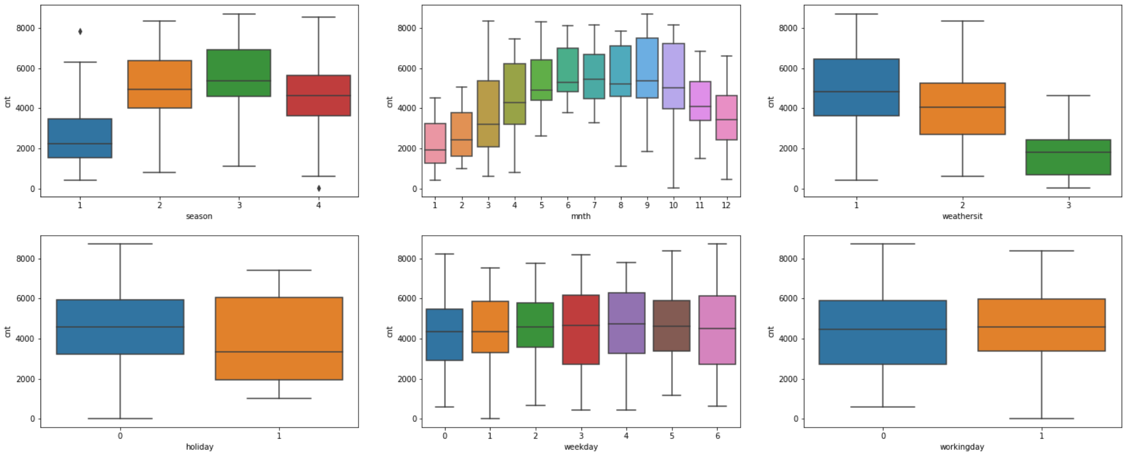
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:



Inference about categorical variables:

Season

Season3 has the most bookings with highest median around 5000
It is clear that season affects the count. Hence it could be a predictor.

Month

Mnth9 has the maximum bookings with a median of 5000
It is clear that mnth affects the count. Hence it could be a predictor.

Weather situation

Weathersit1 has most bokkings with a median of 5000.
It is clear that Weathersit affects the count. Hence it could be a predictor.

Holiday

when it is not a holiday, the median booking is higher. However, when it is a holiday, the 75th percentile is higher. Not sure. Will let the model decide if it is a good predictor.

Weekday

Shows a pattern with median closely distributed.
Will let the model decide if it is a good predictor.

Workingday

Workingday has higher median count than not a working day
It is clear that Workingday affects the count. Hence it could be a predictor.

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

Answer:
If there are two levels to a categorical variable. For example, Gender can have M or F. Pandas get_dummies function will create two columns which will look like the below

| id | M | F |
|---|---|---|
| 200 | 1 | 0 |
| 201 | 0 | 1 |

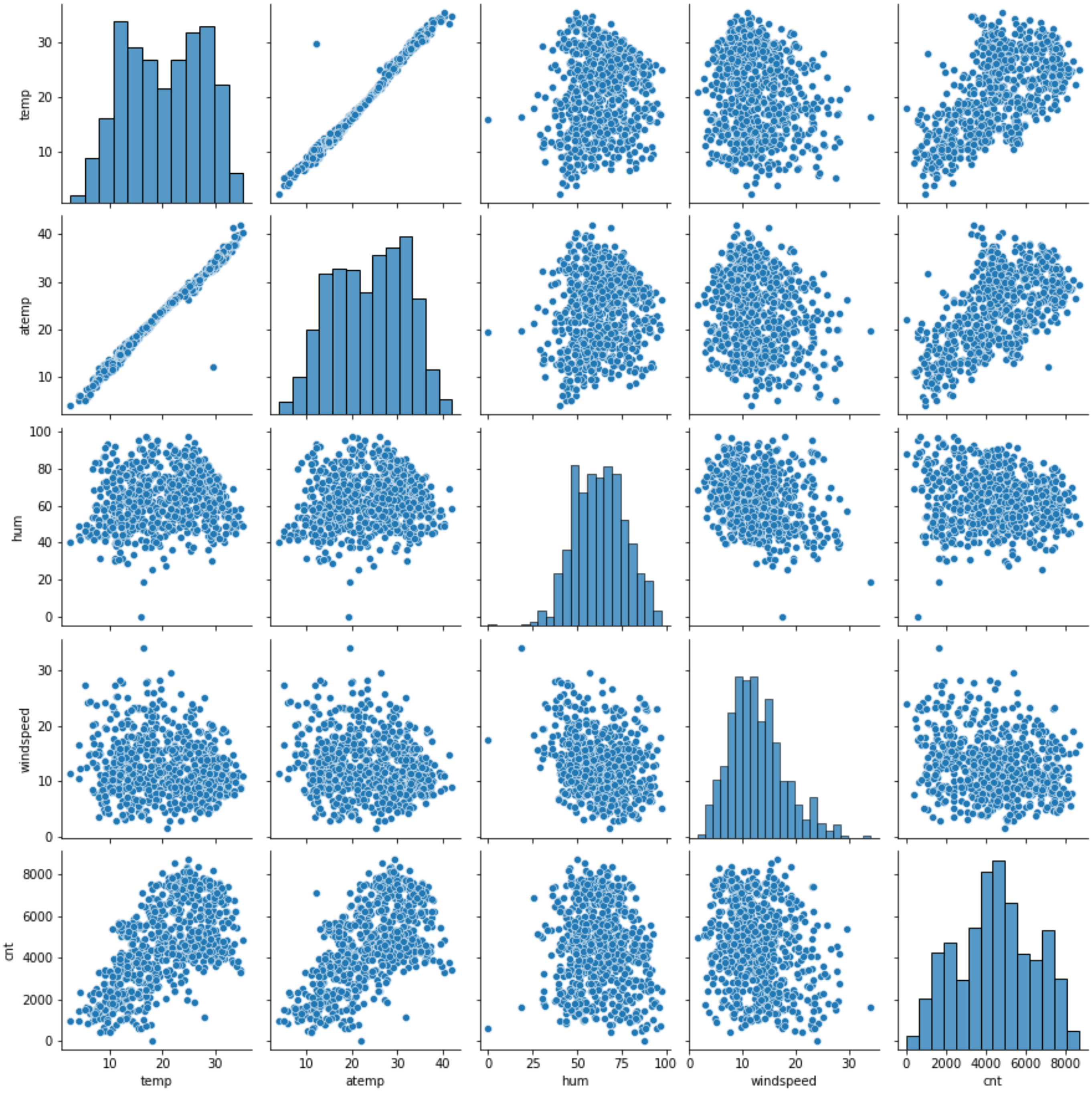
As you can see if M=1 , it implies F=0 and if F=1, it implies M=0. This creates two highly correlated columns which is against one of the assumptions of linear regression. In order to avoid this, we will drop the first column. The dataframe will look like the below.

| id | M |
|---|---|
| 200 | 1 |
| 201 | 0 |

In this M=0, implies that the person is F. We don't need another column to indicate this. Similar logic is applied to categorical variables with n levels.

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

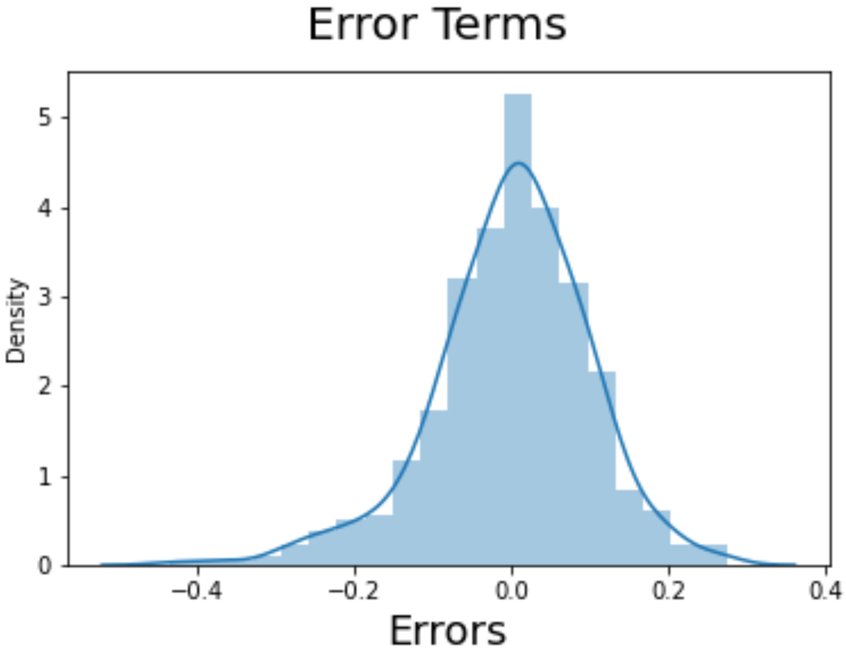
Looking at the plots, temp and atemp is highly correlated with the target variable.



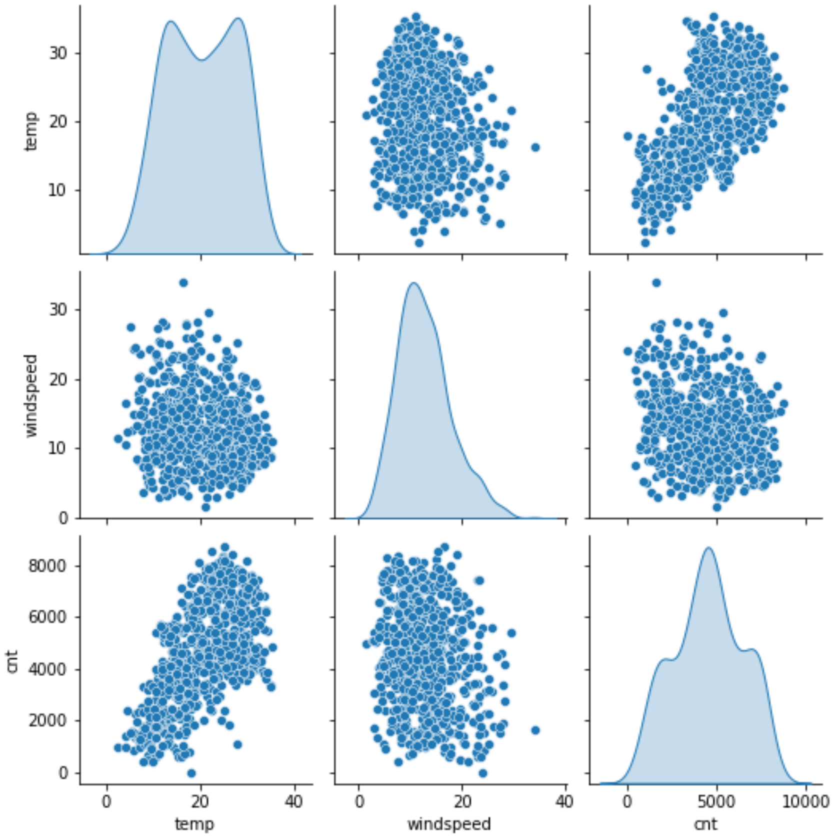
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Assumption 1: Residuals are normally distributed.

This was verified with the below plot.



Assumption 2: There is a linear relationship between X and Y. This was validated by the below plot. It is clear that temp has a linear relationship with cnt.



Assumption 3:
There is No Multicollinearity between the predictor variables.

This is verified by checking the VIF values. In the final model, the VIF values of all the predictors were less than 5.

Assumption 4:
Observations are independent of each other.

This is verified by the low p-value of all the predictors.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

As per our final Model, the top 3 predictor variables that influences the bike booking positively are:

- Temperature
- Year
- Winter

General Subjective Questions

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

Answer:
Linear regression is based on supervised learning. It performs a regression task.
Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

The equation of regression line with x(predictor) and y(target) variable is given by,

$$y = \beta1 + \beta2 * x$$

where β1 is the intercept and β2 is the co-efficient of x.

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best β1 and β2values.

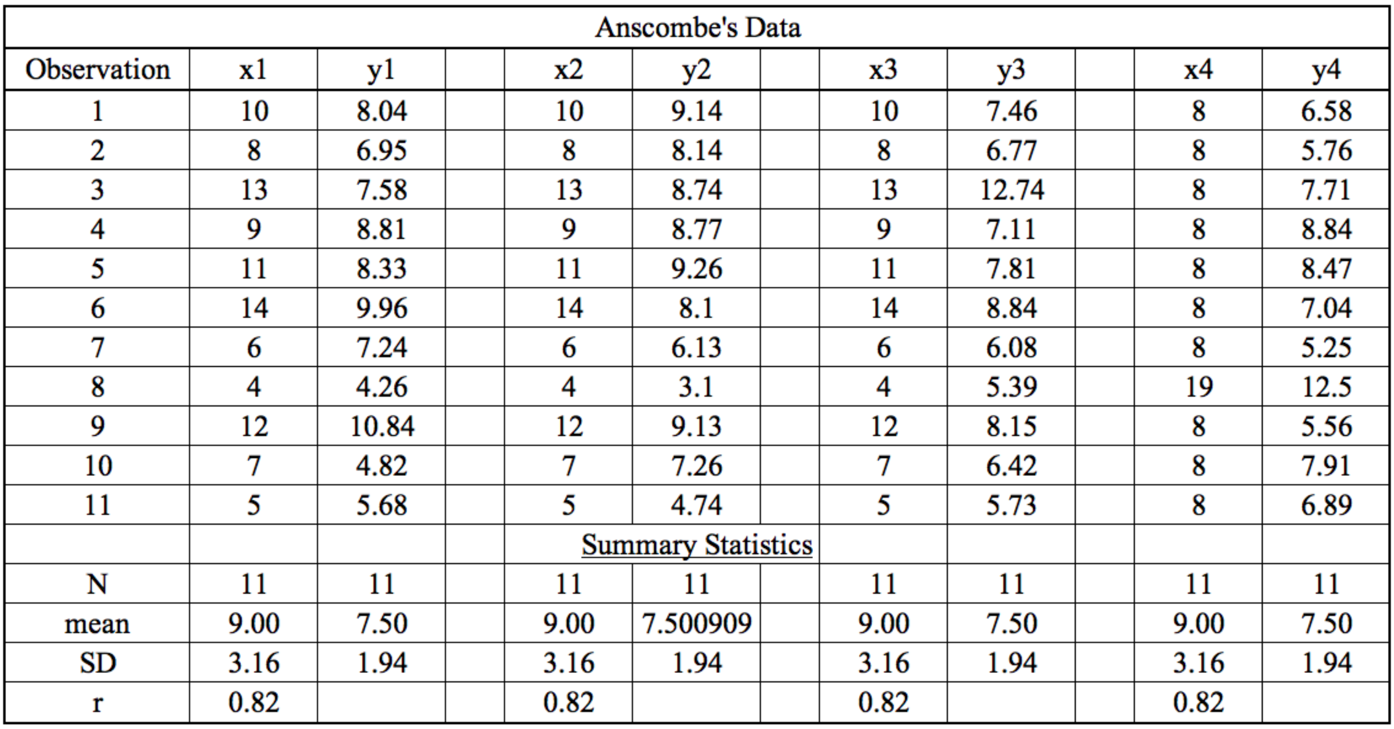
By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum.

Cost function(J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (pred) and true y value (y).

To update θ1 and θ2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the Linear Regression model uses Gradient Descent.

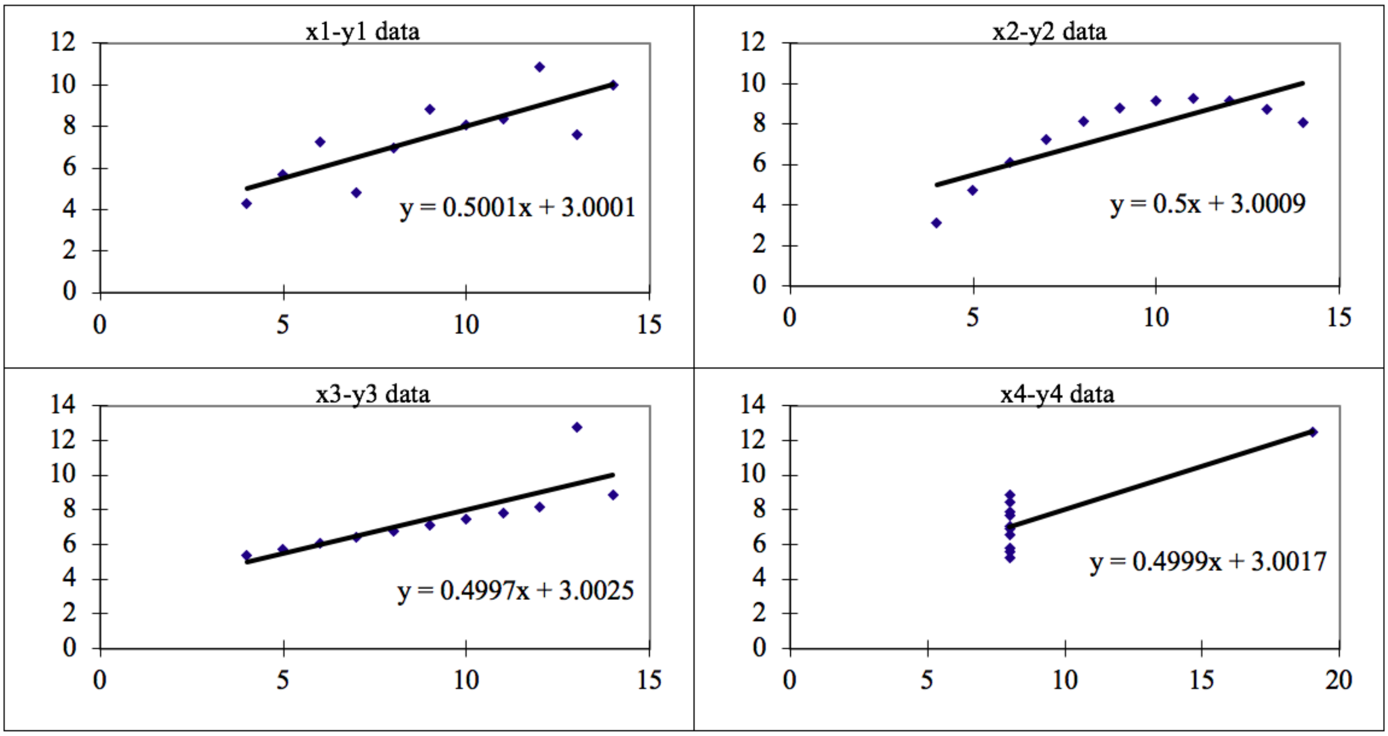
2. Explain the Anscombe’s quartet in detail. (3 marks)

**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**. There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets.
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.



| 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 |
| | | | | | 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 generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:
**Dataset 1:** this **fits** the linear regression model pretty well.
**Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
**Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
**Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

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

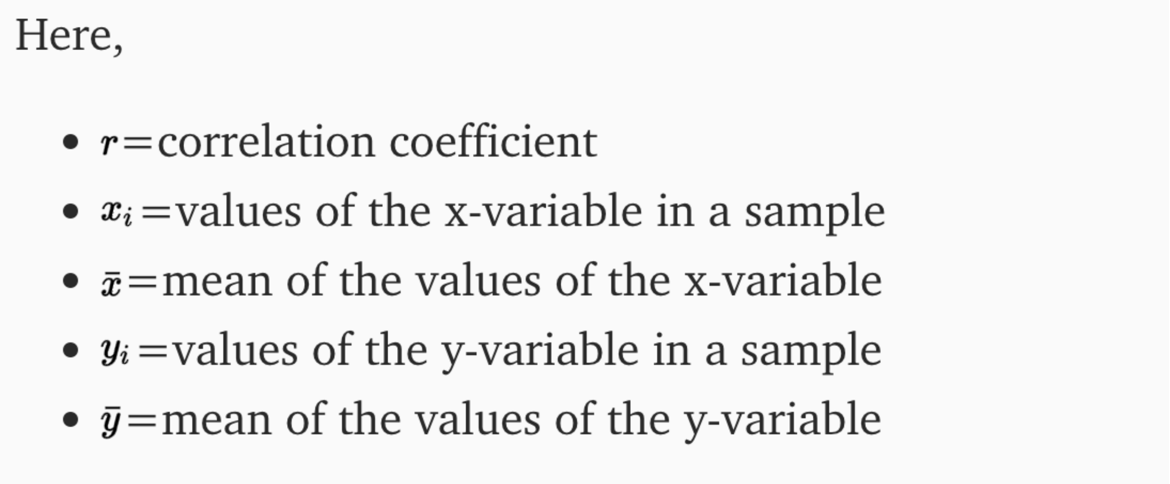
3. What is Pearson’s R? (3 marks)

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 normalised measurement of the covariance, such that the result always has a value between −1 and 1.

| Pearson correlation coefficient (*r*) | Correlation type | Interpretation | Example |
|---|---|---|---|
| Between 0 and 1 | Positive correlation | When one variable changes, the other variable changes in the **same direction**. | Baby length & weight:<br>The longer the baby, the heavier their weight. |
| 0 | No correlation | There is **no relationship** between the variables. | Car price & width of windshield wipers:<br>The price of a car is not related to the width of its windshield wipers. |
| Between 0 and –1 | Negative correlation | When one variable changes, the other variable changes in the **opposite direction**. | Elevation & air pressure:<br>The higher the elevation, the lower the air pressure. |

Pearson r Formula

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



Here,

- $r$=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

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

   Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing.

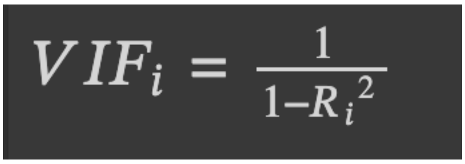
   Normalization is a part of data processing and cleansing techniques. The main goal of normalization is to make the data homogenous over all records and fields. It helps in creating a linkage between the entry data which in turn helps in cleaning and improving data quality.
   Whereas data standardization is the process of placing dissimilar features on the same scale. Standardized data in other words can be defined as rescaling the attributes in such a way that their mean is 0 and standard deviation becomes 1.

   Normalized Data Vs Standardized Data

   - Normalization is used when the data doesn't have Gaussian distribution whereas Standardization is used on data having Gaussian distribution.
   - Normalization scales in a range of [0,1] or [-1,1]. Standardization is not bounded by range.
   - Normalization is highly affected by outliers. Standardization is slightly affected by outliers.
   - Normalization is considered when the algorithms do not make assumptions about the data distribution. Standardization is used when algorithms make assumptions about the data distribution.

6. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.
   The equation of VIF is



$$VIF_i = \frac{1}{1-R_i^2}$$

   R-Squared ($R^2$ or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable.

   In case of perfect multi collinearity, $R^2 = 1$. Hence, the VIF is infinity.

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

- Machine learning algorithms like linear regression and logistic regression perform better where numerical features and targets follow a Gaussian or a uniform distribution.

- It’s an important assumption as normal distribution allows us to use the empirical rule of 68 – 95 – 99.7 and analysis where we can predict the percentage of values and how far they will fall from the mean.

- In regression models, normality gains significance when it comes to error terms. You want the mean of the error terms to be zero. If the mean of error terms is significantly away from zero, it means that the features we have selected may not actually be having a significant impact on the outcome variable.

The power of Q-Q plots lies in their ability to summarize any distribution visually.
QQ plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.
- Skewness of distribution
- In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.