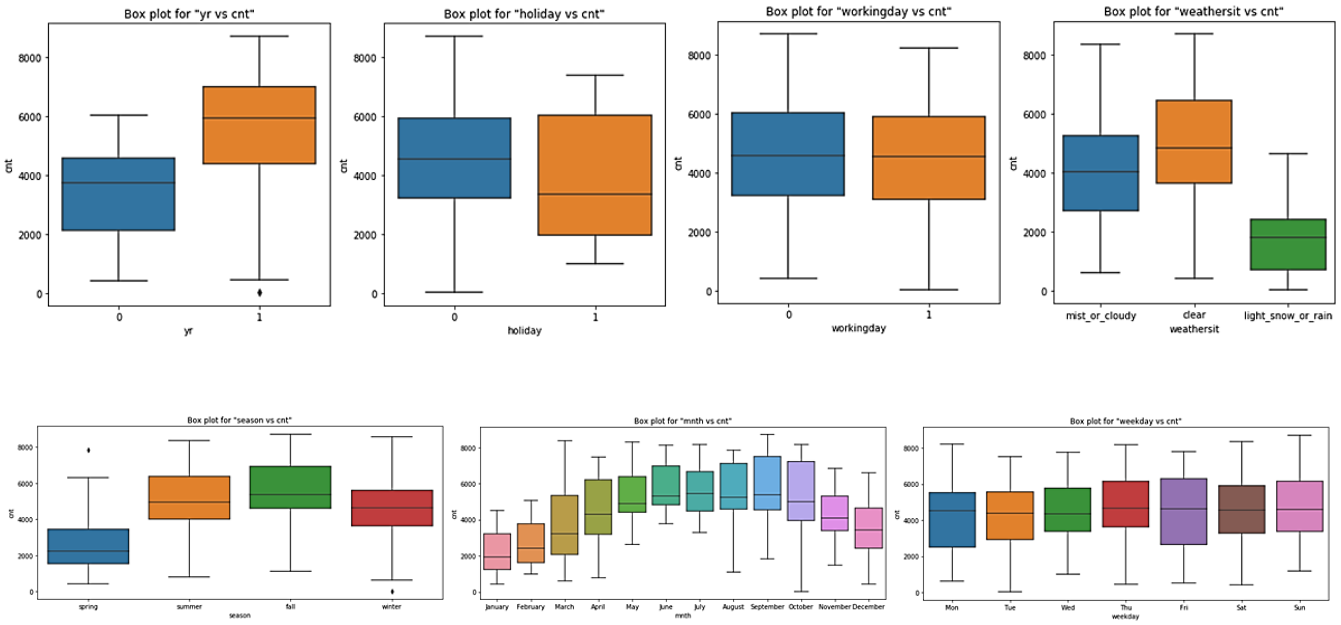
LINEAR REGRESSION REPORT

Jithin Prakash K
jithinprakashk@gmail.com

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?



Year: From the plot, it is clear that for the year=1, i.e. when the year is 2019, the bike rental is more, the 25th percentile, maximum, and median data is much more than the 2018 data

Holiday: The spread/distribution is wide for the number of rentals during holiday, However, the median and 25th percentile is higher for non-holidays. The max value of rental is more for non-holidays and minimum value is far less for non-holidays.

Working day: Distribution and median of number of bike rental are showing similar pattern irrespective of if it is a working day or not, however, the values are slightly lower for all percentiles, minimum and maximum for working day.

Weather Situation: Bike rentals are higher in a clear day. Maximum value and all quantile values are higher for clear day. Rainy or Snowy days has the lowest bike rental counts, with all quantile and median value being lowest.

Season: Bike rentals are higher during fall season. Lowest bike sales are during spring season. During summer and winter, median is almost similar for the rental counts though the count distribution is comparatively less during winter.

Month: The bike rental shows a trend of increase from Jan to September and gradually decreases the median towards December. Most rentals (median) are during July and September, and least during January

Weekday: The Bike rental count medians for all the weekdays are approximately same. There is no specific pattern. Distribution of rental count is more during Friday however the median lies around same count for all days.

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

Nominal categorical features are encoded to different columns using one hot encoding, in which a usual method is by using the pandas.get_dummies function. This function helps to create a column for all the categorical variables and encode it with a 1 or 0 (binary data) if the column has the particular value in it.

