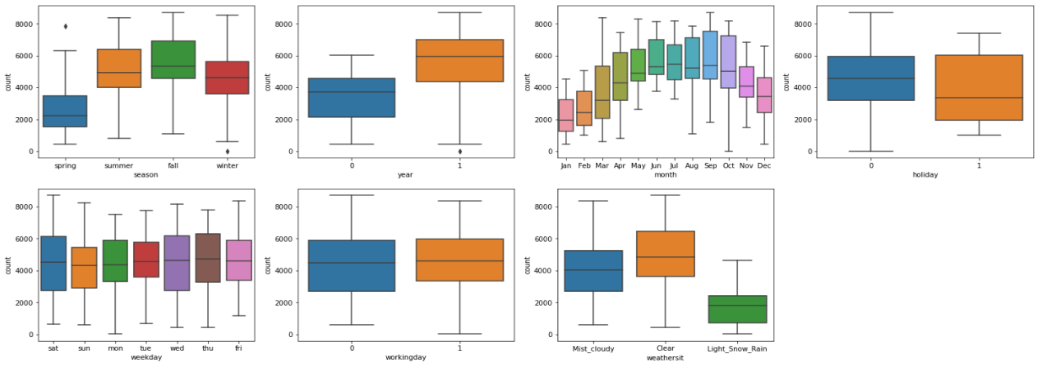
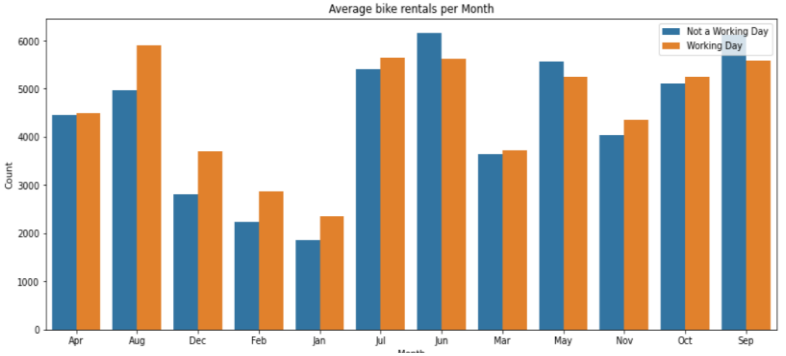
Assignment-based Subjective Questions

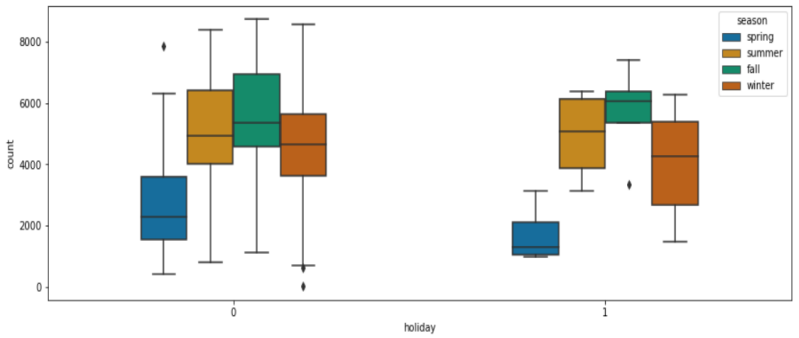
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

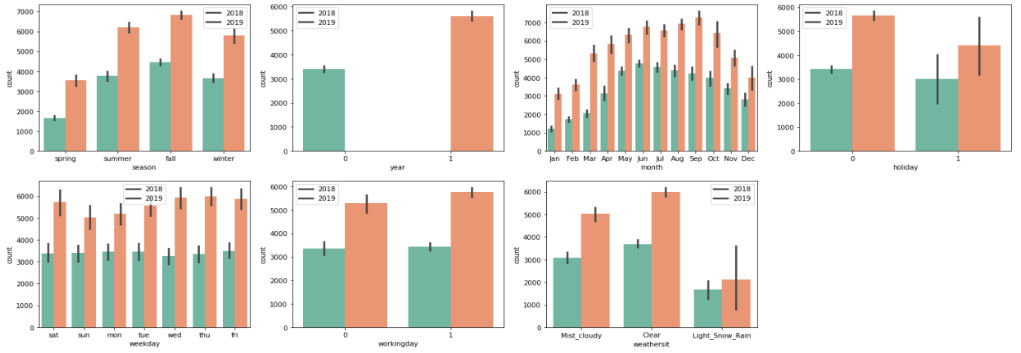
   From the analysis of the categorical variables from the dataset, we can infer that –

