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?

   Answer:

   - The demand of the bikes is lowest during spring and highest during fall.
   - The demand is highest during August, September and October.
   - There is no significant change between Working day and Weekend.
   - The demand is significantly higher in 2019 than 2018
   - The median number of people looking for bikes is higher on holidays
   - There is no significant difference on median number of people booking the bike through out the week
   - There is literally no demand for bikes during heavy rain.
   - The demand for the bike is highest during clear weather
   - The demand drops drastically when there is a light rain

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

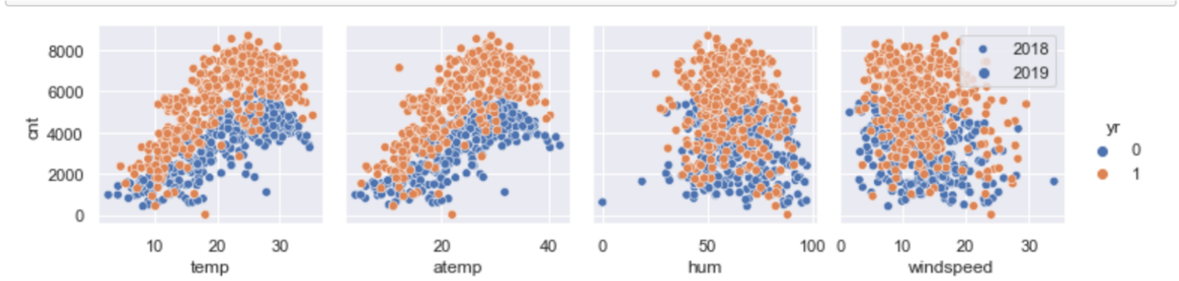
   Answer:

   We remove one of the values from the dummy variables. It is done to avoid **multicollinearity.** Consider there are 2 different options for gender, say Male and Female, If the user didn’t select Male, it becomes Female automatically. So there will be strong relationship between these 2 dummy variables. Hence, we remove one of the variables so that regression model can predict correct values.

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

<u>Answer:</u>

The variables temp, atemp has the highest corelation with the target variable Y.



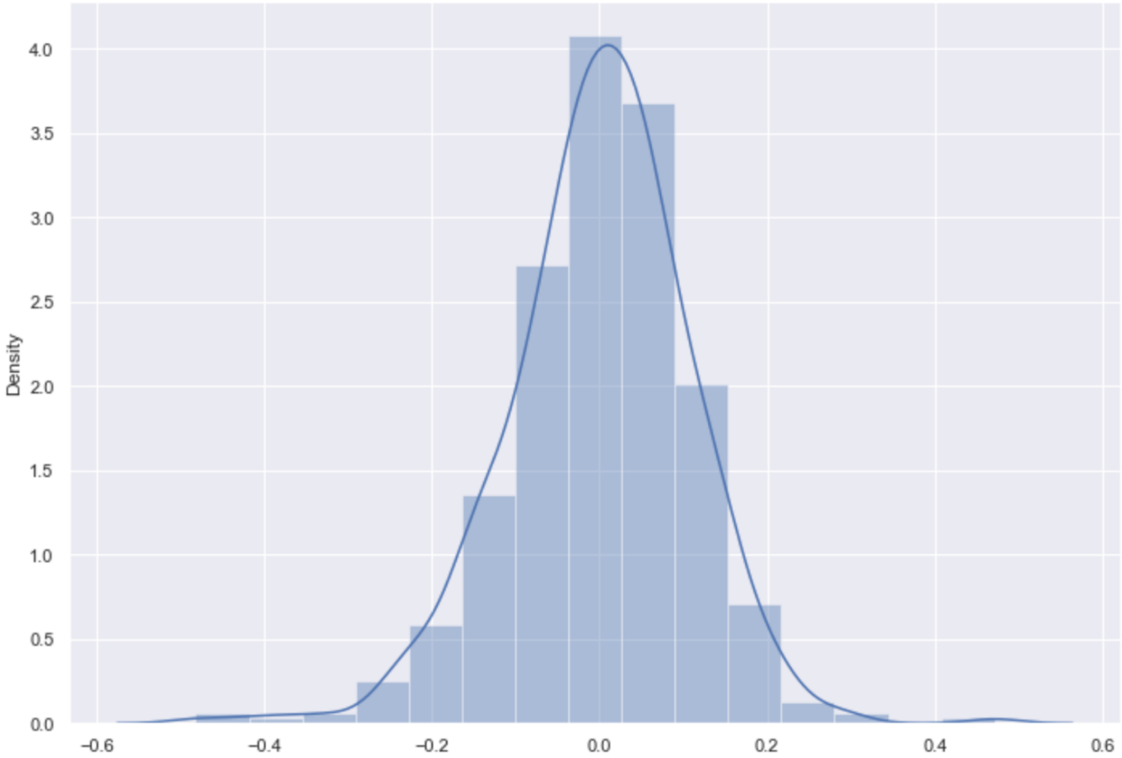
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

<u>Answer:</u>

(a) Normality of residuals:

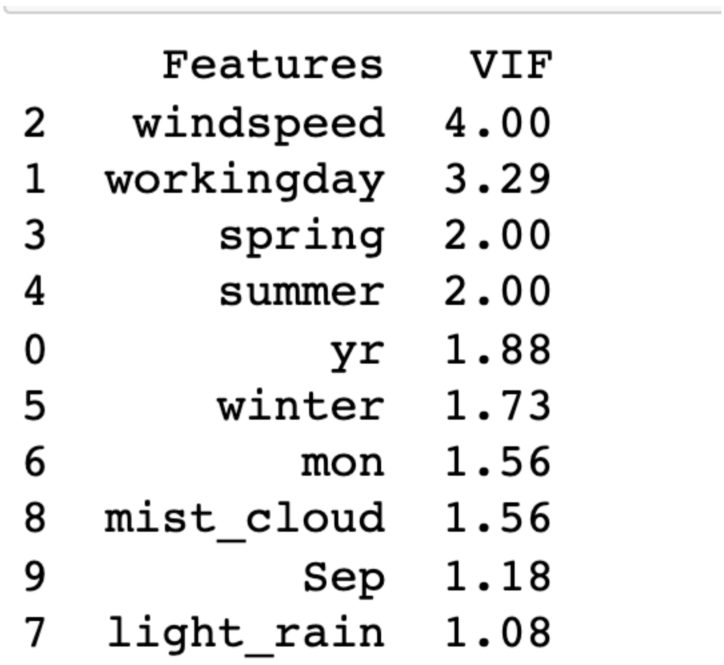
Residuals should follow normal distribution with mean centred around 0. We validated it by plotting the displot of residuals.

The below plot shows that the mean of the residual is centred around 0.



(b) Little or no Multicollinearity:

There should be little or no multi collinearity between the explanatory variables. You can observe the same by the VIF table as below.



```
      Features   VIF
2    windspeed  4.00
1   workingday  3.29
3       spring  2.00
4       summer  2.00
0           yr  1.88
5       winter  1.73
6          mon  1.56
8   mist_cloud  1.56
9          Sep  1.18
7   light_rain  1.08
```

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

   The top 3 variables are follows.

   - yr (0.247 coefficient)
   - light_rain (-0.30 coefficient)
   - spring (-0.29 coefficent)

General Subjective questions:

1. Explain the linear regression algorithm in detail.

   Linear regression is a method to predict the target value based on several independent predictor variables.

   The relationship between independent and dependant variables are determined by plotting a straight line. The straight line is generally known as regression line.

It is based on straight line equation.

y = mx+c.

Here y is the dependant variable we are trying to predict.

m is the slope of the regression line that represents the effect of X on Y

X is the independent variable we are using to make a prediction.

C is the y intercept.

In order to find the best fit line between variables, the error between the predictor and the actual values should be minimized. In other words, the best fit values are obtained by minimising the Residual sum of squares. RSS can be minimised by either differentiation or Gradient descent method.

Assumptions of Linear regression:

(i) There must be a linear relationship between the dependant and independent variables.

(ii) There must be small or no multi collinearity between the dependant and independent variables.

(iii) The error term is the same for all the variables of the independent variables. In other words, there should be .

(iv) Error terms should follow the normal distribution pattern.

(v) Linear regression assumes that there is no auto corelation between error terms.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe quartet is constructed by Francis Anscombe in 2017 to stress the importance of visualising the data (or plotting the data in a graph) before analysing it.

It consists of 4 datasets that contains nearly identical statistical properties but appear very different when plotted in a graph.

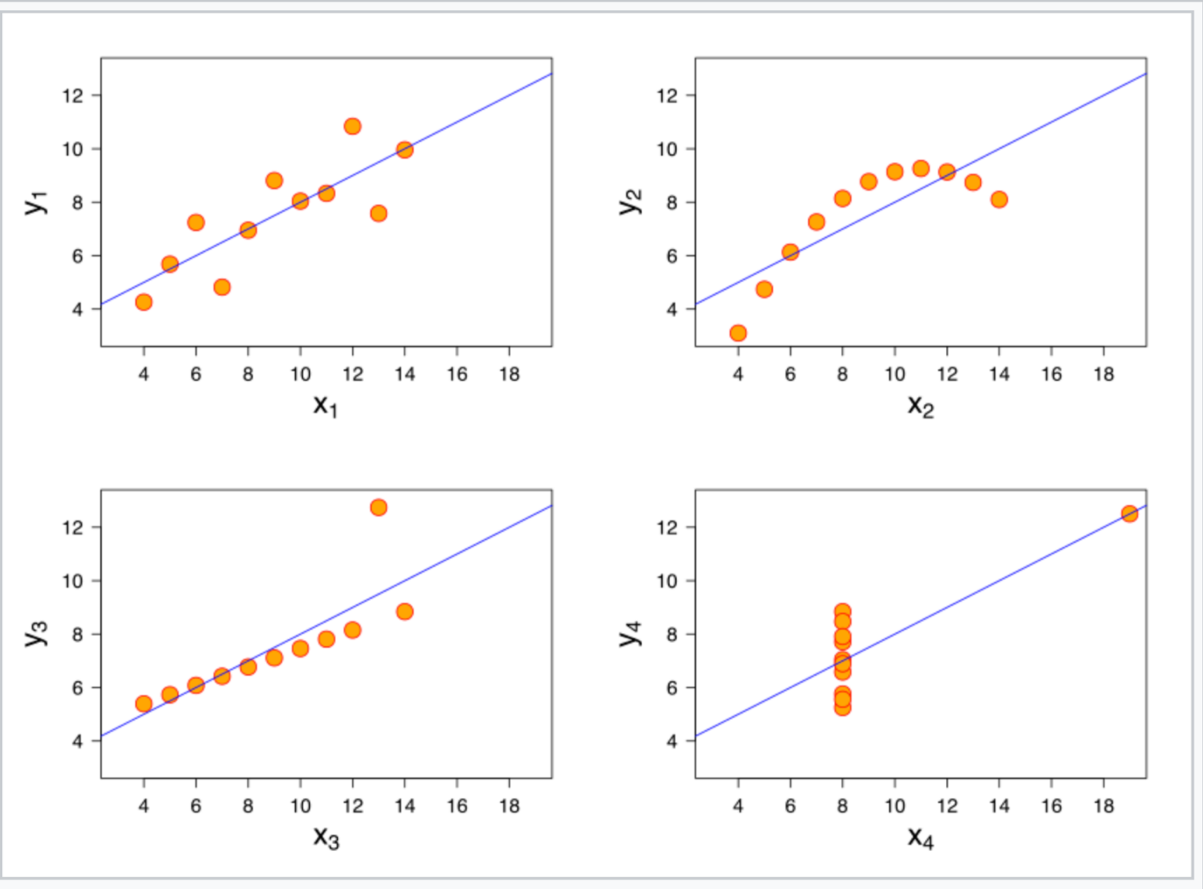


Image is taken from Wikipedia.

(i.) The graph between X1 and Y1 fits the regression model pretty well.
(ii.) The graph between X2 and Y2 is not linear. The data is not normally distributed. Though there is some relationship between the data, the relationship is not linear.
(iii.) The graph between X3 and Y3 shows a linear relationship. However, it should have different regression line. One single outlier has caused the line to be different.
(iv.) The graph between X4 and Y4 shows how one very large outlier is enough to bring in a strong corelation coefficient.

3. What is Pearson's Coefficient R.
   Answer:

   Pearson coefficient is a corelation coefficient that is used to calculate linear relationship between 2 given variables. The corelation coefficient is generally denoted by r. Its value is between -1 and +1.
   If r > 0, it means there is a positive corelation.

If r<0, it means there is a negative corelation.
If r =0, there is no linear association.

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

   Scaling is a technique used to standardize the independent variables present in the dataset within a fixed range.

Normalisation	Standardisation
Rescales the value to a range typically [0,1]	Standardisation scales the data to have a mean of 0 and standard deviation of 1
Normalisation is good if the data does not follow gaussian distribution	Standardisation is good if the data follows Gaussian distribution

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

   Answer:

   VIF becomes infinity when there is a perfect corelation. Incase of perfect corelation, we get the R2 (R square) value as 1 which lead 1/(1-R2) to become infinity. We need to drop one of the variables from the dataset that is causing the perfect multi collinearity.

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

   Answer:

   - QQ plot is used to find if a set of data came from population with similar distribution.
   - It's helpful if we receive test and training datasets separately.
   - Since we are plotting based on quantiles, sample sizes need not be equal.

- Several distribution properties like scale, location, symmetrical changes, changes in outliers can be tested simultaneously.