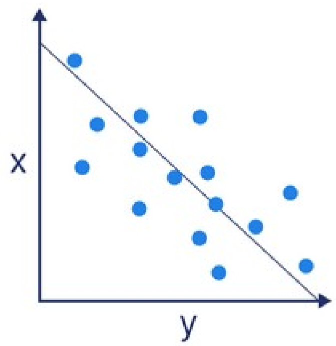
Strong negative correlation



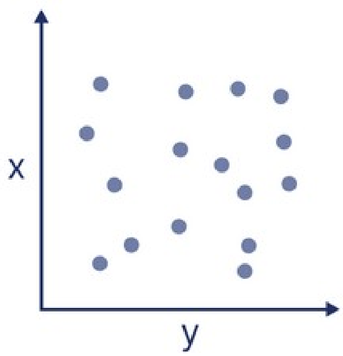
Weak positive correlation



Weak negative correlation

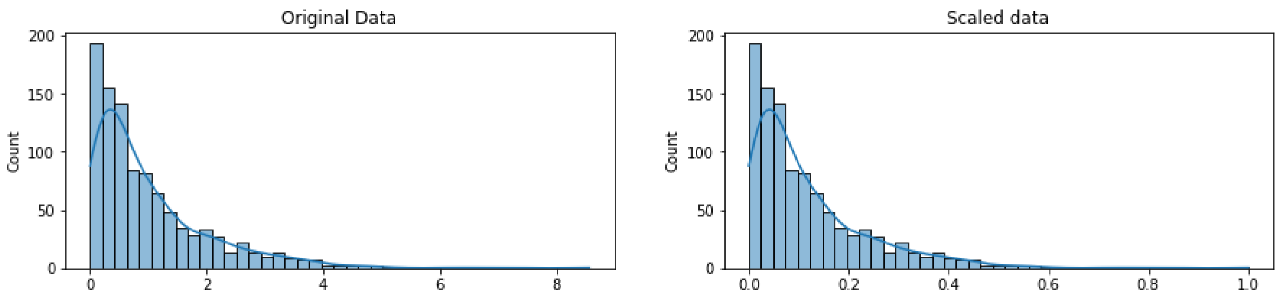


No Correlation



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

ANSWER -->
Scaling: Also known as feature scaling, is Transformation of numerical data so that if fits within a specific scale/range, i.e. -100 to 100, or 0 to 1.



Importance of Scaling
Numerical Data in a dataset are generally very far from each other, thus algorithms will take more time to understand the data resulting in low accuracy.
Hence, data values are brought closed to read other so that we can create properly train our models.

Normalized Scaling
- Values are distributed between minimum and maximum values so that they scale from 0 to 1.
- Minimum value is now 0.
- Maximum value is now 1.
- Useful for data following normal/gaussian distribution.
- Formulae for a (i) element in column(x):
    - $(x_i - min(x))/(max(x)-min(x)$

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardized Scaling

- Values are distributed around mean, as units of standard deviation away.
- Mean value is now 0.
- Values greater than mean are positive deviation away while values less than mean are negative deviation away.
- Can handle outliers also.
- Formulae for a (i) element in column(x):
    - (x_i – mean(x))/standard_deviation(x)

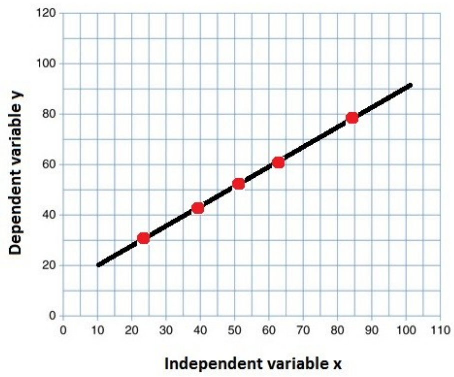
$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

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

ANSWER -->

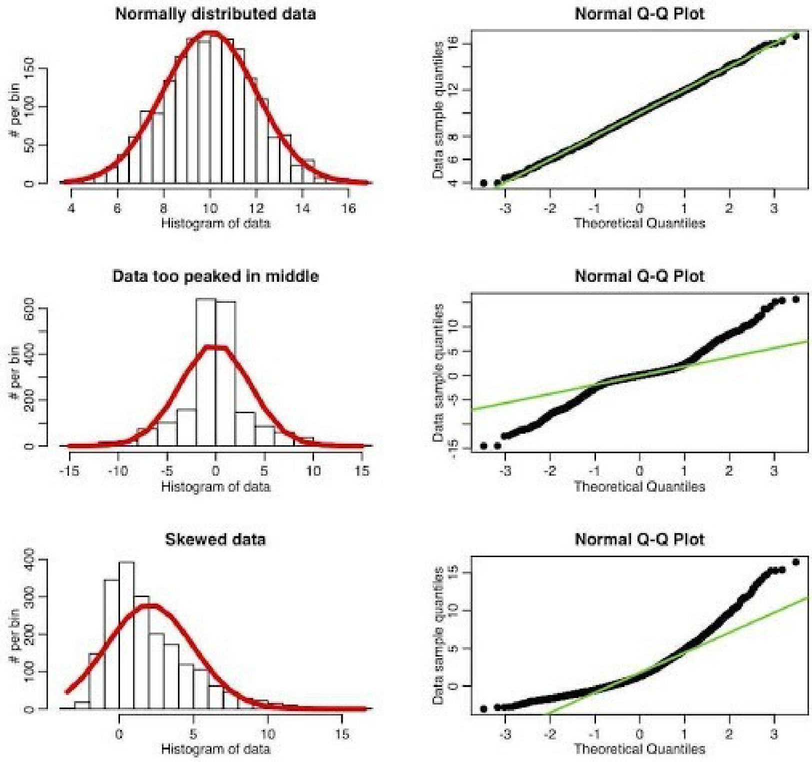
Variance Inflation Factor: It describes multi-collinearity in our model. It selects one of the variable among parameters as target and other variables as feature.
If a variable has a relation with all other variable then, i.e is perfectly collinear and completely defined by all other variables, VIF is very high i.e. infinite (∞).

