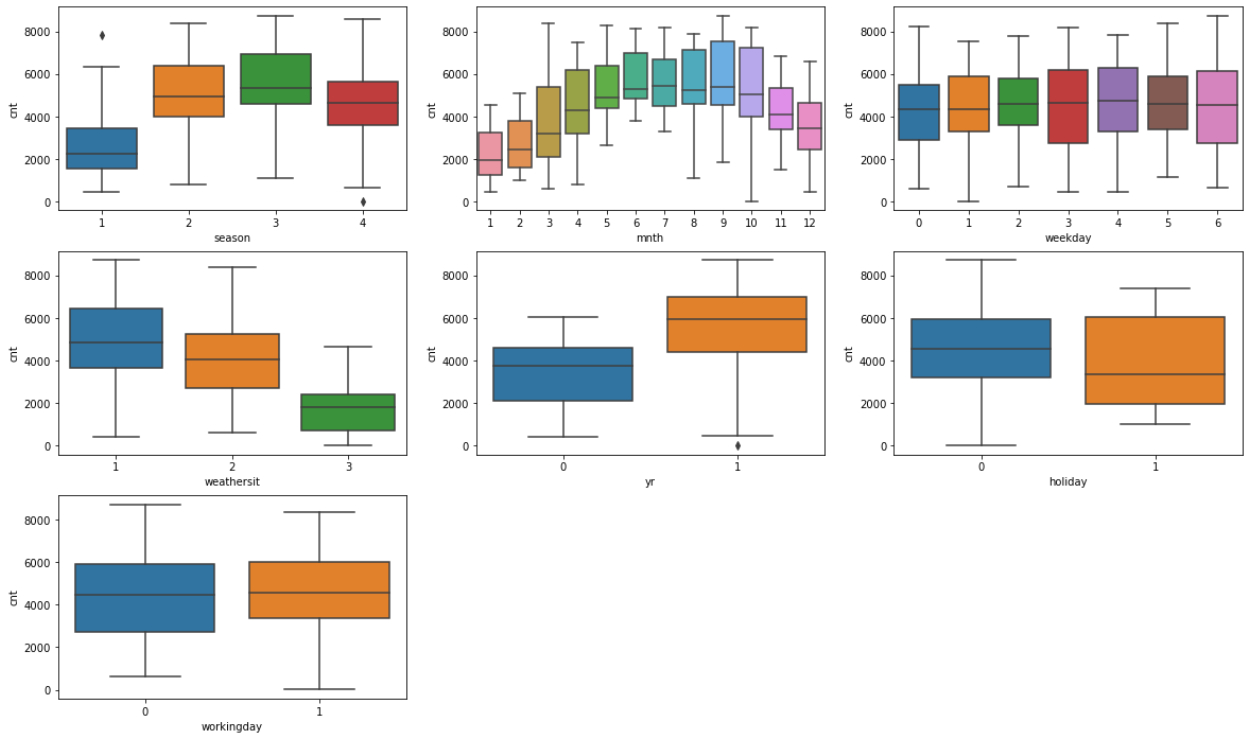
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Season,Month,Weekday,Weathersit,yr,holiday and workingday are categorical variables in the data set. Categorical variables are displayed with box plots.

- Season 2 and 3 shows good booking with mean value greater than 4000 compared to 1 and 4.
- Mnth 8,9,10 show good booking with mean value greater than 4000.
- Weekday shows consistent data no significant observation
- WeatherSit 1 is best weather having mean value greater than 4000 booking
- yr 1 (2019) is better year of booking bike compared to 0 (2018)
- Holiday shows no significant observation.Data is consistent.
- Working day shows no significant observation. Data is consistent.

