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

* Year: The demand for bikes has increased in the year 2019
* Month: Months August to September has highest demand
* Holiday: The demand is in range of (2000-6000) in holidays compared to non-holidays whose demand is in the range of (3500+-6000)
* WeekDay: There is High demand on Thursdays (2500+-6000+) on Thursdays and the higher range is almost in the same range during almost all days except Sundays.
* WorkingDay: There is no major difference in demand between Working day and non-working day.
* Weathersit: The most favourable weather is when it’s clear/partly cloudy whereas the least favourable situation is when it has Light Rains/Thunderstorms
* Season: The demand is high during fall season, whereas the demand is the least during spring season.

**Why is it important to use drop\_first=True during dummy variable creation?**

drop\_first=True option helps us to remove a redundant dummy variable. Which in turn reduces the correlations created among the dummy variables(multi-collinearity).

For ex: If there’s a variable called Working status and it has three options Full-day, Half-day and Leave, without dropping a column the dummy variables looks like :

|  |  |  |
| --- | --- | --- |
| **Full-day** | **Half-Day** | **Leave** |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |

If we use drop\_first=True, it drops a redundant column and Leave can be represented as 0-0

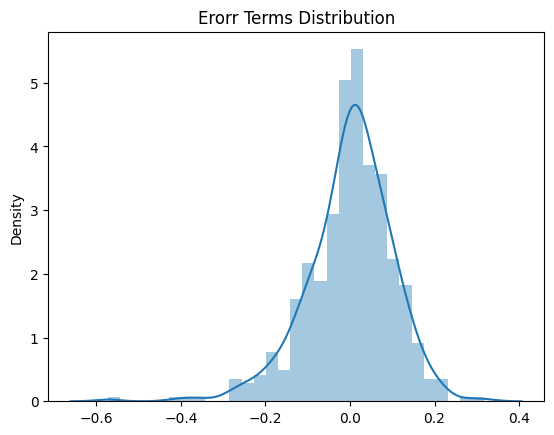
|  |  |
| --- | --- |
| **Full-day** | **Half-Day** |
| 1 | 0 |
| 0 | 1 |
| 0 | 0 |

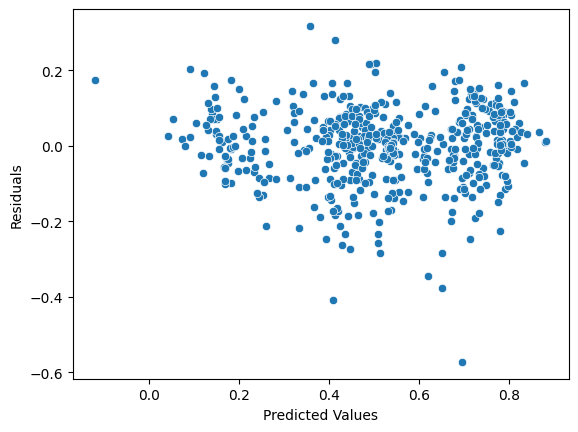
**Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

By looking at the pair plot, temp seems to have the highest correlation.

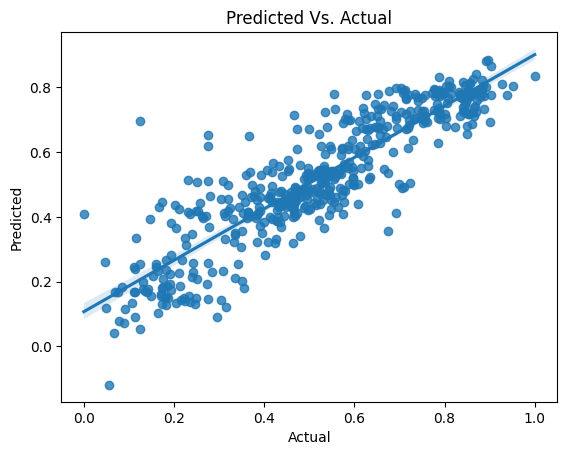
atemp and temp are highly correlated to each other, hence ignoring atemp.

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

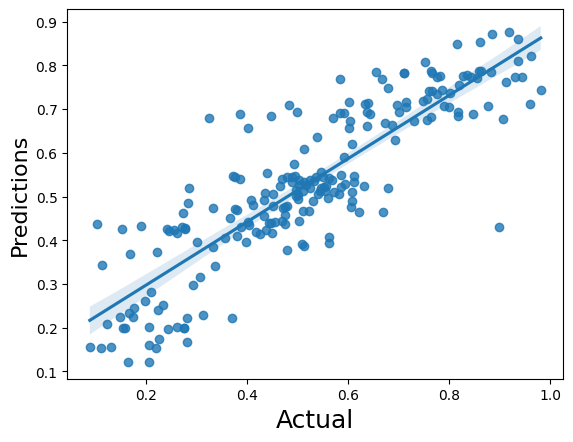
* The Error terms are normally distributed with mean at Zero.
* Error terms are independent of each other since we don’t see any specific pattern.



* There is linear relationship between dependant and independent variables



* Error terms have variance that’s constant



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

Top 3 features that contribute towards the demand of the shared bikes

* weathersit : Weather situations like Rain/Thunderstorm, Misty impacts the demand of shared bikes negatively.
* Season: Seasons like Spring and Winter impacts the demand of shared bikes.
* Yr : We can see that the demand for shared bikes increased in the year 2019

**Explain the linear regression algorithm in detail.**

Linear regression algorithm shows a linear relationship between dependent and independent variables i.e it helps us find how dependent variables change based on the value of independent variables.

It basically provides a sloped straight line that represents the relationship between variables.

Dependent Variables

Independent Variables

There are two types of Linear Regression Algorithm

* Simple Linear Regression
  + Single independent variable is used to predict the value of a dependant variable.
  + Y =
* Multiple Linear Regression
  + More than one independent variable is used to predict the value of a dependant variable.
  + is intercept
  + are slopes

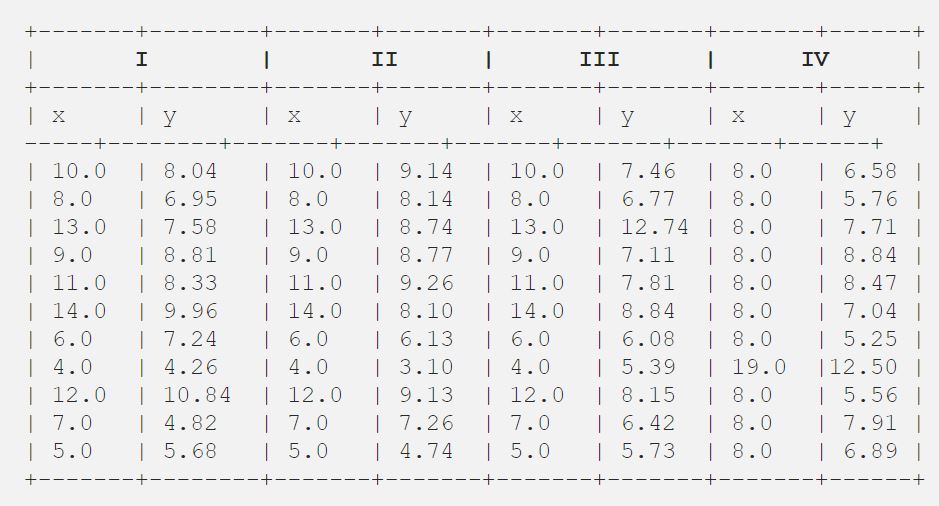
The main goal while using linear regression is to find best fit line i.e error between predicted and actual values should be minimal. This is found by minimising the Residual Sum of Squares:

The Strength of linear regression depends on:

* R² = (1-RSS/TSS)
* Residual Standard Error

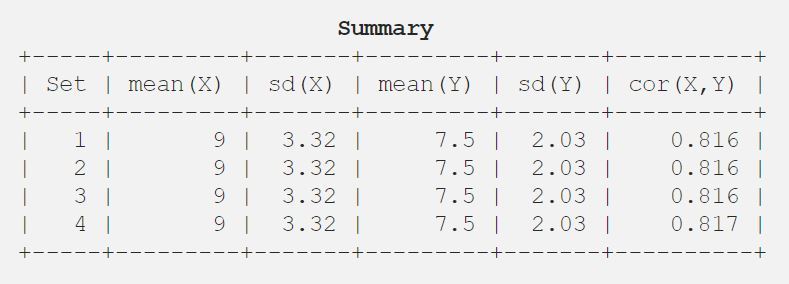
**Explain the Anscombe’s quartet in detail.**

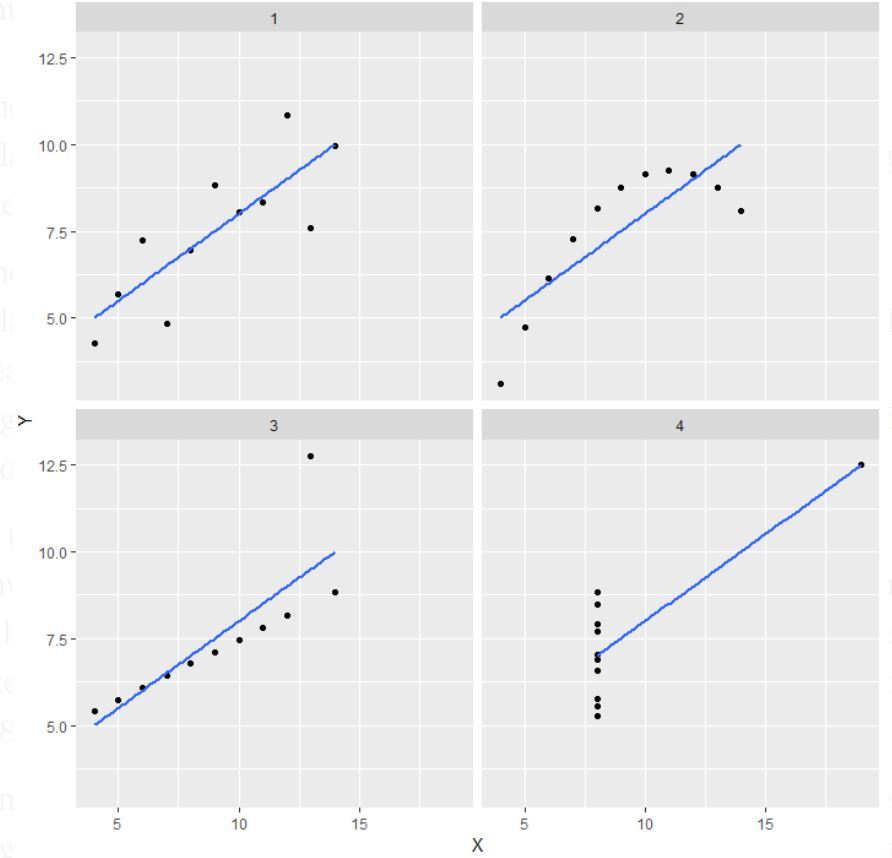
It’s a group of 4 data sets that are nearly identical when used simple descriptive statistics, but they have different distributions and they also appear different when the data sets are put in a graph. Which tells us to visualize the data before applying any algorithm to build models so that various anomalies present in the data such as outliers, diversity of data etc.. can be identified.



The above data sets are the 4 data-sets of 11 points.

When we use descriptive statistics, we can see that all the values of mean, standard deviation are similar.





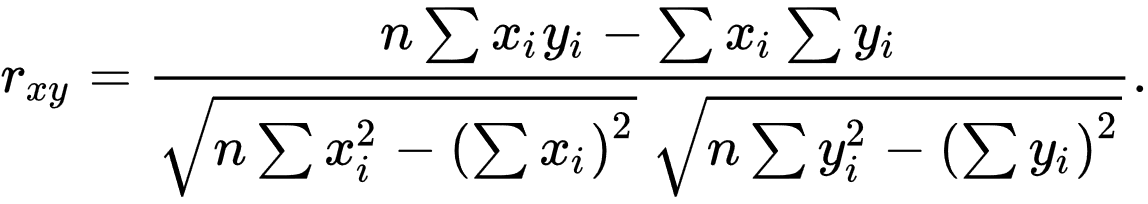
But when data is visualized, we can see that

* In the first plot the data has some sort of linear relationship
* In the second plot there is a non-linear relationship
* In the third plot there is an outlier(far away from the line)
* In the fourth plot shows one high leverage point(far right) is enough to produce a high correlation coefficient

**What is Pearson’s R?**

Pearson’s R also known as Pearson’s correlation Coefficient is a measure of linear correlation between two sets of data. Pearson’s R returns values between -1 and 1. Where

* Coefficient value 1 shows strong relationship
* Coefficient value -1 shows inverse relationship
* Coefficient value 0 shows no relationship



n: sample size

: Individual sample points indexed with i

: and similarly, for

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

Scaling is one of the pre-processing step’s that’s performed on independent variables which helps us normalize data within a particular range.

All independent variables won’t be in the same rage and will have different scales, which inturn will lead to weird coefficients and will be difficult to interpret. Scaling is also important since algorithm takes magnitude into account and not the units.

Normalized Scaling: It brings all of the data in the range of 0 and 1

x= x-min(x)/max(x)-min(x)

Standardized Scaling: It replaces the values by their Z scores. It brings all of the data into a standard normal distribution whose mean is 0 and standard deviation is 1.

Standardization has a disadvantage over normalization since it loses outliers.

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

Formula for VIF =

Based on the above formula, the VIF value will be infinite of the returns 1. i.e When there’s perfect correlation between independent variables i.e. multicollinearity.

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

Also called as quantile-quantile plots. It’s a graphical tool which helps us to see if datasets are from populations with a common distribution.

Many aspects like shift in scale, location, changes in symmetry and presence of outliers can be detected using Q-Q plots.

In linear regression, if we have received train and test data sets separately, Q-Q plots can be used to determine that both are from populations with same distribution.