Assignment-based Subjective Questions

Submitted by: Manas Ranjan Das

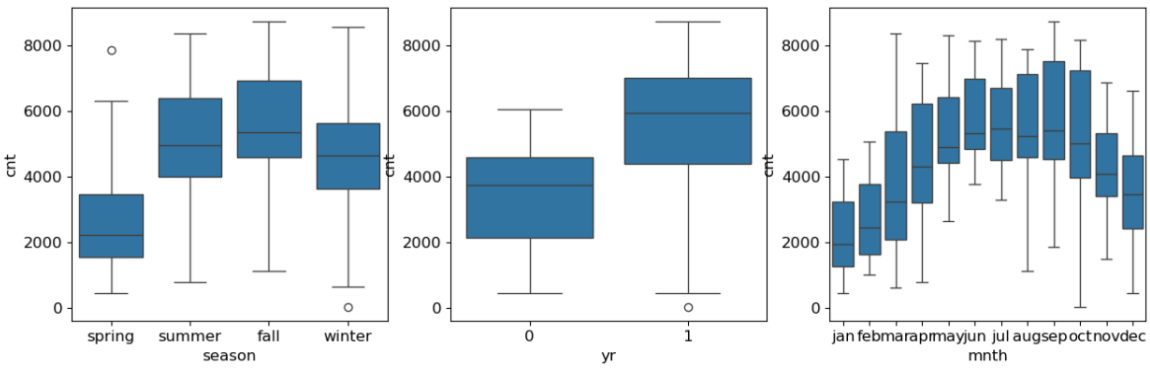
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

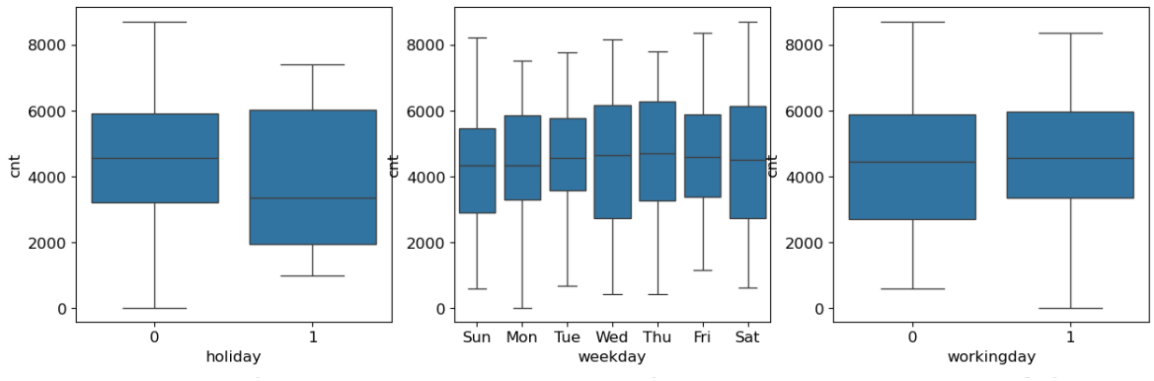
   Answer:

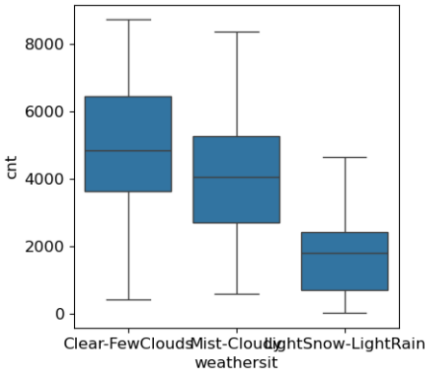
   We have used Boxplot for categorical variables to see demand for
   'season','yr','mnth','holiday','weekday', 'workingday','weathersit'

   Observations:

   1. Season vs cnt comparison: the fall season has highest demand against other seasons towards rental bikes.
   2. Yr vs cnt comparison: for 2019 year, demand has been grown against 2018.
   3. Demand vs cnt comparison: Lowest demand > Jan and Highest demand >September. The demands vs cmt is bellow curve of grow through the year.
   4. Holiday vs cnt comparison: during holiday period, the demand has been decreasing.
   5. Weekday vs cnt comparison: there is not much variance on demand.
   6. Weathershit vs cnt comparison: the (Clear, Few clouds, Partly cloudy, Partly cloudy) weather has highest demand with respect to others.
   7. workingday vs cnt comparison : there is not much variance on demand.