Let's take an example of seasons, there are 4 seasons available in the dataset, namely Spring, Summer, Fall and Winter. When we carry out **one hot encoding** using pd.get_dummies, it creates 4 columns with 1s and 0s in it depending on the data available in the column, i.e. if it is summer for a particular record, it creates 1 in summer column and 0s in all the other columns. That is, indirectly, if all columns have value 0, it means the last column would have the value 1. Considering this scenario, we need to create only 3 columns i.e. p-1 columns when there are p columns in total.

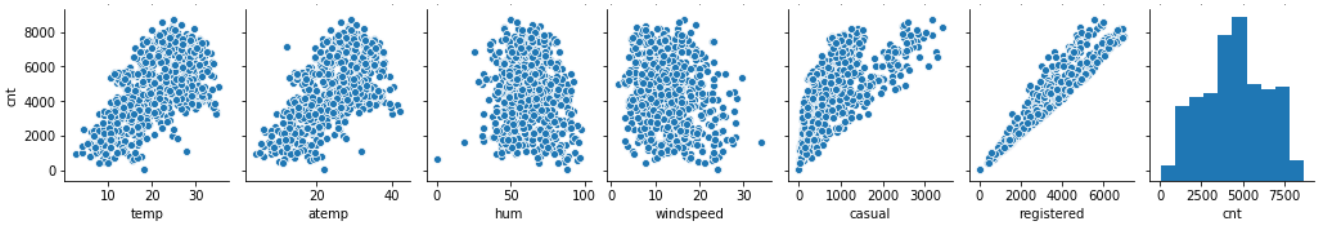
To avoid many numbers of columns, and to obtain an optimized data-frame for the model building, we can get rid of one of the columns. This can be efficiently obtained using drop_first=True parameter in pd.get_dummies.

Thus, essentially pd.get_dummies(pandas.series,drop_first=True) creates p-1 columns for p categories, i.e. when all the values in all the generated columns are 0, this mean the value for the dropped (non-existing) column is 1.

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

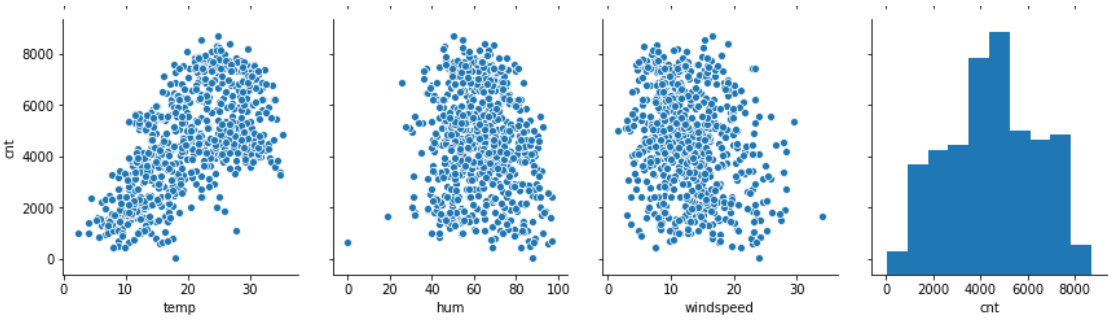
There are 2 pair plots created,

- Before removing highly correlated, or unwanted variables



Before removing variables, registered is highly correlated with bike rental count, but registered by itself is the part of target variable and hence it is removed.

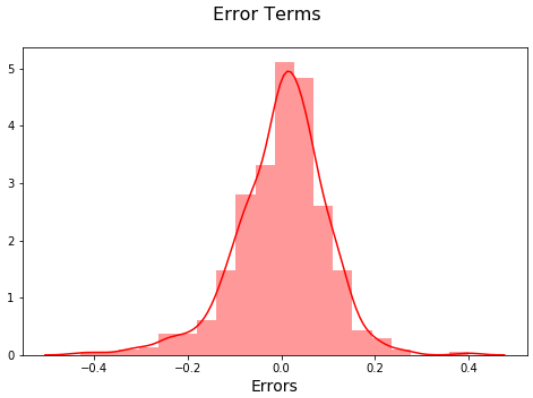
- After removing variables.

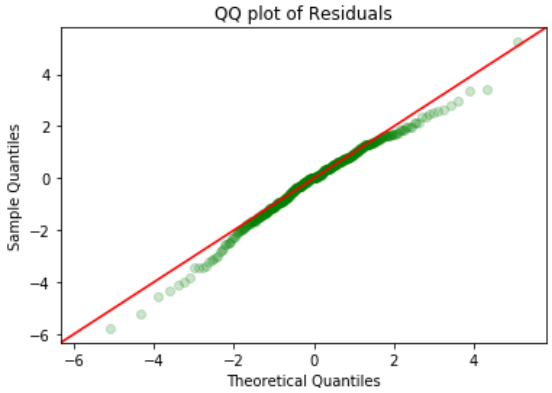


Temperature is highly correlated with the Bike rental count

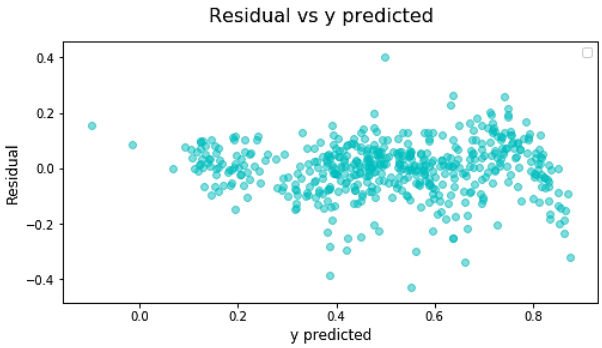
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

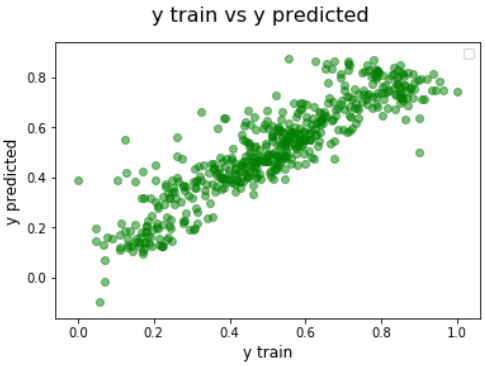
- **Linearity Assumption:** There is a linear Relationship between the independent and dependent variables. This is the primary assumption that there is a linear relationship between predictor and target variable, this can be checked by plotting a scatter plot between actual and predicted values to see if they follow same pattern.
- **Residuals**:
  - **Normality Assumption** – Assumption is that the residual / error terms (y_train - y_train_predicted) are normally distributed.
  - **Zero mean assumption** – The above-mentioned distribution is normal with center at zero (0). Both assumptions are checked by plotting the residuals as a normal plot (seaborn.distplot)





  - **Constant variance assumption (Homoscedasticity)** – i.e. the error terms/residuals should have same variance. This can be obtained by plotting the residuals against the predicted values of train data. Its also assumed that the residuals are independent of each other.





- **Independent Variables**
  - **Multicollinearity** – using VIF (variance_inflation_factor), (statsmodels.stats.outliers_influence.variance_inflation_factor) we can check the multicollinearity within the independent features. During the model building stage, VIF is calculated to check if the independent variables are highly correlated, if so, the feature is eliminated during model building. A VIF score of less than 5 is usually preferred.

	Features	VIF
0	windspeed	4.47
1	temp	3.98
2	yr	1.96
3	spring	1.58
4	mist_or_cloudy	1.47
5	light_snow_or_rain	1.07

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

   The final Linear Regression formula obtained is:

   $BikeRentalCount$=(0.347×$temperature$) + (0.240×$year$) − (0.250×$light\_snow\_or\_rain$) − (0.074×$mist\_or\_cloudy$) − (0.168×$spring$) − (0.141×$wind\_speed$) + 0.329

   i.e. Most affected variables are

   - Temperature - 1 unit increase in temperature increases the Bike Rental by 0.347 unit (keeping other features constant)
   - Year - 1 unit increase in year increases the Bike Rental by 0.240 unit (keeping other features constant)
   - Weather condition (Light Snow or Rain) - 1 unit decrease in light_snow_or_rain increases the Bike Rental by 0.250 unit (keeping other features constant)

     Note: The variables are scaled

GENERAL SUBJECTIVE QUESTIONS

1. **Explain the linear regression algorithm in detail.**

Linear regression is a method of finding the linear relationship between the independent variables and the target variable. There are two types of linear regression.

- **Simple Linear regression**
  Simple linear regression is the linear relationship of target variable with a single predictor variable. The linear regression is in the mathematical form of

$$y = \beta_0 + \beta_1 x_1$$

  - Where y is the target variable
  - β_0 is the intercept
  - β_1 is the slope or the coefficient of x_1
  - x_1 is the independent variable

- **Multiple Linear regression**
  Multiple linear regression is when the target variable is in linear relationship with multiple independent predictor variables, the Multiple Linear regression is in the mathematical form of

$$y = \beta_0 + \beta_1 x_1 + \beta_1 x_1 + \ldots\ldots\ldots + \beta_p x_p$$

  - Where y is the target variable
  - β_0 is the intercept
  - $\beta_1, \beta_2 \ldots, \beta_p$ are the slope or the coefficients of $x_1 \ldots x_p$
  - $x_1, x_2 \ldots x_p$ are the independent variables

Aim of the linear regression is to obtain a best fit straight line fitting the relationship between the predictors and the target variable of any given data.

Algorithms:

Two main types of algorithms are used to build linear regression models.

- Using statsmodels.api (This gives detailed information on coefficients and accuracy)
- Using sklearn.linear_model.LinearRegression

Steps Involved:

- Read and understand the data – use head, info, shape, describe etc.
- Visualize the data – using matplotlib and seaborn, plot and visualize
- Preprocess the data – clean, drop, impute, encode etc.
- Split the data to train and test – Scale the data if needed.
- Build model
  - use RFE – Recursive Feature elimination for feature selection,
  - VIF – for avoiding multicollinearity by dropping correlated variable when VIF>5,
  - p-Value, F-statistics, R^2, Accuracy etc. to iterate and to arrive at the best model.
- Predict the target variable values using the predict function
- Residual analysis to check all linear regression assumptions are satisfied.
- Make prediction using test data and check for accuracy and R^2 etc.
- Evaluate the model – and visualize the test and train data using a scatter plot

2. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet is a set of 4 datasets containing 11 records, i.e. x and y points. It was formulated by Francis Anscombe in 1973. The data clearly depicts the need of visualizing the data to make inferences.

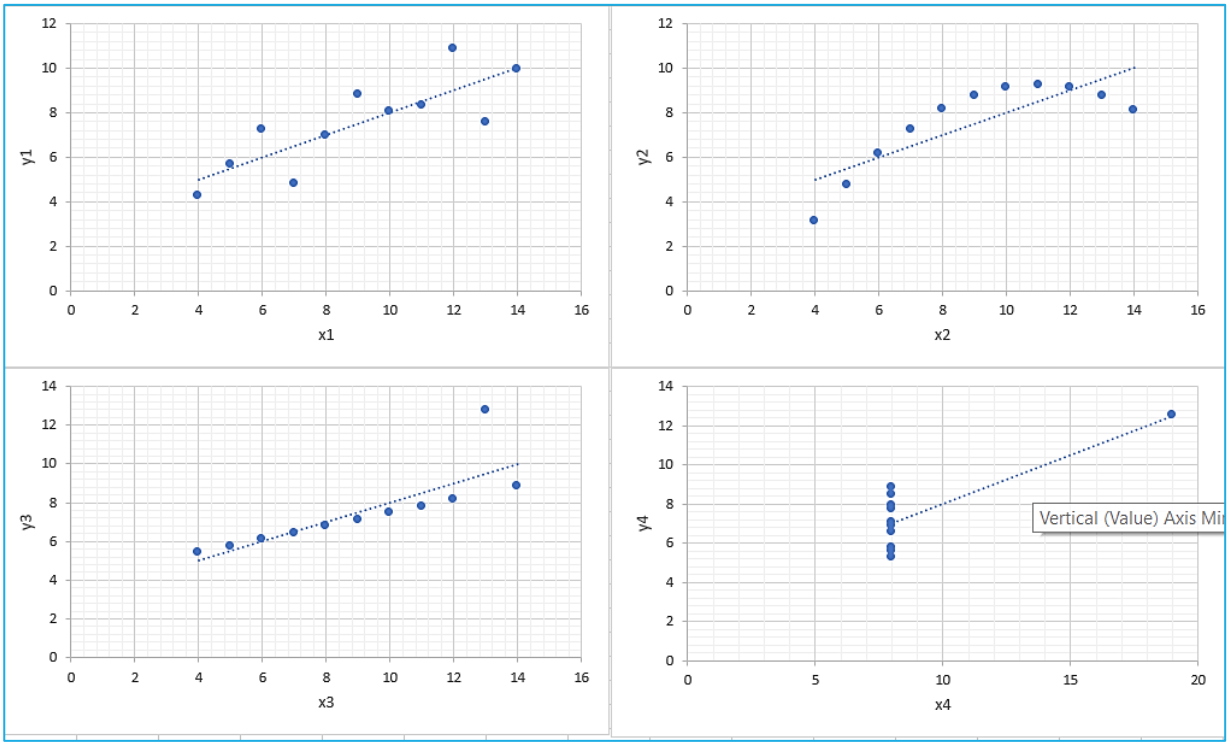
The dataset is as follows:

