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:

The following are the predictors and the corresponding columns obtained from the analysis

const	0.25
yr	0.24
holiday	-0.07
temp	0.43
windspeed	-0.15
spring	-0.13
Light Rain or Snow	-0.25
Jul	-0.07
Oct	0.05

From the above it can be inferred that

1. There is a 43 % chance that as temperature increases the number users for the bike service will increase.
2. Year on Year there is a 24% chance of increase in demand.
3. If it’s holiday there is 7% chance of decrease in demand.
4. As the windspeed increases there is 15% chance that demand will decrease.
5. If the season is spring, there is 13% chance of demand decrease.
6. If the weather situation is ‘Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds’ there is 25% chance of decrease of demand.
7. If the month is Jul there is 7% chance of demand decrease.
8. If the month is Oct there is 5% chance of increase in demand.

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

Ans:

We use **drop_first** in get_dummies function where it indicates whether to drop the first level.

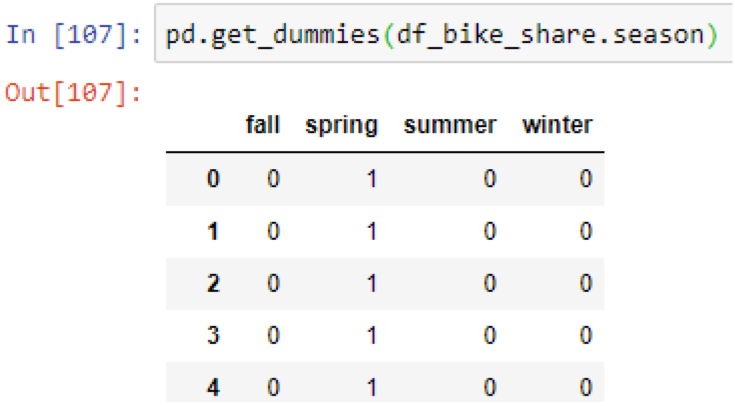
We use it remove the redundant dummy variable.

Ex:

In the assignment we have season categorical variable which has the following values

- season : season (1:spring, 2:summer, 3:fall, 4:winter)

So, if we create dummy variables without using **drop_first= True** then we get the following dummy variables.

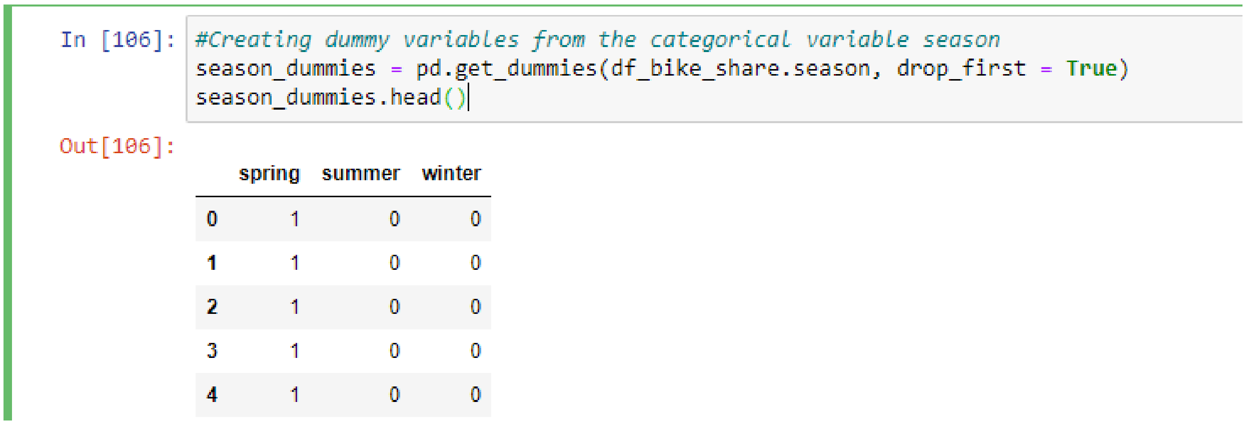


```
In [107]: pd.get_dummies(df_bike_share.season)
Out[107]:
```

	fall	spring	summer	winter
0	0	1	0	0
1	0	1	0	0
2	0	1	0	0
3	0	1	0	0
4	0	1	0	0

But we do not need one of these variables as the combination of other 3 can be used to explain it.

