*Lecturer: Andrew Beatty*

Student: Luminita Done

*Student ID: G00411226*

Pands PROJECT

**Fisher’s Iris data set**

Contents

[About the Iris dataset 2](#_Toc102943967)

[How the program works 2](#_Toc102943968)

[What’s the result of each input? 3](#_Toc102943969)

[Option 1: Save the summary data text. 3](#_Toc102943970)

[How this outcome was possible? 4](#_Toc102943971)

[Option 2: Save histograms to png files 4](#_Toc102943972)

[Option 3: Scatter plot 6](#_Toc102943973)

[Option 4: 7](#_Toc102943974)

[Option 0: 8](#_Toc102943975)

[What libraries were imported: 9](#_Toc102943976)

[Pandas 9](#_Toc102943977)

[Matplotlib 9](#_Toc102943978)

[Seaborn 10](#_Toc102943979)

[Documents and images saved by the program 10](#_Toc102943980)

[References 11](#_Toc102943981)

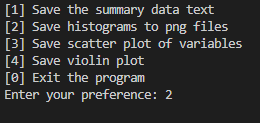
# About the Iris dataset

R.A. Fisher published the first report on the iris data set in the Annals of Human Genetics in 1936. It is a collection of 50 samples collected by the author on each of three Irises species: setosa, versicolor, and virginica. Four attributes of 50 flowers from each of the plants were measured: sepal length, sepal width, petal length, and petal width. According to the author, the lengths and widths of the petal and sepal are features that can be utilized to determine which species they belong to using a linear discriminant model. The linear discriminant model, a statistical, machine learning, and pattern recognition technique used to distinguish between two or more objects, classes, or events, was created by Fischer himself. (Wikipedia, n.d.)

Fischer recorded the findings for the three species in a table containing each of the four measurements, followed by tables of observed means, sums of squares, and other statistics to show how each species could be distinguished from the others. Fischer uses sums of squares and products of deviations from each mean to produce a linear function that best distinguishes between the two species. The ANOVA test is a strong statistical method for determining correlations (such as differences) between variables by comparing the means of the variables. (statisticssolutions.com, n.d.)

# How the program works

The program has a built menu which requires input from a user.



**Option 1**: When the user chooses option one, the program will save the summary data to a text file;

**Option 2**: When the user chooses option two, the program will display and save the histograms to png files;

**Option 3**: When the user chooses option three, the program will display and save a scatter plot to png file;

**Option 4**: When the user chooses option four, the program will display and save the violin plot;

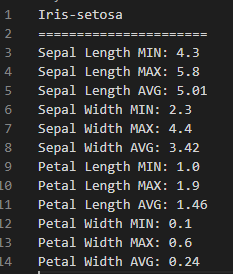
**Option 0:** When the user chooses option zero, the user will exit the program.

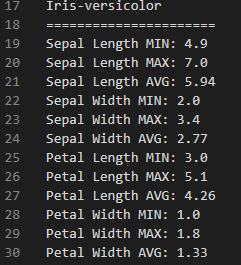
## What’s the result of each input?

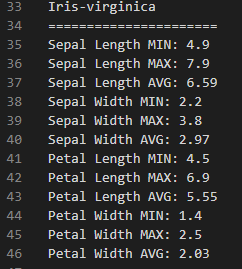
## Option 1: Save the summary data text.

Below we can observe the summary data set of Iris Setosa, Iris Versicolor and Iris Virginica.

All Iris types sepal length min, maximum and average are shown in the summaries below. (adamsmith.haus, 2022)

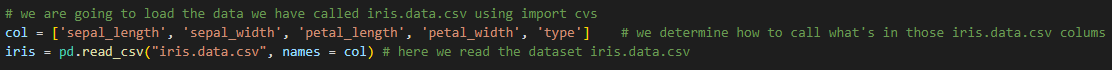






### How this outcome was possible?

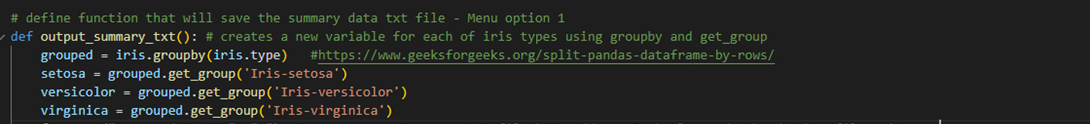
First, we are loading the data from the csv file (analyticsvidhya.com, 2022)



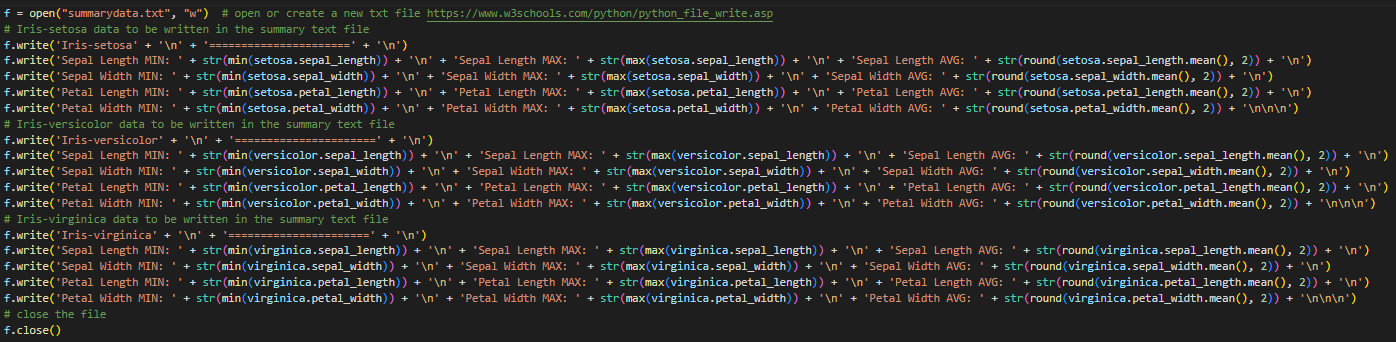
We defined the function that will save the summary data text file, a function named **def output\_histogram()**

This function is divided into two parts:

* The first part is grouping the data from iris variables by type, the last column of the csv file. (geeksforgeeks.org, 2022)



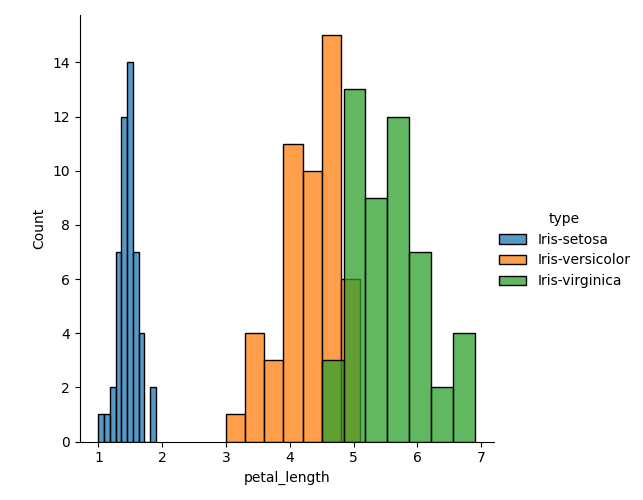
* In the second part, the program opens or creates a new file(**summarydata.txt**) if the file does not exist. After that, the program will write in the new file the values of each iris type: sepal; petal length max, width max, and average. (w3schools.com, 2022)



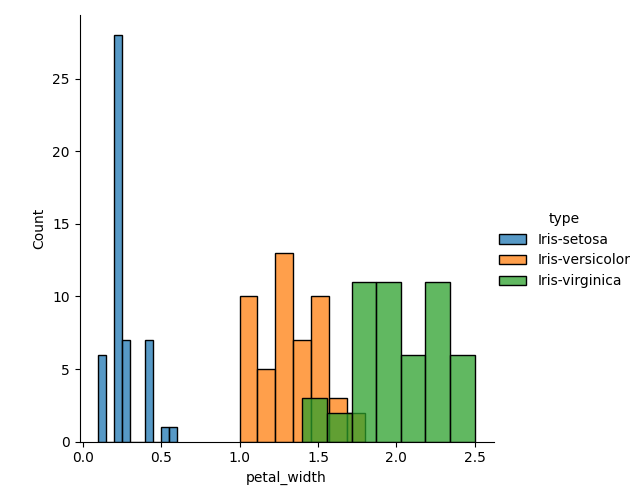
## Option 2: Save histograms to png files

When the user’s input is ‘2’, the following histograms will be displayed and also saved on our computer. (realpython.com, 2022)

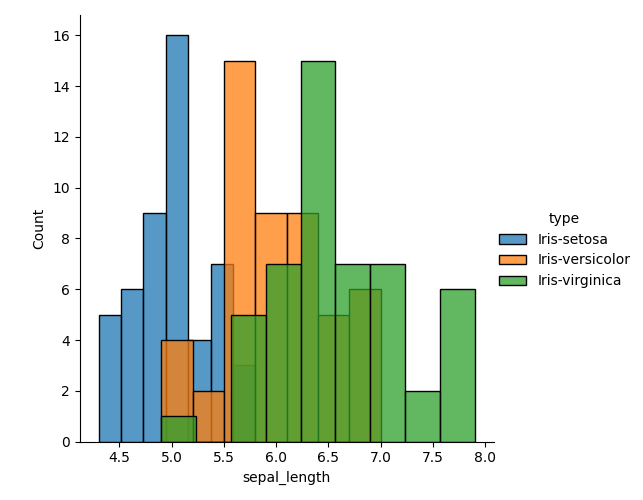
Petal length histogram:



Petal width histogram:



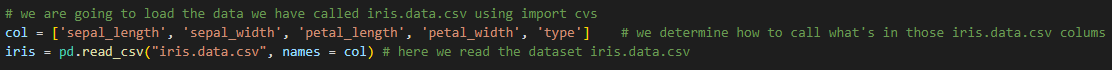
Sepal length:



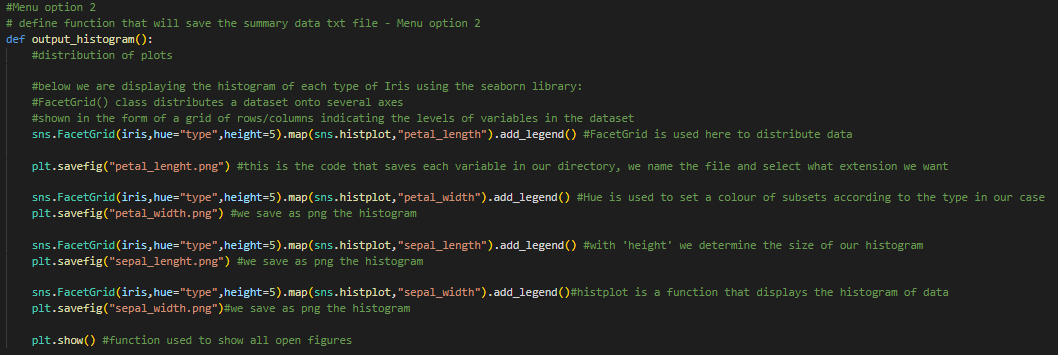
#### How this outcome was possible?

When the user’s input is ‘2’, the function **output\_histogram()** will be executed.

First, we are loading the data from the csv file (analyticsvidhya.com, 2022)

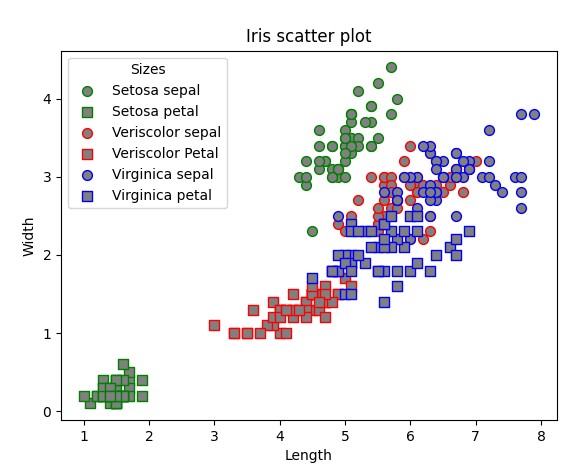


When the function **output\_histogram()**  is executed will read from the values from iris variables and will display and save the histograms to a png file using the seaborn library. (the output of the files displayed above: petal\_lenght.png; petal\_width.png, sepal\_lenght.png, sepal\_width.png). **sns.FacetGrid** is used here to distribute data



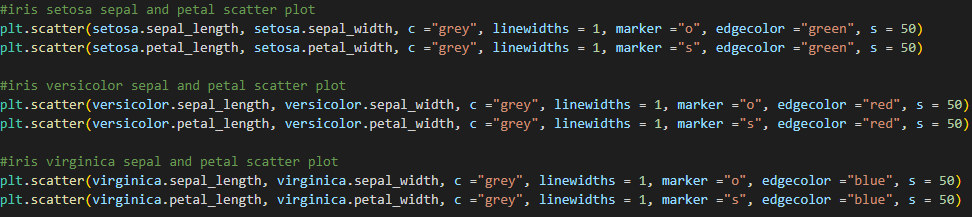
## Option 3: Scatter plot

When the user’s input is ‘3’, the Iris Scatter plot representing the Iris Setosa, Veriscolor and Virginica width and length of petals and sepals is displayed and saved automatically into our computer:



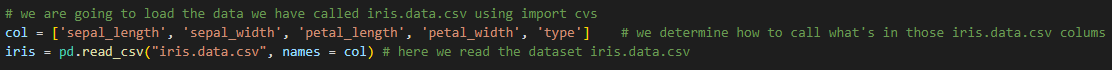
The contour used is to differentiate the types varieties and differences of the Iris. (reneshbedre.com, 2022)

Can be seen in the code below:



#### How this outcome was possible?

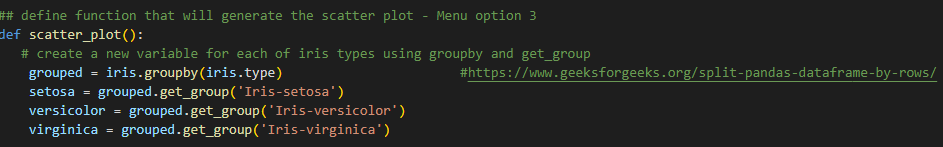
First, the program is loading the data from the csv file (analyticsvidhya.com, 2022)



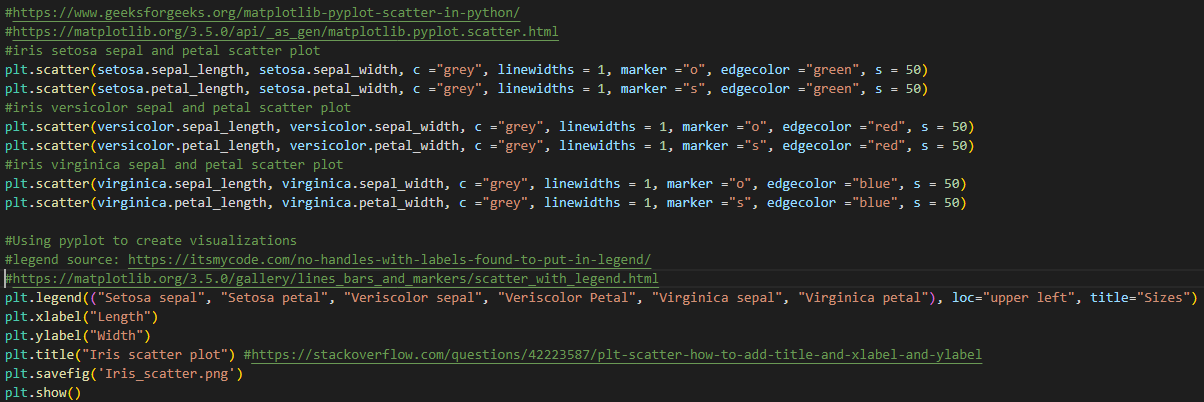
When the user’s input is ‘3’, the function **scatter\_plot()** will be executed.

The function **scatter\_plot()**  is divided into two parts:

- The first part is grouping the data from iris variables by type, the last column of the csv file. (geeksforgeeks.org, 2022)

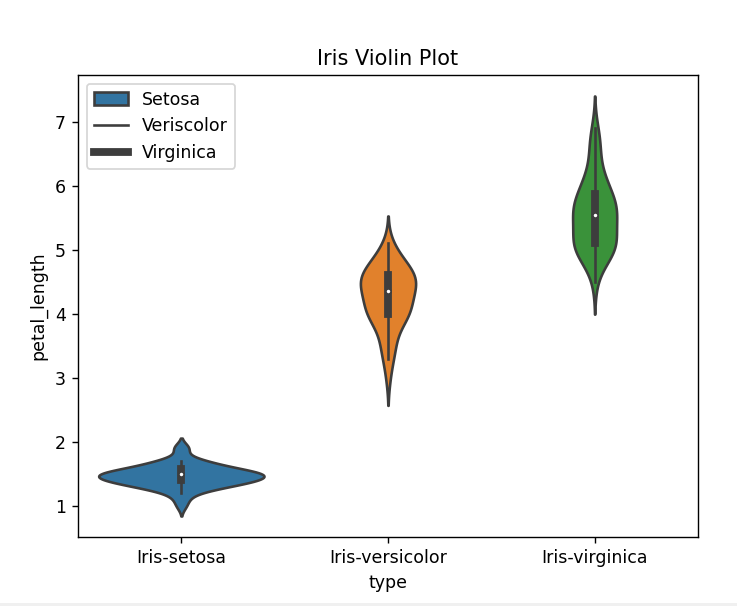


- In the second part the program will create, display and save the scatter plot to a png file(Iris\_scatter.png) using the matplotlib library.



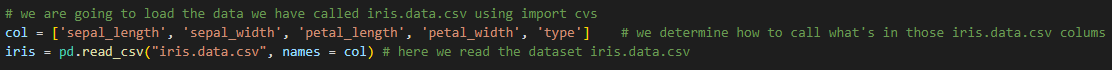
## Option 4:

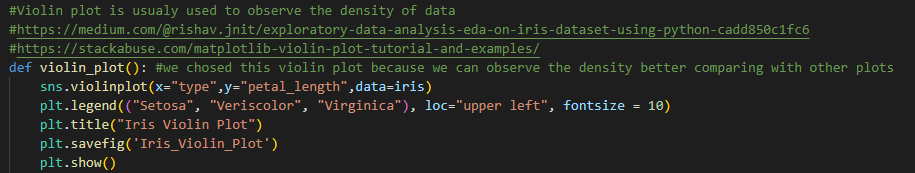
When the user’s input is ‘4’, the Iris Violin plot representing the Iris Setosa, Veriscolor and Virginica is displayed and saved automatically into our computer (geeksforgeeks.org, 2022):



#### How this outcome was possible?

First, the program is loading the data from the csv file (analyticsvidhya.com, 2022)



When the user’s input is ‘4’, the function **violin\_plot()**will be executed. r’s

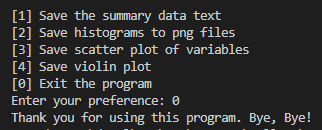
The function **violin\_plot()**  is divided in two parts:

- The first part is grouping the data from the iris variables by type, and petal length .

- In the second part the program will create, display and save the violin plot to a png file(Iris\_Violin\_Plot.png) using the matplotlib library.

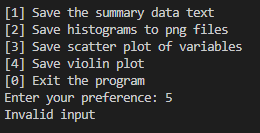
# Option 0:

When the user’s input is ‘0’, the message ‘Thank you for using this program. Bye, bye!” will be displayed.

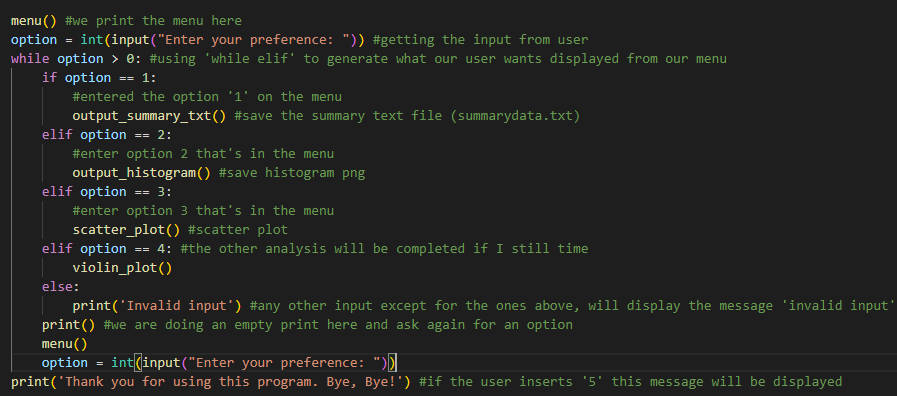


##### Invalid input 5:

Any input other than 1-4, will give the result: ‘Invalid input’



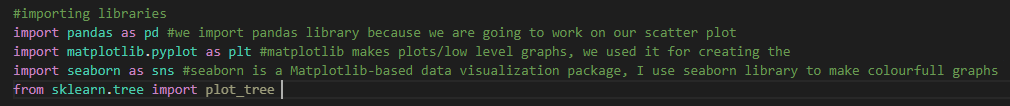
Based on the user’s input option, the program will execute one of the functions:



# What libraries were imported:

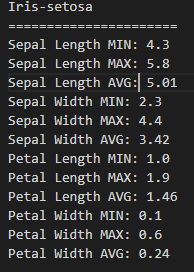
Before we start working on our code, we must draw a plan and research what types of libraries are required to achieve our goal: create and save histograms, scatter plots and summary of variables. (towardsdatascience.com, 2022)

Below are all the libraries I’ve imported. (realpython.com, 2022)



## Pandas

Pandas library was imported because we used it while extracting the min, max and average of sepals and petals.



## Matplotlib

Matplotlib makes plots/low-level graphs, like the histograms we created and the scatter plot.

(matplotlib.org, 2022)

mport matplotlib.pyplot as plt

## Seaborn

Based on matplotlib, Seaborn is a Python data visualization package. It provides a high-level interface for creating visually appealing and useful statistics visuals. (machinelearningplus.com, 2022)

You can read the introductory notes or the paper to get a quick overview of the library's concepts. To learn how to install the package and get started with it, go to the installation page. You can look through the example gallery to see what you can achieve with seaborn, and then read the tutorial or API reference to learn how. (seaborn.pydata, n.d.)

Seaborn is a Matplotlib-based Python data visualization package. The following syntax is the most frequent way to import Seaborn into your Python environment:

import seaborn as sns

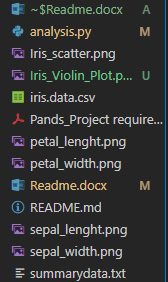
Python is told to introduce the Seaborn library into your present environment via the import seaborn portion of the code.

The as sns section of the code instructs Python to assign Seaborn the alias sns. You can now use Seaborn functions by using sns.function name instead of seaborn.function name.

After Seaborn is imported, we may utilize the built-in methods to visualize data rapidly. (statology.org, n.d.)

### Documents and images saved by the program

The following documents were saved automatically by the program, with just input from a user(0-4).



# References

adamsmith.haus, 2022. *adamsmith.haus.* [Online]   
Available at: https://www.adamsmith.haus/python/answers/how-to-write-a-variable-to-a-file-in-python  
[Accessed 05 05 2022].

analyticsvidhya.com, 2022. *analyticsvidhya.com.* [Online]   
Available at: https://www.analyticsvidhya.com/blog/2021/08/python-tutorial-working-with-csv-file-for-data-science/  
[Accessed 05 05 2022].

geeksforgeeks.org, 2022. *geeksforgeeks.org.* [Online]   
Available at: https://www.geeksforgeeks.org/split-pandas-dataframe-by-rows/  
[Accessed 06 05 2022].

machinelearningplus.com, 2022. *machinelearningplus.com.* [Online]   
Available at: https://www.machinelearningplus.com/plots/top-50-matplotlib-visualizations-the-master-plots-python/  
[Accessed 05 05 2022].

matplotlib.org, 2022. *matplotlib.org.* [Online]   
Available at: https://matplotlib.org/3.5.0/api/\_as\_gen/matplotlib.pyplot.show.html  
[Accessed 06 05 2022].

medium.com, n.d. *medium.com.* [Online]   
Available at: https://medium.com/fintechexplained/why-should-we-use-numpy-c14a4fb03ee9  
[Accessed 19 04 2022].

realpython.com, 2022. *realpython.com.* [Online]   
Available at: https://realpython.com/python-statistics/#choosing-python-statistics-libraries  
[Accessed 06 05 2022].

realpython.com, 2022. *realpython.com.* [Online]   
Available at: https://realpython.com/python-statistics/#histograms  
[Accessed 04 05 2022].

seaborn.pydata, n.d. *seaborn.pydata.* [Online]   
Available at: https://seaborn.pydata.org/  
[Accessed 18 04 2022].

statisticssolutions.com, n.d. *statisticssolutions.com.* [Online]   
Available at: https://www.statisticssolutions.com/anova-analysis-of-variance/  
[Accessed 18 04 2022].

statology.org, n.d. *statology.* [Online]   
Available at: https://www.statology.org/import-seaborn-as-sns/  
[Accessed 19 04 2022].

towardsdatascience.com, 2022. *towardsdatascience.com.* [Online]   
Available at: https://towardsdatascience.com/top-10-python-libraries-for-data-science-cd82294ec266  
[Accessed 06 05 2022].

tutorialspoint, n.d. *tutorialspoint.* [Online]   
Available at: https://www.tutorialspoint.com/seaborn/seaborn\_facet\_grid.htm  
[Accessed 18 04 2022].

w3schools.com, 2022. *www.w3schools.com.* [Online]   
Available at: https://www.w3schools.com/python/python\_file\_write.asp  
[Accessed 06 05 2022].

Wikipedia, n.d. *Wikipedia.* [Online]   
Available at: https://en.wikipedia.org/wiki/Linear\_discriminant\_analysis  
[Accessed 18 04 2022].