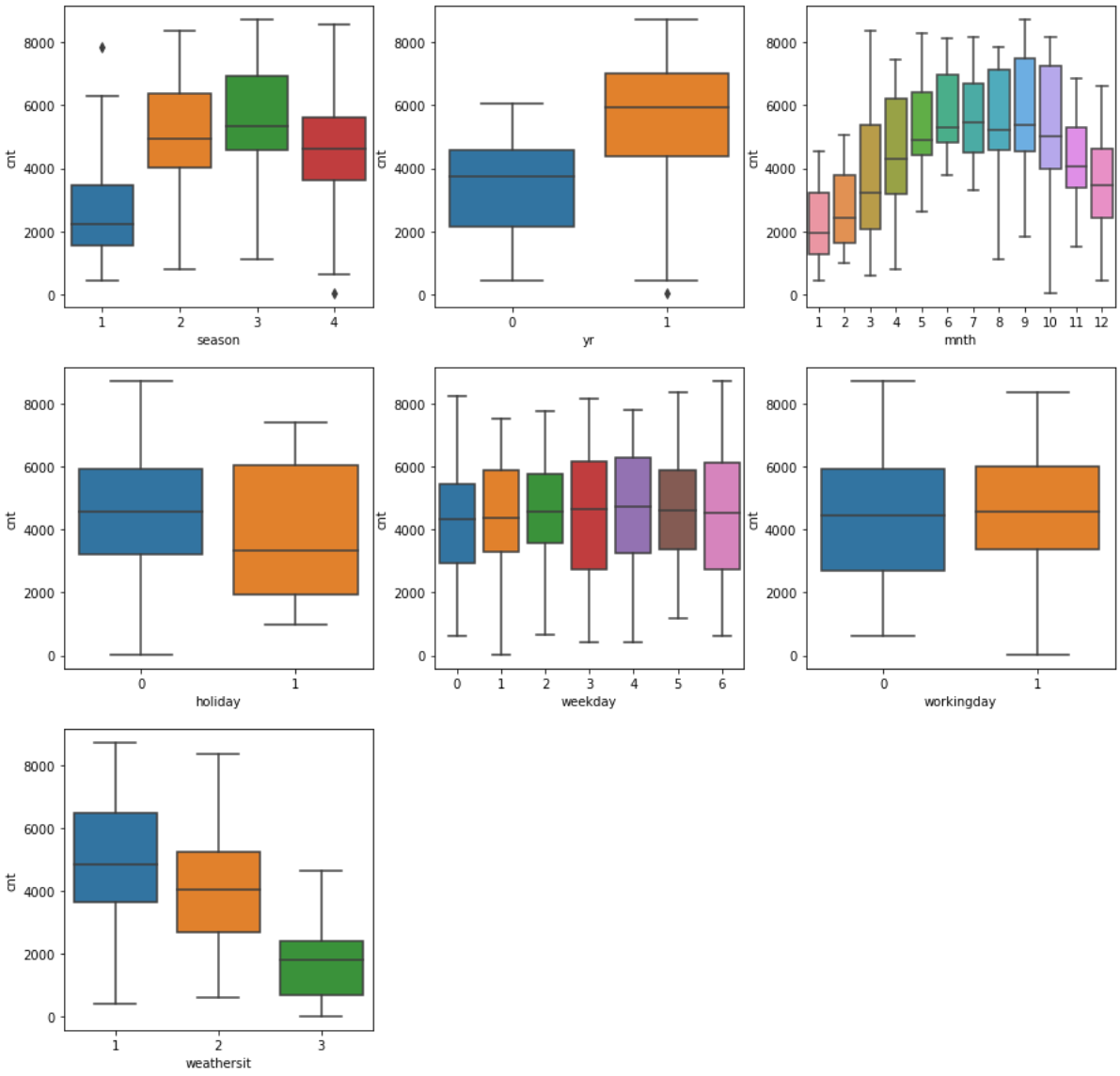
Assignment-based Subjective Questions

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

Ans: We have 7 categorical features in our dataset, following are the effects of them on the dependent variable;

I. **Season** : Season 1 i.e. Spring has the least number of demands, whereas in Fall demand is the highest.
II. **Year** : Year 2019 has significantly higher demands than 2018.
III. **Month** : Month wise we see an increasing trend with June, July, August & September being the highest and then gradually decreasing at the end.
IV. **Holiday**: Holidays have less demands
V. **Weekday**: We see almost similar demand for all the days of the week.
VI. **Working day**: Working day wise also we see similar demand
VII. **Weathersit**: There are no demand on weathersit 4: 'Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog', whereas 1:'Clear, Few clouds, Partly cloudy, Partly cloudy' has the highest.

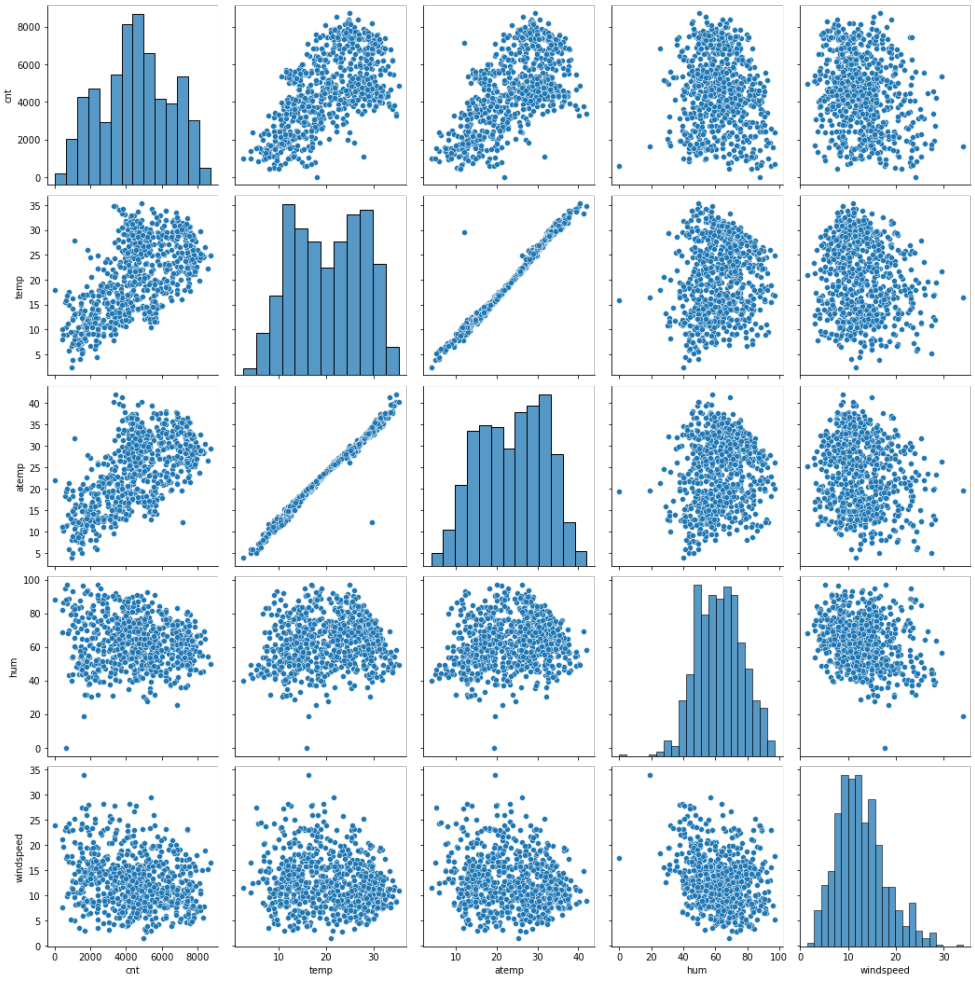


2. Why is it important to use drop_first=True during dummy variable creation?

Ans: It is important to use drop_first = True during dummy variable creation because first of all n-1 columns with the binary combination is enough to represent n categories in a variable. Also we are reducing the number of features by 1 whenever we are using it as it can be correlated and hence it will result in Multi collinearity.
For example if we have 3 colors then 2 column with combination 00,01 & 10 can explain all 3 colors and reduce the column number by 1 which reduce redundancy

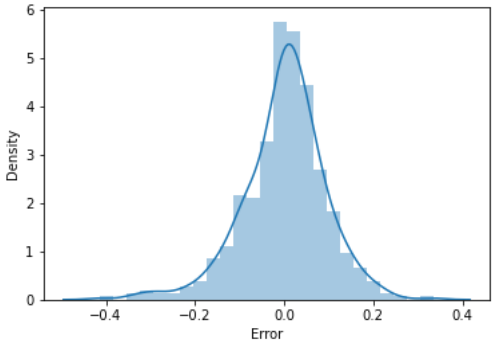
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

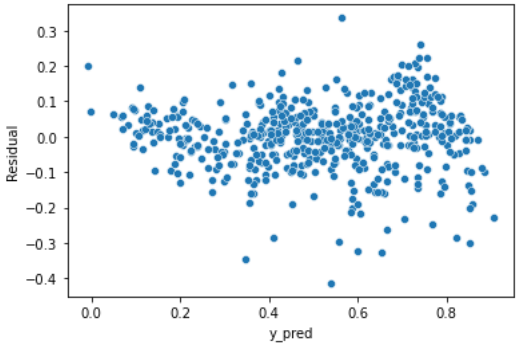
Ans: From the pair-plot and correlation of the numerical variables with target variable. We see ‘atemp’ has the highest correlation coeff as 0.631 and ‘temp’ being the next very close with 0.627.



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

Ans: After plotting the distribution plot of residuals, we see it is normally distributed around 0. Also we can see they are independent of each other and variance remains the same.





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

Ans: Based on the model, 3 features contributing significantly are:

- temp: with coefficient 0.5314
- hum: with coefficient -0.2642
- yr : with coefficient 0.2277

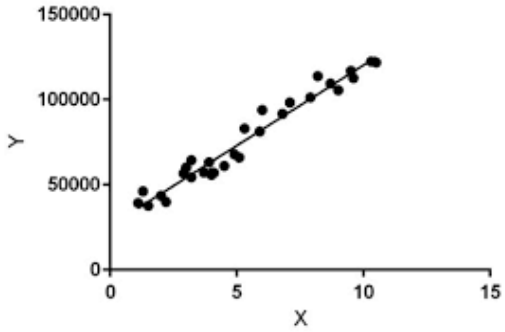
General Subjective Questions

1. Explain the linear regression algorithm in detail

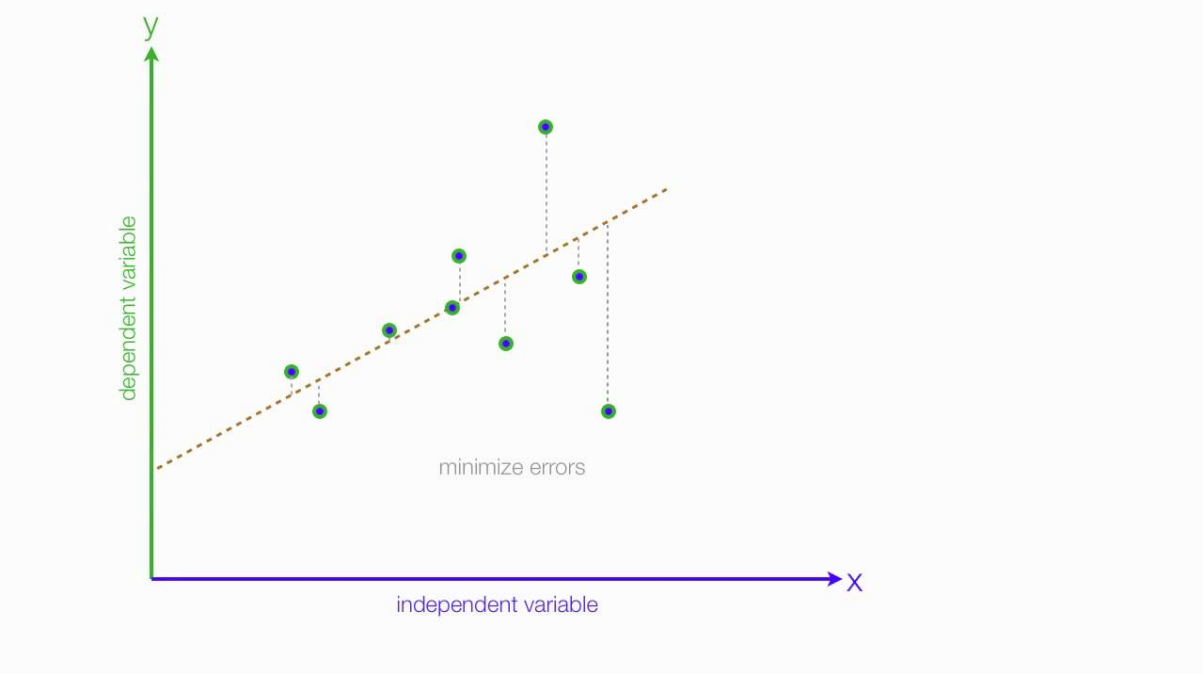
Ans: Linear Regression is an Machine learning algorithm use to analyse the linear relationship between a dependent and a given set of independent variables. It shows how the dependent variables affected by the change in independent variables. For example cost price of apartment can have independent variables .It is generally represented by the following equation:

