zerocode learning

Learning Data Analytics Made Easy

USER GUIIDE

REGRESSION ANALYSIS

TABLE OF CONTENTS

INDEX

1. MODEL- REGRESSION ANALYSIS
2. ALL ABOUT LEFT PANEL
3. DATA INPUT AND OVERVIEW TAB
4. DATA SUMMARY TAB
5. DATA VISUALISAION TAB
6. SUMMARY OLS TAB
7. RESIDUAL-ERROR TAB

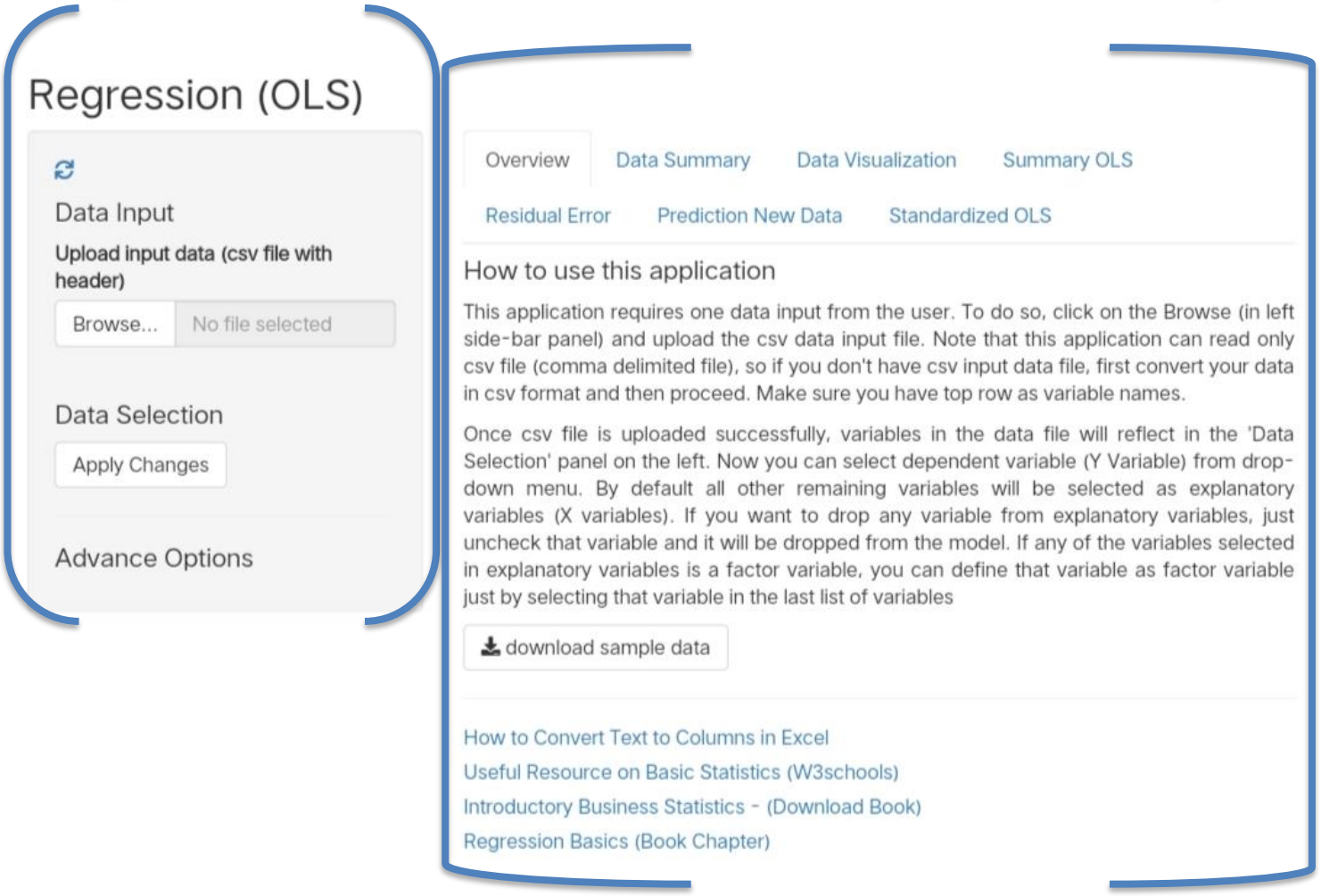
REGRESSION ANALYSIS

01

Regression, Regression analysis is a way of mathematically sorting out which of those variables does indeed have an impact. Here is a simple and useful guide, for you to navigate through various techniques used in Regression, to get the desired outcome, with its intensive methodology.In regression analysis, those factors are called variables. You have your dependent variable — the main factor that you're trying to understand or predict. And then we have independent variables — the factors you suspect have an impact on your dependent variable.

LEFT PANEL (INPUT AREA)

OPERATIONAL ANALYSIS TAB (MAIN PANEL)



LEFT PANEL (INP)

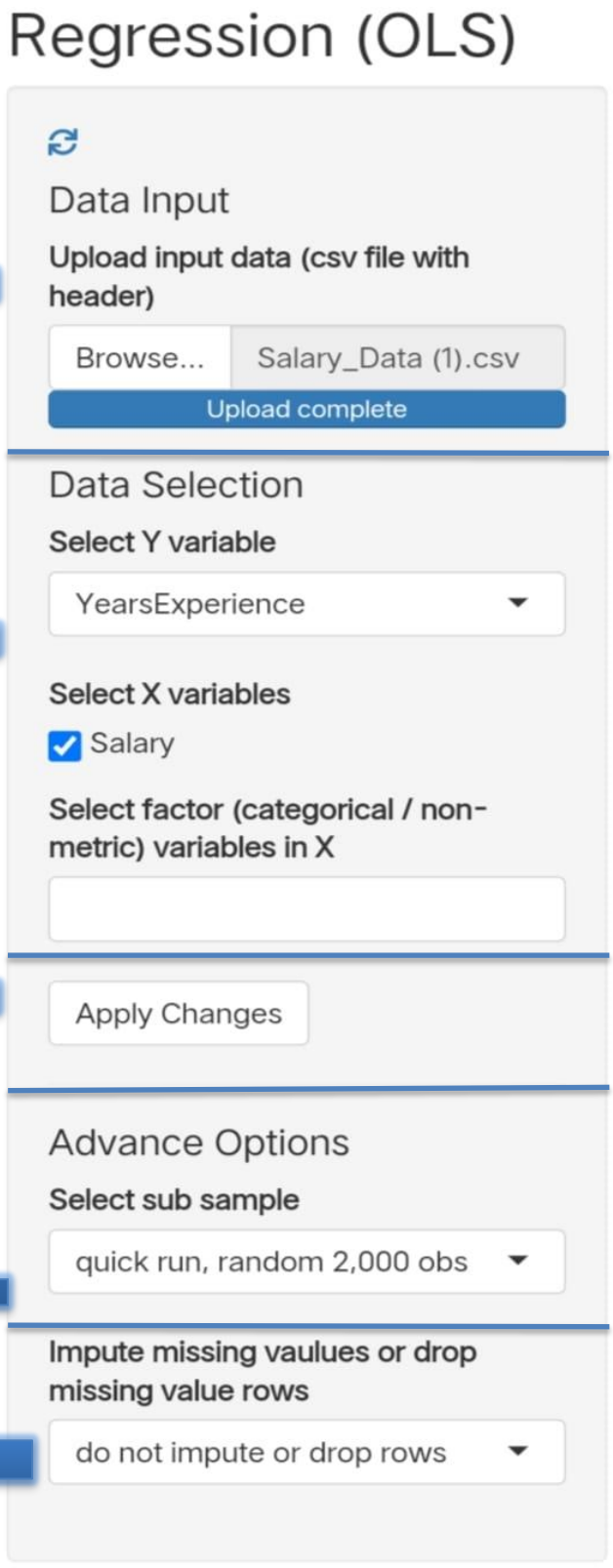
Upload your dataset here

Select your favorable variables required to base the analysis

Apply any changes if you want to do.

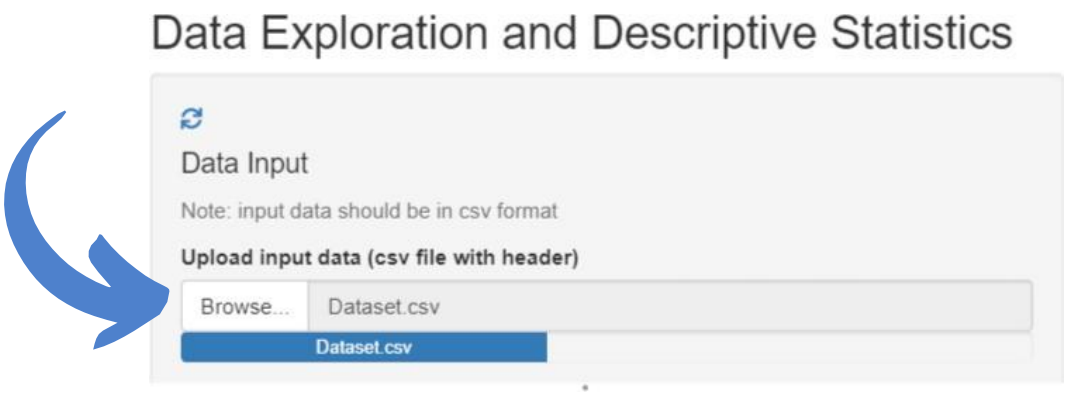
Select the subsamples or the whole data for testing.

Deal with missing values either drop or immute it.



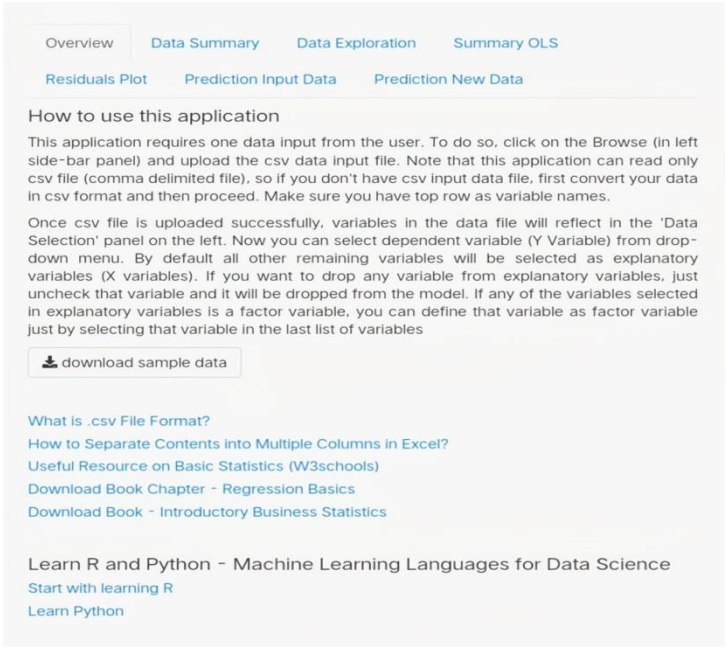
DATA INPUT
(UPLOADING DATASET)

- Click on browse
- Select the datafile that is in the form of csv format.(Ex program.csv)
- Browse the file and select the data to train your model for prediction.
- Top rows of the dataset should be of ‘variable names’.



OVERVIEW TAB

This tab provides you with relevant study resources, tutorials, sample datasets and a short overview to start with, which helps you understand and comprehend your data correctly. This tab also provides you the basic idea about regression analysis , gives sample data and provides the description about regression.



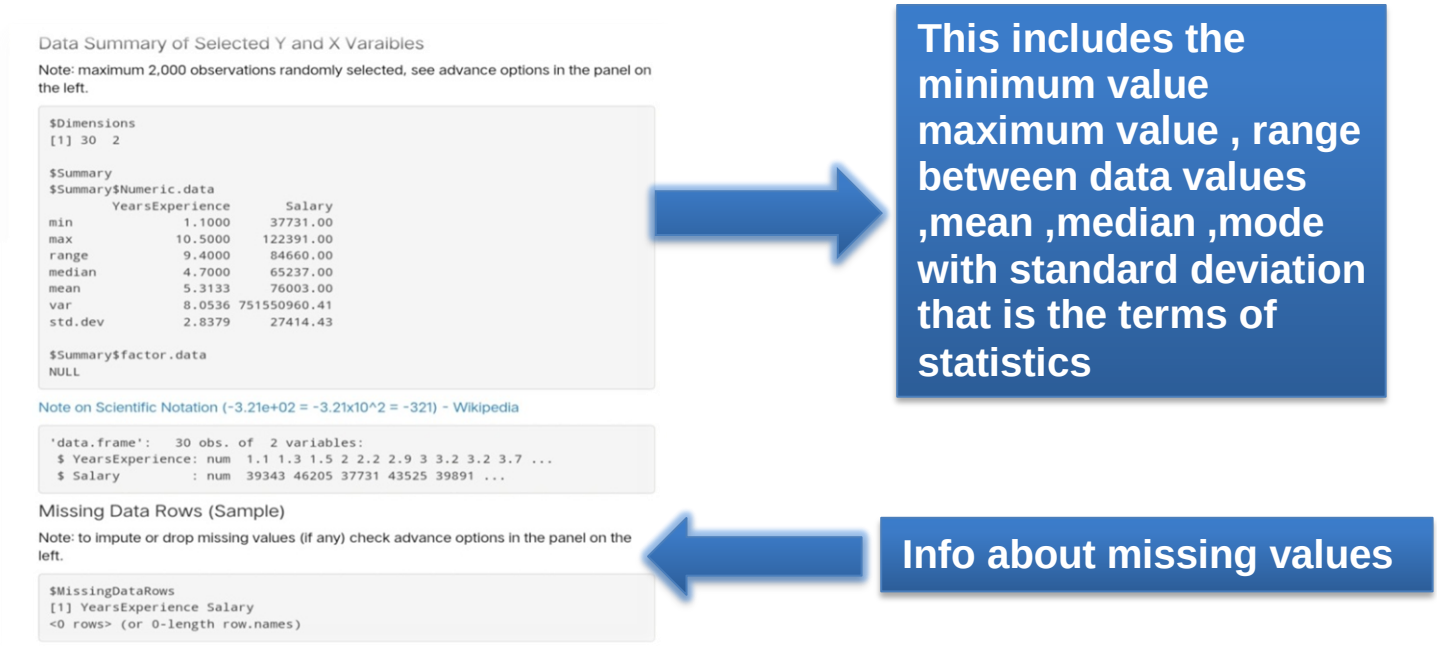
DATA SUMMARY TAB

04

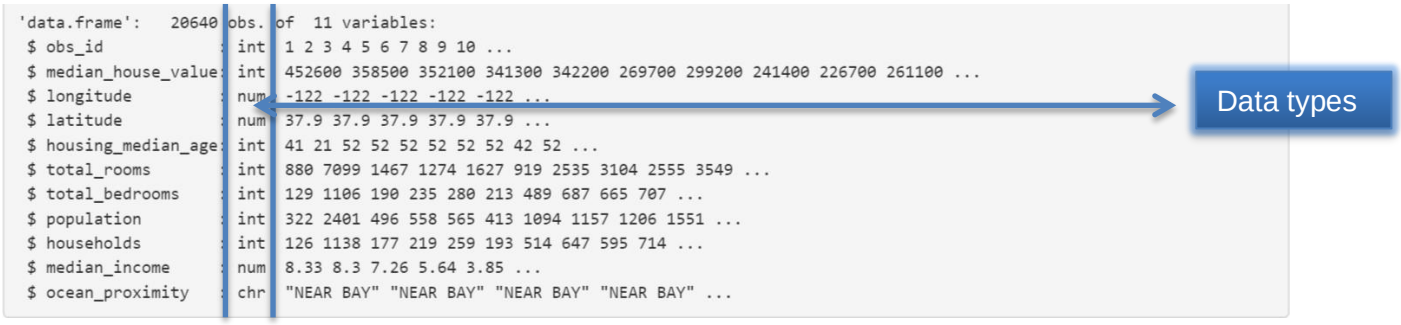
It is very important to understand our data completely to infer meaningful insights and to get an overview of all the data points as a whole, but it is quite impossible to analyze thousand data points manually.

The **'Data Summary'** option enables you to get a comprehensive evaluation through statistical measures that help us form the basis of our analysis.

It will display all the 'descriptive analytics' measures including mean, median, standard deviation, variance etc. for all the data variables present in the dataset. we can review the uploaded data and the contents of it, A brief summary of the data can be seen it includes range of data values, minimum and maximum value missing and null values etc.



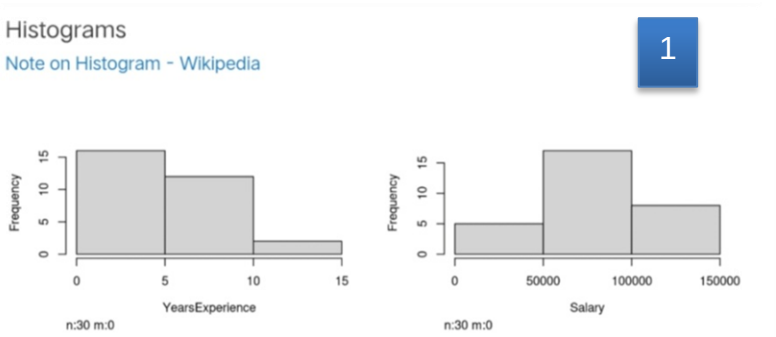
It also segregates dataset variables into respective data types, such as integer, whole numbers, character etc.

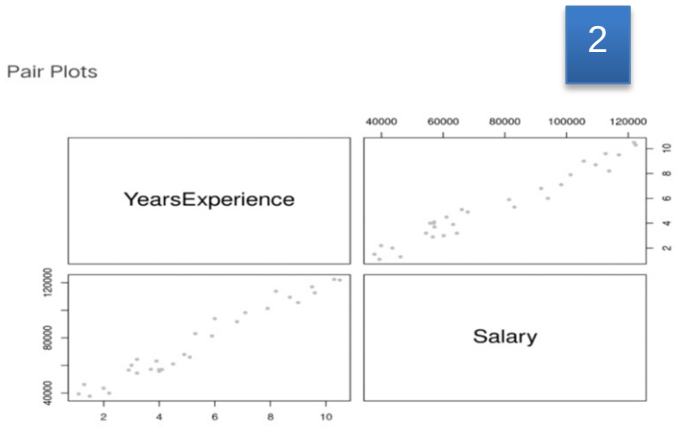


Use the left panel to transform selected variables as per the requirement of analysis ,correspondingly the data summary will also change.

DATA VISUALISATION TAB

Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.





1)HISTOGRAM: A histogram is a graph that shows the frequency of numerical data using rectangles.

The height of a rectangle (the vertical axis) represents the distribution frequency of a variable (the amount, or how often that variable appears). Histograms give a rough sense of the density of the underlying distribution of the data, and often for density estimation,estimating the probability density function of the underlying variable. The total area of a histogram used for probability density is always normalized to 1. If the length of the intervals on the *x*-axis are all 1, then a histogram is identical to a relative frequency plot.

PAIR PLOTS : A pairplot plot a pairwise relationships in a dataset. The pairplot function creates a grid of Axes such that each variable in data will by shared in the y-axis across a single row and in the x-axis across a single column. That creates plots as shown above Pair plot is used to understand the best set of features to explain a relationship between two variables or to form the most separated clusters. It also helps to form some simple classification models by drawing some simple lines or make linear separation in our data-set.

2)CORRELATION TABLE : A two-way tabulation of the relations between correlates; row headings are the scores on one variable and column headings are the scores on the second variables and a cell shows how many times the score on that row was associated with the score in that column.

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

Example: A positive correlation is a relationship between two variables in which both variables move in the same direction. Therefore, when one variable increases as the other variable increases, or one variable decreases while the other decreases. An example of positive correlation would be height and weight.

Use the left panel to modify/deal with the outliers identified here.

06

SUMMARY OLS TAB

The summary tab provides the overall analysis result under few methods like OLS(Ordinary Least Square), RMSE(Root Mean Square Error), Variance Inflation Factor (VIF) etc.

Ordinary Least Squares: Ordinary Least Squares regression (OLS) is a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable (simple or multiple linear regression)

Root Mean Square Error: Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are.RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Based on a rule of thumb, it can be said that RMSE values between 0.2 and 0.5 shows that the model can relatively predict

the data accurately. In addition, Adjusted R-squared more than 0.75 is a very good value for showing the accuracy. In some cases, Adjusted R-squared of 0.4 or more is acceptable as well.

Variance inflation factor : Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable. we can calculate the VIF for the variable points by performing a multiple linear regression using points as the response variable and assists and rebounds as the explanatory variables.

The VIF for points is calculated as 1 / (1 – R Square) = 1

R-Squared : 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 other words, r-squared shows how well the data fit the regression model (the goodness of fit)

The usefulness of R^2 is its ability to find the likelihood of future events falling within the predicted outcomes. The idea is that if more samples are added, the coefficient would show the probability of a new point falling on the line. Even if there is a strong connection between the two variables, determination does not prove causality.

CORRELATION PLOT : The correlation coefficient is the specific measure that quantifies the strength of the linear relationship between two variables in a correlation analysis. The coefficient is what we symbolize with the r in a correlation report. A correlation analysis provides information on the strength and direction of the linear relationship between two variables, while a simple linear regression analysis estimates parameters in a linear equation that can be used to predict values of one variable based on the other.

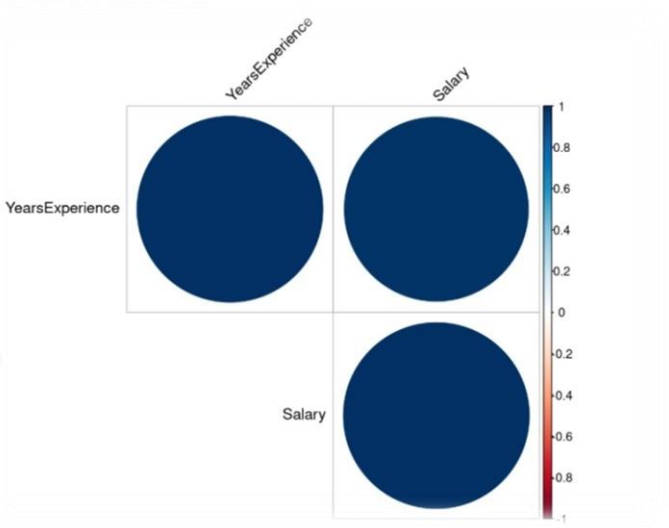
Summary Regression Model (Ordinary Least Square)

Y is YearsExperience

```
Call:
lm(formula = formula, data = test_data())

Residuals:
     Min       1Q   Median       3Q      Max
-1.12974 -0.46457  0.04105  0.54311  0.79669

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.383e+00  3.273e-01  -7.281  6.3e-08 ***
Salary       1.013e-04  4.059e-06  24.950  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5992 on 28 degrees of freedom
Multiple R-squared:  0.957,     Adjusted R-squared:  0.9554
F-statistic: 622.5 on 1 and 28 DF,  p-value: < 2.2e-16
```

Interpreting beta (Estimate) - one unit increase in X variable increases the outcome Y by beta units.

This tab evaluates the correlation coefficients between variables and represents them through a correlation map as shown, where each cell depicts acorrelation between 2 variables. The size and colorof the circles in each cell depict the degree of correlation, the larger the size and darker the color shade; the higher is the correlation.

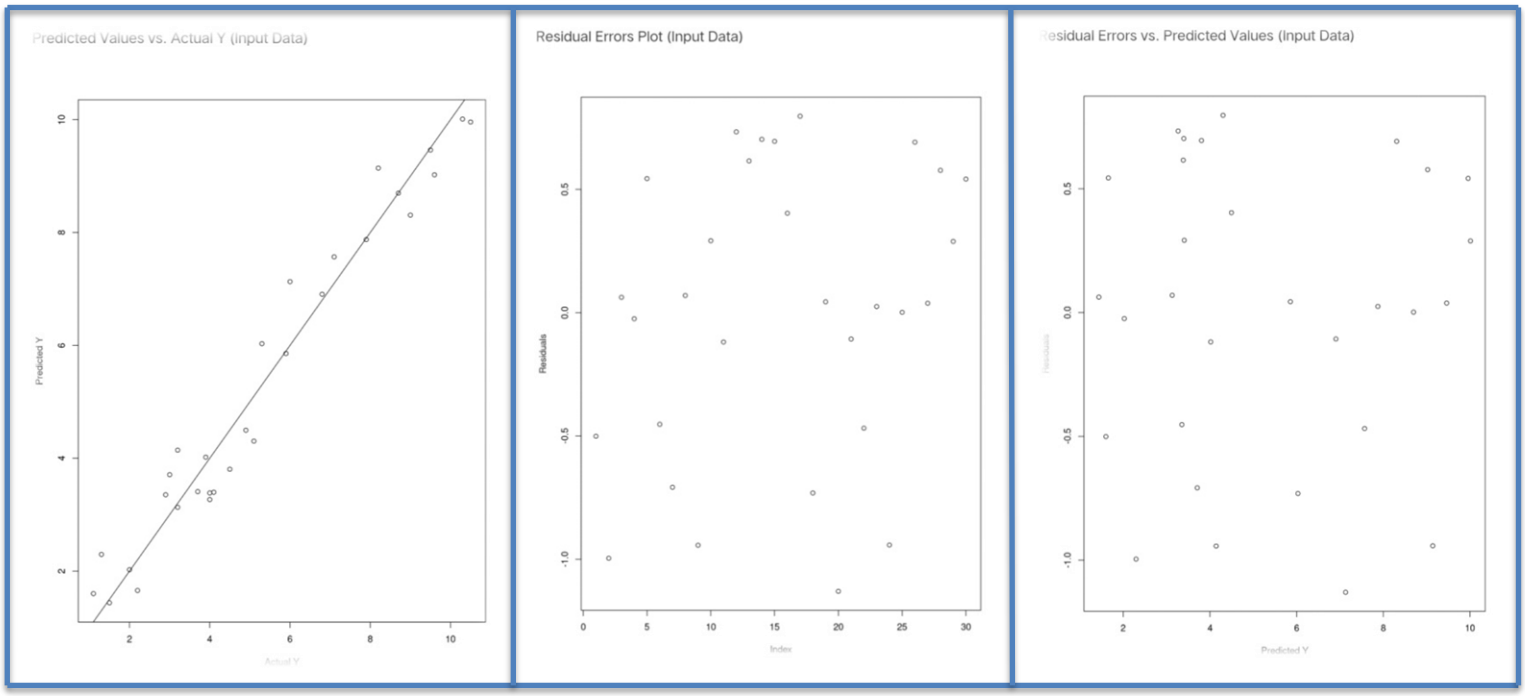


07

RESIDUAL ERROR

A residual is a measure of how far away a point is vertically from the regression line. Simply, it is the error between a predicted value and the observed actual value. A typical residual plot has the residual values on the Y- axis and the independent variable on the x-axis. As residuals are the difference between any data point and the regression line, they are sometimes called“errors.” Error in this context doesn’t mean that there’s something wrong with the analysis; it just means that there is some unexplained difference. In other words, the residual is the error that isn’t explained by the regression line.

Prediction error quantifies one of two things: In regression analysis, it's a measure of how well the model predicts the response variable. In classification , it's a measure of how well samples are classified to the correct category.



EXAMPLE GRAPHS

The equations of calculation of percentage prediction error
(percentage prediction error = measured value - predicted value measured value× 100
or
percentage prediction error = predicted value - measured value measured value ×100) and similar equations have been widely used.

Use the left panel to impute or drop the missing values identified here