Data Science Foundations

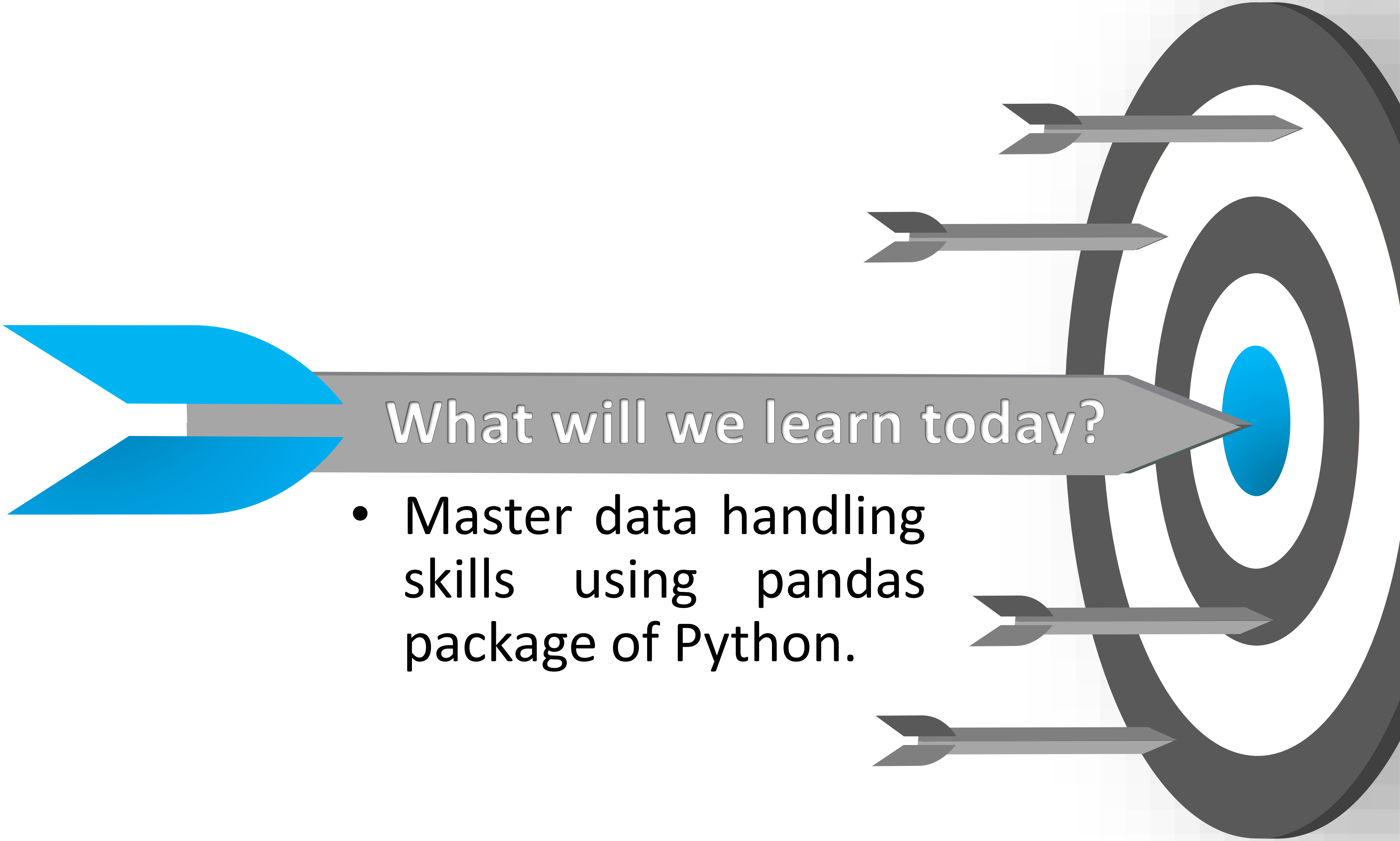
Master in Big Data Solutions 2017-2018

Liana Napalkova
liana.napalkova@bts.tech

Ludovico Boratto
ludovico.boratto@bts.tech

Francisco Guitierres
francisco.gutierres@bts.tech

Today’s Objective



What will we learn today?

- Master data handling skills using pandas package of Python.

Contents

Analysis of homework

Data handling using Titanic data

Data handling using Matplotlib and Seaborn

Today's objective

- Master data handling skills using pandas package of Python.

Contents

Analysis of homework

Data handling using Titanic data

Data handling using Matplotlib and Seaborn

Today's objective

- Master data handling skills using pandas package of Python.

Analysis of errors

Your homework in Session 1

- Create a new script in Jupyter Notebook named “data_handling”.

- Load the data from the file “rural_population_data.csv” into pandas DataFrame.

- Do the following tasks with the data:

1. Find all unique values of the column “Continent”
2. Calculate a grow rate based on this formula:

Growrate = (Population2100 - Population2000) / Population2000

3. Create a histogram of “Growrate”.
4. Check the existence of NaN values in the DataFrame.
5. Find the average population size in the years 2000 and 2010.
6. Check how many rows have UrbanRuralDesignation equal to “Urban”.

- Push your script to your Git repository “DataScienceFoundations”.

Analysis of errors

Your homework in Session 1

- Git pull:

  https://github.com/LianaNapalkova/BTS_MasterInBigData.git

- Go to the folder “Session_2”.

- Take the file “rural_area_assignment_analysis_of_common_errors.ipynb” and import it into your Jupyter Notebook

Analysis of errors

Your homework in Session 1

- Fix errors in your code, if needed and synchronize changes with your repository “DataScienceFoundations”.

- In your repository, please organize the files into folders according to the Session number (3 folders):

  ✓ Session_1
  ✓ Session_2
  ✓ Session_3

Contents

Analysis of homework

Data handling using Titanic data

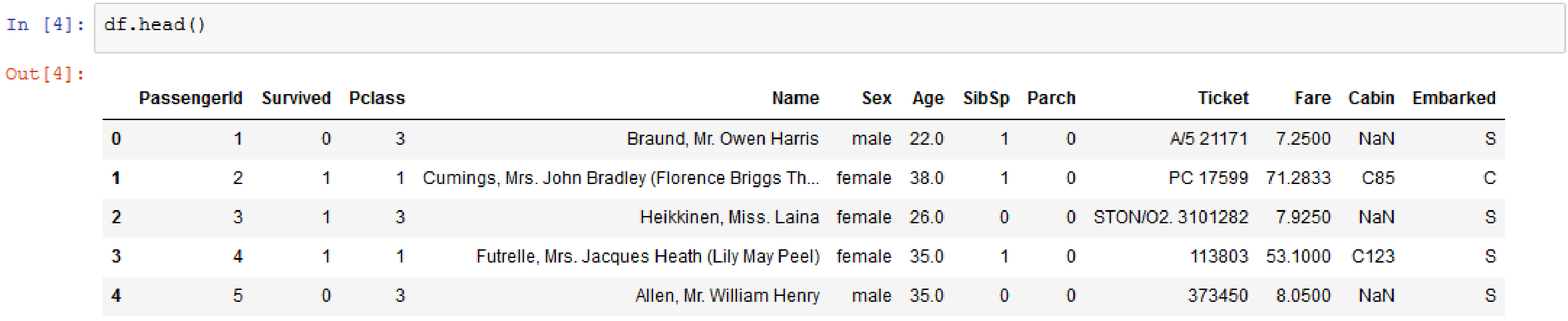
Data handling using Matplotlib and Seaborn

Today's objective

- Master data handling skills using pandas package of Python.

Data handling using Titanic data set

- **Objective:** Mastering skills
- **Data set:** BTS_MasterInBigData/Session_2/**2_titanic_dataset.csv**
- **Data dictionary:** BTS_MasterInBigData/Session_2/2_notes.pdf
- This is the anonymous dataset that defines which persons survived/not survived the Titanic disaster.



```
In [4]: df.head()
```

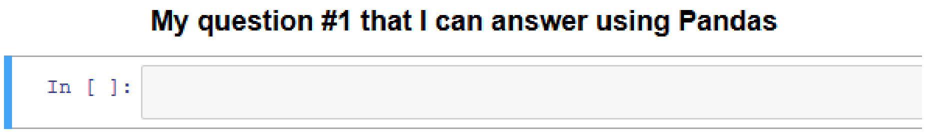
Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Data handling using Titanic data set

1. Define 20 basic questions about your data and answer them using the tools that we learned in Session 1.

2. Clearly state each questions in the comments (Cell type: Markdown).



My question #1 that I can answer using Pandas

```
In [ ]:
```

3. If you create your own function, do not forget to use docstrings.

4. Export your *.ipynb notebook and push it to your repository "DataScienceFoundations"

Data handling using Titanic data set

- Please keep in mind that your objective is to analyze which persons survived/not survived the Titanic disaster. For example, you can formulate and answer the questions similar to:
  - ✓ *What is a total family size?*
  - ✓ *Compute frequencies of Ticket, Cabin and Fare values*
  - ✓ …

- After each question, add a relevant code. Finally put a comment starting with "Knowledge retrieved:" and summarize your finding.

- You have 1 hour in order to do the assignment.

- After this, you will need to quickly (5 minutes) present your code and your findings to other classmates.

Data handling using Titanic data set

```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv("pathtocsvfile/2_titanic_dataset.csv")
```

```
df.info()

df["column_name"]

df.shape

df["column_name"].unique()

df["column_name"].min()

df["column_name"].mean()

df["column_name"].std()

df.isnull().sum()
```

```
df.dtypes

df[column_name'].hist()
plt.show()
```

Additional functions:

- loc
- apply
- fillna
- dropna

Contents

Analysis of homework

Data handling using Titanic data

Data handling using Matplotlib and Seaborn

Today’s objective

- Master data handling skills using pandas package of Python.

Data handling using Matplotlib and Seaborn

```
import matplotlib.pyplot as plt
```

matplotlib:

- python 2D plotting library which produces publication quality figures in a variety of hardcopy formats.
- line plots, scatter plots, barcharts, histograms, pie charts etc.
- relatively low-level; some effort needed to create advanced visualization.

Data handling using Matplotlib and Seaborn

```
import seaborn as sns
```

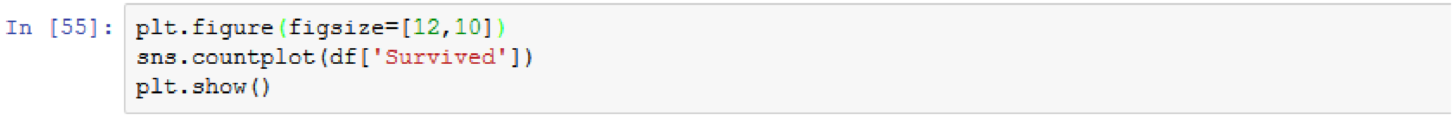
seaborn:

- based on matplotlib.
- provides high level interface for drawing attractive statistical graphics.
- Similar (in style) to the popular ggplot2 library in R.

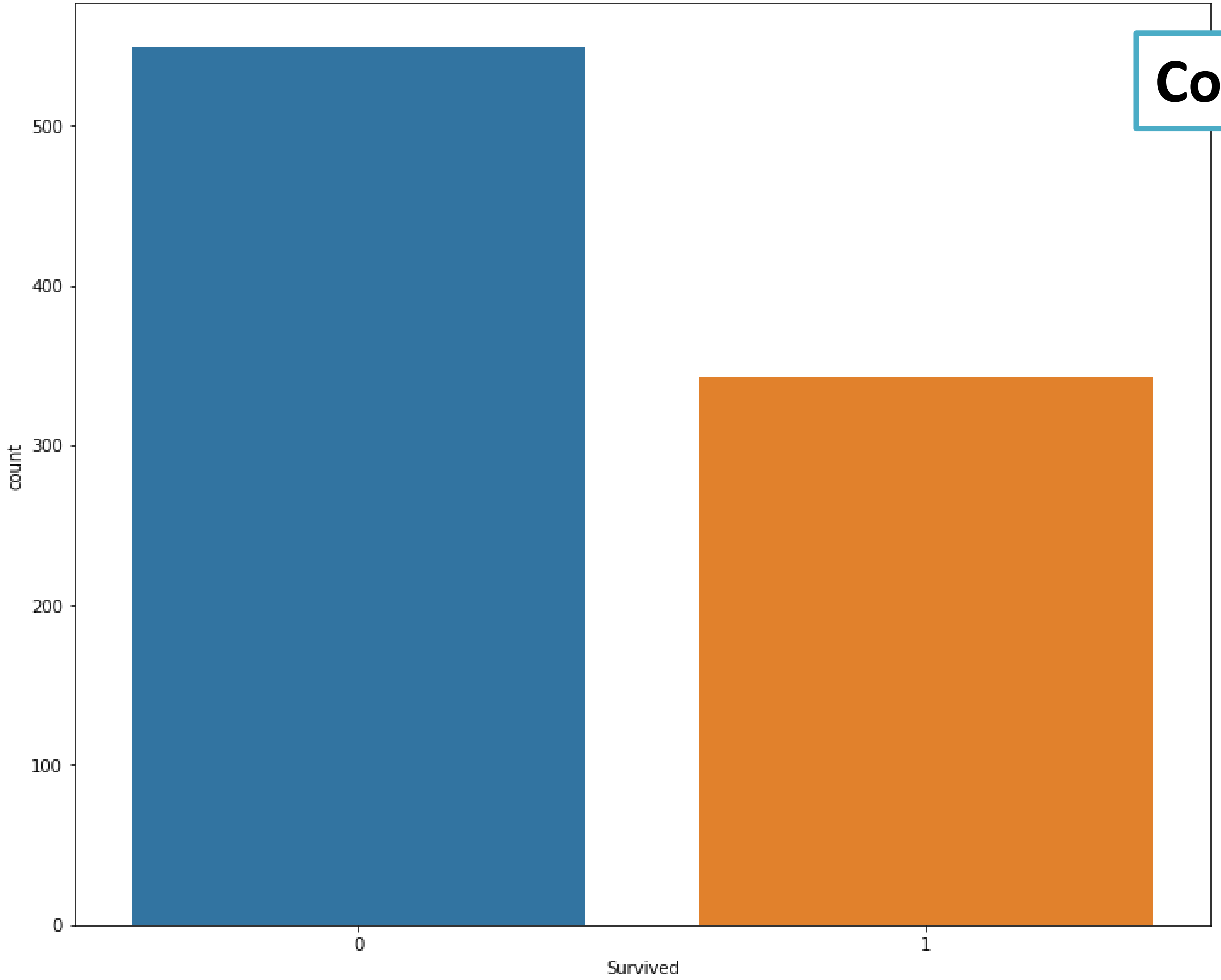
Data handling using Matplotlib and Seaborn

	description
countplot	Counter plot
barplot	Estimate of central tendency for a numeric variable
violinplot	Similar to boxplot, also shows the probability density of the data
distplot	Distribution plot
pointplot	Point or linear plot
pairplot	Pairplot
boxplot	Boxplot
swarmplot	Categorical scatterplot
factorplot	General categorical plot

Data handling using Matplotlib and Seaborn

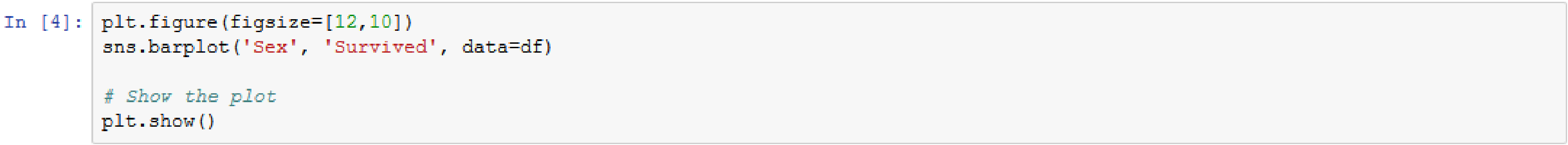


```
In [55]: plt.figure(figsize=[12,10])
         sns.countplot(df['Survived'])
         plt.show()
```

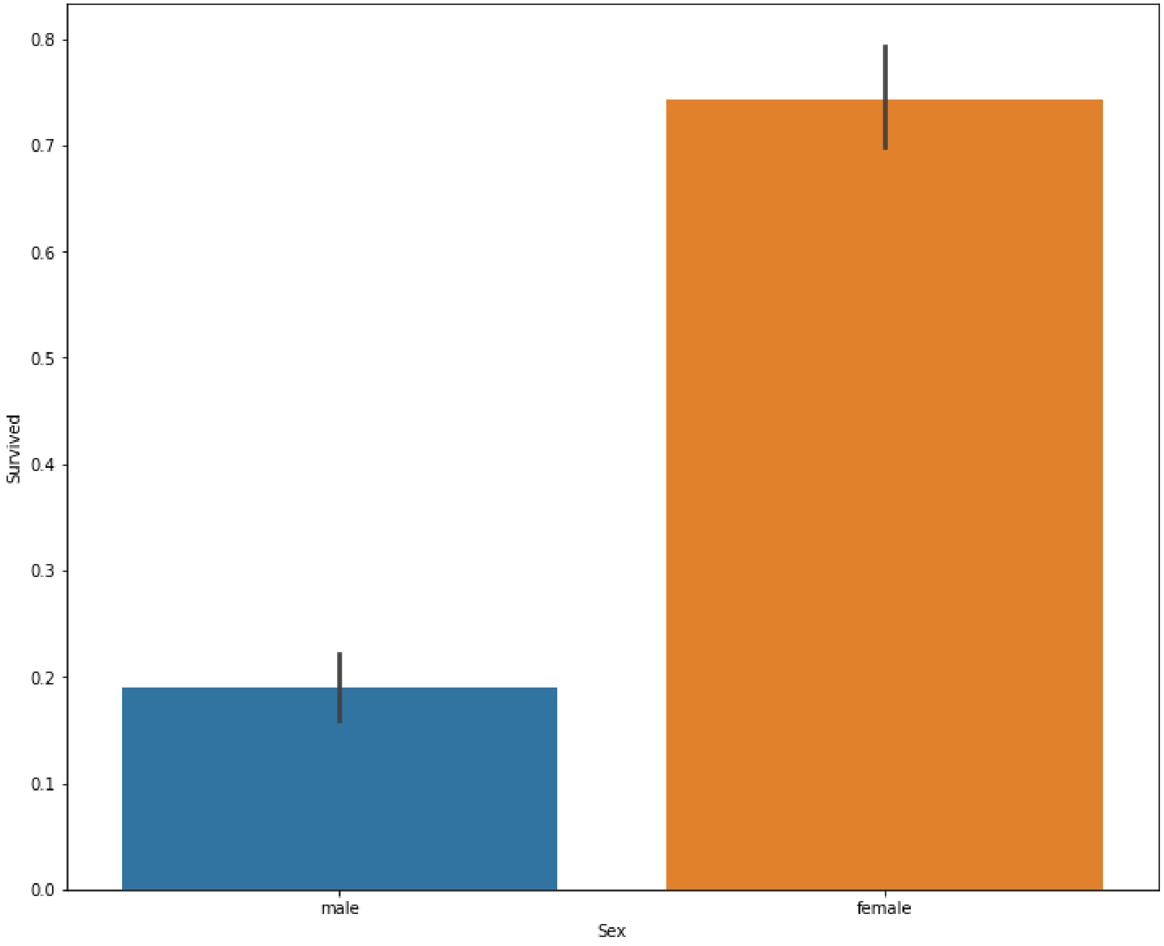


Countplot

Data handling using Matplotlib and Seaborn

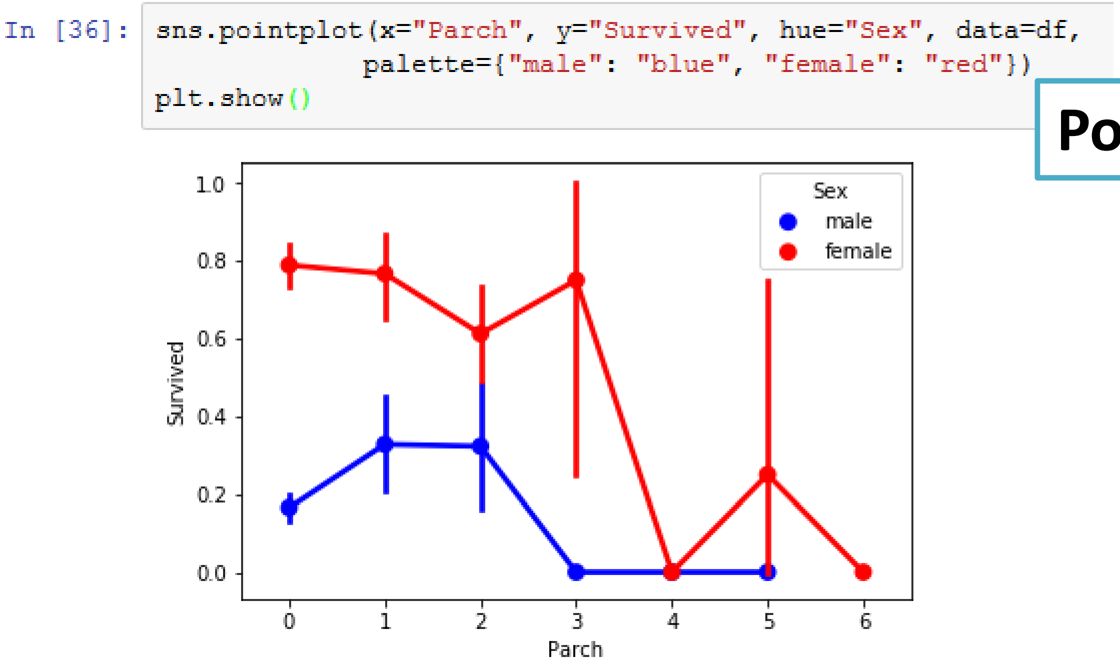


```
In [4]: plt.figure(figsize=[12,10])
        sns.barplot('Sex', 'Survived', data=df)

        # Show the plot
        plt.show()
```



Barchart

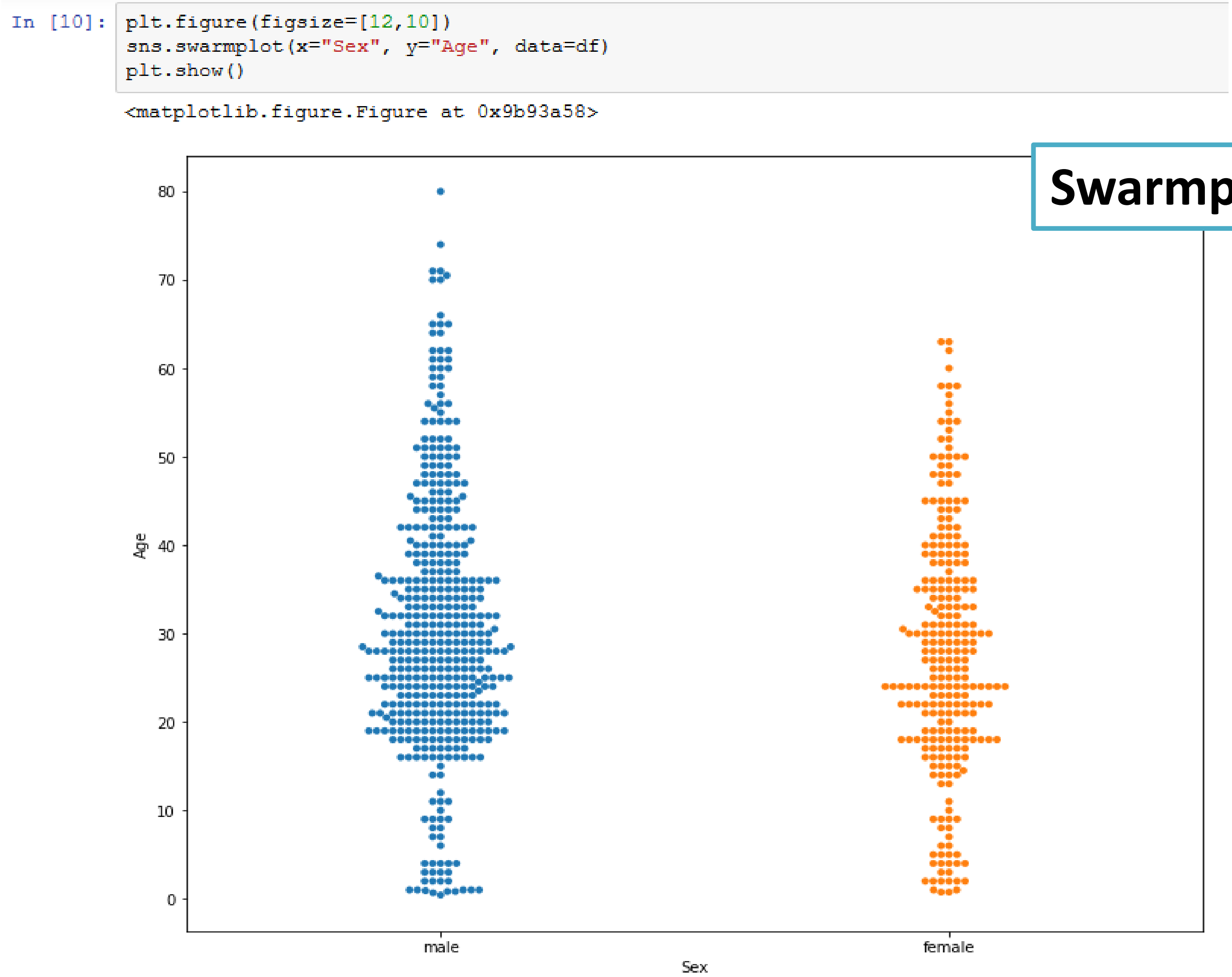
Data handling using Matplotlib and Seaborn



```
In [36]: sns.pointplot(x="Parch", y="Survived", hue="Sex", data=df,
                       palette={"male": "blue", "female": "red"})
         plt.show()
```

Pointplot

Data handling using Matplotlib and Seaborn

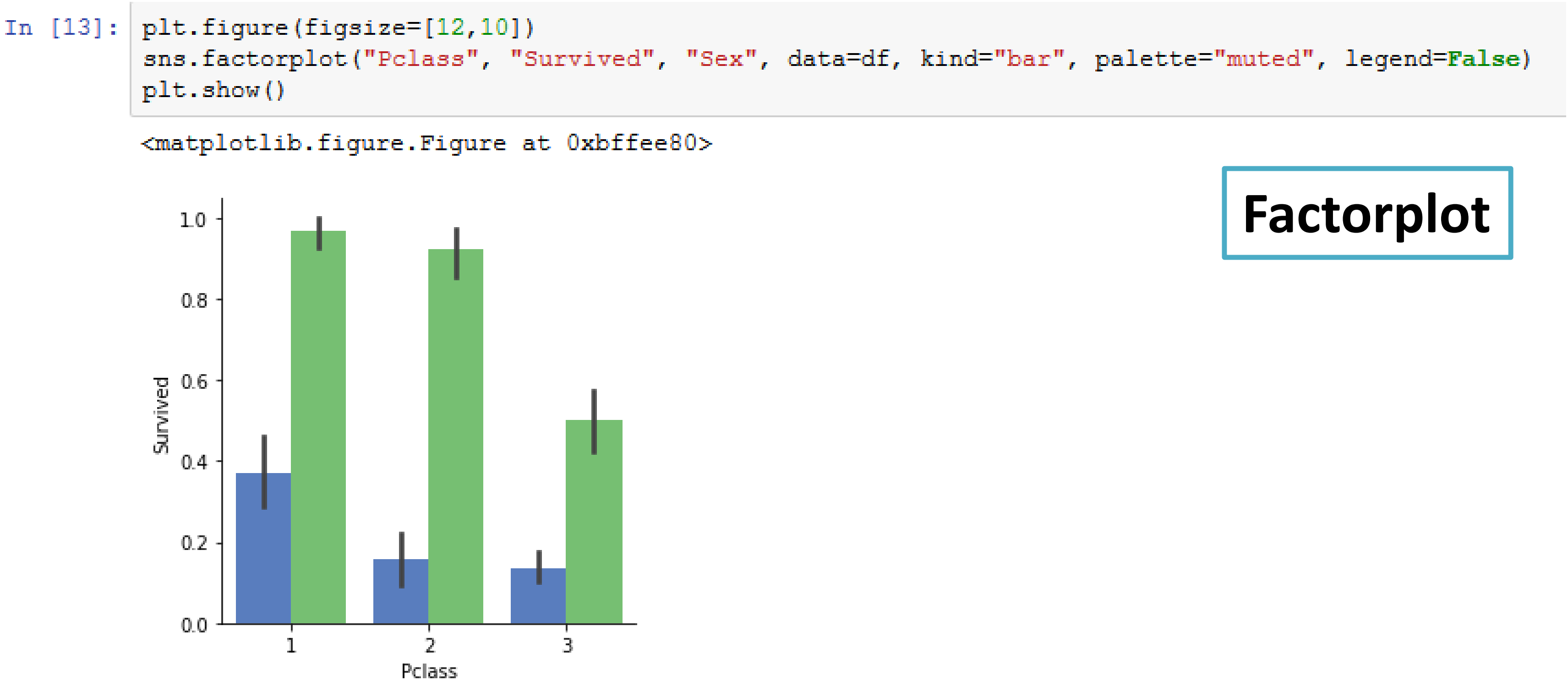


```
In [10]: plt.figure(figsize=[12,10])
         sns.swarmplot(x="Sex", y="Age", data=df)
         plt.show()
```

```
<matplotlib.figure.Figure at 0x9b93a58>
```

Swarmplot

Data handling using Matplotlib and Seaborn



```
In [13]: plt.figure(figsize=[12,10])
         sns.factorplot("Pclass", "Survived", "Sex", data=df, kind="bar", palette="muted", legend=False)
         plt.show()

         <matplotlib.figure.Figure at 0xbffee80>
```

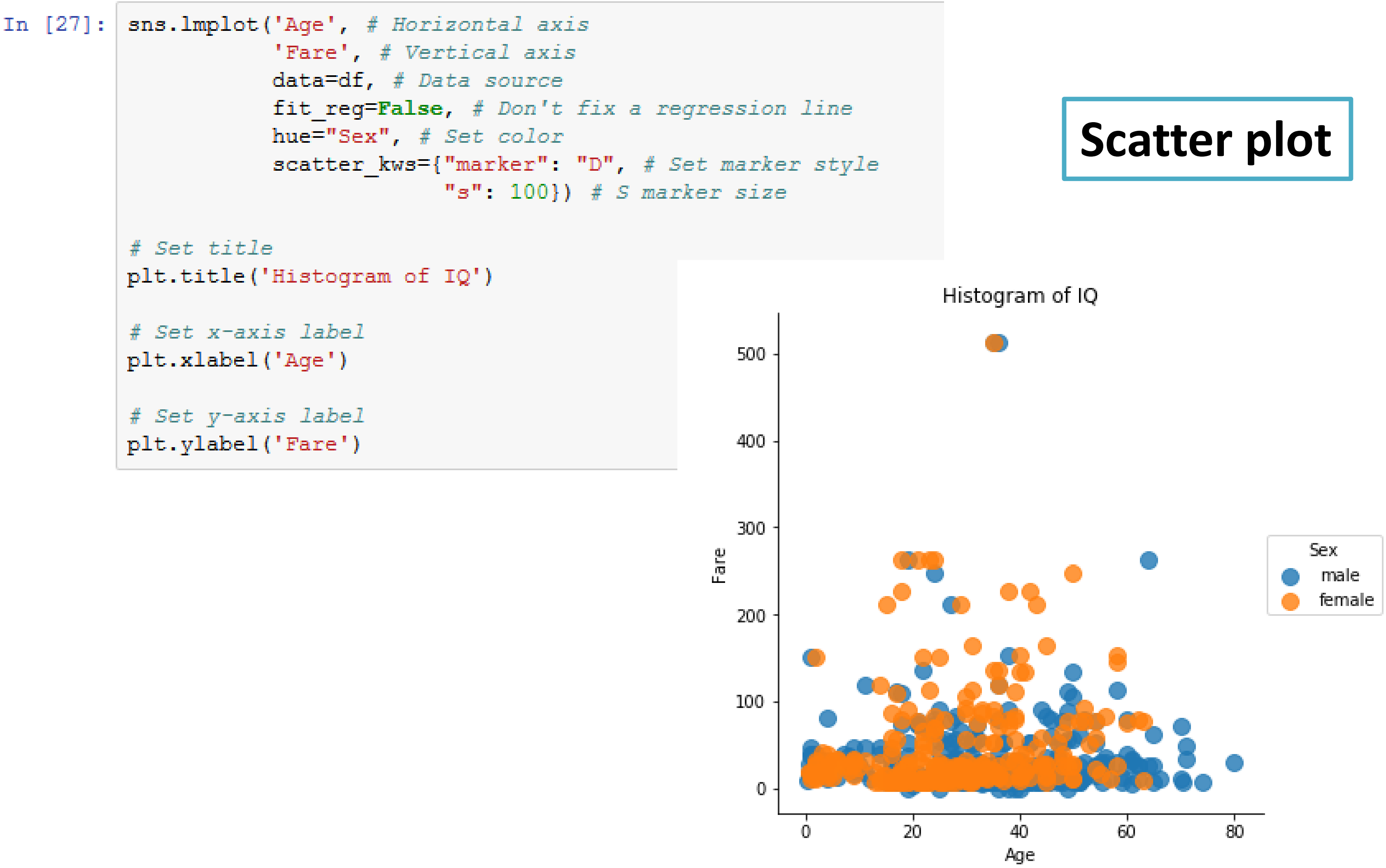
Factorplot

BTS | Barcelona Technology School

eurecat Centre Tecnològic de Catalunya

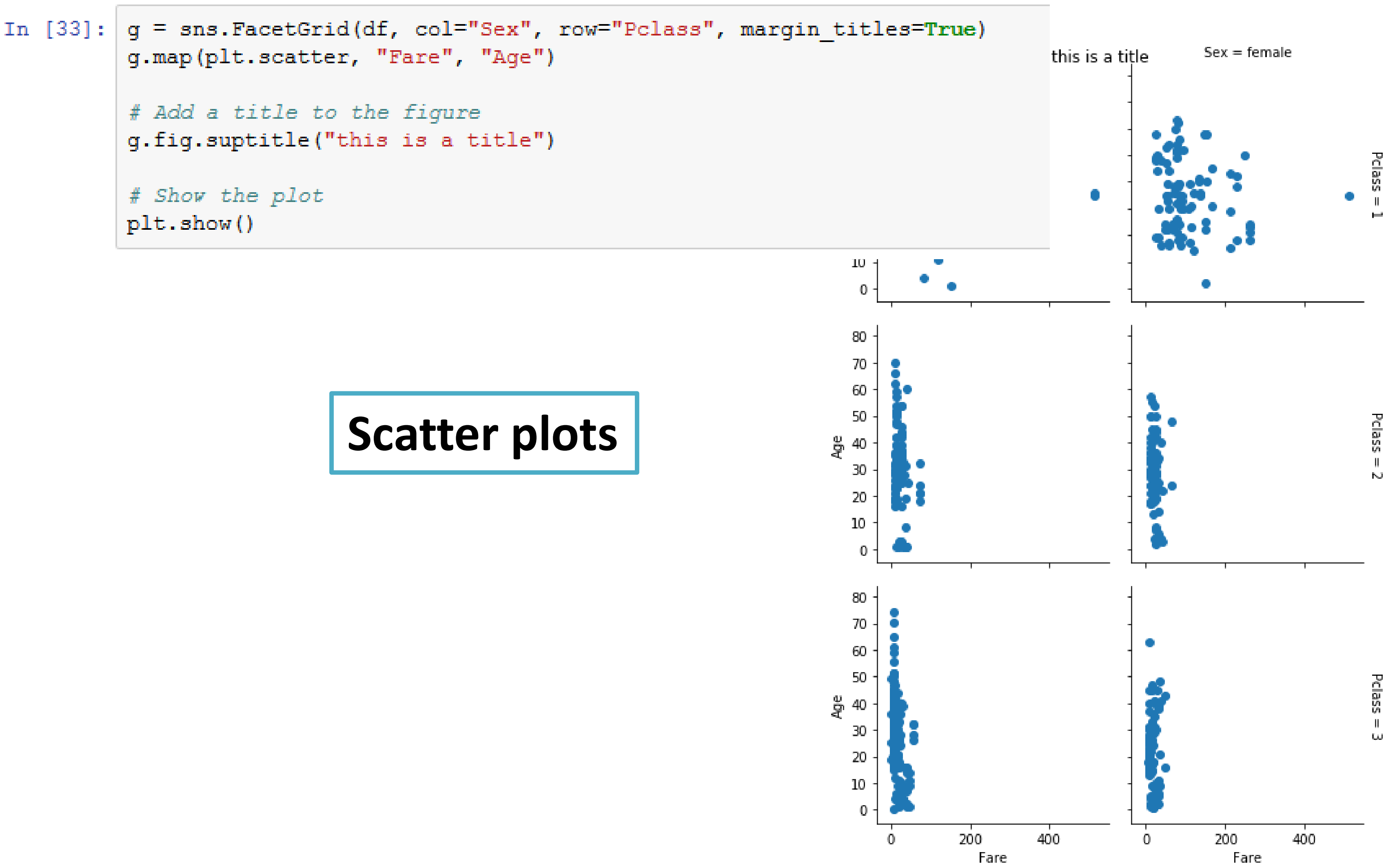
Data handling using Matplotlib and Seaborn

Scatter plot



```
In [27]: sns.lmplot('Age', # Horizontal axis
                    'Fare', # Vertical axis
                    data=df, # Data source
                    fit_reg=False, # Don't fix a regression line
                    hue="Sex", # Set color
                    scatter_kws={"marker": "D", # Set marker style
                                 "s": 100}) # S marker size

         # Set title
         plt.title('Histogram of IQ')

         # Set x-axis label
         plt.xlabel('Age')

         # Set y-axis label
         plt.ylabel('Fare')
```

Data handling using Matplotlib and Seaborn

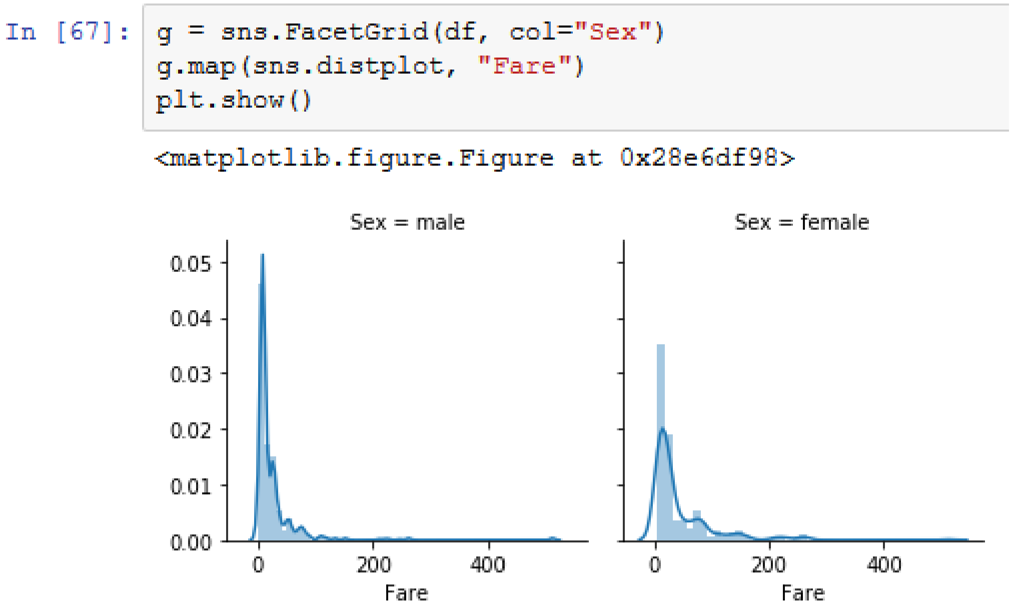


```
In [33]: g = sns.FacetGrid(df, col="Sex", row="Pclass", margin_titles=True)
         g.map(plt.scatter, "Fare", "Age")

         # Add a title to the figure
         g.fig.suptitle("this is a title")

         # Show the plot
         plt.show()
```

Scatter plots

Data handling using Matplotlib and Seaborn

Distplot



```
In [67]: g = sns.FacetGrid(df, col="Sex")
         g.map(sns.distplot, "Fare")
         plt.show()
```

```
<matplotlib.figure.Figure at 0x28e6df98>
```

Data handling using Matplotlib and Seaborn

- Create 10 charts for the dataset “rural_population_data” (Session_1/file_formats).

- Create 5 charts for the dataset “2_titanic_data” (Session_2).

- After each chart, please put a comment starting with “Knowledge retrieved:” and summarize your finding.

- Push your results to the repository “DataScienceFoundations” into the folder “Session_2”.

- You will need to share your findings with classmates (5-10 minutes per person).

- **Please notice that your charts should be informative and should bring useful insight!**

Individual Assignment

- Take a dataset “3_bike_sharing.csv”
- Perform a basic data analysis using Pandas
- Perform visual analysis of data using Matplotlib and Seaborn
- Do not forget to comment your code and specify questions/findings.

BARCELONA
Barcelona Technology School