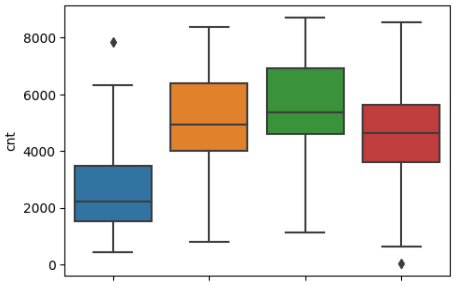
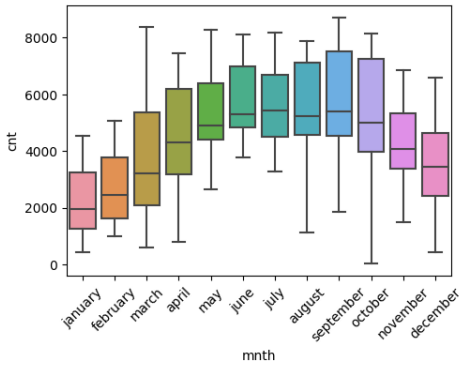
Bike Sharing Assignment-based Subjective Questions

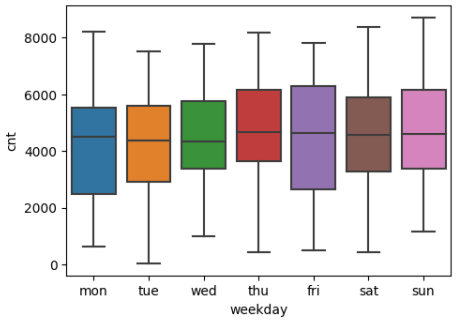
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

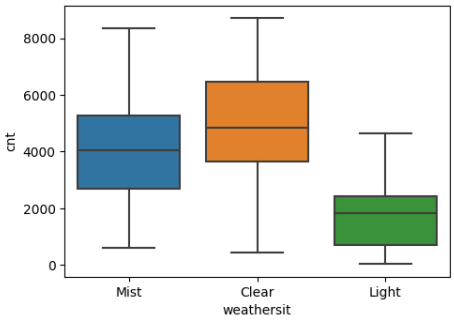
There are a few variables namely season, mnth, yr, weekday, holiday, working day and weathersit. I have carried out visualization by plotting both box plot and bar plot for the above.

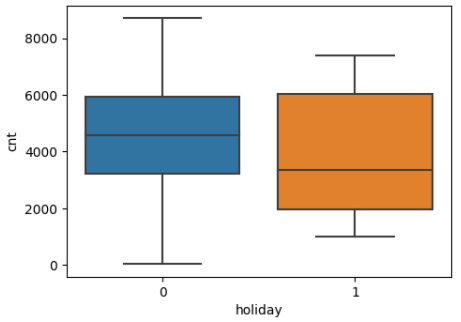
Box plot.