   - Fall season has the highest demand for rental bikes followed by summer
   - Demand in year(1) 2019 has increased as compared to year(0) 2018
   - Demand is continuously growing each month till June whereas, after September, demand is decreasing
   - September month has the highest demand and January month has the lowest demand
   - When there is a holiday, demand has decreased.
   - Number of bookings are higher in clear weathersit whereas least in light snow rain
   - More bookings are there on Saturday as compared to other days
   - Variation can be seen among seasons when day is neither weekend nor holiday (1)
   - Variation can be seen among months when day is neither weekend nor holiday (1) or holiday(0)
   - count is linearly increasing with temp and atemp so demand for bikes is positively correlated to temp and atemp
   - humidity and windspeed values are more scattered and the count is decreasing with an increase in humidity and windspeed
   - Most of the bookings has been done during the month of may, june, july, aug, sep and oct.

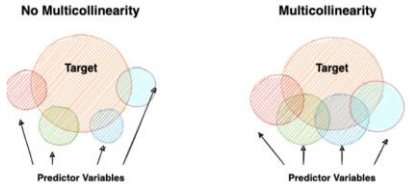




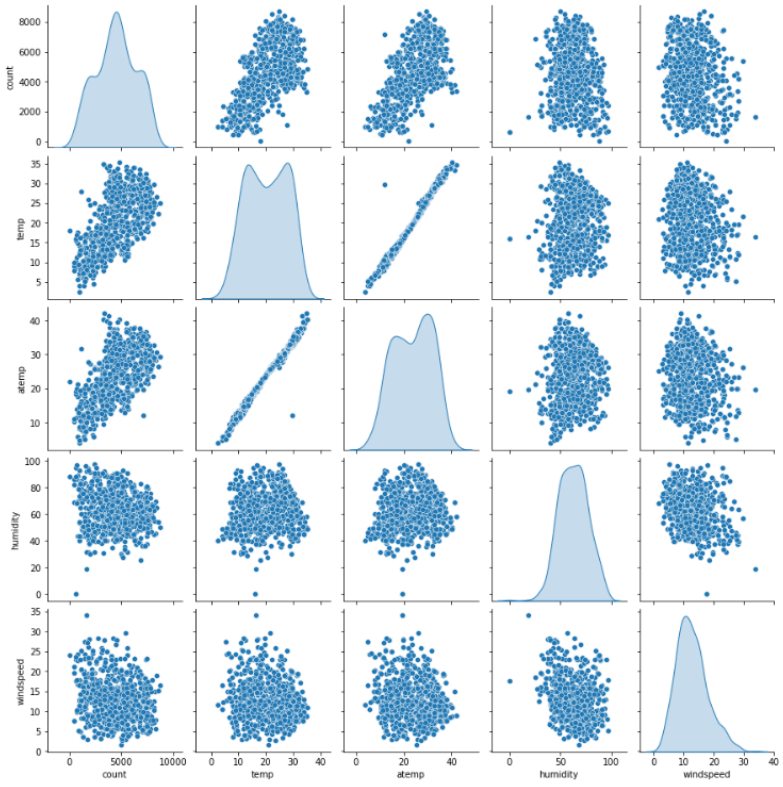




2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   - drop_first=True is important to use as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
   - drop_first: Remove the first level to get k-1 dummies out of k categorical levels.
   - Let's say we have 3 types of values in the Categorical column, and we want to create a dummy variable for that column. If one variable is not furnished and semi-furnished, then It is obvious unfurnished. So, we do not need 3rd variable to identify the unfurnished.
   - Hence if we have a categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.



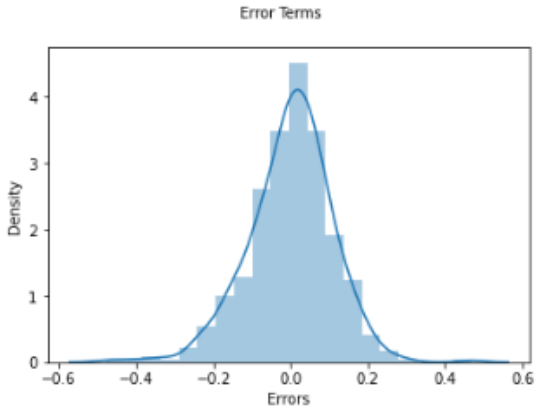
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   We can observe in the pair-plot that, temp and atemp have the highest correlation with the target variable.

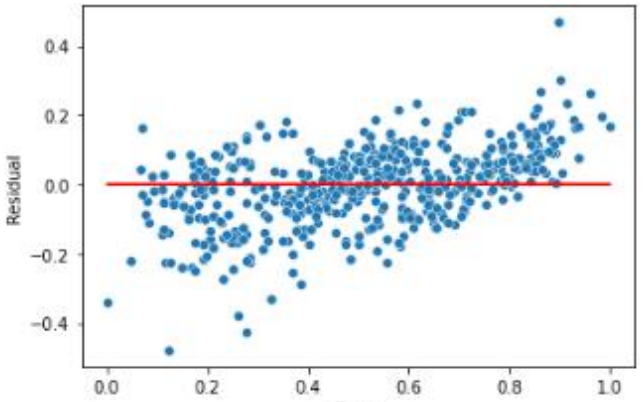


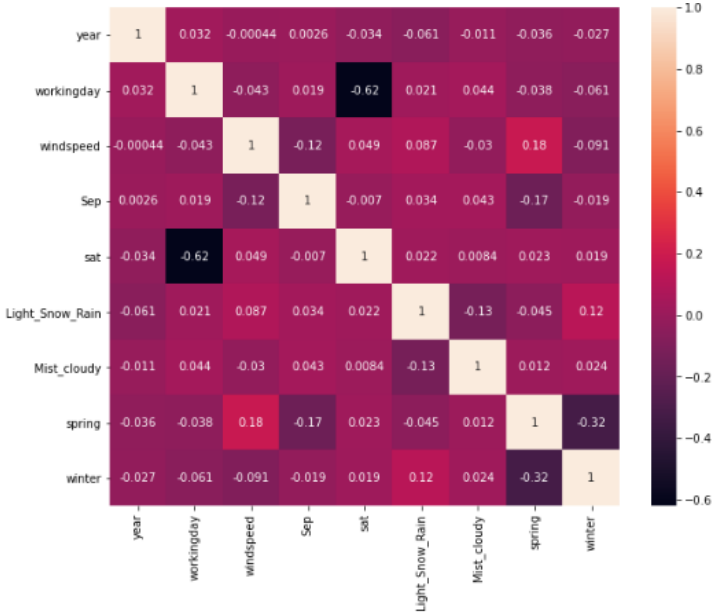
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

The important assumptions were validated in regression analysis:

- There should be **a linear relationship** between a dependent variable and the independent variable(s). A linear relationship suggests that a change in response Y due to one unit change in X^1 is constant, regardless of the value of X^1. An additive relationship suggests that the effect of X^1 on Y is independent of other variables.
- There should be no correlation between the residual (error) terms. The absence of this phenomenon is known as **Autocorrelation**.
- The independent variables should not be correlated. The absence of this phenomenon is known as **multicollinearity**.
- The error terms must have constant variance. This phenomenon is known as **homoskedasticity**. The presence of non-constant variance is referred to as heteroskedasticity.
- The error terms must be **normally distributed**.







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

As per the model, the top 3 significant features are-

- Year
- Sep
- sat

	Features	VIF
2	windspeed	3.42
1	workingday	3.07
0	year	1.87
6	Mist_cloudy	1.53
4	sat	1.52
7	spring	1.50
8	winter	1.39
3	Sep	1.11
5	Light_Snow_Rain	1.08

OLS Regression Results

Dep. Variable:	count	R-squared:	0.764
Model:	OLS	Adj. R-squared:	0.760
Method:	Least Squares	F-statistic:	179.7
Date:	Wed, 12 Jan 2022	Prob (F-statistic):	1.73e-150
Time:	22:34:38	Log-Likelihood:	406.92
No. Observations:	510	AIC:	-793.8
Df Residuals:	500	BIC:	-751.5
Df Model:	9		
Covariance Type:	nonrobust		

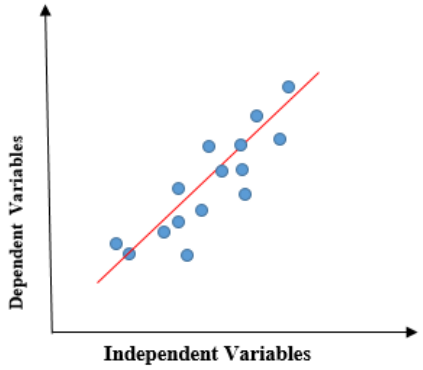
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5132	0.017	29.714	0.000	0.479	0.547
year	0.2480	0.010	25.342	0.000	0.229	0.267
workingday	0.0564	0.013	4.218	0.000	0.030	0.083
windspeed	-0.1861	0.030	-6.297	0.000	-0.244	-0.128
Sep	0.0889	0.018	4.838	0.000	0.053	0.125
sat	0.0642	0.017	3.730	0.000	0.030	0.098
Light_Snow_Rain	-0.2994	0.030	-10.123	0.000	-0.358	-0.241
Mist_cloudy	-0.0935	0.010	-8.998	0.000	-0.114	-0.073
spring	-0.2743	0.012	-22.158	0.000	-0.299	-0.250
winter	-0.0541	0.012	-4.460	0.000	-0.078	-0.030

Omnibus:	34.737	Durbin-Watson:	2.037
Prob(Omnibus):	0.000	Jarque-Bera (JB):	69.799
Skew:	-0.407	Prob(JB):	6.97e-16
Kurtosis:	4.619	Cond. No.	9.86

General Subjective Questions

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

Linear regression is a statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables. The below graph presents the linear relationship between the dependent variable and independent variables. When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.



To calculate best-fit line linear regression uses a traditional slope-intercept form.

$$y = mx+b \implies y = a_0+a_1x$$

y= Dependent Variable., x= Independent Variable., a0= intercept of the line.,a1 = Linear regression coefficient.

The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

- Positive Linear Relationship: If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship.
- Negative Linear Relationship: If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship.

Cost function

- The cost function helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points.
- Cost function optimizes the regression coefficients or weights and measures how a linear regression model is performing. The cost function is used to find the accuracy of the mapping function that maps the input variable to the output variable. This mapping function is also known as the Hypothesis function.

Linear Regression MSE

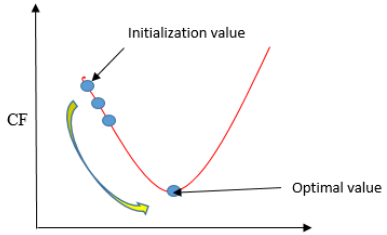
- In Linear Regression, Mean Squared Error (MSE) cost function is used, which is the average of squared error that occurred between the predicted values and actual values.By simple linear equation y=mx+b we can calculate MSE, y = actual values, yi = predicted values

- Using the MSE function, we will change the values of a0 and a1 such that the MSE value settles at the minima. Model parameters xi, b (a0,a1) can be manipulated to minimize the cost function. These parameters can be determined using the gradient descent method so that the cost function value is minimum.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

Gradient descent

- Gradient descent is a method of updating a0 and a1 to minimize the cost function (MSE). A regression model uses gradient descent to update the coefficients of the line (a0, a1 => xi, b) by reducing the cost function by a random selection of coefficient values and then iteratively updating the values to reach the minimum cost function.



Learning Rate

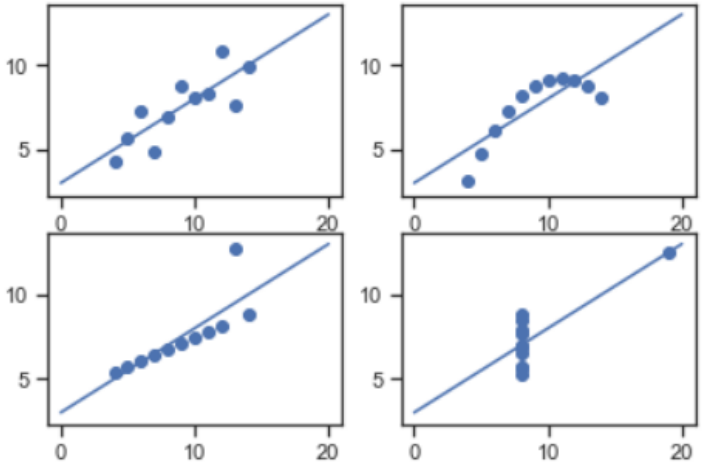
- To update a0 and a1, we take gradients from the cost function. To find these gradients, we take partial derivatives for a0 and a1.

- The blue line represents the optimal value of the learning rate, and the cost function value is minimized in a few iterations. The green line represents if the learning rate is lower than the optimal value, then the number of iterations required is high to minimize the cost function. If the learning rate selected is very high, the cost function could continue to increase with iterations and saturate at a value higher than the minimum value, represented by a red and black line.

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

- Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analyzing it with statistical properties. It comprises four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data sets is that they all share the same descriptive statistics(mean, variance, standard deviation, etc) but different graphical representations. Each graph plot shows the different behavior irrespective of statistical analysis.
- This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of dataset.
- 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.



