Answers to the Assignment-based questions

1. From my analysis, I could conclude the following points:
   a) There are 7 categorical variables in the dataset, namely season, yr, mnth, holiday, weekday, workingday and weathersit.
   b) The bike demand is highest during the Fall and lowest during the Spring season.
   c) There is a significant increase in the bike usage in the year 2019 as compared to the year 2018.
   d) The bike demand pattern increases till September and then it starts declining till the year end.
   e) In the holiday box plot it can be seen that the median of the 'holiday=yes' plot is lesser than the median of the 'holiday=no' plot. So, apparently during holidays, the demand is lesser.
   f) There is not much difference in bike demand during all days of the week and whether it is working day or not.
   g) The demand is the highest during Clear weather situation and significantly low during light snow. There is no bike demand during heavy rains.

2. It is important to use drop_first=True during dummy variable creation as we for a feature that has 'n' categories we need only 'n-1' variables to describe it completely and it is also desirable to have the number of columns as less as possible.
   For example, for season variable, there are 4 categories but we need only 3 dummy variables. When the value of a dummy variable is 1, the corresponding category is selected and when all dummy variables have values 0, then it represents the 4th category that was dropped.

3. Among the numerical variables, temp and atemp seem to have the highest correlation with the target variable cnt.

4. I had validated the assumptions of Linear Regression in the following ways:
   a) Assumption 1 - There is linear relationship between X and y:
   It could be seen from the scatter plots during Bivariate analysis that the numerical variables seem to have linear relationship with cnt.
   b) Assumption 2 – Error terms are normally distributed:
   In the Residual analysis step, I had plotted the distribution plot of the residuals. In the plot I could see that the errors are normally distributed, centred at 0.

c) Assumption 3 – Error terms are independent of each other:
Sorry, I didn't conclude this in my analysis as I don't know how to.

d) Assumption 4 – Error terms have constant variance (Homoscedasticity):
From the normal distribution plotted in the Residual analysis, homoscedasticity property can be seen.

5. Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are:
temp, Light snow, yr.

Answers to General Subjective questions

1. Linear regression algorithm is a popular statistical technique used for predictive modelling. The first thing it assumes is a linear relationship between the dependent variable (target) and one or more independent variables (predictors/features).

Following are some important assumptions in Linear regression.
- There is linear relationship between X and y.
- Error terms are normally distributed.
- Error terms are independent of each other.
- Error terms have constant variance (Homoscedasticity).

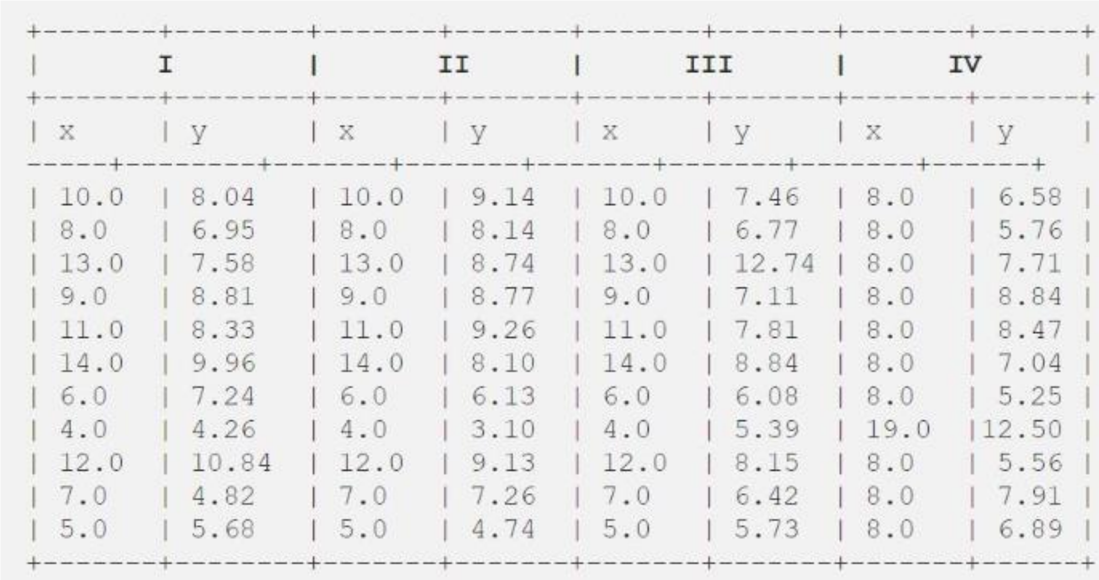
Steps in Linear Regression:
- Data Collection: Gather data on the dependent and independent variables.
- Data Cleaning and Preparation: Handle missing values, outliers, and scale or transform variables if necessary.
- Model Selection: Choose between simple or multiple linear regression based on the number of independent variables and their relationships with the dependent variable.
- Model Training: Includes Parameter Estimation (OLS used) and Cost Function.
- Model Evaluation: Includes R-squared, Adjusted R-squared, Residual Analysis.
- Prediction: Use the trained model to make predictions on new data by plugging in values of independent variables.

Types of Linear Regression:

- Simple Linear Regression -> Involves only one independent variable.
Equation can be given like this -> $y=\beta_0+\beta_1 \cdot x+\epsilon$

- Multiple Linear Regression -> Involves more than one independent variable. Equation can be given like this -> $y=\beta_0+\beta_1\cdot x_1+\beta_2\cdot x_2+\ldots+\beta_p\cdot x_p+\epsilon$

2. Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, yet appear very different when graphically represented. This set of data was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and the potential pitfalls of relying solely 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.
   The 4 datasets of Anscombe's quartet are:



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

   These four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.
   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.

3. Pearson's R, often referred to as Pearson correlation coefficient, is a measure of the strength and direction of the linear relationship between two variables. It quantifies the degree to which a pair of variables are linearly related.

Definition:
Pearson's correlation coefficient R between two variables X and Y is defined as the covariance of X and Y divided by the product of their standard deviations. Mathematically, it is expressed as:

$$R = \frac{\text{cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

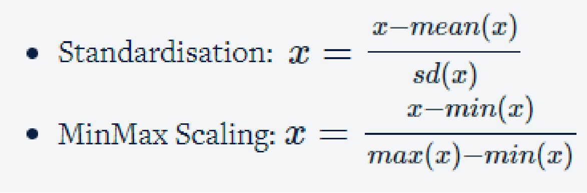
where,
cov(X,Y) is the covariance of X and Y
σ_X and σ_Y are the standard deviations of X and Y, respectively.

Properties:
- Pearson's R ranges from -1 to +1.
- If the sign of R is positive, X and Y tend to increase together. And if the sign is negative, X tends to decrease as Y increases (and vice versa).
- If |R| is close to 1, it indicates strong linear relationship. And if |R| is close to 0, it indicates weak or no linear relationship.

4. Scaling or Feature scaling is a method used to normalize the range of the independent variables or the features.

   Scaling is done so that no feature can unduly influence the variation in the target variable. The features can be of various ranges, so to make the variables fit in a similar range, scaling is done.

   There are two major methods to scale the variables, i.e. Standardisation and Normalization (Min-Max scaling). Standardisation basically brings all of the data into a standard normal distribution with mean zero and standard deviation one. Min-Max scaling, on the other hand, brings all of the data in the range of 0 and 1. The formulae in the background used for each of these methods are as given below:



- Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
- MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

5. VIF formula is given as:
   $1 / (1 – R_i^2)$

It gives a basic quantitative idea about how much the feature/independent variables are correlated with each other.
The value of VIF is infinite when the denominator is 0, i.e., R_i^2 value is equal to 1. This means that there is a perfect (highest possible) correlation between the independent variables.

6. 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

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