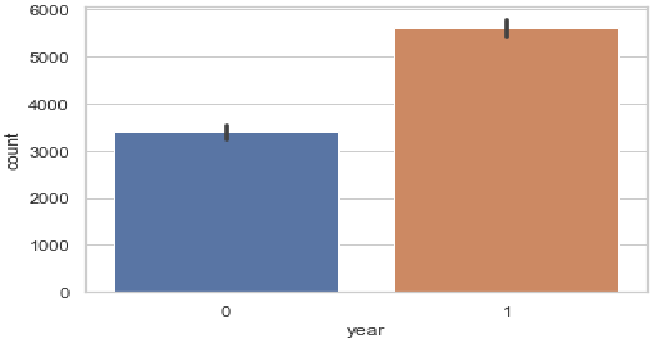
ASSIGNMENT BASED SUBJECTIVE QUESTION

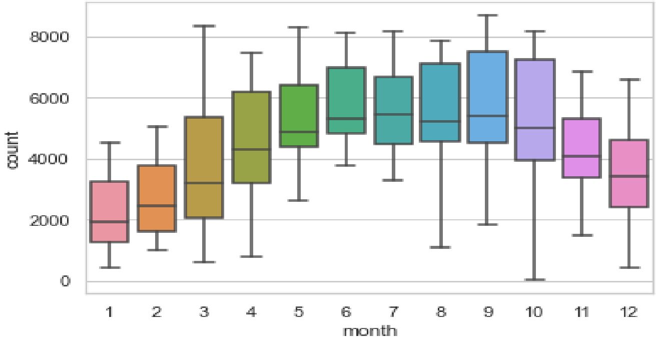
-ARUN TYAGI

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

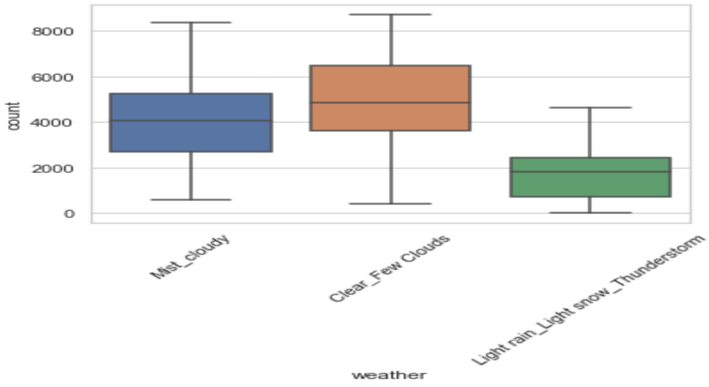
- **Year** - From the above bar plot it can be clearly inferred that the Bike was subscribed more in the year 2019 compared to 2018 hence there was a significant increase in the Bike Sharing business.



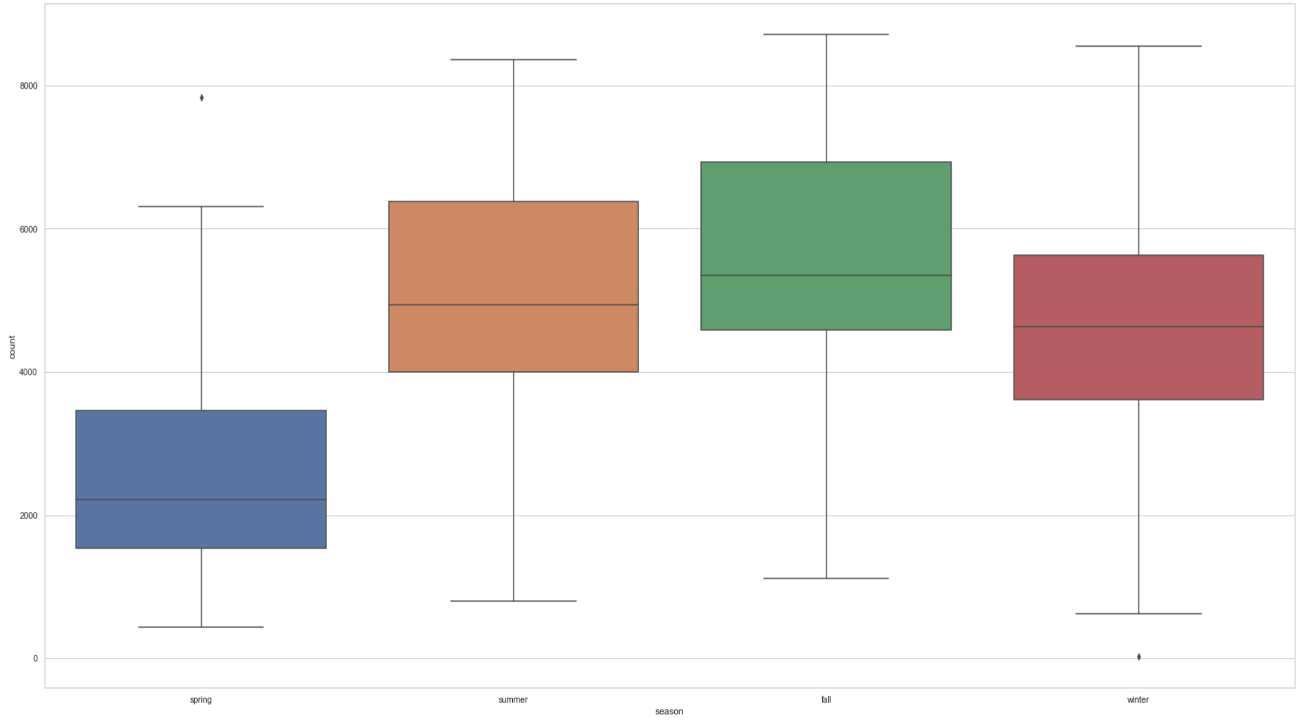
- **Month** – Bike was subscribed more during the month of September specifically between August to October, so he makes sense that season - Fall had more subscribers for the Bikes.