In the above data fall is when spring, summer, winter have values 0. So, we don't this variable, so we use drop_first = True to drop the 1st level.



```
In [106]: #Creating dummy variables from the categorical variable season
          season_dummies = pd.get_dummies(df_bike_share.season, drop_first = True)
          season_dummies.head()
```

Out[106]:

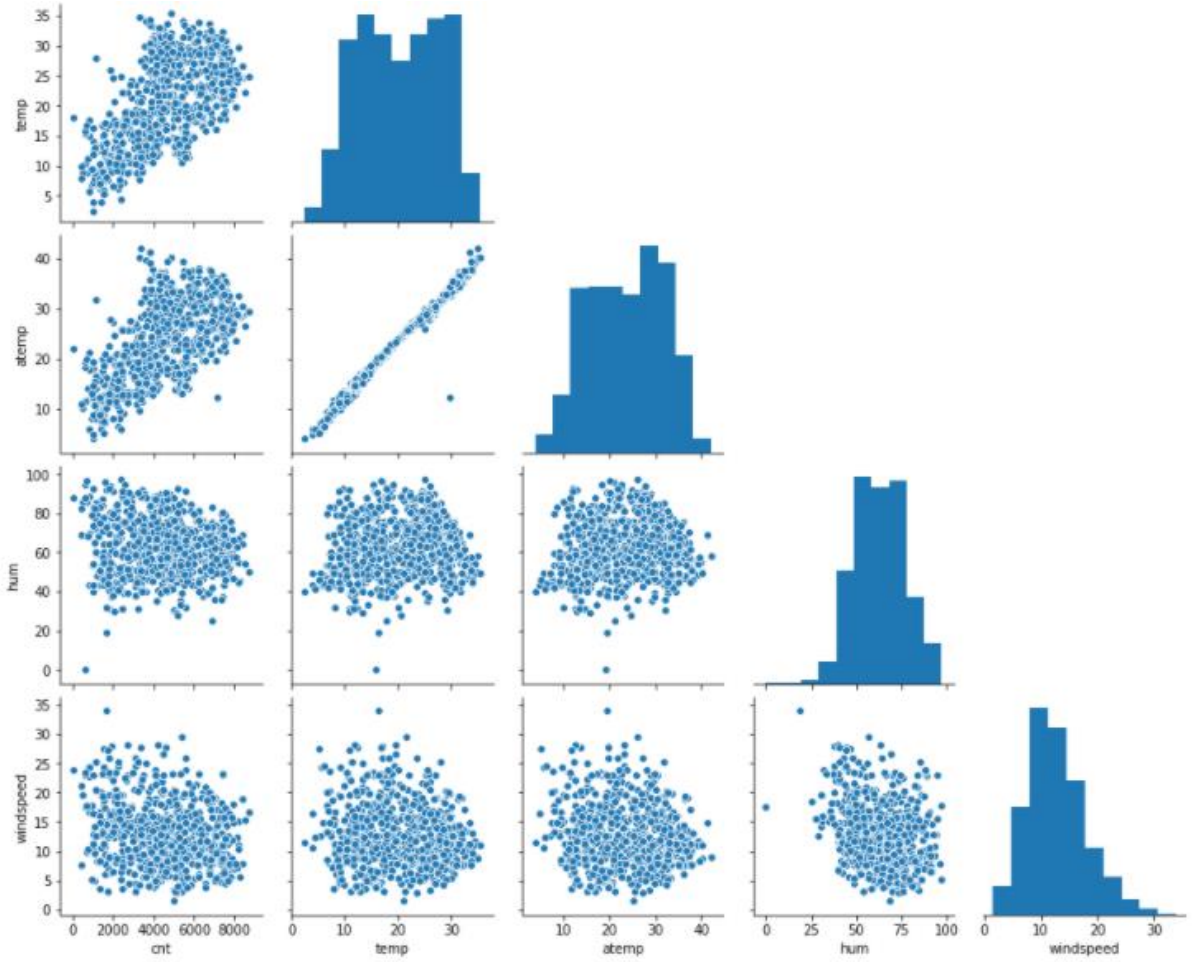
	spring	summer	winter
0	1	0	0
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0

