

Eric De Leenheer 100869527

Here is the link to my Google Drive containing task 1:

https://colab.research.google.com/drive/1RCE_vwXoXcv623ieEDev_bcQDJuWAauY?usp=sharing

and here is the GitHub link:

<https://github.com/Edspeedy1/SOFE-Quality-Lab-5>

Also, this line of code had to be included at the top of the file otherwise the code would run into errors importing pandas:

```
!pip install --no-cache-dir --force-reinstall numpy pandas
```

For task 2, I suspect that the single mislabeled data point is due to the fact that that point can fit into multiple categories (or labels) and thus got tagged multiple times, but it should only belong to one label.

























For task 3,

I suspect that the reason that the first two data points (the setosas) are “suspected anomalies” is due to the fact that they have a high sepal length with a low sepal width.

For the next two (the versicolors) I suspect that they were tagged because of their sepal width to petal width ratio

For the last two, (the virginica) it seems that all of their attributes are smaller than what I would expect

Here is a picture of my findings

Suspected Anomalous Data Points							
Index	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	True Label	Flower	Species
18	5.7 	3.8 	1.7 	0.3 	0	Setosa	Setosa
31	5.4 	3.4 	1.5 	0.4 	0	Setosa	Setosa
68	6.2 	2.2 	4.5 	1.5 	1	Versicolor	Versicolor
82	5.8 	2.7 	3.9 	1.2 	1	Versicolor	Versicolor
106	4.9 	2.5 	4.5 	1.7 	2	Virginica	Virginica
119	6.0 	2.2 	5.0 	1.5 	2	Virginica	Virginica

In almost all cases the petal length was shorter than expected

To see if these are truly anomalies, it helps to plot the original data on a graph (in this case many graphs) and try to see if there are correlations between any of the variables and what the outliers are. Then see if those outliers match up with the ones we just detected