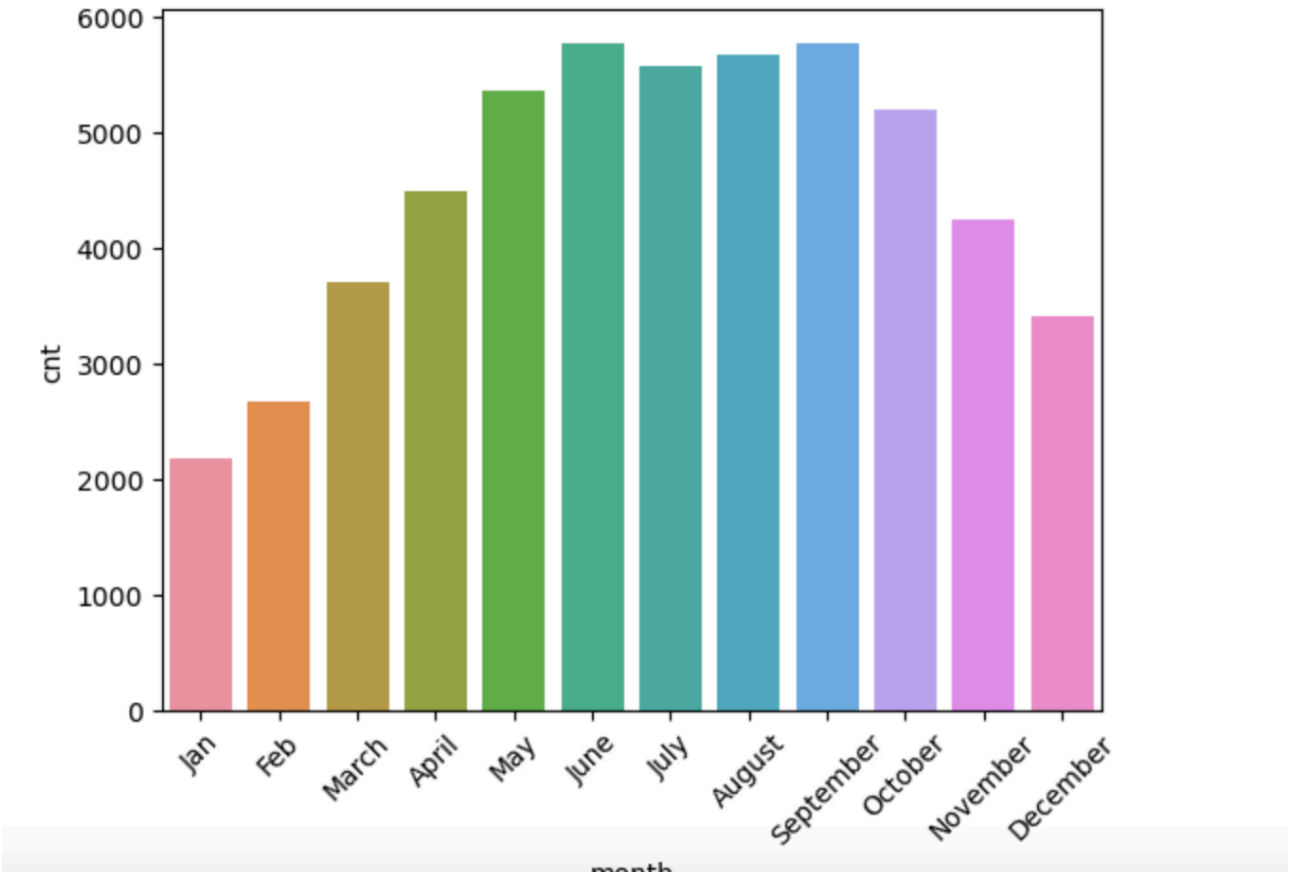
