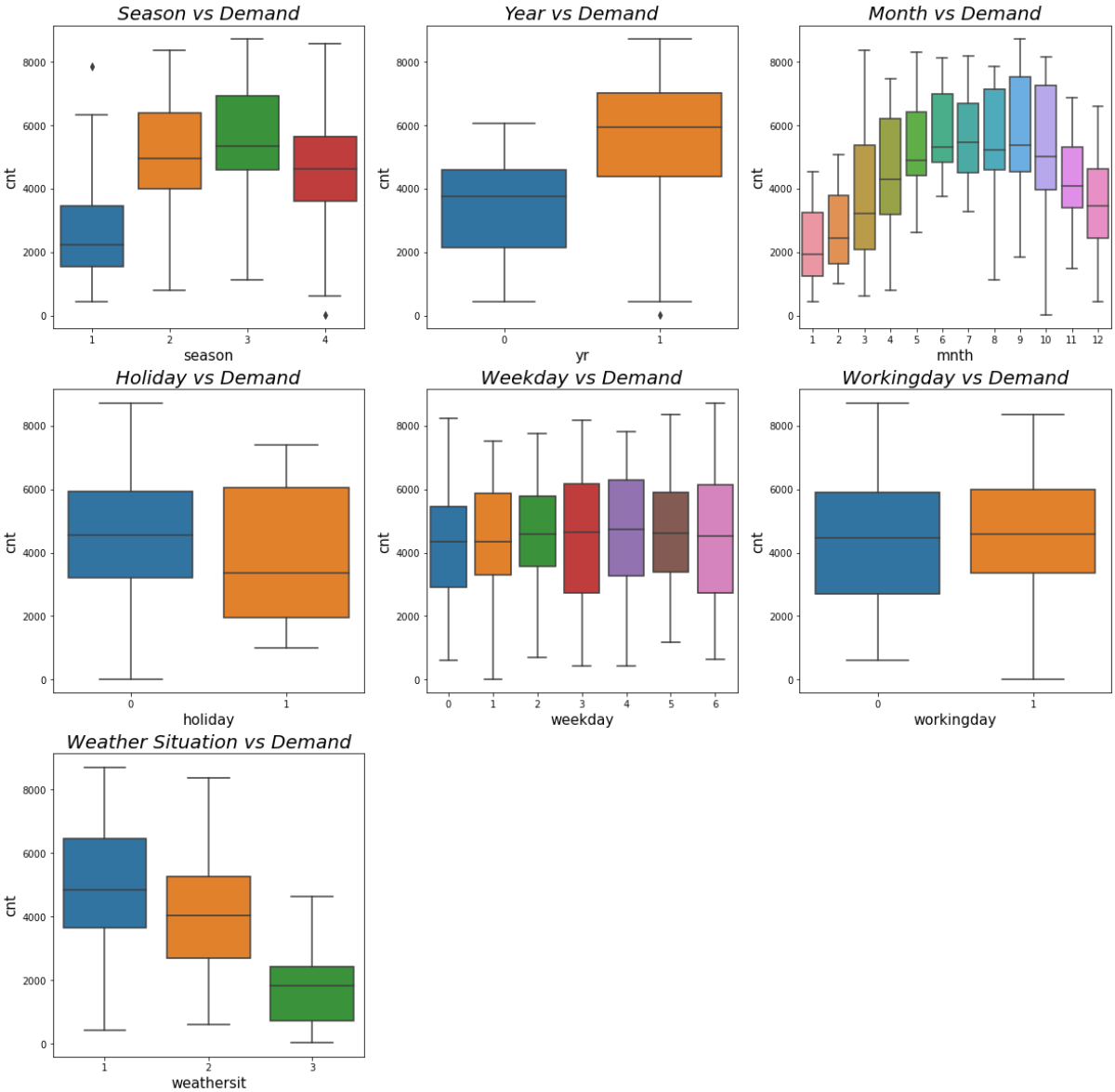
Assignment-based Subjective Questions

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

Answer: Following were marked as categorical variables in our analysis from dataset:

a. 'season'
b. 'yr'
c. 'mnth'
d. 'holiday'
e. 'weekday'
f. 'workingday'
g. 'weathersit'

Here except year rest of the variables are not ordinal.



So keeping that in perspective and referring above visualizations, following points were drawn:

- Demand increases a lot in **Fall** season.
- Overall count increased over year, means, demand/usage is growing year over year.
- Same season pattern is highlighted in months visualization, **July** onwards demand increases and goes till **October.**
- Demand on **holiday** is less indicating that working folks might use it for commute.
- And definitely a **Clear** weather do boost demand.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Answer: Dummy variables creation functions when used without **drop_first=True** they will perform one-hot encoding for each category so if a column has 10 categories then it will result in 10 columns which we can understand if applied to multiple columns will increase the feature set significantly.

But if we ignore one column and we can still get the values by absence(indicated by 0)of rest of columns, reducing feature increase by 1 column, hence those 10 category example above will result in 9 columns. This is what **drop_first=True** does.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Answer: The '**registered'** variable has the highest correlation with target variable(**cnt**)

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Answer: As the assumptions of Linear Regression apart from multicollinearity and linearity among its features, rest all assumptions are around error terms also known as residual.

Hence using Residual analysis we try to determine if model holds good on all those violations.

- **Normality**: We plot a Residual distribution plot to verify if it follow a normal distribution centred around zero.
- **Homoscedasticity**:
- **Independence of error terms**

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Answer:
Based on the coefficients we can say following 3 contribute significantly towards explaining the demand of the shared bikes:

- temp(0.522019) – The demand increases with increase in temperature
- Light_Snow_Light_Rain (-0.282214) – The demand decrease with if with weather situ ation "Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scatter ed clouds"
- yr (0.232789) – As year progress the demand increases.

General Subjective Questions

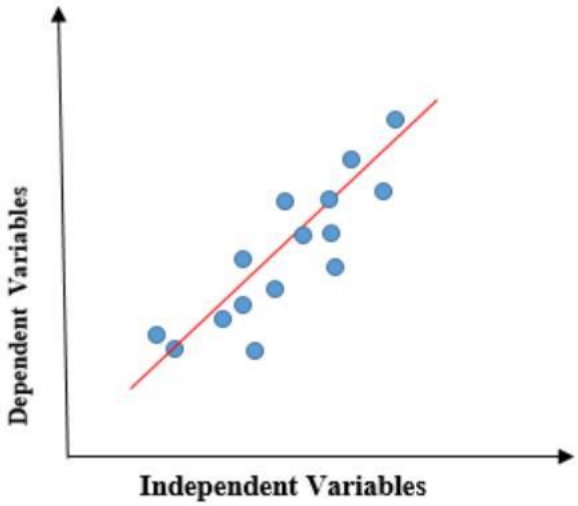
1. Explain the linear regression algorithm in detail. (4 marks)

Answer: Linear Regression algorithm is a type of supervised machine learning algorithm which is used for the prediction of numerical target variables.

It is based on the equation of a straight line

$$y=mx+c$$

It tries to find the **best fit line**(the red line in figure below) which can explain the relationship between the target variable and independent variables.



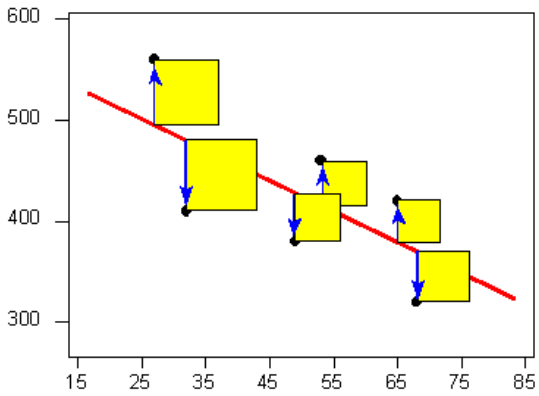
It is broadly classified in into 2 types:

- **Simple Linear Regression**(SLR): when the target variable is predicted using one dependent variable.
- **Multiple Linear Regression**(MLR): when the target variable is predicted using multiple predictors.