	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
Mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Sample Variance	11.000	4.127	11.000	4.128	11.000	4.123	11.000	4.123

Further, the Correlation between x and y is 0.816 (R)

The regression line equation is $y = 3.00 + 0.50x$ for all four datasets.

Thus, all the summary statistics shows similar data for all the four datasets. However, when plotted all four dataset gives four different visualizations as follows:



Interpretation from visualization:

Dataset 1 – Data has high linear relationship between x and y variables within some variance.

Dataset 2 – Data has clear non-linear relationship between x and y variables. Pearson's R becomes insignificant for non-linear relationship

Dataset 3 – Data has high linear relationship between x and y variables with no variance. However, one of the data points is an outlier and that affects the regression line.

Dataset 4 – Data has no relationship between x and y variables; However, the data point has an outlier which, creates a regression line which is not valid.

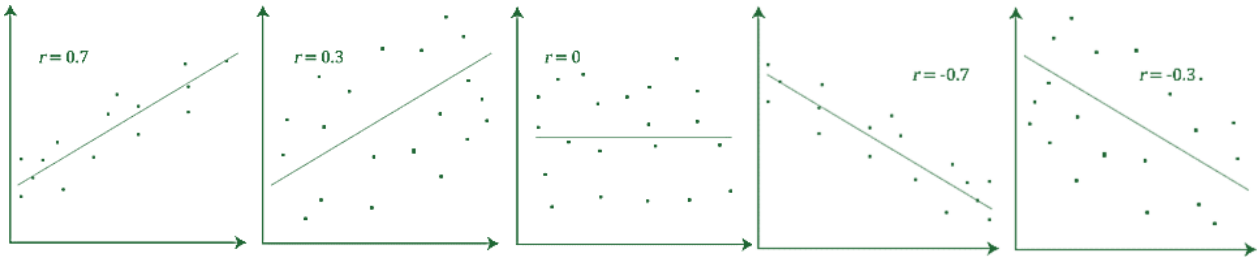
3. What is Pearson's R?

Pearson's R correlation coefficient is the linear correlation coefficient between two variables.
Formula for Pearson's R (r) is given as :

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

- To consider Pearson's r, the features x and y should have a linear relationship.
- The outliers affect the value for Pearson's r.

Consider the following 3 plots.



The above, scatter plots shows that:

- Plot 1 – Features x and y have a high positive correlation, i.e. when the value of x increases, the value of y increases and vice-versa, in this case, Pearson's r has a value of 0.70
- Plot 2 – Features x and y have a weak positive correlation, i.e. when the value of x increases, the value of y increases and vice-versa, with some variance in the data. In this case, Pearson's r has a value of 0.30
- Plot 3 - Features x and y show no correlation, i.e. there is no linear pattern between x and y. There is no linear relationship between x and y, so the Pearson's correlation coefficient is 0.0.
- Plot 4 - Features x and y have a high negative correlation, i.e. when the value of x increases, the value of y decreases and vice-versa. In this case, Pearson's r has a value of -0.70
- Plot 5 – Features x and y have a weak negative correlation, i.e. when the value of x increases, the value of y decreases and vice-versa, with some variance in the data. In this case, Pearson's r has a value of -0.30

Pearson's R ranges from -1 to 1, -1 being high negative correlation, 0 is that there is no correlation between features and 1 meaning that there is high linear positive relationship between variables.

A value outside the range -1 to 1 indicates that there is a measurement error.

