SPAI

An AI Singapore Student Chapter

ML Bootcamp

Day 1

Scan to mark attendance

Scan the QR code to mark your attendance

Attendance

Learning Objectives

☑ Overview of Bootcamp

☑ Introduction to Data Science & Machine Learning

☑ What is Exploratory Data Analysis

☑ A brief Primer on Statistical Concepts

☑ Python for EDA

You should know:

- ☑ Basic Python & Google Colab
- ☑ Today’s content sent to you via Teams Chat
- ☑ 12pm to 1pm: Lunch break
- ☑ Occasional practical sessions daily in addition to quizzes
- ☑ Q&A sessions conducted during breaks and at the end of each day

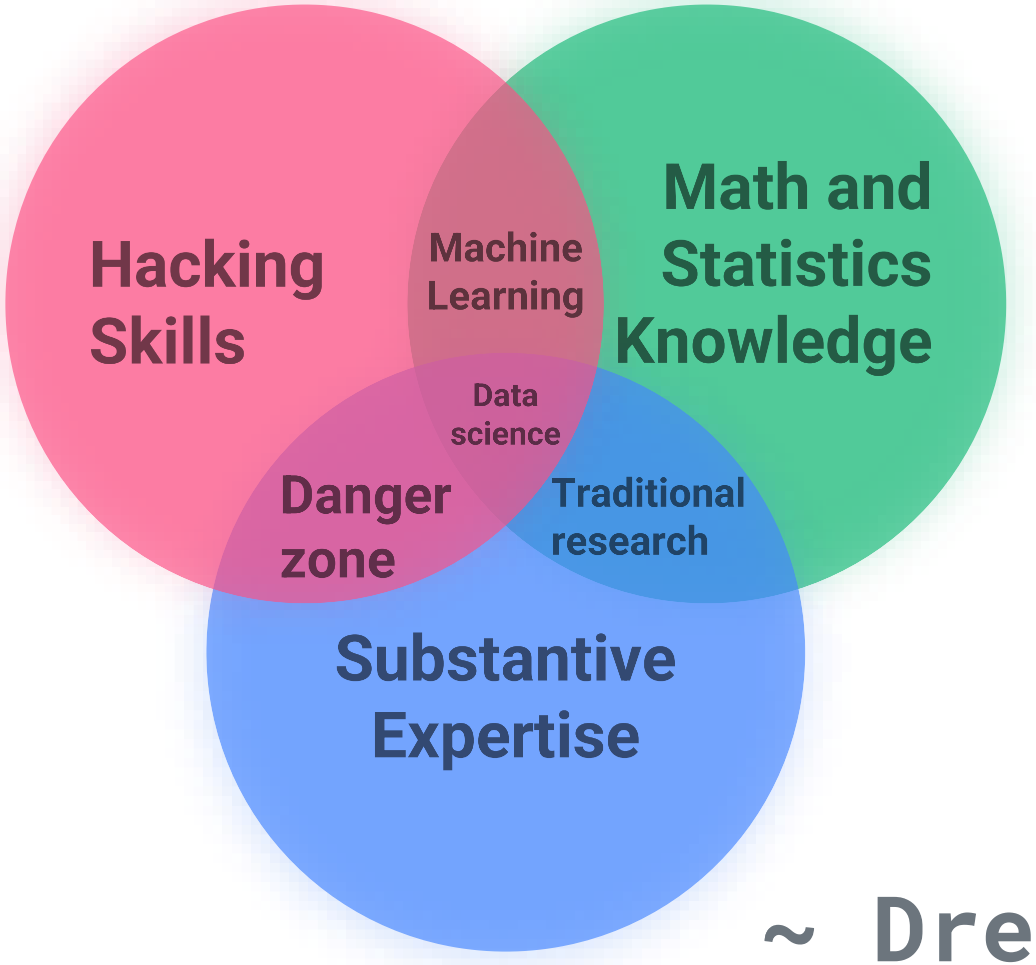
1

Introduction to Data Science & Machine Learning

Data Science lies at the intersection of Computer Science, Statistics, and Substantive Application Domains.

What is Data Science

What is Data Science?



