ASSIGNMENT: LINEAR REGRESSION SUBJECTIVE QUESTIONS

**Declaration: Following document has the response/answers for the questions shared as part Linear Regression assignment. So, in this document 2 sections have been created which are catering the answers to be submitted along-with the case study as part of assignment.*

ASSIGNMENT BASED SUBJECTIVE QUESTIONS:

Q1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans1: Analysis of the categorical variables from the dataset verified in the box plots:

- Season box plots → indicates that more bikes are rented during fall season
- Weather situation box plot → indicates that more bikes are rented on clear, partly clouded weather situation
- Year box plot trends show that more bikes were rented in the year 2019
- Working day & Holiday box plots shows that more bikes are rented on normal working days than on weekends or holidays
- Month box plot indicates → Higher number of bikes rented in the month of September

Q2: Why is it important to use drop_first=True during dummy variable creation?

Ans2: drop_first=True, is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. For example, is we have 3 types of values in Categorical column on house furnished status related. And we want to create dummy variable for that column. If variables value is not indicated as "furnished" or "semi-furnished", then it's understood as unfurnished. So, 3rd variable to identify the unfurnished status is not needed here.

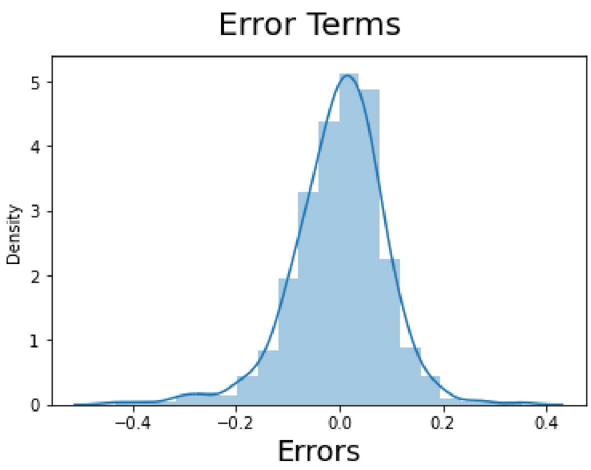
Therefore, if we have categorical variable with n-levels, then we use n-1 columns to represent the dummy variables.

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans3: Pair plot establishes that variable "registered" user has the highest correlation with target variable "cnt" i.e., "0.95" followed by "casual" users & "temp".

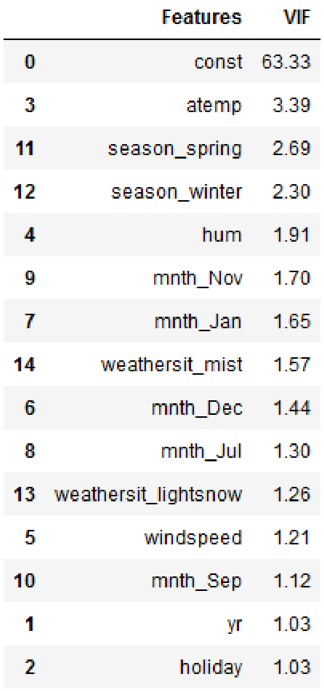
Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans4: Error distribution range has been analyzed.



Error terms is distributed normally with mean 0

Other way to see that using VIF value



	Features	VIF
0	const	63.33
3	atemp	3.39
11	season_spring	2.69
12	season_winter	2.30
4	hum	1.91
9	mnth_Nov	1.70
7	mnth_Jan	1.65
14	weathersit_mist	1.57
6	mnth_Dec	1.44
8	mnth_Jul	1.30
13	weathersit_lightsnow	1.26
5	windspeed	1.21
10	mnth_Sep	1.12
1	yr	1.03
2	holiday	1.03

The VIF are below 5 so there is no multi collinearity exists between predictor variables

Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans5: Based on the final model, following are the 3 features contributing significantly towards the demand of shared bikes:

- "temp" (Positive correlation)

- “yr” (Positive correlation)
- “Weathersit” as “Thunder” (declared as - df_day.loc[(df_day['weathersit'] == 3) , 'weathersit'] = 'Thunder') is negatively correlated

ASSIGNMENT GENERAL SUBJECTIVE QUESTIONS:

Q1: Explain the linear regression algorithm in detail.

Ans1: Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

y = a + bx

Where a and b given by the formulas:

b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^{2}-(\sum x)^{2}}

a(intercept) = \frac{n\sum y - b(\sum x)}{n}

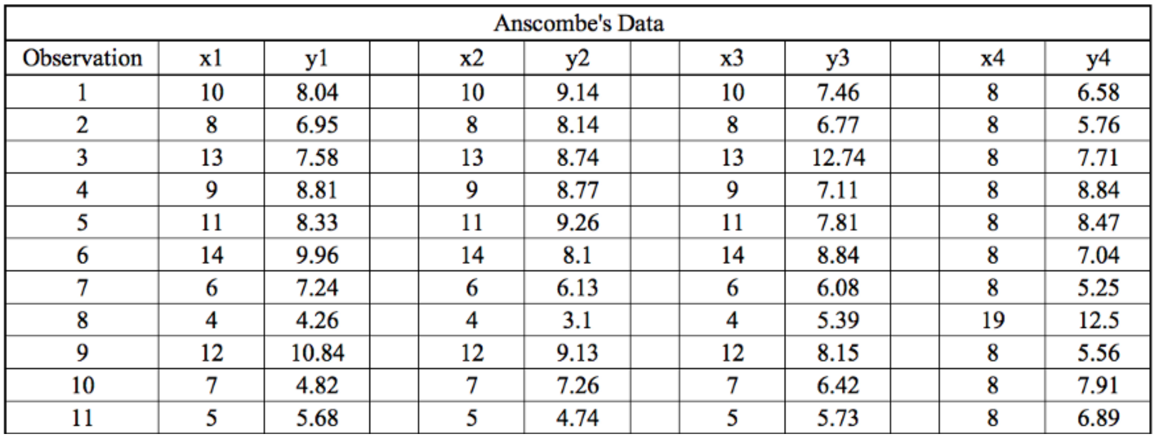
Here, x and y are two variables on the regression line.

b = Slope of the line
a = y-intercept of the line
x = Independent variable from dataset
y = Dependent variable from dataset

Q2: Explain the Anscombe’s quartet in detail.

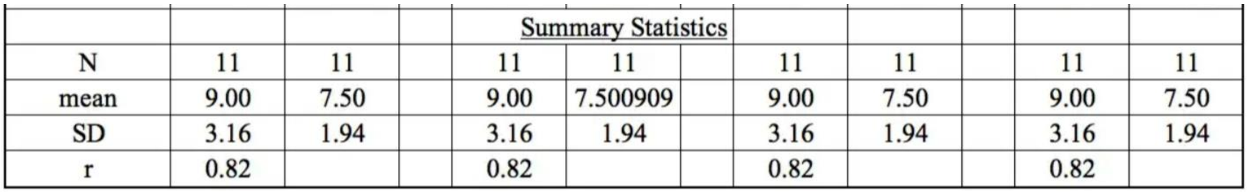
Ans2: Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.



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

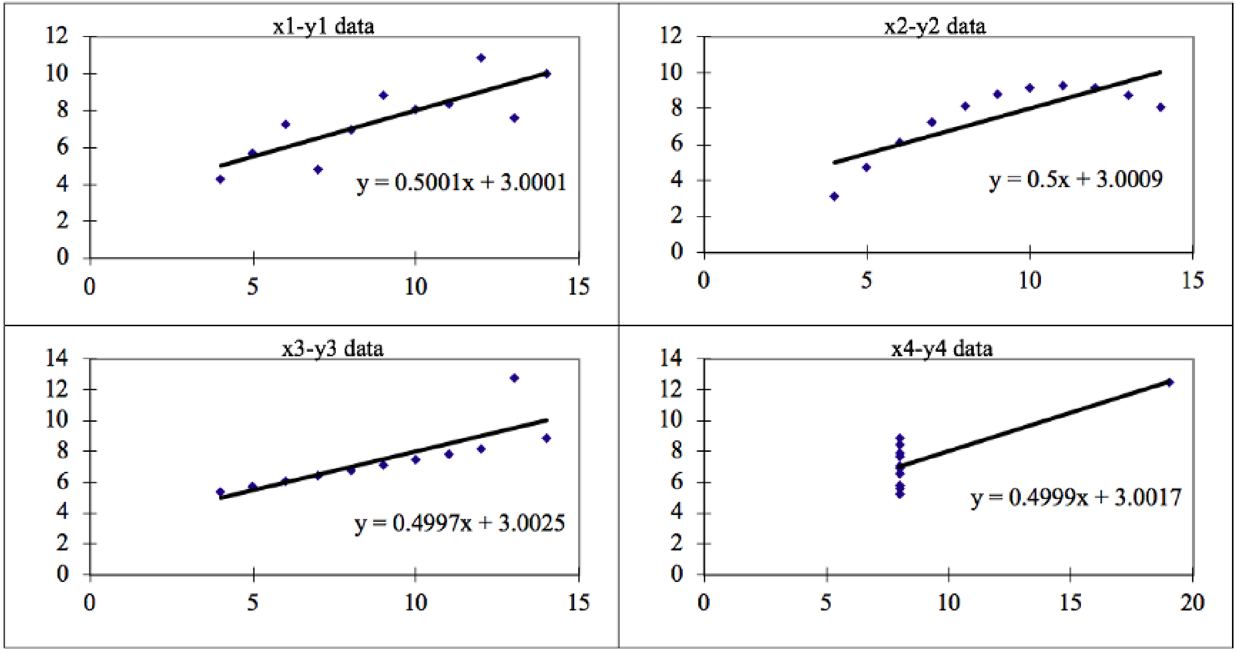
The statistical information for all these four datasets are approximately similar and can be computed as follows:



				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	

When these models are plotted on a scatter plot, all datasets generate different kind of plots which are not interpretable by any regression algorithm (can be seen as follows):

Anscombe's Quartet



The four datasets can be described as:

Dataset 1: Fits Linear regression model.
Dataset 2: Do not fit Linear regression model on the data quite well as the data is non-linear.
Dataset 3 & 4: Shows the outliers involved in dataset which are not handled by Linear regression model

Q3: What is Pearson’s R?

Ans3: In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

Pearson r Formula:

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

- r=correlation coefficient
- x_{i}=values of the x-variable in a sample
- \bar{x}=mean of the values of the x-variable
- y_{i}=values of the y-variable in a sample
- \bar{y}=mean of the values of the y-variable

Q4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans4: Scaling is a step, in data pre-processing stage which is applied to independent variables to normalize the data within a particular range. Most of the times, collected data set contains features highly varying in magnitudes, units, and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

Scaling also helps in speeding up the calculations in an algorithm.

**Note: 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-squared, etc.*

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Normalization	Standardization
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

Q5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

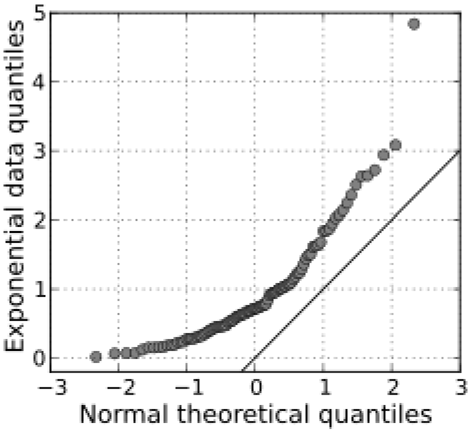
Ans5: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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).

Q6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

Ans6: 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 QQ plots is to find out if two sets of data come from the same distribution. A 45 degrees angle is plotted on the QQ plot; if the two data sets come from a common distribution, the points will fall on that reference line.

Q Q plot showing the 45 degrees reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.