In case of MLR the number of variables are multiple and each one has their own coefficient making the equation look like:

$$y=m_1x_1 + m_2x_2+....+m_nx_n + c$$

Now the difference between the actual value and the one given by the equation of line is called error terms.(the yellow region below)



The algorithm tries to minimize the error terms for each point. So to mathematically approach algorithm should minimize the sum of these error term. Now since we are calculating sum and error can be both positive and negative, so to prevent error terms negating each other the error term aka residual is squared and then sum operation is performed

This sum of squared residuals is known as Cost function and algorithm tries to minimizes this to find the best-fit line:

$$S = Min \sum_{i=0}^{n} (y - \hat{y})^2$$

Some of the methods which does this :

1. Closed form method
2. Gradient descent method

Linear regression algorithm uses Gradient descent method for cost function minimization.

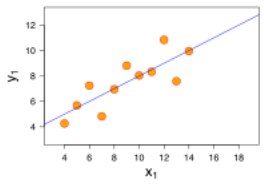
2. Explain the Anscombe's quartet in detail. (3 marks)
Answer: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, i.e., mean, standard deviation etc., yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.

Statistician Francis Anscombe constructed this in 1973 to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."

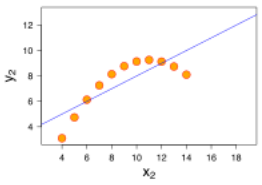
To put in simple words only using statistical summaries to comprehend data only give half picture and actual data can take lot of shapes hence the analysis should be complemented by graphing.

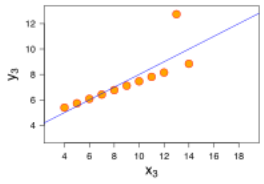
Below is an example of discount and sales(x100) in 4 cities, i.e., Mumbai, Bengaluru, Hyderabad, Kolkata.
The distribution is very different but the mean and standard deviation is same.

Mumbai

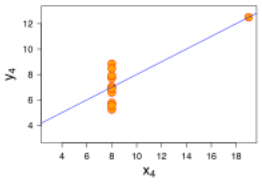


Bengaluru





Hyderabad



Kolkata

	Mumbai		Bengaluru		Hyderabad		Kolkata	
Month	Discount	Sales	Discount	Sales	Discount	Sales	Discount	Sales
January	10	804	10	914	10	746	8	658
February	8	695	8	814	8	677	8	576
March	13	758	13	874	13	1,274	8	771
April	9	881	9	877	9	711	8	884
May	11	833	11	926	11	781	8	847
June	14	996	14	810	14	884	8	704
July	6	724	6	613	6	608	8	525
August	4	426	4	310	4	539	19	1,250
September	12	1,084	12	913	12	815	8	556
October	7	482	7	726	7	642	8	791
November	5	568	5	474	5	574	8	689
Average	**9**	**750.1**	**9**	**750.1**	**9**	**750.1**	**9**	**750.1**
Std. Dev.	**3.16**	**193.7**	**3.16**	**193.7**	**3.16**	**193.7**	**3.16**	**193.7**

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

Answer: Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

- The pandas function of **corr()** uses the Person method for determination of the correlation coefficient.
- It is defined as the ratio between the covariance of two variables and the product of their standard deviations.
- It has a value between -1 and 1 and indicates the strength and direction of linear relation between two datasets. It does not indicate any info on any other type of correlation.
- It is a good estimate to study multicollinearity between two variables during Linear Regression Modelling of a dataset.

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

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

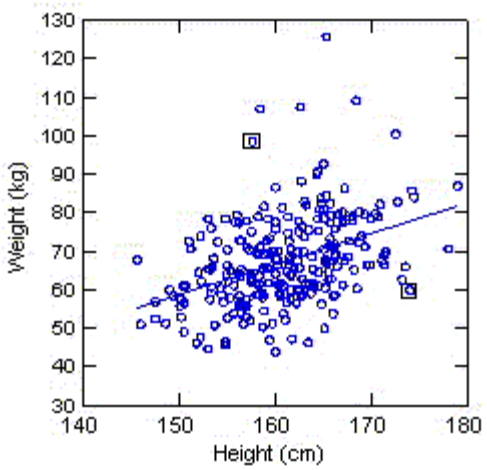
$\bar{y}$ = mean of the values of the y-variable

As it is a numerical summary of the strength of the linear association between the variables.

- So if the variables tend to go up and down together, the correlation coefficient will be positive.
- And if the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows.

As shown for plot below The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrate.



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

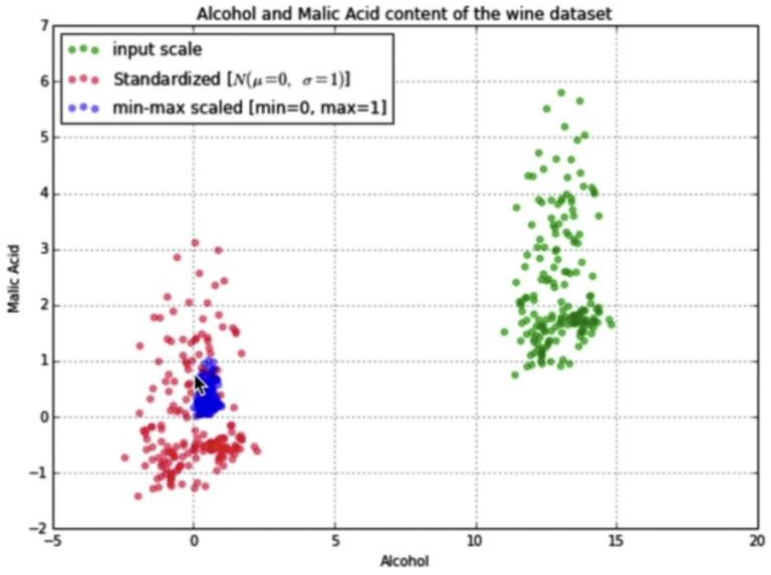
Answer: Scaling : It is a technique to standardize various variables having varied units of measurement such as kilogram, Rupees, Years, etc into unitless measures making them comparable. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

Importance of Scaling: We can speed up gradient descent by having each of our input values in roughly the same range. This is because θ will descend quickly on small ranges and slowly on large ranges, and so will oscillate inefficiently down to the optimum when the variables are very uneven.

Also if scaling not done on a dataset having varied ranged, a variable with a higher magnitude will have a small coefficient than that with a smaller magnitude. This could distort the appreciation of the coefficients that are evaluated by the model.

Difference between normalized scaling and standardized scaling:

Normalized Scaling (Min Max Scaling)	**Standardized Scaling**
Normalization involves dividing the input values by the range (i.e. the maximum value minus the minimum value) of the input variable, resulting in a new range of just 0 to 1. $x_i := \frac{x_i - x_{min}}{x_{max} - x_{min}}$	Standardization involves subtracting the average value from the values resulting in a new average value of zero. $x_i := \frac{x_i - \mu_i}{s_i}$
It compresses the data between a particular range of 0 to 1	The data will be centred around 0 mean but will have similar variance and spread.



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

Answer: Variance Inflation Factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity.

It measures how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related.

The standard error of an estimate in a linear regression is determined by four things:

- The overall amount of noise (error). The more noise in the data, the higher the standard error.
- The variance of the associated predictor variable. The greater the variance of a predictor, the smaller the standard error (this is a scale effect).
- The sampling mechanism used to obtain the data. For example, the smaller the sample size with a simple random sample, the bigger the standard error.
- The extent to which a predictor is correlated with the other predictors in a model.

$$\textbf{VIF} = 1/(1-R^2)$$

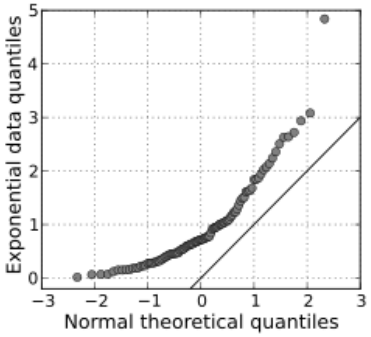
So if VIF is infinite, then it means $1 - R^2$ is zero which implies $R^2 = 1$.

It means 100% of the variance in the y-variable is explained by the x-variable. It represents a perfect straight line with every y-variable falling on the line itself.

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

Answer: 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 is shown below:



If the two distributions which we are comparing are exactly equal, then the points on the Q-Q plot will perfectly lie on a straight-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.

The Q-Q plot helps us determine:

- Whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behaviour.

In Linear Regression we can use Q-Q plot to assess for **Normality** assumption of our model. We can use any continuous distribution as a comparison, as long as we can calculate the quantiles. In fact, a common procedure is to test out several different distributions with the Q-Q plot to see if one fits your data well.