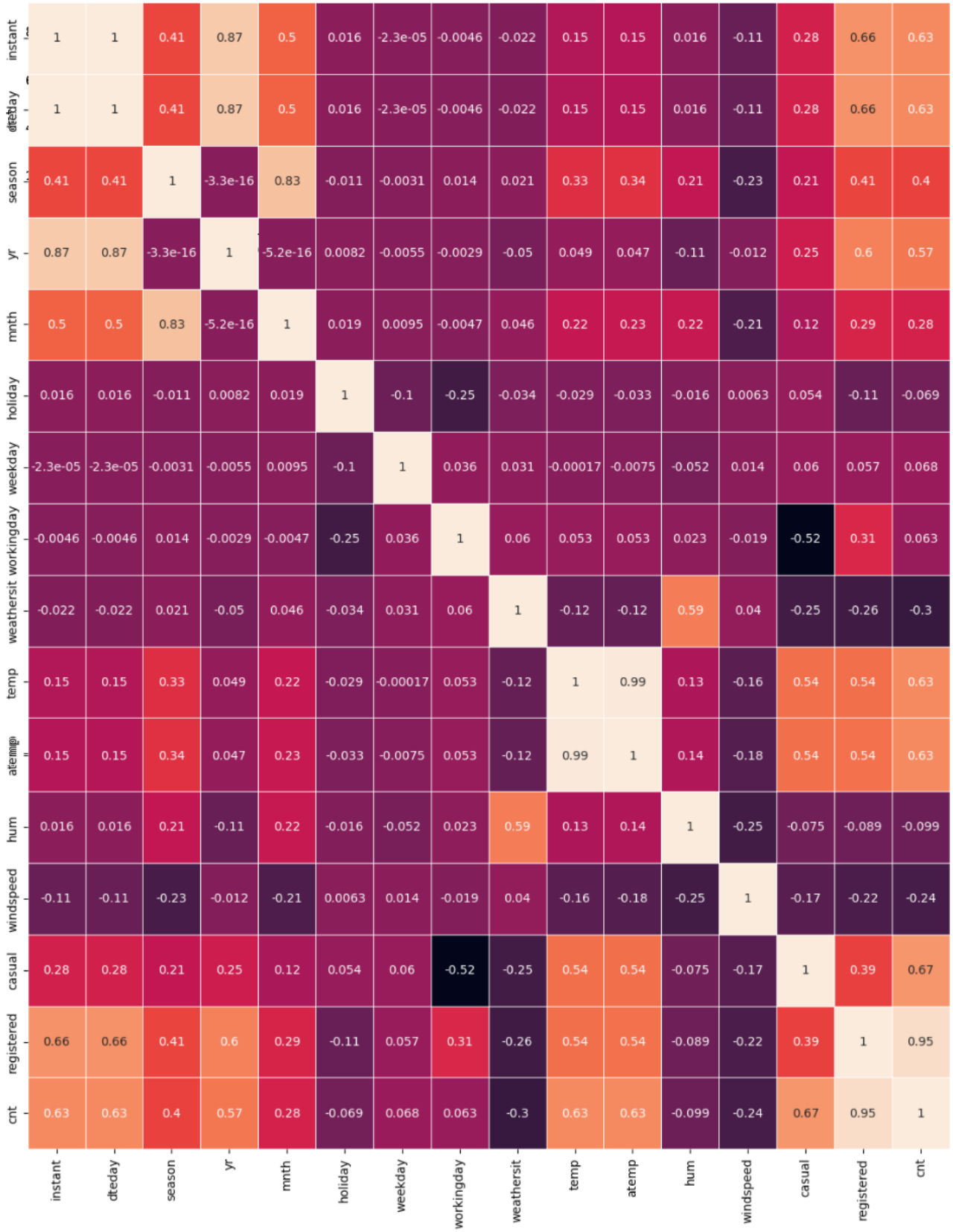
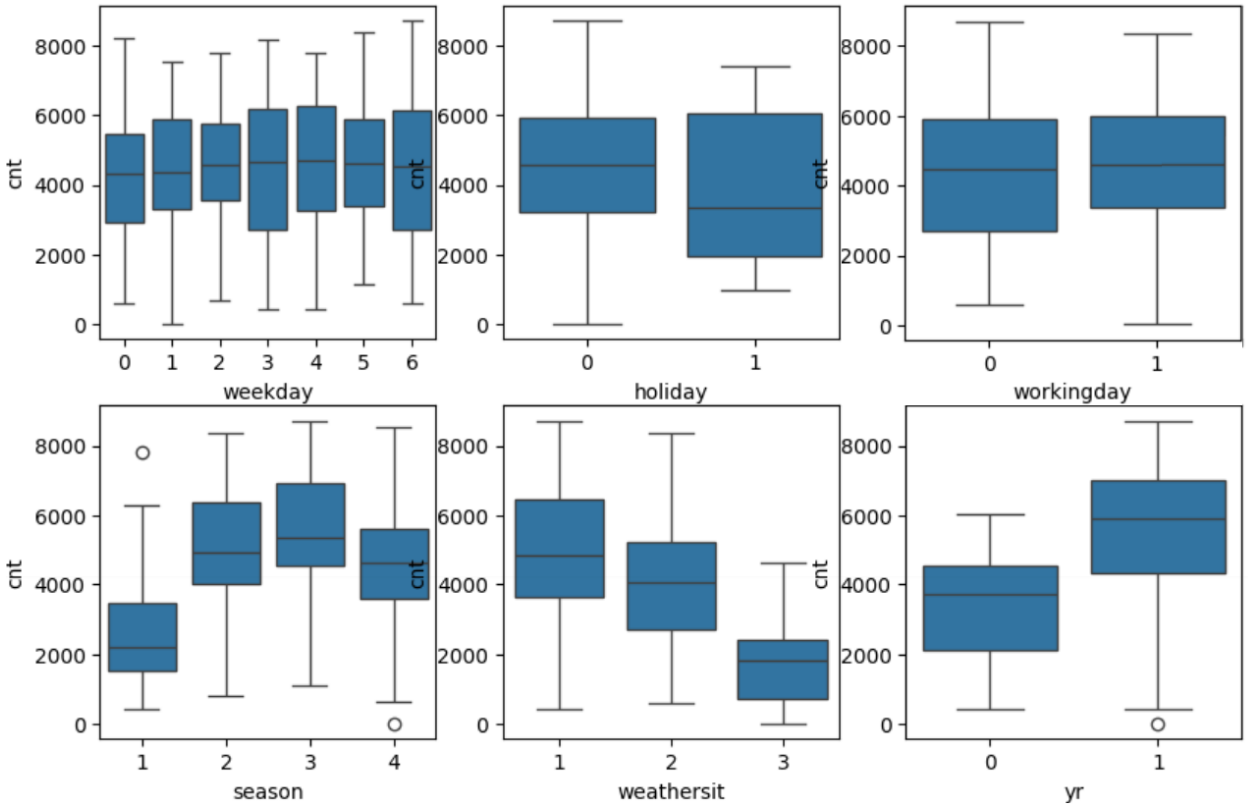
Assignment-based Subjective Questions

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



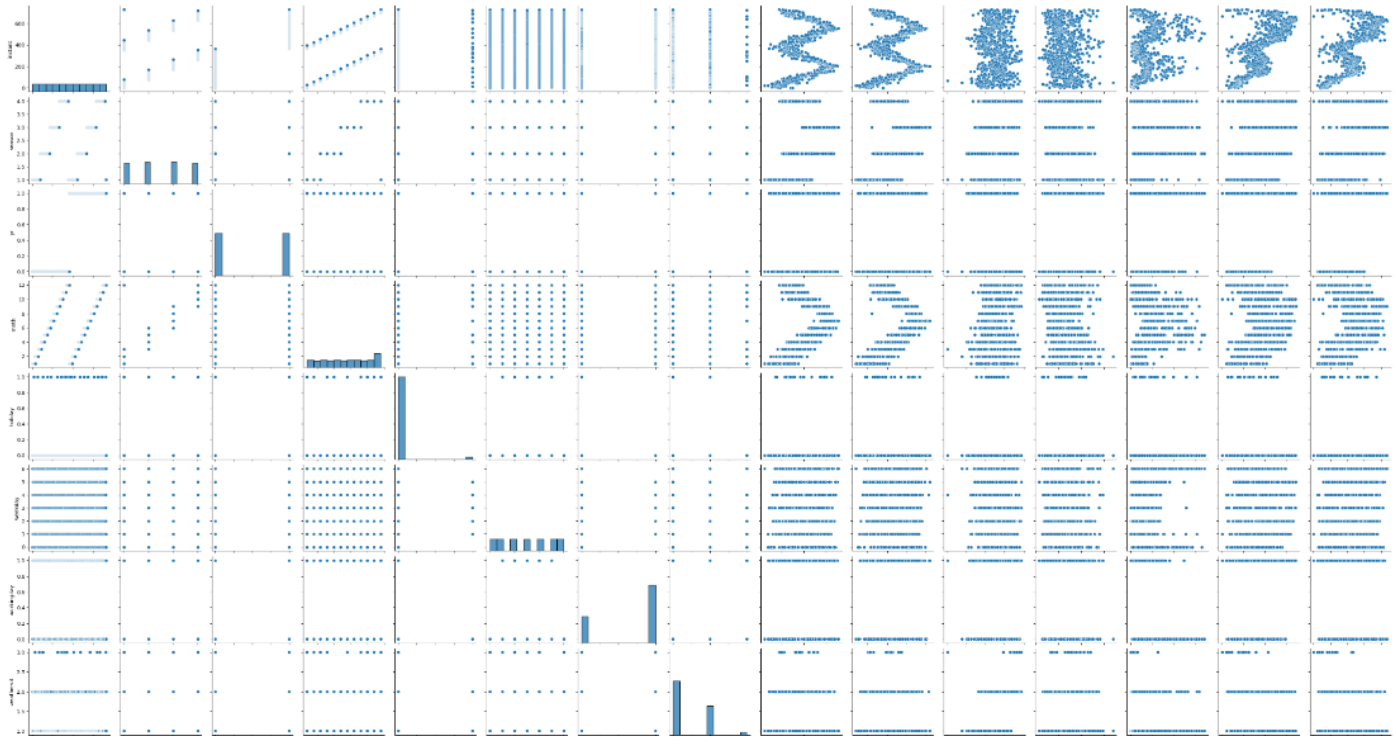


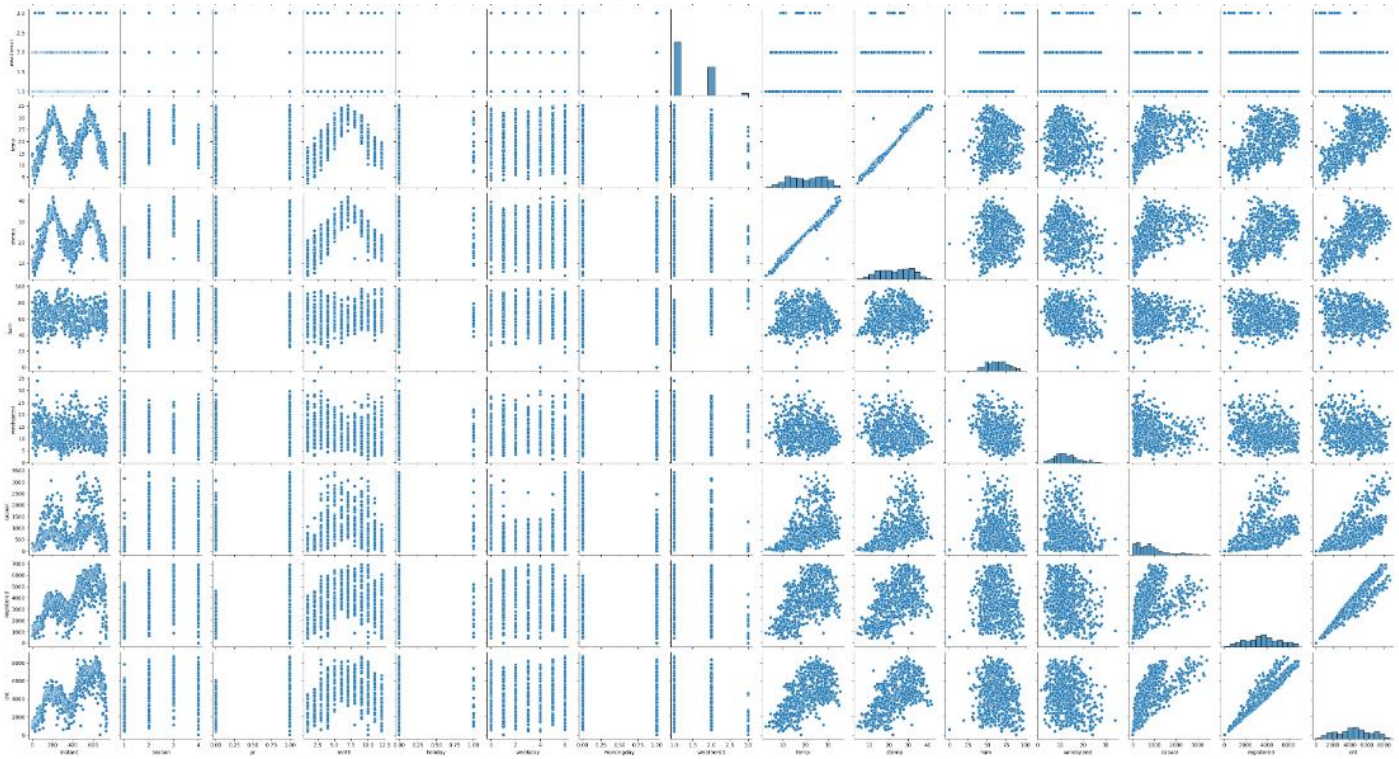
We can see that registered, casual, holiday and weekday have a good impact.

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

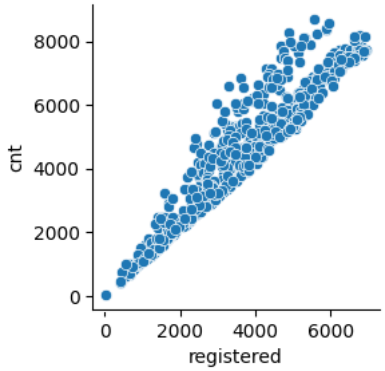
We use drop_first = True, to avoid creation of n dummy variables. If there are n dummy variables, n-1 dummy variables can help to predict the value of the nth dummy variable, so one dummy variable can be dropped to avoid redundancy and most importantly multicollinearity.

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





Upon verifying the pair plot we can see that the highest corelation with cnt is with registered



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

After splitting the data into a 70-30 ratio for training and test purpose, I evaluated the linear regression model summary on the training data and verified that the p-values are mostly zero's and also checked that the R-squared is a decent value that explains the prediction, Prob (F-statistic) is zero. Then I ran the model on the test data to compare the R2 value of training data and test data and could see that it almost matched .

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

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.970
Model:	OLS	Adj. R-squared:	0.969
Method:	Least Squares	F-statistic:	1451.
Date:	Wed, 17 Apr 2024	Prob (F-statistic):	0.00
Time:	08:38:27	Log-Likelihood:	930.46
No. Observations:	510	AIC:	-1837.
Df Residuals:	498	BIC:	-1786.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0558	0.021	2.705	0.007	0.015	0.096
holiday	-0.0459	0.011	-4.000	0.000	-0.068	-0.023
workingday	-0.1130	0.004	-26.797	0.000	-0.121	-0.105
temp	0.1269	0.016	8.121	0.000	0.096	0.158
hum	-0.0325	0.017	-1.967	0.050	-0.065	-3e-05
windspeed	-0.0418	0.011	-3.657	0.000	-0.064	-0.019
registered	0.9242	0.011	80.746	0.000	0.902	0.947
Spring	0.0038	0.009	0.441	0.659	-0.013	0.021
Summer	0.0280	0.006	4.790	0.000	0.017	0.039
Winter	-0.0013	0.007	-0.185	0.854	-0.016	0.013
Misty	0.0079	0.011	0.703	0.483	-0.014	0.030
Partly cloudy	0.0118	0.012	0.989	0.323	-0.012	0.035

Omnibus:	80.454	Durbin-Watson:	1.950
Prob(Omnibus):	0.000	Jarque-Bera (JB):	155.464
Skew:	0.897	Prob(JB):	1.74e-34
Kurtosis:	5.025	Cond. No.	27.8

Based on the above model, the top features contributing to the towards explaining the demand of the shared bikes are

1. If the day is a Holiday or not
2. If it is a working day or not
3. Temperature on that day

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear Regression is a supervised learning methodology, where the predicted output will be continuous in nature. For example, things like marks prediction, price prediction, and so on.

   Linear Regression is a fundamental statistical and machine learning technique used for modelling the relationship between a dependent variable (also known as the target or response variable) and one or more independent variables (predictors or features).

   Its objective is to establish a linear equation that best represents the association between these variables, allowing us to make predictions and draw insights from the data.

The primary aim of linear regression is to find the "best-fit" line (or hyperplane in higher dimensions) that minimizes the difference between the predicted values and the actual observed values.

This best-fit line is defined by a linear equation of the form:

$Y = b0 + b1X1 + b2X2 +...+ bnXn$

In this equation:

Y represents the dependent variable we want to predict.

X1,X2,...,Xn are the independent variables also known as features.

b0 is the intercept (the value of Y when all X values are zero).

b1,b2,...,bn are the coefficients that determine the relationship between each independent variable and the dependent variable.

Linear regression assumes that there is a linear relationship between the predictors and the target variable.

The goal of the model is to estimate the coefficients (b0,b1,...,bn) that minimize the sum of the squared differences between the predicted values and the actual values in the training data. This process is often referred to as "fitting the model."

2. Explain the Anscombe's quartet in detail.

   Anscombe's Quartet is the modal 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 analysing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyse 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 behaviour irrespective of statistical analysis.

   - Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
   - Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
   - Data-set III — looks like a tight linear relationship between *x* and *y*, except for one large outlier.
   - Data-set IV — looks like the value of *x* remains constant, except for one outlier as well.

   The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

   Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

   Anscombe's Quartet emphasizes the significance of exploratory data analysis (EDA). By thoroughly examining our data, conducting descriptive statistics, and visualizing relationships, we can find hidden insights and avoid making oversimplified conclusions.

3. What is Pearson's R?

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

The Pearson correlation coefficient (r) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when all of the following are true:

1. Both variables are quantitative: You will need to use a different method if either of the variables is qualitative.
2. The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
3. The data have no outliers: Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
4. The relationship is linear: "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

Below is a formula for calculating the Pearson correlation coefficient (*r*):

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

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

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. 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 units hence incorrect modelling. 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.

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

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

- 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.

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

   VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.
   If all the independent variables are dependent on each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.
   If VIF is large and multicollinearity affects your analysis results, then you need to take some corrective actions before you can use multiple regression.
   VIF can be calculated by the formula below:

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

   Where **Ri2** represents the unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones. The reciprocal of VIF is known as **tolerance**. Either VIF or tolerance can be used to detect multicollinearity, depending on personal preference.
   If Ri2 is equal to 0, the variance of the remaining independent variables cannot be predicted from the ith independent variable. Therefore, when VIF or tolerance is equal to 1, the ith independent variable is not correlated to the remaining ones, which means multicollinearity does not exist in this regression model. In this case, the variance of the ith regression coefficient is not inflated.

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.
   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.
   A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.