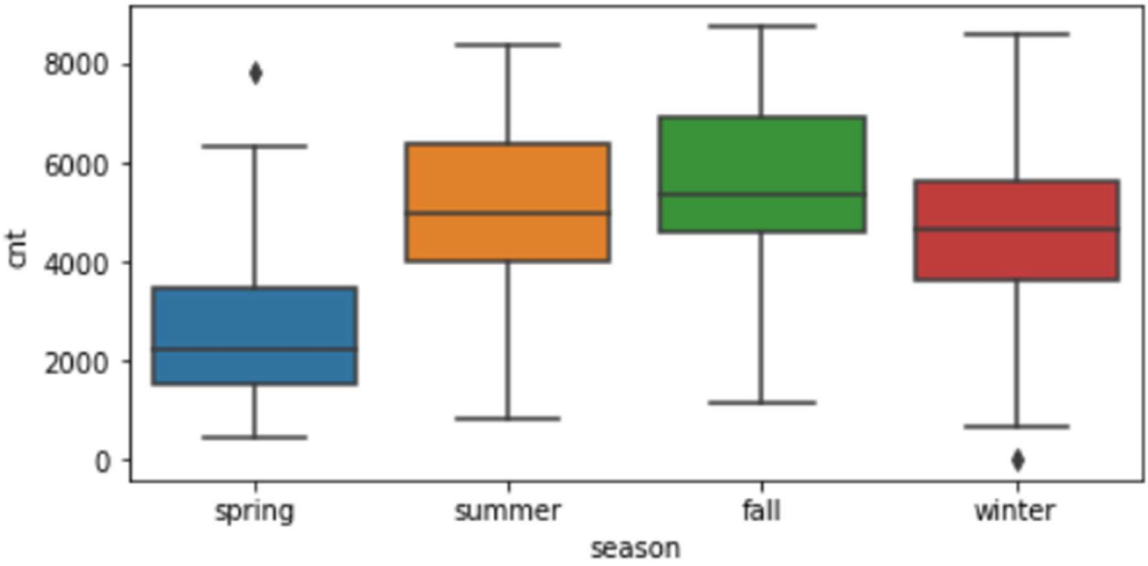
Name: Manish Kumar Singh

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)

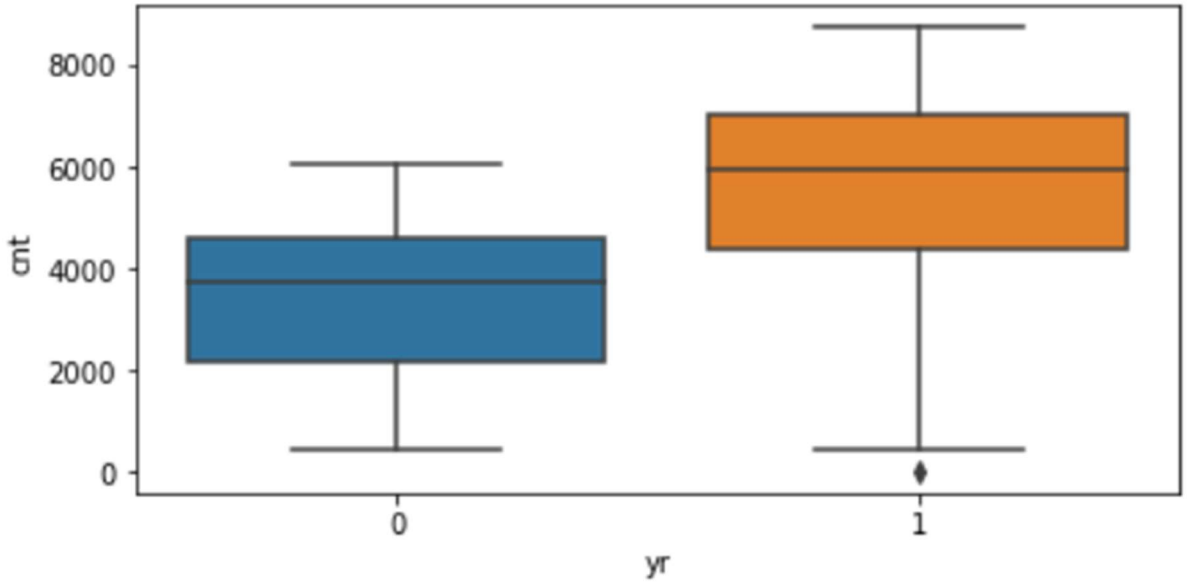
Ans: Dataset we have total seven types of categorical variables:

I) Season



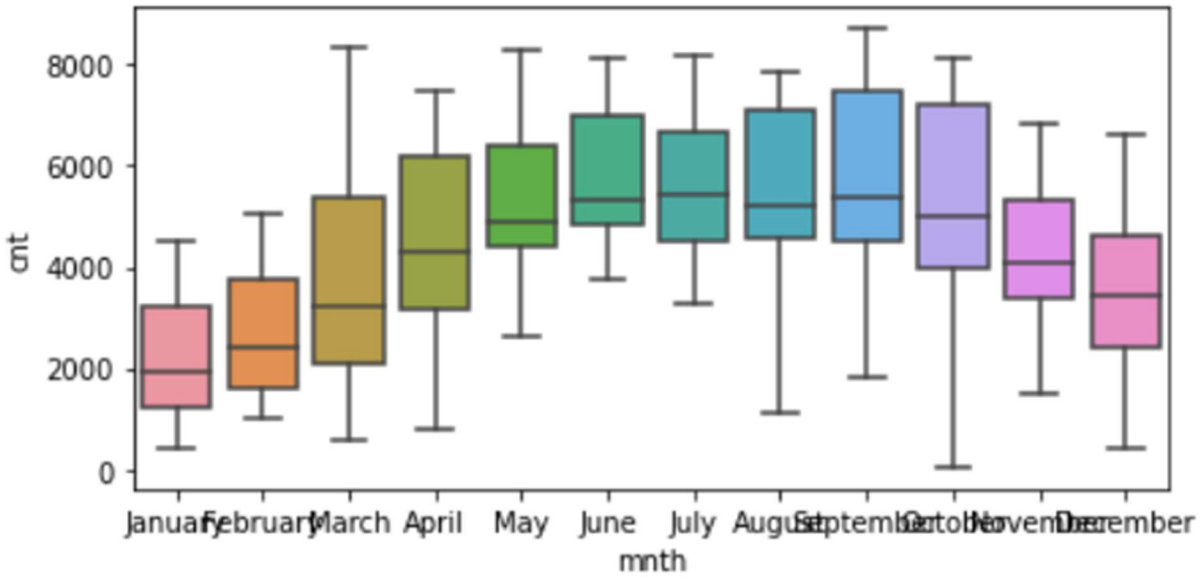
Renting of bikes is maximum in fall and then followed by summer, winter & spring.

II) Year



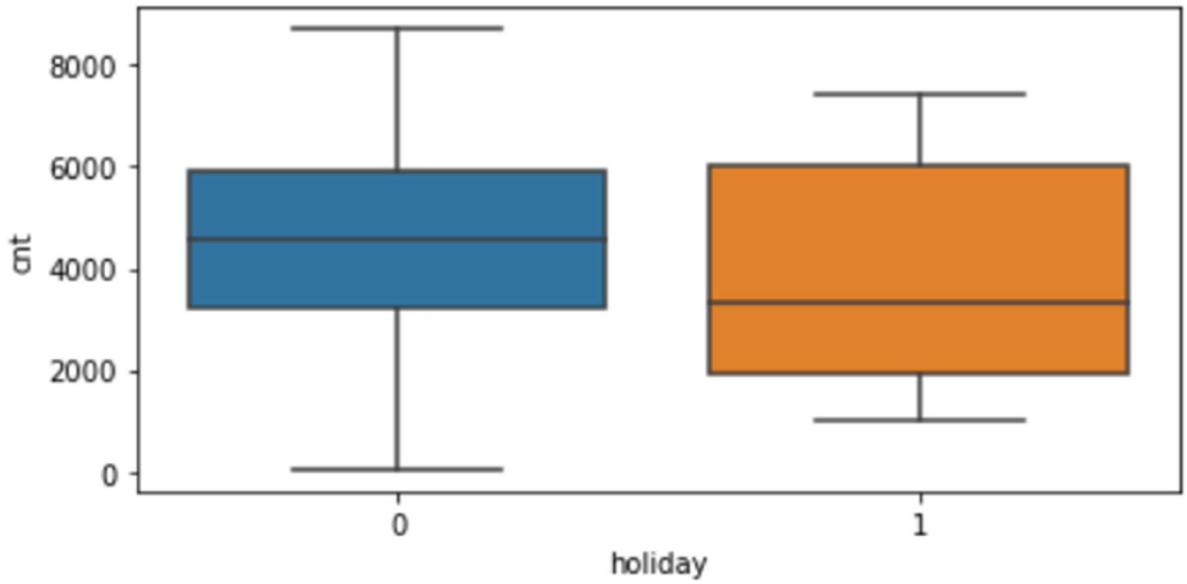
Renting of bikes has increased significantly in 2019 as compare to 2018
(0 = 2018 & 1=2019)

III) Month



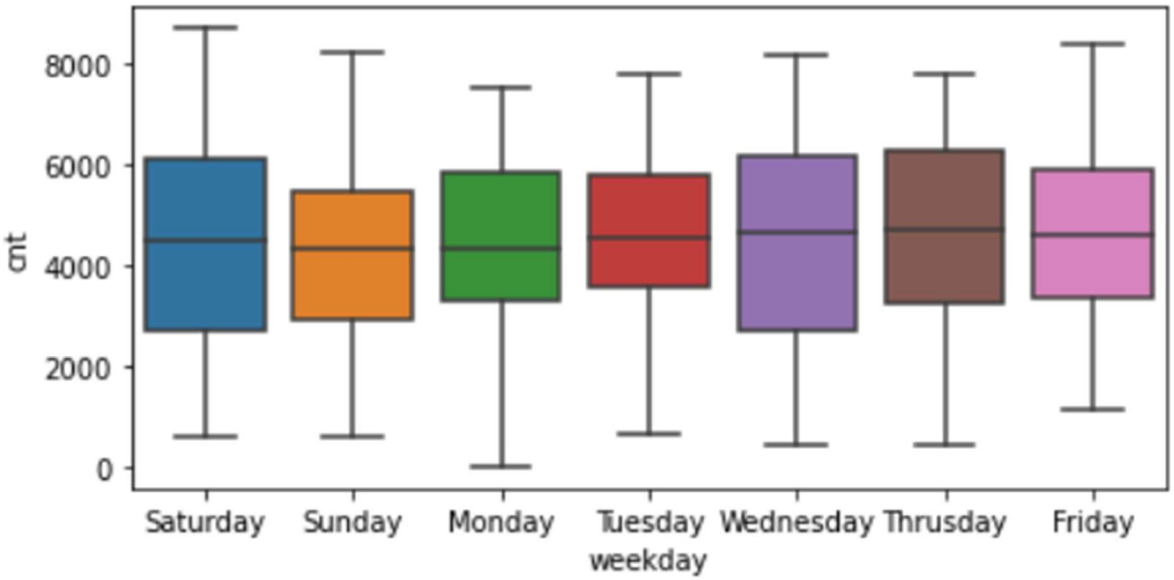
Renting of bikes seems to have more from May to October as compare to other months.

IV) Holiday



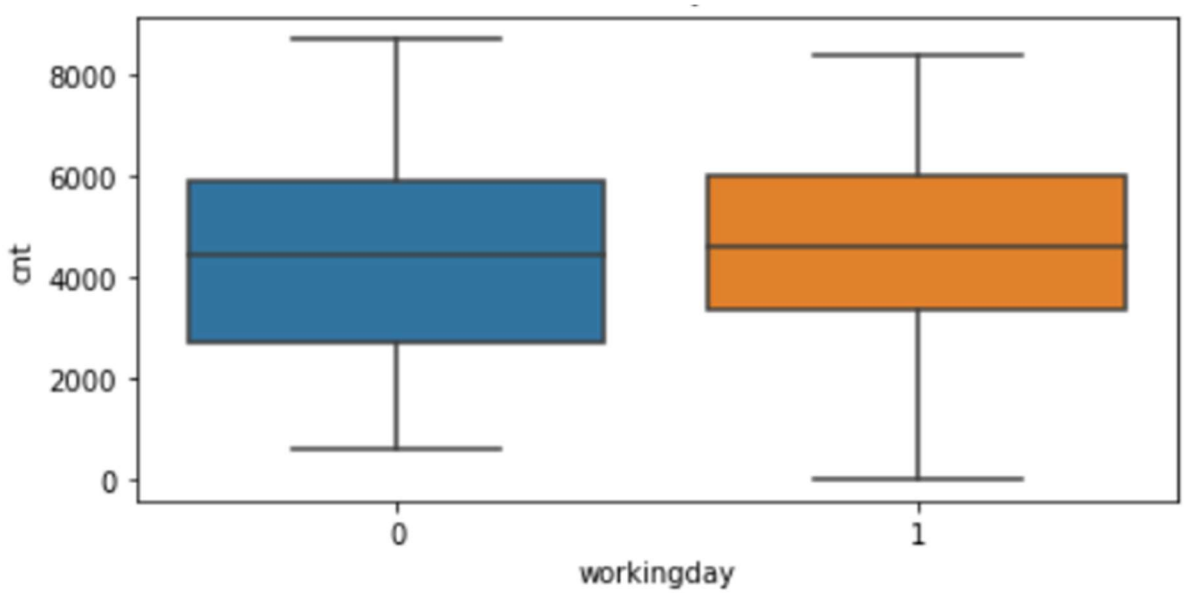
Renting of bikes seems to have more on holidays as compare to any other day. (0=holiday, 1=not holiday)

V) Weekday



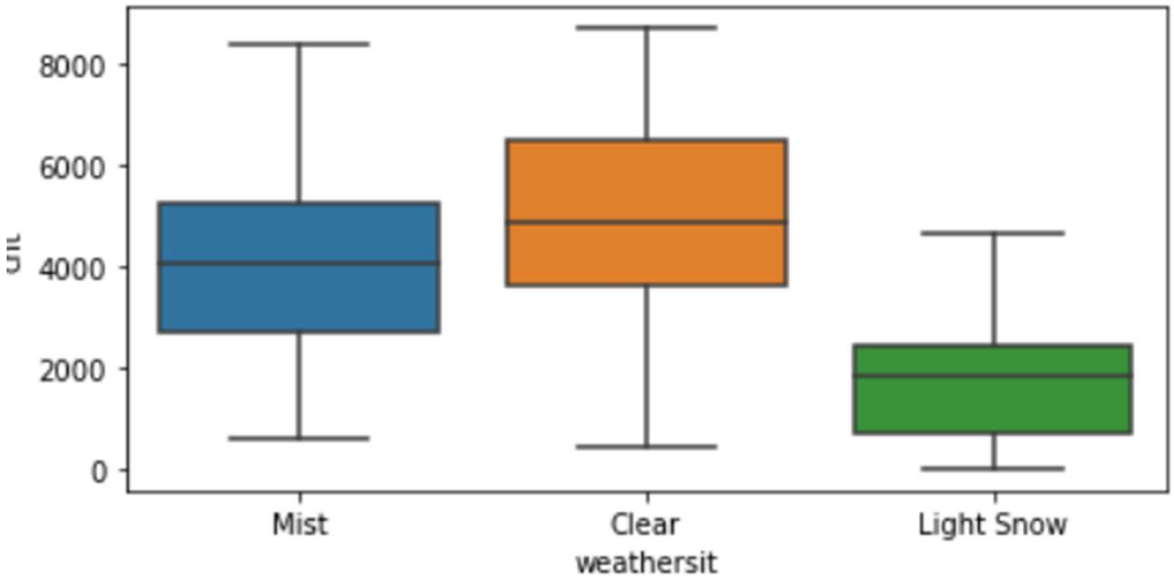
Renting of bikes seems similar on the days of week without any considerable variation.

VI) Working day



Renting of bikes seems almost similar on the working/non-working day.

VII) Weathersit



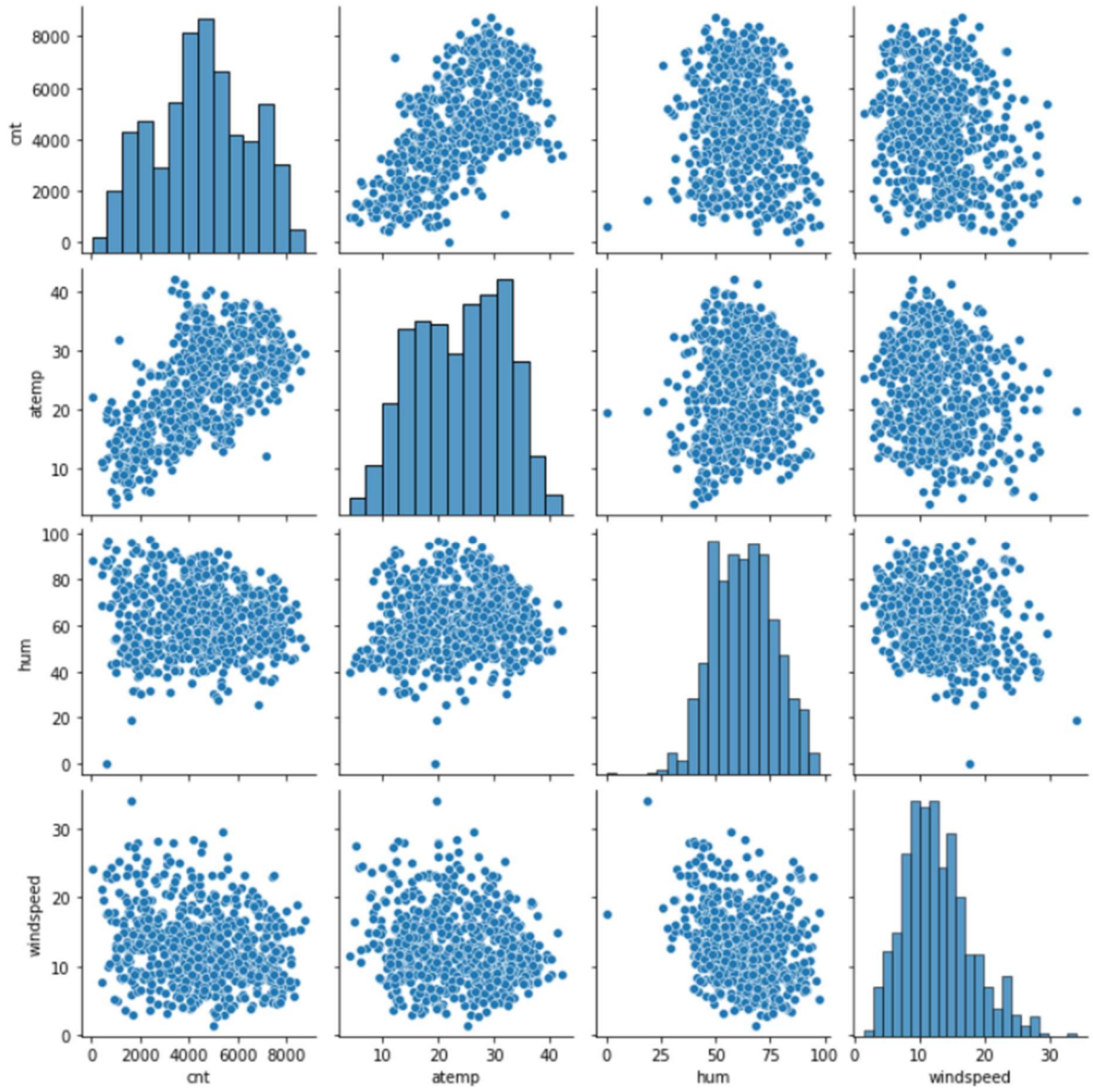
Renting of bikes is more on clear weather and then followed by Mist & Light Snow

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

(2 mark)

Ans: This is not needed since one of the combinations will be uniquely representing this redundant column. Hence, it's better to drop one of the column. This Overall approach reduces multi-collinearity in the dataset, which is one of the prime assumption of Multiple Linear Regression.

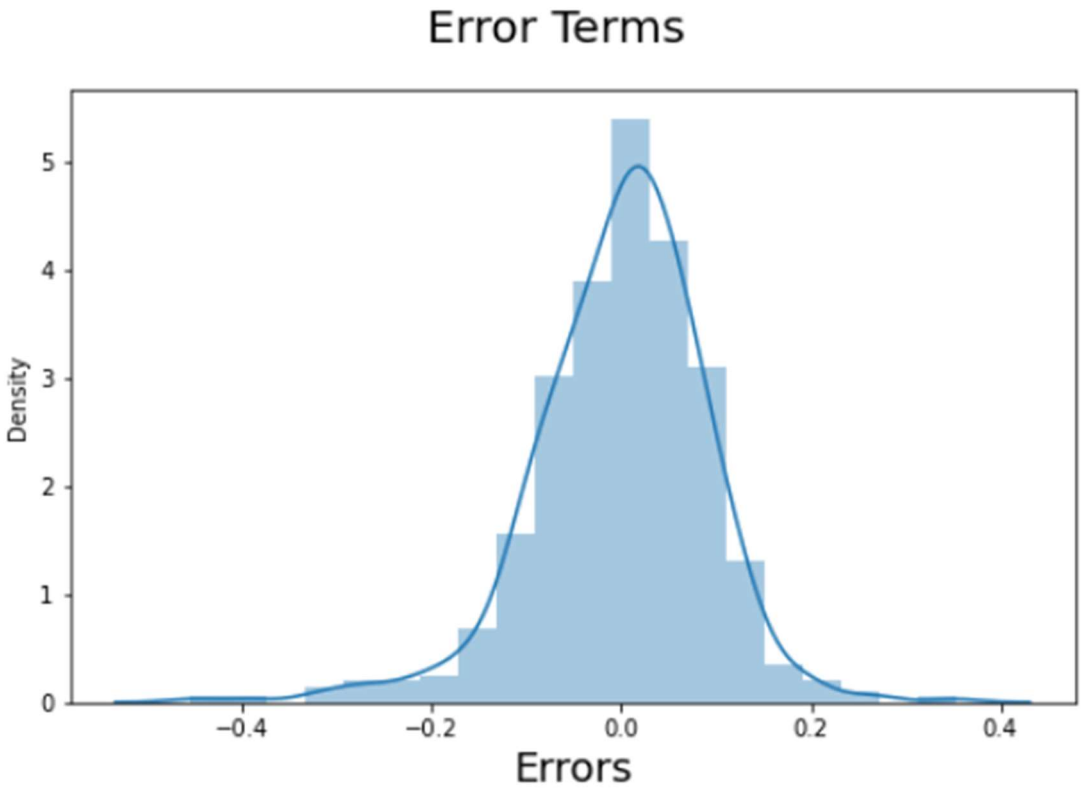
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)



atemp is highly corelated with cnt.

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

Ans:



Residuals distribution should follow normal distribution and centred around 0.(mean = 0). We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not. The above diagram shows that the residuals are distributed about mean = 0.

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

Ans: The top 3 features are

- atemp - coefficient: 0.479
- weaathersit_Light Snow - coefficient: -0.242
- yr – coefficient: 0.252

General Subjective Questions

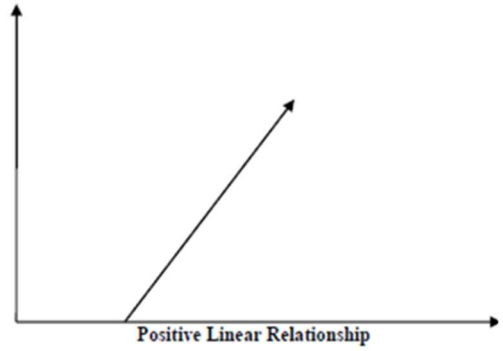
1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (Increase or decrease).

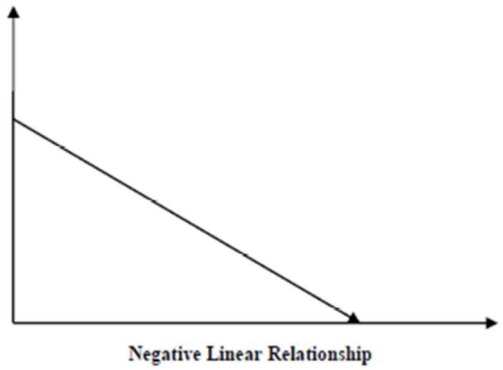
Mathematically the relationship can be represented with the help of following equation – y =mx +c

Here, y is the dependent variable we are trying to predict.
x is the independent variable we are using to make predictions.
m is the slop of the regression line which represents the effect x has on y
c is a constant, known as the y-intercept. If x = 0,y would be equal to c.
The linear relationship can be positive or negative in nature.

Positive Linear Relationship - A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph −



Negative Linear Relationship- A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph –



Types of Linear Regression:

- **Simple Linear Regression** - SLR is used when the dependent variable is predicted using only one independent variable.
- **Multiple Linear Regression** - MLR is used when the dependent variable is predicted using multiple independent variables.

Assumptions:

The following are some assumptions about dataset that is made by Linear Regression model −

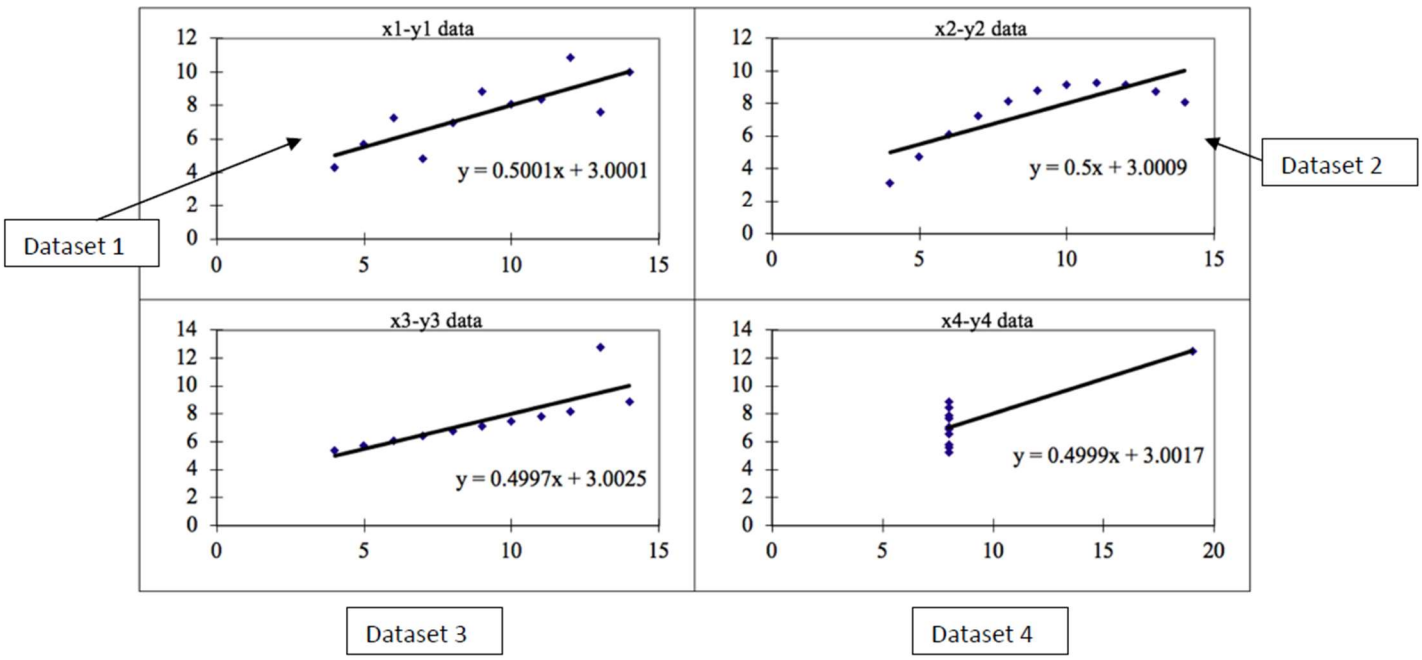
Multi-collinearity − Linear regression model assumes that there is very little or no multicollinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

Auto-correlation − Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

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

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

Ans: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics. They have very different distributions and appear differently when plotted on scatter plots. It was constructed by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. The four data sets which provides same statistical information that involves variance, and mean of all x,y points in all four datasets. 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 datasets. These four plots can be defined as follows:



The four datasets can be described as:
1. Dataset 1: This fits the linear regression model pretty well.
2. Dataset 2: this could not fit linear regression model on the data quite well as the data is nonlinear.
3. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model.
4. Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

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

Ans: In statistics, Pearson's r or Pearson's correlation coefficient is defined as the measurement of the strength of the relationship between two variables and their association with each other. The relationship of the variables is measured with the help Pearson correlation coefficient calculator. This linear relationship can be positive or negative and its value ranges between -1 to +1. It shows the linear relationship between two sets of data
1) r = 1 means the data is perfectly linear with a positive slope.
2) r = -1 means the data is perfectly linear with a negative slope.
3) r = 0 means there is no linear association.

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

Ans: Feature scaling is a method used to normalize the range of independent variables or
features of data. In data processing, it is also known as data normalization and is generally
performed during the data pre-processing step.
Normalization: Also known as min-max scaling or min-max normalization, it is the simplest
method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively.
Standardization: Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Here, σ is the standard deviation of the feature vector, and x̄ is the average of the feature vector.

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

Ans: Variance Inflation Factors (VIFs) provide a one-number summary description of collinearity for each model term and is defined as $VIF_i = 1/1-R^2_i$ where R^2_i is the coefficient of determination of a regression model where the ith factor is treated as a response variable in the model with all of the other factors. If there is perfect correlation, then VIF = infinity. Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1.So, VIF = 1/(1-1) which gives VIF = 1/0 which results in infinity.

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

Ans:The quantile-quantile (Q-Q) plot is a graphical representation for determining if two data sets come from populations with a common distribution. A Q-Q plot of the quantiles of the first data set against the quantiles of the second data set. Here quantile, the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. If both the sets of quantiles came from the same distribution, we should see the points forming a line that is roughly straight.