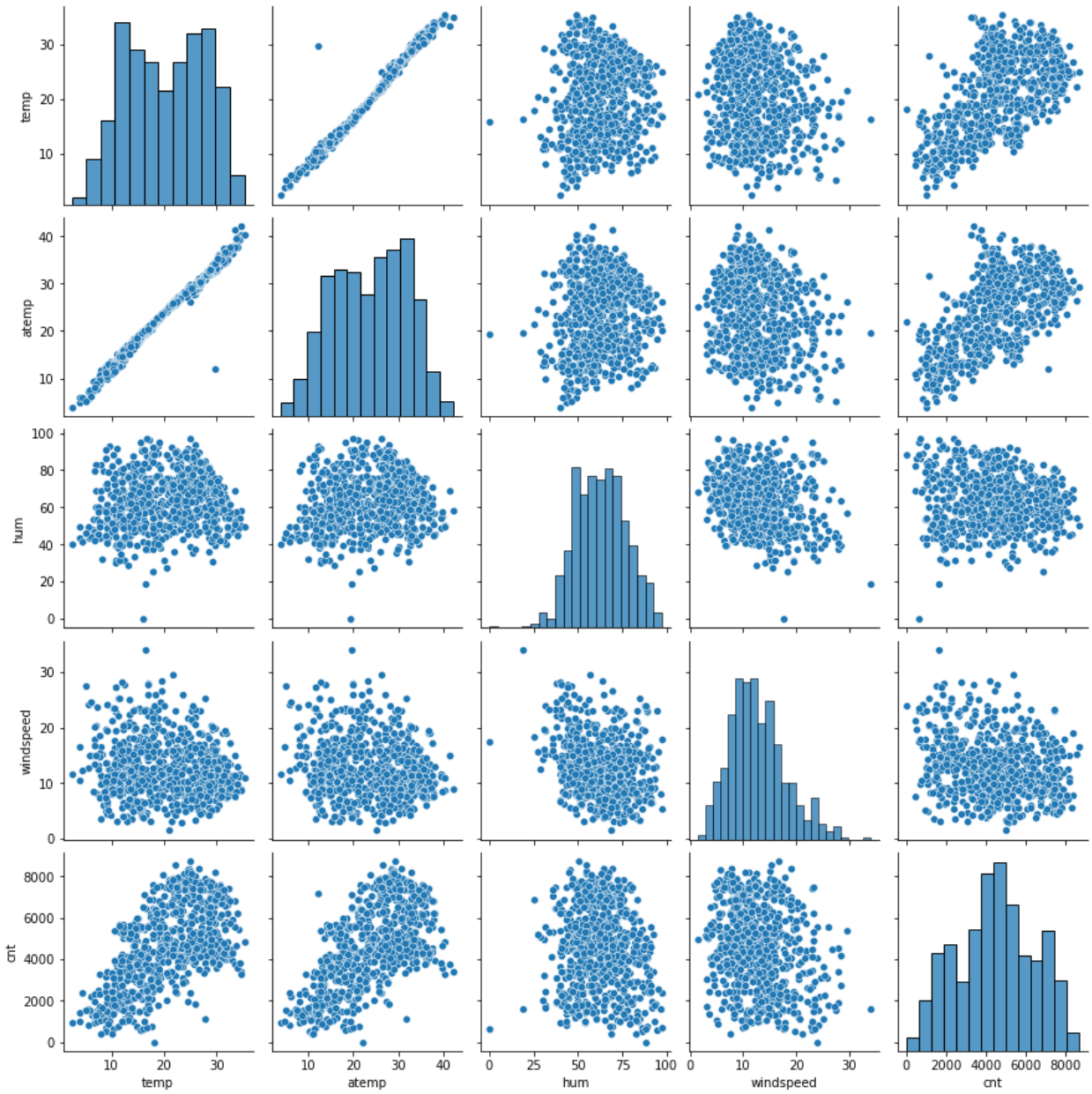


2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Drop first is used to reduce the multicollinearity in Multiple Linear regression. Since k columns for k levels of a categorical variable is a good idea, there is a redundancy of one level, which is a separate column. Since one of the combinations will uniquely represent the redundant column. So it's good to drop one of the columns and keep k-1 columns to present k levels.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

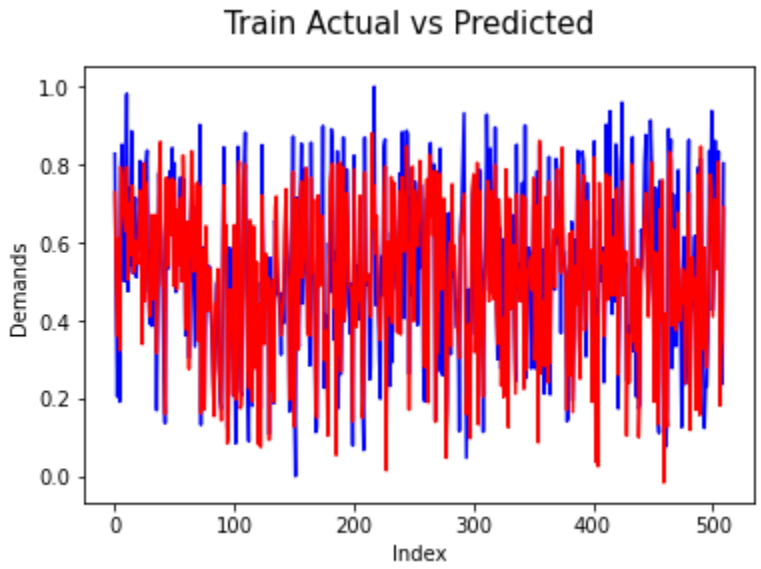
Based on a pair plot temp and atemp variable has highest correlation with cnt target variable.