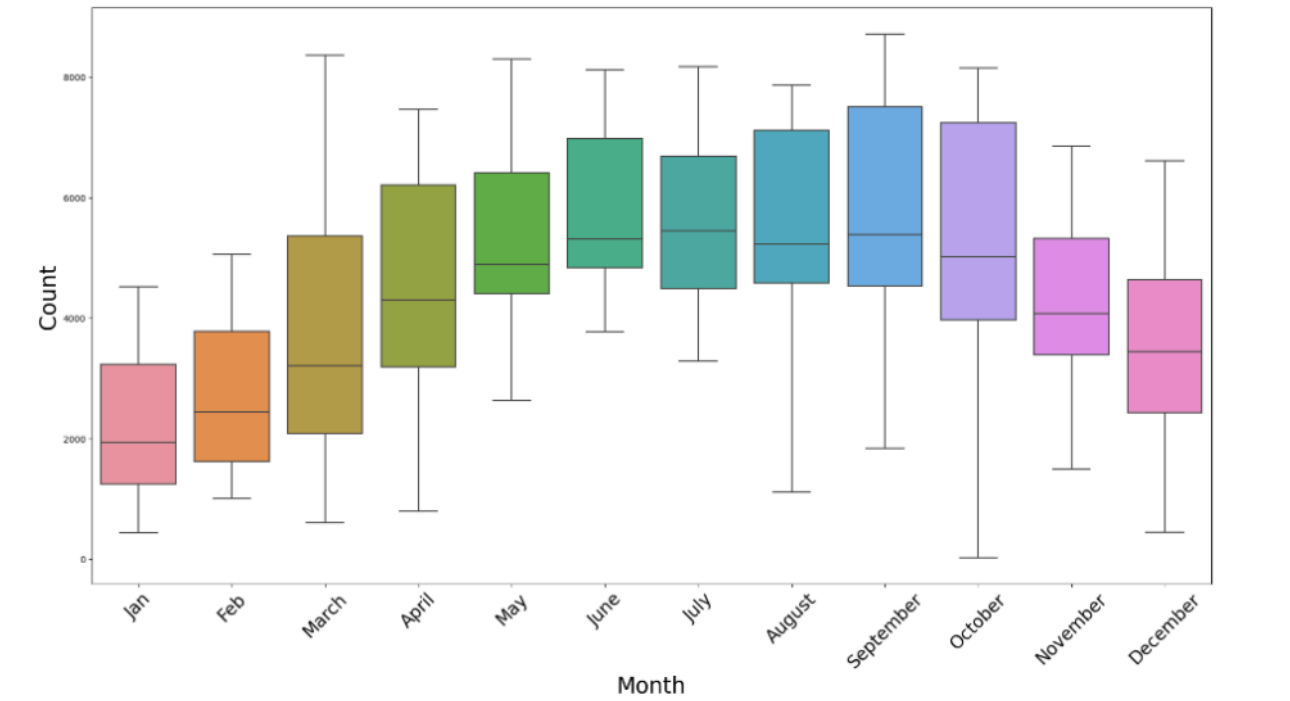
Assignment-based Subjective Questions

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

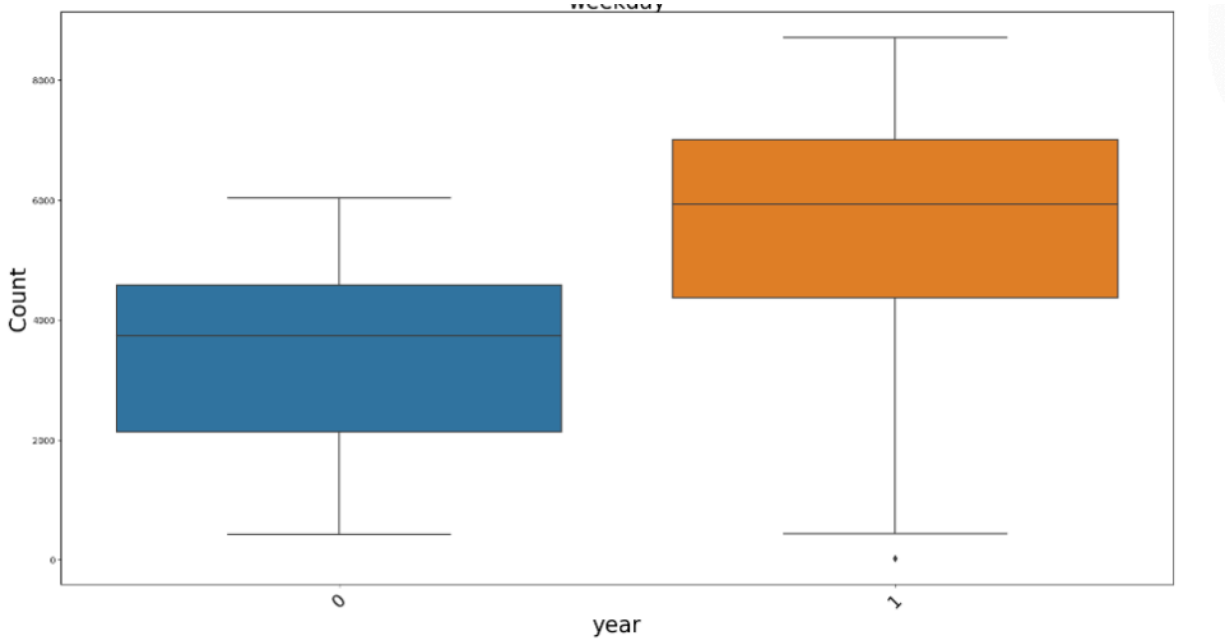
Ans . Based upon the analysis of the categorical variables like month ,season ,weather , weekdays had some impact on the dependent variable cut as we can seen as below :

1) June and September month had the highest count and the same can be seen on the model as September had a positive impact of 0.0730 the variable cnt which means sales cnt increase in September. Also December , January and November had a negative impact on the bike sharing rental count which can also be seen using the plots below.

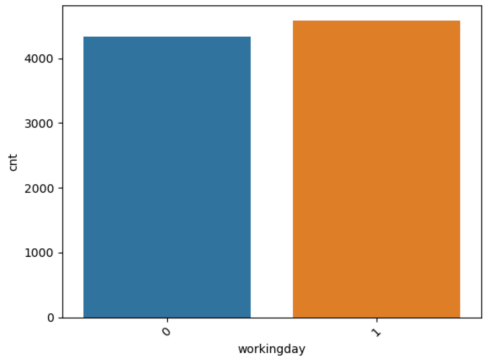




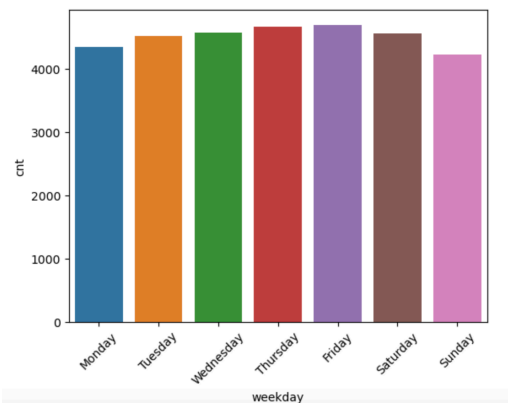
2) Year had a positive impact on the cnt variable which means sales increase with increase in year



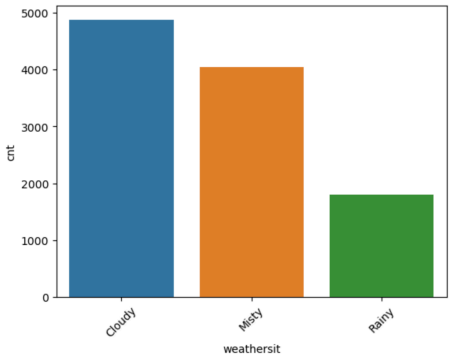
3) Working day had a positive impact which means if it is a working day then cnt will increase.



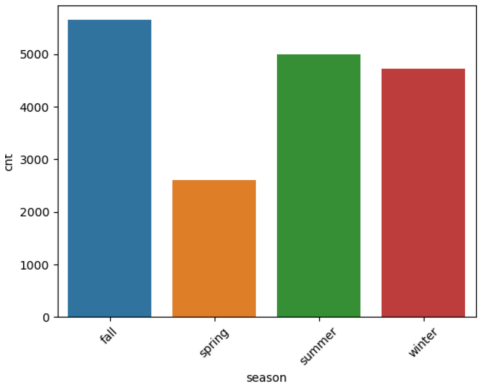
4) Saturday weekday had a positive impact on the cnt variable
Which means cnt will be more on Saturday.



5) Misty weather and Rainy weather has a negative impact on the cnt .



6) Spring season had a negative impact on the cnt.



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

Ans : It is important to use drop_first=True in the dummy variable creation because it reduces the extra redundant column since the value of the first column can be found out using the other variables. This helps in reducing the multicollinearity amount the variables.

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

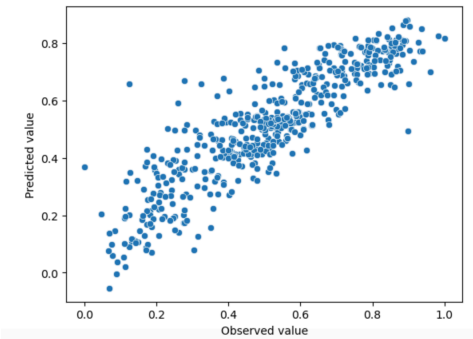
The variable temp and atemp has the highest correlation with the target variable.

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

Ans :

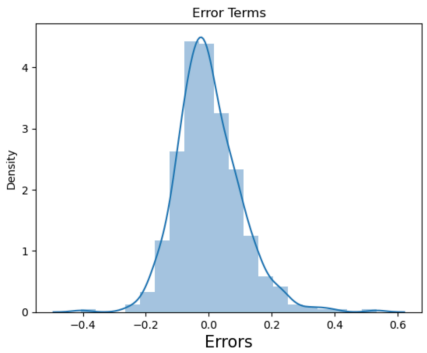
Assumption 1 : Linear Relationship

To test Linear Relationship we need to draw a plot of y_predicted vs y_observed value . Here we can see the relationship is linear



Assumption 2 : Error Terms is Normally Distributed

To Verify this we have created a distribution plot which is normally distribution with mean at 0.



Assumption 3 : Mean of Error terms is Zero

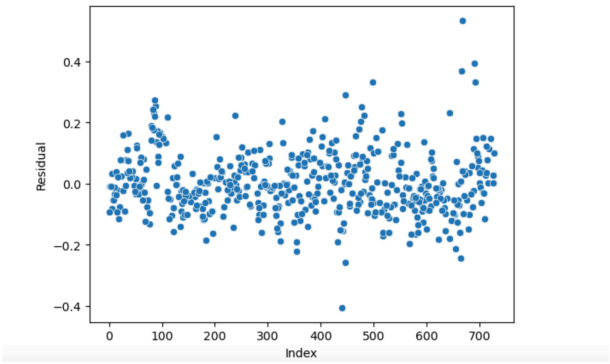
We have proved this assumption by using the mean function on error terms

```
In [259]: residual = y_train_cnt-y_train
          residual.mean()

Out[259]: -3.030255784859251e-16
```

Assumption 4 : Error terms have a constant variance

This we have proved by plotting a scatter plot of residuals vs index which shows a horizontal line distribution with constant variance .



Assumption 5 : No autocorrelation among the error terms - DW test

This we have proved by doing a GW test on the residuals which has a value of around 2 , hence there is no autocorrelation between error terms.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans : Rainy weather , yr and spring season are top 3 features that significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans : Linear Regression is a type of supervised model in machine learning whose aim is to find out best fit line between target and input variables**.**Linear regression algorithm is used to find out a linear relationship between a target variable and one or more predictor or input variables. By using linear regression we can get the value of predicted value of target variable by using input variables.

There are 2 types of linear regression

1. Simple linear Regression
2. Multiple linear regression

1. **Simple Linear Regression** is where we have one independent or predictor variable and one dependent or target variable.

Simple Linear Regression is denoted by following equation

$Y = b0 + b1x$

Where b0 is the intercept and b1 is the coefficient or slope of the best line

.

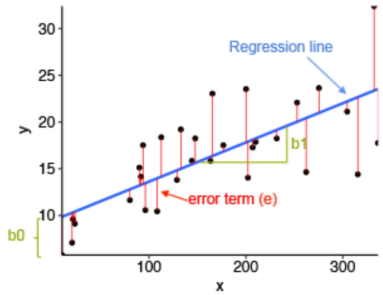
2. **Multiple Linear Regression** is where we have one dependent variable and more than one independent variable.
It is denoted by

y= b0+b1x1+b2x2+b3x3+b4x4..
Where b0,b1,b2,b3,b4.. are the coefficients of variables x1,x2,x3,x4.. respectively and b0 is the intercept.

A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.

Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.

Let's understand this with the help of a diagram.



In the above diagram,

- x is our dependent variable which is plotted on the x-axis and y is the dependent variable which is plotted on the y-axis.
- Black dots are the data points i.e the actual values.
- bo is the intercept which is 10 and b1 is the slope of the x variable.
- The blue line is the best fit line predicted by the model i.e the predicted values lie on the blue line.

The vertical distance between the data point and the regression line is known as error or residual. Each data point has one residual and the sum of all the differences is known as **the Sum of Residuals/Errors.**

Mathematical Approach:

Residual/Error = Actual values – Predicted Values

Sum of Residuals/Errors = Sum(Actual- Predicted Values)

Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))2

2) Explain the Anscombe's quartet in detail.

Anscombe's Quartet can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset

that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.
It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the **importance** of **plotting the graphs** before analyzing and model building, and the effect of other **observations on statistical properties**.There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets.
This tells us about the importance of visualising 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 datasets. The statistical information for all these four datasets are approximately similar and can be computed 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	

Image by Author

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:

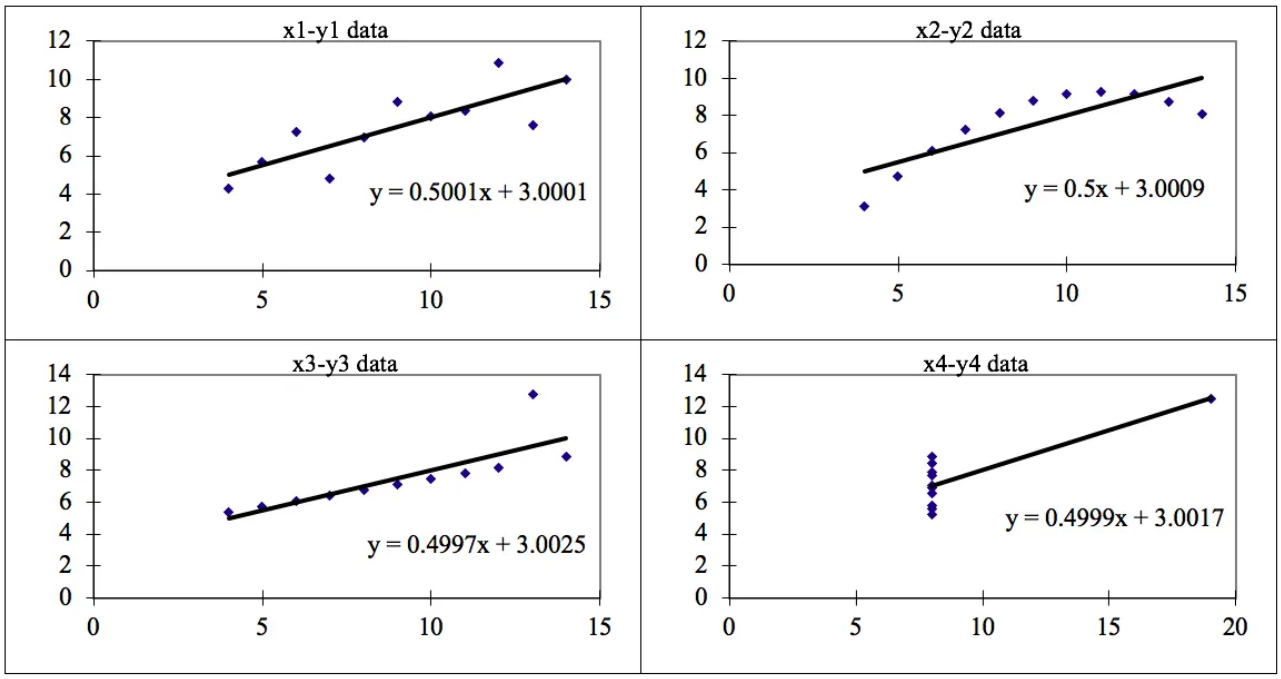


Image by Author

The four datasets can be described as:

- **Dataset 1:** this **fits** the linear regression model pretty well.
- **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
- **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
- **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

3. What is Pearson's R?

Pearson relation coefficient R is used to measure the relationship between the variables. It measures the strength of the relationship between 2 continuous variables.The relationship can be positive or neagtive. The coefficient not only measure the presence of positive or negative relationship , it also tells the extent to which the variables are related to each other.It is independent of unit of measurement and have the range of values between -1 and +1 where -1 denotes the prefect negative relationship between two variables and +1 denotes perfect positive relationship between 2 variables.

Formula for Pearson Coefficient is

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

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

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

Ans : Scaling is a method which is used to standardise the range of independent variables or features of data. In data processing, it is also known as data normalization or standardization.

Feature scaling is generally performed during the data pre-processing stage, before training models using machine learning algorithms. The goal is to transform the data so that each feature is in the same range (e.g. between -1 and 1). This ensures that no single feature dominates the others, and makes training and tuning quicker and more effective. Feature scaling can be accomplished using a variety of linear and non-linear methods, including min-max scaling, z-score standardization, clipping,, taking logarithm of inputs before scaling, etc

Normalised scaling : In normalised scaling , the features are transformed on a a similar scale . The new point is calculated as:

```
X_new = (X - X_min)/(X_max - X_min)
```

This scales the range to [0, 1] or sometimes [-1, 1].Normalization is useful when there are no outliers as it cannot cope up with them. Usually, we would scale age and not incomes because only a few people have high incomes but the age is close to uniform.

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

```
X_new = (X - mean)/Std
```

Standardization can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Geometrically speaking, it translates the data to the mean vector of original data to the origin and squishes or expands the points if std is 1 respectively. We can see that we are just changing mean and standard deviation to a standard normal distribution which is still normal thus the shape of the distribution is not affected.

Standardization does not get affected by outliers because there is no predefined range of transformed features.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: The VIF value of some variables were infinite which meant that there was perfect collinearity considering the variable as dependent variable and all other variables as independent variable.

This can be proved by the formula of VIF = 1/1-R2

If the variable is perfectly collinearated then R2 will be 1 hence VIF = 1/1-1 = infinite

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

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.
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:
a) It can be used with sample sizes also
b) 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 following scenarios:
If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. 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.
a) ***Similar distribution****: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis*
b) ***Y-values < X-values:*** *If y-quantiles are lower than the x-quantiles.*

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

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