Graphical Representation of Anscombe's Quartet

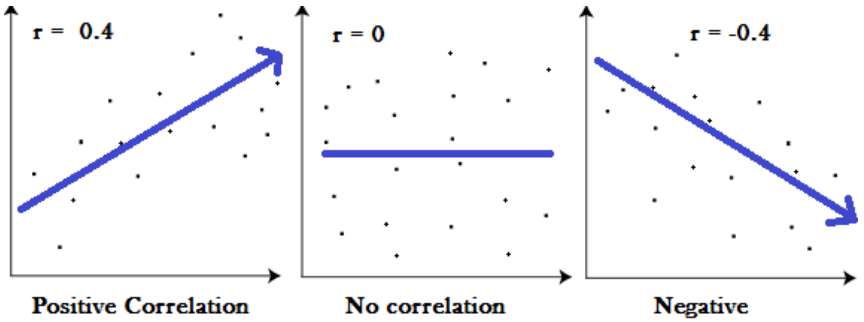
3. What is Pearson's R? (3 marks)

- Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

- The Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.
- Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:
  - 1 indicates a strong positive relationship.
  - -1 indicates a strong negative relationship.
  - A result of zero indicates no relationship at all.



- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.
- There are certain requirements for Pearson's Correlation Coefficient:
    - Scale of measurement should be interval or ratio
    - Variables should be approximately normally distributed
    - The association should be linear
    - There should be no outliers in the data

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

- Most of the time, the collected data set contains features highly varying in magnitudes, units, and range. If scaling is not done then the algorithm only takes magnitude into account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
- 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-squared, etc.
- **Normalization** is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
- It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0. On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1. If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1.
- **Standardization** is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

$$X' = \frac{X - \mu}{\sigma}$$

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
- sklearn.preprocessing.scale helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

- 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.
- The value of VIF is calculated by the below formula:

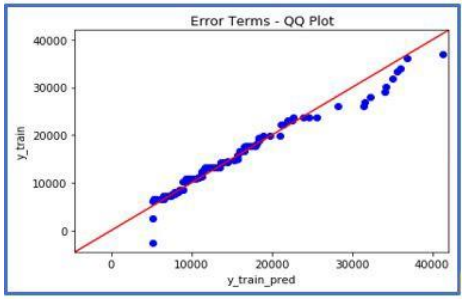
$$VIF_i = \frac{1}{1 - R_i^2}$$

Where 'i' refers to the ith variable.

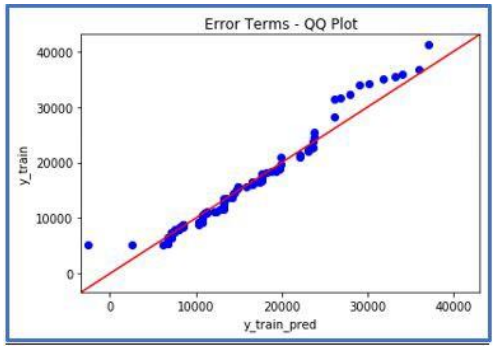
- If there is a perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing 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).
- If VIF is large and multicollinearity affects your analysis results, then you need to take some corrective actions before you can use multiple regression. Here are the various options:
    - One approach is to review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model. For example, if your inputs are measuring the weight in kgs and lbs then just keep one of these variables in the model and drop the other one. Dropping the term with a large value of VIF will hopefully, fix the VIF for the remaining terms and now all the VIF factors are within the threshold limits. If dropping one term is not enough, then you may need to drop more terms as required.
    - A second approach is to use principal component analysis and determine the optimal set of principal components that best describe your independent variables. Using this approach will get rid of your multicollinearity problem but it may be hard for you to interpret the meaning of these "new" independent variables.
    - The third approach is to increase the sample size. By adding more data points to our model, hopefully, the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.
    - The fourth approach is to transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.
    - Finally, you can use a different type of model call ridge regression that better handles multicollinearity.
- In conclusion, when you are building a multiple regression model, always check your VIF values for your independent variables and determine if you need to take any corrective action before building the model.

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

- Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential, or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
- 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.
- 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.
- This helps in a scenario of linear regression when we have the training and test data set received separately and then we can confirm using a Q-Q plot that both the data sets are from populations with the same distributions.
- Few advantages:
    - It can be used with sample sizes also
    - 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 the following scenarios:
- If two data sets —
    - come from populations with a common distribution
    - have common location and scale
    - have similar distributional shapes
    - have similar tail behavior
- Interpretation:
    - A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
- Below are the possible interpretations for two data sets.
    - Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
    - Y-values < X-values: If y-quantiles are lower than the x-quantiles.



- X-values < Y-values: If x-quantiles are lower than the y-quantiles.



- Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis