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?

ANS: There are 7 categorical variables in the dataset. Let's look at the effect of the variable on each of them.

- The demand for shared bikes is highest during Fall/Autumn and Summer as it's the best season to travel through bikes
- The demand for the bikes has increased significantly in 2019 compared to 2018
- The demand for bikes on the average are high from May till Oct
- The demand seems to be higher during non-holiday days which implies people love to spend time at home or like to drive cars during holidays. It also implies that the demand for bikes is for office-going/commercial purposes
- There is almost no effect on the count of bikes shared during weekdays as they all seem to be approximately same
- The demand for shared bikes is on an average same for working days and non-working days
- The demand is highest during if the weather is Clear, having few clouds and partially cloudy

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

ANS: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation and reduces the correlations created among dummy variables.

For e.g., Let's say we have 3 levels of values of categorical variables for weathersit. These are nominal variables and the levels do not represent any order in them.

Weathersit
Misty
Clear
Cloudy

So, we drop the first column while creating dummy variables

Weathersit_Clear	Weathersit_Cloudy
0	0
0	1
1	0

So, when the values for Clear and Cloudy are 0, it should be "Misty" which will have the value 1 which is not specified explicitly.

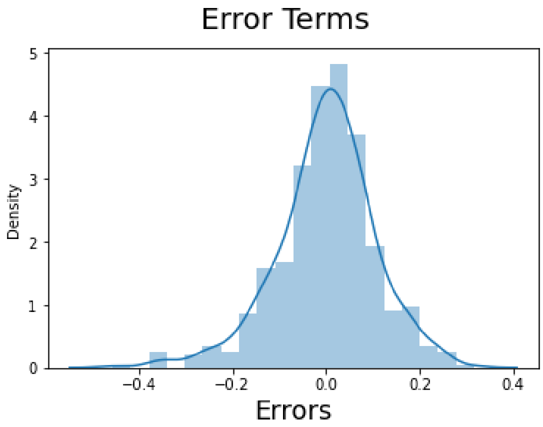
When we have n levels of categories in a categorical variable, n-1 dummies are created.

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

ANS: The numeric variable 'registered' is the highest correlated with the target variable followed by 'casual', we know that the sum of these two variables value onto 'cnt', our target variable. So these two variables should be dropped and should not be considered for our model building.

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

ANS: After building the model, we do residual analysis and calculate the error terms by plotting a histogram of the residuals to see if the error terms follow a normal distribution or not and see if the mean value of the error terms is equal to 0. This way we validate our assumptions of Linear Regression. Given below the histogram of the residuals for the model we built.



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

ANS: Based on the final model, the following features contribute significantly towards explaining the demand of shared bikes:

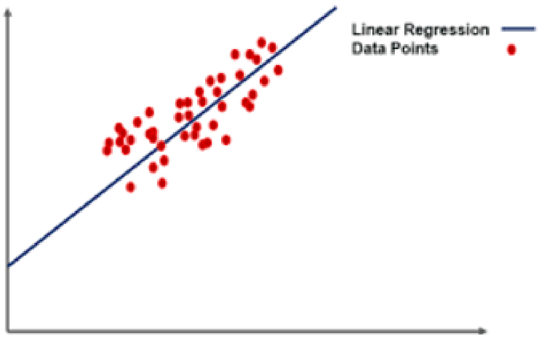
- The variable ‘temp’ has a coefficient of ‘0.5854' representing unit increase in target variable, increases the shared bikes numbers by 0.5854 units
- The variable ‘yr’ has a coefficient of ‘0.2329’ representing unit increase in target variable, increases the shared bikes numbers by 0.2329 units
- The variable ‘weathersit_Snowy/Rainy’ has a coefficient of ‘0.2524’ representing unit increase in target variable, decreases the shared bikes numbers by 0.2524 units
- The variable ‘holiday’ has a coefficient of ‘0.0875’ representing unit increase in target variable, decreases the shared bikes numbers by 0.0875 units

GENERAL SUBJECTIVE QUESTIONS:

1. Explain Linear Regression algorithm in detail.

ANS: Linear Regression is a machine learning algorithm under Supervised learning. Linear Regression shows linear relationship between the independent variable(s) and the dependent variable. It is used for predictive analysis and shows the relationship between the continuous variables. There are two types of Linear Regression:

- ***Simple Linear Regression*** *– It attempts to explain the relationship between one independent and one dependent variable using a straight line*
- ***Multiple Linear Regression*** *– It explains the relationship between more than one independent variable and one dependent variable using a straight line*



Mathematically we can represent linear regression as:

$y = mx + c$ (or)

$y = \beta_0 + \beta_1 x$

Where y is the dependent variable;

x is the independent variable

β_1 is the slope of the straight line

β_0 – intercept of the line

BEST-FIT LINE

In regression, the main goal is to find the best fit line which is the line which fits the given scatter plot of datapoints in the best way.,i.e., the error between predicted values and actual values should be minimized. The best fit line will have the least error

Best fit line is obtained by minimizing a quantity called Residual Sum of Squares(RSS) – a line that reduced the sum of squares of residuals.

The different values for weights or the coefficient of lines (β_0, β_1) gives a different line of regression, so we need to calculate the best values for β_0 & β_1 to find the best fit line, so to calculate this we use cost function.

COST FUNCTION

Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.

For Linear Regression, we use the Mean Squared Error (MSE) cost function, which is the average of squared error occurred between the predicted values and actual values. A regression model uses gradient descent to update the coefficients of the line by reducing the cost function. It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

MODEL EVALUATION:

R-squared method: It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.

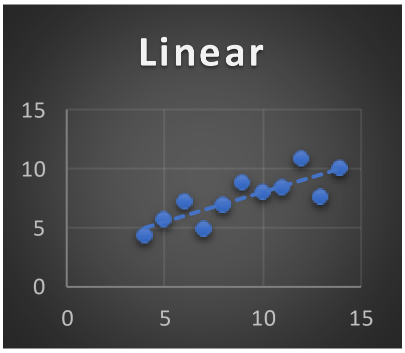
R-squared = 1 – TSS/RSS

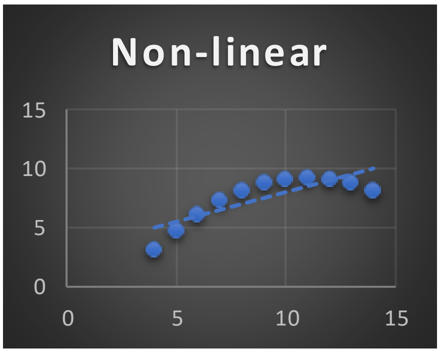
2. Explain Anscombe’s quartet in detail.

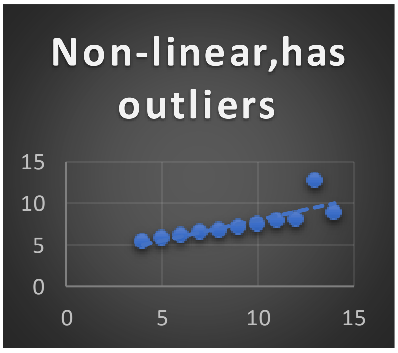
ANS: Anscombe’s quartet is a group of four datasets which nearly have same statiscal observations but the datapoints are distributed differently. It emphasizes the importance of visualizing a dataset before building a model. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.)

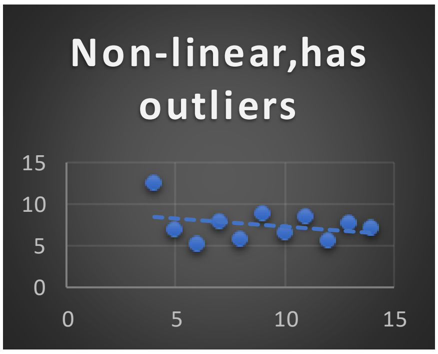
For e.g., consider the below data points:

OBS	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		10	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		13	7.71
4	9	8.81		9	8.77		9	7.11		9	8.84
5	11	8.33		11	9.26		11	7.81		11	8.47
6	14	9.96		14	8.1		14	8.84		14	7.04
7	6	7.24		6	6.13		6	6.08		6	5.25
8	4	4.26		4	3.1		4	5.39		4	12.5
9	12	10.84		12	9.13		12	8.15		12	5.56
10	7	4.82		7	7.26		7	6.42		7	7.91
11	5	5.68		5	4.74		5	5.73		5	6.89
SUMMARY STATISTICS											
N	11	11		11	11		11	11		11	11
mean	9	7.5		9	7.5		9	7.5		9	7.5
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	







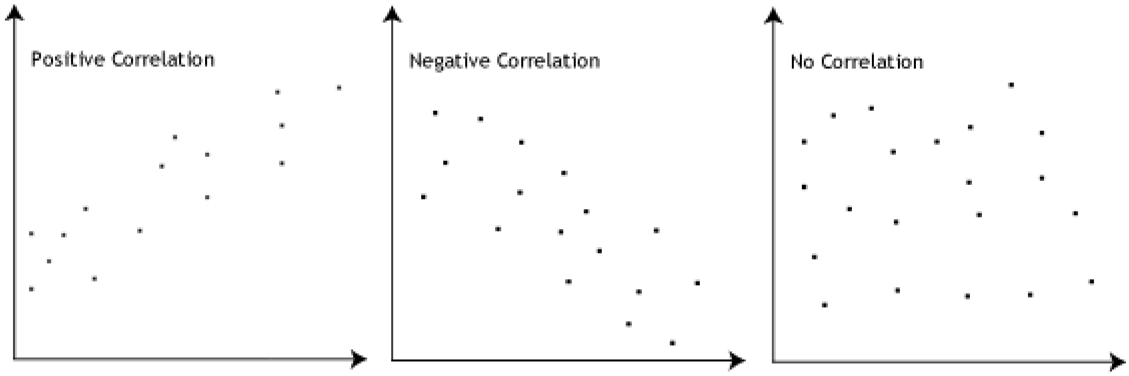


3. What is a Pearson's R?

ANS: It is a measure of the strength of a linear association between two variables and is denoted by r. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

The Pearson correlation coefficient, r, can take a range of values from +1 to -1.

- A value of 0 indicates that there is no association between the two variables.
- A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.
- A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.



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

ANS: Scaling is a pre-processing step applied to variables to normalize the data within a particular range. It also helps in speeding up the calculations in the algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and ranges. If scaling is not done, then algorithm only takes magnitude into account and not units hence models incorrectly. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-stat, F-stat, p-values, r-squared.

There are two types of Scaling:

- **MinMax / Normalization:** It brings all of the data in the range of 0 and 1. It is mathematically represented as $x = \frac{x - min(x)}{max(x) - min(x)}$
- **Standardization Scaling:** It replaces the values by their z-scores. It brings all the data into a standard normal distribution which has mean as zero and Standard Deviation as one. It is mathematically represented as $x = \frac{x - mean(x)}{sd(x)}$

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

ANS: 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.

For e.g., consider in our dataset, the variables ‘atemp’ and ‘temp’. Both variables are highly correlated and that’s the reason we had to drop ‘atemp’ to avoid 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).

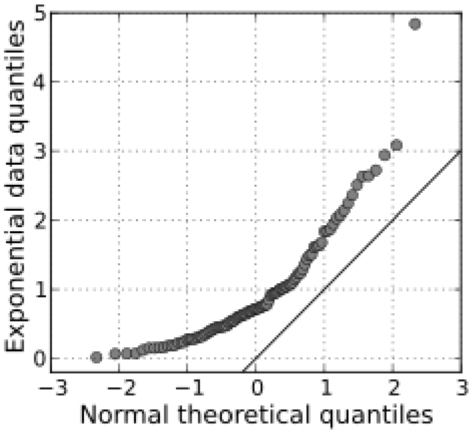
6. **What’s a Q-Q plot, explain the use and importance of a Q-Q plot in Linear Regression?**
ANS: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

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.



THANK YOU,

by, Yoga Sree Durga K