$$VIF = 1 / (1-R^2), \text{ if } R^2 = 1, VIF = \infty.$$

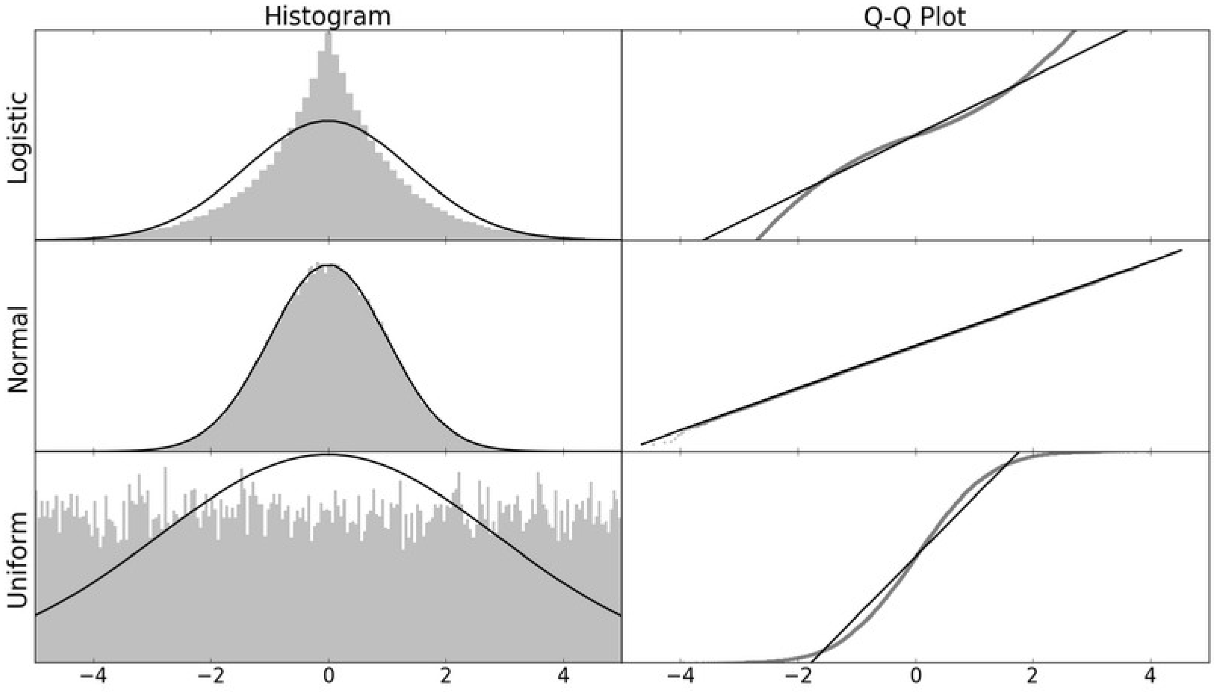


6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

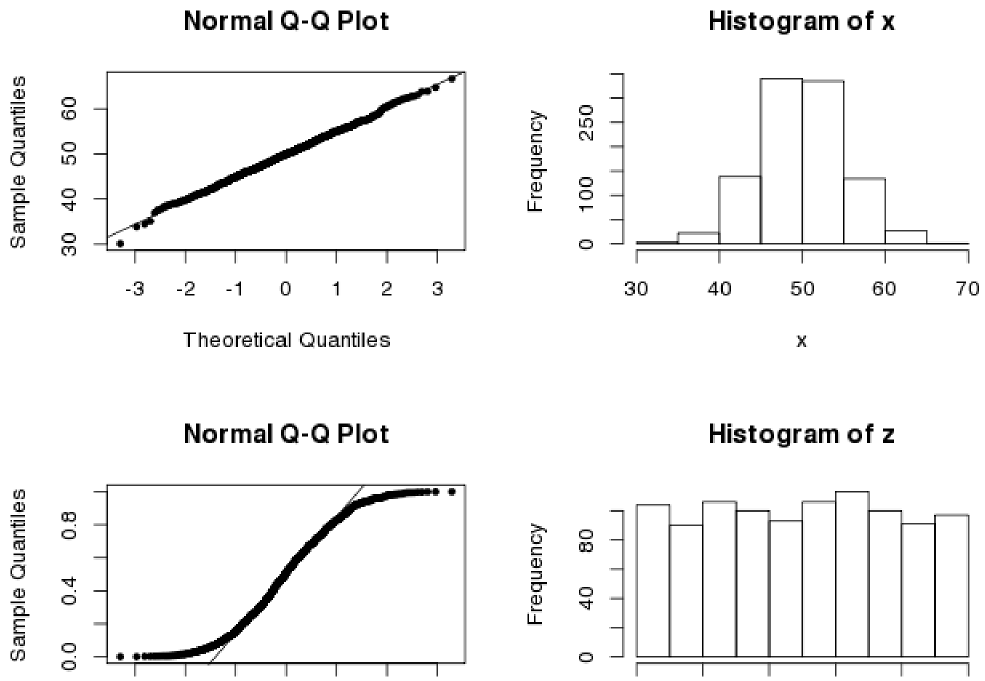
Answer-->
Quantile-Quantile plot or Q-Q plot is a scatter point distribution between the data points and the quantile values.

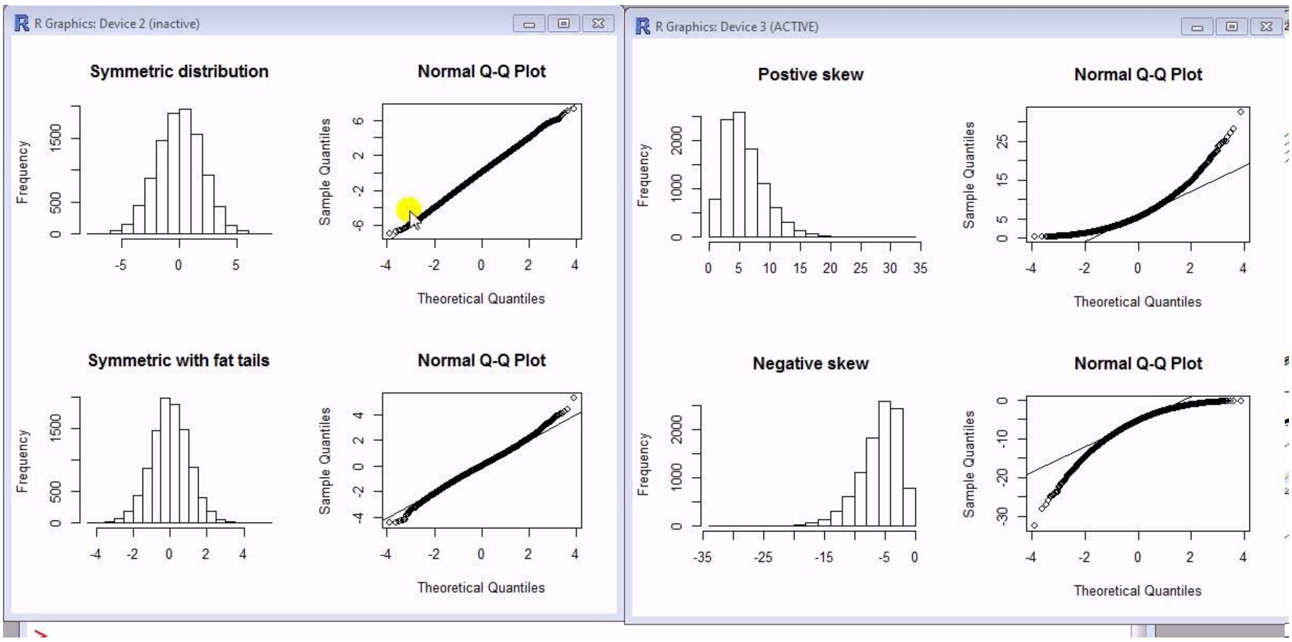


If the scatter points fits in a single line, we may assume our distribution assumption normal/uniform/skewed is correct.



Steps:
1. plot the data on the chart.
2. assign quantile values to each data point.
3. assume a uniform/normal/skewed distribution.
a. uniform distribution – distance between each quantile value is constant.
b. normal distribution – points at the median position are closer to each other whereas points faraway fro mean/median are faraway from each other.
c. skewed distribution – points at one of the end are closer to each other whereas points at the other end are faraway from each other.
4. plot a scatter graph with data points and their quantile values. If all of these points lie on a line, the distribution assumed is correct, if wrong assume different distribution.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Histogram of x
Frequency
x
Normal Q-Q Plot
Sample Quantiles
Histogram of z
Frequency


R Graphics: Device 2 (inactive)
Symmetric distribution
Frequency
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Symmetric with fat tails
Frequency
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
R Graphics: Device 3 (ACTIVE)
Postive skew
Frequency
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Negative skew
Frequency
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles