Lab 4: Visualizing Number with Python
Due: next class meeting day

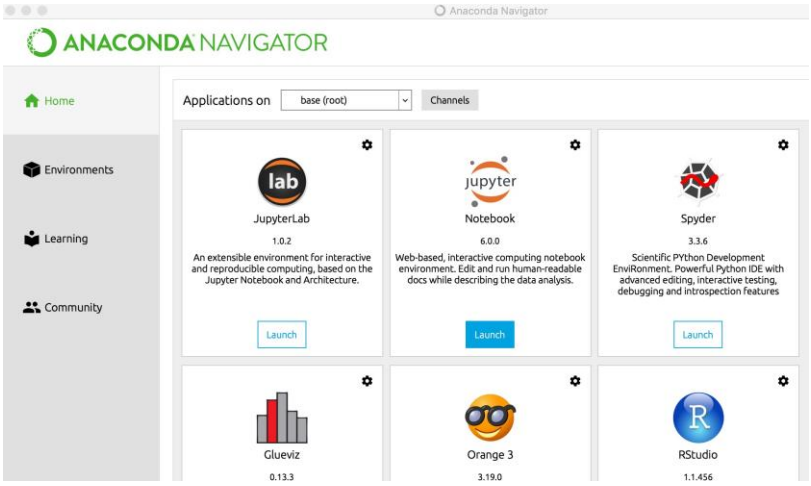
Description
This lab will allow you to practice using popular packages in Python for visualization. You will have 3 tasks: 1) installing Python, and Jupyter Notebook with Anaconda; 2) visualize the Iris dataset with Python; and 3) create 4 visualizations using Tableau

Deliverables
- The modified ipython notebook
- This document with corresponding sections filled.

If you execute pandas.tools.ploting() for one of the visualizations and it does not work, simply remove tools from the command

Task 1: Installing Python and Jupyter Notebook with Anaconda

Steps:
1. Install Anaconda (If you are using your own computer, or the lab machine happens to miss Anaconda)
    a. Install on a Windows machine: https://docs.anaconda.com/anaconda/install/windows/
    b. Install on a Mac machine: https://docs.anaconda.com/anaconda/install/mac-os/
2. Launch Anaconda
    a. Windows: go to ***Start*** menu, and find Anaconda from the application list
    b. Mac: go to ***Applications*** -> ***Anaconda***
3. Installing **Jupyter Notebook**: as shown below, Jupter Notebook will appear on the main screen of Anaconda. You need to click **Install** in order to install it on your computer **(**Mine has been installed, so it shows ***Launch***).
    a. Alternatively, you can also install Jupyter Notebook using pip:
        i. Mac: Type pip install notebook in Terminal window
        ii. Window: Type pip install notebook to Command Prompt window



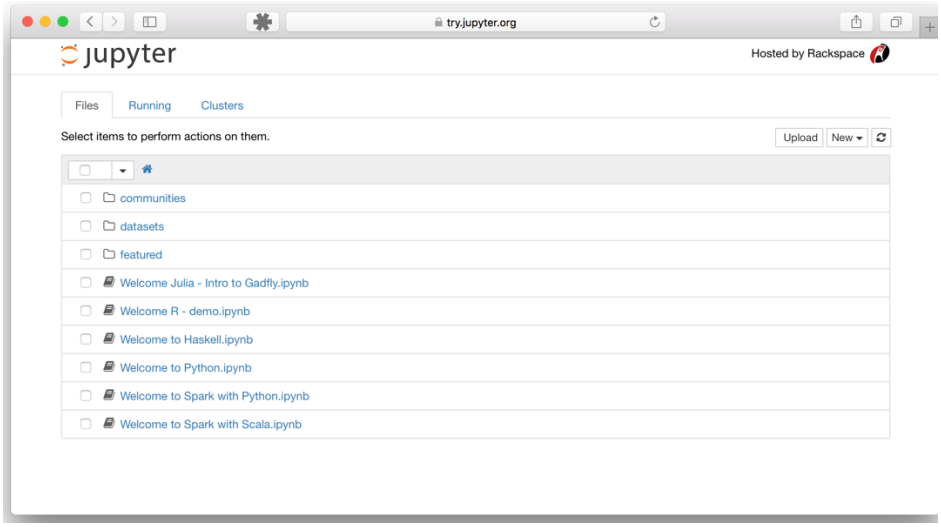
4. Once installed, you can open Jupyter Notebook by clicking ***Launch*** from the Anaconda main screen. This will open a new webpage in your browser. The new page is the python notebook working page.
    a. Alternatively, you can also use command to start Notebook:
        i. Windows: type juypter notebook in Command Prompt window
        ii. Mac: type juypter notebook in Terminal window

5. This will print some information about the notebook server in your terminal, including the URL of the web application (by default, http://localhost:8888):

```
$ jupyter notebook
[I 08:58:24.417 NotebookApp] Serving notebooks from local directory: /Users/catherine
[I 08:58:24.417 NotebookApp] 0 active kernels
[I 08:58:24.417 NotebookApp] The Jupyter Notebook is running at: http://localhost:8888/
[I 08:58:24.417 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to
```

It will then open your default web browser to this URL.

6. When the notebook opens in your browser, you will see the Notebook Dashboard, which will show a list of the notebooks, files, and subdirectories in the directory where the notebook server was started. Most of the time, you will wish to start a notebook server in the highest level directory containing notebooks. Often this will be your home directory.



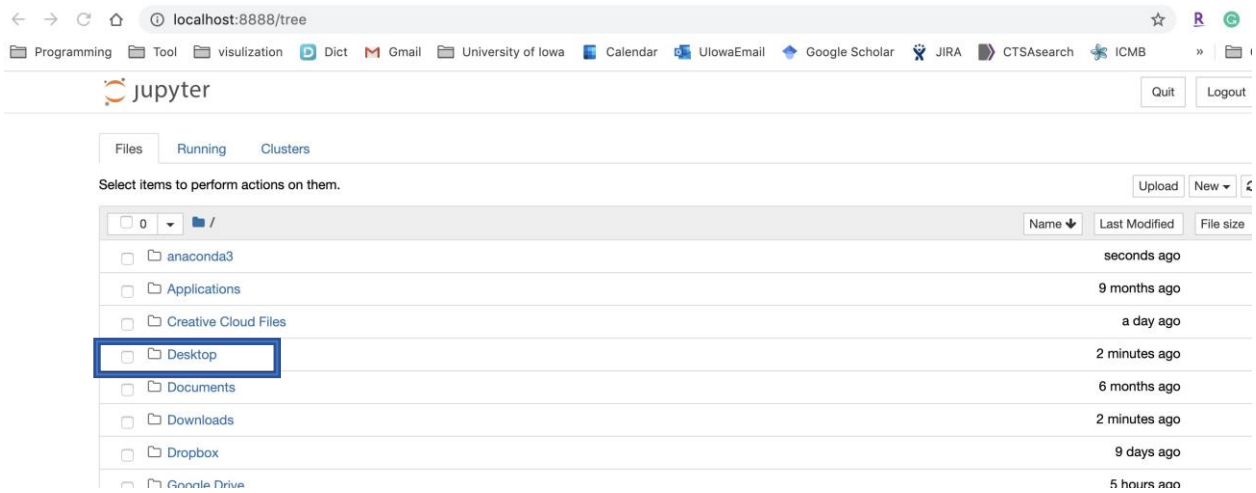
Task 2: Visualizing Iris data

You will be using Python to visualize the Iris dataset we mentioned in class.

Steps:

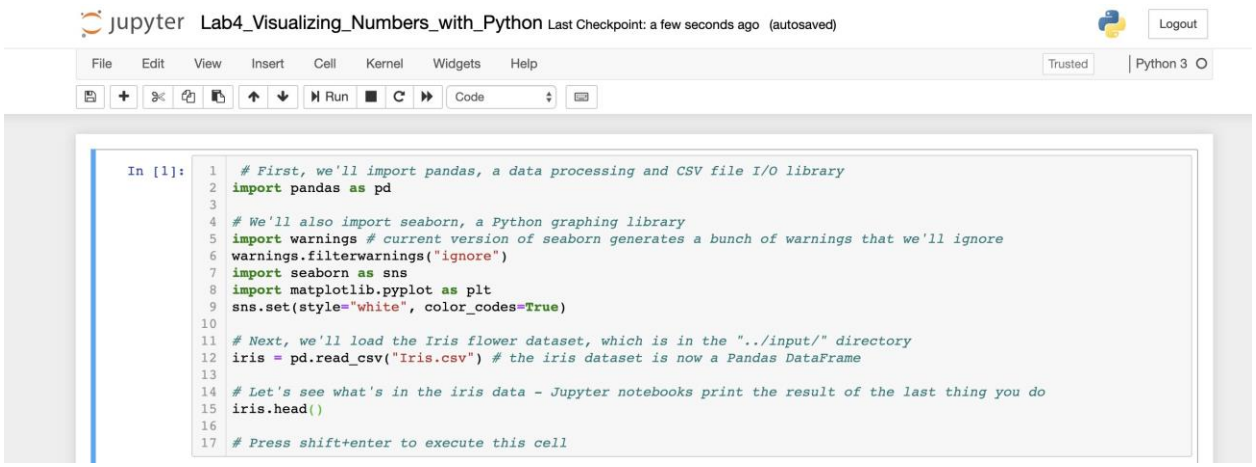
1. Please download Iris.csv and Lab4_Visualizing_Numbers_with_Python.ipynb file from ICON and put them to your Desktop (Make sure you unzip the files from Lab4.zip and put these two files directly on Desktop. The file directory is very important for the following steps.)

2. From the Jupyter notebook window you opened from the last step, please click Desktop to navigate to you Desktop. Then click the Lab4_Visualizing_Numbers with Python.ipynb file. See blow.



3. Once you switch to Desktop, please open the Lab4_Visualizing_Numbers_with_Python.ipynb file by clicking it. You will see all the codes and my comments to finish the task.

The ipython notebook for the lab opened:

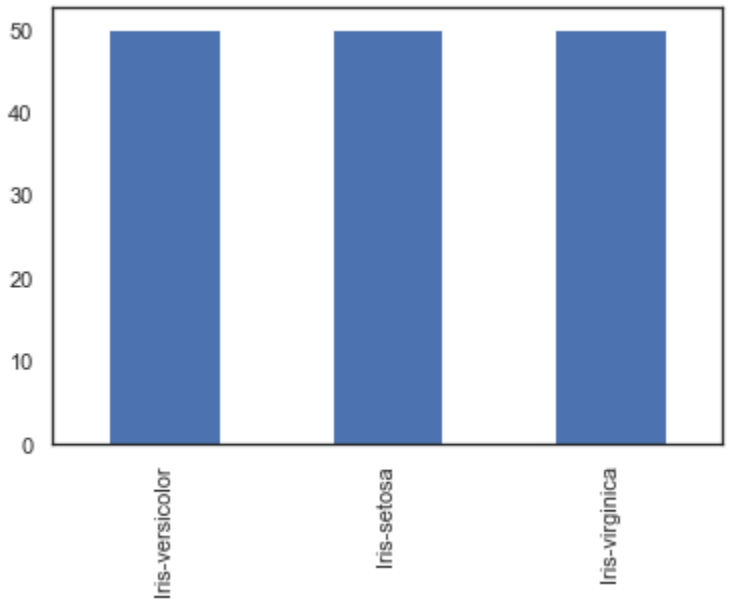


4. Please follow my instruction in the notebook to finish the assignment.

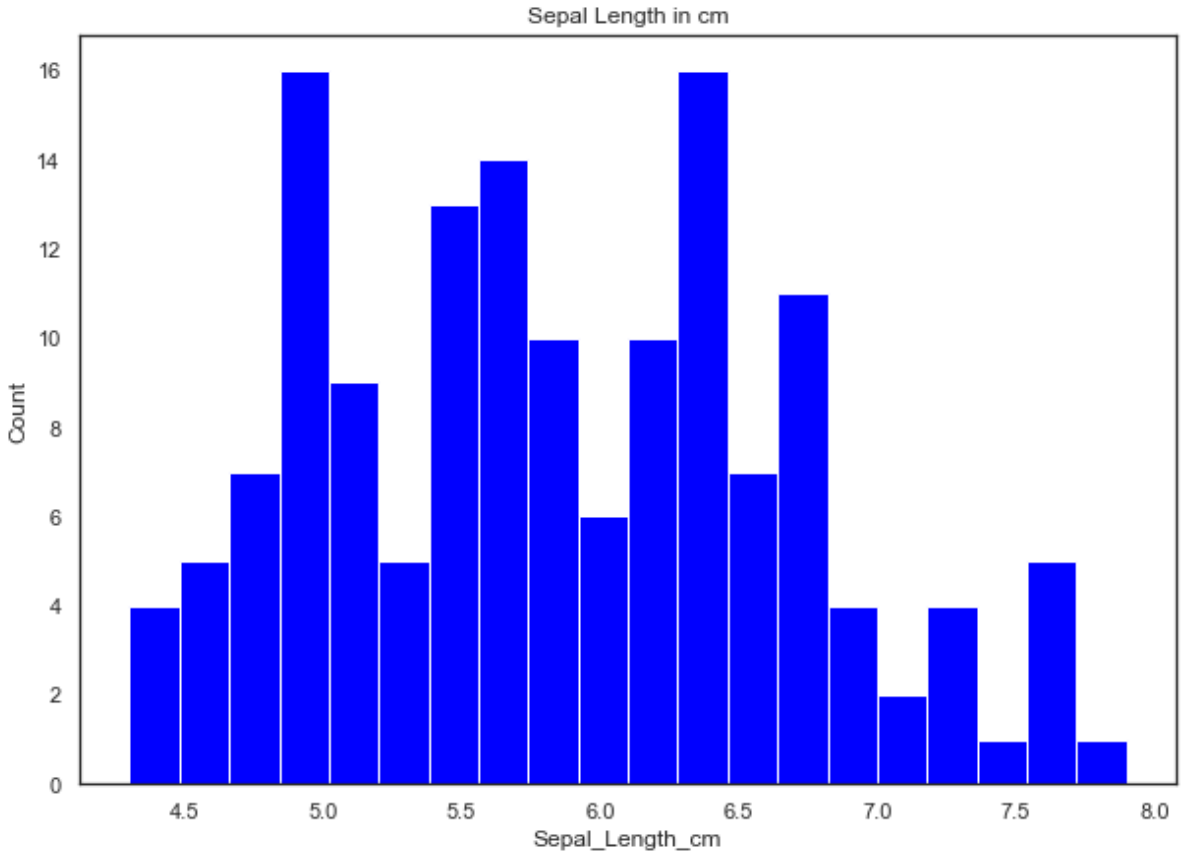
5. Among those visualizations provided to you, or created by you, which ones are appropriate? Which ones are not? Why?

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

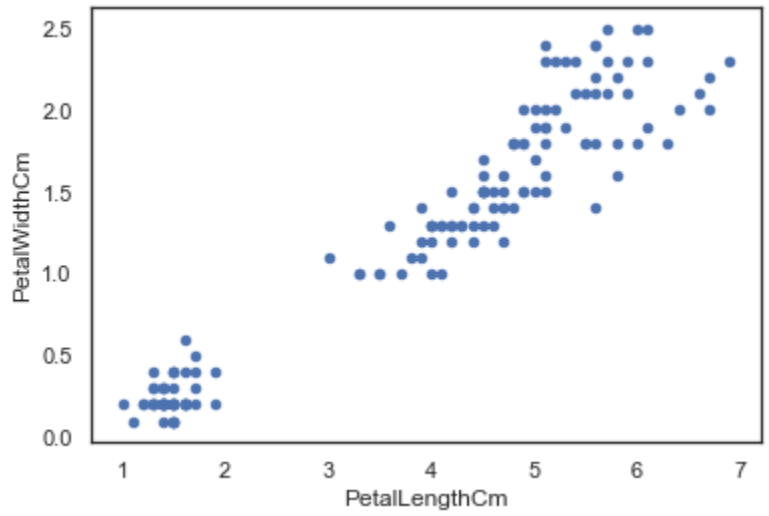
Good: Actual numbers are clearly displayed and easy to find.
Bad: No visualization to give a sense of correlation between variables.



Good: Simple and easy to read.
Bad: No need to make a bar chart when every value is the same. It's better to display the info that there were 50 samples for each species in a simple table.

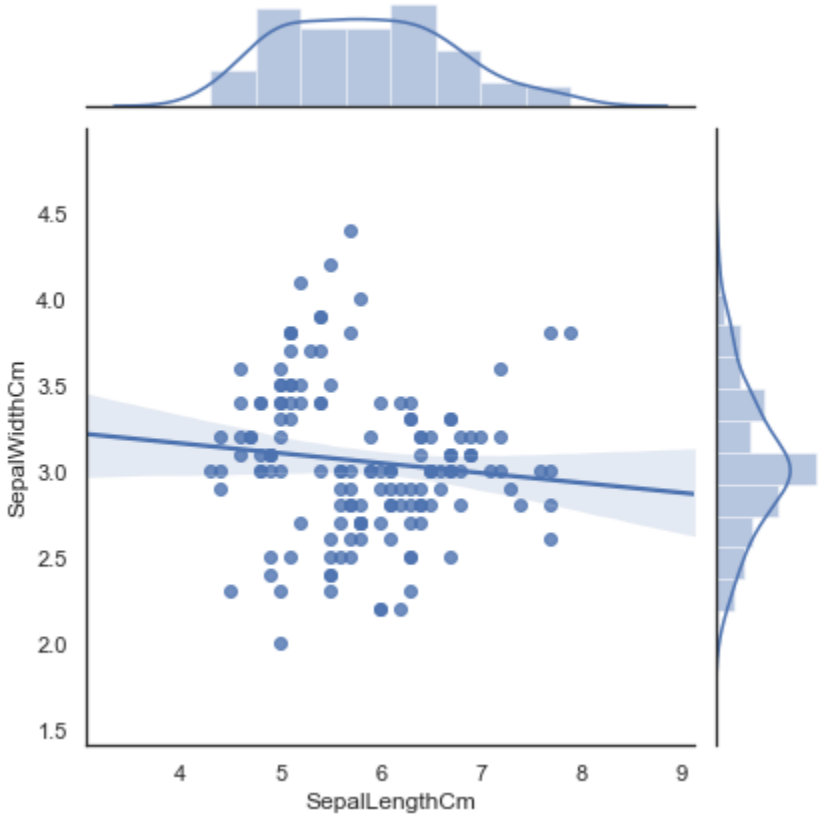


Good: Suggests that sepal length has a roughly normal distribution, while making it clear that there are outliers that are much longer or shorter than the average.
Bad: No differentiation by species. No information to account for the outliers around 5.0 cm and 6.25 cm.

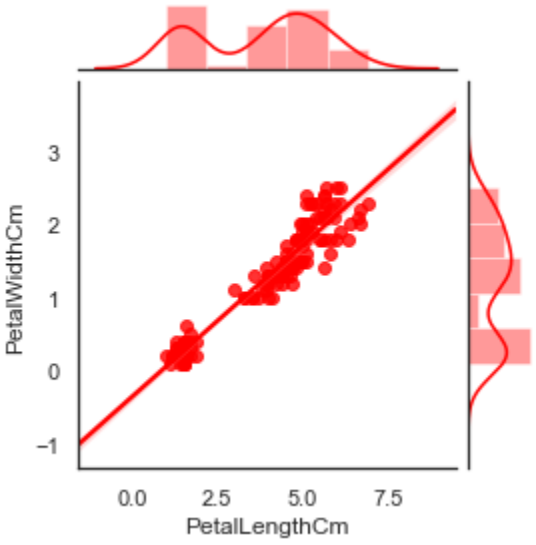


Good: Clearly shows a very strong positive correlation between petal length and petal width. Also shows that there are outliers that don't follow the trend quite as closely.

Bad: No differentiation by species. The gap between 2-3 petal length is unexplained.

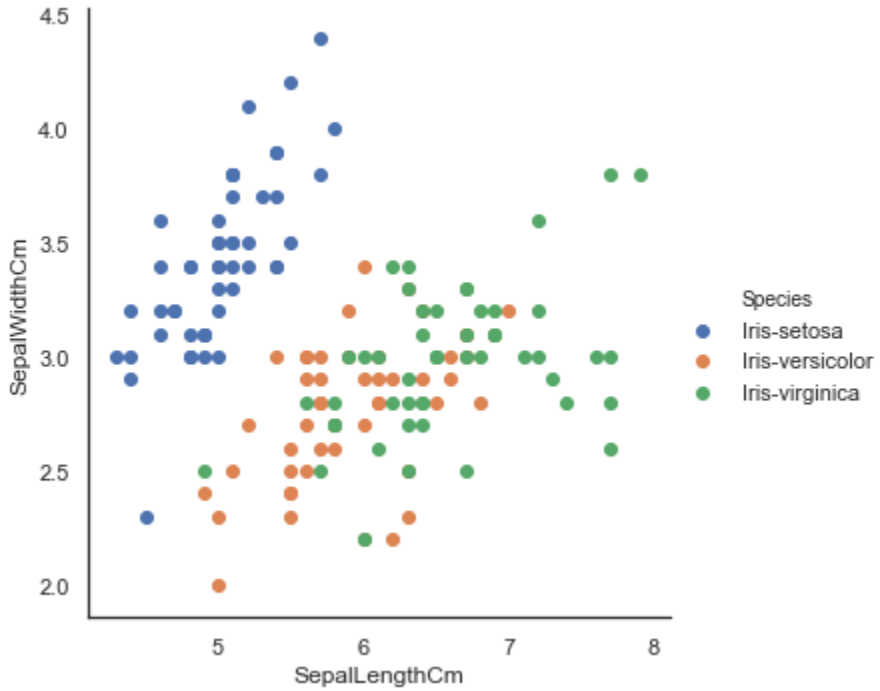


Good: Histograms for x and y axes allow viewer to see that both axes have roughly normal distribution. The main regression line shows the negative correlation between sepal length and sepal width. The visualization has high information density.
Bad: No differentiation by species. Somewhat difficult/strange to read a histogram sideways.



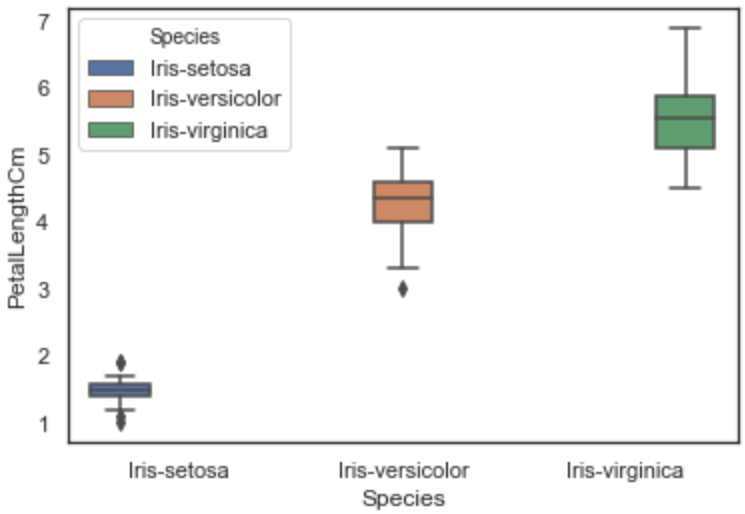
Good: Similar to the previous chart, but is showing the positive correlation between petal width and petal length. The change of color is good so the two visualizations don't get mixed up.

Bad: The scatter plot points are so densely clustered it is hard to read them. They should be made smaller, or the image should be made larger.

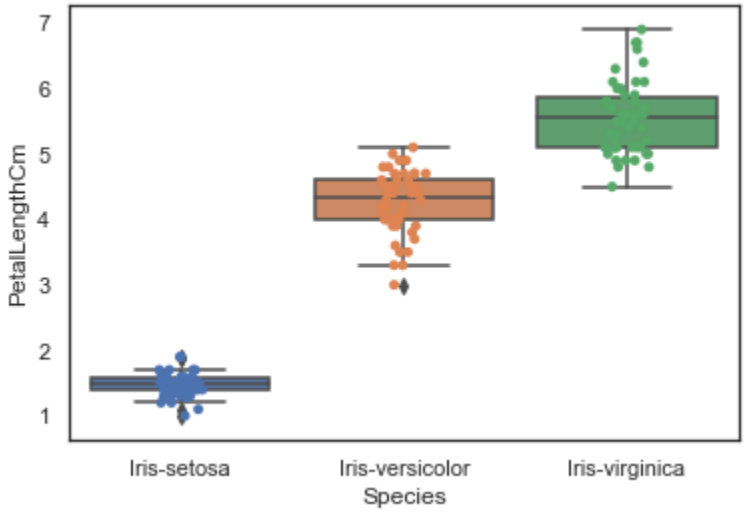


Good: Can compare the correlation between sepal length and sepal width between each species. Reveals that Iris-setosa has a stronger positive correlation than the other two species. Iris-virginica seems to have the most outliers.

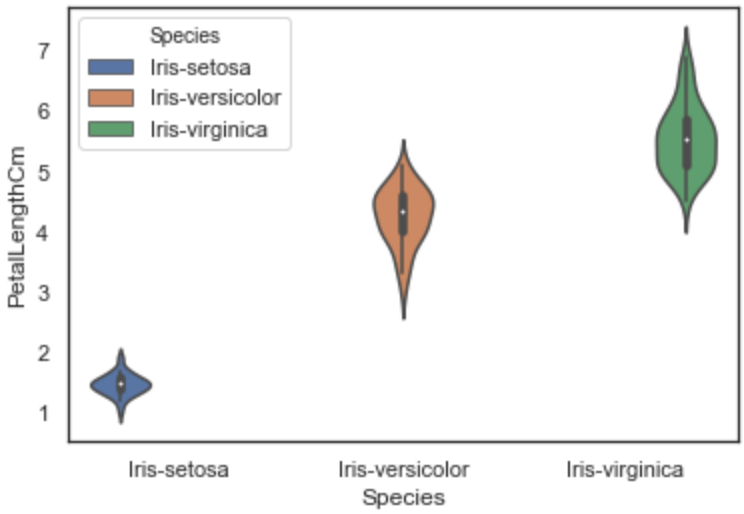
Bad: Adding a line for each correlation might make it easier to view, especially where the brown and green dots overlap.



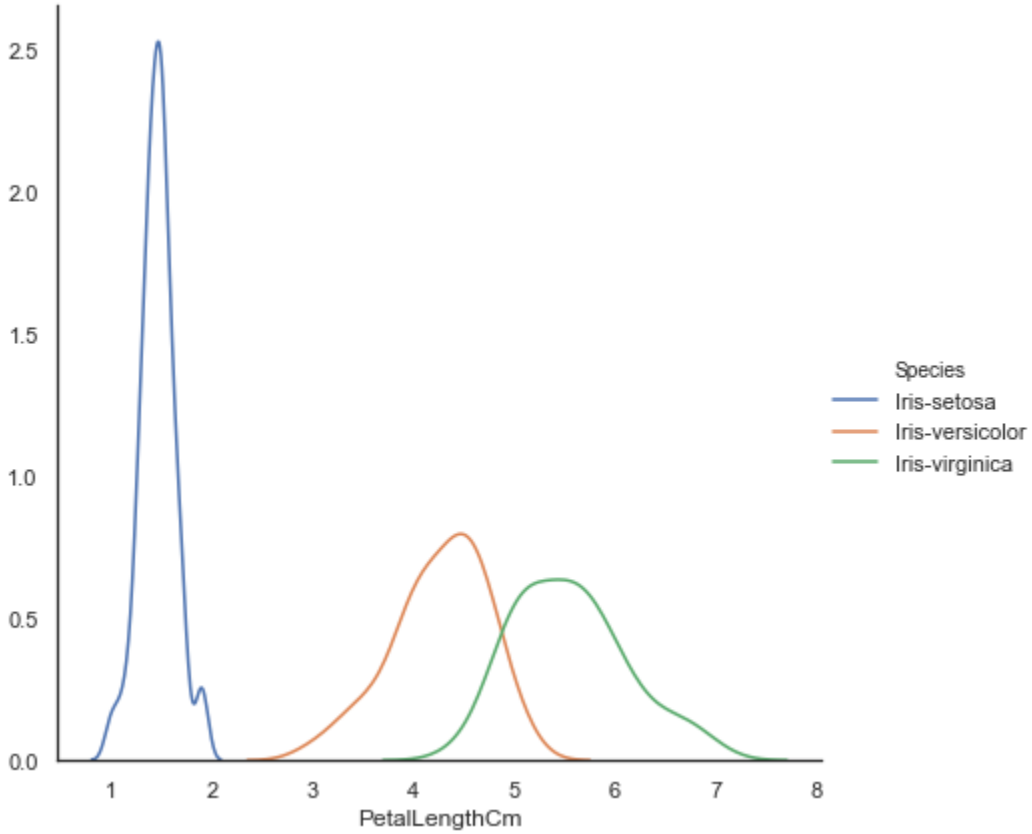
Good: Easy to compare the box plots between species.
Bad: Hard to know what the numbers are for each box plot, as there are no guiding lines on the image.



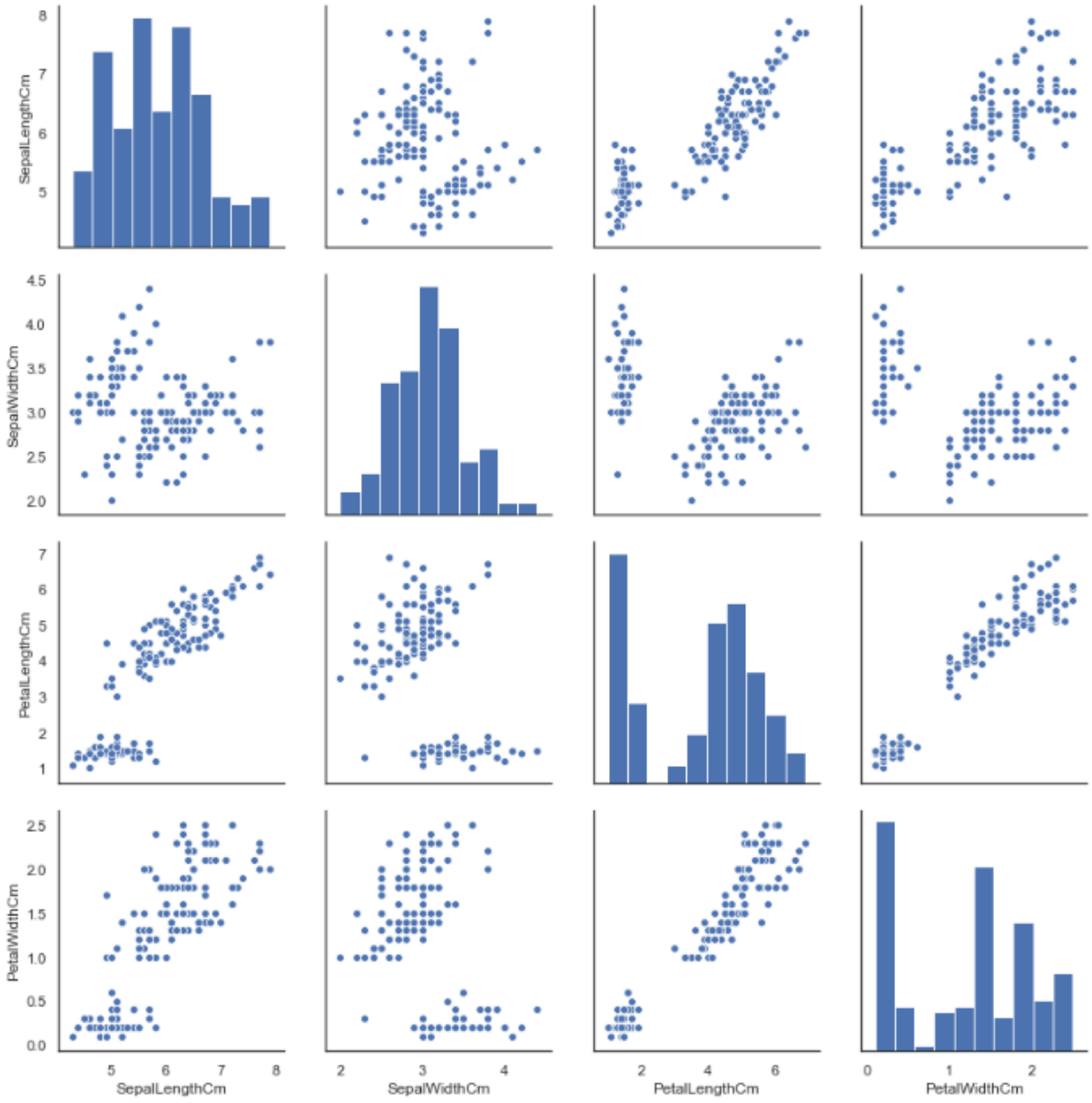
Good: Easy to compare the box plots between species.
Bad: Hard to know what the numbers are for each box plot, as there are no guiding lines on the image. Also, the dots are a bit too big and overlap with each other.



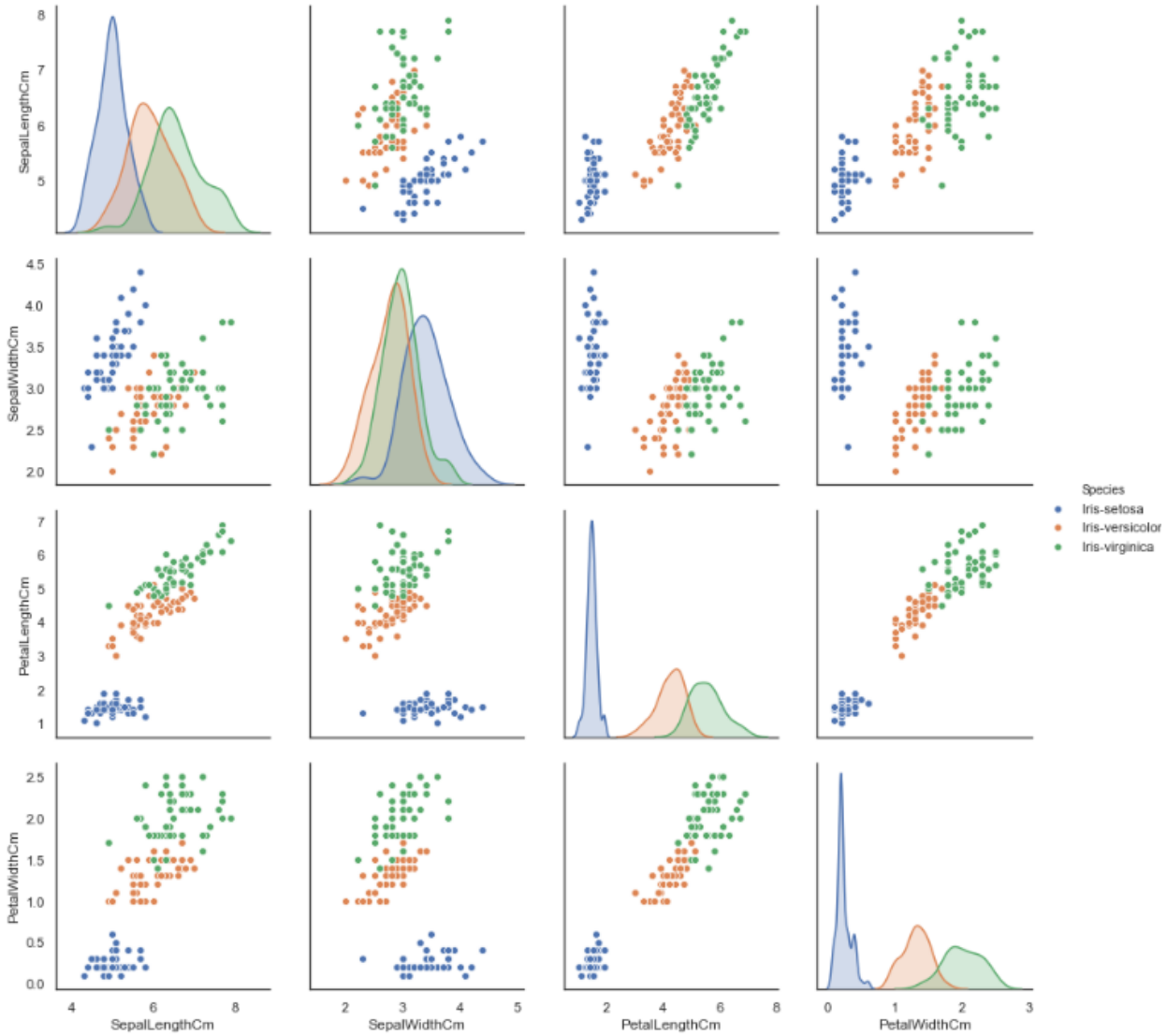
Good: Easy to compare the box plots between species.
Bad: Hard to guess the exact numbers.



Good: Can tell that Iris-setosa has the highest petal length with the smallest distribution. The distribution is larger for the other two species.
Bad: Labels indicating that this is a kernel density plot would be helpful.

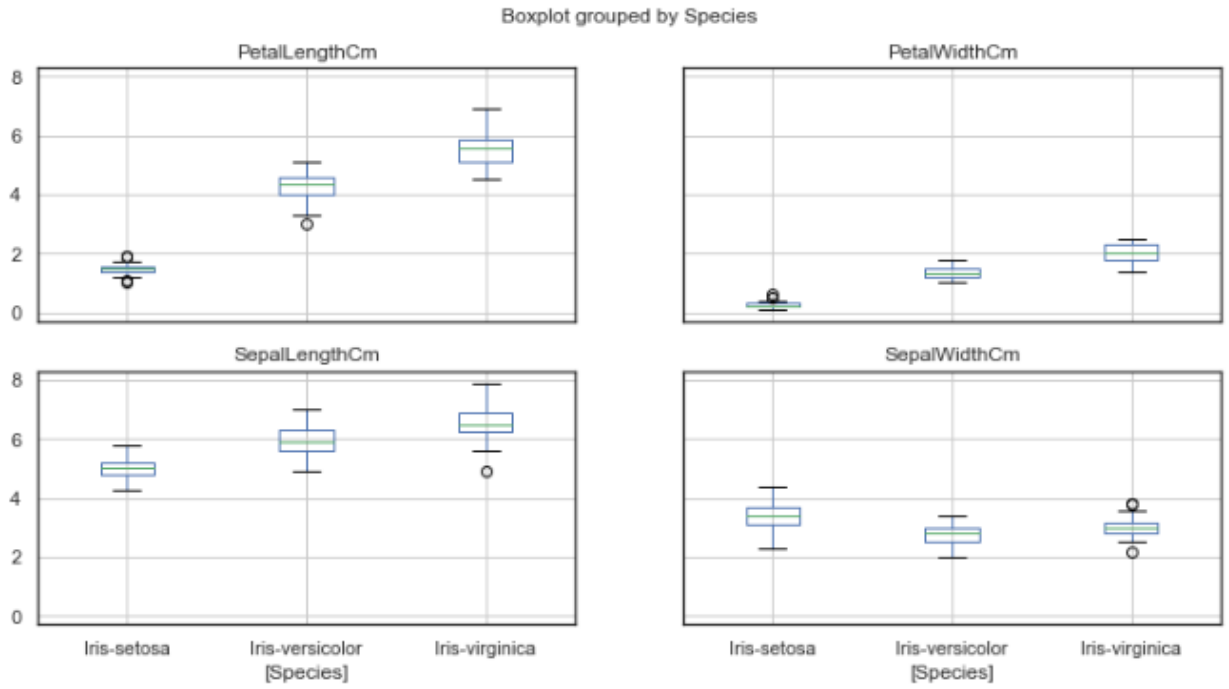


Good: Many different charts displayed on same axes; very high information density.
Bad: Somewhat difficult to take all the information in at a glance. This seems to be for more detailed analysis.



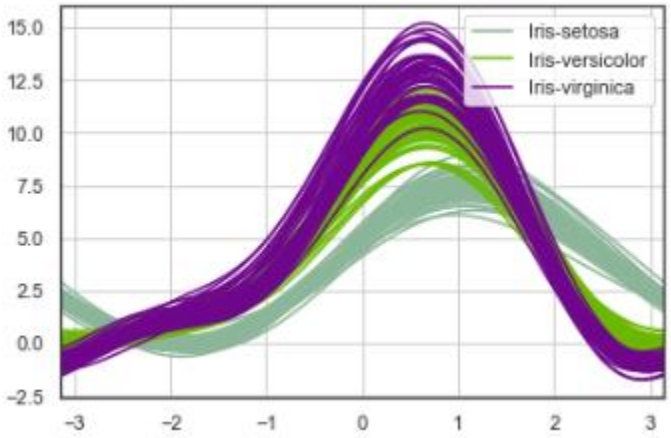
Good: Color coding is easy to read. High information density with the many different types of charts.

Bad: Difficult to understand at a glance.



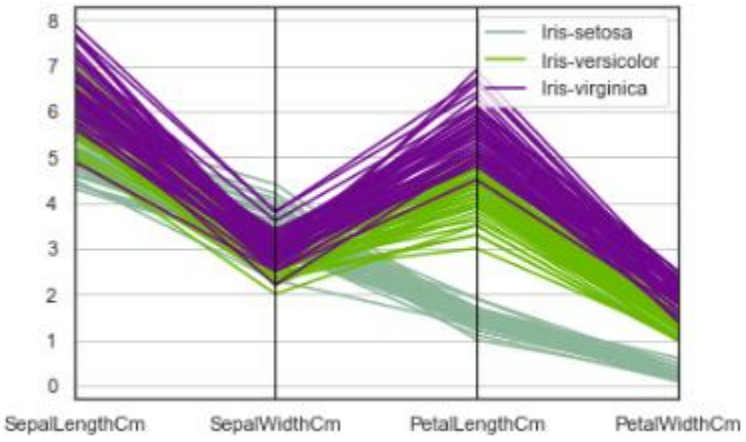
Good: Immediately get a sense of the relative lengths for each measurement in each species.

Bad: The box plots get squished so small it is difficult to read/guess the numbers they represent.

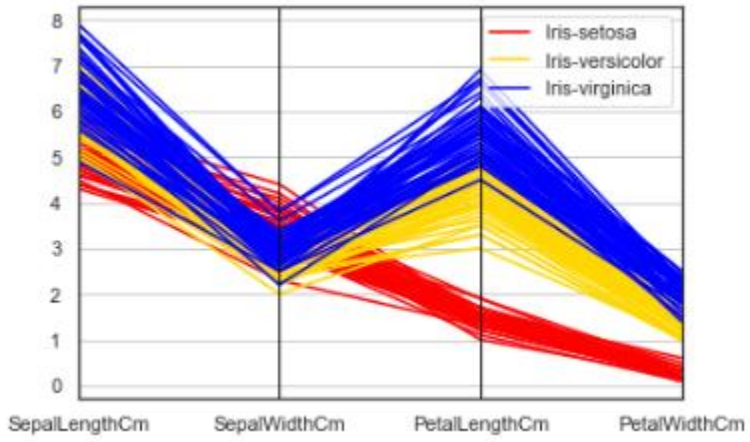


Good: Gives sense of the average length of each species, while still showing that the samples have some variance.

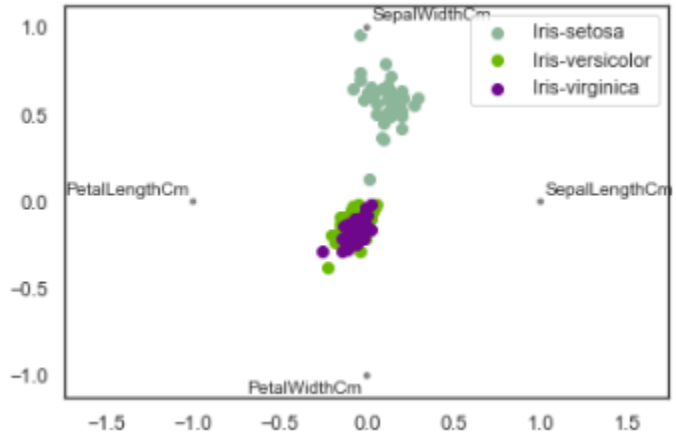
Bad: The purple lines cover up the light green lines and some of the grey-green lines.



Good: Gives sense of the average length of each measurement, while still showing that the samples have some variance.
Bad: The purple lines cover up the green lines.



Good: Gives sense of the average length of each measurement, while still showing that the samples have some variance.
Bad: The red and yellow lines are partially obscured.



Good: Interesting way to show how the Iris-setosa tends towards wide sepals while Iris-versicolor and Iris-virginica tend towards wider petals.
Bad: The green dots are partially obscured by the purple dots.

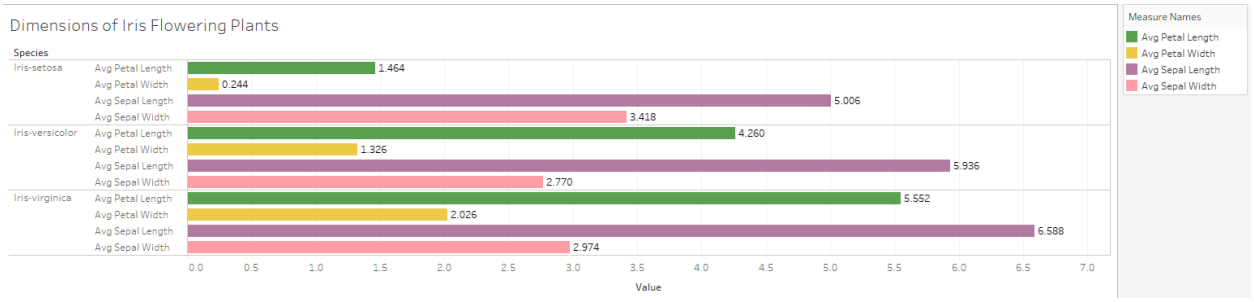
Additional Tutorials for using Jupyter Notebook:

(1) https://www.learndatasci.com/tutorials/python-pandas-tutorial-complete-introduction-for-beginners/
(2) https://www.datacamp.com/community/tutorials/tutorial-jupyter-notebook#WhatIs
(3) https://www.learnpython.org/en/Pandas_Basics
(4) https://realpython.com/tutorials/data-viz/

Task 3: Create 4 different visualizations using Tableau and summarize your major findings from each visualization.

Please copy and paste your Tableau visualization below, along with your summaries.

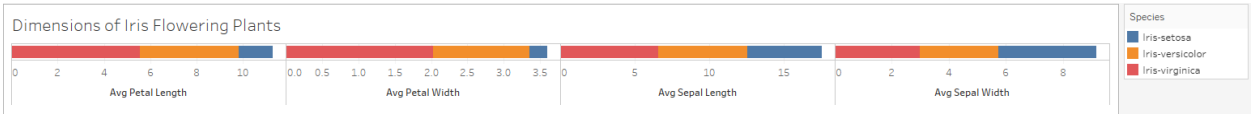
Visualization 1



Summary 1:

Good: This is a bar chart of the average petal/sepal measurements for each species, with the measurement types color-coded. The four measurements are grouped by the species they were taken from, so it is relatively easy to compare the sizes between species. We also quickly notice a pattern that sepal length the longest measurement, and petal width is the shortest. I chose to use a horizontal bar chart rather than vertical since the bars are so long; it is easier to compare the long bars when they are vertically stacked, especially since there are 12 different bars.
Bad: It is a little bit difficult to track the top of the bars over to the numbers on the bottom, so I added the measurement numbers right above each bar. This graph could possibly be improved by including a table with the numbers next to it in a dashboard, and removing the numbers from the bars, as they are a bit distracting from the image and make it look cluttered.

Visualization 2

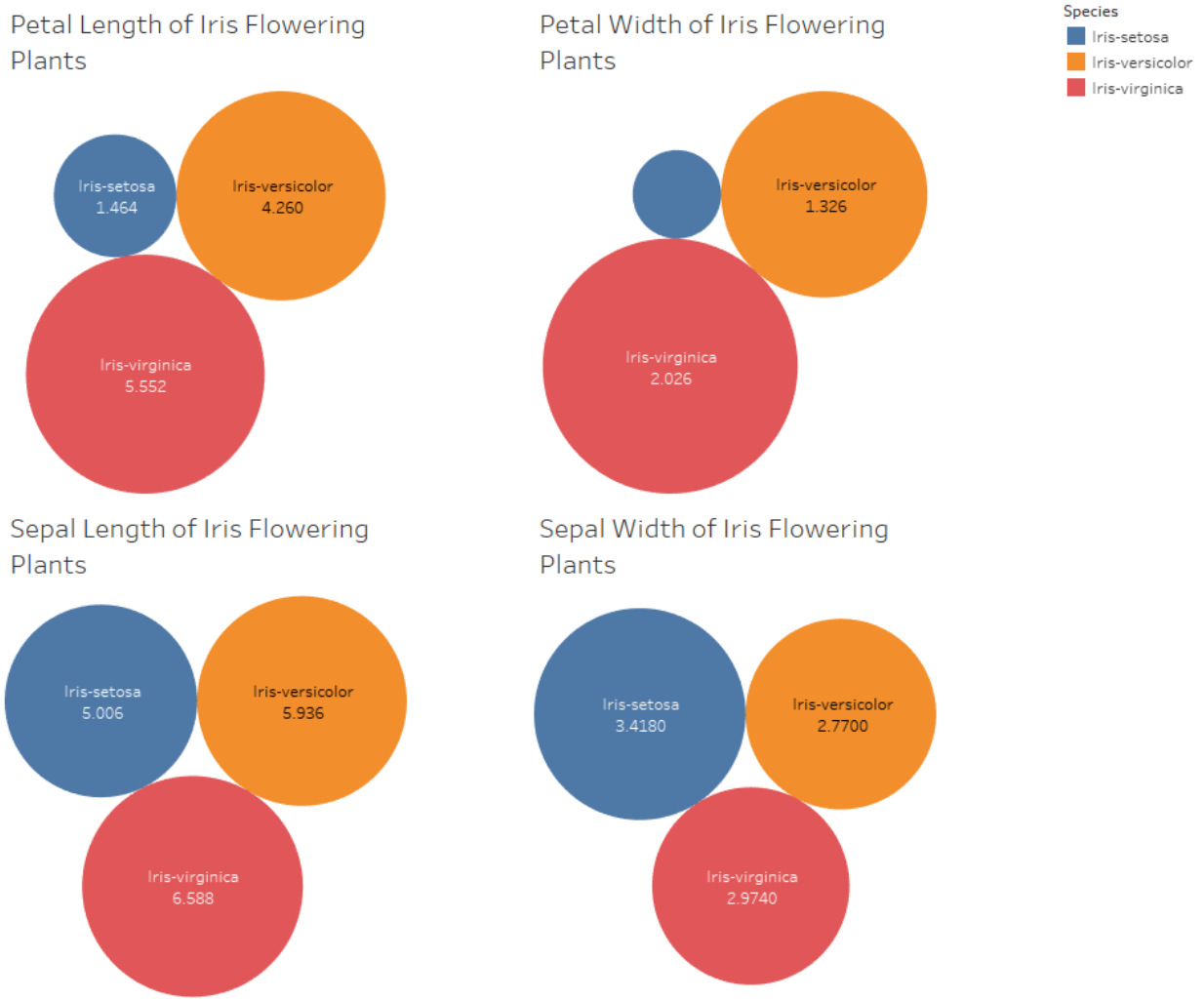


Summary 2:

Good: These horizontal bars show the relative size of each species for the four different measurements. The viewer can quickly get a sense of the relative size of each plant species compared to the others. For example, Iris-virginica obviously has larger petals than the other two species, but its sepal sizes are more average.

Bad: This visualization makes it difficult to know the exact number of the average length for the four measurements.

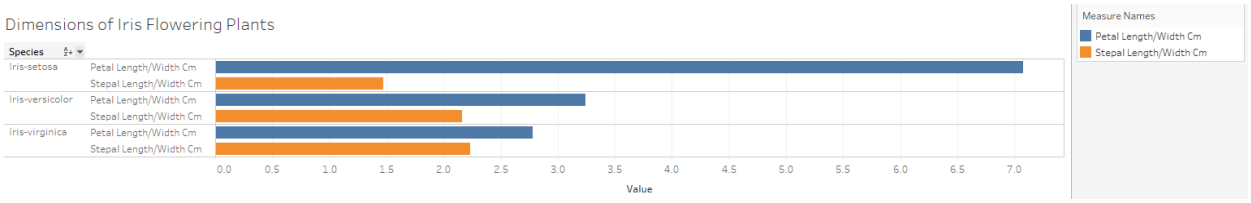
Visualization 3



Summary 3:

Good: These bubbles are scaled by the average for each measurement, and the bubbles are color-coded by the species. I decided to make this visualization to give a more physical-feeling comparison of the sizes of the four measurements between the three species. Comparing the sizes of bars can make the viewer forget that these are measurements of physical objects (plants). Having the bubbles right next to each other feels more natural to me. I was also able to include the precise measurement number inside of each bubble along with the species name for easier readability.
Bad: The bubble for Iris-setosa petal width is so small that the name/number disappear. Also, bubbles are a rather imprecise way to compare the sizes, since the colors and blank space can create visual illusions that make small bubbles appear larger than they really are, and vice-versa.

Visualization 4



Summary 4:
Good: I created two new calculated fields with the formulas AVG([Petal Length Cm]/[Petal Width Cm]) and AVG([Sepal Length Cm]/[Sepal Width Cm]) to measure the ratio of length/width for both petal and sepals. I thought it would be interesting to see which species has the most extremely long and narrow or short and broad petals/sepals. The visualization shows us that Iris-setosa has the highest length/width ratio for petals and the lowest length/width ratio for sepals. This could be useful information for botany researchers.
Bad: The information in this visualization is limited to this one question of length/width ratio.