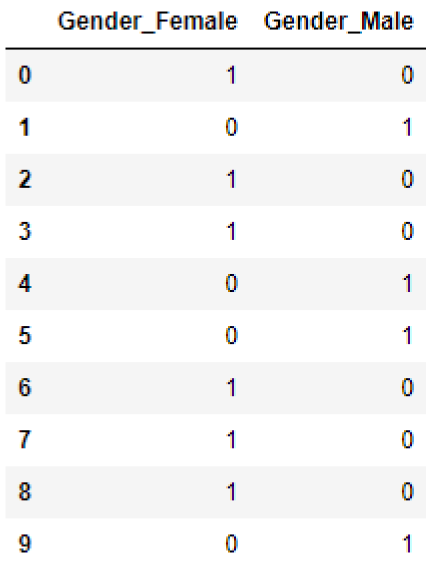
Assignment-based Subjective Questions

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

   For the season categorical variable it was observed that Summer and Winter had a positive impact on the number of bike rides. For the weathersit categorical variable, bike rides decreased in Cloudy and during LightRain weather conditions.

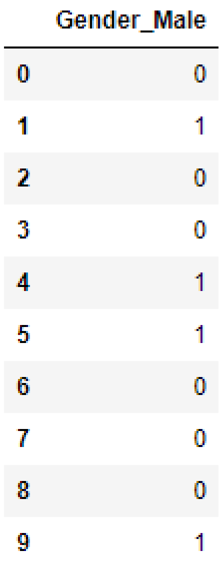
2. **Why is it important to use drop_first=True during dummy variable creation?**

   It is important to use drop_first = True because the function pd.get_dummies() returns N number of variables where N is the number of categories in the categorical variable. It is possible to decode the category name using only N-1 variables and that's exactly what drop_first = True does. Consider an example where we have a gender column where the values can be Female or Male. If we pass the gender column to pd.get_dummies() we will get something like below.



	Gender_Female	Gender_Male
0	1	0
1	0	1
2	1	0
3	1	0
4	0	1
5	0	1
6	1	0
7	1	0
8	1	0
9	0	1

On the left side, pd.get_dummies() returns two columns one for Female and one for Male. However, it is possible to categorize with using only 1 column (let's take Gender_Male) where a value of 1 signifies that the gender is Male and 0 signifies that the gender is Female. Using drop_first=True, we get exactly that, which can be seen on the right side image.



	Gender_Male
0	0
1	1
2	0
3	0
4	1
5	1
6	0
7	0
8	0
9	1

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

   The columns Temp and ATemp have the highest correlation with Cnt (target variable).

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

   After making the predictions on the train and test set, the following tests verify that the assumptions are met.

   - The error terms are calculated by using the formula (y_actual – y_predicted). Plotting a histogram of the previous calculation should give a normal distribution curve.
   - A scatter plot of the residual terms and the predicted values should not show any patterns.

That is how we can verify that the linear regression assumptions are valid and we can use the Linear Regression Algorithm.

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

   Temp (Temperature), LightRain (Light Rain) and Yr (Year) are the top three features which contribute significantly towards the count of shared bikes. Temp and Yr are positively impacting the count of shared bikes and LightRain is negatively impacting the count of shared bikes on any given day.

General Subjective Questions

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

   Linear Regression is a Machine Learning Algorithm wherein we predict a continuous variable using a single/multiple independent variables. It is a supervised ML Algorithm, which simply means that we use labelled data to learn the coefficients of the model and then predict on unseen data.

   Linear Regression makes use of the simple mathematic equation of a straight line.

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots\cdots\cdots\cdots\cdots + \beta_i x_i$$

   For Simple Linear Regression, we only have a single independent variable.

   To find out the best line (with optimal slope and coefficient) we define our cost function and use Gradient Descent algorithm to find the values of β_o and β_1.

   Loss Function:

$$RSS = \sum_{i=1}^{n} (y_i - y_{pred})^2$$

   Where
   RSS is the Residual Sum of Squares
   y_i is the actual y value
   y_{pred} is the predicted y value

$$RSS = \sum_{i=1}^{n} (y_1 - \beta_o - \beta_1 x_1)^2$$

   Now we have to the find the values of β_o and β_1 such that the RSS is minimized and for that we use the Gradient Descent algorithm. Following is the steps involved in the algorithm.

   - Initially let $\beta_o = \beta_1 = 0$
   - Calculate the derivate of the loss function wrt β_o and β_1.
   - Plug in the values of β_o, β_1 , x_1 and y_1 to find the partial derivative values (D_{β_o} and D_{β_1}).
   - Now using the below equation we update the values of β_o and β_1.
   - $\beta_o = \beta_o - 0.0001 D_{\beta_o}$ and $\beta_1 = \beta_1 - 0.0001 D_{\beta_1}$, where 0.0001 is the learning rate.

- We repeat the above steps till RSS becomes close to 0.

Assumptions of Simple Linear Regression:

- There must be a linear relation between the independent and dependent variables.
- Error terms, $\varepsilon_i = y_i - y_{pred}$ should be normally distributed.
- Error terms should be independent of each other.
- Error terms should have constant variance.

Multiple Linear Regression: We use more than one independent variables to predict the dependent variable.

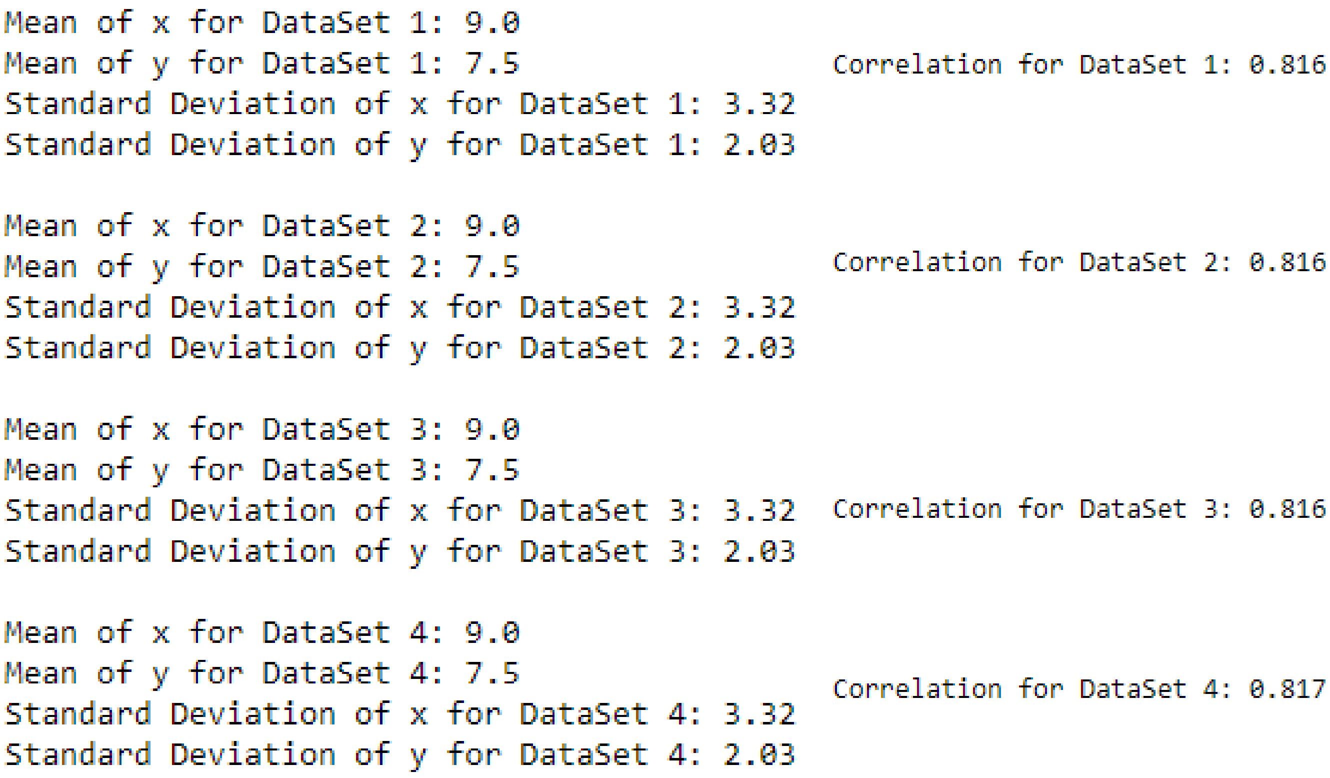
$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots\cdots\cdots\cdots\cdots + \beta_i x_i$$

Interpretation of the coefficients: Change in output(y) per unit increase in any one independent variable when others are held constant.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet has 4 datasets which have identical descriptive statistics however when plotted visually, they are very different.

It was constructed by Francis Anscombe to illustrate the importance of data visualization (EDA) before starting with the process of model building.

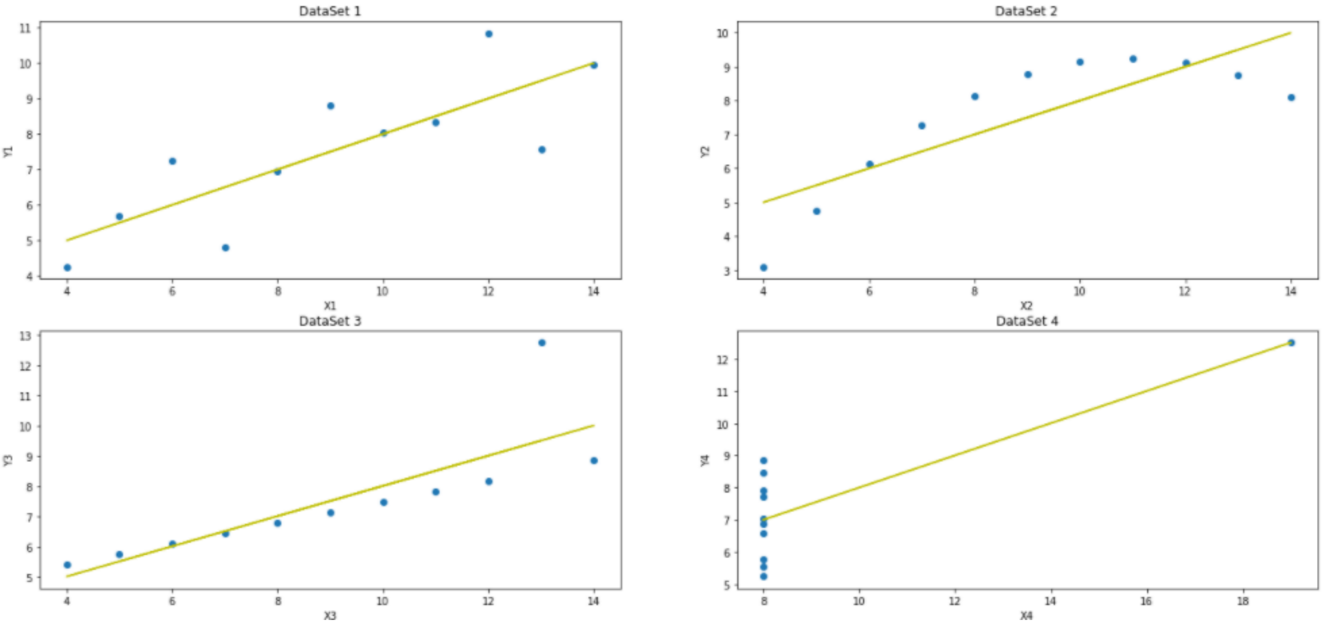


```
Mean of x for DataSet 1: 9.0
Mean of y for DataSet 1: 7.5                    Correlation for DataSet 1: 0.816
Standard Deviation of x for DataSet 1: 3.32
Standard Deviation of y for DataSet 1: 2.03

Mean of x for DataSet 2: 9.0
Mean of y for DataSet 2: 7.5                    Correlation for DataSet 2: 0.816
Standard Deviation of x for DataSet 2: 3.32
Standard Deviation of y for DataSet 2: 2.03

Mean of x for DataSet 3: 9.0
Mean of y for DataSet 3: 7.5
Standard Deviation of x for DataSet 3: 3.32     Correlation for DataSet 3: 0.816
Standard Deviation of y for DataSet 3: 2.03

Mean of x for DataSet 4: 9.0
Mean of y for DataSet 4: 7.5
Standard Deviation of x for DataSet 4: 3.32     Correlation for DataSet 4: 0.817
Standard Deviation of y for DataSet 4: 2.03
```



As we can deduct from the above plots, 4 very different datasets have almost same mean, standard deviation. Fitting a linear line while looks acceptable for Dataset 1 and Dataset 3, but not for Dataset 2 and Dataset 4. Therefore simply looking at the descriptive statistics and without seeing how the data looks like, might not be a very good decision for model building.

3. What is Pearson's R?

Pearson's R is a type of correlation coefficient which is used in Linear Regression. It measures how strong the linear relationship exists between two variables and reports a value between -1 and 1, where:

- 1 signifies perfect positive linear correlation
- -1 signifies perfect negative linear correlation
- 0 signifies no linear correlation

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

where:

- x_i is the value of x variable.
- $\bar{x}$ is the mean of values of x variable.
- y_i is the value of y variable.
- $\bar{y}$ is the mean of values of y variable.
- r is the correlation coefficient.

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

Scaling is a technique to standardize the independent features limited to a fixed range which can then be fed to the ML algorithm.

Scaling is essential when we need to understand the coefficient (how each independent variable impacts the dependent variable). Without scaling, the ML algorithm tends to assign weights (coefficients) based on the range of the independent feature. By adding constraint on the values taken by the independent features, we can then make meaningful interpretation of their coefficients.

Scaling can be performed using the following 2 algorithms.

- MinMax Scaler(Normalized Scaling)
- Standard Scaler(Standardization)

In Normalized Scaling technique, we limit the feature value between 0 and 1. The following computation is used to calculate the scaled value.

$$X_{i,sc} = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

In Standardization Scaling technique, we rescale such that the mean of the distribution is 0 and the standard deviation is 1. (Standard Normal Distribution and Z-Score). The following computation is used to calculate the scaled value.

$$X_{i,sc} = \frac{X_i - \bar{X}}{\sigma_X}$$

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

Variance Inflation Factor or VIF is a measure of multicollinearity in the independent features in our dataset. A high VIF means that, that feature is explained well by other features in the dataset. It is calculated using the following formula.

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

VIF for any independent feature is calculated by running an ML model with that feature as target variable and all other features as independent features. The R-Squared metric returned by our model is used to calculate the VIF of the independent feature (which was the target variable).

VIF is infinite when ${R_i}^2 = 1$
A value of 1 for R-Squared means that, the variance in that feature is explained perfectly by all other independent features.

As a general heuristic, it is recommended to drop features which have VIF > 5.

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

   Q-Q plots (Quantile-Quantile plots) are plots of two quantiles against each other. It plots the quantiles of first dataset against a second dataset. Q-Q plot helps us determine if two datasets come from a similar distribution.

   In linear regression, Q-Q plots are used to verify the assumption if the error terms are normally distributed. We plot the quartiles of the error terms against quartiles of a normal distribution. If the error terms are normally distributed, in the Q-Q plot we will see the points along a straight line (at an angle of 45 degrees from X axis).