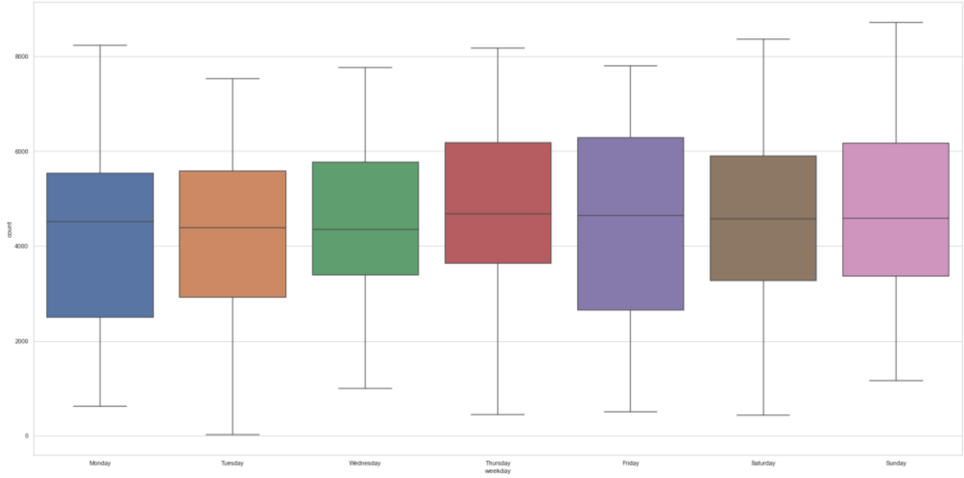
- **Weather -** When the weather is clear the count for total bike users increases. And when there's Light rain and snow the count for bike users decreases which is intuitive as well.



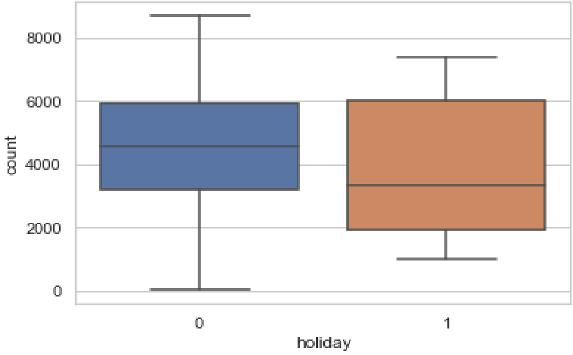
- **Season -** Fall season is best season in terms of business as count for bike users is the highest in fall. Spring season on the other hand has the lowest count for bike users



- **Days -** Median value is approximately same for all the days which is above 4000 which tells us that on all 7 days in a week the bike was on an average subscribed to more than 4000 times. Sunday records the highest count of Bike subscribers and on the Contrary Tuesday records the lowest count.



- Holiday - We can infer that on the Working Day bike was on an average subscribed more than 4000 times compared to a Holiday when it was subscribed to approximately 3000 times. The maximum times it was subscribed also happens to be a Weekday and the max. times it was subscribed on a Holiday is slightly more than 7000 times.

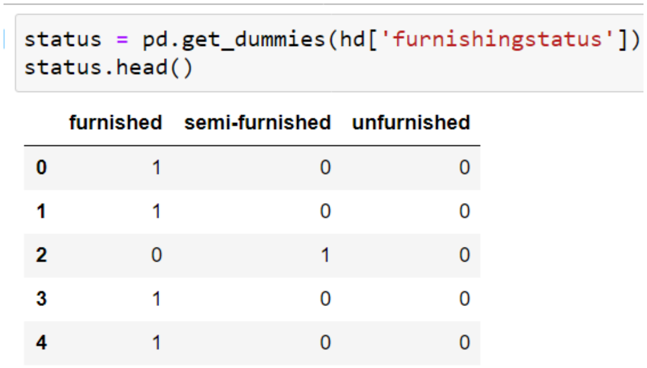


2. Why is it important to use drop first=True during dummy variable creation?

- '**<u>drop_first = True</u>**' is an argument which is passed while creating dummy variables during the Pre-Processing of the data phase in Machine Learning Modelling. It is a One-Hot encoding method which is a way to include 'Categorical Variables' in the Regression Model. To include it in the Regression Model, we must convert it into a number as Model understands only numbers and not nominal values.

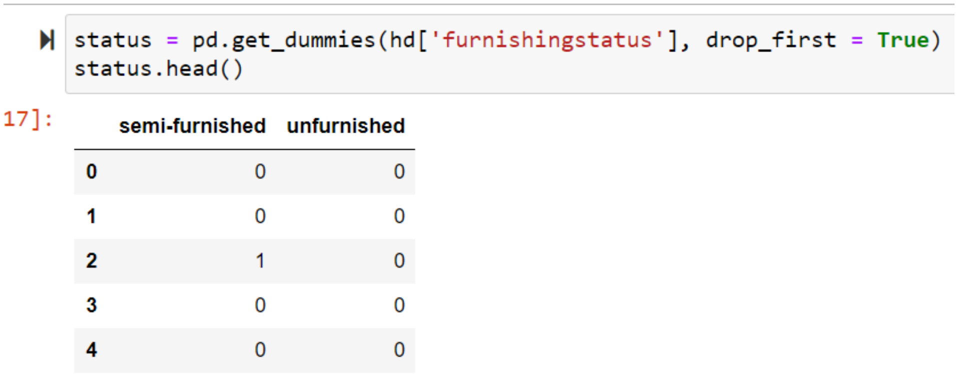
- When we use this argument, it drops the redundant column. If we have 'K'-variables, then they can be represented by using 'K-1' variables. For instance, if we have 3 variables then they can be represented by 2 variables.

- **<u>Example</u>**: Following code snippets & tables generated respectively when **<u>drop_first = True</u>** is not used & Default = False, meaning that the reference is not dropped and k dummies created out of k categorical levels.



```
status = pd.get_dummies(hd['furnishingstatus'])
status.head()
```

	furnished	semi-furnished	unfurnished
0	1	0	0
1	1	0	0
2	0	1	0
3	1	0	0
4	1	0	0



```
status = pd.get_dummies(hd['furnishingstatus'], drop_first = True)
status.head()
```

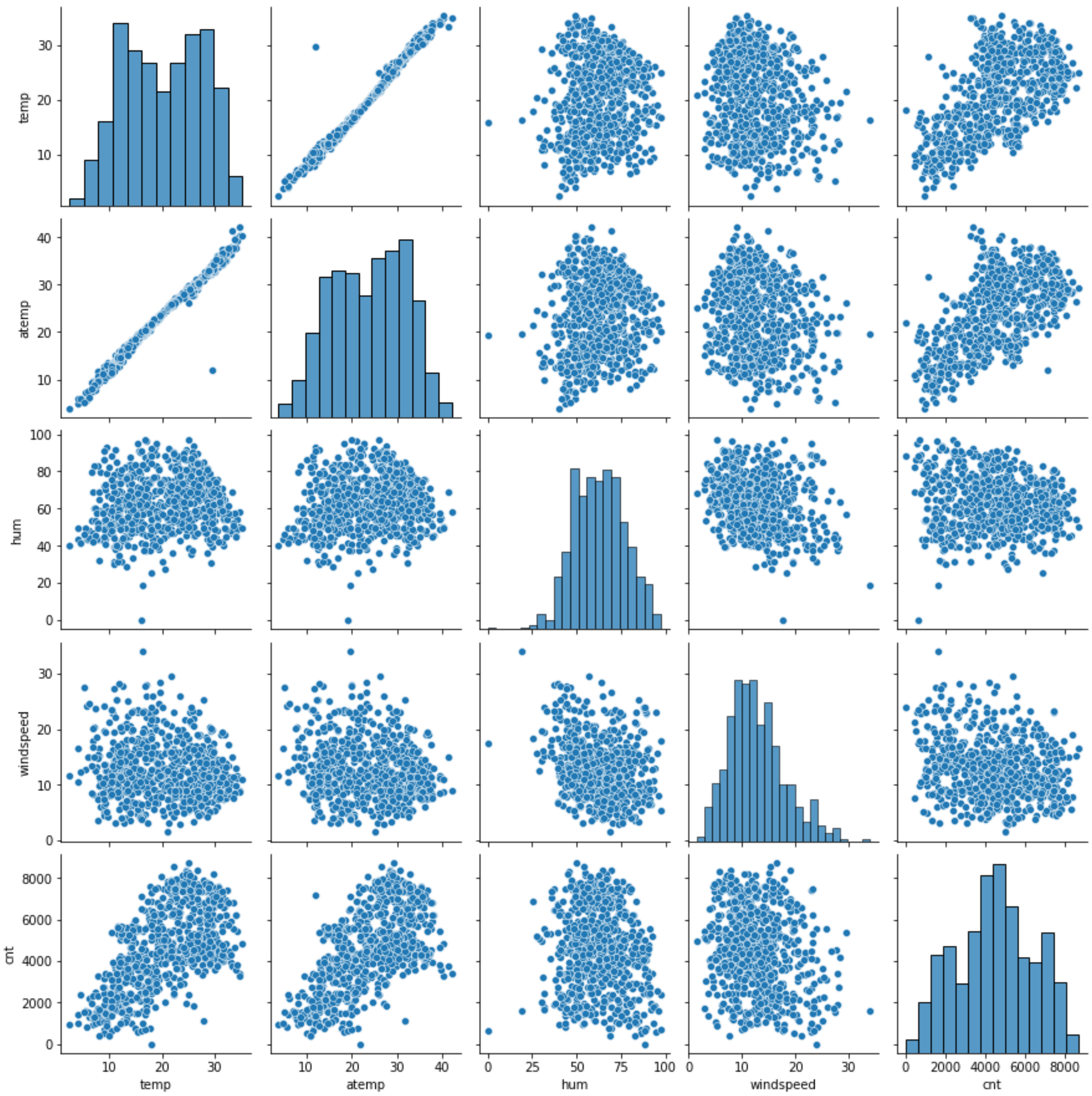
17]:

	semi-furnished	unfurnished
0	0	0
1	0	0
2	1	0
3	0	0
4	0	0

- The importance of using this argument is to reduce the correlation between the variables. Having high collinearity in the dataset will affect the interpretation & inferences which are drawn from the final model thereby impacting the decision-making ability. This also goes against our Assumptions of Multi Linear Regression.

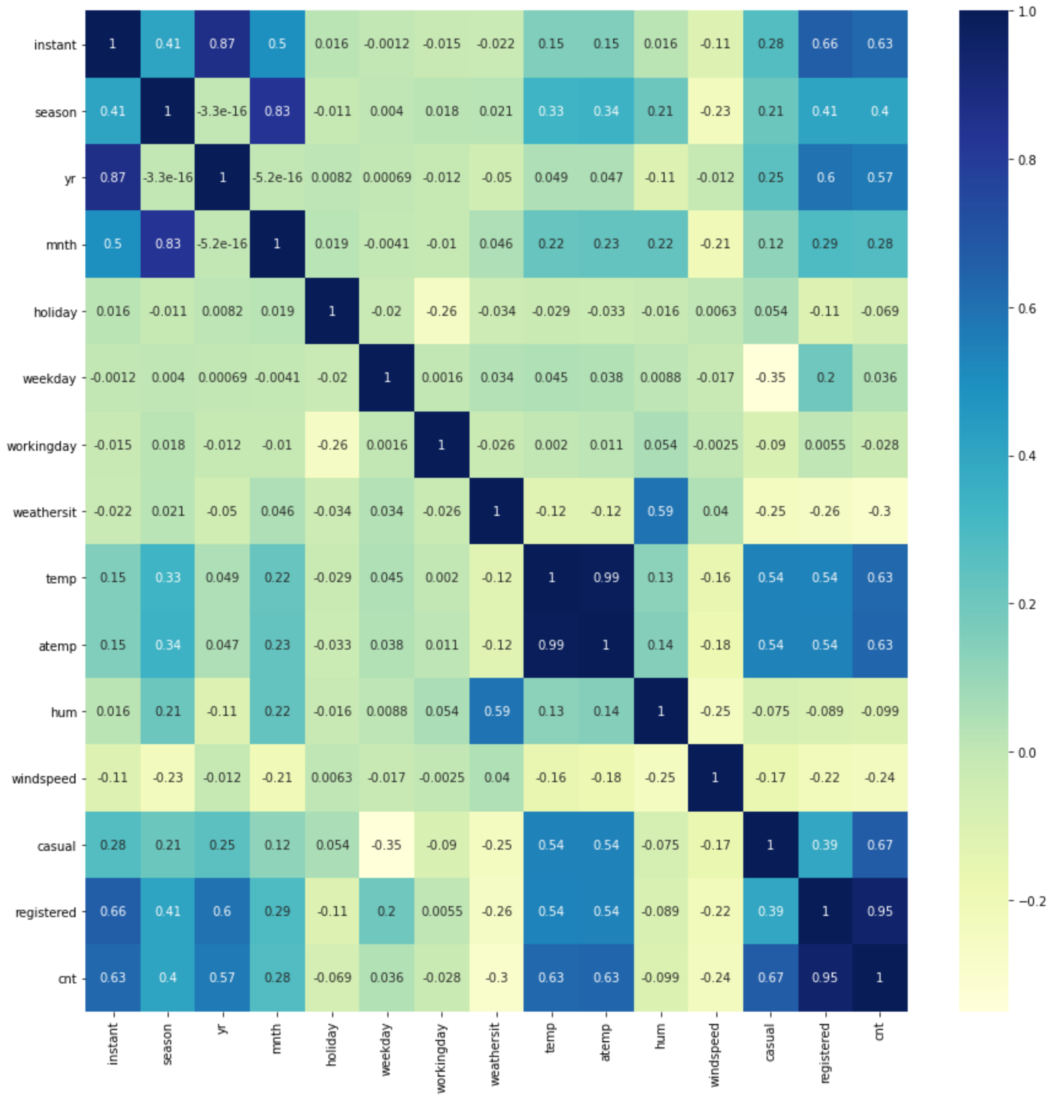
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

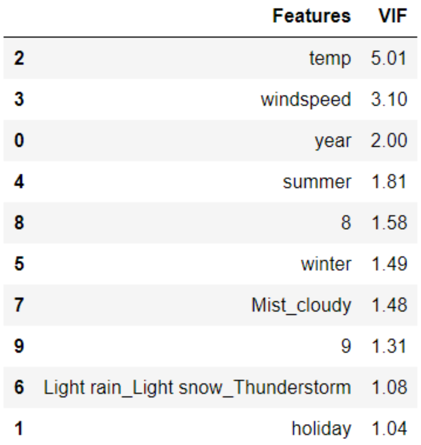
- Both temp and atemp have same correlation that is 0.63 with the target variable cnt.



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

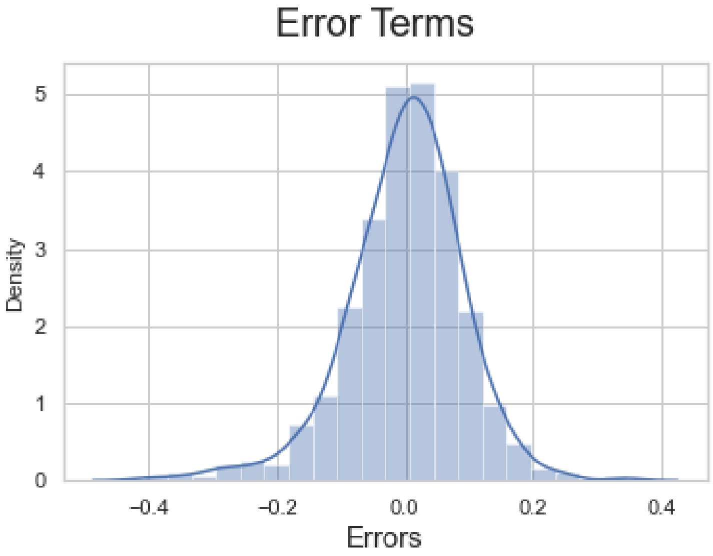
- **Linear Relationship between the features and target** – We checked the there is a linear relationship between the Target Variable Count and the Temperature Variable.
- **Little or no Multicollinearity between the features** – We validated this using Pair Plot and Heatmap. We dropped correlated values like atemp, registered, casual, dteday to ensure the dataset is free from any Multicollinearity. We also validated the same for independent variables through the Variance Inflation Factor (VIF) which is shared below.





	Features	VIF
2	temp	5.01
3	windspeed	3.10
0	year	2.00
4	summer	1.81
8	8	1.58
5	winter	1.49
7	Mist_cloudy	1.48
9	9	1.31
6	Light rain_Light snow_Thunderstorm	1.08
1	holiday	1.04

- **We did Residual Analysis to check the distribution of error terms.**



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

Top 3 features which are contributing significantly towards explaining the demand of the shared bikes are as follows –

1. Temperature - A coefficient value of '0.5174' indicates that a unit increase in temp variable increases the bike hire numbers by '0.5174' units.

2. Light rain_Light snow_Thunderstorm-A coefficient value of '-0.2828' indicates that a unit increase in Light rain_Light snow_Thunderstorm variable decreases the bike hire numbers by '0.2828' units

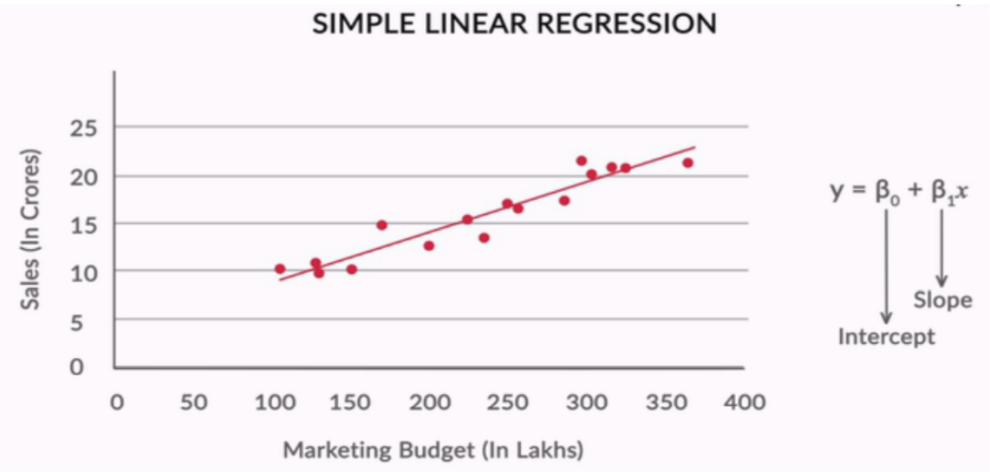
3. Year - A coefficient value of '0.2325' indicates that a unit increase in year variable increases the bike hire numbers by '0.2325' units.

- $count=0.2325\times year-0.0971\times holiday+0.5174\times temp-0.1497\times windspeed+0.1000\times summer+0.1383\times winter+0.0542\times August(8)+0.1162\times September(9)-0.2770\times LightrainLightsnowThunderstorm-0.0825\times Mistcloudy$

General Subjective Questions

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

    - Linear Regression is a Machine Learning algorithm based on the Supervised Learning Method where historic data is labelled and used to determine the value of the output variable.
    - It is a statistical tool that is used to find out the relationship between the outcome variable also known as the dependent variable, and one or more independent variables.
    - This method is mostly used for forecasting and finding out cause-and-effect relationships between variables.
    - There 2 types of Linear Regression based on upon the number of independent variables involved-
        - Simple Linear Regression (SLR)
        - Multi Linear Regression (MLR)
    - In Simple Linear Regression, we have only one independent variable which is used to predict the value of the dependent or the target variable. It's represented by the equation of the straight line that is **y = mx + c.** Here 'c' represents the constant and 'm' represents the slope of x which is nothing but change in x (independent variable) and y is the target or the dependent variable.

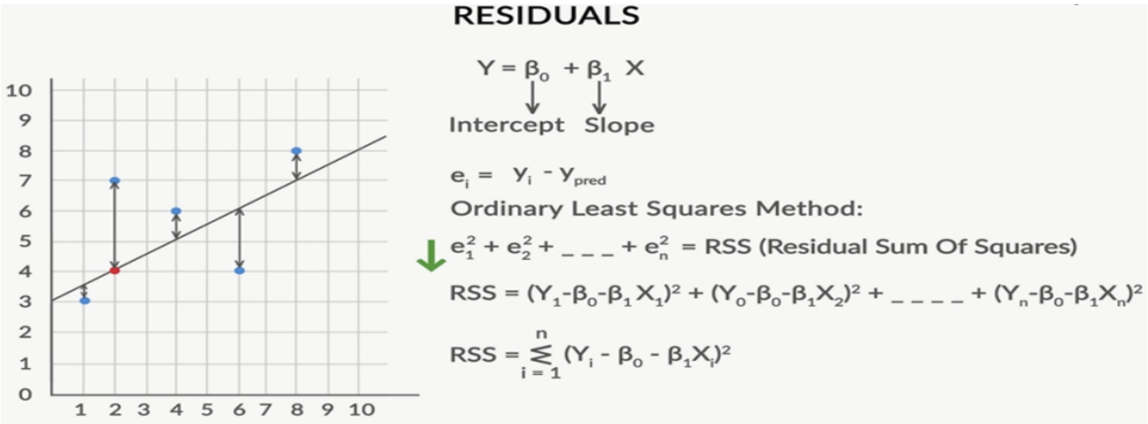


    - In Multi Linear Regression we have more than one independent variable which is used to predict the values of the dependent or the target variable. It's also represented by the equation of a straight-**line y = c +** $\mathbf{m_1x_1 + m_2x_2 + m_3x_3 + \ldots + m_nx_n}$ where 'n' represents the n^{th} independent variable. m_1, m_2, m_3....m_n represents the slope/coefficients of the corresponding independent variables, c represents the constant which is also known as Intercept (In simple words, it's the point on the y-axis where the line meets) and y is the target or the dependent variable.

- In Linear Regression we try to find the optimal values of m and c and fit the best straight line through the data points.

- The Best Line is the one whose Residual Sum of Squares of error terms is minimum that is nothing but the sum of squares of difference of actual y value and the predicted y value.

- The Residual Sum of Squares (RSS) is a cost function that is minimized by different techniques but here in our Assignment, we have used Gradient Descent.


RESIDUALS
$Y = \beta_0 + \beta_1 X$
Intercept
Slope
$e_i = y_i - y_{pred}$
Ordinary Least Squares Method:
$e_1^2 + e_2^2 + \_\_\_ + e_n^2$ = RSS (Residual Sum Of Squares)
$RSS = (Y_1 - \beta_0 - \beta_1 X_1)^2 + (Y_0 - \beta_0 - \beta_1 X_2)^2 + \_\_\_\_ + (Y_n - \beta_0 - \beta_1 X_n)^2$
$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$
10
9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10

- While building the Linear Regression Model for prediction we consider 5 assumptions which are as follows-

  - **Autocorrelation** – There is no autocorrelation in data. It takes place when residual errors are dependent on each other in one or another way.

  - **Multi-Collinearity** – It doesn't exist or is present scarcely. It happens when independent features or variables show some dependency.

  - **Linear Relationship -** There exists a linear relationship between dependent and independent variables.

  - **Normal Distribution -** The error terms follow Normal Distribution.

  - **Homoscedasticity –** The error terms have constant variance.

- Uses of Linear Regression –

  - To estimate the relationship between target and independent variables.

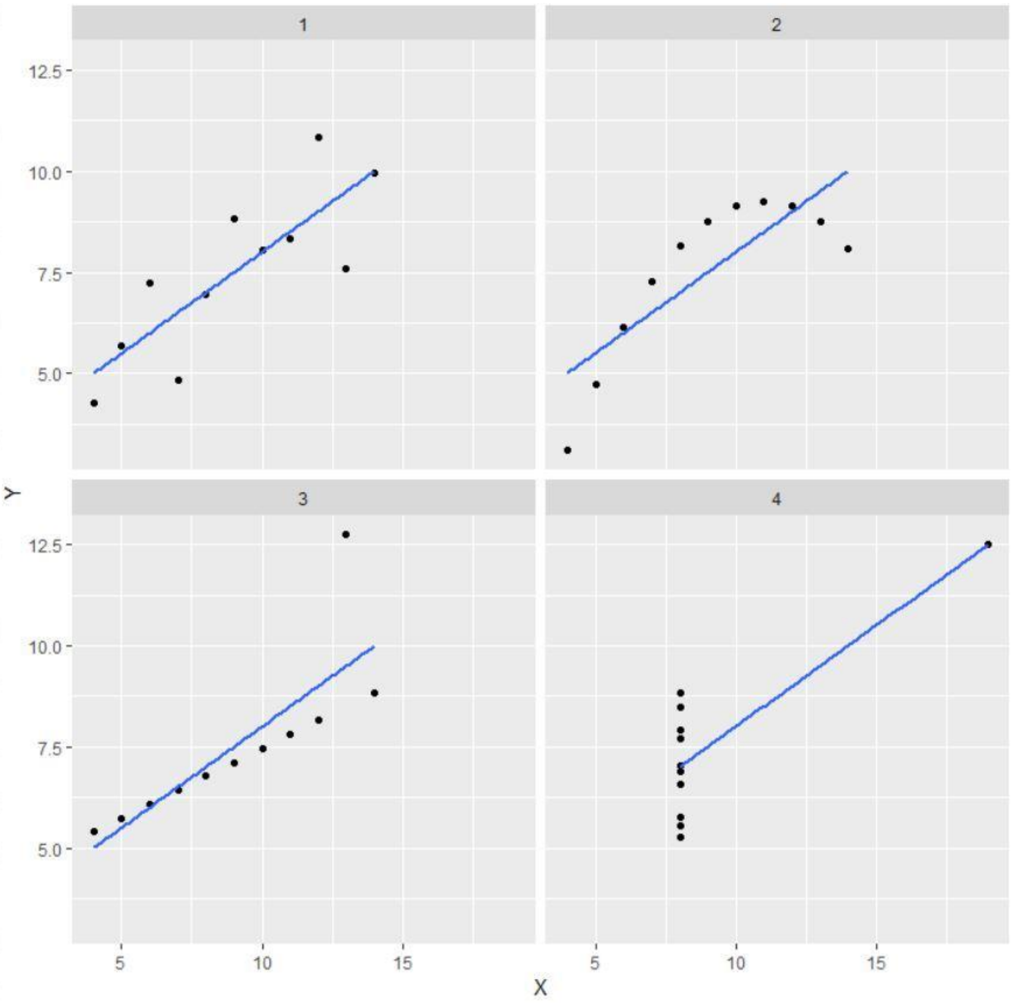
  - To find trends in data.

  - To help in predicting real/continuous values.

  - To determine the most important factor, the least important factor, and how each factor is affecting the other factors.

2. Explain Anscombe’s quartet in detail.

- Anscombe’s quartet consists of four datasets that have nearly identical simple statistical properties but appear different when plotted on the graph.
- Each dataset consists of eleven (x, y) points.
- They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.
- Anscombe's Quartet reminds us that graphing data before analysis is good practice, outliers should be removed when analyzing data, and statistics about a data set do not fully depict the data set in its entirety.
- Statistics are great for describing general trends and aspects of data, but they alone can't fully depict any data set.
- In the given graph shown below, all four datasets have the same variance in x, variance in y, mean of x, mean of y and linear regression but they are different from one another.
- In the first one (top left) scatter plot there seems to be a linear relationship between x and y.
- In the second one (top right) scatter plot we can conclude that there is a non-linear relationship between x and y.
- In the third one (bottom left) figure we can say that there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated to be far away from that line.
- Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.



- Anscombe’s Quartet can be connected to the Exploratory Data Analysis and Data Handling that we do before building the Machine Learning model to avoid the issues mentioned above.

- Application: 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.

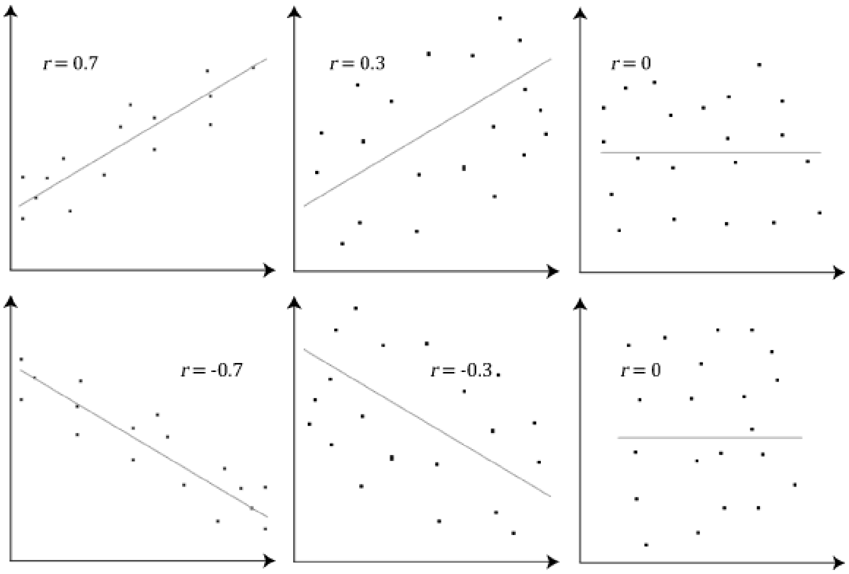
3. What is Pearson’s R?

- Pearson’s correlation coefficient is used to measure the strength of a linear association between two variables. It’s denoted by R.

- It draws a line of best fit through two variable’s data. It indicates how far away all these data points are to this line of best fit.

- Pearson’s correlation coefficient variables can be measured in different units based on the dataset. For example, in the Bike Sharing dataset we could correlate the count of Bike subscribed in a day to the temperature which is in degree Celsius/Fahrenheit or humidity and windspeed which are completely in different units.

- Pearson's correlation coefficient(r) is a unitless measure of correlation and doesn't change in the effect of origin or scale shift measurement.
- It doesn't take into consideration whether a variable has been classified as a dependent or independent variable. It treats all variables equally. We might want to find out whether basketball performance is correlated to a person's height. But if we determine whether a person's height was determined by their basketball performance (which makes no sense), the result will be the same.
- Properties:
  - The range of r is between [-1,1].
  - The computation of r is independent of the change of origin and scale of measurement.
  - r = 1 (perfectly positive correlation), r =-1 (perfectly negative correction)
    r = 0 (no correlation)



- Pearson's correlation coefficient formula is represented by –

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

where x_i, y_i, are the variables and x bar, y bar is the mean, respectively.

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 present in the dataset in a fixed range. It’s performed during Data Pre-Processing.

- Scaling is performed for two reasons:

  1. Ease of interpretation

  2. Faster convergence for gradient descent method

- There are two methods to scale the features:

  1. **Standardizing**: The variables are scaled in such a way that their mean is zero and standard deviation is one. It means after Standardization features will have mean = 0 and std. dev. = 0. Standardizing is also known as Z-Score Normalization.

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

  2. **Normalization**: 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. Normalization is also known as min-max normalization or min-max scaling.

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

- Difference between Normalization and Standardization:

  - If you have outliers in your feature (column), normalizing your data will scale most of the data to a small interval, which means all features will have the same scale and hence it will not handle outliers well.

  - Standardization is more robust to outliers, and in many cases, it is preferable over Max-Min Normalization.

  - Normalization is good to use when your data does not follow a Normal distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

  - Standardization, on the other hand, can be helpful in cases where the data follows a Normal distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

- VIF stands for Variance Inflation Factor. It helps explaining the relationship of one independent variable with all the other independent variables.
- The common heuristic for VIF is that while a VIF greater than 10 is high, a VIF of greater than 5 should also not be ignored and inspected appropriately.
- VIF is given by the formula-

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

- 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 $1/(1\text{-}R^2)$ infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

- The quantile-quantile plot or Q-Q plot is a graphical tool to validate if two datasets are coming from populations with common distribution.
- We assume that given data to be normally distributed for ease of inferring useful information. One way to assess our assumption's correctness is to use Q-Q plot. Not just Normal distribution, we can test for other distributions (for example uniform distribution etc.) as well.
- Quantiles are the breakpoints that divide the ordered numerical data into equal sized bins.
- Percentiles are a type of quantiles that divide the data into 100 equal bins, quartiles divide the data into 4 equal parts and so on.
- Q-Q plot compares the quantiles of 2 datasets. We can make Q-Q for any 2 datasets if the quantiles can be calculated for both.
- 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.
- Importance of Q-Q plot in Linear Regression:

- Two datasets/sample can be of different size.

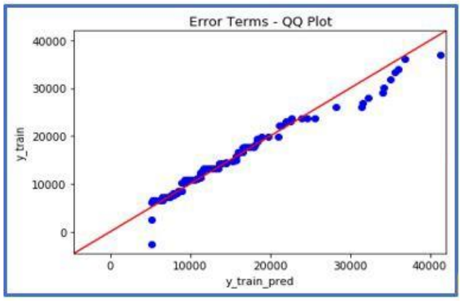
- Q-Q plot can detect outliers, shifts in scale, location, symmetry etc. simultaneously.

- One of the important assumptions of Linear Regression is that the residual of the model is normally distributed. This can be assessed using Q-Q plot

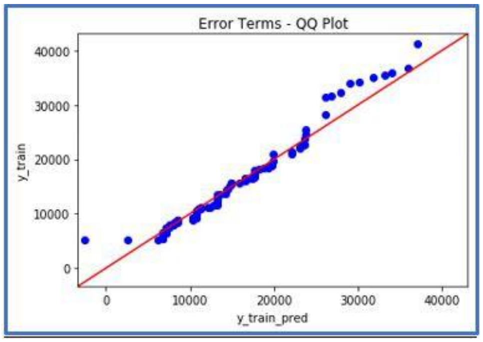
- Interpretation:

  - 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