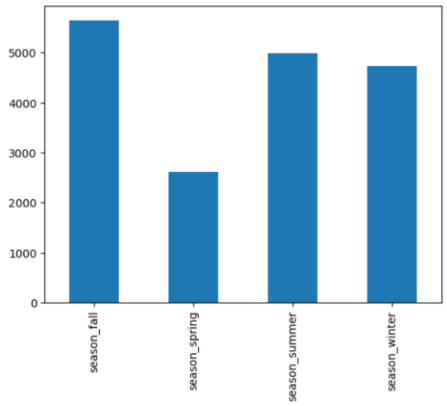
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

Ans:

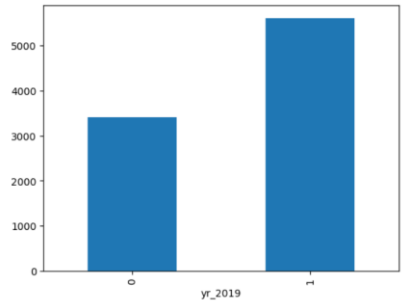
I have visualized the impact of categorical variables on the dependent variable using bar graph and we can infer the following from it.

- **Season :**



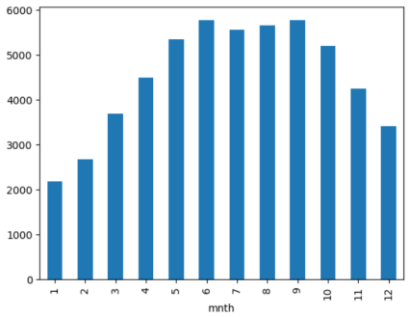
The value of **cnt** is more for **season_fall** and is less for **season_spring.**

- **Year :**



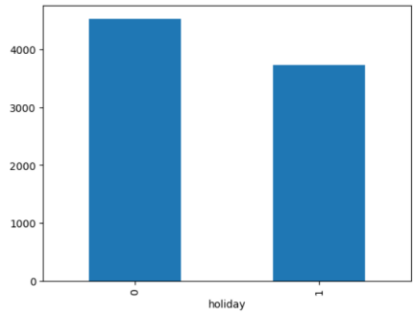
The average value of **cnt** is more in year 2019 than in 2018.

- **Mnth:**



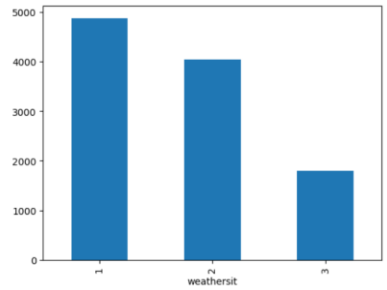
The average value of **cnt** is more for months 5,6,7,8,9 and 10.

- **Holiday**:



The average value of **cnt** is more when it is not holiday.

- **Weathersit**:



The average value of **cnt** is more when weathersit is 1 and less when it is 3.

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

Ans:

- Using **drop_first=True** prevents creation of an extra variable.
- When you have a categorical variable of n levels, the idea of dummy variable creation is to build (n-1) variables instead of n variables as shown below.

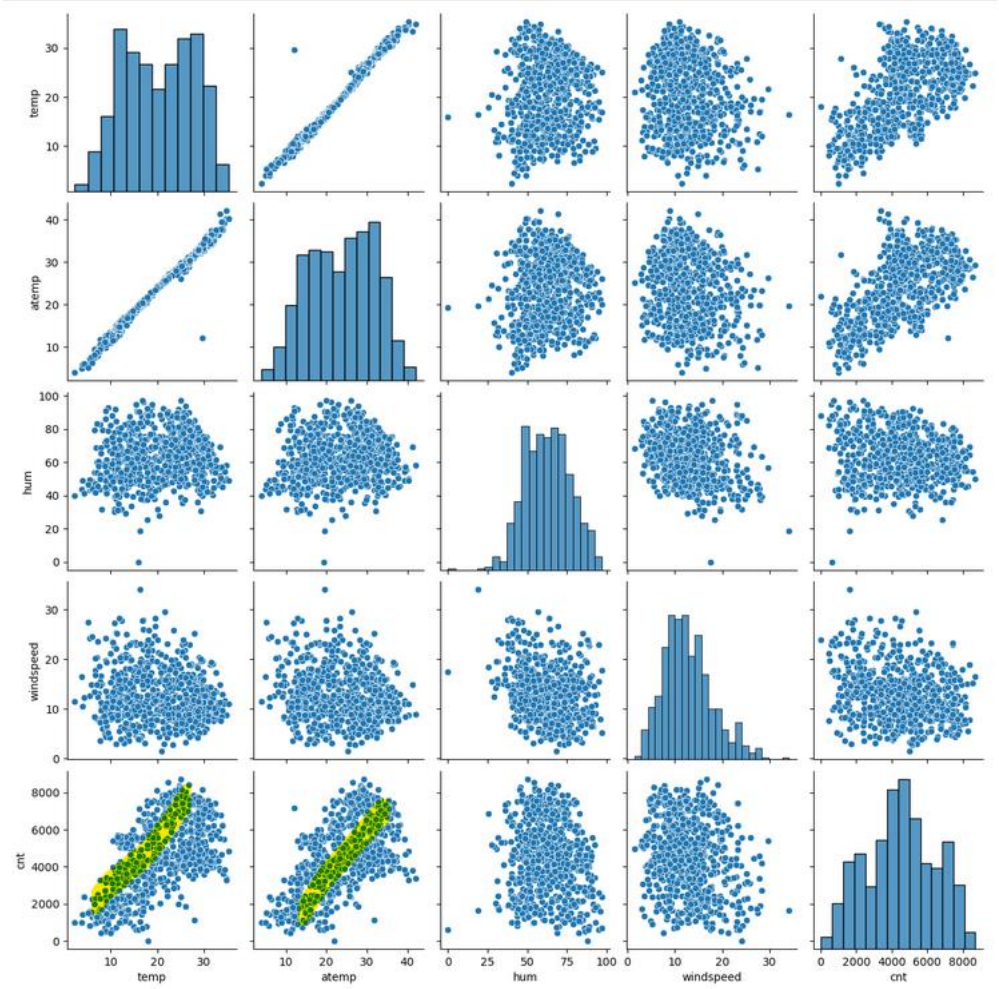
Relationship status	In a relationship	married
Single	0	0
In a relationship	1	0
married	0	1

Clearly the variable Relationship status has 3 levels, but we created 2 dummy variables treating if a person is not in a relationship and not married is obviously single.

- To achieve this (creating n-1 dummy variables instead of n), we can use **drop_first=True.**

3. Looking at the pair-plot among the numerical variables, which one has the highest correlationwith the target variable? (1 mark)

Ans:

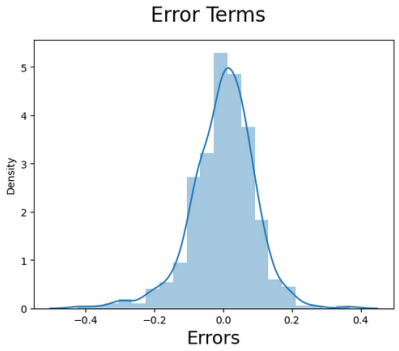


- By looking at the pairplot, we can clearly see the independent variables **temp** and **atemp** has high correlation with the dependent variable **cnt.**

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

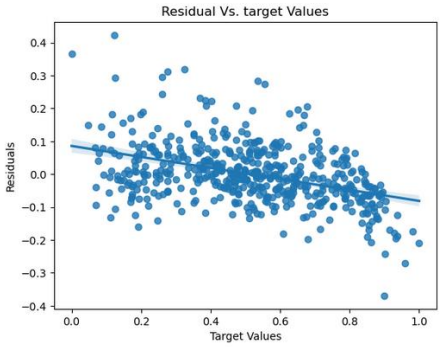
Ans:

- After building the model, I have checked the following assumptions.
  1. Error terms are normally distributed (not X, Y).



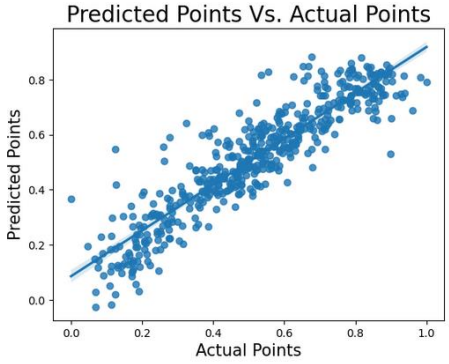
Error terms (y_pred(i) – y(i)) are normally distributed with mean 0.

2. Error terms are independent of each other.



I have plot a graph between error values and the y_train values, I did not see any patterns/relation between them.

3. Error terms have constant variance (homoscedasticity)



I have plot a graph between predicted y values and actual y values. I can see a constant variance between them.

4. Linear relationship between X and Y.
   For this – I have eliminated features using RFA and manual approach and only selected the independent variables which have relationship with Y and does not have multicollinearity between themselves.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

Ans:

- The final model which we built is –

  cnt = 0.2519 - 0.0986 X holiday + 0.4515 X temp - 0.1398 X windspeed - 0.1108 X season_spring + 0.0473 X season_winter + 0.2341 X yr_2019 - 0.0727 X month__July + 0.0577 X month__September - 0.0811 X weathersit_2 - 0.2864 X weathersit_3

- The top 3 features contributing significantly towards explaining the demand of the shared bikes are:
  1. **temp**
  2. **weathersit_3**
  3. **yr_2019**

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans:

- Linear regression is a supervised machine learning algorithm where we train a model to predict the dependent variable of your data based on some independent variables.
- In the case of linear regression as you can see the name suggests linear that means dependent variable must be linearly correlated with independent variables.
- The standard equation of the regression line is given by the following expression:

  $\mathbf{Y = \beta_o + \beta_1 X}$

  Here,

  Y -> dependent variable (target variable to predict)

  X -> independent variable

  β_1 -> slope of regression line (Used for interpretation)

  β_o -> constant

- Linear regression can be of two types –
  1. Simple linear regression -> one independent variable
  2. Multiple linear regression -> more than one independent variable
- The assumptions of simple linear regression are:
  1. Linear relationship between X and Y
  2. Error terms are normally distributed (not X, Y)
  3. Error terms are independent of each other
  4. Error terms have constant variance (homoscedasticity)

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans:

- 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.
- The four datasets of Anscombe's quartet are:

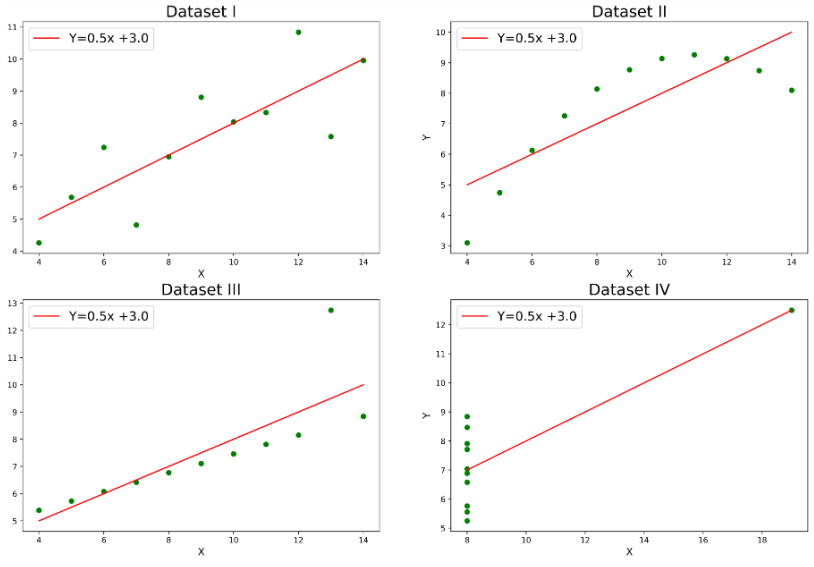
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

- The summary statistics of all 4 datasets are:

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727

We can see it is same for all the four datasets.

- When we plot the scatter plot for X and Y for the above four datasets



Anscombe's quartet Plot

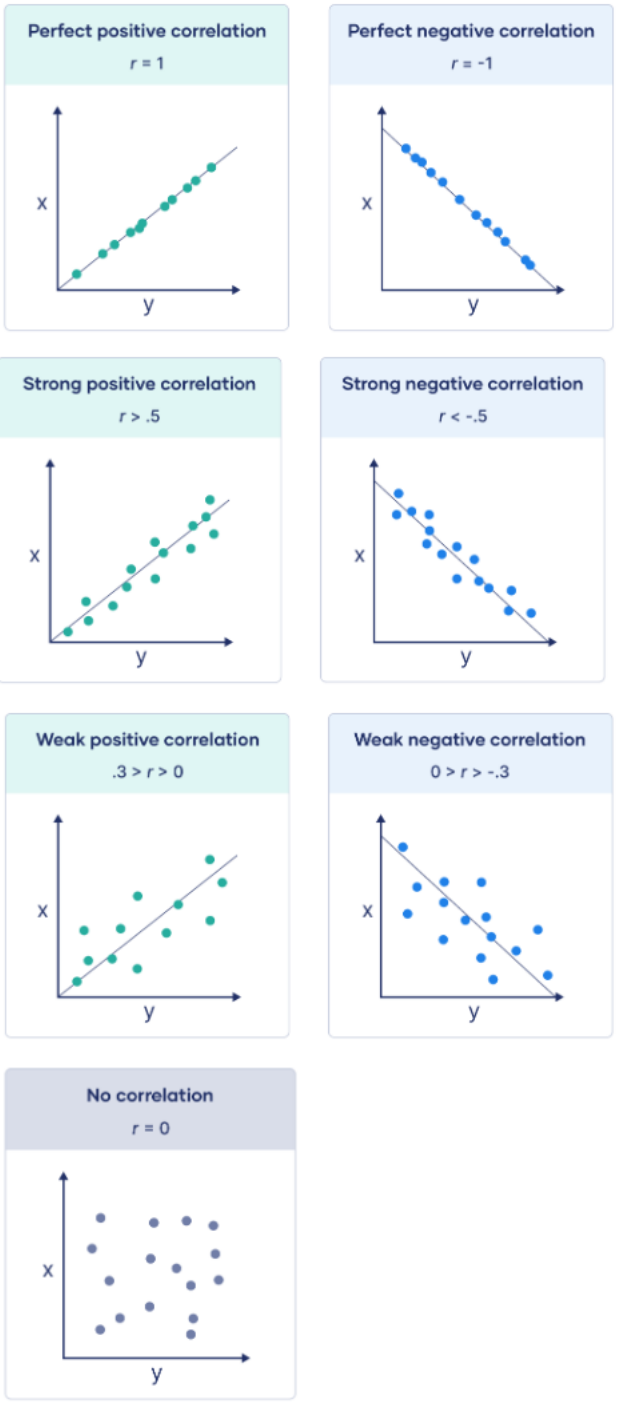
1. In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
2. In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
3. In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
4. Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

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

3. What is Pearson's R? (3 marks)

Ans:

- 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.
- **Visualization:**



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans:

- When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons
    1. Ease of interpretation
    2. Faster convergence for gradient descent methods
- You can scale the features using two very popular method:
    1. **Standardizing:** The variables are scaled in such a way that their mean is zero and standard deviation is one.

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

    2. **MinMax Scaling**: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans:

- **Variance Inflation Factor (VIF):** By looking at correlations might not always be useful as it is possible that just one variable might not completely explain some other variable but some of the variables combined might be able to do that. To check this sort of relations between variables, we use VIF. VIF basically helps explaining the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below :

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

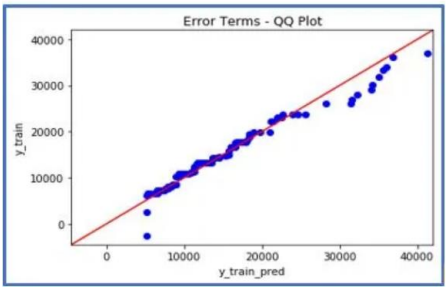
- If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.
- To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

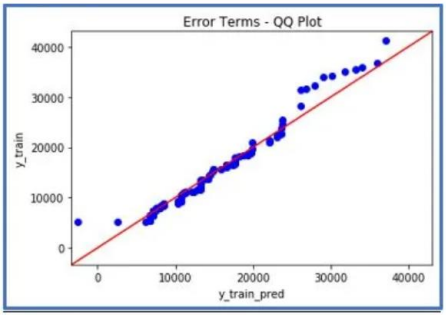
Ans:

- 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.

- **Advantages:**
  - It can be used with sample sizes also.
  - 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 —
    - Come from populations with a common distribution.
    - Have common location and scale.
    - Have similar distributional shapes.
    - 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.
    - **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis.
    - **Y-values < X-values**: If y-quantiles are lower than the x-quantiles.



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



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