Data Visualization

Prerequisites:

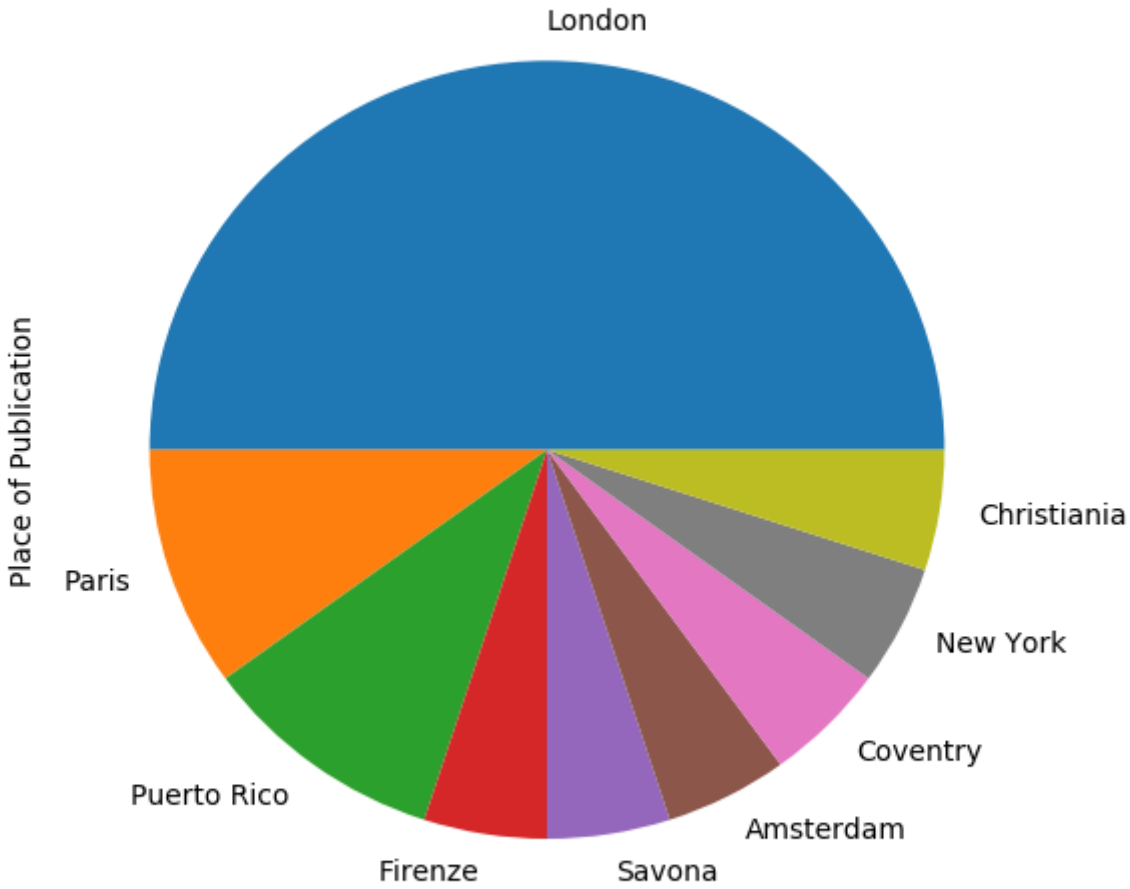
It is assumed that you will take a look at the following packages in python before heading to activities:

- pandas

This lab makes use of two datasets: Books.csv as you are already familiar with, and the iris dataset (https://raw.githubusercontent.com/mysilver/COMP9321-Data-Services/master/Week4_Visualization/iris.csv) . This dataset has four features including sepal_length, sepal_width, petal_length, and petal_length of three species of flowers: setosa, versicolor, and virginica.

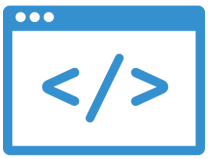
Activity-1:

Description : Plot a pie chart which illustrate the percentages of books published in each place (see the Books.csv (https://github.com/mysilver/COMP9321-Data-Services/blob/master/Week3_Data_Cleansing/Books.csv) which is used in previous labs). Here is what your chart should be like:



Steps :

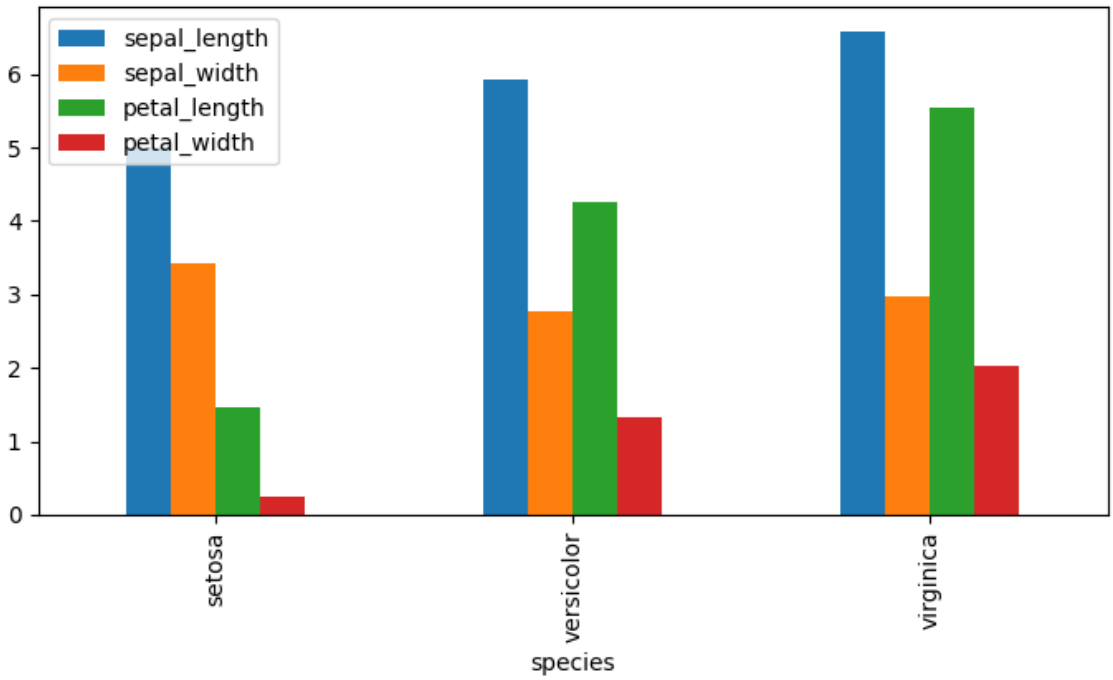
1. Load the CSV file into a dataframe
2. Select the "Place of Publication" column (https://pandas.pydata.org/pandas-docs/stable/indexing.html)
3. Optional: Clean the dataframe as you did in Lab-3 (https://webcms3.cse.unsw.edu.au/COMP9321/18s2/resources/19959)
4. Count the values in the selected column to know how many times each unique value has appeared in the column (https://pandas.pydata.org/pandas-docs/version/0.22/generated/pandas.Series.value_counts.html)
5. Plot a pie chart (https://pandas.pydata.org/pandas-docs/version/0.23/generated/pandas.DataFrame.plot.pie.html) for the value counts calculated in the previous step
6. To show the chart you need to ask matplotlib (https://stackoverflow.com/questions/34347145/pandas-plot-doesnt-show)



(https://github.com/mysilver/COMP9321-Data-Services/blob/master/Week4_Visualization/activity_1.py)

Activity-2:

Description : This activity is based on the iris dataset. You need to create a bar chart as below which indicates the average length and widths of sepal and petals based on species:



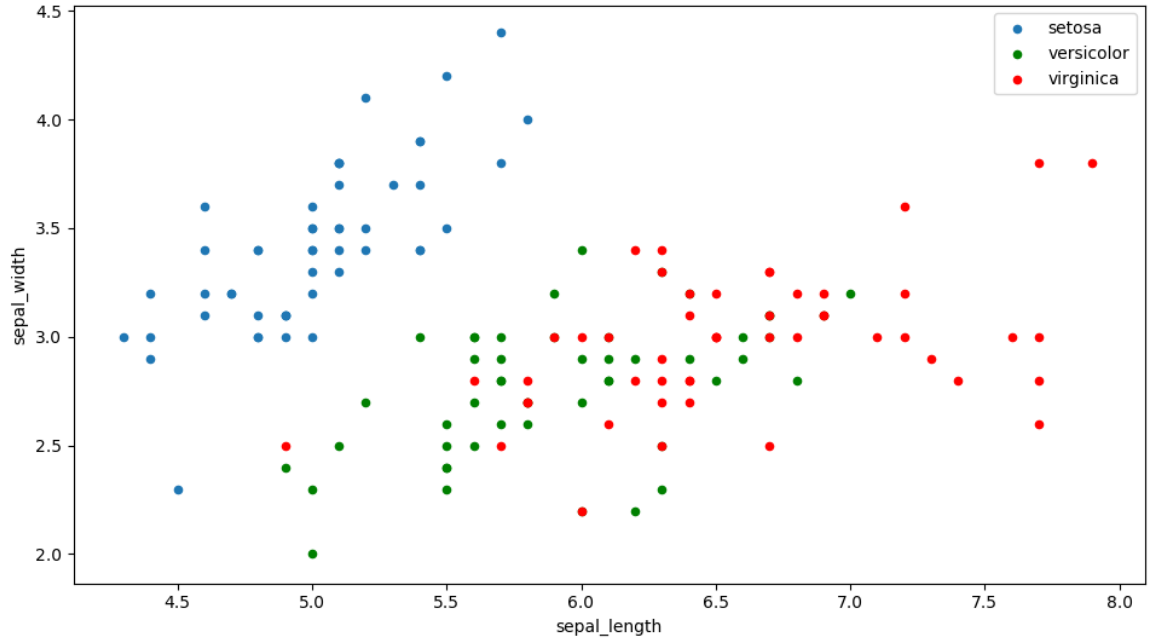
Steps :

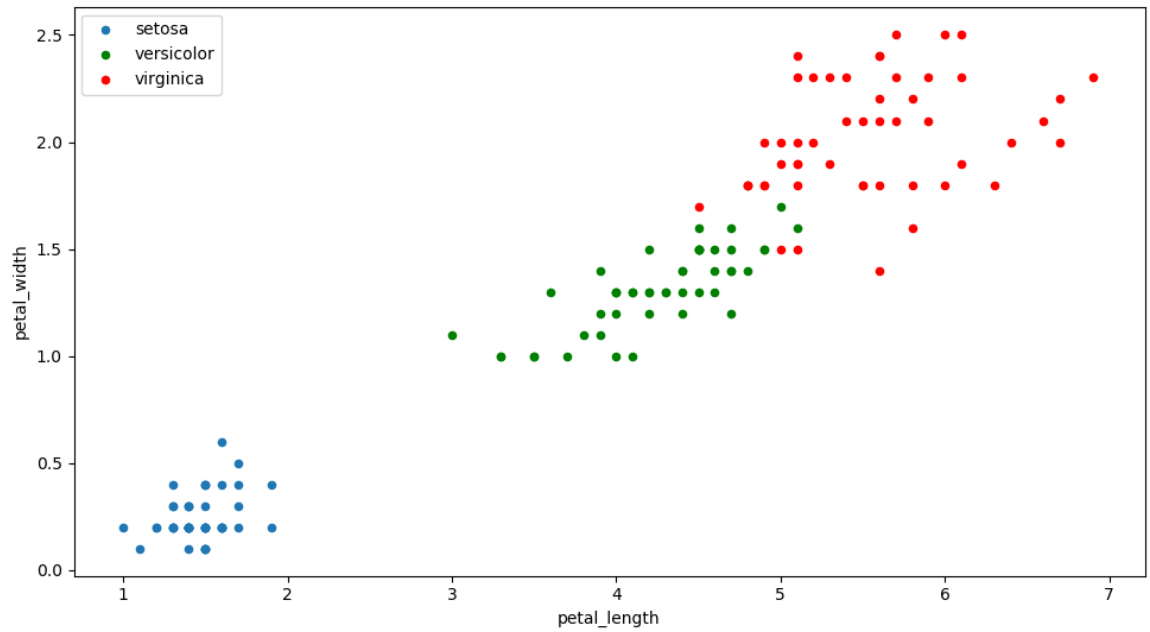
1. Load the CSV file into a dataframe
2. Group by the dataframe based on the "species" and calculate the mean (https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.groupby.html)
3. Plot a bar chart (https://pandas.pydata.org/pandas-docs/version/0.23/generated/pandas.DataFrame.plot.bar.html)
4. To show the chart you need to ask matplotlib (https://stackoverflow.com/questions/34347145/pandas-plot-doesnt-show)

(https://github.com/mysilver/COMP9321-Data-Services/blob/master/Week4_Visualization/activity_2.py)

Activity-3:

Description : This activity is also based on the iris dataset. You need to plot two scatter charts as shown as blow:





Steps :

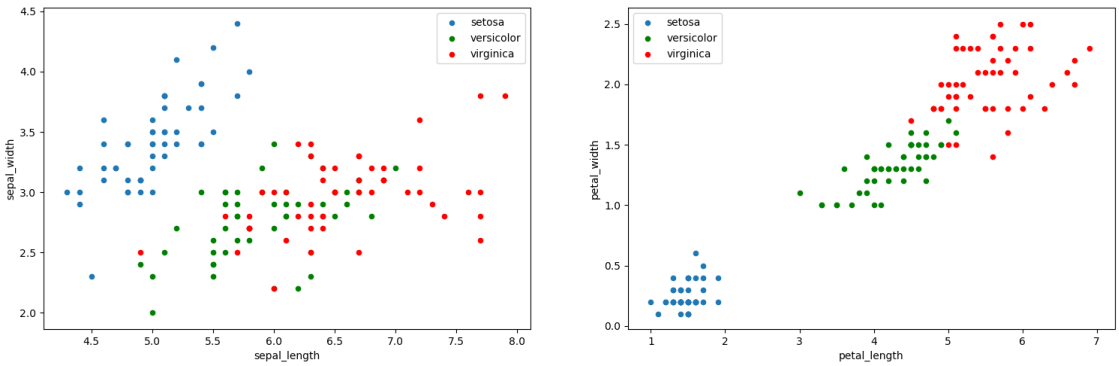
1. Load the CSV file into a dataframe
2. Split the dataset into three dataframes based on the three species; you can use panda's query method (https://pandas.pydata.org/pandas-docs/version/0.22/generated/pandas.DataFrame.query.html) to filter rows as you did in the previous labs.
3. Plot a scatter chart using x='sepal_length', y='sepal_width', and separate colors for each of the three dataframes (https://pandas.pydata.org/pandas-docs/version/0.23/generated/pandas.DataFrame.plot.scatter.html)
4. You need to link the three scatter charts to plot them in a single figure (https://stackoverflow.com/questions/13872533/plot-different-dataframes-in-the-same-figure/45225366#45225366)
5. Repeat the last two steps with x='petal_length', y='petal_width'

6. To show the chart you need to ask matplotlib (https://stackoverflow.com/questions/34347145/pandas-plot-doesnt-show)

(https://github.com/mysilver/COMP9321-Data-Services/blob/master/Week4_Visualization/activity_3.py)

Activity-4:

Description : In the previous activity, you created two separate figures; In this activity, you must plot both charts into one figure; in other words, you need to show the charts side by side as follows:



Steps :

1. Copy and paste the code you wrote for the previous activity
2. Manually create the subplots with matplotlib (https://stackoverflow.com/questions/22483588/how-can-i-plot-separate-pandas-dataframes-as-subplots#22484249) While creating the subplots, set the number of rows to 1 and columns to 2
3. To show the chart you need to ask matplotlib (https://stackoverflow.com/questions/34347145/pandas-plot-doesnt-show)

(https://github.com/mysilver/COMP9321-Data-Services/blob/master/Week4_Visualization/activity_4.py)

Resource created 2 months ago (Monday 14 February 2022, 01:12:13 PM), last modified 2 months ago (Monday 07 March 2022, 10:30:56 AM).

Comments

(/COMP9321/22T1/forums/search?forum_choice=resource/73144)
(/COMP9321/22T1/forums/resource/73144)

Add a comment

Harsimran Saini (/users/z5208912) 2 months ago (Tue Mar 08 2022 19:35:29 GMT+0800 (China Standard Time)), last modified 2 months ago (Tue Mar 08 2022 19:45:58 GMT+0800 (China Standard Time))

I'm getting the following error for Activity 1 on Vlab:

```
Traceback (most recent call last):
  File "activity1.py", line 32, in <module>
    counts.plot.pie(subplots=True)
  File "/usr/lib/python3/dist-packages/pandas/plotting/_core.py", line 2908, in pie
    return self(kind='pie', **kwds)
  File "/usr/lib/python3/dist-packages/pandas/plotting/_core.py", line 2741, in __call__
    **kwds)
  File "/usr/lib/python3/dist-packages/pandas/plotting/_core.py", line 2002, in plot_series
    **kwds)
  File "/usr/lib/python3/dist-packages/pandas/plotting/_core.py", line 1804, in _plot
    plot_obj.generate()
  File "/usr/lib/python3/dist-packages/pandas/plotting/_core.py", line 258, in generate
    self._compute_plot_data()
  File "/usr/lib/python3/dist-packages/pandas/plotting/_core.py", line 363, in _compute_plot_data
    "timedelta"])
  File "/usr/lib/python3/dist-packages/pandas/core/frame.py", line 3077, in select_dtypes
    include_these = Series(not bool(include), index=self.columns)
  File "/usr/lib/python3/dist-packages/pandas/core/series.py", line 275, in __init__
    raise_cast_failure=True)
  File "/usr/lib/python3/dist-packages/pandas/core/series.py", line 4149, in _sanitize_array
    value, len(index), dtype)
  File "/usr/lib/python3/dist-packages/pandas/core/dtypes/cast.py", line 1201, in construct_1d_arraylike_from_scalar
    subarr = np.empty(length, dtype=dtype)
TypeError: Cannot interpret '<attribute 'dtype' of 'numpy.generic' objects>' as a data type
```

it seems to be originating from line 32 of code below:

```
28  if __name__ == "__main__":
29      books_df = pd.read_csv("Books.csv")
30      books_df = clean(books_df)
31      counts = books_df['Place of Publication'].value_counts()
32      counts.plot.pie(subplots=True)
33
34      plt.show()
35
```

However, my solution is very similar to the sample solution above and when I search the error it seems to be due to an incompatibility of pandas and people recommend going to an earlier version but I can't do try that on vlab so I wanted to ask if anyone else also had this and how they fixed it.

EDIT:

I realised I could install the newer version of pandas onto Vlab so I ran the following:

pip3 install pandas==1.0.5

and that fixed the issue. Not sure what the exact versions we should have for the labs and assignment, does anyone have the answer to that?

Reply

Gordon Chen (/users/z5161163) 2 months ago (Tue Mar 08 2022 21:38:15 GMT+0800 (China Standard Time))

maybe you have an outdated version of pandas in your cse account. to get the latest version:

pip3 install pandas --upgrade

Reply