Assignment#5 Data Reproducibility
Points: 10%

Due Date: 07 Dec 2020

Problem Statement:

Students are provided with four datasets. The four dataset has to be iterated over using glob(). For each of the dataset three visualization and a corresponding message has to be printed as demonstrated under Instructions.

This Assignment evaluates students' knowledge on Python Functions, Transformations using Python , git and github.

Instructions:

Step 1: Use the following code to get started.

```
import glob
import numpy
import matplotlib
import matplotlib.pyplot

dataset = sorted(glob.glob('data*.csv'))

for datasets in dataset[:4]:
    print(datasets)
    visualize(datasets)
    identify_issues(datasets)
```

Step 2: From the above code we can infer that we need to create two functions.

visualize( ) and

identify_issues( )

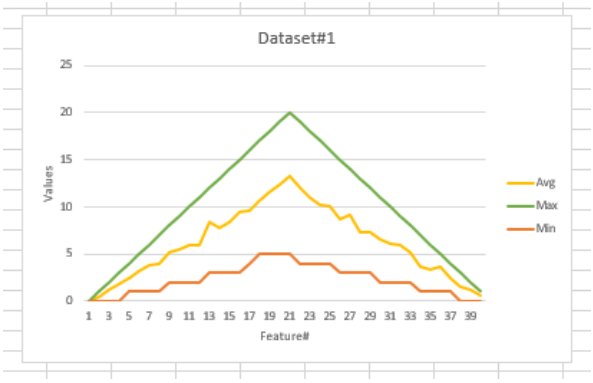
Step 3: visualize( )

visualize( ) must produce a sub plot of average of each features, max of each features and min of each features.

The following is the visualization you get for the dataset data-01

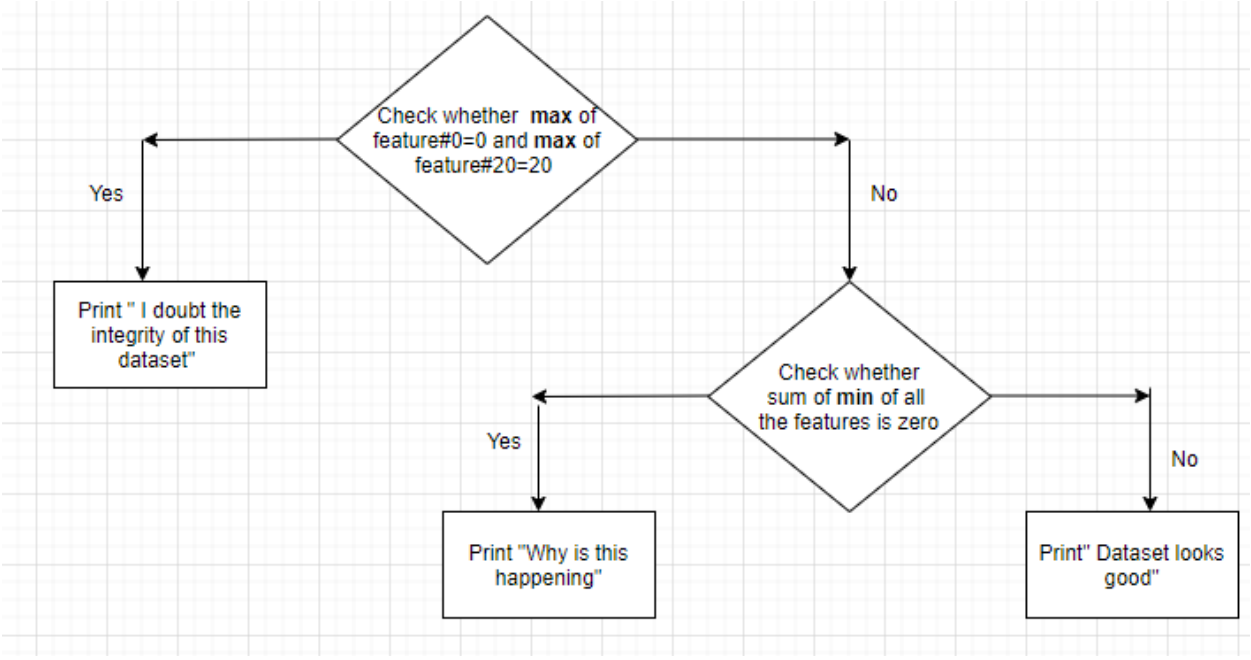
OR

A single plot with the following - average of each features, max of each features and min of each features.



Step 4: identify_issues( )

This function should work as shown in below flowchart



Step 5: Create a github account

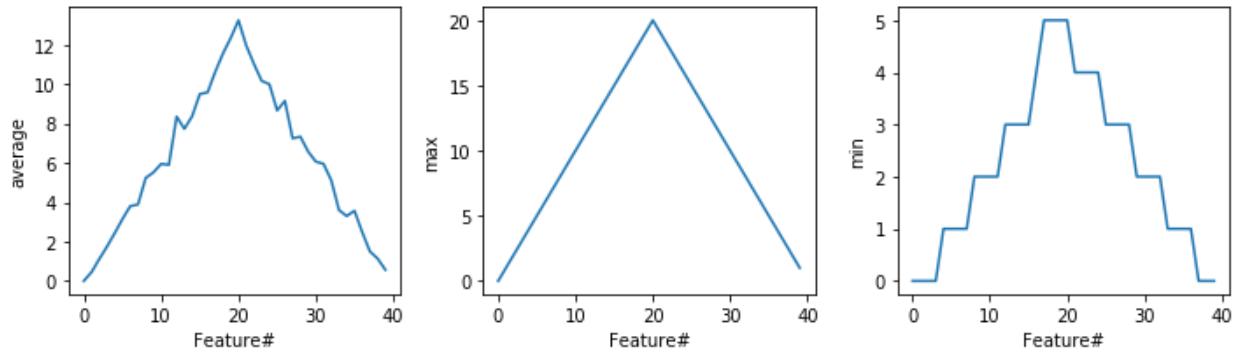
Step 6: Create a new repository in your github called Assignment-5

Step 7: Push the jupyter notebook along with the four dataset into the newly created Assignment-5 repo.

Expected Result

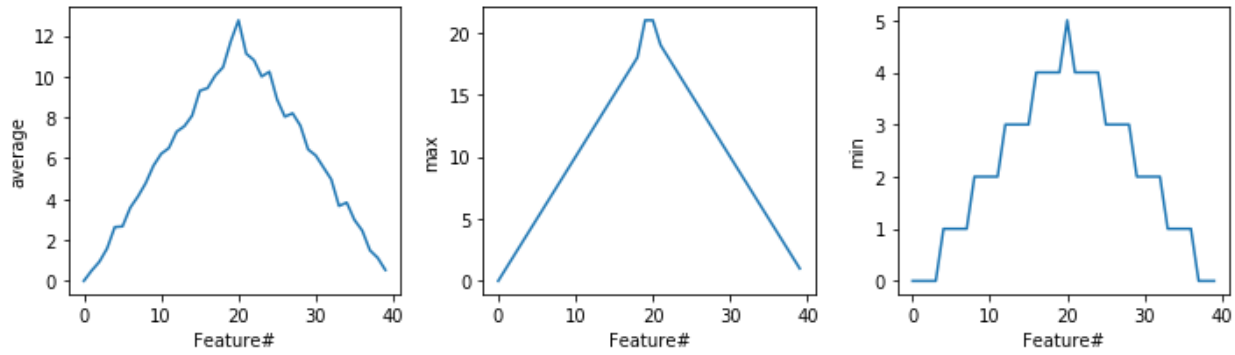
If your code run without error and you have managed to fully complete the assignment this is how the result should look like.

`data-01.csv`



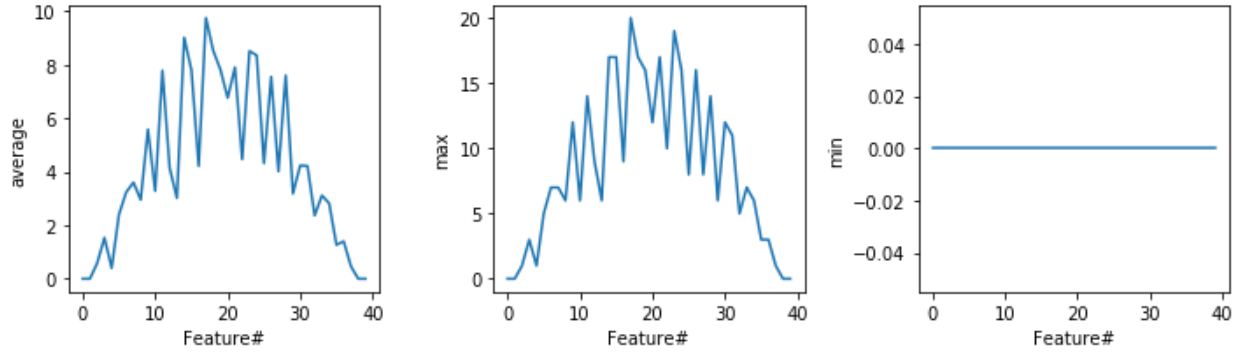
`I doubt the integrity of this dataset`

`data-02.csv`



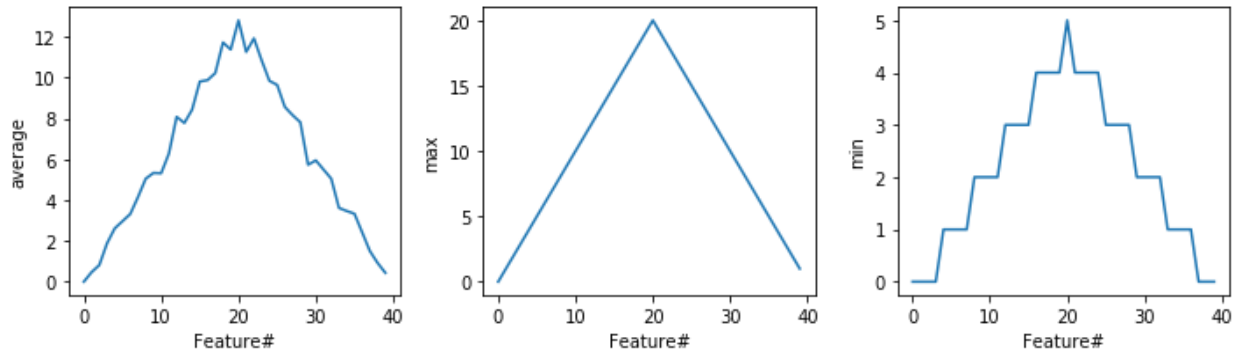
`Dataset looks good`

data-03.csv



Why is this happening?

data-04.csv



I doubt the integrity of this dataset

Submission Format:

In the DC Connect under Assignment#5, post the link to the github repo hosting your assignment folder.

Academic Integrity and Late submission:

Late assessments will be subject to a 20% per calendar day late penalty unless otherwise stated by the professor. Students should communicate with the professor in advance of a due date for any requests for an extension as a result of exceptional circumstances.

Any violation of academic integrity will not be accepted and will be given a grade of zero (0). Please watch this video on academic integrity.
https://www.youtube.com/watch?v=BnEw72e_YYo&feature=youtu.be

Rubrics:

1. A working code (ie code without error) that gives the desired output.
2. Code should be commented properly.
3. Appropriate headings should be given to each cell.
4. Student should develop an Optimized Code.
5. Demonstration on how well the concept of data reproducibility is used while designing functions.
6. Do the functions have docstring?
7. Student have successfully uploaded all four datasets and the jupyter notebook file into their github and shared the link to their repo.
8. Has student completed the assignment with minimum or no help from the instructor in correcting errors?