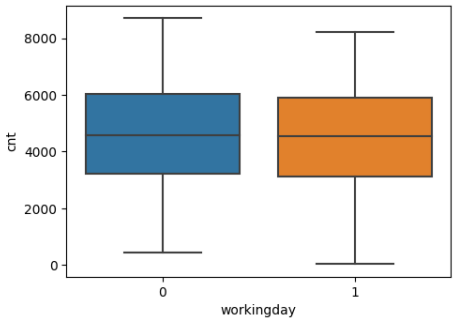


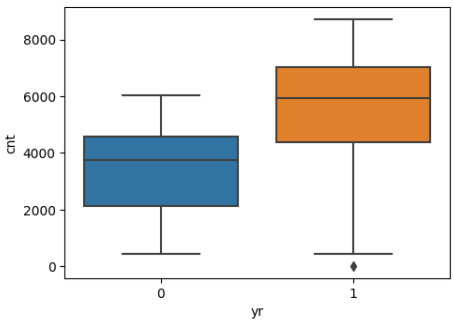


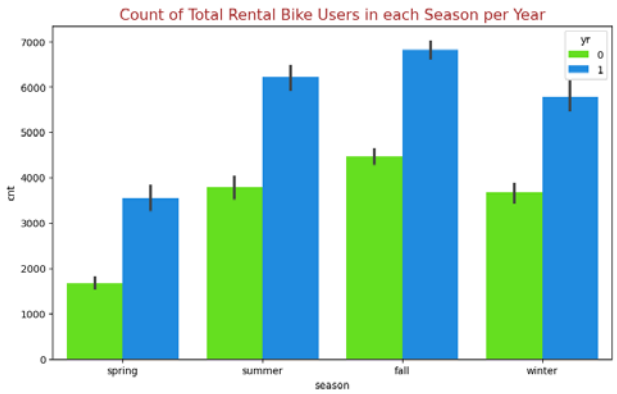


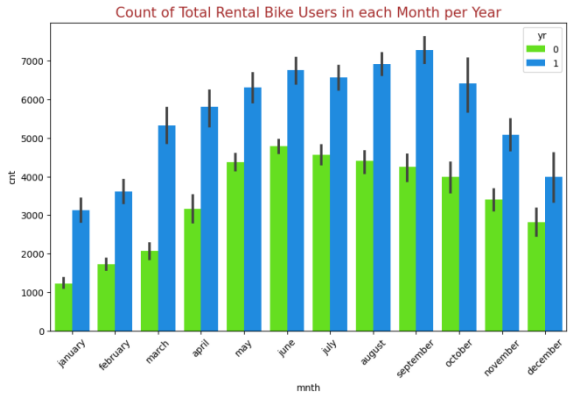


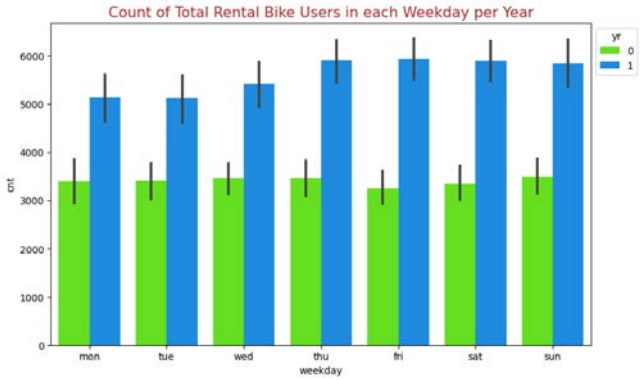


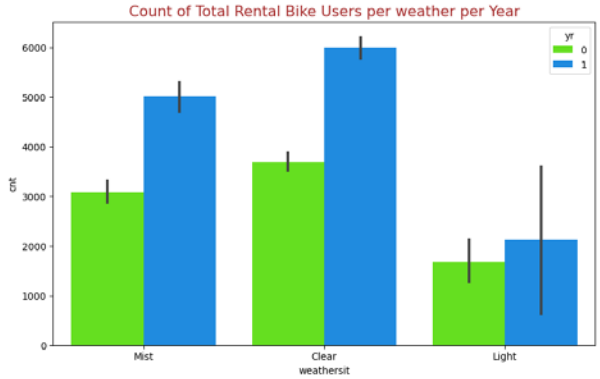


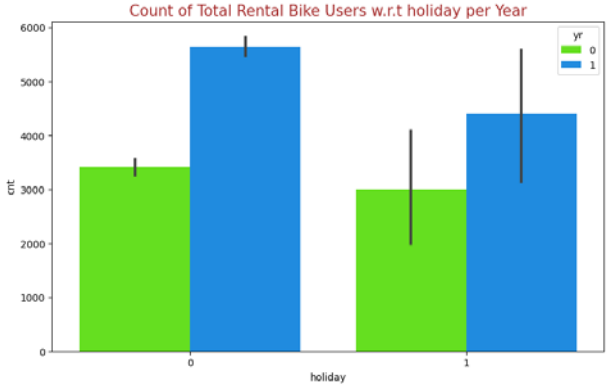


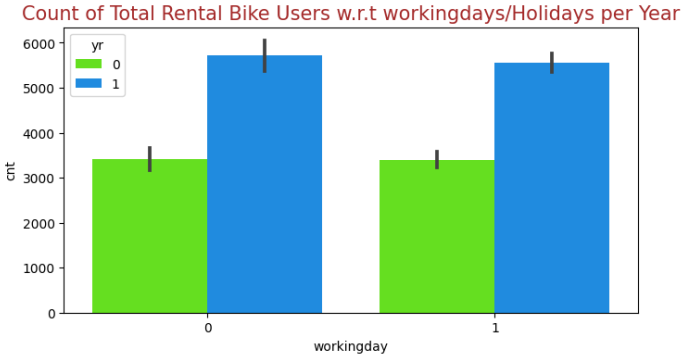












Season - Season 'fall' has the highest demand where as season 'spring' has the lowest demand for rental bikes. There is a substantial increase in the demand as the year progressed from 2018 to 2019.

Month - The plots showed an increasing trend from starting of the year till mid of the year with most of the bookings done during the month of may, june, july, aug, sep and oct and then it started decreasing as we approached the end of year. August has the highest demand for rental bikes. There is a substantial increase in the demand per month as the year progressed from 2018 to 2019.

Weekdays - Thursdays, Fridays, Saturdays and Sundays has more demand compared to other weekdays.

Weather - Clear weather attracted more booking which is natural.There is a increase in the demand for each weather situation as the year progressed from 2018 to 2019.

Holiday - Holidays have more bookings compared to working days.

Working Days - No of bookings seemed to be almost similar either on working day or non-working day. But the count increased from 2018 to 2019.

Year - 2019 attracted more number of booking compared to 2018, which shows good progress in terms of business.

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

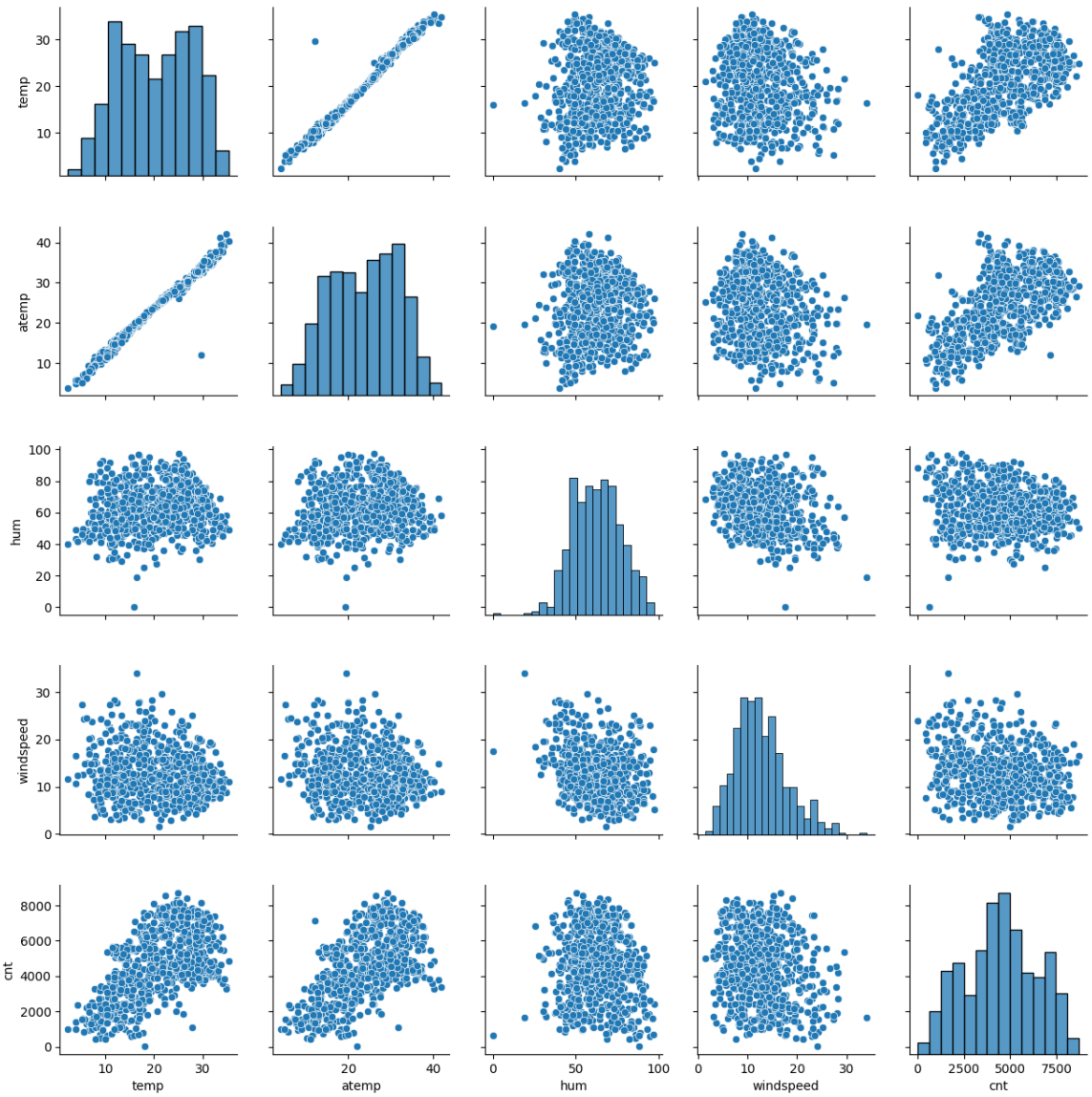
If we do not use drop_first = True, then n dummy variables will be created, and these predictors (n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

For Example, you have a column for gender that contains 3 variables- "Male", "Female", and “Other ". So a person is either "Male", or "Female. If they are not either of these 3, then their gender is "Other".

Hence if we have 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?

Using the below pair plot it can be seen that, “temp” and “atemp” are the two numerical variables which are highly correlated with the target variable (cnt)



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

The following tests were done to validate the assumptions of linear regression:

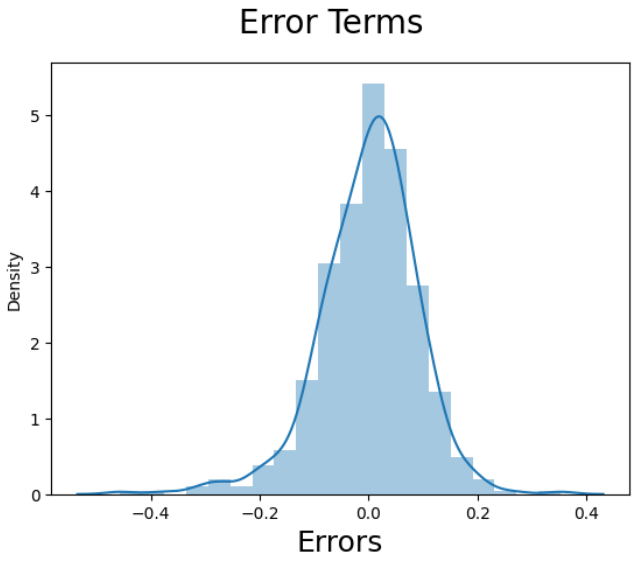
1. Linear regression needs the relationship between the independent and dependent variables to be linear. We visualized the numeric variables using a pair plot to see if the variables are linearly related or not. Refer notebook for more details.

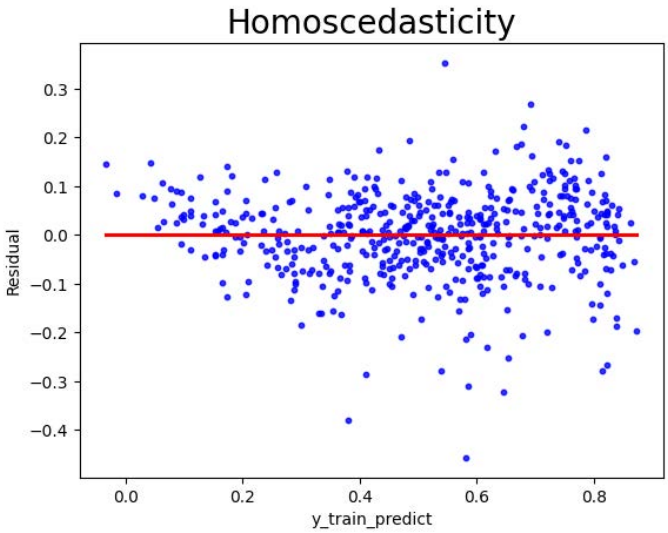
2. Residuals distribution should follow normal distribution and centered around 0 (mean = 0). We validated this assumption about residuals by plotting a dist plot of residuals and saw if residuals

are following normal distribution or not. The diagram below shows that the residuals are distributed about mean = 0.

3. Linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to get the quantitative idea about how much the feature variables are correlated with each other in the new model. Refer notebook for more details.

4. Homoscedasticity: From the linear regression model fit plot that the residuals are distributed equally along both sides of the best fit line (predicted values). There is no high or low concentration of residuals in certain regions which proves Homoscedasticity 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?

Top 3 features that has significant impact towards explaining the demand of the shared bikes are temperature, year and weather situation.

1) temp: coefficient = 0.4019

2) year: coefficient = 0.2350

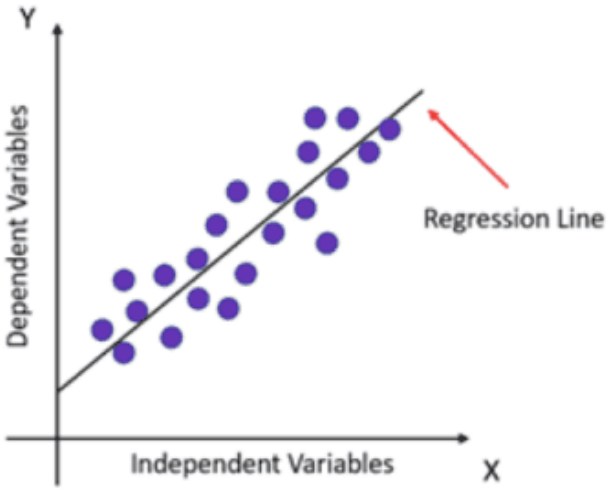
3) weather situation(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds): coefficient = -0.2939

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is an algorithm that belongs to supervised Machine Learning. It tries to apply relations that will predict the outcome of an event based on the independent variable data points. The relation is usually a straight line that best fits the different data points as close as possible. The output is of a continuous form, i.e., numerical value. For example, the output could be revenue or sales in currency, the number of products sold, etc. In the above example, the independent variable can be single or multiple.

a) Linear Regression Equation



Linear regression can be expressed mathematically as:

$Y = \beta_0 + \beta_1 X$

Here,

- Y= Dependent Variable
- X= Independent Variable
- β 0= intercept of the line
- β1 = Linear regression coefficient (slope of the line)

b) Linear Regression Model

Since the Linear Regression algorithm represents a linear relationship between a dependent (y) and one or more independent (y) variables, it is known as Linear Regression. This means it finds how the value of the dependent variable changes according to the change in the value of the independent variable. The relation between independent and dependent variables is a straight line with a slope.

Types of Linear Regression

Linear Regression can be broadly classified into two types of algorithms:

1. Simple Linear Regression

A simple straight-line equation involving slope (dy/dx) and intercept (an integer/continuous value) is utilized in simple Linear Regression. Here a simple form is:
y=mx+c where y denotes the output x is the independent variable, and c is the intercept when x=0. With this equation, the algorithm trains the model of machine learning and gives the most accurate output

2. Multiple Linear Regression

When a number of independent variables more than one, the governing linear equation applicable to regression takes a different form like:

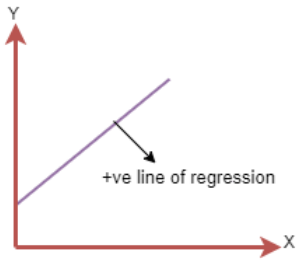
y= c+m1x1+m2x2… mnxn where represents the coefficient responsible for impact of different independent variables x1, x2 etc. This machine learning algorithm, when applied, finds the values of coefficients m1, m2, etc., and gives the best fitting line.

Linear Regression Line

A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

- **Positive Linear Relationship:**

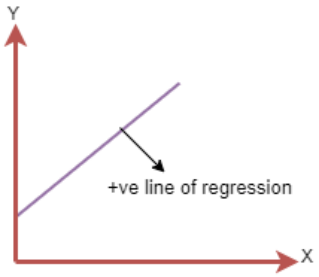
If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: $Y= a_0+a_1x$

- **Negative Linear Relationship:**

If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



The line equation will be: $Y= a_0+a_1x$

Assumption for Linear Regression Model

Linear regression is a powerful tool for understanding and predicting the behavior of a variable, however, it needs to meet a few conditions in order to be accurate and dependable solutions.

Linearity: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion.

Independence: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.

Homoscedasticity: Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors.

Normality: The errors in the model are normally distributed.

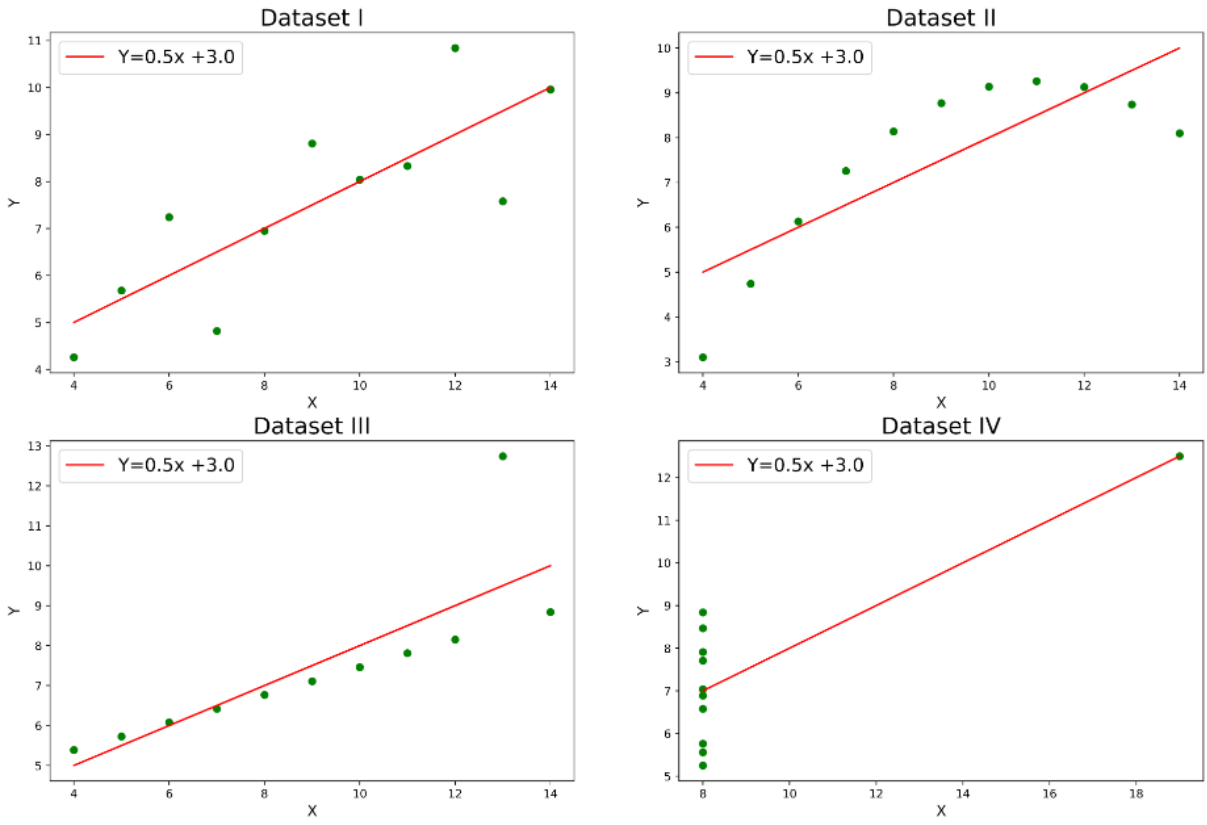
No multicollinearity: there is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is the model 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 of 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 representation. Each graph plot shows the different behavior irrespective of statistical analysis.

Anscombe's Data									
Observation	x1	y1		x2	y2	x3	y3	x4	y4
1	10	8.04		10	9.14	10	7.46	8	6.58
2	8	6.95		8	8.14	8	6.77	8	5.76
3	13	7.58		13	8.74	13	12.74	8	7.71
4	9	8.81		9	8.77	9	7.11	8	8.84
5	11	8.33		11	9.26	11	7.81	8	8.47
6	14	9.96		14	8.1	14	8.84	8	7.04
7	6	7.24		6	6.13	6	6.08	8	5.25
8	4	4.26		4	3.1	4	5.39	19	12.5
9	12	10.84		12	9.13	12	8.15	8	5.56
10	7	4.82		7	7.26	7	6.42	8	7.91
11	5	5.68		5	4.74	5	5.73	8	6.89
				Summary Statistics					
N	11	11		11	11	11	11	11	11
mean	9.00	7.50		9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94		3.16	1.94	3.16	1.94	3.16	1.94
r	0.82			0.82		0.82		0.82	

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behavior.



We can describe the four data sets as:

ANSCOMBE'S QUARTET FOUR DATASETS

- Data Set 1: fits the linear regression model pretty well.
- Data Set 2: cannot fit the linear regression model because the data is non-linear.
- Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

As you can see, Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

3. What is Pearson's R?

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.

Assumptions:

Independent of case: Cases should be independent to each other.
Linear relationship: Two variables should be linearly related to each other. This can be assessed by plotting the value of variables on a scatter diagram, and check if the plot yields a relatively straight line.
Homoscedasticity: the residuals scatter plot should be roughly rectangular-shaped.

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 of the 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?

It is a step of data Pre-Processing which is applied to independent variables to transforming the values of features or variables in a dataset to a similar scale. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not unit's 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.

Difference between Normalization and Standardization

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

VIF - Variance Inflation Factor

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

If value of VIF is infinite then it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead 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.

When R2 is equal to 0, and therefore, when VIF is equal to 1, the independent variable is not correlated to the remaining ones, meaning that multicollinearity does not exist.

In general terms,

- VIF equal to 1 = variables are not correlated
- VIF between 1 and 5 = variables are moderately correlated
- VIF greater than 5 = variables are highly correlated

The higher the VIF, the higher the possibility that multicollinearity exists, and further research is required. When VIF is higher than 10, there is significant multicollinearity that needs to be corrected.

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

The quantile-quantile or q-q plot is an exploratory graphical device used to check the validity of a distributional assumption for a data set. In general, the basic idea is to compute the theoretically expected value for each data point based on the distribution in question. If the data indeed follow the assumed distribution, then the points on the q-q plot will fall approximately on a straight line.

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. Q–Q plots can be used to compare collections of data, or theoretical distributions. The use of Q–Q

plots to compare two samples of data can be viewed as a non-parametric approach to comparing their underlying distributions. A Q–Q plot is generally more diagnostic than comparing the samples' histograms, but is less widely known. Q–Q plots are commonly used to compare a data set to a theoretical model. This can provide an assessment of goodness of fit that is graphical, rather than reducing to a numerical summary statistic. Q–Q plots are also used to compare two theoretical distributions to each other. Since Q–Q plots compare distributions, there is no need for the values to be observed as pairs, as in a scatter plot, or even for the numbers of values in the two groups being compared to be equal.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.