Analysis:

1. Basic information:

In the exploratory data analysis, we mostly use python as the preferable scripting language to perform all sorts of statistical operations and the data analysis and the data visualization. To carry all these steps we can make use of few python libraries such as pandas, numpy, matplotlib, seaborn etc,. Each of the packages have different functionalities.

Numpy is also known as numerical python which is used to carry-out wide variety of statistical operations. Multi-dimensional arrays are created, and data reshaping and data handling actions can be carried out using these arrays. Pandas is another huge python libraries which performs data reading action and data storing. Data frames can be created using pandas and these data frames are later used for multiple data analysis actions.

Matplotlib on otherside used for data visulaisation and data plotting. Bar graphs, histograms, pie charts are visualized with the help of this library and various plots like line plot, scatter plot, multi line plot etc,. Seaborn with the statistical inference methods in it is used for the dislay of graphical structures which are better visualized in comparison with the matplolib.

2. Data reading:

There are different types of data formats available in the world of data science. File formats like csv(Comma Separated Values), tsv(Tab Separated Values), json files, dictionaries, xml etc,. All these formats are supported documents for the data storage and we mostly make use of csv files to store the data.

We now use pandas to read all these data files to perform various data operations. There are built in functions available in pandas library which are used for reading csv files. For this, we have to create a variable to hold this method and this method requires parameters like csv filename and any na values available in that dataset. This can be read as a data frame and we then use it for displaying the data.

Data can be displayed in various formats and few of them are listed here. Head method is used to display the top five rows as a default limit. One more is location method, here we choose a particular location with the row number or the data column to display the desired data among the data available.

3. Data handling:

There are few data handling techniques available in python for the data analysis. To know the total count of the rows and columns in the dataset we use shape method which reflects only the count of total count of rwos x columns. but count method will revert the individual column length whether the data exist or not.

The main steps included in the data handling is filling the null values which will lead to errors in accuracy of the model in the future. Before that we have dropped the duplicates which are available in the assetset and we took the count again. We didn’t find any duplicates available in the data frame. Data handling for the null values is that, we find the null valued rows first. We got 201 rows with null values for bmi column and we replaced those null values with the mean value. And after filling those mean values in place of the null values, we got the null values count as 0 and it states that all the columns are having data.

Another hazard for the extensive error causing factor in the dataset is with the outliers that exist in the dataset.

Outlier - 1

Chart, box and whisker chart

Description automatically generated

Here for this avg\_glucose\_level column, we have the following outliers and the data above 150 are out of the maximum outlier range and it is safe to remove the data points which are not in the boxplot.

Outlier - 2:

A picture containing text, antenna

Description automatically generated

When it comes to the bmi column, we have got little amount of outliers while lie out of the maximum range of the boxplot and which is around 50 on the x axis as the bmi range.

Here we have removed the data points which are apart from the inter quartile range which means the data points that are available below to the minimum quartile 25% and the data points available maximum to the 75% quartile and displayed the data below.

4. Label Encoding:

Yet this label encoding is another key pattern which is required to perform any premium statistical calculations like finding out the p-values, f stats etc,. Label encoding is carried out for the catergorical columns to represent them in numerical forms. We have performed this action for the below columns. They are gender, work\_type, smoking\_status, ever\_married. And if we display the data we can find the data as numerical form only.

A screenshot of a computer

Description automatically generated with medium confidence

5. Data analysis:

We now entered the main sequence where the actual data statistics begins and we then carry forward this data to the model building. The statistical dependencies between the variables in comparison to the stroke column is visualized below.

A picture containing text

Description automatically generated

But each factor has the equal and variant effect with the gender also. It can be displayed and interpreted as follows.

Here gender has the data for the stroke value and it is highly variant among the available data. Nearly female has stroke for about 1500 people and male has the remaining data count for the stroke value.

Chart, histogram

Description automatically generated

Smoking status is another variable which has label encoded and each of the encoded value has the stroke value and which has different effect and can be visualized easily using the seaborn.

Chart, histogram

Description automatically generated

As we know there exist people with multiple age groups and are having stroke with all different age groups and the count can be displayed as calculated.

Chart, bar chart, histogram

Description automatically generated

Another factor that leads to the stroke is work type that people prefer to do so and it is statistically displayed which can be interpreted for various results.

Chart, histogram

Description automatically generated

The clear evidence for the each column depency with the stroke variable and it is witnessed through the heatmap. Here the highlighted colours are having the high correlation between the variables.

Analysis:

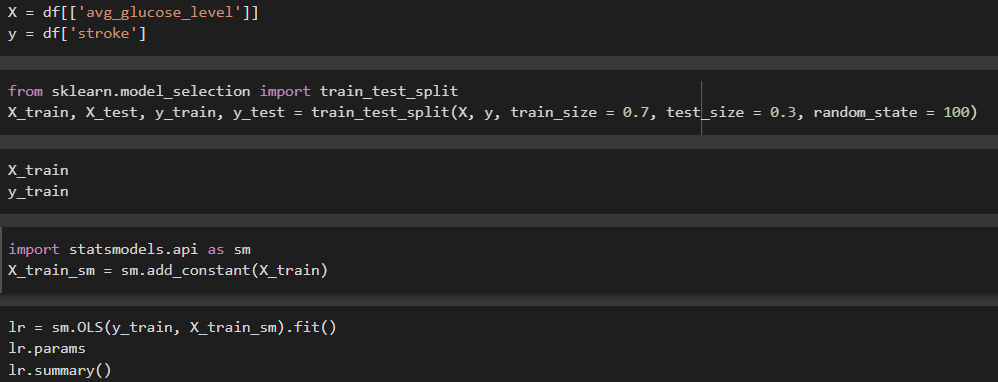
Statistical interpretation is the heart of the data analysis and her we use ordinary least squares method. From the ols summary table we can interpret the follow statistics. Dependent variable is stroke grade level and it is measured here. We know the data and the time variable which are when the data is created.

Total number of observations we took for the statistics here are displayed and 700 and the degree of freedom is 698 which is n-k-1 are we know that it is 2. Rsquared can be calculated as the dependency of independent variable with the dependent variable and it is shown as 0.934 in our case. The next factor is adjusted r squared which is important to calculate the multiles dependant variable. We can interpret the r squared as the less value will refere as few variables have no purpose in the statistical contributions.

Another factor in the ols summary table is the intercept, which is change in variable effects the independent variable. Here we got 5.64 which refers to that one unit of x will change to that many times as an intercept value. But it is note worthy about intercept that when it is negative, then we can assume that relationship would be inverse and one factor increase will effect in other direction.

Std error depicts that the estimates of standard-deviation of co-efficient which is the measurement of variation in data points. Here p-value is less than 0.05 which is less than the significance levela and the null hyposthesis is rejected. We also got the f stats intervals where 0 can not be in the range of interval we got and we can conclude that f-test is rejected.

Preliminary results:



As the results we got, we can interpret that the p-value is comparatively less than the threshold of the significant level and we can reject the null-hypothesis with the 0 value. And for the f-test results we never found the 0 value fit in the confidence-interval which is 4.394 to 6.888. The dependency of the variables is highly dependent on the independent variables.