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

   Answer:

   The purpose of drop_first = True is usually to avoid multicollinearity. It helps reducing the extra column created during dummy variable creation. N-1 dummy variables can be used to describe a categorical variable with N levels. Dummy variables are mostly used when there are fewer levels. Example : Marital status

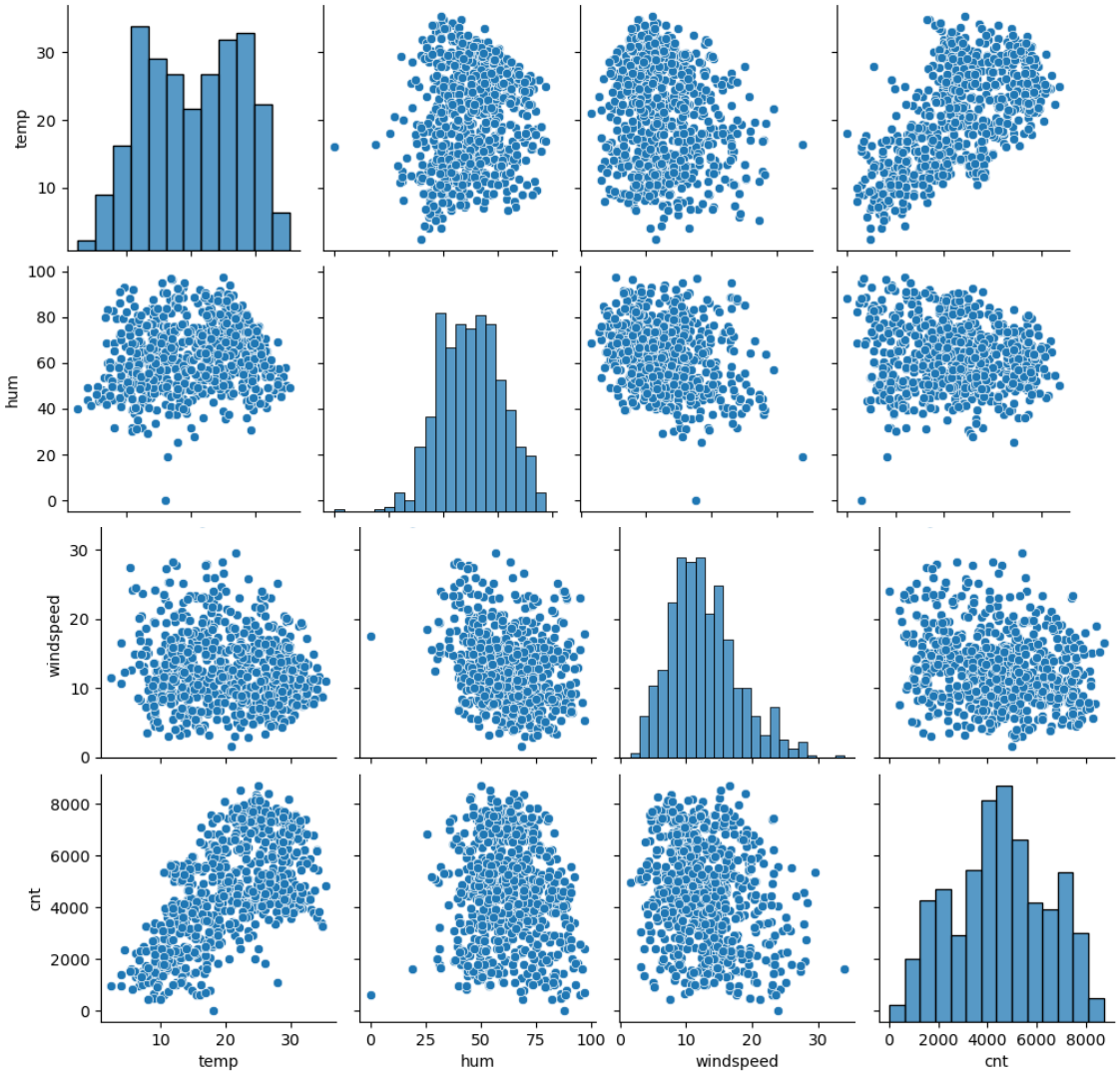
   Consider when a categorical variable is a factor of 15 levels. The number of dummy variables will be required to represent this categorical variable when developing a linear regression model = 15-1 -> 14.

   Command: dummy vars = pd.get_dummies(bike_df[category_vars].drop_first = True)

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

   Answer:

   Based on our bivariate analysis, we can see that the **temp** has highest positive correlation with variable **cnt**. Please find the screenshot from my pair-plot as below:



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

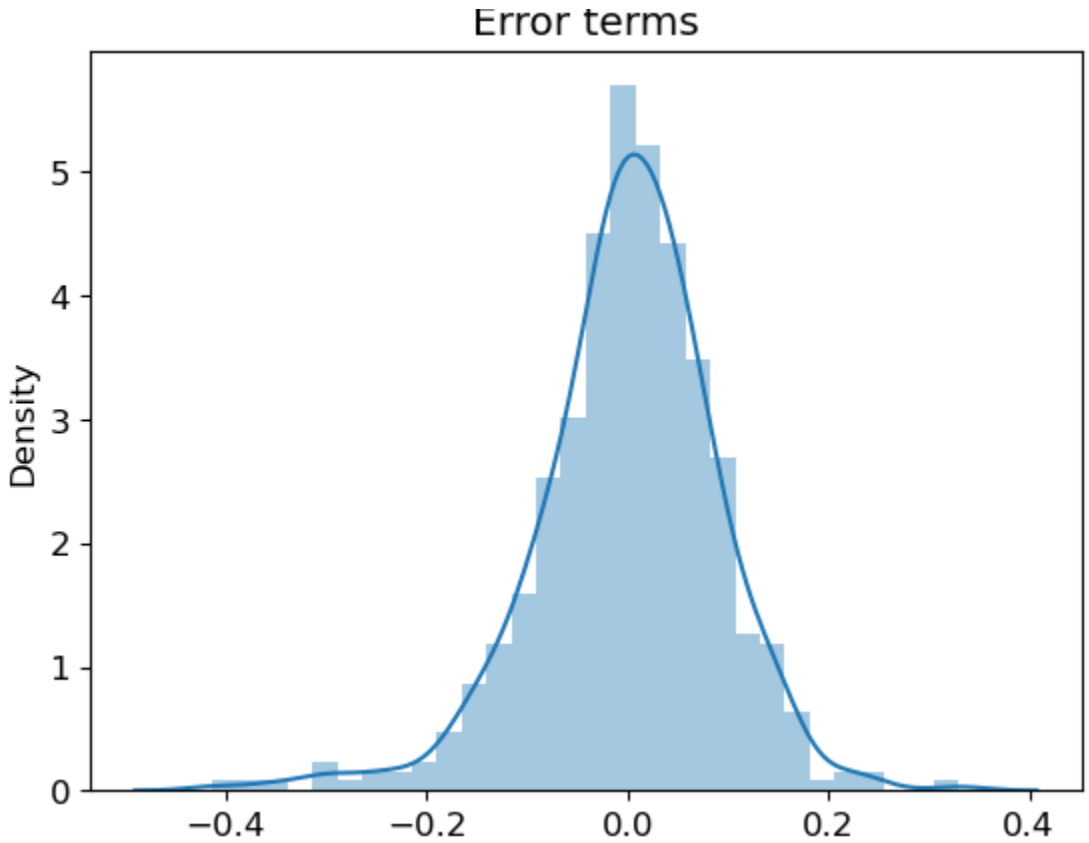
Answer:

The key assumptions of Linear Regression are as follows:

1. The error terms should be normally distributed.
2. Normality
3. Homoscedasticity
4. Multicollinearity
5. The error terms should have constant variance

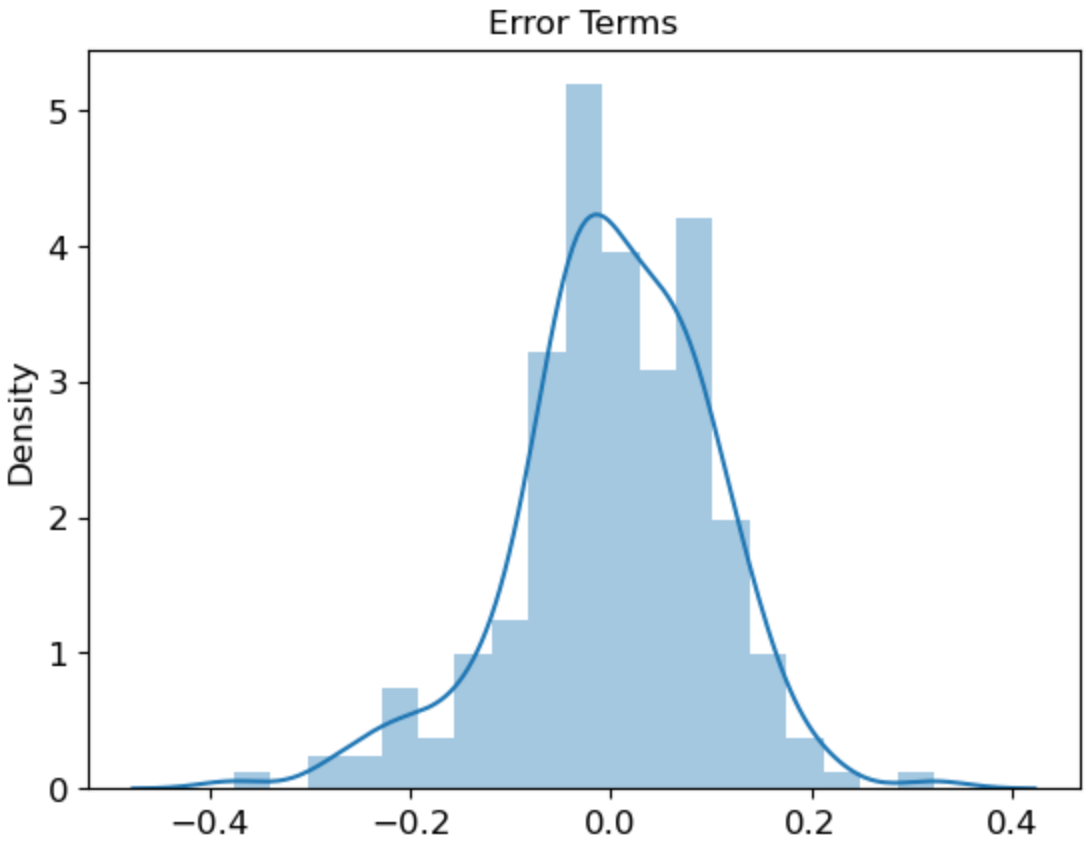
Based on our Model, we have achieved all the necessary assumptions require to be linear regression:

The distribution plot shows normal distribution with mean at Zero.



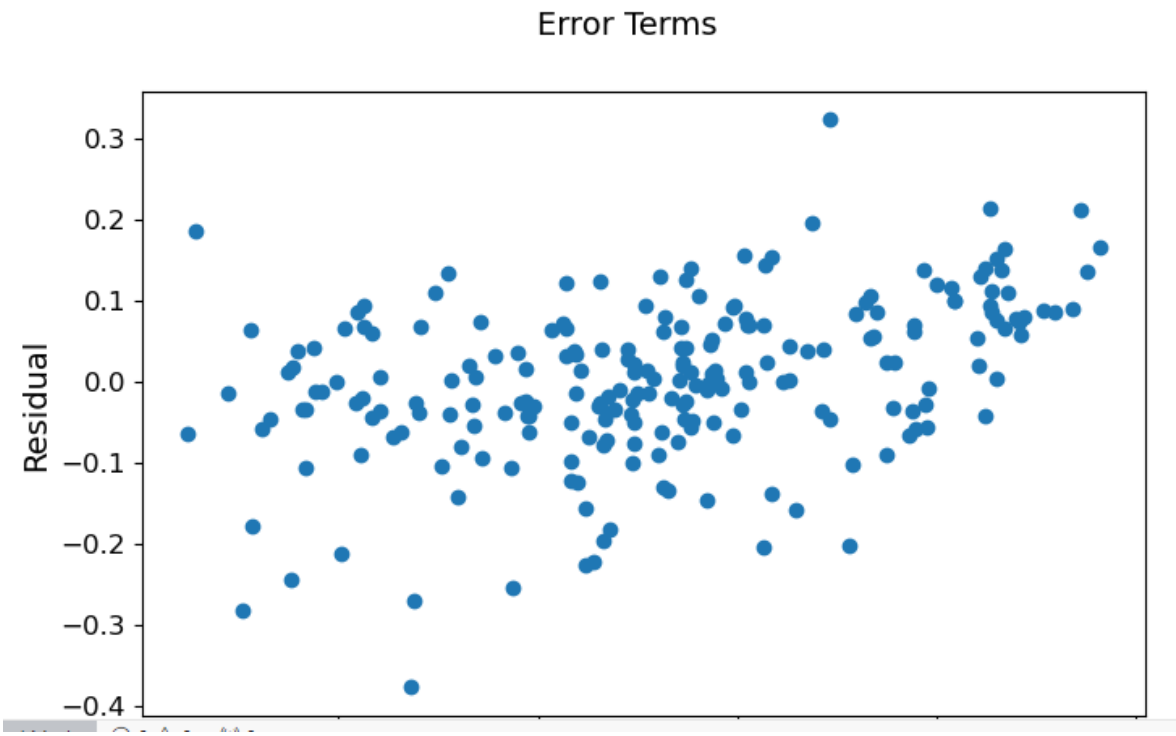
1. The distribution plot above for error term showed the normal distribution with mean zero.

We have performed Residual analysis and validated the distribution plot of Residue for test data set.



2. the above distribution plot for error terms shows Gaussian distribution with mean at Zero.

3. Homoscedasticity(There should be no visible pattern in residual values): We have observed the residual plot for test data and please find the screenshot as below:



In this above image, we found that the residual plot is random ones and the error terms are satisfying the rule of having reasonable constant variable.

4. **Multi-collinearity**: Based on our model analysis, we found that the linear regression model assumes that there is little or no multi-collinearity in the data. During Model building, we used the technique of detecting multi-collinearity using VIF and our model has all P-values nearby 0 and as well VIF < 5. Hence proved to have multi-collinearity in the data.

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:

Based on final model , the top three features that significantly contributing towards explaining the bile sharing demand model as follows:

1. temp

2. weathersit:Light Rain + Thunderstorm + Scattered clouds , Light Rain + Scattered Clouds

3. year (yr)

General Subjective Questions

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

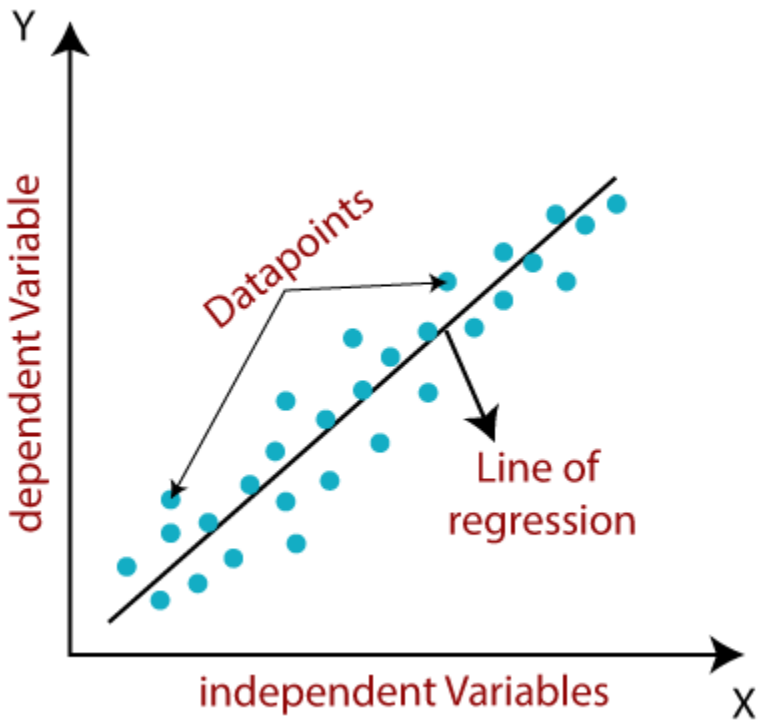
   Answer:

   Linear regression is a data analysis technique or an algorithm that predicts the value of unknown data by using another related and known data set. It is nothing but a linear relationship between an independent and dependent variables to predict the outcome of future events as a linear equation.

   In the mathematical models, the relationship can be represented as equation as : y = mX + c

   Where

   1. y is a dependant variable
   2. X is a Independent variable
   3. m is the slope of regression line = slope -> tan @ (where @ represents angel)
   4. c is interception



   The linear regression model is simple and easy to interpret mathematical formula to generate predictions. The greater the linear relationship between independent and dependent variables, the more accurate is the prediction.

   There are two types of linear regression:

1. Simple Linear regression
2. Multi Linear regression

Simple linear regression is defined by the linear function: **$Y= \beta0*X + \beta1 + \varepsilon$**
Where β0 and β1 are two unknown constants representing the regression slope, whereas ε (epsilon) is the error term.
This simple linear regression model can be used for the relationship between two variables,
Examples:

- Rainfall and crop yield
- Age and height in children

Multiple linear regression: The dataset contains one dependent variable and multiple independent variables. The linear regression line function changes to include more factors as follows: **$Y= \beta0*X0 + \beta1X1 + \beta2X2+\ldots\ldots \beta nXn+ \varepsilon$**
As the number of predictor variables increases, the β constants also increase correspondingly.
Examples:

- Rainfall, temperature, and fertilizer use on crop yield
- Diet and exercise on heart disease

The key assumptions of Linear Regression are as follows:

1. The error terms should be normally distributed.
2. Normality
3. Homoscedasticity
4. Multicollinearity
5. Relationship between variables

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

Answer:

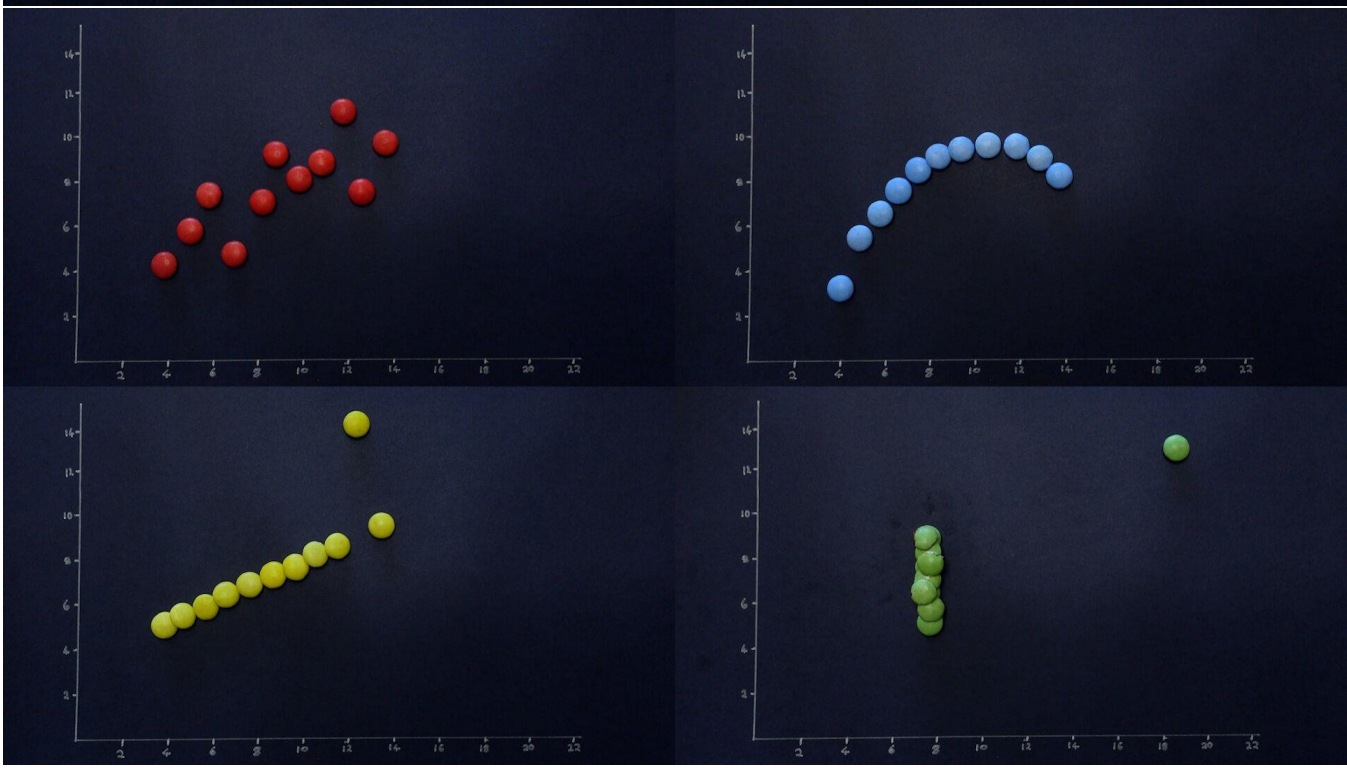
Anscombe's quartet can be defined as a group of four data sets those are **nearly identical in simple descriptive statistics**, but there are peculiarities that **fool the regression model** once you plot each data set. The data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build the model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets.

Anscombe's Quartet through example:

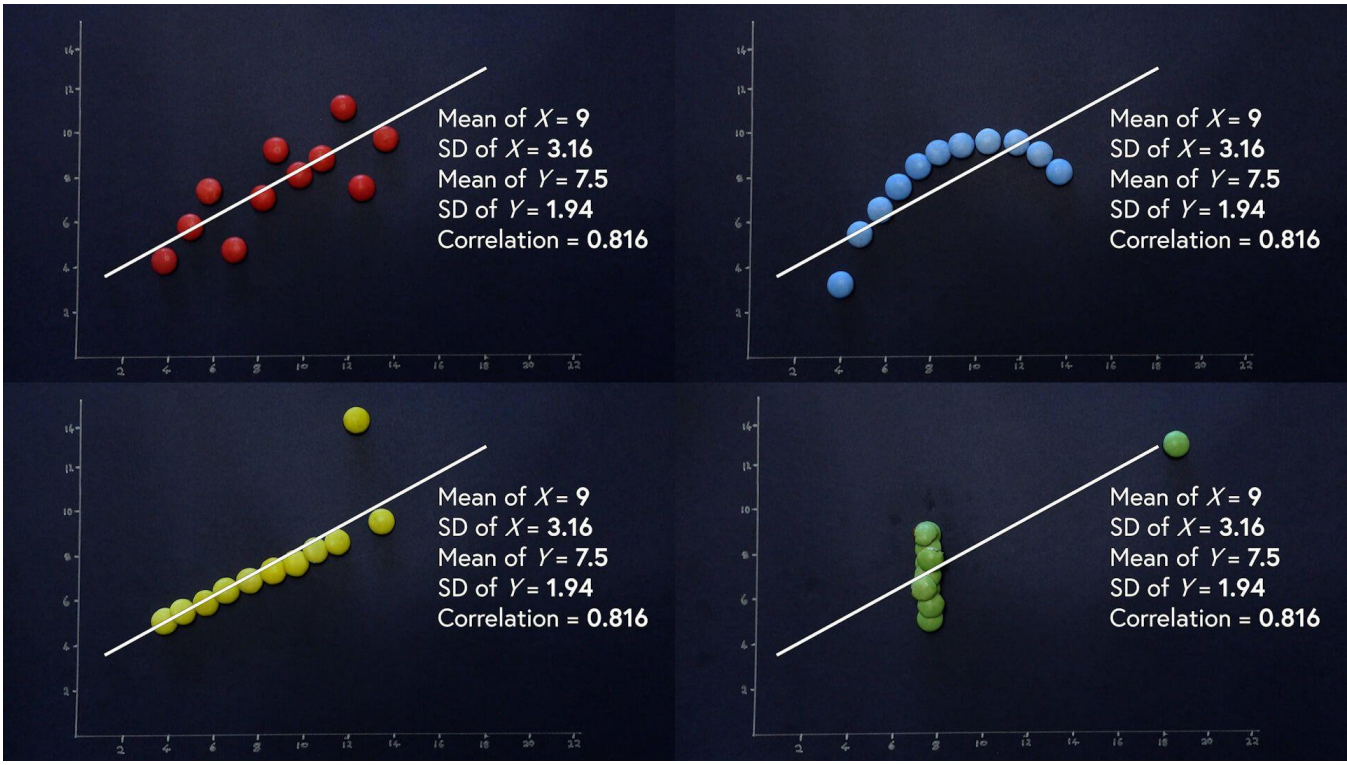
Anscombe's Quartet shows how four entirely different data sets can be reduced down to the same summary metrics. Here are the data sets from Anscombe's Quartet – both as raw data, and plotted on a chart:

Red		Blue		Yellow		Green	
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



You can tell just by looking that these data sets are very different. However, each data set is practically identical when you calculate the following summary metrics:

- Mean of *X* = 9
- Standard deviation of *X* = 3.16
- Mean of *Y* = 7.5
- Standard deviation of *Y* = 1.94
- Correlation between *X* & *Y*= 0.816
- The linear regression (the line of best fit) is also the same



Conclusion: It is important to use these as just one tool in a large data set analysis process. Visualizing our data set allows us to revisit our summary statistics and re-contextualize them as needed.

3. What is Pearson's R? (3 marks)

Answer:

Pearson's R is known as Pearson Correlation Coefficient (r) and the most widely used correlation coefficient.

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) 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

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

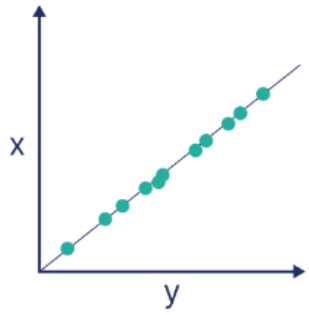
Pearson correlation coefficient (r) value	**Strength**	**Direction**
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

When r is 1 or −1, all the points fall exactly on the line of best fit:

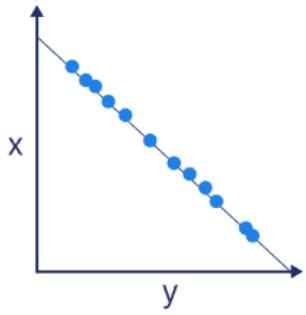
Perfect positive correlation

$r = 1$



Perfect negative correlation

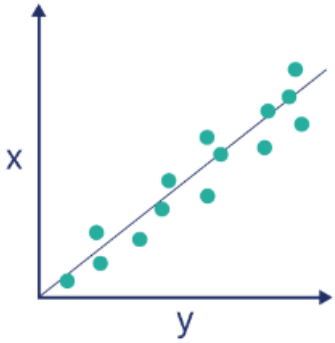
$r = -1$



When r is greater than .5 or less than −.5, the points are close to the line of best fit:

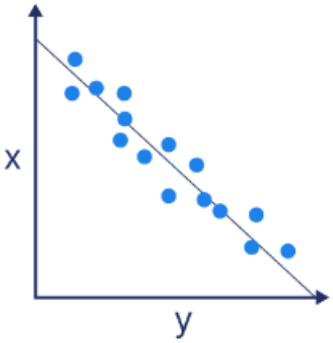
Strong positive correlation

$r > .5$

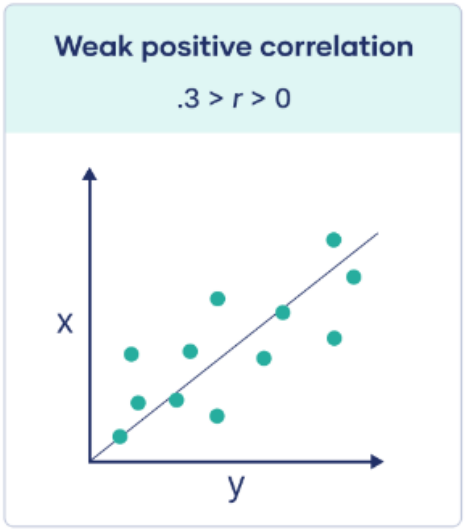


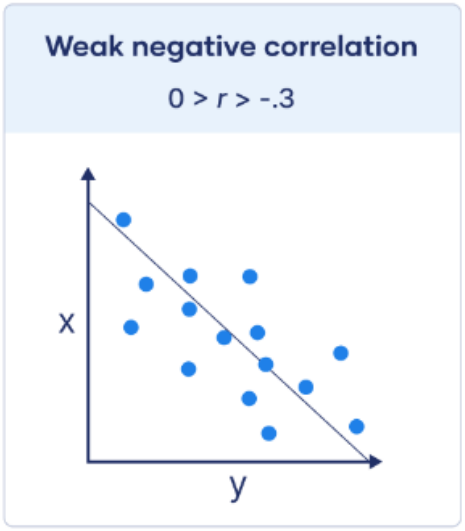
Strong negative correlation

$r < -.5$

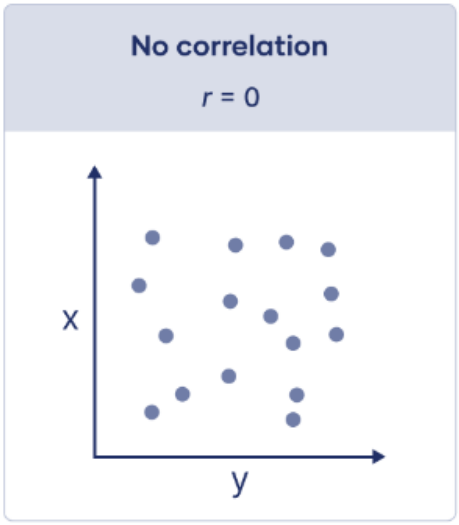


When *r* is between 0 and .3 or between 0 and –.3, the points are far from the line of best fit:





When *r* is 0, a line of best fit is not helpful in describing the relationship between the variables:



Calculating the Pearson correlation coefficient

Below is a formula for calculating the Pearson correlation coefficient (*r*):

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

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

Answer:

The Scaling is known as data normalization and is a data pre-processing step. Scaling of variables is a crucial step as it is a method used to normalize the range of independent variables of features of data. It is important to have everything on the same scales for the model to be easily interpretable.

There are two types of popular scaling:

1. Min-Max Scaling
2. Standardisation (mean = 0 and sigma = 1)

NOTE: It is important to note that **scaling just affects the coefficients** and none of the other parameters like t-statistics, F- statistic, p-values, R-squared etc.

Why Scaling!

The different models in the world, assigns weights to the independent variables according to their data set and conclusion for output. In that case, the difference between the data points are high, the model will need to provide more significant weight to the primary points and in the results, the model with large weight value assigned to underserving feature is often unstable. This means the model can produce poor results or can perform poorly during learning.

Scaling should always be done after the test-train split since you do not want the test dataset to learn anything from the train data. Hence, if you are performing the test – train split earlier, the test data will have information regarding the data like the minimum and maximum values.

Difference between normalized and standardized scaling:

1. Standardised scaling will affect the values of dummy variables, but MinMax scaling wiil not.
2. The advantage of standardisation over the other is that it does not compress the data between a particular ranges as Min-Max scaling. This is useful, especially if there are extreme data points (Outiler).
3. standardisation basically brings all the data into a standard normal distribution with mean zero and standard deviation as one. MinMax scaling , on the other hand, brings all of the data on the range of 0 to 1.
4. standardisation – method used to make sure that data is internally consistent where as MinMax scaling – Method used for making sure that data is internally consistent.

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

Answer:

VIF is a critical parameter that provided a measure of how much the variance of an estimated regression coefficient increase due to collinerity.IN order to determine VIF, we fit a regression model between independent and dependent variables.

The value of VIF is calculated by below Formula:

VIF = 1/ (1- R2)

If R-square (R2) from above formula is = 1, then the denominator will be zero and the overall value will become infinite. It denotes perfect correlation in variables.

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

(3 marks)

Answer:

The Q-Q plot or quantile-quantile plot is a graphical tool to help us access if a set of data possibly came from some theoretical distribution such as normal , uniform and exponential distribution. This plot helps for determining if two data sets are came from population with a common distribution.

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against.

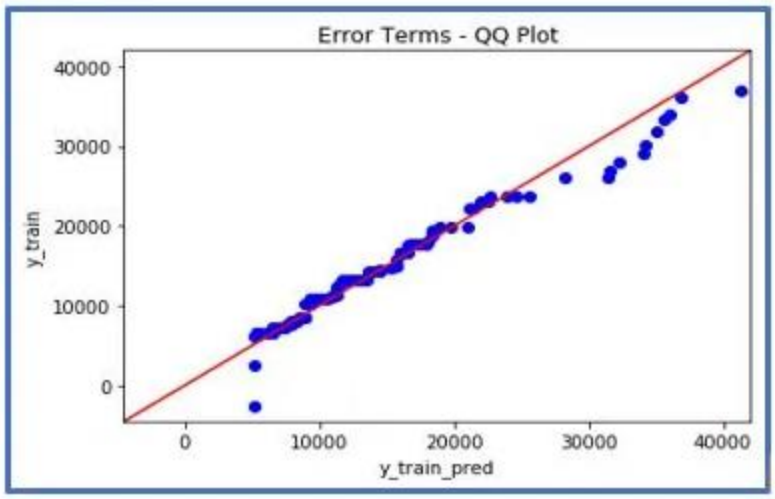
Interpretation:

A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

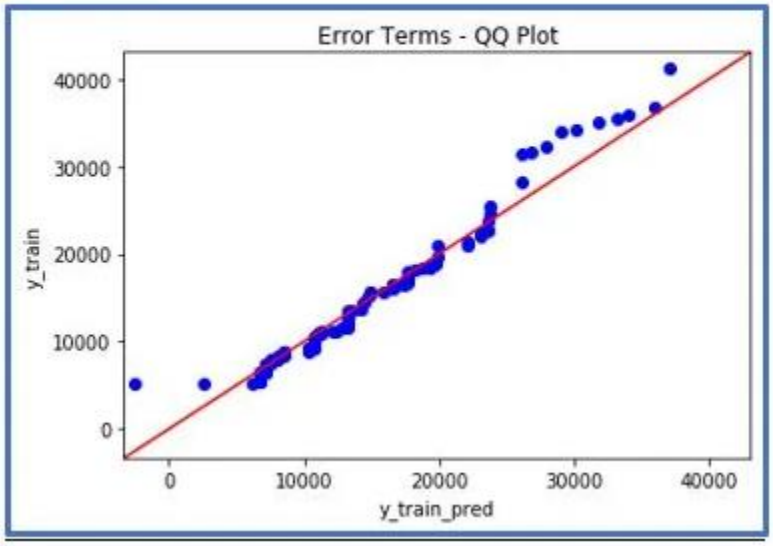
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x –axis.

Y-values < X-values: *If y-quantiles are lower than the x-quantiles.*



X-values < Y-values: *If x-quantiles are lower than the y-quantiles.*



Use of Q-Q plot in Linear Regression:

The Q-Q plot is used to see if the points lie approximately on the line. If they do not , it means, our residuals are not Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of a Q-Q plot in linear regression:

1. The sample sizes do not need to be equal.
2. 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.
3. The Q-Q plot can provide more insight into the nature of the difference than analytical methods.