Assignment-based Subjective Questions

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

Ans:-

a) **Year**:- Bike Rentals have a substantial increase in counts from year 2018 to 2019.

b) **Weathersit**:- Bike rentals for Light + snow have not seen a great increase from 2018 to 2019 but has grown significantly with Clear + weathersit and Mist.

c) **Season**:- Bike rentals are highest in fall, followed by summer. It is lowest in spring season

d) **Month**:- Bike rentals are highest in september,followed by june and august. It is lowest in January.

e) **Holiday**:-Bike rentals reduced during holidays.

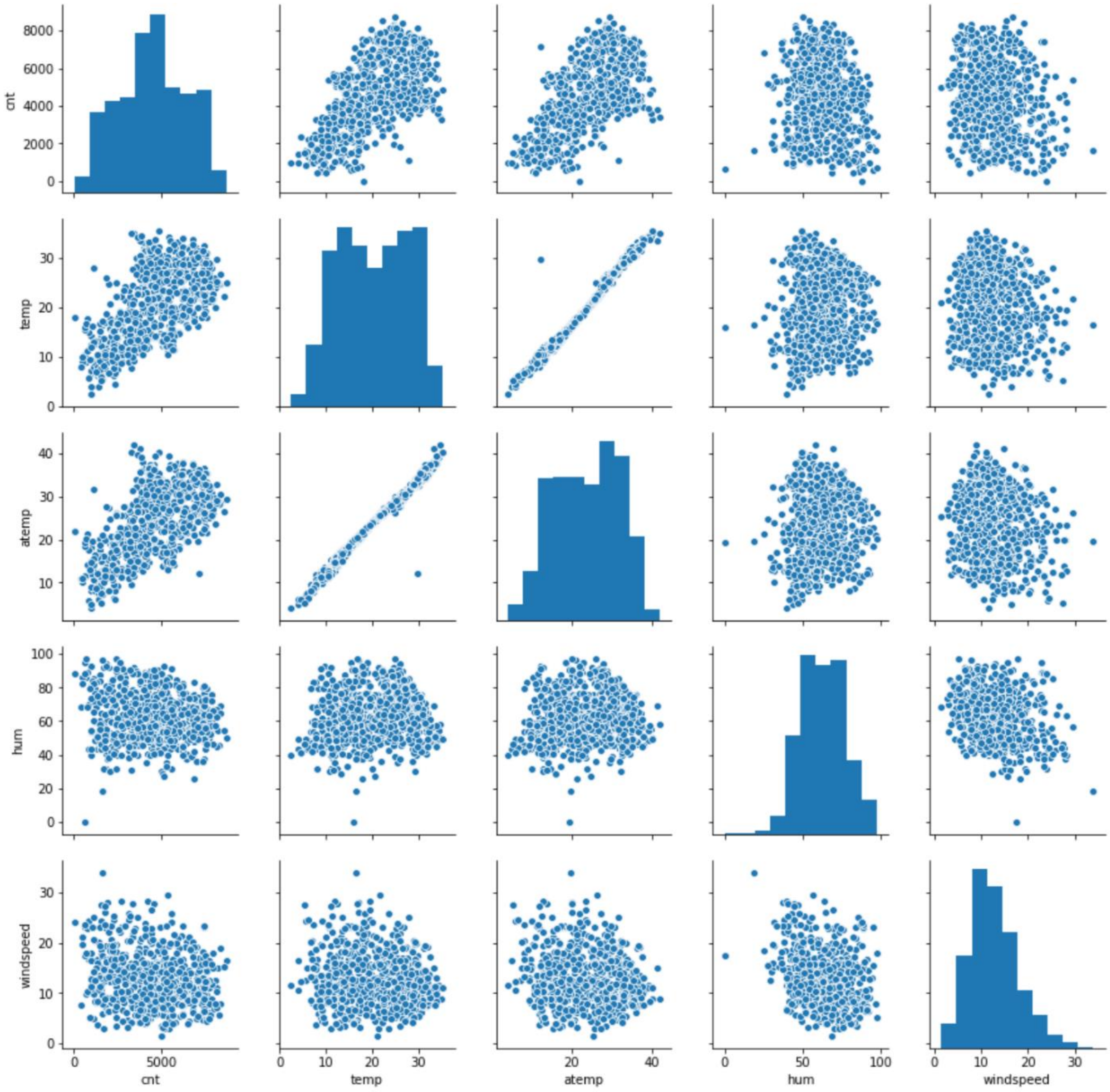
2. **Why is it important to use drop_first=True during dummy variable creation?**

Ans:-

- If you don't use "drop_first=True",you will get a redundant feature, let's see an example.

- If you have a feature "Is_male", you use "get_dummies" you will get two features "Is_male_0" and "Is_male_1", but if you look carefully they are redundant actually you just need one of them, the other one will the exact opposite of the other.

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

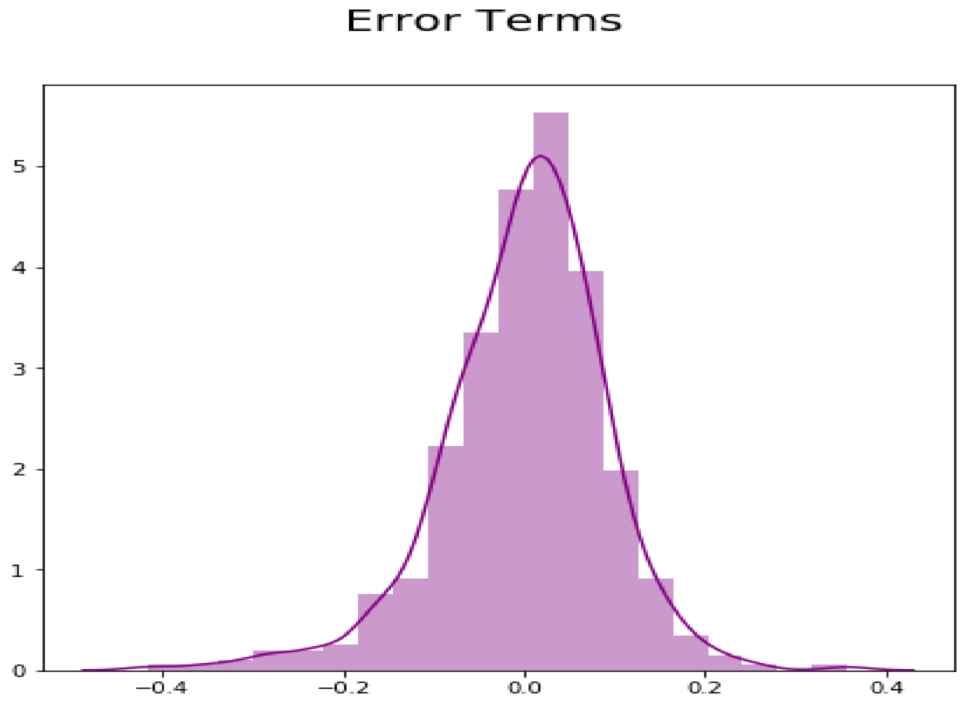


Ans:-

After looking at the pair-plot, we can conclude that “temp” and “atemp” are highly correlated with target variable “cnt”

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

Ans:-



- Residuals should follow a normal distribution and centered around zero(mean=0). We validate this by plotting a distribution plot of error terms and see if they follow a normal distribution or not. The results that we have retrieved and displayed above also follows a normal distribution.

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

Ans:-

The top 3 features after the evaluation of final models are as follow:-

- Temp:- positive coefficient value of 0.491508
- year:- positive coefficient value of 0.233482
- winter- positive coefficient value of 0.083084

General Subjective Questions

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

Ans:-

Linear Regression is a type of Supervised Machine Learning Algorithm. It is a statistical measure which is used to find the relationship between two or more variables, where one variable is Dependent variable and other variables are independent variables. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). It is based on the straight line equation “y = mx + c”.

In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable.

Regression is broadly divided into simple linear regression and multiple linear regression.

1. Simple Linear Regression : SLR is used when the dependent variable is predicted using only one independent variable.

2. Multiple Linear Regression : MLR is used when the dependent variable is predicted using multiple independent variables.

The equation of MLR is given as:-

y = B0 + B1 * x1+ B2 * x2+ B3 * x3

where,

B0=Intercept

B1,B2,B3-Coeffieient of independent variables X1,X2, and X3.

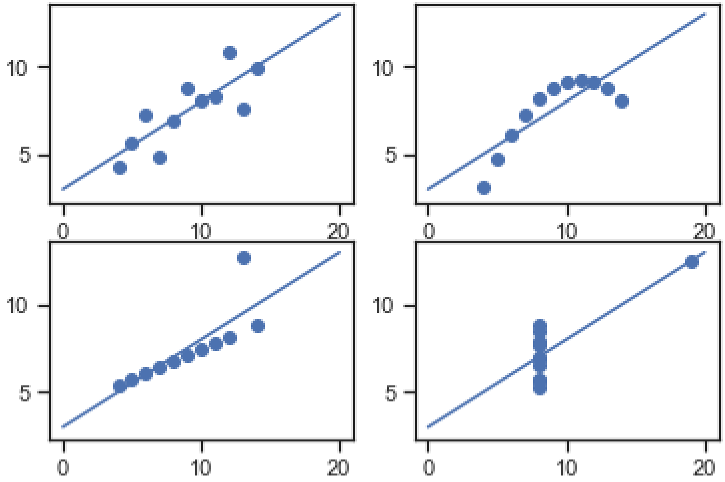
The linear regression works on the principle of Ordinary Least Square.

2. Explain the Anscombe's quartet in detail.

Ans:-

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph.

The basic thing to analyze 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 behavior irrespective of statistical analysis.



- Data-set I represents a linear relationship with some variance.

- Data-set II shows a curve shape but doesn't show a linear relationship.

- 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.

3. **What is Pearson's R?**

Ans:-

- Pearson's r is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other.

- In simple words, Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

- Its value ranges between -1 to +1 where,

  r = 1 means the data is perfectly linear with a positive slope
  r = -1 means the data is perfectly linear with a negative slope
  r = 0 means there is no linear association.

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

Ans:-

- Feature Scaling is a technique to standardize or normalize the independent features present in the data in a fixed range.
- It is performed during the data pre-processing to handle highly varying Data points or values.
- 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.

There are two types of 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.

2. **MinMax Scaling**: The variables are scaled in such a way that all the values lie between 0 and 1 using the maximum and the minimum values in the data.

- 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-square, etc.

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

<u>Ans:-</u>

- Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables.
- (VIF) =1/(1-R_1^2).
- If there is perfect correlation, then VIF = infinity.
- 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 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.

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

<u>Ans:-</u>

- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.
- A quantile is a fraction where certain values fall below that quantile.
- For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.
- The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.
- It is used in 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.
- From a Q-Q plot, we can get the following possible interpretations about the two data:-

1) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis.
2) Y-values < X-values: If y-quantiles are lower than the x-quantiles.
3) X-values < Y-values: If x-quantiles are lower than the y-quantiles.
4) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis