

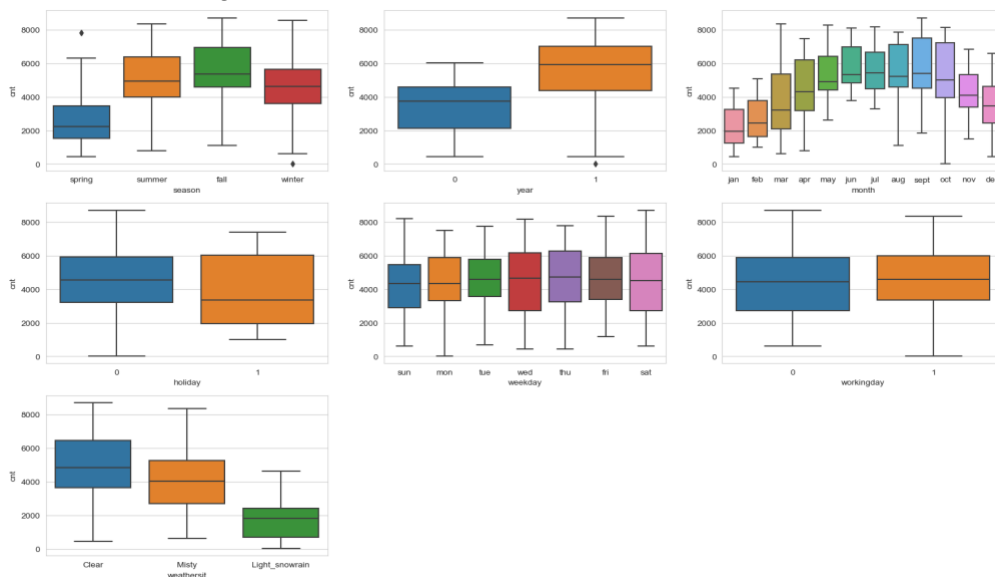
Assignment-based Subjective Questions

by Shrinivas Nandkishor Dhamone

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

Answer: According to the analysis done there are these categorical variables present in the dataset, “season”, “yr”, “mnth”, “holiday”, “weekday”, “workingday”, “weathersit”. And from the model chosen for prediction these are the categorical variables with their effects on the dependent variable:

1. mnth - This categorical variable represents 12 months from Jan to Dec and according to analysis September had the highest demand also in November and December the demand is least which is totally logical as the weather situation is severe as these months have heavy snow fall.
2. yr - This categorical variable represents two years 2018 and 2019 and according to analysis the demand was more in 2019 than in 2018 maybe because the company gained more popularity across the year 2018 so business increased in 2019.
3. holiday - This categorical variable represents whether it is a holiday or not, and the analysis shows that during holiday the demand increases, and that might be due to people going for outing etc.
4. season - This categorical represents different seasons namely, spring, winter, summer, fall, according to analysis the demand is the least in spring and the most in fall and its almost medium in summer and winter, the company should focus on increasing their sales in fall.
5. weathersit – This variable represents the weather condition i.e., Clear, Misty, Light_snowrain
6. Heavy_snowrain, from the analysis it was visible that during good weather condition i.e., Clear, few clouds, partly cloudy, partly cloudy the demand is highest and during bad weather condition (Light_snowrain / Heavy_snowrain) the demand is lowest.
7. BoxPlots for Catg Variables



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

Answer: We use **drop_first=True** to drop unnecessary columns because we just need p-1 levels for p levels of a categorical variable. Example- For a variable say, 'Relationship' with three levels namely, 'Single', 'In a relationship', and 'Married', you would create a dummy table like the following:

Relationship Status	Single	In a relationship	Married
Single	1	0	0
In a relationship	0	1	0
Married	0	0	1

But you can clearly see that there is no need of defining three different levels. If you drop a level, say 'Single', you would still be able to explain the three levels.

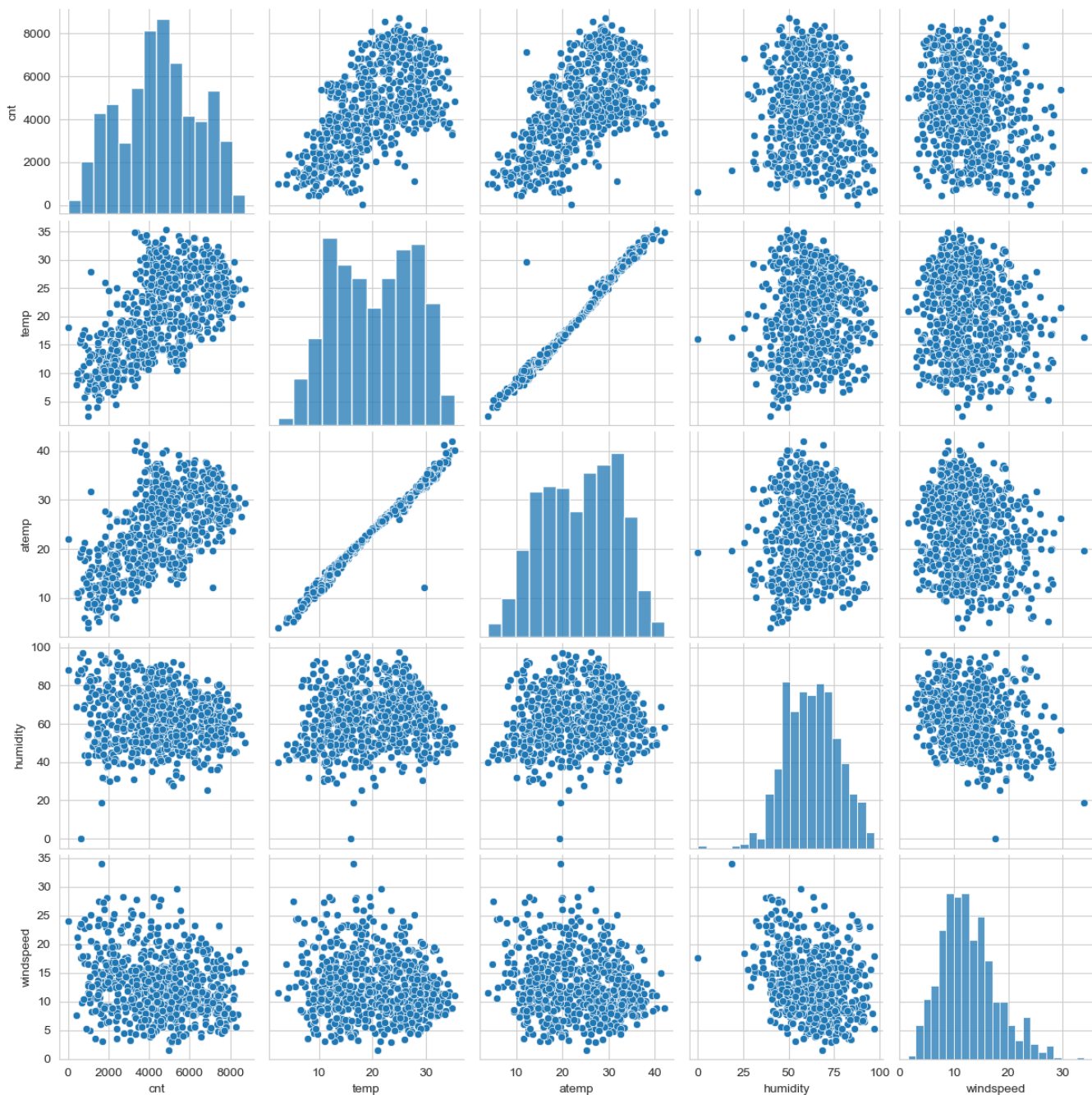
Let's drop the dummy variable 'Single' from the columns and see what the table looks like:

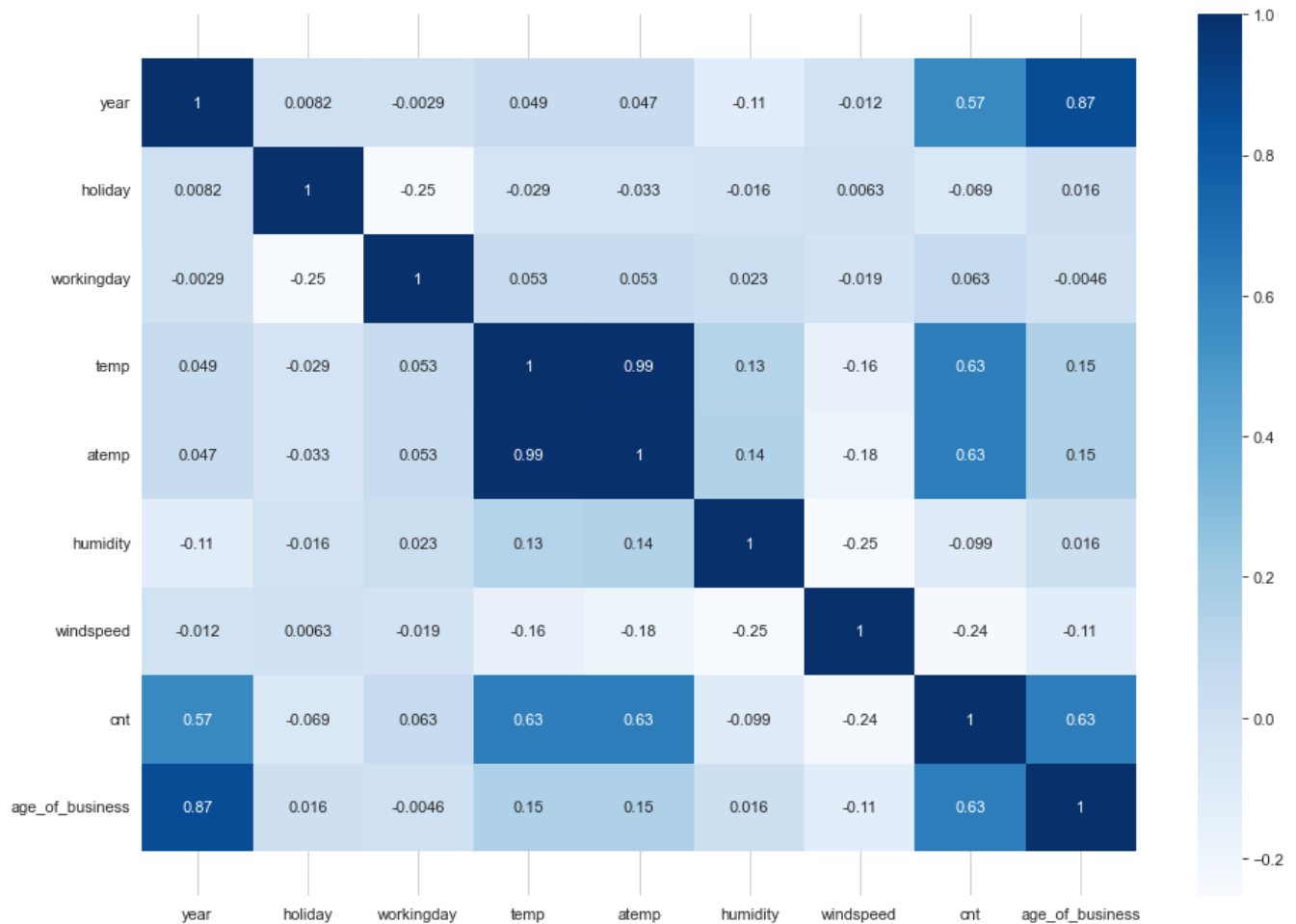
Relationship Status	In a relationship	Married
Single	0	0
In a relationship	1	0
Married	0	1

If both the dummy variables namely 'In a relationship' and 'Married' are equal to zero, that means that the person is single. If 'In a relationship' is one and 'Married' is zero, that means that the person is in a relationship and finally, if 'In a relationship' is zero and 'Married' is 1, that means that the person is married.

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

Answer: From the given graph, we can clearly see that “temp” and “atemp” are the two variables which show highest correlation with the variable “cnt” which is the target variable and this can also be seen from the heat map as show below. From correlation heatmap value is 0.63.



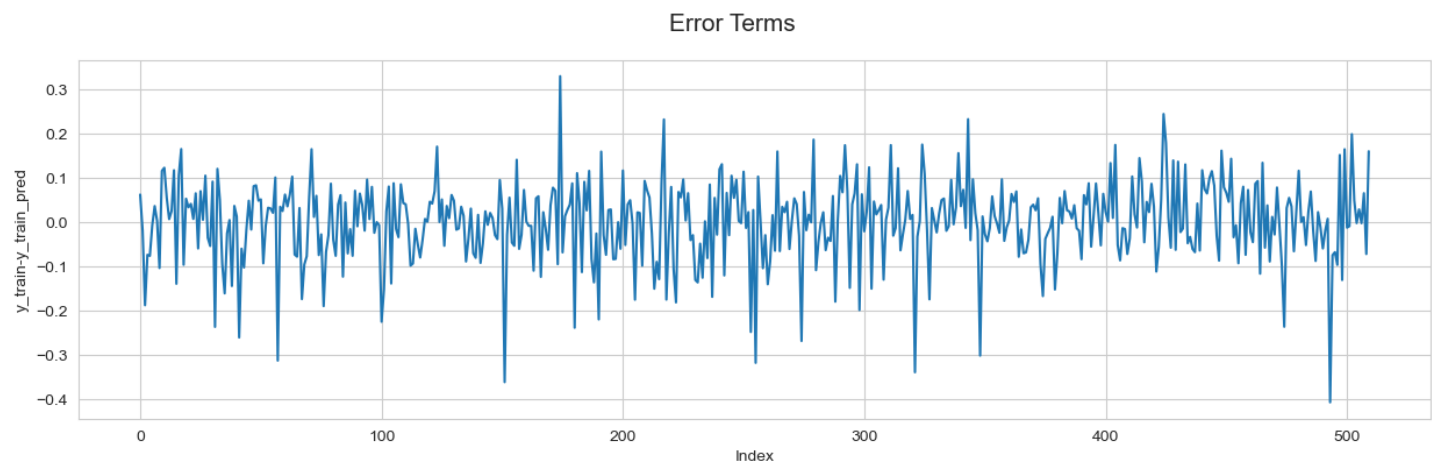
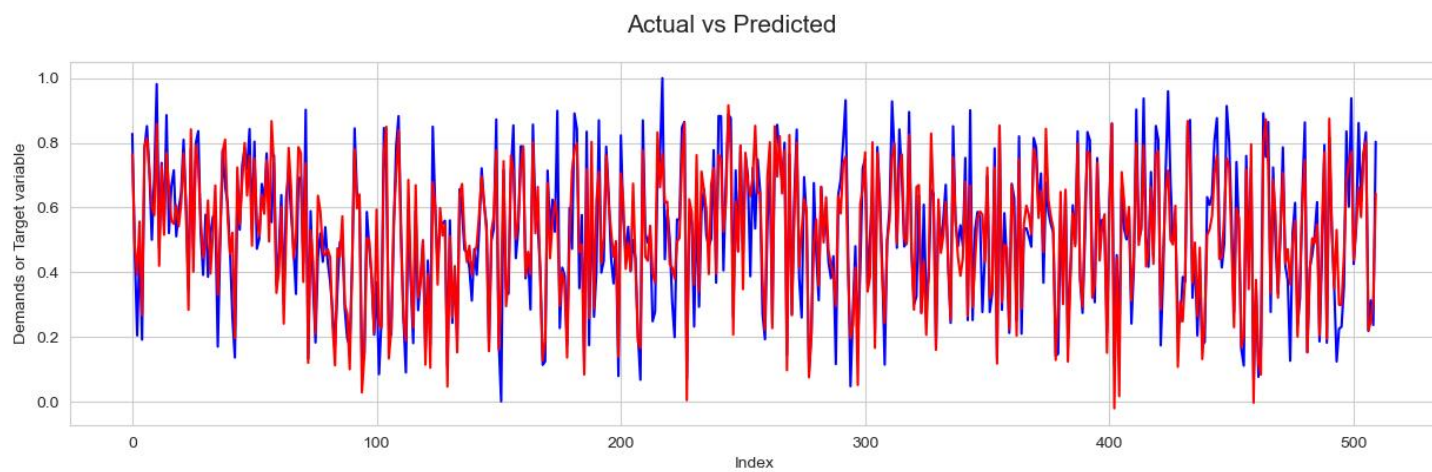
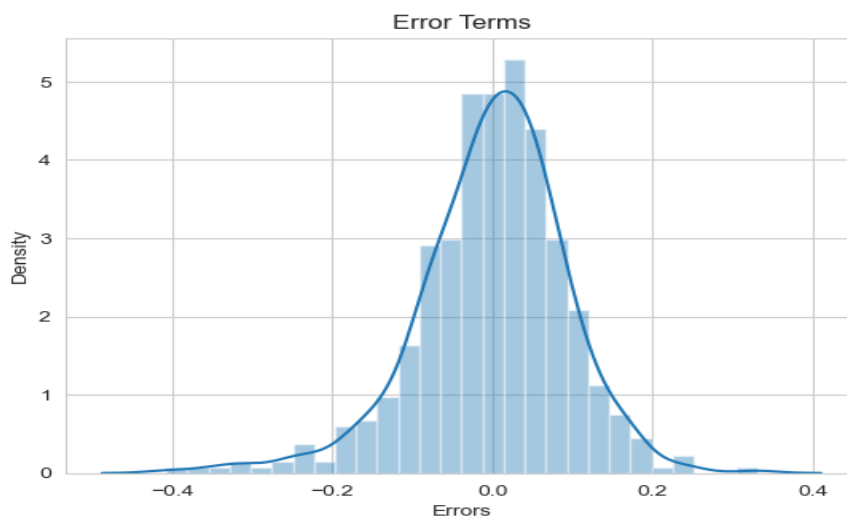


Q.4 How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: I did the following.

- If a model is good, it should have its residuals normally distributed and should be centred at 0 and also should be independent of each other as seen from the below two graphs:
- We can also prove that a model is valid by plotting the actual v/s predicted line graph and check how well they overlap as given below:
- And it's also safe and full proof to compare the r-squared value that we got for our model and the actual value.
- VIF of the features of the final model < 5.0
- Multicollinearity - Variables like temp is linearly related to cnt d.

f. Homoscedasticity – The variance of the residual is almost constant.



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

Answer: Coefficients of the final model are below

	Variables	Coefficient value
index		
3	temp	0.548008
1	year	0.232861
6	season_winter	0.129345
0	const	0.125926
7	month_sept	0.101195
5	season_summer	0.088080
8	weathersit_Misty	-0.078375
2	holiday	-0.098685
4	windspeed	-0.153246
9	weathersit_Light_snowrain	-0.282869

As per our final Model 7, the top 3 predictor variables that affects the demand or target variable are:

1. Temperature (temp) - A coefficient value of '0.5480' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5480 units.
2. weathersit_Light_snowrain - A coefficient value of '-0.2828' indicates that with unit increase in weathersit_Light_snowrain the demand of bikes will decrease by 0.2828 units.
3. Year (yr) - A coefficient value of '0.2328' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2328 units.

General Subjective Questions

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

Answer: Linear regression is a Supervised Machine learning algorithm in which the output variable is continuous in nature. Idea here is to predict the value of a dependent variable based on different independent variables.

It is of 2 types:

1. Simple linear regression – Single variable is involved

$$y = mx + c$$

y – dependent variable which needs to be predicted x – independent variable

c – intercept

m- slope or coefficient of dependent variable x.

2. Multiple linear regression – More than one variable is involved and hence equation changes to

$$y = m_1x_1 + m_2x_2 + m_3x_3 + \dots + m_nx_n + c$$

y – dependent variable which needs to be predicted x_i – independent variables

c – intercept

m_i - slope or coefficient of dependent variable x_i .

To Find the dependent variable we have below steps:

Step 1. Reading, understanding and Visualizing the data

- a. Here we read the data and look for its statistics values like mean, max, min etc.
- b. We visualize the data using pairplot for continuous variables and boxplot for categorical variables

Step 2. Prepare the data for modelling (test/train split, rescaling etc)

- a. Encoding
 - a. Convert binary categorical variables to 0 and 1
 - b. Convert other categorical variables to dummy variables
 - i. For a variable with p levels we will need p-1 dummy variables.
- b. Splitting into test and train dataset
- c. Rescaling of the variable so that unit is same for all.

Step 3. Training the model

- a. Here we create the object of LinearRegression.
- b. To choose variables we have 3 approaches
 - i. Take one by one
 - ii. Take all
 - iii. Use RFE (Recursive Feature elimination – need to specify the count at the start)
- c. Based on VIF and p-value we eliminate the unnecessary variables.
- d. Look for R-Squared value. Higher the value of it better the model.

Step 4. Residual analysis of the train data

- a. Here is check for assumptions of Linear regression using
 - a. Error terms should be a normal distribution
 - b. Independent variables should not be related to each other
 - c. Variance of the residual should be constant.

Step 5. Making predictions

- a. We make predictions on test dataset using the model we created with train dataset.

Step 6. Model evaluation against test data

- a. We compare the R-Squared values for train and test dataset. It should be close.

After above 6 steps we get the final model in terms of an equation.

Q.2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when plotted on scatter plots.

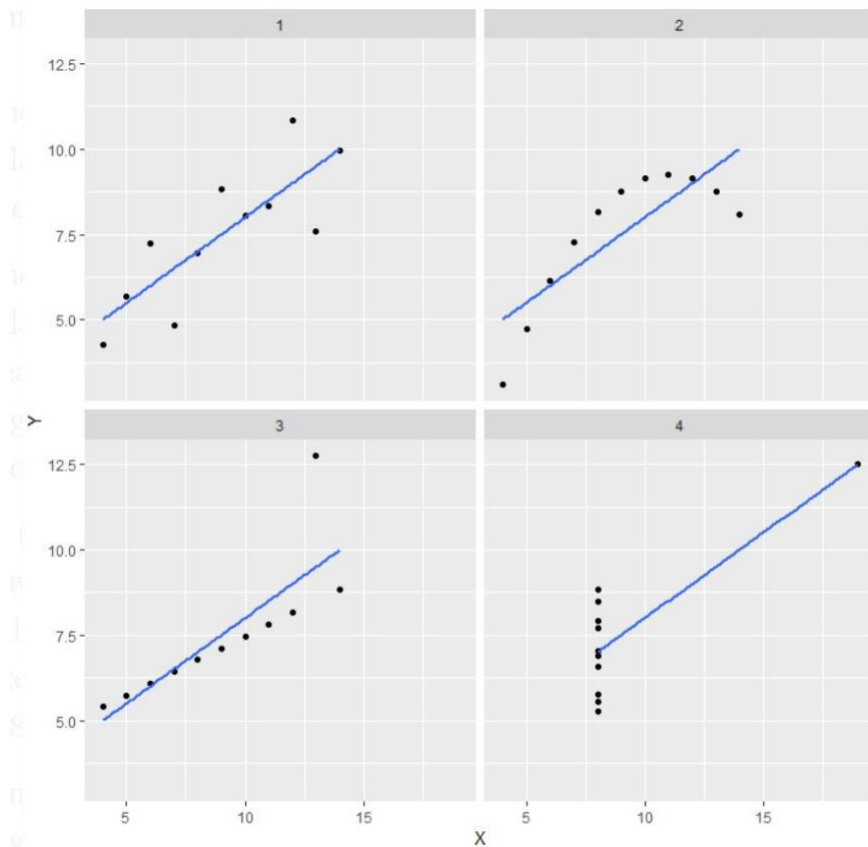
This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets.

Those 4 sets of 11 data-points are given below.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

When they were plotted we got



Explanation of this output:

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high- leverage point is enough to produce a high correlation coefficient.

Use: The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

References:

<https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2#:~:text=Anscombe's%20Quartet%20can%20be%20defined,when%20plotted>

<https://www.geeksforgeeks.org/anscombes-quartet/>

Q.3. What is Pearson's R?

Answer: Pearson's correlation coefficient or Pearson's R is used to measure the relationship between two continuous variables. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between - 1.0 and +1.0.

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

How is it calculated:

There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

Formula is:

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

Where,

N = the number of pairs of scores

$\sum xy$ = the sum of the products of paired scores **$\sum x$** = the sum of x scores

$\sum y$ = the sum of y scores

$\sum x^2$ = the sum of squared x scores

$\sum y^2$ = the sum of squared y scores

The positive value of Pearson's correlation coefficient implies that if we change either of these variables, there will be a **positive effect** on the other.

Reference:

<https://www.analyticssteps.com/blogs/pearsons-correlation-coefficient-r-in-statistics>

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

Answer:

Before any model building, we first need to perform the test-train split and **scale the features**.

Scaling of variables is an important step because, a variable might be on a different scale with respect to all other numerical variables. Also, the categorical variables that we encode may take either 0 or 1 as their values. Hence, it is important to have everything on the same scale for the model to be easily interpretable.

Scaling doesn't impact the model. It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So, it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale.

There are two common ways of rescaling:

- Min-Max scaling (Normalization) --> compresses all the data between 0 and 1
 - Normalization = $(x - x_{\min}) / (x_{\max} - x_{\min})$
- Standardisation (mean-0, sigma-1)
 - Standardisation = $(x - \mu) / \sigma$, $\mu = \text{mean}$

As a general rule of thumb -- we should use Min-Max scaling, as it takes care of outliers. But there are advantages of Standardisation too.

The advantage of Standardisation over the other is that it doesn't compress the data between a particular range as in Min-Max scaling. This is useful, especially if there are extreme data point (outlier).

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

Answer:

The higher the value of VIF the higher correlation between this variable and the rest. If there is perfect correlation, then $VIF = \text{infinity}$. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get $R^2 = 1$, which lead to $1/(1-R^2)$ infinity.

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

Answer:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

Few advantages:

- a) It can be used with sample sizes also
- b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

- i. come from populations with a common distribution
- ii. have common location and scale
- iii. have similar distributional shapes
- iv. have similar tail behaviour