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?**

Answer – a. Year categorical variable is one of the important variables which affect Bike rental. I found that year 2019 has more rentals than 2018. I think it is because Bike sharing is becoming popular over time due to many reasons.

b. People rent more on working days than holidays and weekends. So the working day variable is also an important variable.

c. Weather is also important to factor in as if weather is clear then there are more rentals of bikes compared to bad weather conditions.

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

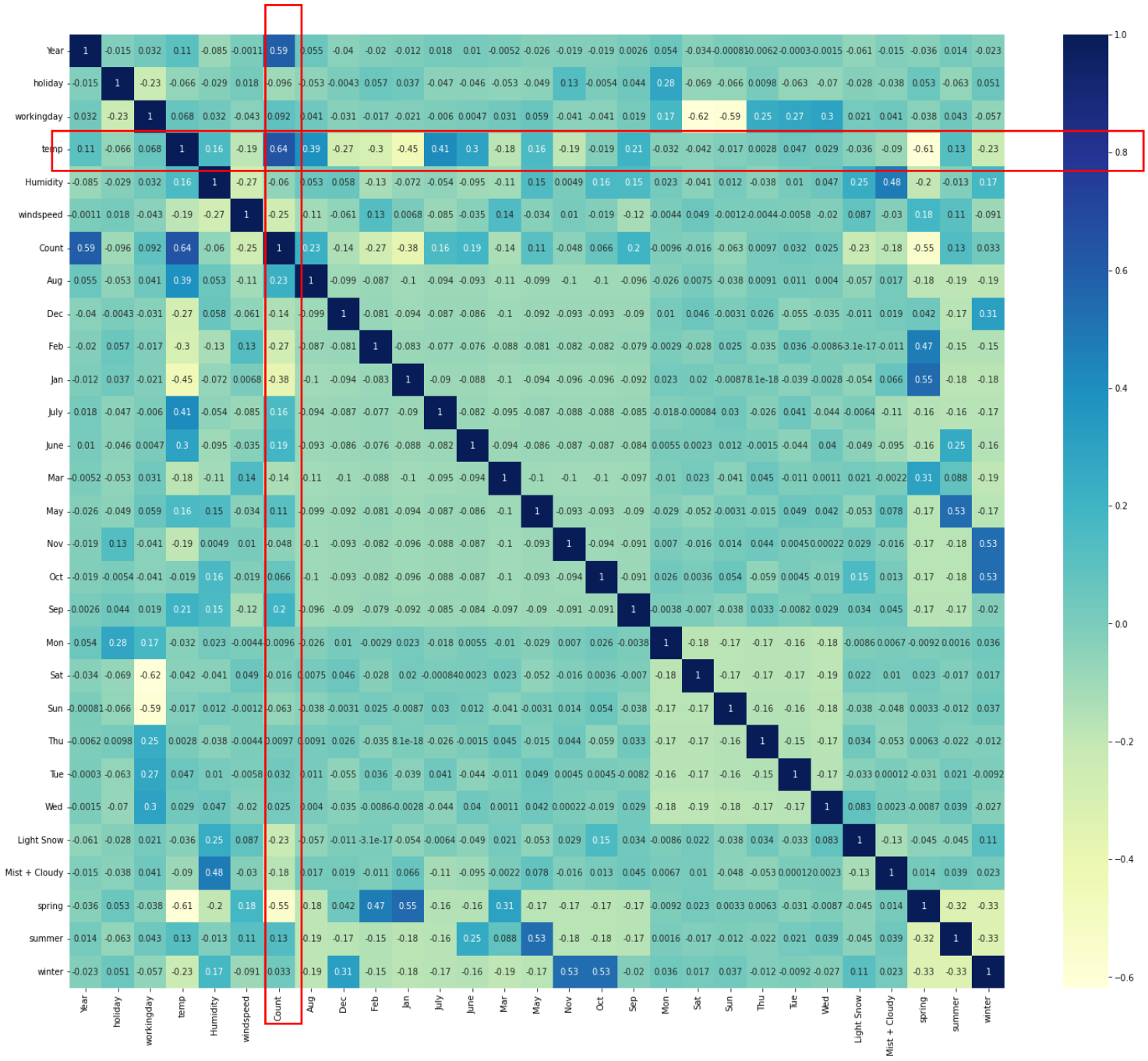
Answer – It is important because it will reduce the extra column created by the dummy variable. For example, if you have a variable gender, you don't need both a male and female dummy. Just one will be fine. If male=1 then the person is a male and if male=0 then the person is female. But, if you have a category with hundreds of values, I suggest not dropping the first column. That will make it easier for the model to "see" all the categories quickly during learning (and the adverse effects are negligible).

It also depends on the model. If you don't drop the first column then your dummy variables will be correlated. Hence it reduces the correlations created among dummy variables. Hence if we have a 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?**

Answer- By looking at the pair plot temp variable has the highest (0.64) correlation with the target variable 'cnt'.

Below is the heatmap from the Bike-sharing assignment.



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

Answer – a. Residual Analysis of the train data - The residual errors should be normally distributed.

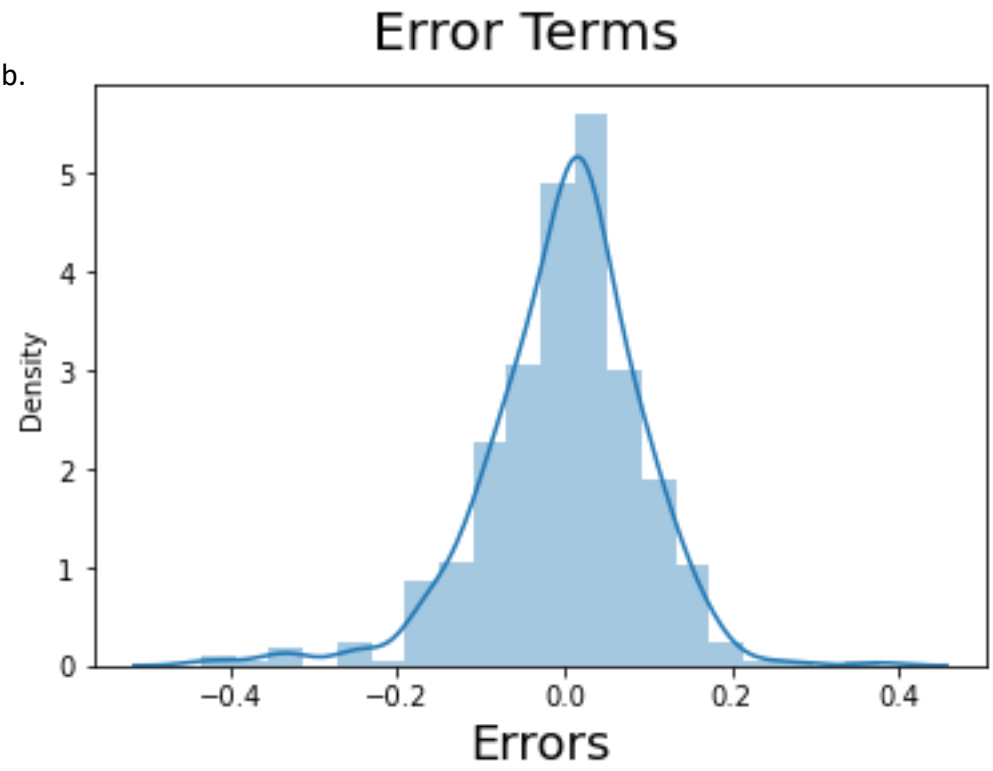
Below I attached a histogram of Error terms. We can see that error terms are normally distributed.

b. Residual errors should be homoscedastic: The residual errors should have constant variance.

c. The training and testing accuracy are nearly equal hence there is no Overfit/Underfit situation.

d. The predicted values have a linear relationship with the actual values.

b.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes?

Answer – a. Temperature could be a prime factor for deciding for the company.

b. We can see demand for bikes was more in 2019 than in 2018 so the year is also an important variable.

c. Working days as they have a good influence on bike rentals. So it would be great to provide offers to the working individuals.

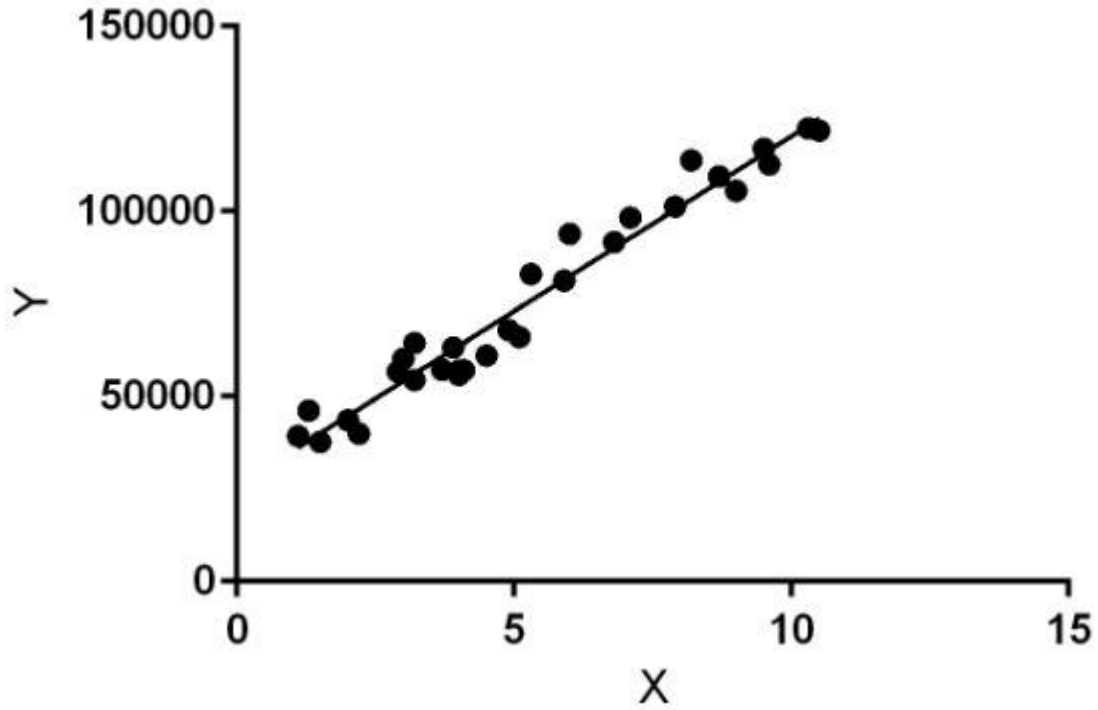
The top features contributing significantly towards explaining the demand of the shared bikes are: -

- temperature
- year
- working day

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Answer- **Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables i.e. between independent and dependent variables. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.
If one independent variable is used then it is **Simple linear regression.**
If more than one independent variable is used then it is **Multiple linear regression.**



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis function for Linear Regression :

$$y = \theta_1 + \theta_2.x$$

While training the model we are given:
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.
θ1: intercept
θ2: coefficient of x
Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our prediction model, it will predict the value of y for the input value of x.
How to update θ1 and θ2 values to get the best fit line?
Cost Function (J):
By achieving the best-fit regression line, the model aims to predict the y value such that the error difference between the predicted value and the true value is minimum. So, it is very important to update the θ1 and θ2 values, to reach the best value that minimizes the error between the predicted y value (pred) and the true y value (y).

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).
Gradient Descent:
To update θ1 and θ2 values to reduce Cost function (minimizing RMSE value) and achieve the best fit line the model uses Gradient Descent. The idea is to start with random θ1 and θ2 values and then iteratively update the values, reaching minimum cost.

$$\begin{aligned}
&\text{repeat until convergence: } \{ \\
&\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_0^{(i)} \\
&\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_1^{(i)} \\
&\theta_2 := \theta_2 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_2^{(i)} \\
&\dots \\
&\}
\end{aligned}$$

2. Explain Anscombe's quartet in detail.

Answer – In Anscombe's quartet, four data sets have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Some peculiarities fool the regression model once you plot each data set.

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

Purpose of Anscombe's quartet-

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. 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.

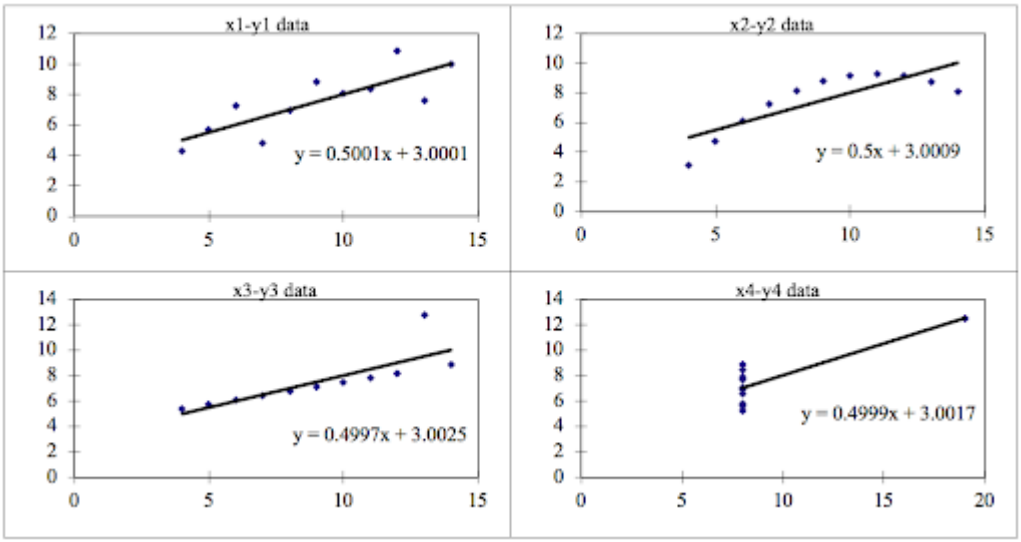
We can define these four plots as follows:

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 is 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

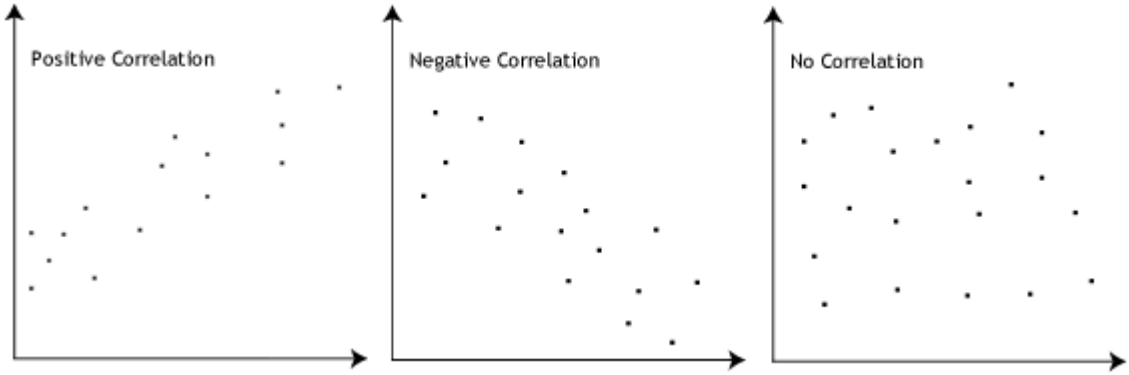
As you can see, Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set to help build a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

3. What is Pearson's R?

Answer - In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association



Pearson r Formula

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

Where 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?

Answer –Scaling is a general term that means changing the range of each feature/variable/predictor. These methods are usually used as pre-processing steps all over the data science field. Before modeling data, they need to be pre-processed. It also helps in speeding up the calculations in an algorithm.

Most of the time, the collected data set contains features highly varying in magnitudes, units, and ranges.

If scaling is not done then the algorithm only takes magnitude into account and not units 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.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1

Scikit learn library (sklearn.preprocessing.MinMaxScaler) helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

Standardization replaces the values with their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation of one (σ).

Scikit learn library (sklearn.preprocessing.scale) helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

Answer - If there is a perfect correlation, then VIF = infinity. This only happens when the denominator has zero value.

$$VIF_i = \frac{1}{1 - R_i^2}$$

And the denominator will be zero if R2 is one. It means that there is very high collinearity.

VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

If VIF is large and multicollinearity affects your analysis results, then we can take some corrective actions before you can use multiple regression. Here are the various options:

- One approach is to review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model. For example, if inputs are measuring the weight in kgs and lbs then just keep one of these variables in the model and drop the other one. Dropping the term with a large value of VIF will hopefully, fix the VIF for the remaining terms and now all the VIF factors are within the threshold limits. If dropping one term is not enough, then you may need to drop more terms as required.
- A second approach is to use principal component analysis and determine the optimal set of principal components that best describe your independent variables. Using this approach will get rid of your multicollinearity problem but it may be hard to interpret the meaning of these “new” independent variables.
- The third approach is to increase the sample size. By adding more data points to our model, the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.
- The fourth approach is to transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.

In conclusion, when we are building a multiple regression model, always check VIF values for independent variables and determine if we need to take any corrective action before building the model.

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

Answer – Q-Q plot - The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

The key to making insightful charts and useful models is finding out the nature of the data. Most of the data we collect fits into the bell curve of a Normal Distribution.

But, The Q-Q here stands for Quantile-Quantile. This means points of the sample data are plotted against quantiles of a normal distribution. Quantiles are breakpoints that divide the numeric data into proportioned buckets.

Important things about Q-Q plot-
1. Q-Q plots can be used to compare the sample data against any kind of distribution (Poisson, Exponential, etc.)
2. They can also be used to check whether two separate samples are distributed in the same manner.
3. Q-Q plots fail to work when the data points are too few. So, use a large dataset.

Components of a Q-Q plot -
> Green line: This line denotes the normal distribution.
> Blue points: These are the numeric entries from our dataset plotted on the y-axis

Thus, if the sample data is normally distributed, it will fit seamlessly on the Q-Q plotline. If not, it is enough to know that the data is skewed.