Women Entrepreneurship and Labor Force Data Analysis

(using Python 3.7)

Project Delivery

(Documentation, Implementation)

Instructor: Dr. Mahmuod Mounir

Program: Software Engineering

Course Name: Statistical Analysis

Course Code: CIS240

Delivery Date: 29 /03/2021

Team members: Fayrouz Tarek

Noran Amr

Samaa Alaa El-Din

Khadija Tarek

Salma Hany

Table of Contents

Abstract

It’s common knowledge that Entrepreneurship is the engine of economic development of each country, which is why this dataset helps provide a visual representation of how women entrepreneurship has been on the rise recently. The dataset analysis shows us that it is notable that female entrepreneurship is expanding in all countries, and the increasing number of companies whose owners/ managers are women, have been reported in many countries. Every day female entrepreneurship becomes a source of new employment in developing countries and it could be said that it also contributes to the balancing of the economic development of the countries as shown in the dataset. This paper goes through explaining the dataset, the program developed flow, the methodology of gathering the data and there is a user interface guide for the users of the program. It shows also the pieces of code provided in addition to test samples. The discussion and hypothesis could be found as the last section of the document along with the conclusion.

1. Women Entrepreneurship and Labor Force

1.1. Dataset

The FEI* tries to find out what are the factors that enable the thriving of women entrepreneurs. The FEI systematic strategy collects the data from different countries which allows us to see variation in gender inequalities conditions that could impact the development of female entrepreneurship. The FEI focuses on recognising a country's pros and cons when it comes to providing desirable conditions that could lead the development of female entrepreneurship. In 2015, FEI covered 77 countries collecting the dataset that consists of women entrepreneurship rate, female labor force rate, inflation rate, etc.

**FEI: Female Entrepreneurship Index*

1.2. Project/ Program Content

The program/ application was constructed using Python programming language. There are three tabs the user could transfer from, the "User Data" tab is designed for the user to select the desired dataset and perform the procedures upon. Next is , the "Numerical data"tab along with the last one , the "categorical data" tab.

In the first tab contains or is for:
The user can upload a dataset that will be used in the analysis. Then the user will choose the column they want to perform the analysis on by typing the column name in the textbox. The user can then choose to calculate any value for the chosen column or display a graph.

In the other two tabs they are working on fixed data, there are three sections for:

- First section: The user should choose the column they want to perform the analysis on.
- Second section: The user selects any value they want that could be calculated for this column.
- Third section: The user picks a graph they want to be displayed for the column they chose.

1.3. Methodology of Collecting the Data

Upon searching for datasets on the internet, this was the dataset chosen from 'kaggle.com'. The dataset collection is mostly based on stratified sampling, the population is divided into female and male gender. The dataset is using only one stratum which focuses on females. A sample is taken from this stratum based on systematic sampling as the creator specified. The data used for the Female Entrepreneurship Index encompass both individual and institutional level data. For the individual level data it is collected from the GEM* dataset, it is said to be specifically gathered from the 2010-2012 Adult Population Survey.

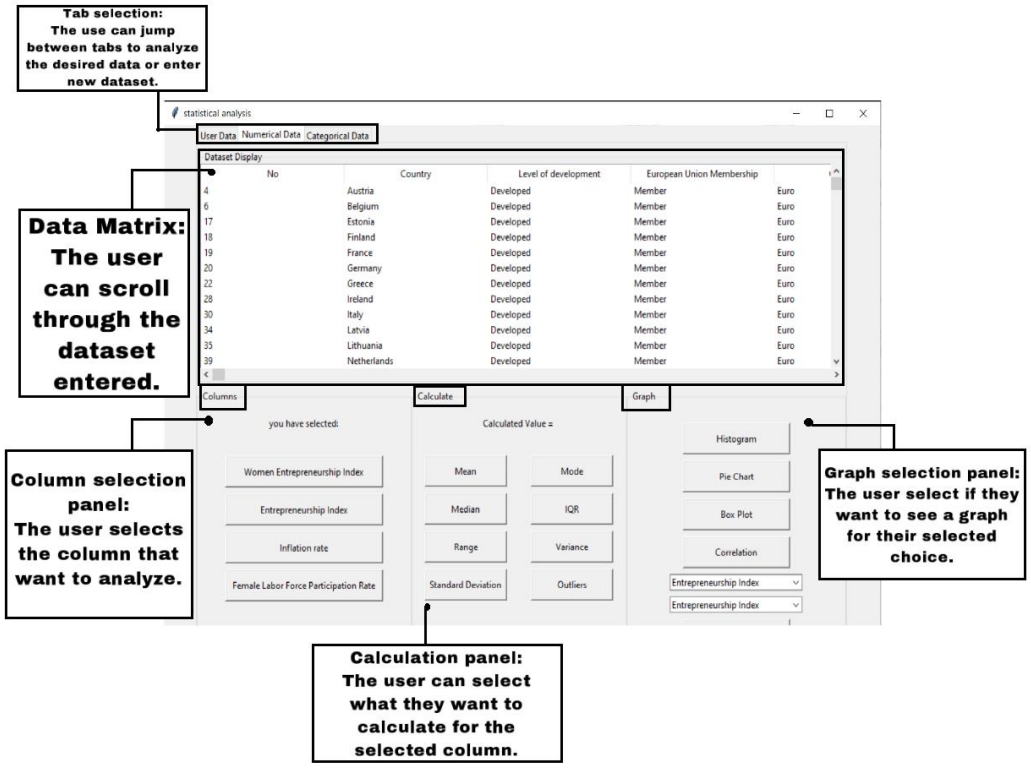
Individual level data is based on attitudes and observations upon the feedback from women of ages between 18 and 64 out of the population. Institutional level data was collected through three levels of analysis, the first level includes the systemic foundations that the entrepreneurs are affected by, which include the Business Freedom collected by the Heritage Foundation and based on the WEF*. The second stage of analysis included institutions involving unequal gender rights or roles. The third level consists of areas where women have limited access to services than men, to capture the overcrowding of the labor force by men and women.

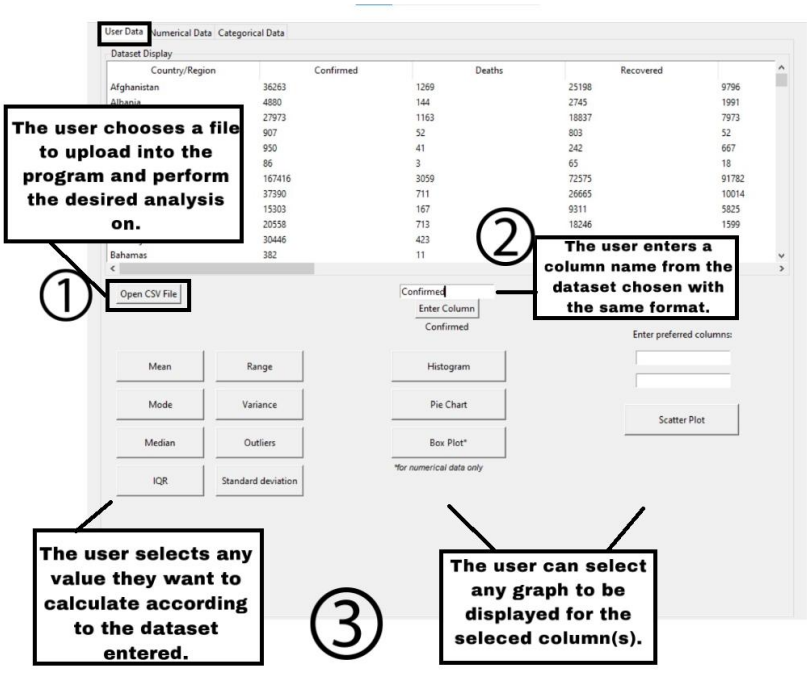
**GEM: Global Entrepreneurship Monitor*
**WEF: World Bank's Forum*

1.4. Objectives

This data analysis report aims to highlight all the needed information regarding the dataset that the user provides or the fixed data already available. And in our case we work on The Women Entrepreneurship & Labor Force dataset that has information on how women have impacted the labor force in the developing countries and the number of Female Entrepreneurs is and continues to be on the rise as seen in the Female Labor Force Participation Rate .

1.5. User Interface Guide





2. Code Implementation and materials used

2.1. Information about the Code and Materials

- Python libraries used:
  (tkinter, pandas, numpy, matplotlib, seaborn, scipy, sklearn).

- Functions implemented to get the:
  (Mean, Mode, Median, IQR, Range, Variance, Standard Deviation, Outliers).

- Graphs displayed:
  (Histogram, PieChart, Box Plot, Correlation, Scatter Plot).

- Connecting the GUI with the code and the functions created:
  Using tkinter from the python GUI library where we used widgets (labels, button, listboxes) so the user can navigate through the functions and display the outputs through labels. Along with the treeview from tkinter to view the columns of the dataset easily and listboxes to make them choose what they want.

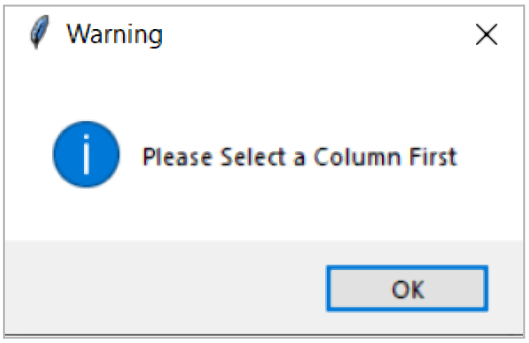
- Code editors used to develop this program:
  - JupyterLab online notebook python language
  - Microsoft Visual Studio 2019 Version 16.6.2*

2.2. Code Implementation

Note: the following section includes testing data on the "inflation rate" column for numerical data and the "level of development" column for categorical data to keep the results consistent.

Validation

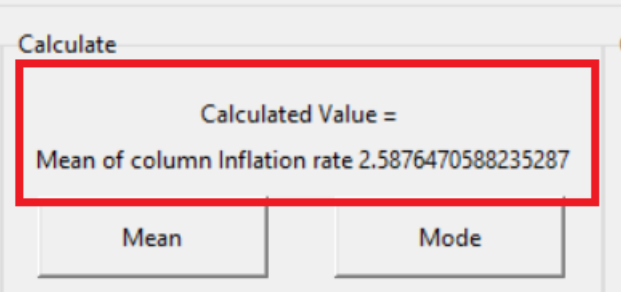
If the user didn't select any column from the "Columns" panel the program will display a warning pop-up message.



Measures of Central Tendency:

- To get the mean of any selected column:

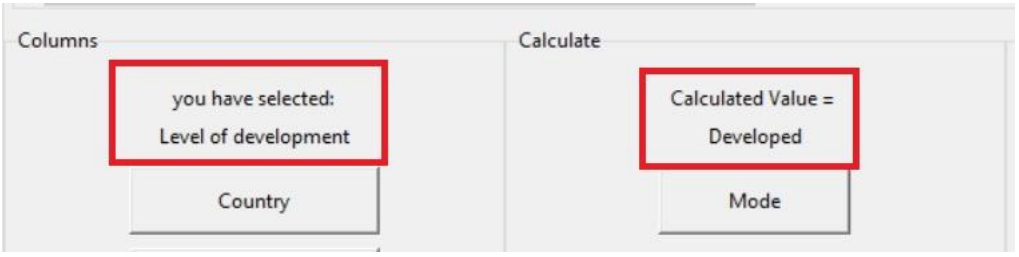
```
def getMean(columnName, dataset, LabelAns):
    if isPressed(columnName):
     print(columnName)
     columnData = dataset[columnName]
     print("Mean of column " + columnName + " " +
str(np.mean(columnData)) + "\n")
      LabelAns.config(text = "Mean of column " +
columnName + " " + str(np.mean(columnData)))
```



- To get the mode for categorical data:

```
def getMode2(columnName):
```

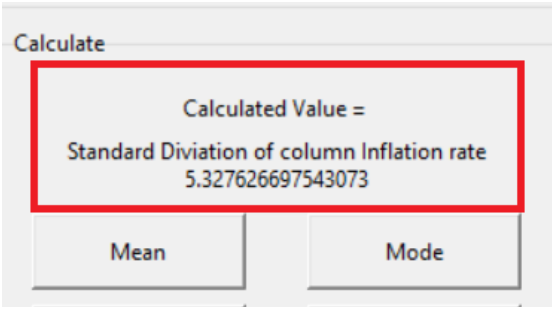
```
    if isPressed(columnName):
     print(columnName)
     columnData = data[columnName]
     label_answer2.config(text
=str(stat.mode(columnData)))
```



Measures of Dispersion/Variability:

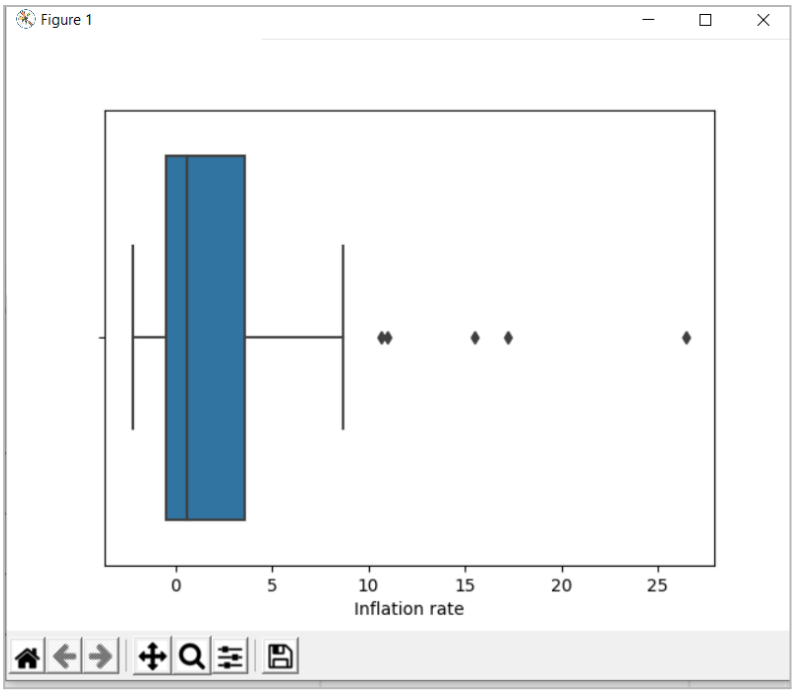
- To get the standard deviation of any selected column:

```
def getSD(columnName, dataset, LabelAns):
    if isPressed(columnName):
     print(columnName)
     columnData = dataset[columnName]
     standD = np.std(columnData)
     LabelAns.config(text= "Standard Deviation of column
"  + columnName + " "  + str(standD))
```

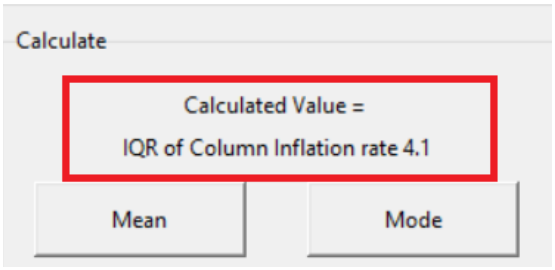


- Box Plot - IQR - Outliers - Range

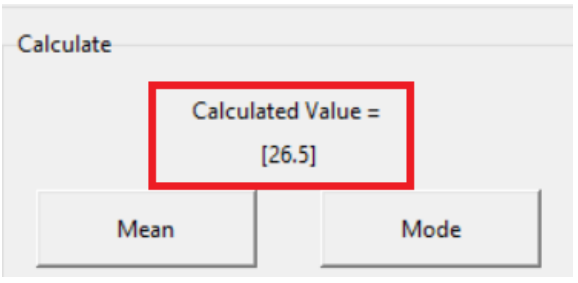
Box Plot:



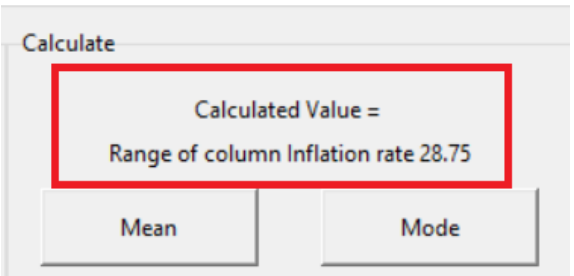
IQR:



Outliers:



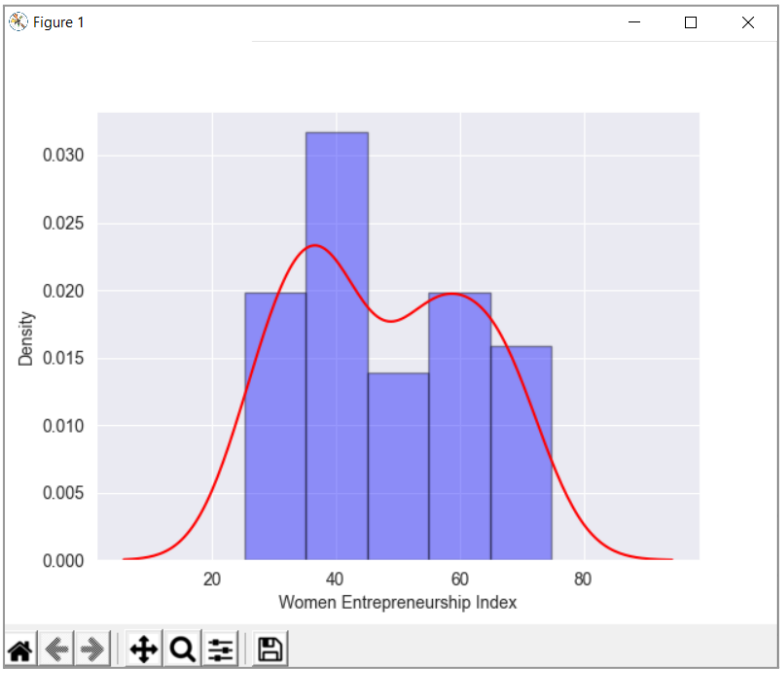
Range:



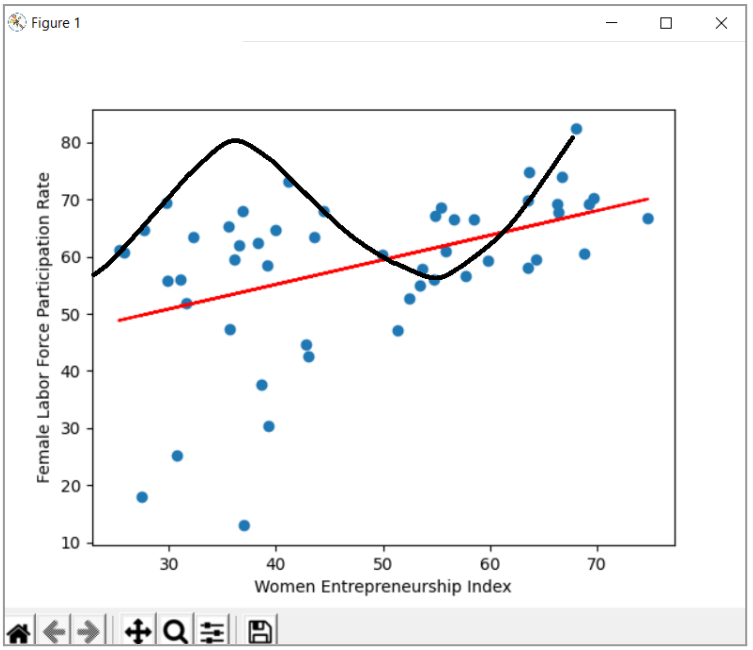
3. Discussion and Hypothesis

Discussion Topic 1: Normal Distribution

The Graph that represents the normal distribution of the "Women Entrepreneurship Index" is displayed through the program as follows:



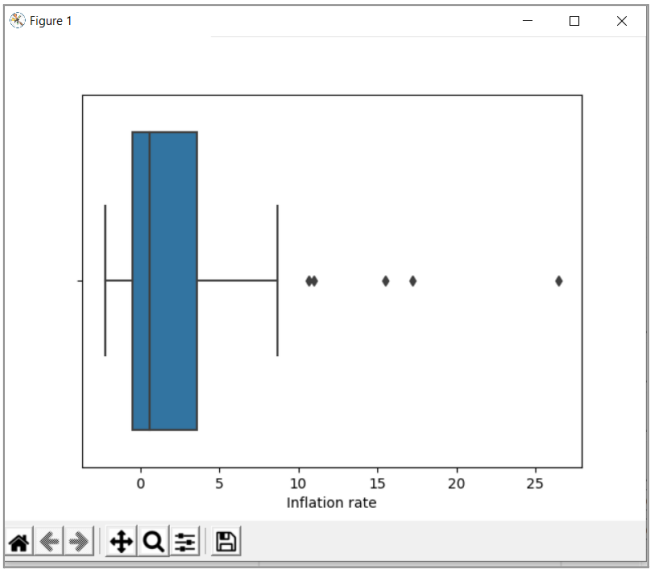
The graph shows our distribution with two bumps/ peaks from which we can tell it is a Bimodal distribution. This is different from the bell-shaped curve we usually see, this type of data distribution indicates that there are different means and different standard deviation for our analysis, simply means that you data that was analyzed have got two different groups with a fair gap between them to make that sudden decrease and return to the increasing in values of data. Also it indicates that is this data have got a sinusoidal wave, to check this information we can look on the scatter plot:



We can see that it forms somehow a sinusoidal wave, it could be visually clearer if the sample was more broad.

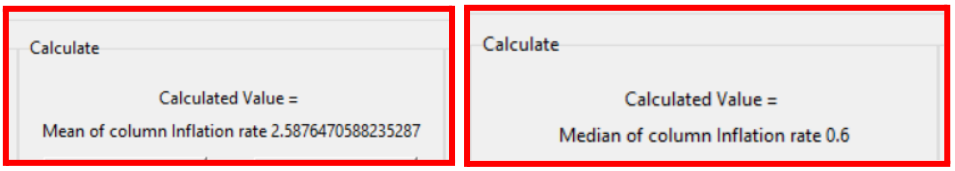
Discussion Topic 2: Skewness

The box plot and the normal distribution for the “Inflation Rate” is a good example for what we are going to discuss.

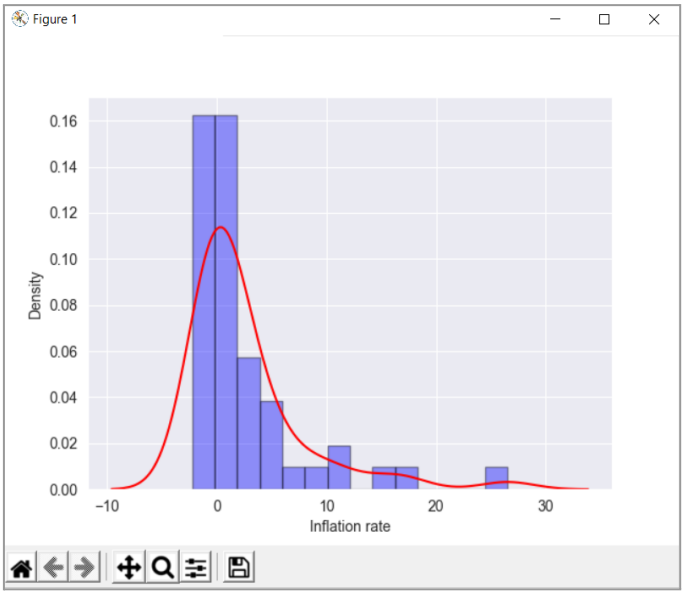


The box plot shows that the data values of this column are right/ positively skewed which means that in comparison with rest of the data, this column’s data lower bounds are extremely weak. Also that provide us with another interpretation that is the mean, median and mode values are

distinct with the condition that the mean > median > mode, demonstrated down below:

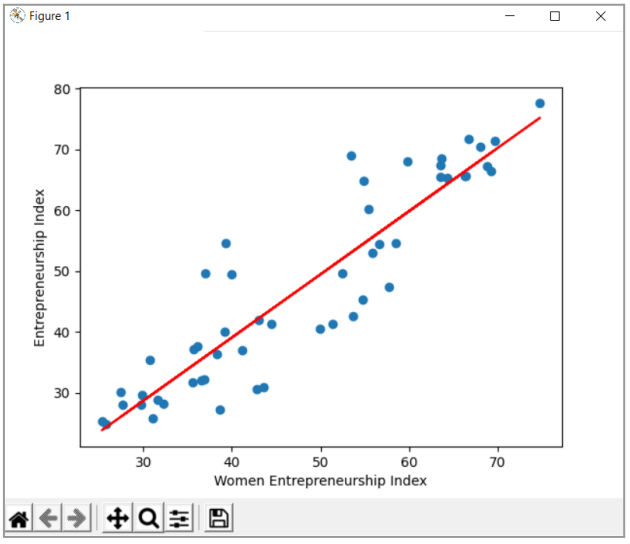


Also if we take a look at the normal distribution of this column:

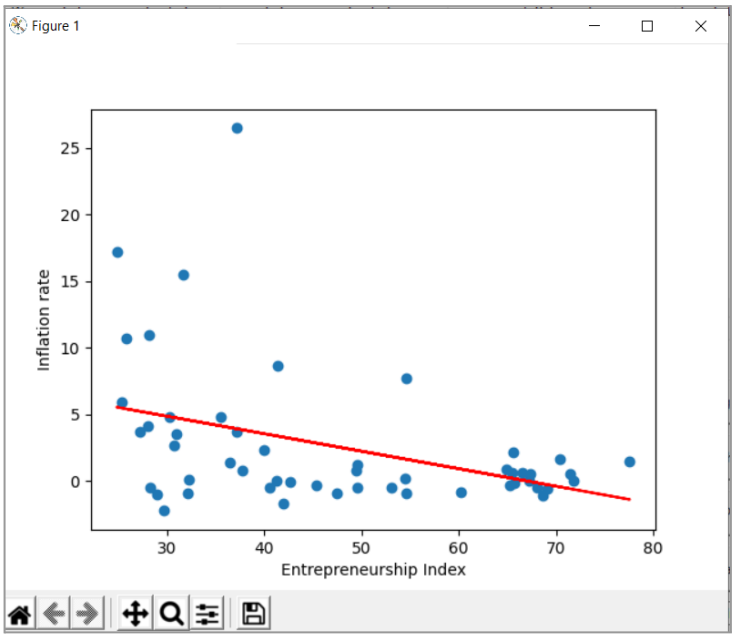


Discussion Topic 3: Scatter Plot Relationships

The scatter plot which shows the relationship between "Women Entrepreneurship Index" and "Entrepreneurship Index", we can deduce from it that they have a moderate relationship/ positive correlation.



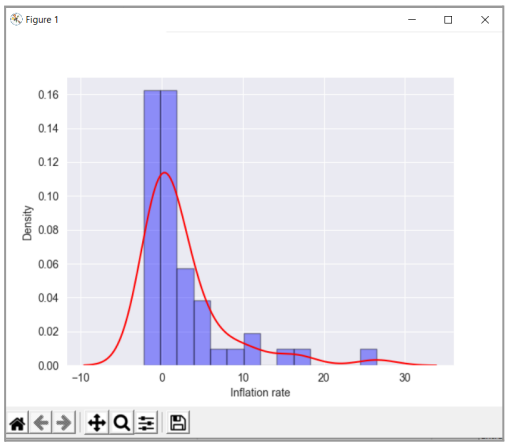
Another scatter plot between the “Entrepreneurship Index” and “inflation rate”, it indicates that there is a relationship between these two columns of type moderate/ negative correlation.

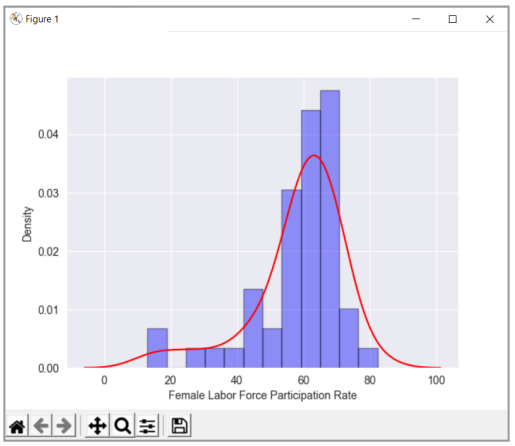


Hypothesis: Inflation rate and Female Labor Force Participation rate

We made a hypothesis that in countries where the inflation rate is high, the female labor force participation rate is low, and we will try to see if this assumption will be correct or not through the graphs and the data we get from our program.

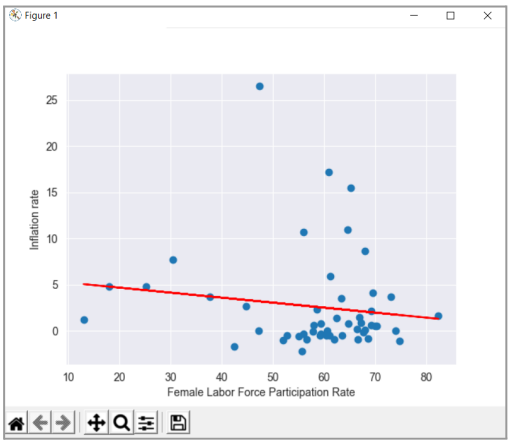
- First: the normal distribution

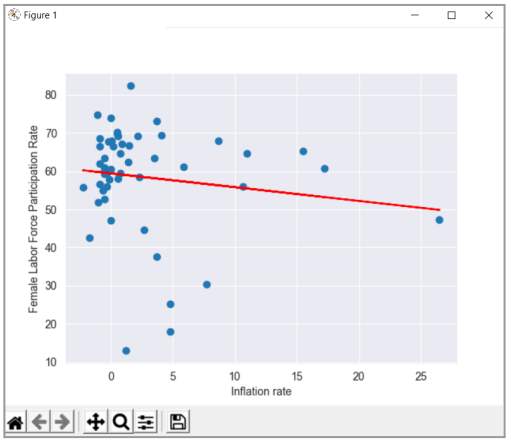




We can visually see that the higher values in the “Inflation Rate” reflect the lower values or nearly no values at all of the “Female labor force participation rate”, this takes us a step higher in proving our hypothesis could be right.

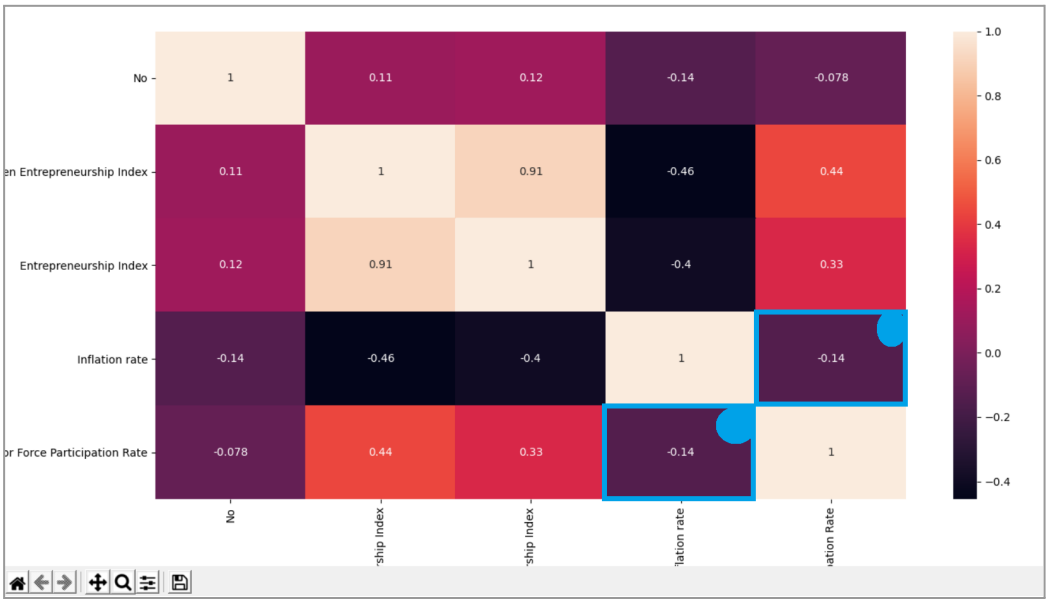
- Second: Scatter Plot





These two scatter plots represent their relationship wherein both plots we can indicate that there is nearly no association between these two attributes and if there was a relationship it would be a weak one, negative correlation. Indicating another information which is that these two values are inversely proportional which signifies that when one of them increases, the other would decrease.

- Third: heat map



The Person “r” value for their relationship is equal to -0.14 and by looking at the color gradient that indicates the strength of the correlation, we will find it near the zero (no linear relationship) that concludes the negative weak correlation. That gives us our last step into proving our hypothesis is correct.

4. Conclusion

Our project's aim is to analyse the dataset using certain graphical and numerical procedures given in the form of functions written in python programming language. The user can access and display the desired analysis of the raw data by choosing any button and it shows the information they want in text fields whether in the form of graphs (heatmaps, histogram, box-plots) or numerical values (mode, mean, IQR etc). The dataset could be of the **user’s choice** or fixed (**Women Entrepreneurship & Labor Force)** and we’ve successfully provided both options depending on what the user chooses. The dataset is a visual representation on how women are currently playing a major role in the labor force in the developed and developing countries given. The dataset works on a sample for each country to measure the women labor rate and study the conditions that could prevent women in developing countries to be the majority of their country’s labor force. The result of our dataset analysis upon exploring the dataset has revealed that actually the developing countries with high inflation rate could affect the female labor force participation negatively.

Our answer or recommendation for future research on how we could decrease the percentage of unemployment would definitely be female entrepreneurship and giving women more chances in different fields. Female labor represents nearly the same driving force as male employees, holding onto this mindset women in developing countries could release their creativity and a lot of countries would be able to touch their impact on our economy.

5. References

Dataset Reference:

Kübra Ersoy. (2020, November). *Women Entrepreneurship and Labor Force, Version 1*. Retrieved January 09, 2021 from: https://www.kaggle.com/babyoda/women-entrepreneurship-and-labor-force

Other References:

I. edureka!.(2020, April 29). *Exploratory Data Analysis (EDA) Using Python|Python Data Analysis|Python Training*[Video]. Youtube. https://youtu.be/-o3AxdVcUtQ

II. Stack Overflow.(2019). *I want to Print my output on pop up window? how should I approach?*. Retrieved from Stack Overflow website: https://stackoverflow.com/questions/59433386/i-want-to-print-my-output-on-pop-up-window-how-should-i-approach

III. The Pandas Development Team.[no date]. *pandas.DataFrame.describe*. Retrieved from pandas website: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html

IV. Martin Campbell. (2016). *How can I globally change a variable from a function in Python?*. Retrieved from Quora website: https://www.quora.com/How-can-I-globally-change-a-variable-from-a-function-in-Python

V. Grepper. (2020). *"python change variable outside function"*. Retrieved from Grepper website:

https://www.codegrepper.com/code-examples/dart/python+change+variable+outside+function

VI. Python Examples. [no date]. *How to change Button Text Dynamically in Tkinter?* .Retrieved from Python Examples website: https://pythonexamples.org/python-tkinter-change-button-text-dynamically/

VII. "w3cschools.com". [no date]. *Machine Learning - Mean Median Mode*. Retrieved from w3schools.com website: https://www.w3schools.com/python/python_ml_mean_median_mode.asp

VII. "w3cschools.com". [no date]. *Machine Learning - Standard Deviation*. Retrieved from w3schools.com website: https://www.w3schools.com/python/python_ml_standard_deviation.asp

IX. Stack Overflow.(2020). *How do I create a pie chart using categorical data in matplotlib?*. Retrieved from Stack Overflow website: https://stackoverflow.com/questions/63687789/how-do-i-create-a-pie-chart-using-categorical-data-in-matplotlib

X. Ramon Williams.(2020, July 26). *How to view Excel File or Pandas DataFrame in Tkinter (Python GUI)* [Video]. Youtube. https://youtu.be/PgLjwl6Br0k

XI. The Teclado Blog. (2020, August 22). Tkinter's Grid Geometry Manager. Retrieved from Teclado website: https://blog.tecladocode.com/tkinters-grid-geometry-manager/

XII. RIP Tutorial. [no date]. tkinter place() .Retrieved from RIP Tutorial: https://riptutorial.com/tkinter/example/29714/place--

XIII. codebasics. (2018, July 1). *Machine Learning Tutorial Python - 2: Linear Regression Single Variable* [Video]. Youtube. https://youtu.be/8jazNUpO3lQ