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?(3marks)

Answer:-

- Bike Rentals are more during the fall season and then in summer

- Almost 99% of the bike bookings increased in year with median of close to previous year booking (for the period of 2 years). This indicates, yr can be a good predictor for the dependent variable

- Almost 97.6% of the bike booking were happening when it is not a holiday, which means this data, is clearly biased. This shows holiday cannot be a good predictor for the dependent variable.

- Almost 10% of the bike booking were happening in the months may, jun, jul, aug & sept with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
- Bike Rentals are more on Saturday,wednesday and thursday

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

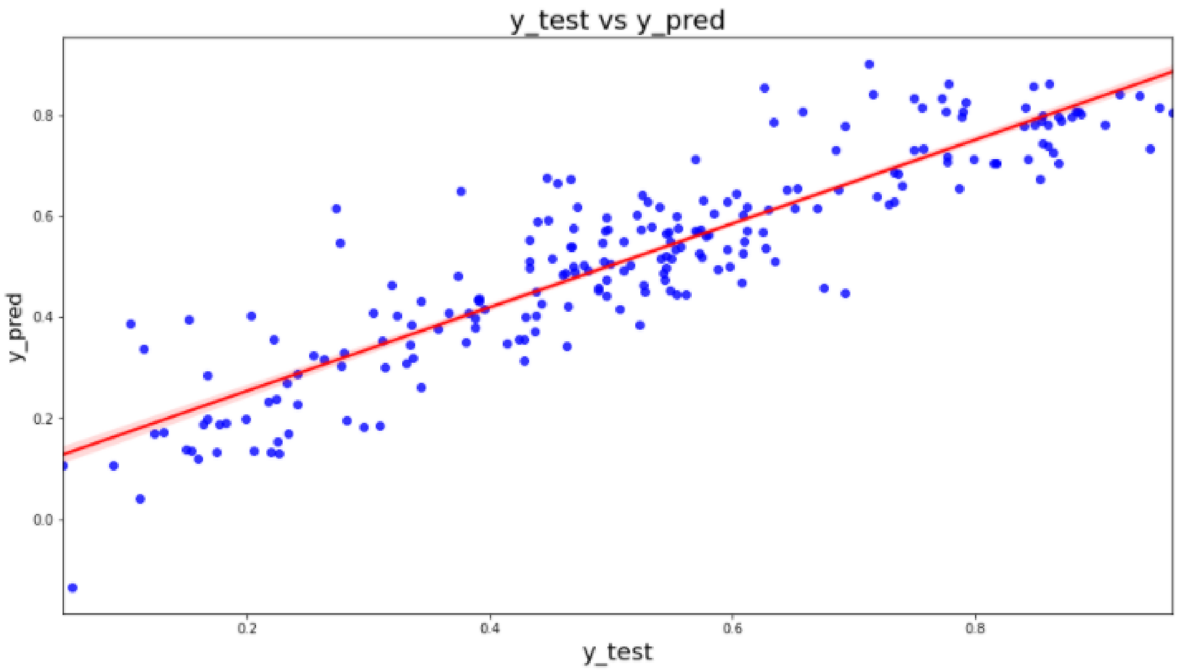
Answer:- It avoids the multi collinearity among the categorical variables. If a categorical variable has n levels, we need to represent the dummy variables in n-1 levels as with the knowledge of n levels, we can deduce the status of n-1th level and it does not need to be mentioned in a new column of dummy variable.

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

Answer:- Registered as the highest correlation with count(target)

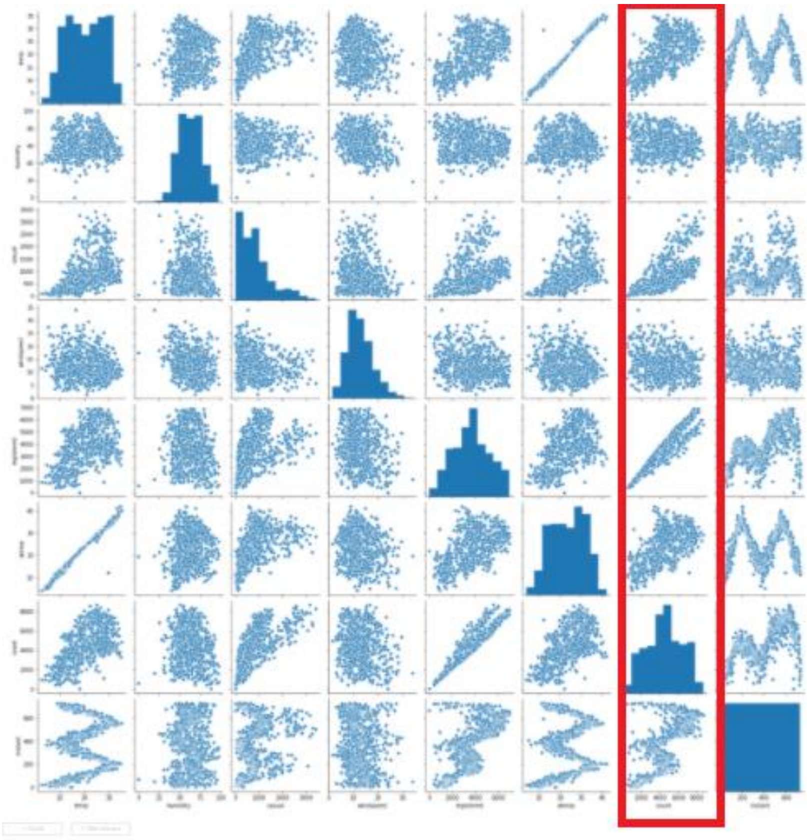
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

→ a) The error terms after the model should have constant variance



From the figure above we can see that the density of true observations are evenly spread out across the regression line hence it satisfies the **homoscedasticity**

b) **Linear and Additive** :- There is almost linear inclination between the target variable (count) and most of the independent variables. It is shown in pair plots below

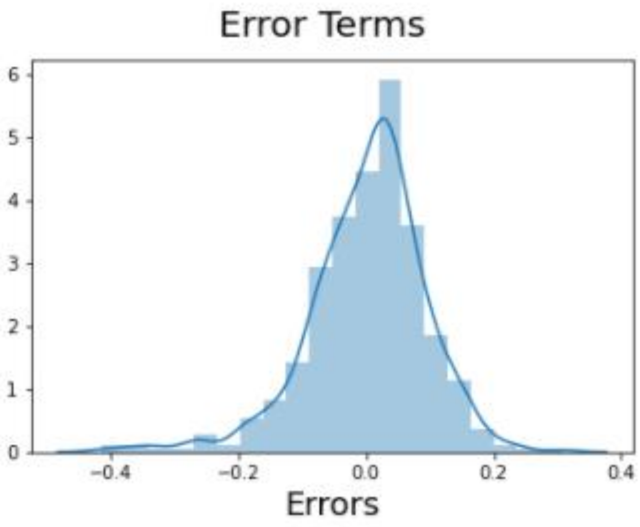


c) The VIF is checked after every iteration, which eliminated the **multicollinearity** which might have existed. While creating dummy variables n-1 levels were taken. This also stopped unnecessary **multicollinearity**.

d) **Autocorrelation**:-

Durbin-Watson parameter was evaluated for each REF stage and 2<DW<4 is followed to check if there is any autocorrelation.

e) Error terms are normally distributed



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

Answer:-

1)Weather plays a major role wherein “Light Snow” puts negative influence on the sale of bikes.

2)Season also has quite an impact on the sale of shared bikes with (Summer and Spring) putting positive weight on model.

3)Month (September) has positive weight on the model

General Subjective Questions

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

Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.

Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called Ordinary Least Squares. It is common to therefore refer to a model prepared this way as Ordinary Least Squares Linear Regression or just Least Squares Regression

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

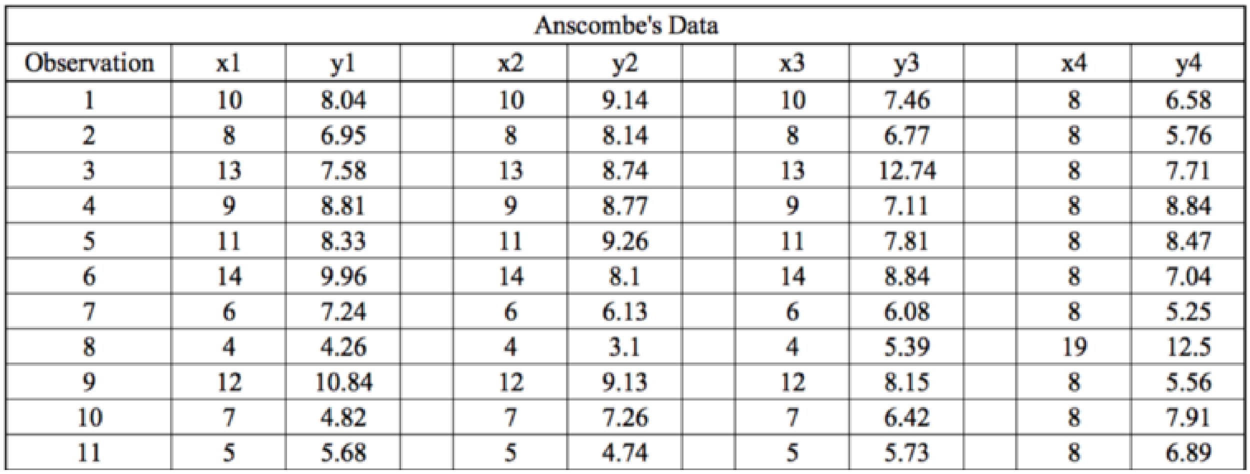
y = B0 + B1*x

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).

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

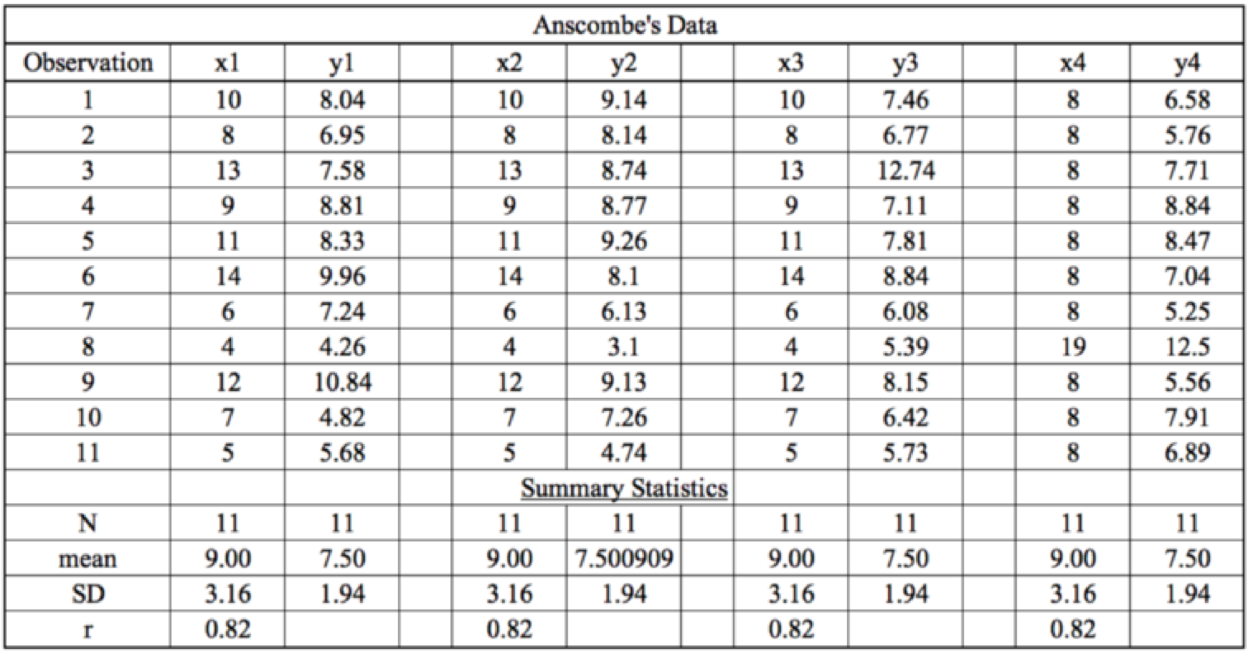
Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

With example



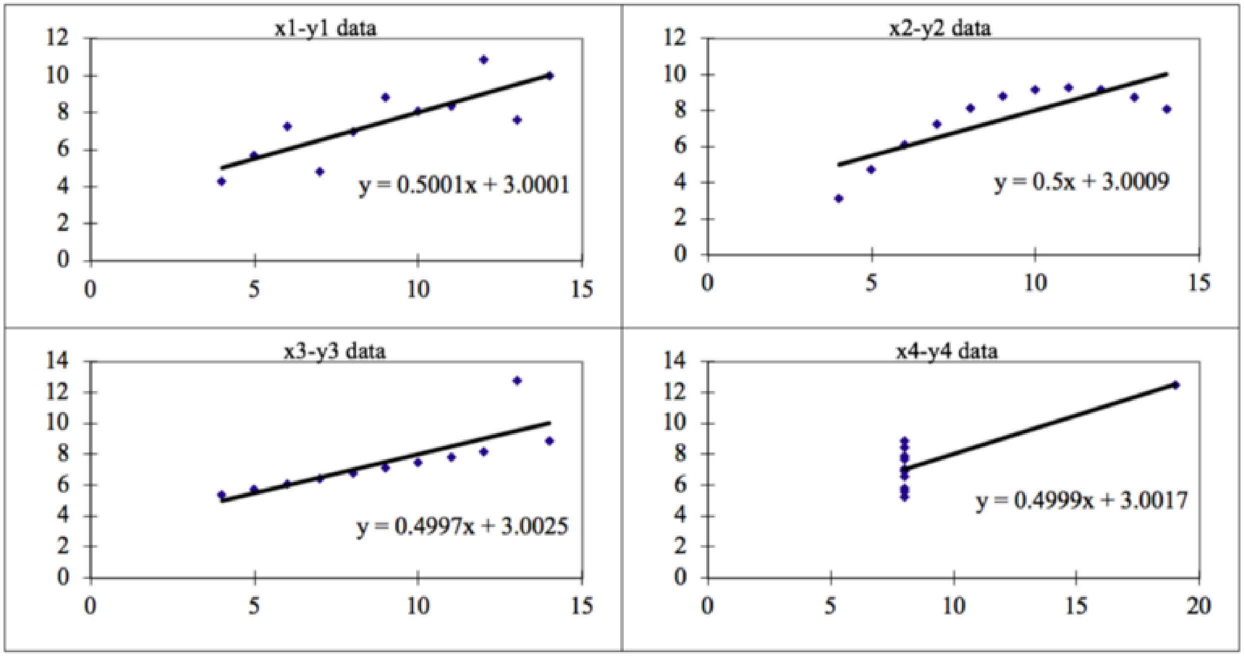
Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

Above table will give below descriptive statistics



Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

The stats for above data is almost similar but the regression will be different



In a way the data has fooled the regression and linear regression is not supposed to be used in the x2 x3 and x4.

IMAGES TAKEN FROM (towardsdatascience.com)

3.What is Pearson’s R?(3marks)

Pearson correlation coefficient or Pearson’s correlation coefficient or 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. Pearson’s correlation coefficient calculates the effect of change in one variable when the other variable changes.

The correlation coefficient formula finds out the relation between the variables. It returns the values between -1 and 1. Use the below Pearson coefficient correlation calculator to measure the strength of two variables.

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2- (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

Pearson correlation coefficient or the Pearson coefficient correlation r, determines the strength of the linear relationship between two variables. The stronger the association between the two variables, the closer your answer will incline towards 1 or -1. Attaining values of 1 or -1 signify that all the data points are plotted on the straight line of ‘best fit. The closer the answer lies near 0, the more the variation in the variables.

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

Answer:-

Scaling is the process of fitting or mapping numerical observations into a limited set or limited range of observations. This is used to speed up the processing power of algorithms

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X’ is 0

- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1
- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation

Mu is the mean of the feature values and sigma is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range

Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution

Standardization is mostly used where the data follows a Gaussian 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?(3marks)

Answer:- Infinite VIF shows that the variable can be expressed exactly by the linear combination of other variables.

VIF = 1 /(1 – R2). When R2 reaches 1, VIF reaches infinity. Which shows perfect collinearity

An R2 of 1 indicates that the regression predictions perfectly fit the data.

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

Answer:- The q-q or quantile-quantile is a scatter plot which helps us validate the assumption of normal distribution in a data set. Using this plot, we can infer if the data comes from a normal distribution. If yes, the plot would show straight line. Absence of normality in the errors can be seen with deviation in the straight line.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.