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

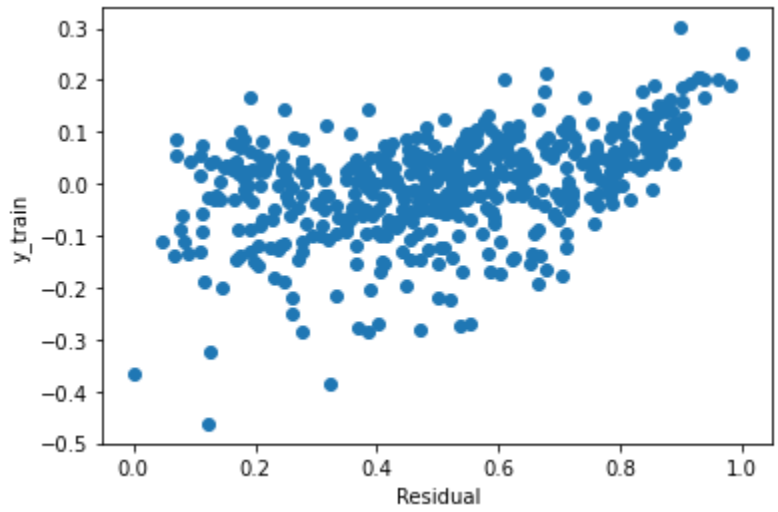
If a model with a fairly high R-squared (0.8076137158576726) is obtained, it might seem that the task is done. But one or more important explanatory variables could still be missing. Thus, we need to assess the model.

Let's first see the graph between the predicted and actual



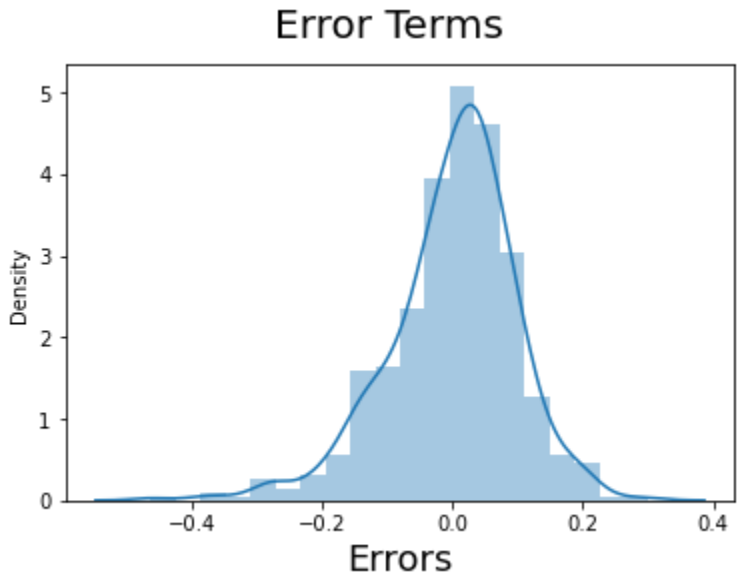
The actual and predicted views significantly overlapped, thus indicating that the model is able to explain the demand very well.

Let's see the error term (difference between predicted and actual values) plot.



Observe that the errors (the differences between the actual values and the values predicted by the model) are randomly distributed.

We can validate the Linear regression by plotting the Residuals distribution. It should follow normal distribution where mean = 0



After we determined that the coefficient is significant, using p-values, we need some other metrics to determine whether the overall model fit is significant. To do that, we need to look at a parameter called the F-statistic.
So,the parameters to assess a model are:
1. t statistic: Used to determine the p-value and hence, helps in determining whether the coefficient is significant or not
2. F statistic: Used to assess whether the overall model fit is significant or not. Generally, the higher the value of F statistic, the more significant a model turns out to be
3. R-squared: After it has been concluded that the model fit is significant, the R-squared value tells the extent of the fit, i.e. how well the straight line describes the variance in the data. Its value ranges from 0 to 1, with the value 1 being the best fit and the value 0 showcasing the worst.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.808
Model:                            OLS   Adj. R-squared:                  0.804
Method:                 Least Squares   F-statistic:                     209.5
Date:                Tue, 01 Feb 2022   Prob (F-statistic):          1.81e-171
Time:                        21:53:34   Log-Likelihood:                 459.20
No. Observations:                 510   AIC:                            -896.4
Df Residuals:                     499   BIC:                            -849.8
Df Model:                          10
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                       0.1552      0.031      4.929      0.000       0.093       0.217
yr                          0.2342      0.009     26.332      0.000       0.217       0.252
holiday                    -0.0851      0.028     -3.027      0.003      -0.140      -0.030
temp                        0.5128      0.036     14.325      0.000       0.442       0.583
windspeed                  -0.1403      0.027     -5.175      0.000      -0.194      -0.087
season_spring              -0.0620      0.023     -2.713      0.007      -0.107      -0.017
season_summer               0.0438      0.016      2.663      0.008       0.011       0.076
season_winter               0.0827      0.019      4.433      0.000       0.046       0.119
mnth_7                     -0.0421      0.020     -2.094      0.037      -0.082      -0.003
mnth_9                      0.0682      0.018      3.718      0.000       0.032       0.104
weathersit_lightsnow       -0.2543      0.026     -9.605      0.000      -0.306      -0.202
==============================================================================
Omnibus:                       66.613   Durbin-Watson:                   1.978
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              121.747
Skew:                          -0.779   Prob(JB):                     3.66e-27
Kurtosis:                       4.817   Cond. No.                         16.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
                Features   VIF
2                   temp  4.97
3              windspeed  4.60
5          season_summer  2.18
0                     yr  2.07
4          season_spring  2.00
6          season_winter  1.71
7                 mnth_7  1.58
8                 mnth_9  1.33
9  weathersit_lightsnow  1.06
1               holiday  1.04
```

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

yr(1) 2019,temp and weathersit_lightsnow are 3 features contributing significantly towards explaining the demand of the shared bikes.

Variable **Co-efficient**

yr	0.234194
temp	0.512806
weathersit_lightsnow	-0.254302

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Regression is the most commonly used predictive analysis model.
Linear regression is an important class of supervised learning algorithms. It is a form of technique which tells us the relationship between dependent (target) and independent (predictor) variables.
There are 2 types of Linear regression :

1. Simple Linear Regression
   It explains the relationship between dependent variables and 1 independent variable.
   The standard equation of the regression line is given by the following expression:
   $Y = \beta_0 + \beta_1.X$
   Where β_0 is Intercept and β_1 is Slope
2. Multiple Linear Regression
   It explains the relationship between 1 dependent variable and several independent variables.The equation of multiple linear regression would be as follows:

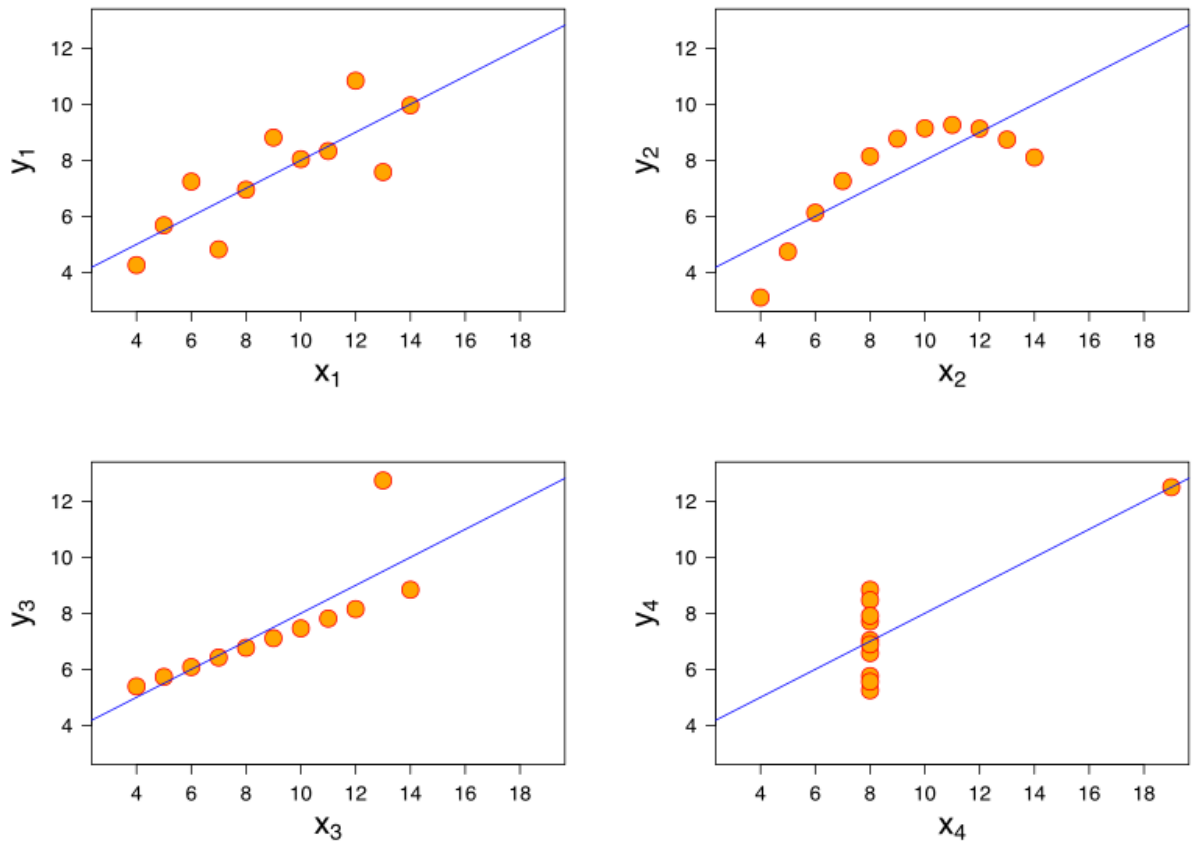
   $$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p ,$$

   where Y is the predicted or expected value of the dependent variable,
   X_1 through X_p are p distinct independent or predictor variables,
   b_0 is the value of Y when all of the independent variables (X_1 through X_p) are equal to zero, and b_1 through b_p are the estimated regression coefficients.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet was 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.

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.



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear..
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for).
- Finally, the fourth graph (bottom right) shows an example 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.

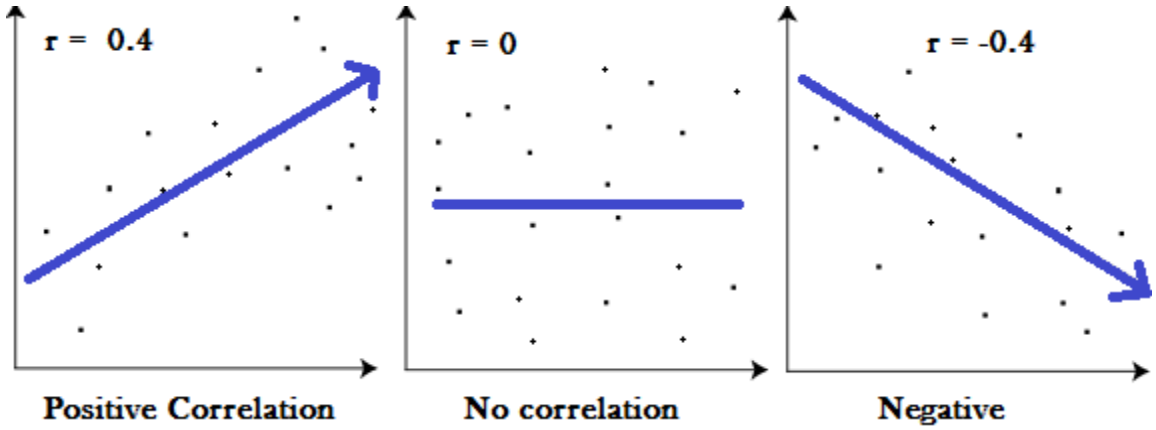
The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets

3. What is Pearson's R? (3 marks)

Pearson's R is a measure of linear correlation between two sets of data. It's the ratio between the covariance of 2 variables and the product of their standard deviations;thus its normalized measurement of the covariance range between -1 to 1 .

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.

- A result of zero indicates no relationship at all.



Basic Pearson R formula

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

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

When we have a lot of independent variables in a model, a lot of them might be on very different scales which will lead to a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:
1. Ease of interpretation
2. Faster convergence for gradient descent methods

You can scale the features using two very popular method:
1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

VIF = infinity when there is perfect correlation between two independent variables. Since R2 =1, which leads to 1/(1-R2) = infinity.

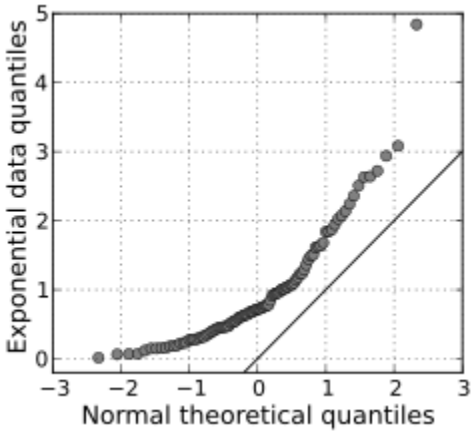
To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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).

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.