This is applicable for any categorical variable. A generic statement would me for

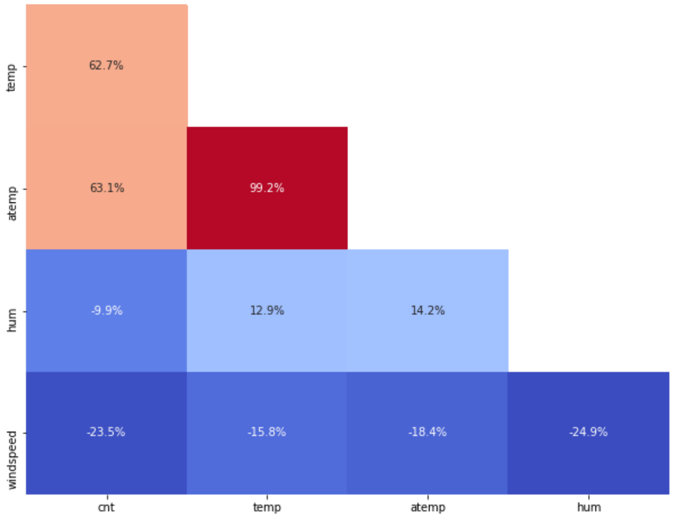
k categorical levels of a categorical variable can be explained by k-1 dummy variables.

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 we can understand that temp and atemp seem to have high correlation with our target variable cnt.
Note: casual and registered are not considered as they are used to derived target variable.



To further confirm the same point, used heat map.



From the above heat map we can conclude that the target variable cnt has

1. Positive co-relation of 62.7% with temp.
2. Positive co-relation of 63.1% with atemp.

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

Ans:

1. Once the model was finalized prediction was performed on training set. We have now y_actual and y_train_pred. The difference of these 2 are the error terms.

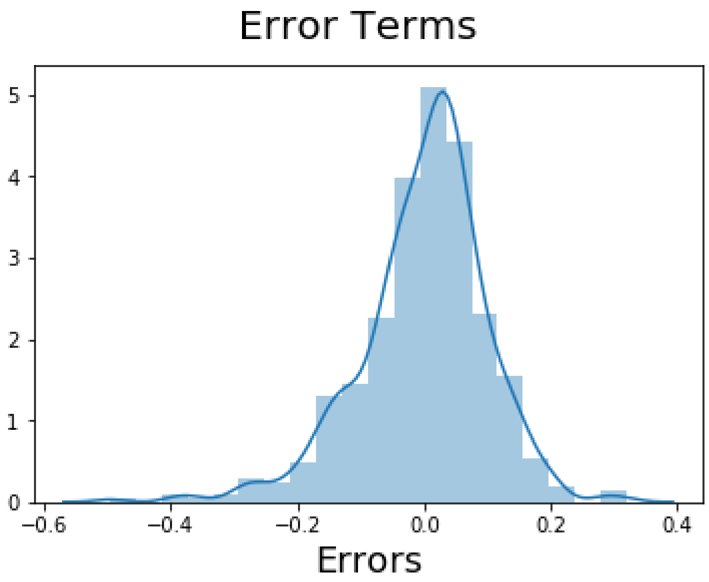
The following are our assumptions in linear regression.

Multivariate Normality:
It means the error terms are normally distributed and the mean of error terms is zero.

No Multicollinearity:
The predictor variable do not depend on each other.

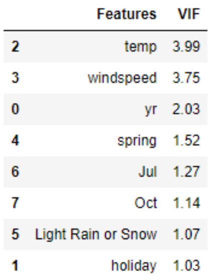
Homoscedasticity:
Variance of Error terms is constant.

2. First, we plot a dist plot of error terms.



The above graph proves that our error terms are normally distributed, and variance is constant.

3. To prove there is not multicollinearity we have used Voluntary Inflation Factor (VIF) during feature elimination. Below are the VIF values of our predictors and none of them are > 5.



	Features	VIF
2	temp	3.99
3	windspeed	3.75
0	yr	2.03
4	spring	1.52
6	Jul	1.27
7	Oct	1.14
5	Light Rain or Snow	1.07
1	holiday	1.03

Thus, all assumptions of our linear regression are validated.

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

Ans:

The top 3 features contributing significantly towards explaining the demand of the shared bikes are

1. temp : temperature in Celsius
2. weather situation - 'Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds'.
3. yr : year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

Linear regression algorithm is one of the Supervised Machine learning algorithms which is used to determine the linear relationship between a dependent variable (also called target variable) and independent variables.

There are mainly 2 types of Linear Regression algorithms

1. Simple Linear Regression – Model with 1 independent variable.
2. Multiple Linear Regression – Model with more than 1 independent variable.

Before we go into the algorithm, we have a few assumptions in Linear Regression

1. **No Multicollinearity**
   Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

2. **Relationship between variables**
   Linear regression model assumes that the relationship between response and feature variables must be linear.

3. **Homoscedasticity of Error Terms**
   Error terms have constant variance.

4. **Multivariate Normality**
   Error terms are normally distributed. The mean of the error terms is 0.

   These assumptions allow us to make inferences.

Note: There are no assumptions on the distribution of X and Y.

Algorithm Steps:

1. **Noise Removal:**

Linear regression assumes that independent and target variables are not noisy. We use exploratory data analysis operations to clean and handle outliers etc in the data.

2. **Rescaling:**

Linear regression makes more reliable predictions if you rescale input variables using standardization or normalization.

3. **Collinearity Removal:**

Overfitting is of model is classic case of existence of Collinearity.

4. **Model Building**

Hypothesis of Linear Regression:

The linear regression model can be represented by the following equation:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

where,

Y is the predicted value

θ_0 is the constant term.

θ_1,…,θn are the model parameters

x_1, x_2,…,xn are the feature values.

To find the best fit model we need to minimize the cost function.

Cost function:

The model aims to predict y value such that the error difference between predicted value and true value is minimum. So, the best model parameters must reach best value to minimize the error predicted between y value predicted and y value true. The formulation would be

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

J – Cost function

Cost function is Root Mean Squared Error between predicted y value (pred) and true y value (y).

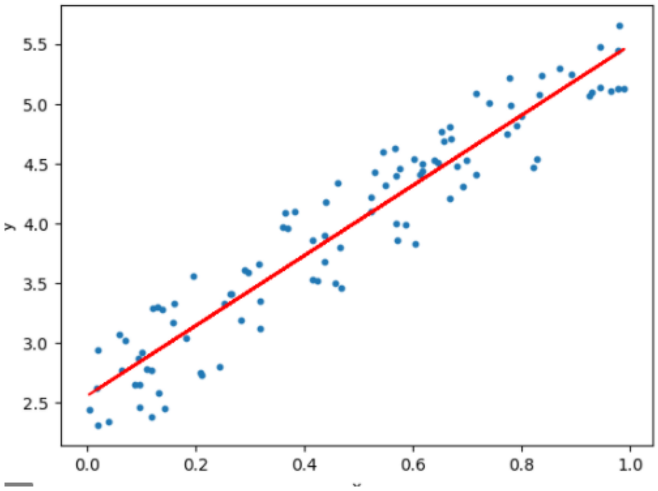
To minimize cost function, we use gradient descent

Gradient Descent:

The Gradient Descent algorithm works by starting with random values for each coefficient. The sum of the squared errors are calculated for each pair of input and output values. A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error. The process is repeated until a minimum sum squared error is achieved or no further improvement is possible.

Once the best model parameters are determined we predict using this model and verify our assumptions with **Residual analysis.**

Below is an example of a model built using linear regression



2. Explain the Anscombe's quartet in detail.

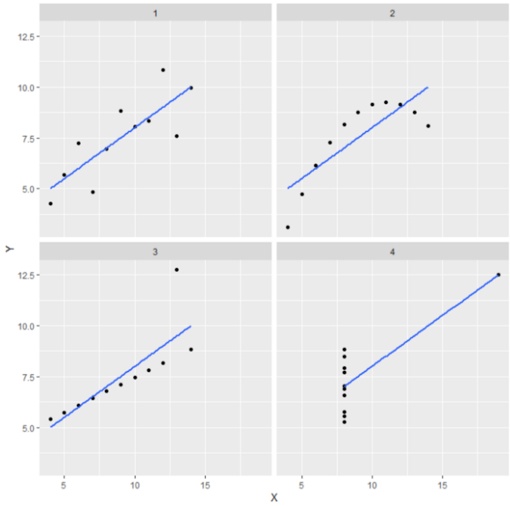
Ans:

To describe data in general terms we use variance and standard deviation which gives us the idea roughly about data consistency. However as demonstrated by Francis Anscombe in 1973, knowing variance alone does not give you the full picture of data's true form.

Anscombe's quartet:

Anscombe's Quartet showcases that four datasets that have identical statistical properties can indeed be very different.

Several data sets with several identical statistical properties were created to illustrate this as shown below



All the four data sets plotted have same variance in x, variance in y, mean of x, mean of y, and linear regression. From the above graph it can be clearly stated that they are dissimilar from each other.

This is a great demonstration of importance of graphing. Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets. For example, In graph number 3 if the one outlier point (i.e. y > 12.5) didn't exist the statistical properties would no longer be identical to the other graphs.

These below observations demonstrate the value in graphing your data before analyzing it.

1. Graph 2 should have never been analyzed with linear regression because of its curvature.
2. Graph 1 should be analyzed with a linear regression because it's a scatter plot that moves in a roughly linear manner.

Anscombe's Quartet conveys that

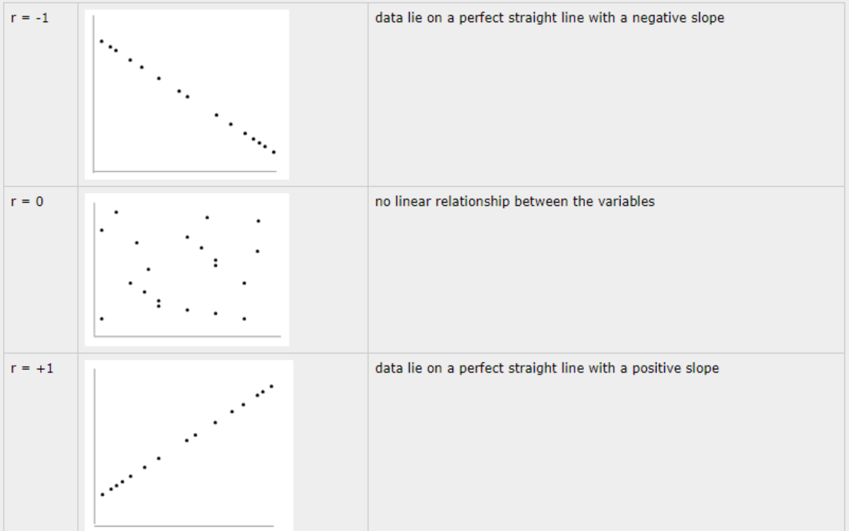
1. Graphing data prior to analysis is good practice.
2. Outliers should be removed when analyzing data.
3. Statistics about a data set do not fully depict the data set.

3. What is Pearson's R?

Ans:

Pearson's correlation coefficient (r) is a measure of the strength of the association between the two continuous variables.
Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1:

When a scatter plot is drawn between 2 continuous variables



r = -1		data lie on a perfect straight line with a negative slope
r = 0		no linear relationship between the variables
r = +1		data lie on a perfect straight line with a positive slope

1. If the variables tend to go up and down together, the correlation coefficient will be positive.
2. If the variables tend to go up and down in opposition, the correlation coefficient will be negative.