$$Y\_pred = b0 + b1*X$$

Where y is the dependent variable, X is the independent variable, b0 is the intercept of the equation which represent the value of y when X=0 and b1 is the slope which represent the effect of change in X over y.



The Linear Regression tries to find the b0 and b1 values such that it results in a best fit line. Ordinary least square (OLS) method is used to find the best fit line by minimizing the sum of squares of the distance between data point and regression line.



The difference between the actual data point and the regression line at a given X is the residual which refers to the error terms.
OLS method use the gradient decent algorithm to minimize the sum of the squares so that we get the best fit line.

Assumption of linear regression:
1.There exist a linear relation between the dependent and independent variable.
2.Error terms are normally distributed.
3.Error terms are independent of each other.
4.Error terms have constant variance i.e. Homoscedasticity.

Types of linear Regression:

1. Simple Linear regression:
   Where there is a single independent variable which is contributing to the change in independent variable. For eg windspeed is independent and rotation of windmill is dependent variable. It is represented as:
   $y\_pred = b0 + b1*X$

2. Multiple Linear Regression:
   Where there are multiple independent variables affecting a dependent variable. For eg house price is affected by area, rooms, stories, bedrooms etc etc. It is represented as:
   $Y\_pred = b0 + b1*X1 + b2*X2 + b3*X3+\ldots\ldots\ldots\ldots bn*Xn$

Finally we can analyse the model by the following metrics:

1. R squared: Tells us how much the variance of the data is explained by the model. It can be represented as
   $(1 – RSS/TSS)$

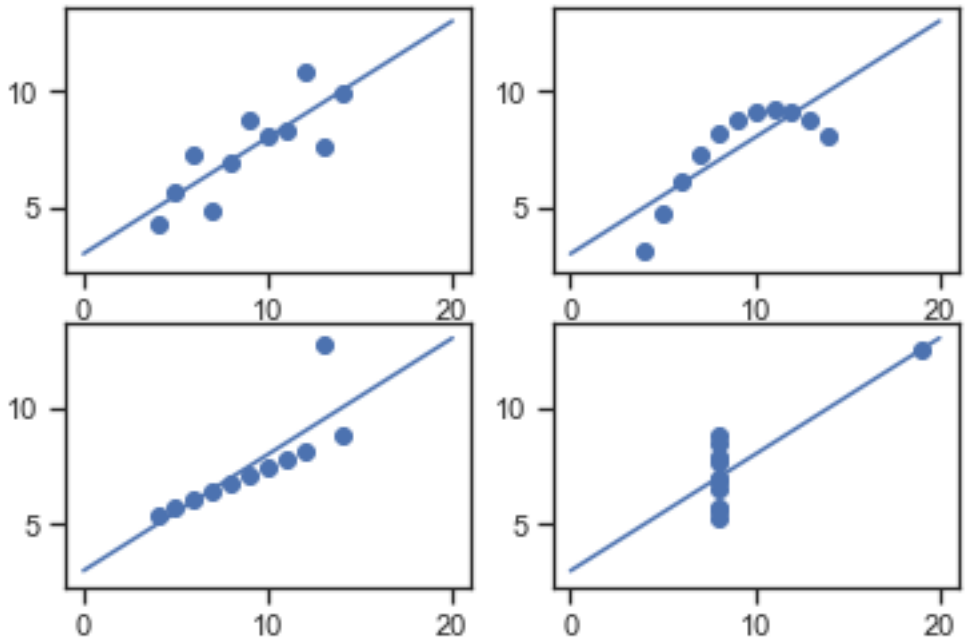
2. P-value : P-value of the coefficients tells us whether the coefficient is significant or not.

3. RMSE : It can be used to see the difference between actual data points and regression line.

2. Explain the Anscombe’s quartet in detail.

Ans: Anscombe’s Quartet signifies both the importance of plotting data before analysing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyse about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behaviour irrespective of statistical analysis.

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behavior.

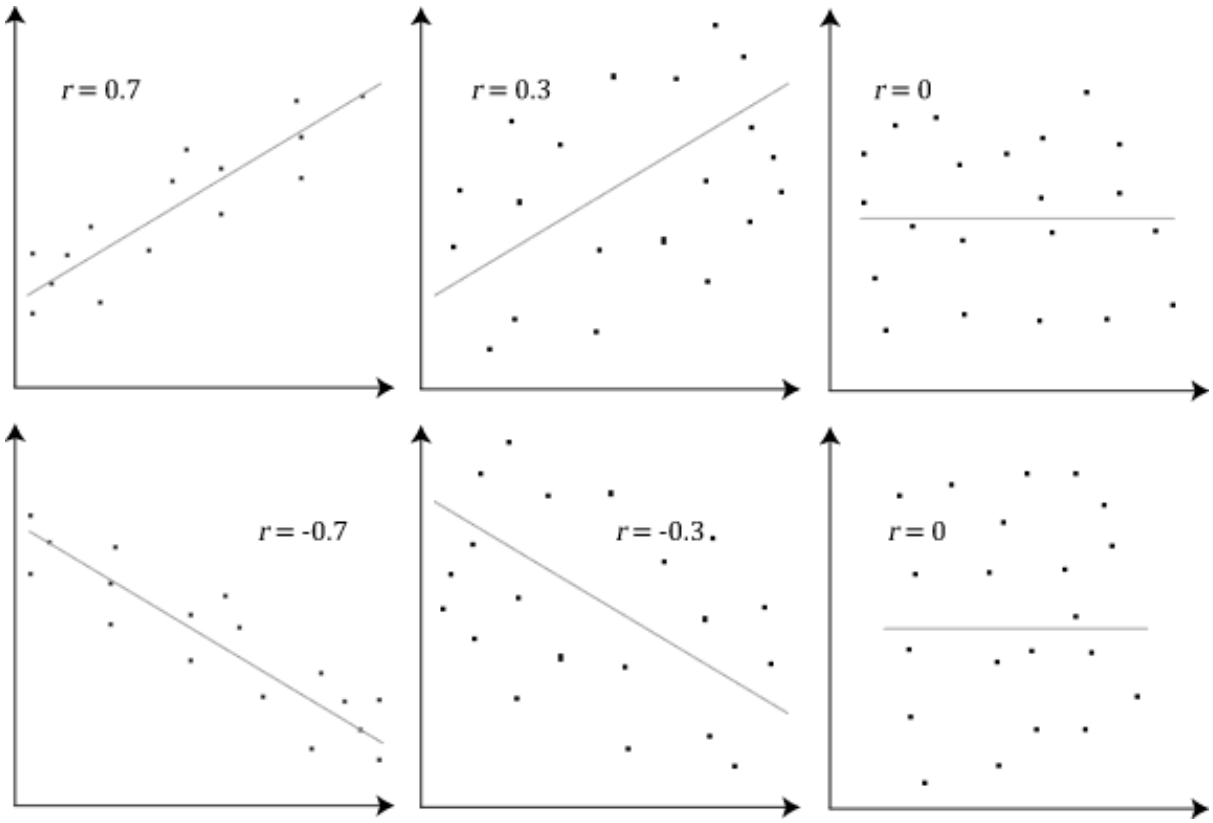


Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.
Data-set IV — looks like the value of x remains constant, except for one outlier as well.

Data-sets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data. This isn't to say that summary statistics are useless. They're just misleading on their own. It's important to use these as just one tool in a larger data analysis process. Visualizing our data allows us to revisit our summary statistics and re-contextualize them as needed.

3. What is Pearson's R?

Ans: The Pearson's R is a type of correlation coefficient that represents the relationship between two variables that are measured on the same interval or ratio scale. The Pearson coefficient is a measure of the strength of the association between two continuous variables. The Pearson coefficient is a mathematical correlation coefficient representing the relationship between two variables, denoted as X and Y. Pearson coefficients range from +1 to -1, with +1 representing a positive correlation, -1 representing a negative correlation, and 0 representing no relationship. The Pearson coefficient shows correlation, not causation.



Above figures show the values of Pearson's R respectively, we can see the positive, negative and zero correlation.

The stronger the association of the two variables, the closer the Pearson correlation coefficient, r, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for r between +1 and -1 (for example, r = 0.8 or -0.4) indicate that there is variation around the line of best fit.

It is important to realize that the Pearson correlation coefficient, r, does not represent the slope of the line of best fit. Therefore, if you get a Pearson correlation coefficient of +1 this does not mean that for every unit increase in one variable there is a unit increase in another. It simply means that there is no variation between the data points and the line of best fit.

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

Ans: Scaling is an important step during data pre-processing to standardize the independent features present in the dataset. By standardizing, we mean to scale the features to bring them in the same range. There are multiple techniques to perform feature scaling.

It is important to do feature scaling as real life datasets have many features with a wide range of values like for example let's consider the house price prediction dataset. It will have many features like no. of. bedrooms, square feet area of the house, etc. As you can guess, the no. of

bedrooms will vary between 1 and 5, but the square feet area will range from 500-2000. This is a huge difference in the range of both features. Many machine learning algorithms that are using Euclidean distance as a metric to calculate the similarities will fail to give a reasonable recognition to the smaller feature, in this case, the number of bedrooms, which in the real case can turn out to be an actually important metric.

Types of Scaling

- Normalized Scaling:
  Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

  X` = (x-min(x))/(max(x)-min(x))

  Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively. Impact of Outliers is very high in Normalization.

- Standardized Scaling:
  Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:
  X` = (x-x_bar)/sigma

  Where x_bar is the mean and sigma is the standard deviation.
  It performs well in case of outliers.

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

Ans: VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

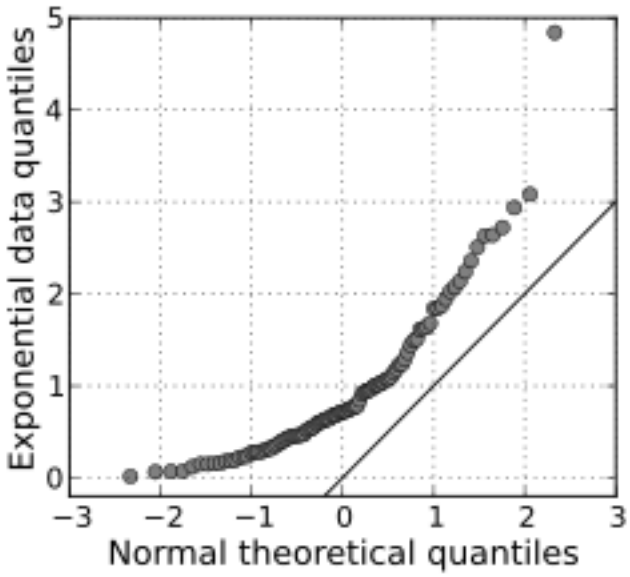
Even during our assignment I saw during first model that we are getting infinite assignment for few of the variables. A large value of VIF indicates that there is a correlation between the variables. But when there is perfect correlation then VIF is equal to infinity which means we have redundant variables.

VIF is given by

VIF = 1/(1-r2)

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

Ans: Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. In other words we can say plot quantiles against quantiles. Whenever we are interpreting a Q-Q plot, we shall concentrate on the 'y = x' line. We also call it the 45-degree line in statistics. It entails that each of our distributions has the same quantiles. In case if we witness a deviation from this line, one of the distributions could be skewed when compared to the other.



The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It's being compared to a set of data on the y-axis. This particular type of Q-Q plot is called a normal quantile-quantile (Q-Q) plot. The points are not clustered on the 45 degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.

It is useful in linear regression. Q-Q plots perform well even for small sample size. As most of our statistical methods depend on the normality assumptions, checking normality for the sample data is important. If we violate this assumption then the inference driven from the analysis may not be precise. In an advanced treatment, the q-q plot can be used to formally test the null hypothesis that the data are normal. This is done by computing the correlation coefficient of the n points in the q-q plot. Depending upon n, the null hypothesis is rejected if the correlation coefficient is less than a threshold. The threshold is already quite close to 0.95 for modest sample sizes.

Q-Q plot is used to determine if the 2 datasets come from the population with a common distribution. Also if the 2 datasets have common location and scale, similar distributional shape or similar tail behaviour