Consider a simple linear regression equation

$$y = \beta_0 + \beta_1 x_1$$

For a high positive correlation, the β_1 (slope/coefficient of x) value would be positive i.e. $r \approx 1$

For a high negative correlation, the β_1 (slope/coefficient of x) value would be negative i.e. $r \approx -1$

For a non-linear relationship, there are chances that Pearson's R value is high, but that is unreliable.

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

Scaling is a technique used to reduce the scale of the data and to standardize multiple variables to a standard range. Scaling do not make any change in the distribution; it brings the data to a standard scale depending on the method/technique chosen.

Use of Scaling:

- Scaling helps to bring the data to a standard scale for all the features i.e. handling the high magnitude feature values, to a standardized value for Linear Regression.
- Scaling helps in faster convergence of gradient descend.

There are two major techniques of scaling data.

- Min-Max scaler (Normalization)
- Standard scaler (Standardization)

Min-Max scaler: Converts the feature within a range 0 to 1 i.e. the minimum value is converted to 0 and maximum value is converted to 1. The formula is given as

$$\boldsymbol{x_{scaled}} = \frac{\boldsymbol{x_i} - \boldsymbol{x_{min}}}{\boldsymbol{x_{max}} - \boldsymbol{x_{min}}}$$

Standard Scaler: Converts the features within a range with mean as 0 and standard deviation as 1. The formula is given as

$$\boldsymbol{x_{scaled}} = \frac{\boldsymbol{x_i} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}$$

Both of these scaling methods are used during the data preprocessing stages of the model building and this is done after the train test split, so that the fitting of the data is performed on train set and transformed to test set.

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

VIF (Variance Inflation Factor) is the factor that is used to check the correlation within the independent/predictor variables (Multicollinearity)

VIF formula is given as

$$\mathrm{VIF} = \frac{1}{1 - \mathrm{R}^2}$$

Mathematically, VIF is infinity when the denominator is 0, i.e. 1-R^2 is 0 which means the value of R^2 is 1.

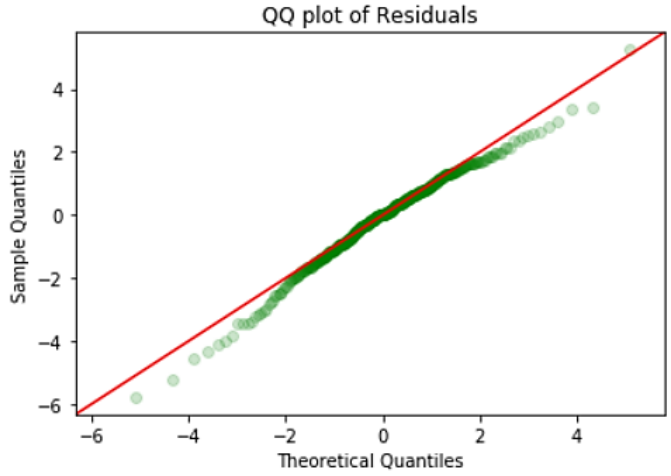
R^2 =1 means that the correlation between the features is too high, i.e. the independent features are highly correlated (positive or negative).

In short, the Infinite VIF means that the independent variables are highly correlated. We can express one variable using the other variable with a linear equation of the format $\mathrm{X1} = \mathrm{m} * \mathrm{X2} + \beta$ i.e. a linear relationship between predictor variables. We need to remove one of the variables while building the model to avoid multicollinearity.

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

Q-Q plot is Quantile-Quantile plot. This is plotted taking each point from the quantile data of a normal distribution and plots it in 2D (x-y) plane, along the x and y axis. Q-Q plot helps to determine if two data sets come from populations with a common distribution. We can plot the residual distribution on to a quantile-quantile plot to understand if they follow the normal distribution.

```
import scipy.stats as stats
res=(y_train - y_train_pred)
sm.qqplot(res, stats.t, fit=True, line='45', color='g', alpha = 0.2)
plt.title('QQ plot of Residuals')
plt.show()
```



The Q-Q plot is implemented in python using statsmodels.api.qqplot
We are plotting the residuals on the Q-Q plot using stats.t attribute and then drawn a 45° reference line to check if the data follow the distribution. Ideally all the points should fall along the straight line.

The advantages of the q-q plot are:

- Equal sample sizes are not mandatory.
- Q-Q plot can simultaneously test distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers etc. The points should lie along a straight line that is displaced either up or down from the 45-degree reference line.