**Analysis Report**

**Exploratory Data Analysis of the dataset**

Master’s in computer science

Student: Mohit Kumar

Email: mohit.kumar@iu-study.org

Matriculation: 92124758

University: International University of Applied Sciences

Juri-Gagarin-Ring 152 · D-99084 Erfurt

University Supervisor: Dr. Cosmina Croitoru

c.croitoru@iubh-fernstudium.de

Dr. Lino Antoni Giefer

[l.giefer@iubh-fernstudium.de](mailto:l.giefer@iubh-fernstudium.de)

Submission Date: January 2023

Contents

[**Abstract.** - 3 -](#_Toc110198341)

[**1.** **Introduction** - 3 -](#_Toc110198342)

[**2.** **Exploratory data analysis** - 3 -](#_Toc110198343)

[**3.** **Conclusion** - 24 -](#_Toc110198344)

[**4.** **Bibliography** - 25 -](#_Toc110198345)

[**5.** **GitHub Page** - 25 -](#_Toc110198346)

[**6.** **Appendix** - 25 -](#_Toc110198347)

**Abstract.**

In this investigation report, we present exploratory analysis for a given dataset (Train function) and pattern recognition for a collection of datasets (Ideal function) concerning the Train dataset, identifying the behavior of the test dataset through mapping.

**“Every block of stone has a statue inside it, and it is the task of the Sculptor to discover it.” – Michelangelo** [1]

# **Introduction**

The training dataset is analyzed, and behavioral similarity is to be identified from the set of ideal functions, the test data is to be investigated by mapping concerning the ideal function, A pattern and behavioral analysis of the training dataset is part of the investigative project.

The project is divided into three stages of investigation according to the dataset:

1. Understating characteristics of training data function.
2. Identifying the best fit ideal function to train data function.
3. Mapping test data points to ideal function within max deviation and observe characteristics

The analysis code is developed with the python programming language taking support from the relevant python library Pandas (Data Analysis Library) for data processing, and Bokeh (Graph and Plotting Library) for data visualization.

The SQLite relational database management system is used to store data, and the SQLAlchemy Python toolkit is used to perform database operations.

# **Exploratory data analysis**

* 1. **Data Collection**
     1. **Understanding data**

Three datasets are available for the investigative project in CSV format,

Visual validation of the dataset is performed to check for irregularities and impurities of data, and it’s validated that the dataset has no irregularities in value,

Below is the list of datasets available for investigation:

1. Training dataset (4 different datasets).

**Task**: Analyzing the given dataset for the behavioral and characteristic pattern.

Based on the visual observation of the dataset, we know it has an independent variable (X), and four dependent variables (Y1, Y2, Y3, Y4)



1. Collection of 50 Datasets (Ideal function).

**Task**: Estimating the line of best fit to train the dataset.

Based on the visual observation of the dataset, we know it has an independent variable (X), and fifty dependent variables (Y1, Y2, Y3 …Y50)



1. Test dataset (1 dataset).

**Task**: Mapping of test dataset on the ideal function within the max deviation

Based on the visual observation of the dataset, we know it has an independent variable (X), and one dependent variable (Y)



* + 1. **Data storing**

The datasets available for the project are in plain text (CSV format), and the data must be stored in a database.

SQLite is the choice, because of its characteristics and suitability as follows.

a) SQLite is an open-source file-based relational database management system.

b) SQLite doesn't require installation, unlike PostgreSQL.

c) High performance and one read/write operation at a time

* + 1. **Accessing data**

The dataset is stored and retrieved from SQLite using the SQLAlchemy Python module, which is a python SQL toolkit, and Object-relational Mapper.

Following are the characteristics of SQLAlchemy that are suitable for project requirements.

1. Stable and High performance
2. Supports SQLite Database
3. SQL queries can be written in a pythonic way
4. Pythonic SQL queries for different SQL databases (SQLite or SQL Server)
   * 1. **Result of data stored in a database**

Below are the tables to be created in the SQLite database.

**Table 1:** The training data's database table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y1 (training func) | Y2(training func) | Y3(training func) | Y4(training func) |
| x1 | y11 | y21 | y31 | y41 |
| ... | ... | ... | ... | ... |
| xn | y1n | y2n | y3n | y4n |

**Table 2:** The ideal functions’ database table:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | Y1 (ideal func) | Y2 (ideal func) | ... | Ym (ideal func) | ... | Y50 (ideal func) |
| x1 | y11 | y21 | ... | ym1 | ... | y50\_1 |
| ... | ... | ... | ... | ... | ... | ... |
| Xn | y1n | y2n | ... | Ymn | ... | y50\_n |

**Table 3:** The database table of the test data, with mapping and y-deviation

|  |  |  |  |
| --- | --- | --- | --- |
| X (test func) | Y (test func) | Delta Y (test func) | No. of ideal func |
| x1 | y11 | y21 | N1 |
| ... | ... | ... | ... |
| Xn | y1n | y2n | y3n |

* 1. **Pattern analysis**
     1. **Data visualization**

Exploratory and Explanatory(declarative) are two essential categories of

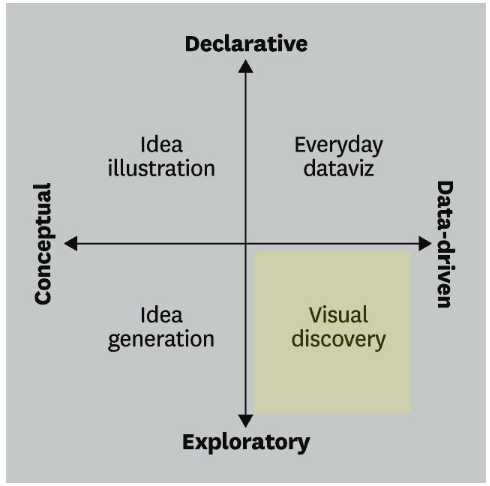
Information design, explains Scott Berinato, author of the book "Good

Charts: The HBR Guide to Making Smarter, More Persuasive Data

Visualizations"[2] Based on these four quadrants, explains Berinato

information design can be divided into four aspects,

**Figure** 2.1: Four types of Visualization chart



Quadrant I: Data-driven & Declarative (Everyday dataviz)

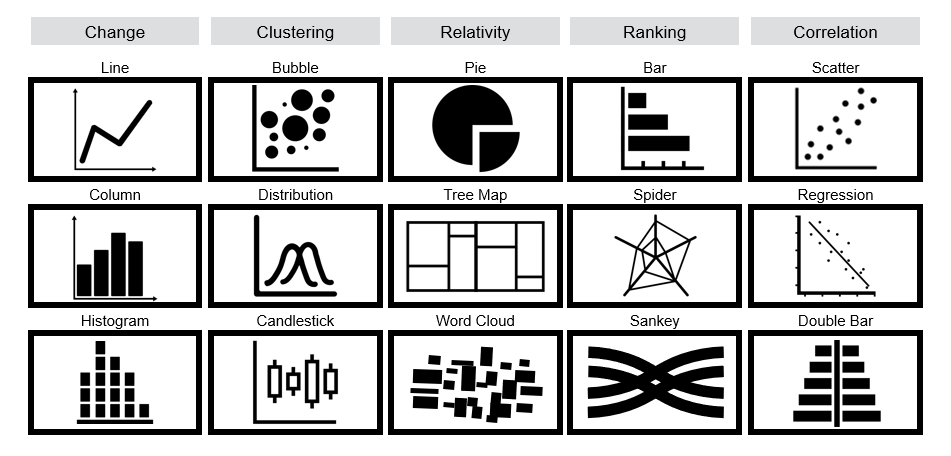
Quadrant II: Conceptual & Declarative (Idea illustration)

Quadrant III: Conceptual & Exploratory (Idea generation)

**Quadrant IV:** Data-driven & Exploratory (Visual discovery)

**The investigation demands exploration of the dataset, so it's a data-driven project and visual discovery of data is required.**

**Figure** 2.2: Visualization charts for data-driven pattern,



**Correlation is best achieved with a regression chart, which analyses the relationship between independent and dependent variables.**

**Explore patterns using exploratory analysis**

Exploratory data analysis is an approach to [analyzing](https://en.wikipedia.org/wiki/Data_analysis) [data sets](https://en.wikipedia.org/wiki/Data_set) to summarize their main characteristics, often using [statistical graphics](https://en.wikipedia.org/wiki/Statistical_graphics) and other [data visualization](https://en.wikipedia.org/wiki/Data_visualization) methods

During the exploratory analysis of the training dataset, we will explore the relationship between the dependent and independent variables without making any assumptions about the dataset.

* + 1. **Regression analysis**

Regression analysis allows for investigating the relationship between variables. Independent and dependent variables are two datasets required for building correlation.

Independent variables are part of the input, factors that affect the outcome of dependent variables, an example of an independent variable is time. As time increases, a lot of changes are encountered in the world (increase in the number of electronic devices)

Dependent variables are part of the output or outcome, factors affected by the independent variables. examples of dependent variables are the number of electronic devices in the world since 1950 vs Years since 1950.

The regression analysis has four primary purposes: description, estimation, prediction, and control [4].

**Description**: Describes the relationship between dependent and independent variables.

**Estimation**: When building a relationship, it's possible to estimate the outcome through observation.

**Prediction**: predicting outcome with changing independent variables.

**Control**: controlling the effect of the dependent variable with varying different independent variables.

Types of regression analyses are linear, logistical, and multiple regression.

**Linear regression analyses are best suited for project requirements with one independent and dependent variable.**

* + 1. **“Line of best fit” in linear regression**

The regression line is the line of best fit, provided independent and dependent variables are in a linear relationship,

The regression line is defined with the equation.

Y’ = bx + A

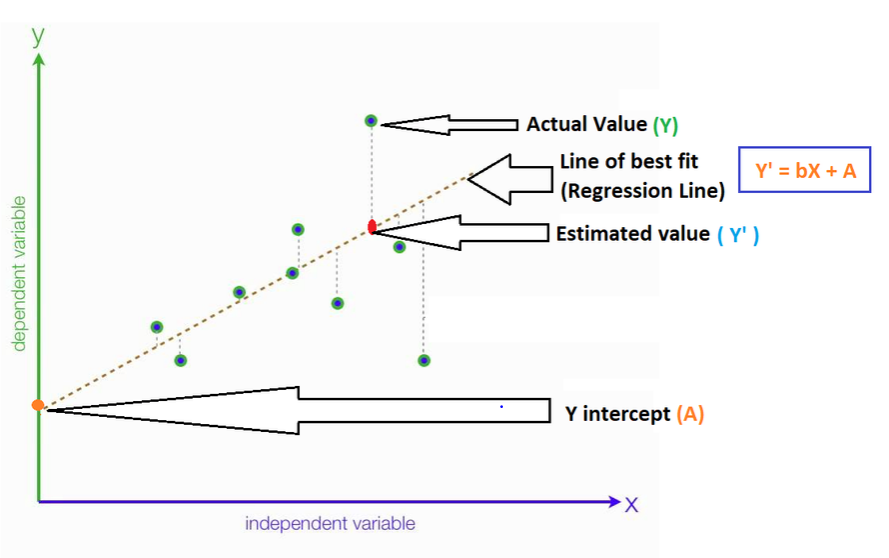
y = point along the Y-axis

b = Slope of a line

x = point along the X-axis

a = Point of the intersection at the Y-axis.

**Figure** 2.3: linear regression equation



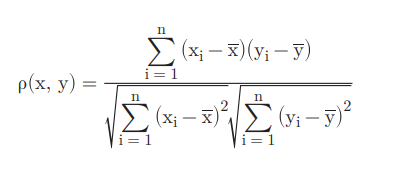
The least-squares method is used for finding the best-fitting curve or line of best for the dataset,

The Least Squares Regression Line is the line that makes the vertical distance from the data points to the regression line as small as possible. It’s called a “least square” because the best line of fit is one that minimizes the variance (the sum of squares of the errors).

**Least squares regression is used to predict the behavior of dependent variables**

* + 1. **Correlation coefficient r and R-squared Linear Regression**

Correlation coefficient r is a statistical measurement of the degree relationship between two variables and the value ranges from -1 to 1, below show the formula for the Correlation coefficient



A value of r close to -1: indicates a negative correlation between two variables.

A value of r close to 0: indicates no correlation between two variable

A value of r close to 1: indicates a positive correlation between two variable

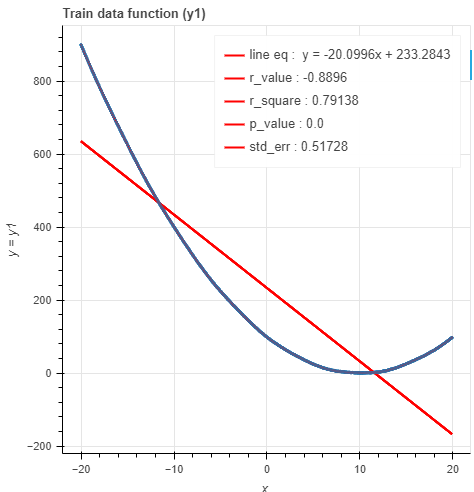
R**-**squared is a measure of how well a linear regression model fits the data. It can be interpreted as the proportion of variance of the outcome Y explained by the linear regression model.[4]

**R Squared is the square of the correlation coefficient, r** (hence the term r squared).

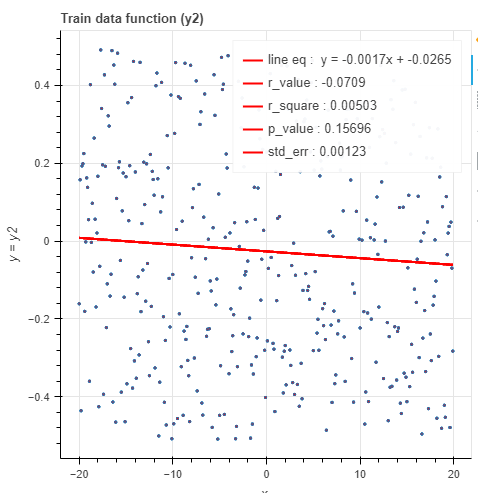
The value of the R-Squared is always between 0 to 1 (0% to 100%).

* A high R-Squared value means that many data points are close to the linear regression function line.
* A low R-Squared value means that the linear regression function line does not fit the data well.
  + 1. **Result of the best-fit line for the train data**

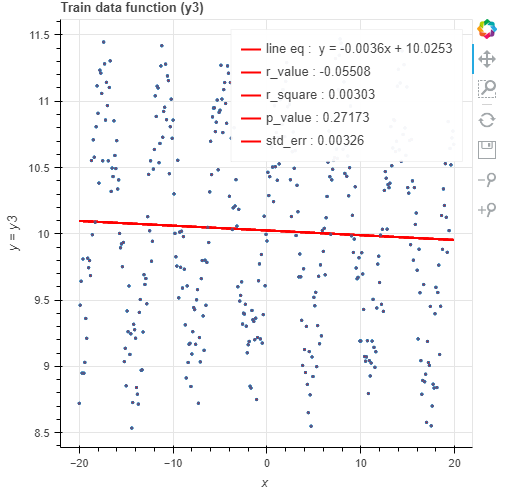
**Figure** 2.4: Best fit line for Train dataset y1



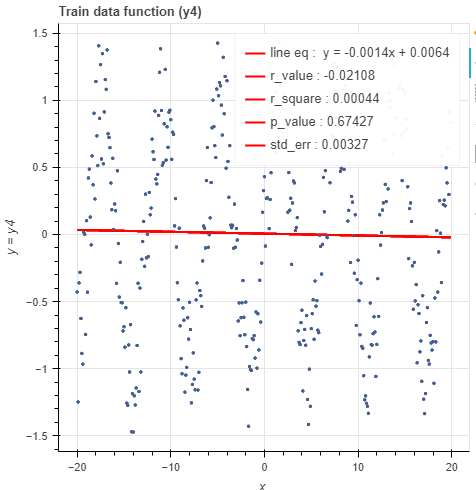
**Figure** 2.5: Best fit line for Train dataset y2



**Figure** 2.6: Best fit line for Train dataset y3



**Figure** 2.7: Best fit line for Train dataset y4



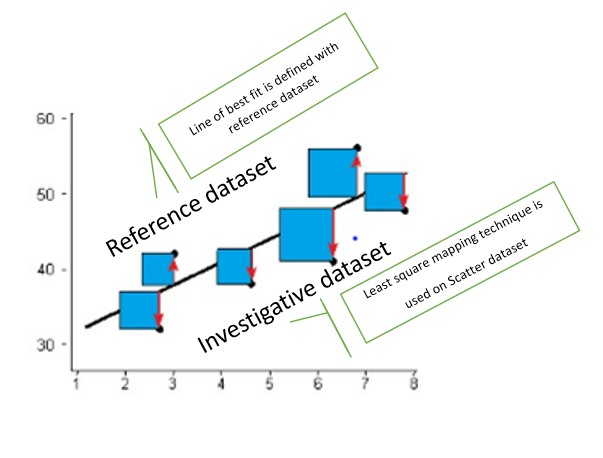
* 1. **Pattern recognition**
     1. **Pattern recognition concept**

The least-squares method used for pattern recognition,

the reference dataset, line of best fit is taken as reference for comparison with the investigative dataset.

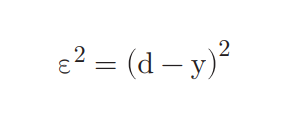
The reference dataset is the training dataset and the investigative dataset is an ideal function

**Figure** 2.8: Pattern recognition using the least square method



The Sum of the Squared Residuals as the Loss Function evaluates how well a line fits the data.

Loss function(ε), where the model output(y) and desired output (d), we find the difference between the model’s output and the desired output.

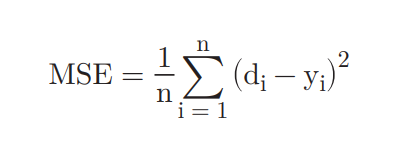


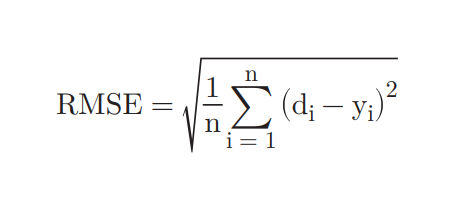
Below, we have a sample representation of the sum of squared residuals,

Sum of squared residuals = (1.4 - (intercept + 0.64 x 0.5))2 + (1.9 - (intercept +0.64 x 2.3)2 + (3.2 - (intercept + 0.64 x 2.9))2

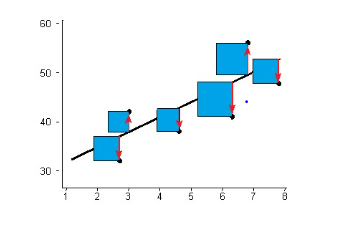
**The ideal function with the least square would be considered the best-fit line**.

Mean Square Error (MSE) and Root Mean Square Error (RMSE) are calculated for easy representation of data on a graph, below is the respective formula.





**Figure** 2.9: Line of best fit using “least square”

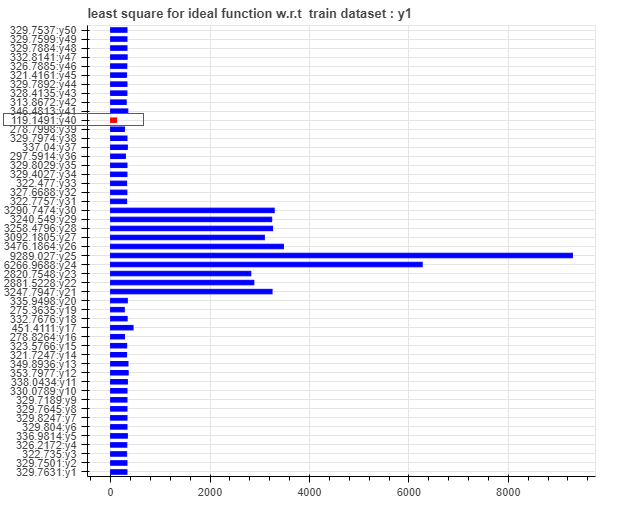


Least-squares are to be calculated for the investigative dataset, ideal function (Y1, Y2 … Y50) concerning each reference dataset (train dataset (Y1, Y2, Y3, Y4)).

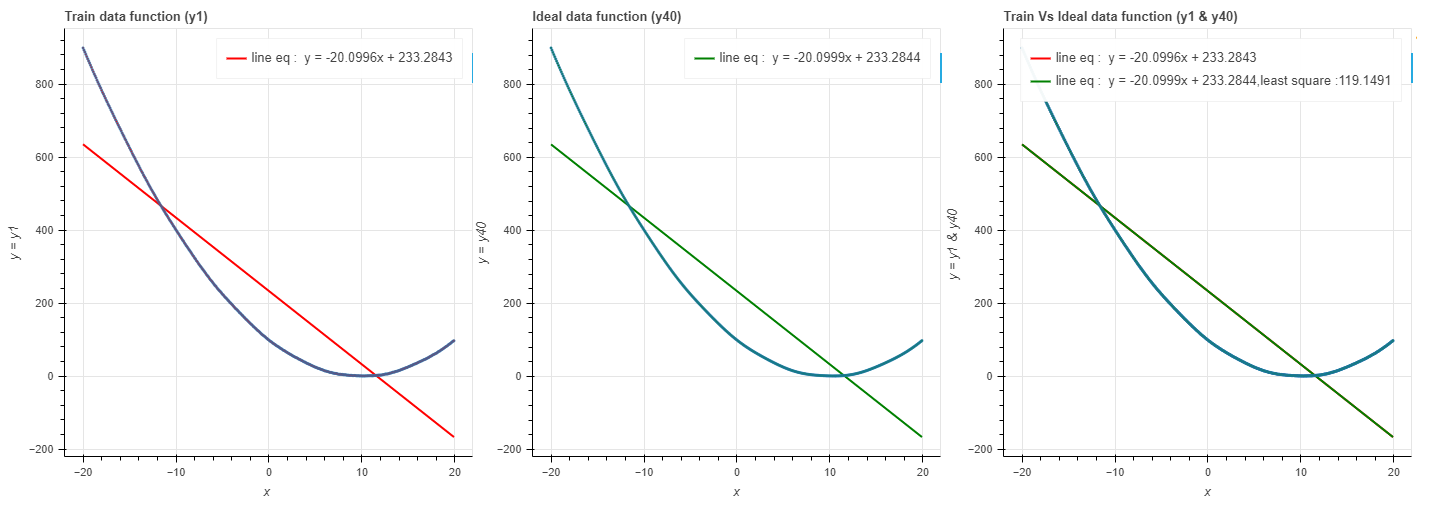
The dataset with the least residual error will be considered near the reference dataset (Y1, Y2, Y3, Y4)

* + 1. **Result of a least square method for ideal function**

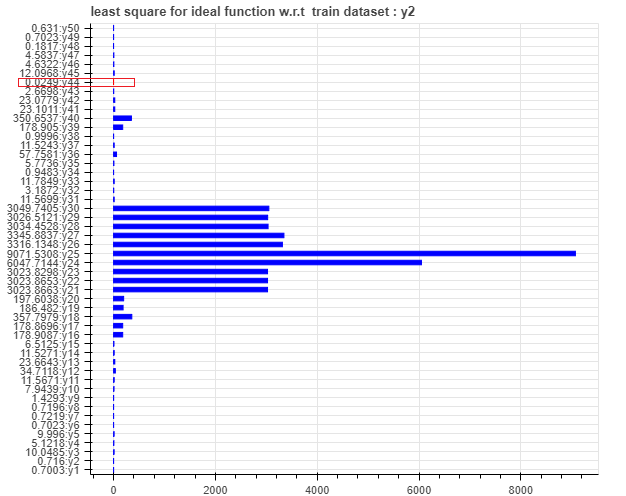
**Figure** 2.9: Least square method for ideal functions and train dataset (y1)

****

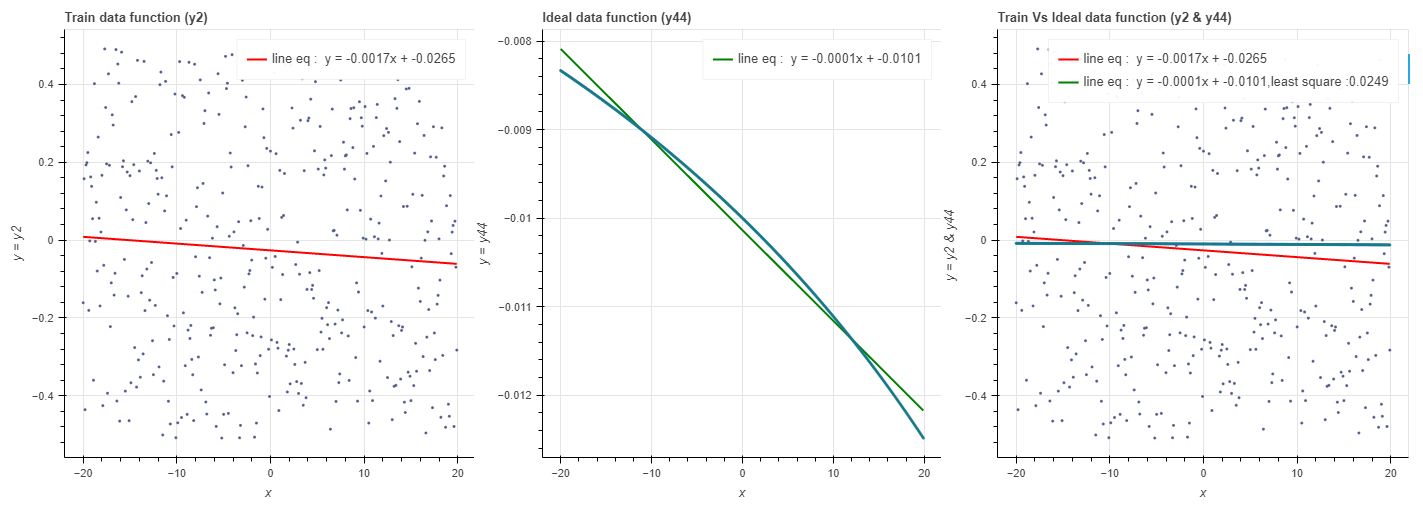
**Figure** 2.10: Scatter plot for train data(y1) and ideal function(y40)

****

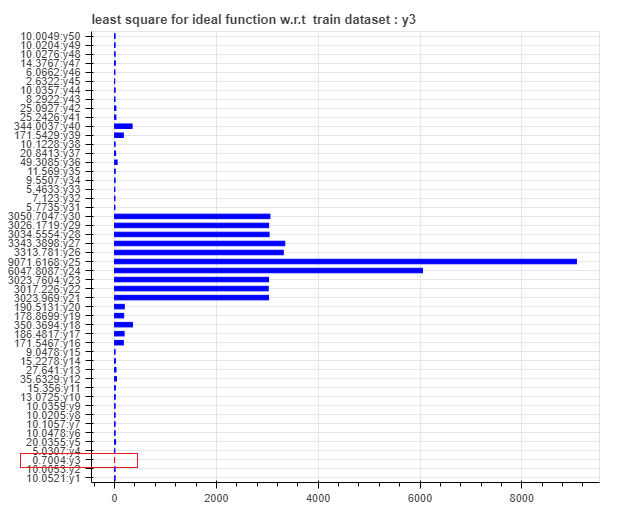
**Figure** 2.11: Least square method for ideal functions and train dataset (y2)

****

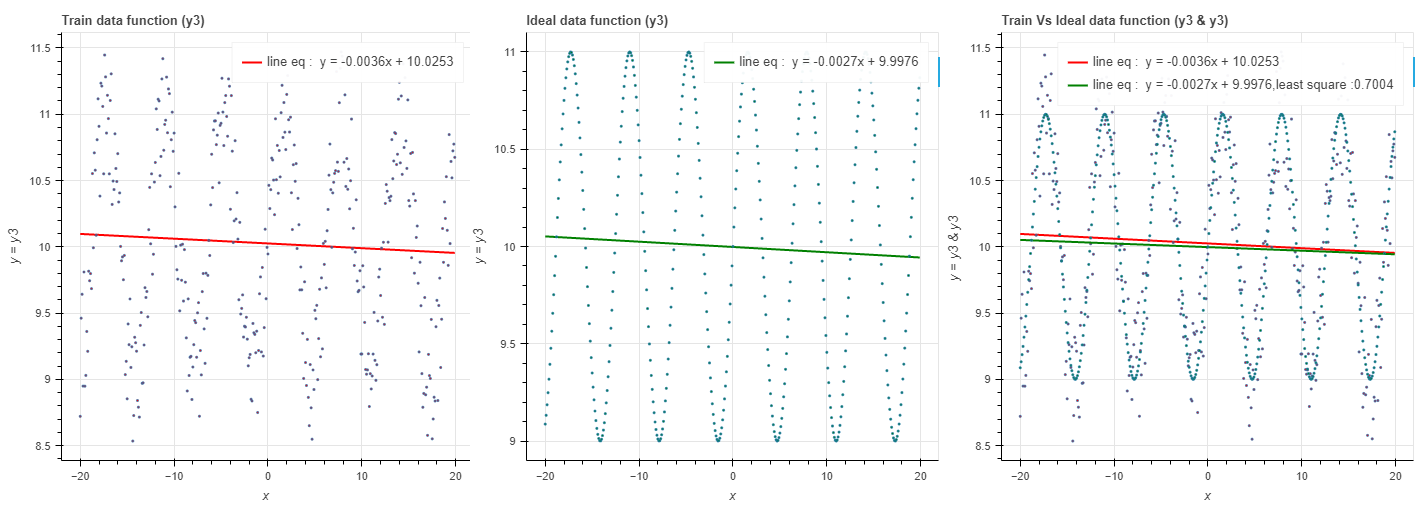
**Figure** 2.12: Scatter plot for train data(y2) and ideal function(y44)

****

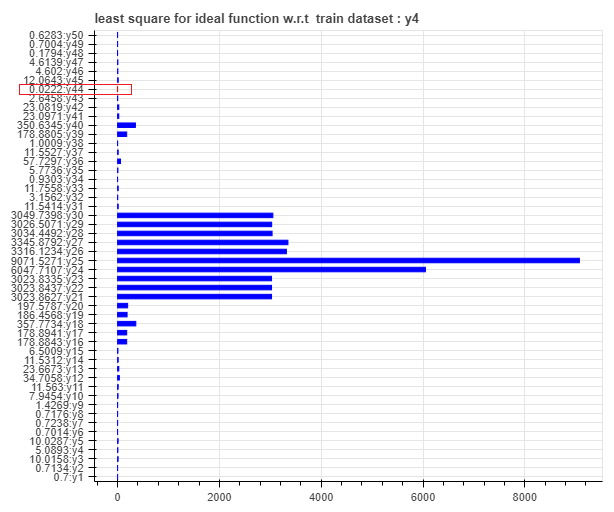
**Figure** 2.13: Least square method for ideal functions and train dataset (y3)



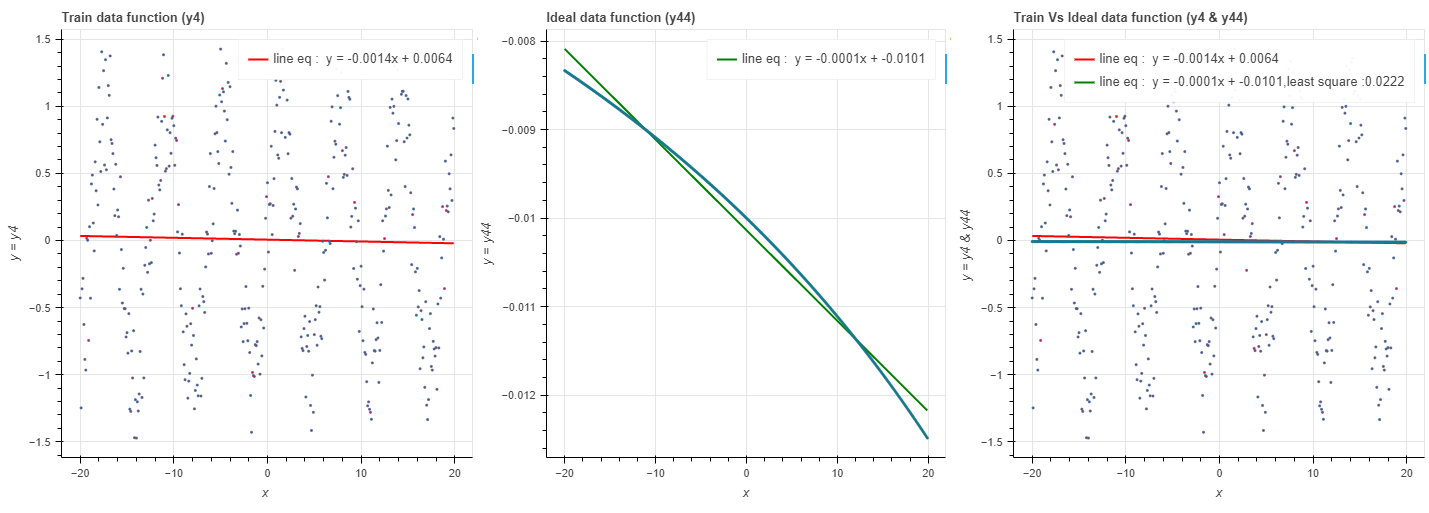
**Figure** 2.14: Scatter plot for train data(y3) and ideal function(y3)



**Figure** 2.15: Least square method for ideal functions and train dataset (y4)



**Figure** 2.16: Scatter plot for train data(y4) and ideal function(y44)



* 1. **Data mapping**
     1. **Understanding test data**

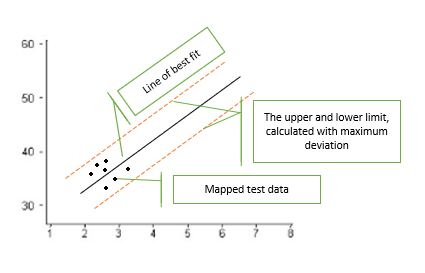
Investigative test data are to be mapped on the ideal function within the limit, the limit is calculated based on the maximum deviation calculated from the training dataset and the selected ideal function,

The below figure describes the mapped test data within the limit of maximum deviation of the ideal and train dataset.

Max deviation is calculated with the difference between the best-fit line of the training dataset and respective ideal function, find the max deviation, and multiply by the square root of 2.

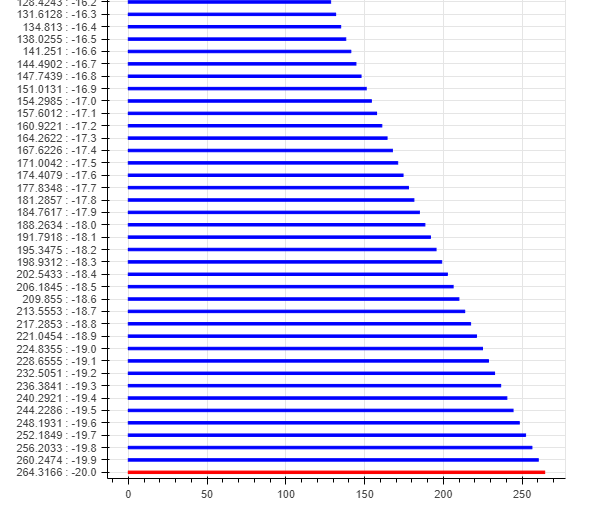
Max deviation can be used to create an upper and lower band and test data can be mapped within this range.

**Figure** 2.17: mapping of the dataset within the limits (maximum deviation)

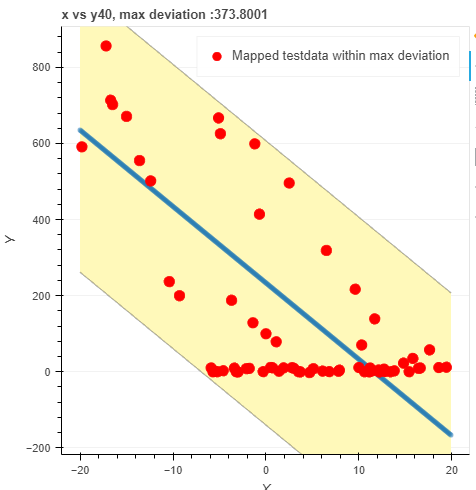


* + 1. **Result for mapping test data within the max deviation**

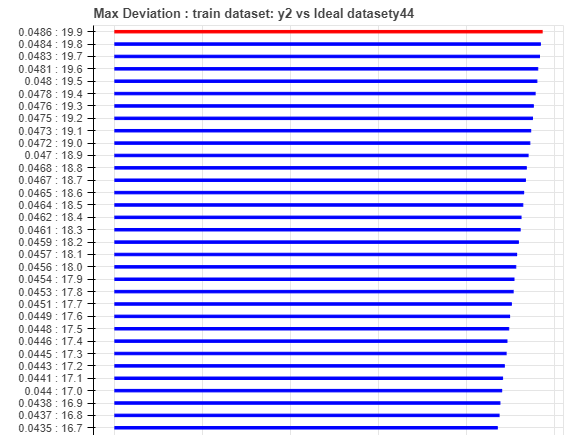
**Figure** 2.18: Horizontal bar chart for deviation of y1(train function) and y40(ideal function)



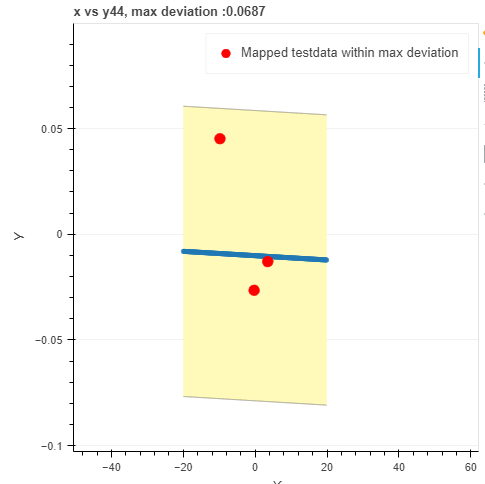
**Figure** 2.19: Mapping of the dataset within the limits (maximum deviation) for ideal function(y40)



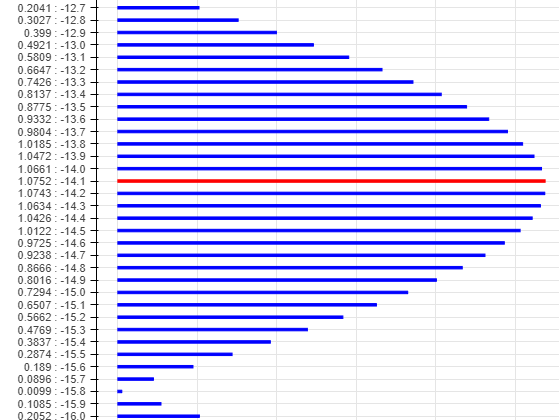
**Figure** 2.20: Horizontal bar chart for deviation of y2(train function) and y44(ideal function)



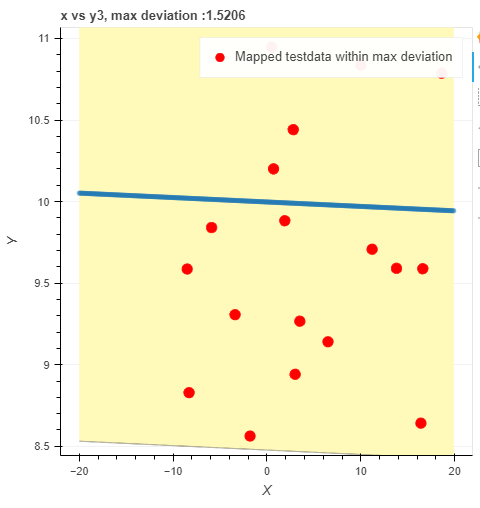
**Figure** 2.21: Mapping of the dataset within the limits (maximum deviation) for ideal function(y44)



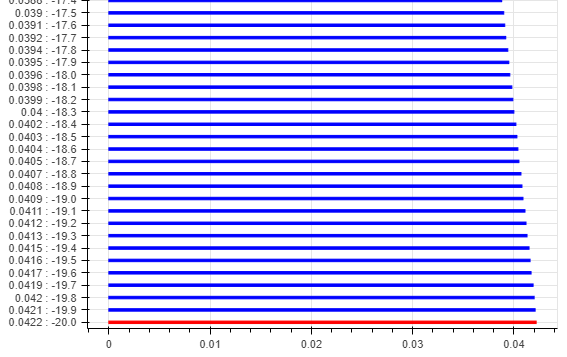
**Figure** 2.22: Horizontal bar chart for deviation of y3(train function) and y3(ideal function)



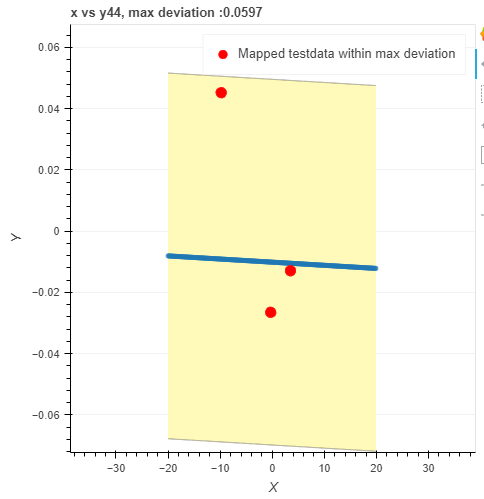
**Figure** 2.23: mapping of the dataset within the limits (maximum deviation) for ideal function(y3)



**Figure** 2.24: Horizontal bar chart for deviation of y4(train function) and y44(ideal function)



**Figure** 2.25: mapping of the dataset within the limits (maximum deviation) for ideal function(y44)



# **Conclusion**

We can conclude the investigation for 4 training function (y1, y2, y3, y4), as described below

3.1. Train function (y1):

* R value (-0.888) negative value, strong negative correlation
* R square (0.7913) close to 1, indicating most of the data is close to the best-fit line.
* Ideal function (Y40) is close match to train function(y1), as per least square method
* 100% of the test point is above the best-fit line and within max deviation (373.8)

3.2. Train function (y2):

* R value (-0.070) negative value and close to 0, slightly negative correlation
* R square (0.005) close to 0, indicate most of the data are not close to the best-fit line.
* Ideal function (Y44) is close match to train function(y1), as per least square method
* 66% of the test point is below the best-fit line and within max deviation (0.068)

3.3. Train function (y3):

* R value (-0.055) negative value and close to 0, slightly negative correlation
* R square (0.003) close to 0, indicate most of the data are not close to the best-fit line.
* Ideal function (Y3) is close match to train function(y3), as per least square method
* 73% of the test point is below the best-fit line and within max deviation (1.5206)

3.4. Train function (y4):

* R value (-0.02) negative value and close to 0, slightly negative correlation
* R square (0.000) close to 0, indicate most of the data are not close to the best-fit fit line.
* Ideal function (Y44) is close match to train function(y3), as per least square method
* 66% of the test point is below the best-fit line and within max deviation (0.068)

**The consolidated conclusion from the above dataset is data is mostly negatively correlated and scattered.**

# **Bibliography**

1. Every block of stone has a statue inside it - ReflectandRespond. <https://reflectandrespond.com/every-block-of-stone-has-a-statue-inside-it-michelangelo/>

2. Cheng, Samantha, et al. “Strengthen Causal Models for Better Conservation Outcomes for Human Well-Being.” PLoS One, vol. 15, no. 3, Public Library of Science, Mar. 2020, p. e0230495.

3. Exploratory Data Analysis Using Python Functions. - Medium. <https://medium.com/illuminations-mirror/exploratory-data-analysis-using-python-functions-67faf725f08>

4. Relationship Between r and R-squared in Linear Regression. https://quantifyinghealth.com/relationship-between-r-and-r-squared/

# **GitHub Page**

1. <https://github.com/Moeet/Analysis>

2. README.md to execute the project

# **Appendix**

1. Code is attached as a Zip file

