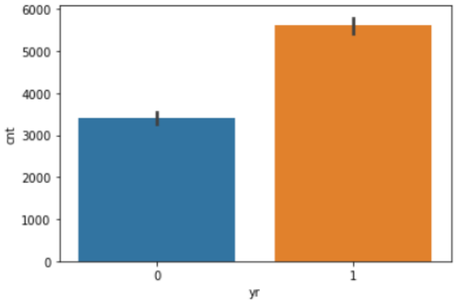
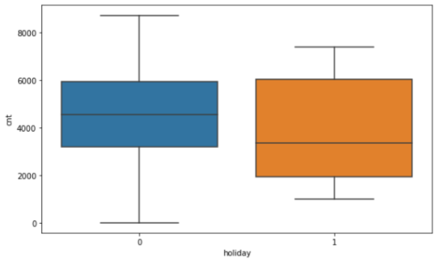
Assignment-based Subjective

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)

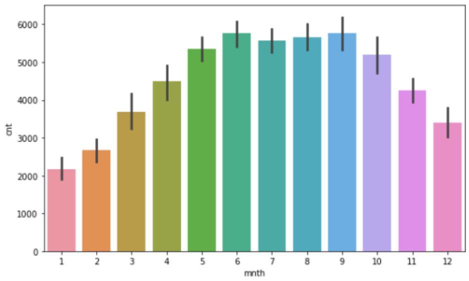
1. There is 64% growth in average daily user of bike in year 2019 compared to 2018.



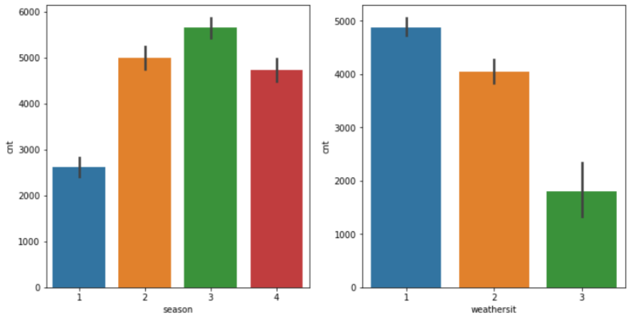
2. People are using more bikes on Non-holidays.



3. Usage of bike is more during month of June to Sept.



4. Usage of bike is high during fall season & during clear sky. As the climate changes to have snow & rain people are avoiding bikes.



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

Answer – While working on on-hot encoding on categorical variable, we create as many columns as many category we have. Example , we have column as season which takes 3 categories

1. Spring 2.Summer 3.fall 4. Winter

So whenever wherever we have season as ‘1’, the spring column will have value as ‘1’ others will have value as ‘0’.

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

Setting drop_first = True causes get_dummies to exclude the dummy variable for the first category of the variable you're operating on.

season_dummy = pd.get_dummies(day['season'],drop_first = True) will give us below.

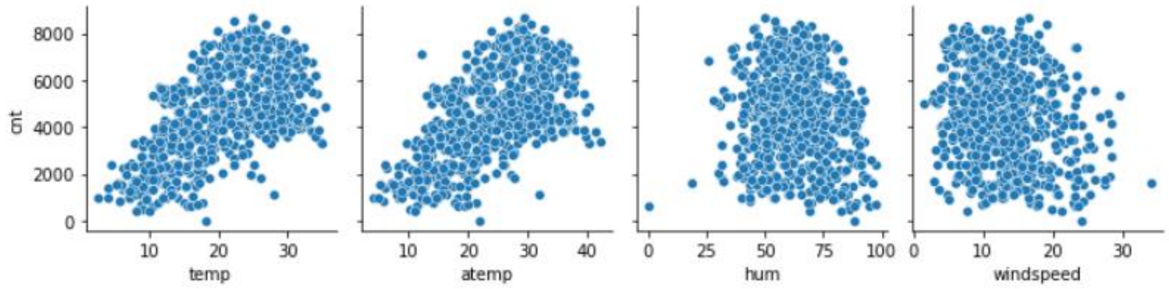
This will help us to reduce the number of columns being produced while performing One-hot encoding.

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

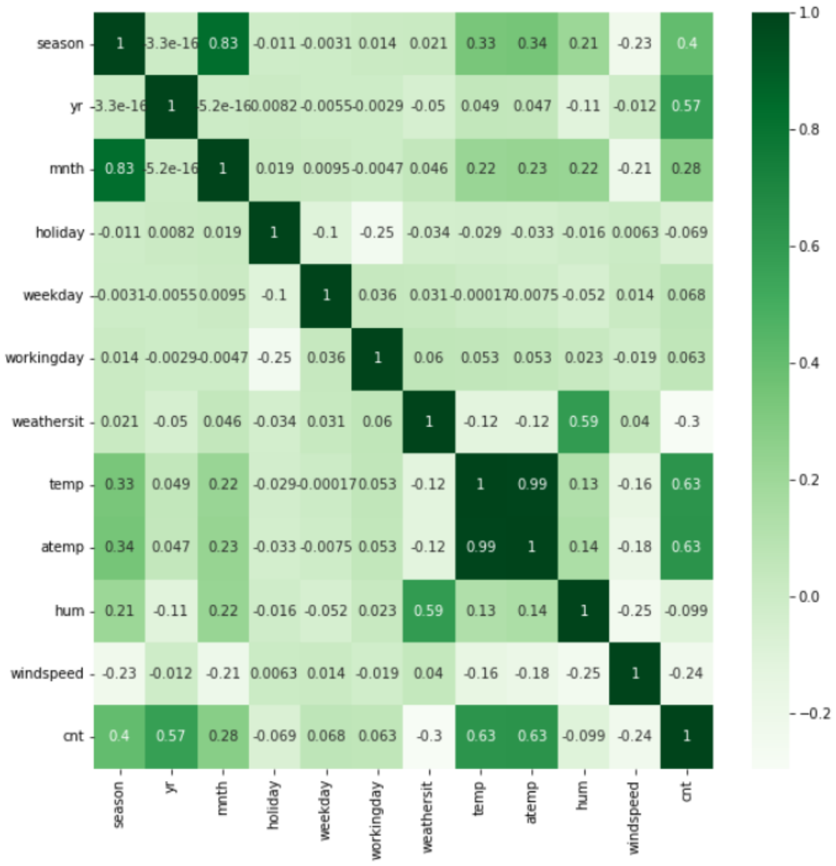
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Based on pair plot, ‘temp’ variable has highest correlation with target variable. See below pair plot.

```
<Figure size 1440x864 with 0 Axes>
```



Also based on correlation matrix, we can see ‘temp’ has highest correlation of 0.63 with target variable.



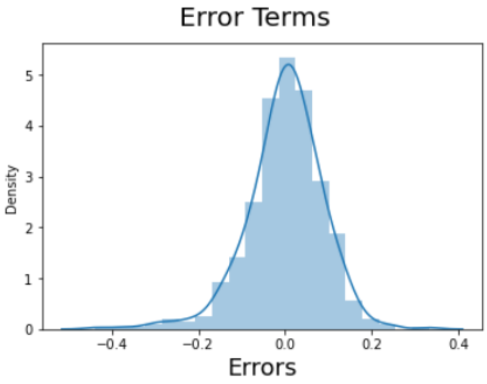
	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	cnt
season	1	3.3e-16	0.83	-0.011	-0.0031	0.014	0.021	0.33	0.34	0.21	-0.23	0.4
yr	-3.3e-16	1	5.2e-16	0.0082	-0.0055	-0.0029	-0.05	0.049	0.047	-0.11	-0.012	0.57
mnth	0.83	5.2e-16	1	0.019	0.0095	-0.0047	0.046	0.22	0.23	0.22	-0.21	0.28
holiday	-0.011	0.0082	0.019	1	-0.1	-0.25	-0.034	-0.029	-0.033	-0.016	0.0063	-0.069
weekday	-0.0031	-0.0055	0.0095	-0.1	1	0.036	0.031	-0.00017	-0.0075	-0.052	0.014	0.068
workingday	0.014	-0.0029	-0.0047	-0.25	0.036	1	0.06	0.053	0.053	0.023	-0.019	0.063
weathersit	0.021	-0.05	0.046	-0.034	0.031	0.06	1	-0.12	-0.12	0.59	0.04	-0.3
temp	0.33	0.049	0.22	-0.029	-0.00017	0.053	-0.12	1	0.99	0.13	-0.16	0.63
atemp	0.34	0.047	0.23	-0.033	-0.0075	0.053	-0.12	0.99	1	0.14	-0.18	0.63
hum	0.21	-0.11	0.22	-0.016	-0.052	0.023	0.59	0.13	0.14	1	-0.25	-0.099
windspeed	-0.23	-0.012	-0.21	0.0063	0.014	-0.019	0.04	-0.16	-0.18	-0.25	1	-0.24
cnt	0.4	0.57	0.28	-0.069	0.068	0.063	-0.3	0.63	0.63	-0.099	-0.24	1

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

Answer : We are validating with below assumption on linear regression.

1. **Error terms are normally distributed with mean as zero.**

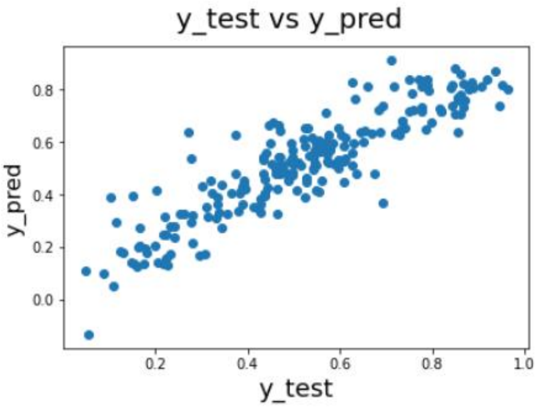
```
Text(0.5, 0, 'Errors')
```



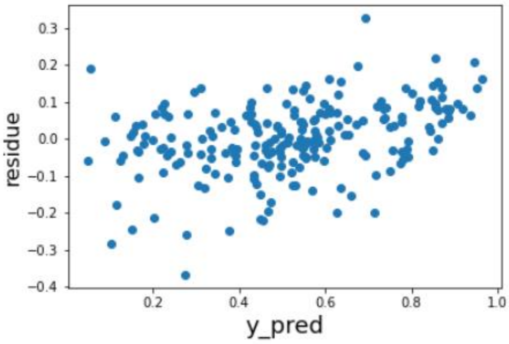
2. **There is linear relationship between X & Y and Homoscedasticity.**

The relation between X & Y is linear. Also there is NO patter in error terms and they have constant variance.

```
Text(0, 0.5, 'y_pred')
```



3. **Error terms are independent of each other.**



Question 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 – Below are top 3 contributors towards explaining demand of bikes.

1. **Temperature** – With positive co-efficient of 0.51 it remains the top contributing feature. People are preferring to ride bike when temperature rises. High users were recorded with temperature between 20 to 30 degree celcius.

2. **Weather with Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds** with negative co-efficient of -0.2942. **It appears that users avoiding the bikes in such weather, hence affects demand negatively.**

3. **Year** with positive co-efficient of 0.2329. The bike demand has grown by almost 64% in 2019 compared to previous year.

Apart from above 3 variables below variables (with co-efficient) are significant contributors for prediction

windspeed	(-0.1532)	People are avoiding such weather & negatively affecting bike usage.
winter	(0.1267)	People use bikes more in winter hence positively affecting bike usage.
sept	(0.1189)	Highest users were recorded in month of sept.

General Subjective Questions

Question 1 : Explain the linear regression algorithm in detail.

Answer : Linear regression is a statistical method for modelling the relationship between a scalar response of and one or more explanatory variables (also known as dependent and independent variables).

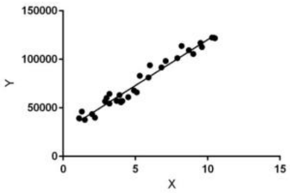
If the relation-ship is being established between response variable (Y) with one independent variable, then it is called as **Simple linear regression. (SLR)**

If the relationship is being established between response variable (Y) based on more than one independent variables, then it is called as **Multiple Linear regression (MLR)**.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

In machine learning, linear regression is part of **supervised learning methods** and is used to predict the value of target variable based on independent variables.

Example – In business, 'Profit' would be dependent variables (Y) and it can be predicted based on independent variables like Money invested on marketing campaign, raw material cost, economic factors, competitor sales, overall sales etc.



While training the model we are given :
x: input training data
y: labels to data (supervised learning)

Simple linear regression model equation –
Y = a + bX where a – intercept i.e. value of Y when X is zero ,
& b is slope of graph i.e. change in one unit of Y with change in 1 unit of X.

Multiple linear regression model equation –
Y = a + b1X1+ b2X2 +…+bnXn where a – intercept i.e. value of Y when X is zero ,
& b1 to bn are co-efficient of independent variables X which states that change in value Y based on one unit change in X1 provided other variables are constant.

Assumption for linear regression:
1. X and Y should display some sort of a linear relationship; otherwise, there is no use of fitting a linear model between them.
2. Error terms are normally distributed with mean zero(not X, Y)
3. Error terms are independent of each other.

4.Error terms have constant variance (homoscedasticity)

In machine learning, linear regression algorithm steps are below-

1. Reading and understanding the data
Cleaning and manipulating data to make it up to the standards that exploratory data analysis can be performed by treating null values if any, updating to necessary formats, changing data types if needed, removing unwanted rows or columns etc. The raw data in whatever condition you get must be squeaky cleaned of any muck before assessing it for visualization.

2. Visualizing the data (Exploratory Data Analysis)
a. Visualizing numerical variables using scatter or pairplots in order to interpret business /domain inferences.

b. Visualizing categorical variables using barplots or boxplots in order to interpret business/domain inferences.

3. Data Preparation
Converting categorical variables with varying degrees of levels into dummy variables (numerical in nature) so that these variables can be represented during model building in order to contribute to the best fitted line for the purpose of better prediction.

4. Splitting the data into training and test sets
Splitting the data into two sections in order to train a subset of dataset to generate a trained (fitted) line that will very well generalize how new and unknown data (test set or new dataset) will be evaluated, and how the fitted line will be able to accurately estimate new or unknown datasets. Generally, the train-test split ratio is 70:30 or 80:20.
Rescaling the trained model: It is a method used to normalize the range of numerical variables with varying degrees of magnitude.

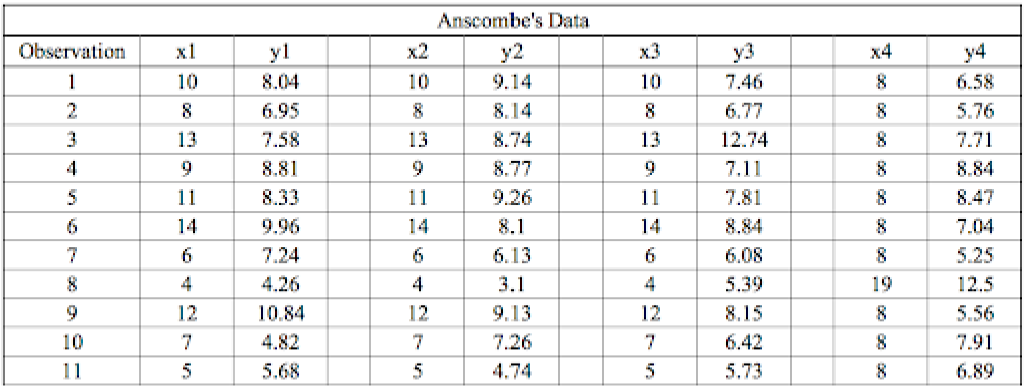
5. Building a linear model
Either using manual approach (forward or backward selection) , we finalize the important features that are required to build the model with VIF. In case of automated approach, we use RFE to select the variables.

6. Residual analysis of the train data:
It tells us how much the errors (y_actual — y_pred) are distributed across the model. A good residual analysis will signify that the mean is centred around 0.

7. Making predictions using the final model and evaluation
We will predict the test dataset by transforming it onto the trained dataset. Divide the test sets into X_test and y_test and calculate r2_score of test set.

Question 2 : Explain the Anscombe's quartet in detail.

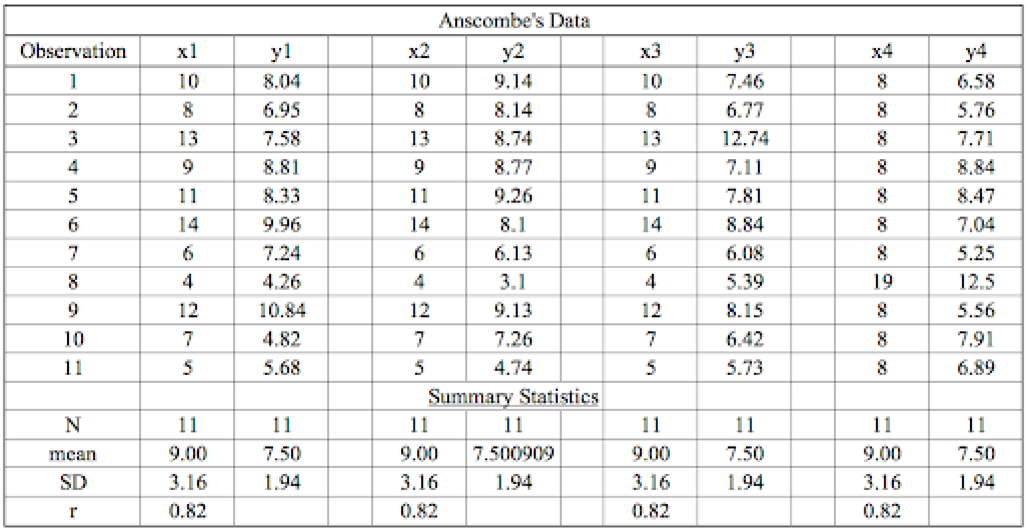
Answer : The **Anscombe's quartet** was constructed to highlight the importance of data visualization to analyse the data before concluding. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.
We can define these four plots as follows:



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

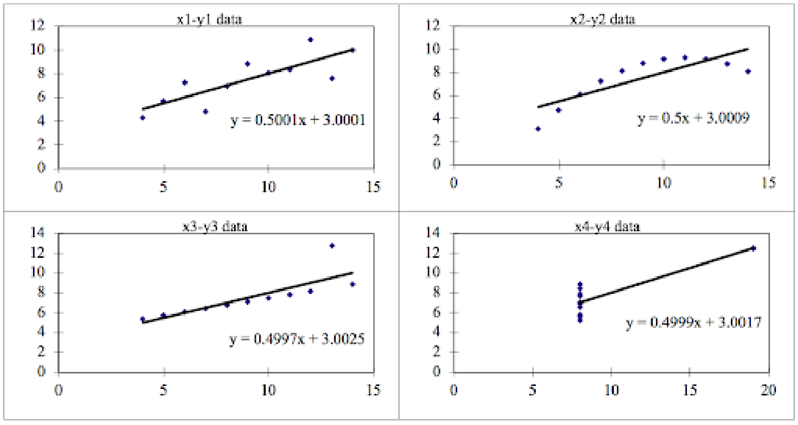
The statistical information for these four data sets are approximately similar. We can compute them as follows:



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, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



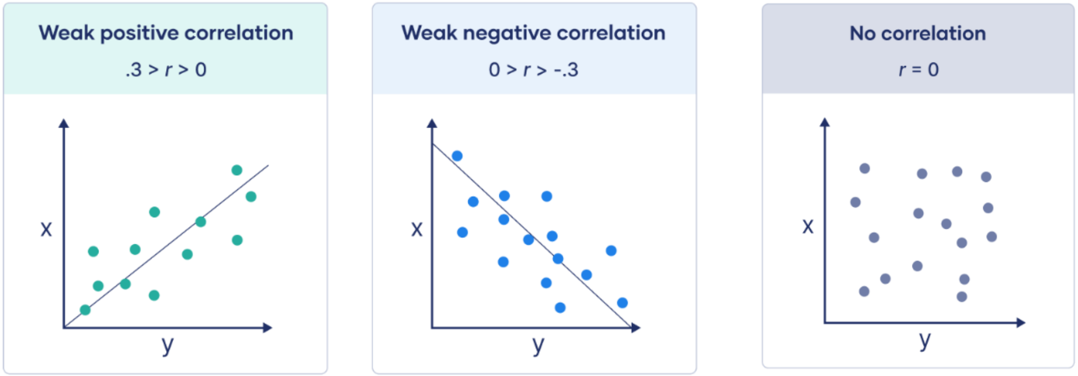
We can describe the four data sets as:

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.

Question 3 : What is Pearson's R?

Answer : In statistics, the Pearson correlation coefficient — also known as Pearson's r, is a measure of linear correlation between two sets of data. It describes the strength and direction of the linear relationship between two quantitative variables. It takes value between -1 to 1.
The greater than zero value denotes that there is positive correlation between two sets of data i.e. if increase in one quantity will cause increase in another quantity.

Less than zero value denotes that there is negative correlation between two sets of data i.e. if increase in one quantity will cause decrease in another quantity.



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

Answer :
Scaling : In multiple Linear regression, feature scaling is a method which is used to normalize the data in same range. This step is important because various columns has various range of values.

Example- if in your model, you have 2 independent variables called as **age** and **income** , then range of values can be in different range. Age can take value from 0 to 100 and income can take value from 100000 to 2500000. In this case, the coefficient will have very different value, the co-efficient of age will have very high value which will make age as more important feature and co-efficient of income can have very low value which will make income as less important feature and model may give incorrect score.

With scaling we bring both variables in same range so that co-efficient can have values in same range and model will give better result

There are two major methods to scale the variables, i.e. standardisation and MinMax scaling. Standardisation basically brings all of the data into a standard normal distribution with mean zero and standard deviation one. MinMax scaling, on the other hand, brings all of the data in the range of 0 and 1.
Standardisation: x = x – mean(x)/ sd(x)

MinMax scaling x = x- min(x) / max(x) – min(x)

Standardised scaling will affect the values of dummy variables but MinMax scaling will not.

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

Answer : VIF , Variance Inflation factor is metric which is used to denote correlation between two independent variables. If one independent variable can be explained by another independent variable, such situation is called as co-linearity. Higher the correlation, higher is VIF.

If there is 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 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 multi-collinearity.

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).

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

Answer: Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

The power of Q-Q plots lies in their ability to summarize any distribution visually. QQ plots is very useful to determine following

- If two populations are of the same distribution If residuals follow a normal distribution.
- Having a normal error term is an assumption in regression and we can verify if it’s met using this.
- Skewness of distribution

In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.