Properties:

Limit: Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists.

Pure number: It is independent of the unit of measurement. For example, if one variable's unit of measurement is in inches and the second variable is in quintals, even then, Pearson's correlation coefficient value does not change.

Symmetric: Correlation:

Coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

Degree of correlation:

Perfect: If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).

High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.

Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.

Low degree: When the value lies below + .29, then it is said to be a small correlation.

No correlation: When the value is zero.

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

Ans:

Scaling also known as data normalization is a data preprocessing step which is used to normalize the range of variables.

Need for Scaling:

Machine learning algorithms like linear regression, logistic regression, neural network, etc. that use gradient descent as an optimization technique require data to be scaled.

If we take an example of gradient descent

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

The presence of feature value X in the formula will affect the step size of the gradient descent. The difference in ranges of features will cause different step sizes for each feature. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.

There are majorly two kinds of scaling

1. Standardization
   Re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

2. Normalization (Max-Min)
   This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

The major differences are

S.No	Normalization	Standardization
1.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
2.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
3.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
4.	Scales values between [0, 1].	It is not bounded to a certain range.
5.	Outliers have huge impact on this method.	It is much less affected by outliers.
6.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Ans:

The variance inflation factor (VIF) measures the extent of correlation between one predictor and the other predictors in a model. It is used for detecting multicollinearity.

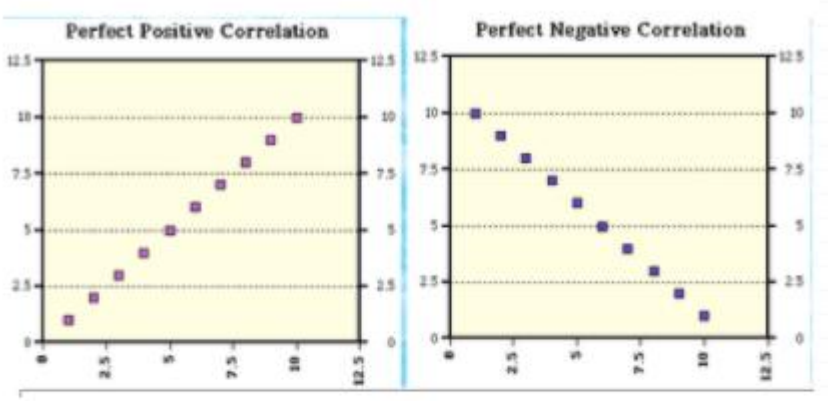
The variance inflation for a variable is computed as:

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

The value of VIF will be infinite when the predictor variables has a perfect co-correlation.

$$VIF = 1/(1-1)= 1/0 = \text{infinity}$$

This means that the corresponding predictor variables has a perfect co-correlation (± 1).



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

Ans:

Quantile-Quantile Plot is a plot When the quantiles of two variables are plotted against each other. It helps to assess if a set of data plausibly came from a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

Advantages of Q-Q plot:

1. It can be used with sample sizes also.
2. Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
   It is used to check following scenarios:
   If two data sets —

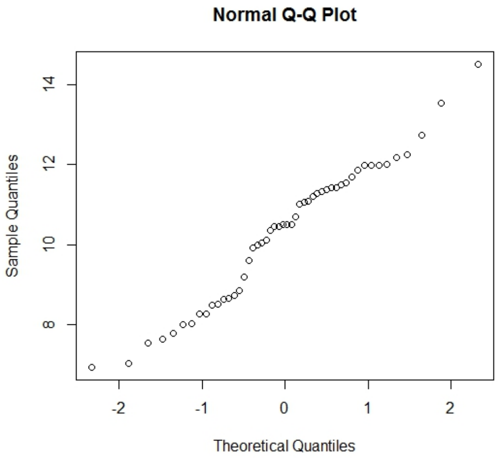
   i. come from populations with a common distribution

   ii. have common location and scale

   iii. have similar distributional shapes

   iv. have similar tail behavior

   This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles