Pandas

Session 2

TECH TALENT ACADEMY

WOMEN IN DATA ACADEMY

WOMEN IN DATA ACADEMY

Session Content

Viewing Data

Calculations

Using Booleans

Filtering

Viewing Data – Head, Tail and Shape

When reading from a file you can use various commands to view your data.

- **DataFrame.shape** will show the number of rows and columns within the data set.

```
In [17]: import pandas as pd

         data = pd.read_csv("ign.csv")

         data.shape

Out[17]: (18625, 11)
```

You will see that the ign.csv data set has 18625 rows and 11 columns

DataFrame.head() prints the **first** X amount of rows in your dataframe.

```
In [16]: import pandas as pd

         data = pd.read_csv("ign.csv")

         data.head(3)
```

Out[16]:

	Unnamed: 0	score_phrase	title	url	
0	0	Amazing	LittleBigPlanet PS Vita	/games/littlebigplanet-vita/vita-98907	F
1	1	Amazing	LittleBigPlanet PS Vita -- Marvel Super Hero E...	/games/littlebigplanet-ps-vita-marvel-super-he...	F
2	2	Great	Splice: Tree of Life	/games/splice/ipad-141070	

DataFrame.tail() prints the **last** X amount of rows in your dataframe.

```
In [14]: import pandas as pd

         data = pd.read_csv("ign.csv")

         data.tail(4)
```

Out[14]:

	Unnamed: 0	score_phrase	title	url
18621	18621	Amazing	LEGO Star Wars: The Force Awakens	/games/lego-star-wars-the-force-awakens/ps4-20...
18622	18622	Mediocre	Star Ocean: Integrity and Faithlessness	/games/star-ocean-5/ps4-20035681
18623	18623	Masterpiece	Inside	/games/inside-playdead/xbox-one-121435
18624	18624	Masterpiece	Inside	/games/inside-playdead/pc-20055740

Using () will automatically retrieve 5 rows. To specify another amount enter this within the parentheses – e.g. (10) will retrieve 10 rows.

Pandas Series Objects

```
In [7]: import pandas as pd

        data = pd.read_csv("ign.csv")

        data["platform"]
```

```
Out[7]: 0        PlayStation Vita
        1        PlayStation Vita
        2                    iPad
        3                 Xbox 360
        4            PlayStation 3
                        ...
        18620               Wii U
        18621        PlayStation 4
        18622        PlayStation 4
        18623             Xbox One
        18624                   PC
        Name: platform, Length: 18625, dtype: object
```

There is a 3rd way to retrieve information from a column in Pandas . You can specify the column name in single square brackets []

The information you retrieve is called a series object – this is a single column. In comparison to the Data Frame that stores tabular data this stores single column or row.

Verify single column is a Series

```
In [10]: import pandas as pd

         data = pd.read_csv("ign.csv")

         type(data["platform"])
```

```
Out[10]: pandas.core.series.Series
```

Creating Dataframes – another method

You can use Pandas to create dataframes.

```
In [3]: import pandas as pd

        frame = pd.DataFrame([[8.96, 1884], [7.87, 1149], [7.13, 428]],
                             index = ["Copper", "Iron", "Zinc"],
                             columns = ["Density g/cm3", "Melting Point BC"])

        print(frame)
```

```
        Density g/cm3  Melting Point BC
Copper           8.96              1884
Iron             7.87              1149
Zinc             7.13               428
```

You can then use the loc function to isolate specific data.

```
In [5]: print(frame.loc["Copper"])

Density g/cm3          8.96
Melting Point BC    1884.00
Name: Copper, dtype: float64
```

DataFrame Methods

There are multiple types of calculations that can be conducted using Pandas, such as calculating the mean:

This will find the mean of a series (column)

```
In [2]: import pandas as pd

        data = pd.read_csv("ign.csv")

        data["score"].mean()

Out[2]: 6.950459060402666
```

```
In [4]: import pandas as pd

        data = pd.read_csv("ign.csv")

        data.mean()

Out[4]: Unnamed: 0       9312.000000
        score               6.950459
        release_year     2006.515329
        release_month       7.138470
        release_day        15.603866
        dtype: float64
```

This will find the mean of each numerical column in the DataFrame (this is known as the pandas.DataFrame.mean method)

DataFrame Maths

```
In [11]: data["score"] / 2

Out[11]: 0        4.50
         1        4.50
         2        4.25
         3        4.25
         4        4.25
                  ...
         18620    3.80
         18621    4.50
         18622    2.90
         18623    5.00
         18624    5.00
         Name: score, Length: 18625, dtype: float64
```

This will divide every value in the score column by 2

```
In [13]: data["score"] * 10

Out[13]: 0         90.0
         1         90.0
         2         85.0
         3         85.0
         4         85.0
                  ...
         18620     76.0
         18621     90.0
         18622     58.0
         18623    100.0
         18624    100.0
         Name: score, Length: 18625, dtype: float64
```

This will multiply every value in the score column by 10.

Boolean Indexing

```
In [8]: import pandas as pd

        data = pd.read_csv("ign.csv")

        myfilter = data["score"] > 8

        myfilter

Out[8]: 0         True
        1         True
        2         True
        3         True
        4         True
                 ...
        18620    False
        18621     True
        18622    False
        18623     True
        18624     True
        Name: score, Length: 18625, dtype: bool
```

Boolean operators allow you to generate values that you can use for comparison.

This shows us where each row either meets or does not meet the condition that the score is greater than 8.

Boolean Indexing – creating filters

Once you have created your Boolean comparison, you can use this to create a filter and select rows where a True value applies:

```
In [13]: highscore = data[myfilter]

         highscore. head()
```

Out[13]:

	Unnamed: 0	score_phrase	title	url	platform	score	genre	e
0	0	Amazing	LittleBigPlanet PS Vita	/games/littlebigplanet-vita/vita-98907	PlayStation Vita	9.0	Platformer	
1	1	Amazing	LittleBigPlanet PS Vita -- Marvel Super Hero E...	/games/littlebigplanet-ps-vita-marvel-super-he...	PlayStation Vita	9.0	Platformer	
2	2	Great	Splice: Tree of Life	/games/splice/ipad-141070	iPad	8.5	Puzzle	
3	3	Great	NHL 13	/games/nhl-13/xbox-360-128182	Xbox 360	8.5	Sports	
4	4	Great	NHL 13	/games/nhl-13/ps3-128181	PlayStation 3	8.5	Sports	

Filtering with Multiples

You can use multiple conditions to filter:

```
In [10]: import pandas as pd

         data = pd.read_csv("ign.csv")

         data[(data.score > 8) & (data.platform == "iPad")]
```

Out[10]:

	Unnamed: 0	score_phrase	title	url	platform	score	ge
2	2	Great	Splice: Tree of Life	/games/splice/ipad-141070	iPad	8.5	Pu
26	26	Amazing	Bastion	/games/bastion/ipad-140874	iPad	9.0	Ac F
52	52	Amazing	The World Ends with You: Solo Remix	/games/the-world-ends-with-you-solo-remix/ipad...	iPad	9.5	F
137	137	Amazing	The Walking Dead: The Game -- Episode 3: Long ...	/games/the-walking-dead-season-1-episode-3/ipa...	iPad	9.0	Adven
247	247	Amazing	The Walking Dead: The Game --	/games/the-walking-dead-	iPad	9.5	Adven

Creating Visualisations

In Pandas you can also create visualisations. These include:

- Line Plot
- Scatter Plot
- Area Plot
- Bar Chart
- Pie Chart
- Histogram
- Kernel Density Function
- Box Plot
- Scatter Matrix Plot

Setting up the Data

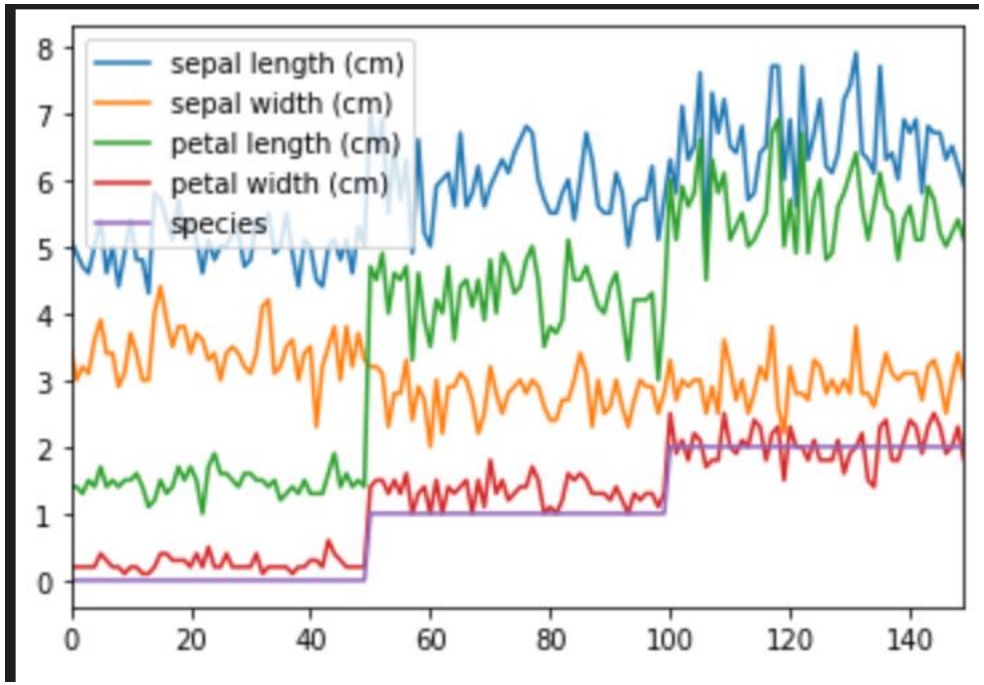
We will be using the package Sci-Kit Learn to create the data set

```
from sklearn.datasets import load_iris

data = load_iris()
df   = pd.DataFrame(data['data'], columns=data['feature_names'])

df['species'] = data['target']

df.head()
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0

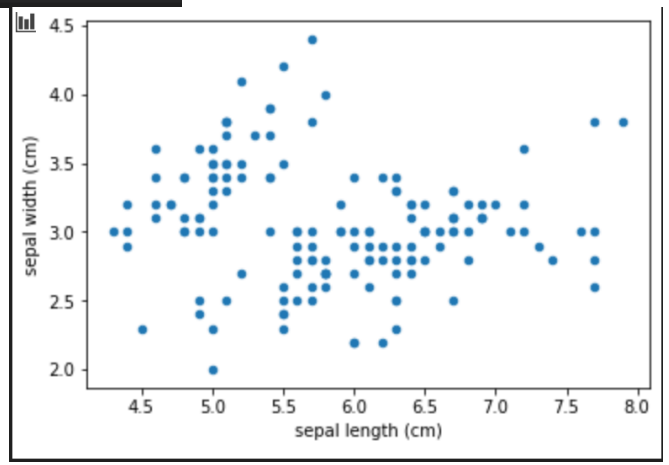
Line Plot

```
df.plot()
```

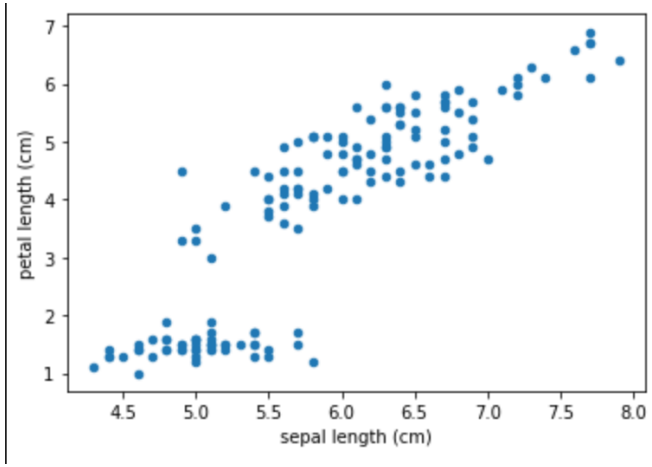


Scatter Plot

```
df.plot.scatter(x='sepal length (cm)', y='sepal width (cm)')
```

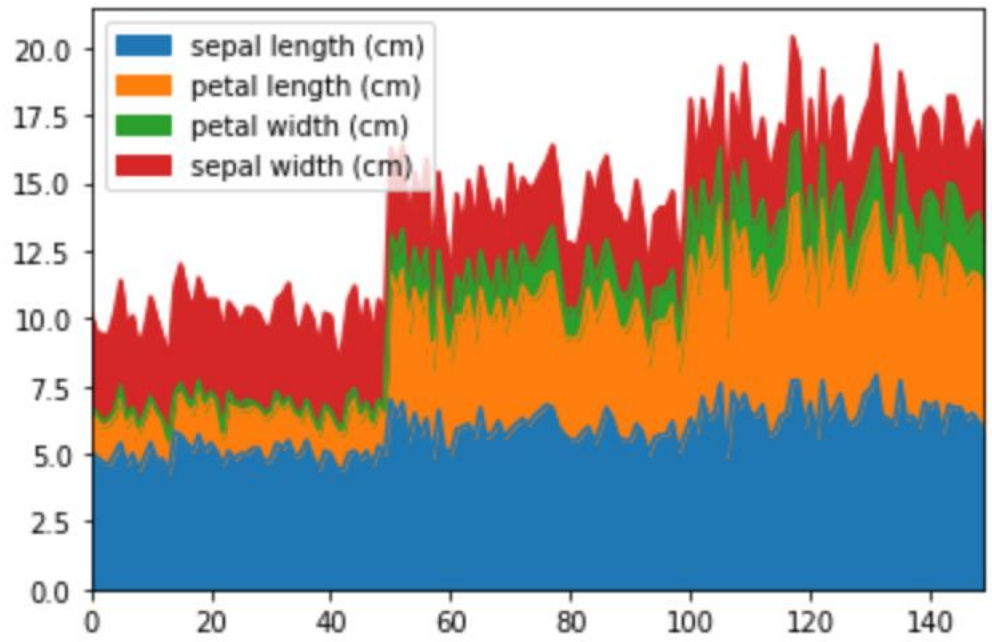



```
df.plot.scatter(x='sepal length (cm)', y='petal length (cm)')
```



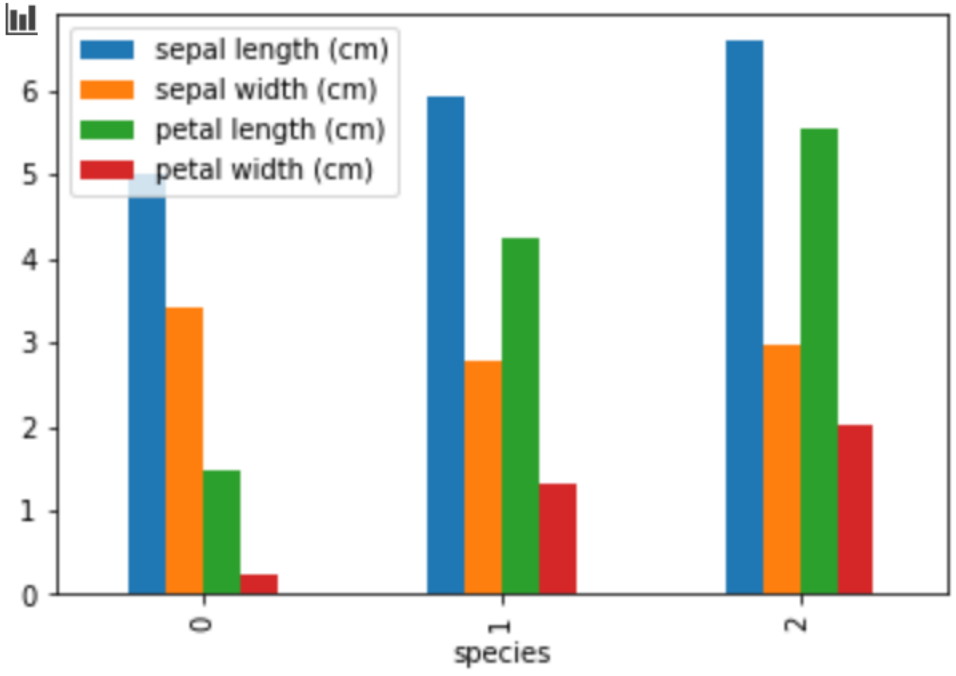
Area Plot

```
columns = ['sepal length (cm)', 'petal length (cm)', 'petal width (cm)', 'sepal width (cm)']
df[columns].plot.area()
```



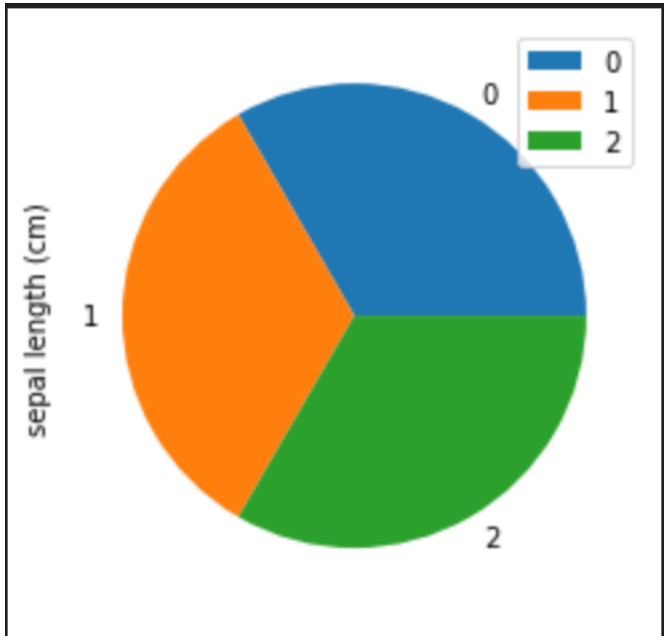
Bar Chart

```
df.groupby('species').mean().plot.bar()
```



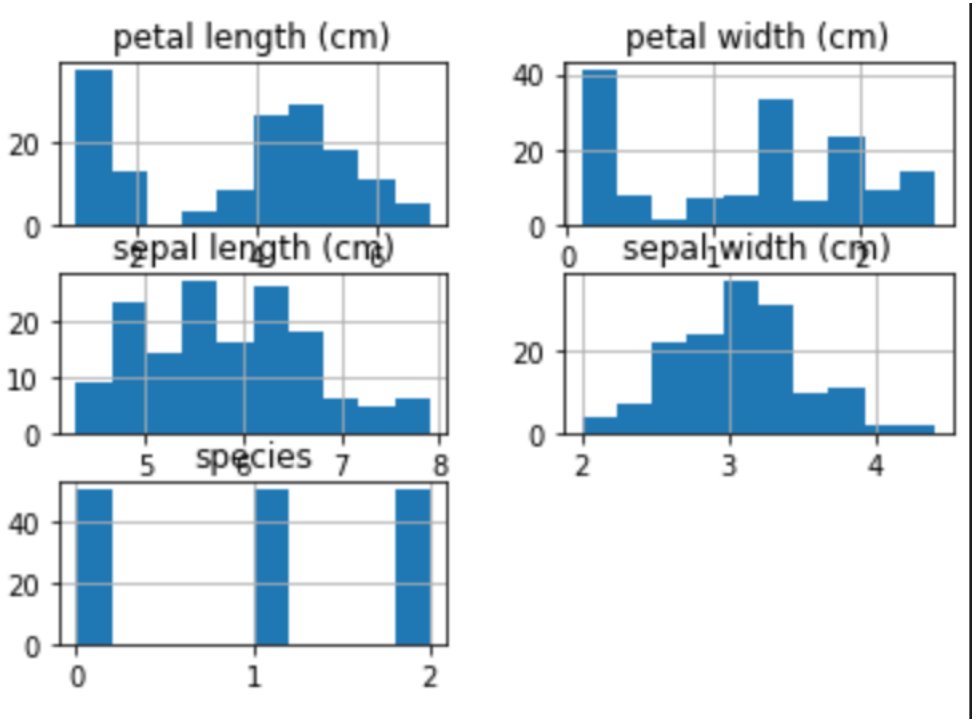
Pie Chart

```
df.groupby('species').count().plot.pie(y='sepal length (cm)')
```



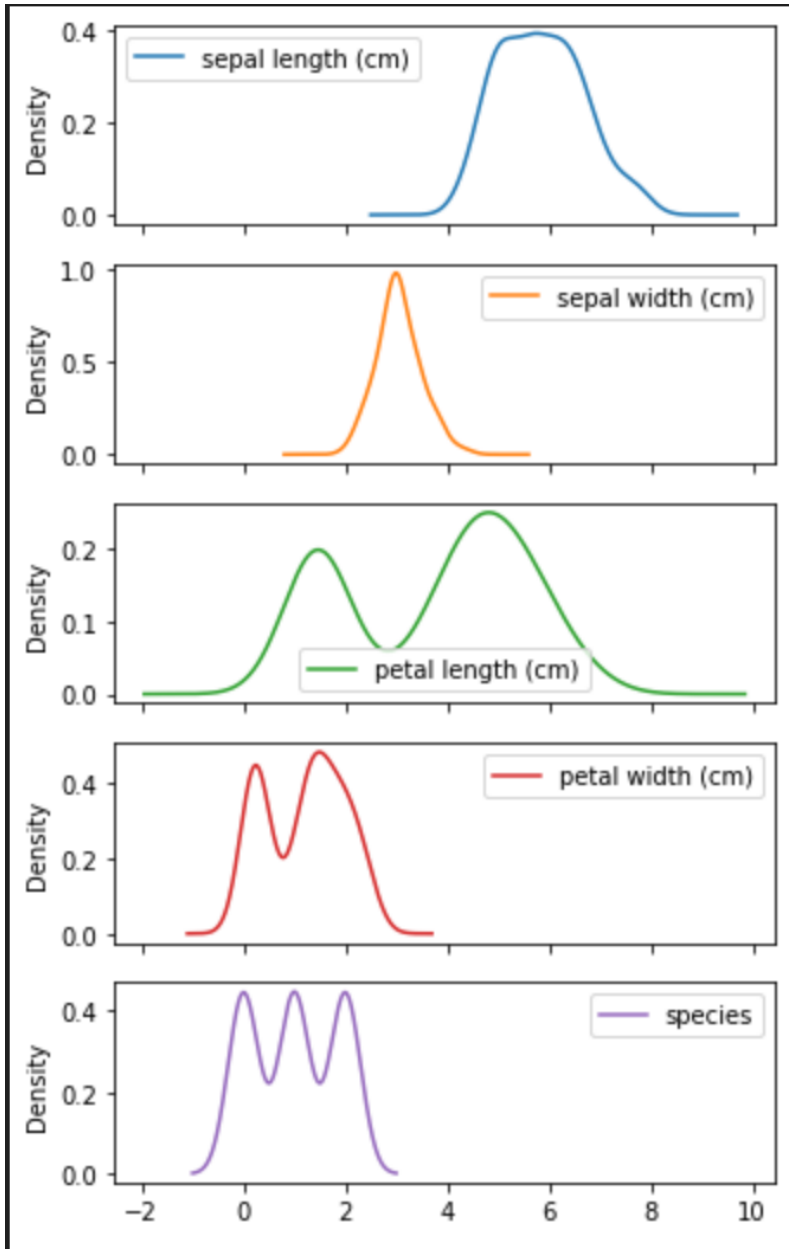
Histogram

```
df.hist()
```



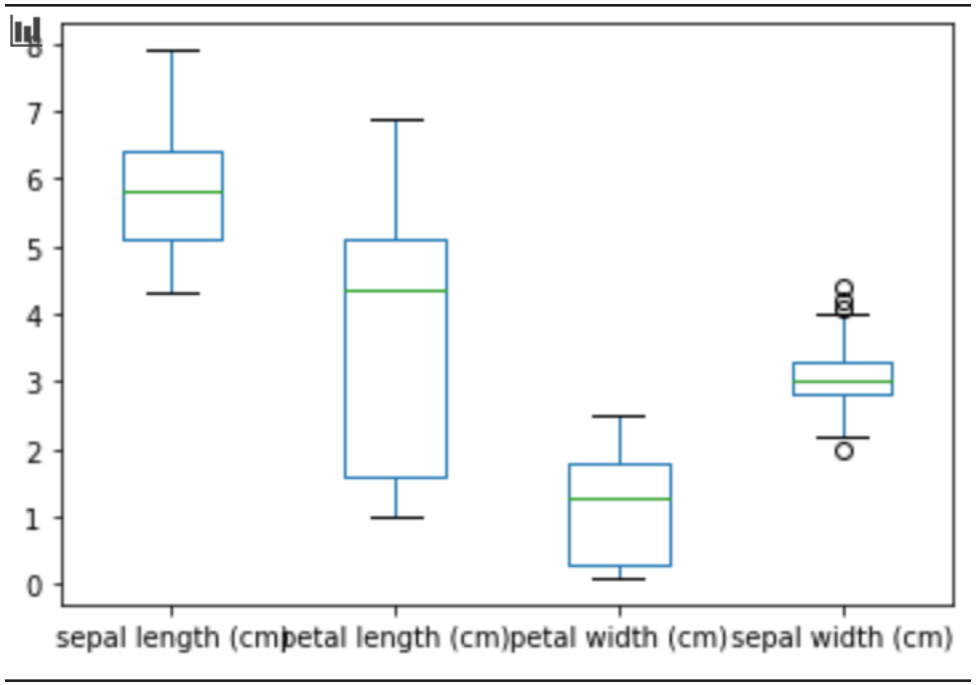
Kernel Density Function

```
df.plot.kde(subplots=True, figsize=(5,9))
```



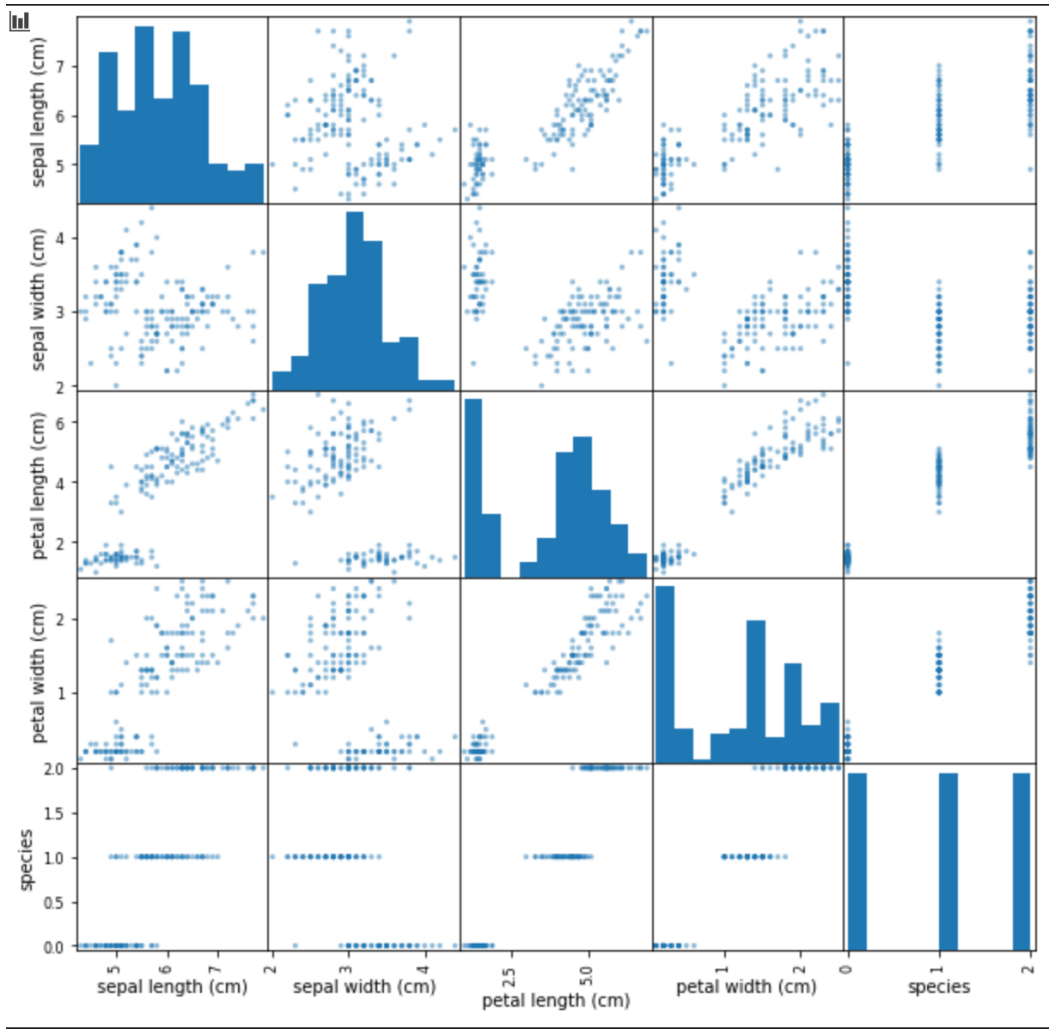
Box Plot

```
columns = ['sepal length (cm)', 'petal length (cm)', 'petal width (cm)', 'sepal width (cm)']
df[columns].plot.box()
```



Scatter Matrix Plot

```
from pandas.plotting import scatter_matrix

scatter_matrix(df, figsize=(10, 10))
```



Home Learning Tasks & Random

Lesson Tasks

Create a CSV file of 15 holiday destinations for a website

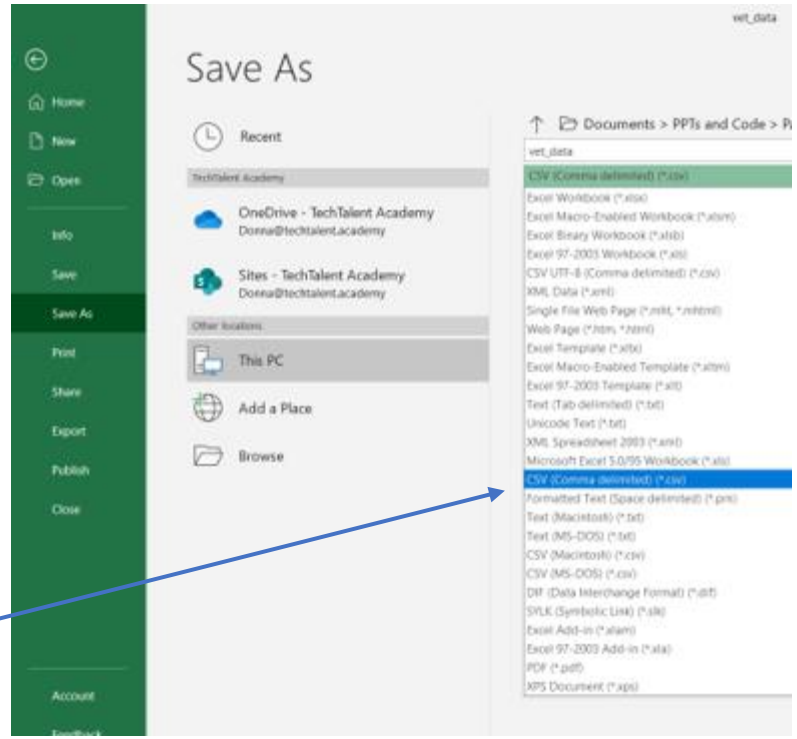
1. Add in a column of destinations
2. Add in a column that shows feedback score out of 10 for that destination
3. Add in a column for average hotel star rating for those destinations
4. Add in a column for number of all-inclusive hotels within each destination
5. Add in the most visited city in each destination

Note: the data that you use to create your csv file can be hypothetical.

Creating a csv file

To create your csv file, add your data into a blank excel workbook

	A	B	C	D	E	F
1	Owner_Surname	Pet_Name	Pet_Age	Last_Visit	Type	Chipped
2	Adams	Fluffy	2	11/05/2020	Cat	Yes
3	Smith	Zuko	8	30/01/2019	Dog	Yes
4	Radcliffe	Nala	4	24/11/2019	Dog	Yes
5	Holland	Mr Chips	10	15/06/2019	Dog	No
6	Potter	Daisy	9	07/04/2020	Rabbit	Yes
7	Sorola	Oscar	1	27/02/2020	Hamster	No
8	Pike	Pepsi	6	18/09/2019	Cat	Yes
9	Murray	George	17	12/02/2020	Tortoise	No
10	Aston	Monty	3	09/03/2020	Dog	Yes
11	Waller	Flo	7	06/05/2019	Cat	Yes
12	De la Force	Anton	2	31/08/2019	Cat	No
13	Reed	Farah	5	02/03/2020	Horse	No
14	Martinez	Homer	6	24/11/2019	Dog	Yes
15	Li	Iggy	5	14/12/2019	Dog	Yes
16	Rodriquez	Bobby	3	28/02/2020	Rabbit	No
17						



Then **Save As,** and from the file type drop down box locate **CSV** from the list.

Practical

Complete the following data analysis, taking screenshots/evidence for each one , save into a document prepare a presentation for the customer on (you need to show the output for each one and also be able to show how you reached the result) :

1. How many rows and columns are there in your file?
2. Print row 3-8 (using iloc/loc).
3. Find the mean number of all-inclusive hotels across all destinations.
4. Find the lowest scoring destination.
5. Find the highest scoring destination.
6. Find all the destinations where there are more than 9 all-inclusive hotels.
7. Filter the data by destination and score above 8.
8. Filter the data by destination and score below 2 (I need to know if these destinations should be removed or there is a problem)
9. Is there a correlation between number of all-inclusive hotels and score?
10. Create a data visualisation diagram to show destination and highest scores?

WOMEN IN DATA ACADEMY

TECH TALENT ACADEMY