~ Drew Conway

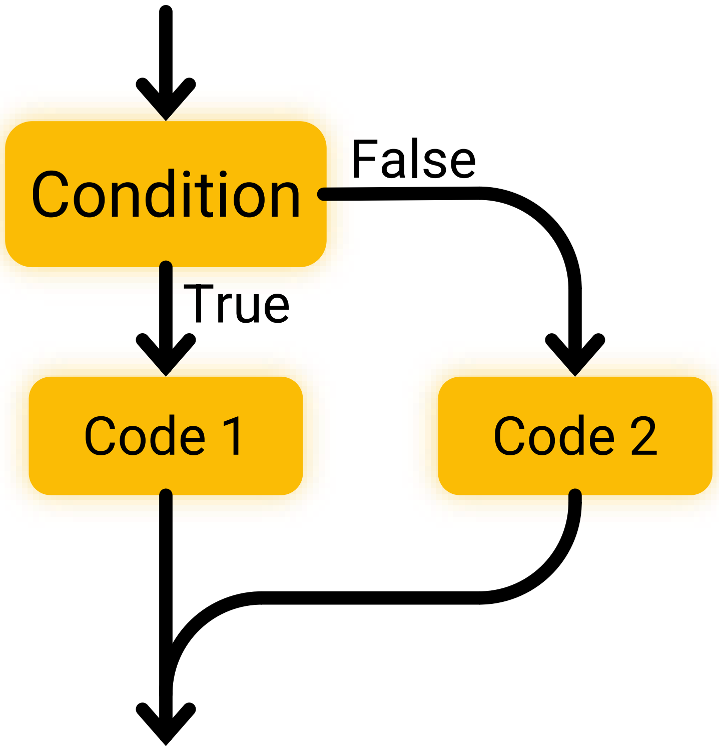
SPAI

> Machine Learning is a field of study that gives computers the ability to learn without being explicitly programmed.
>
> ~ Arthur Samuel(1959)

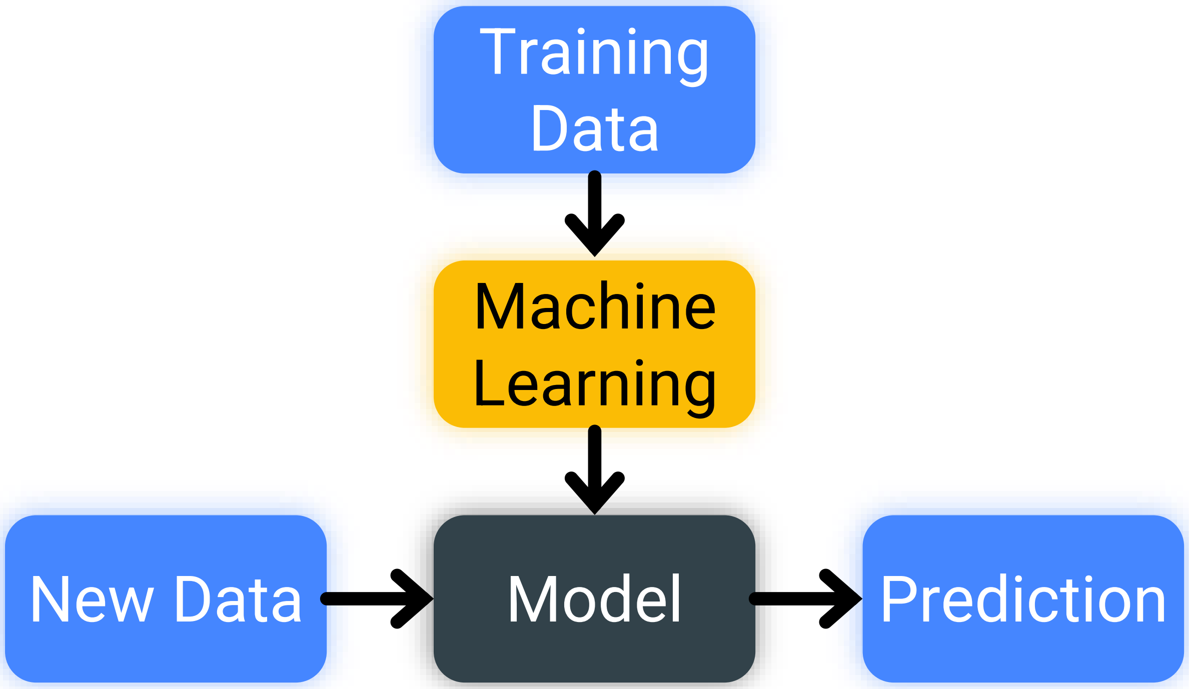
What is Machine Learning

Machine Learning

Rule-based approach



Machine Learning

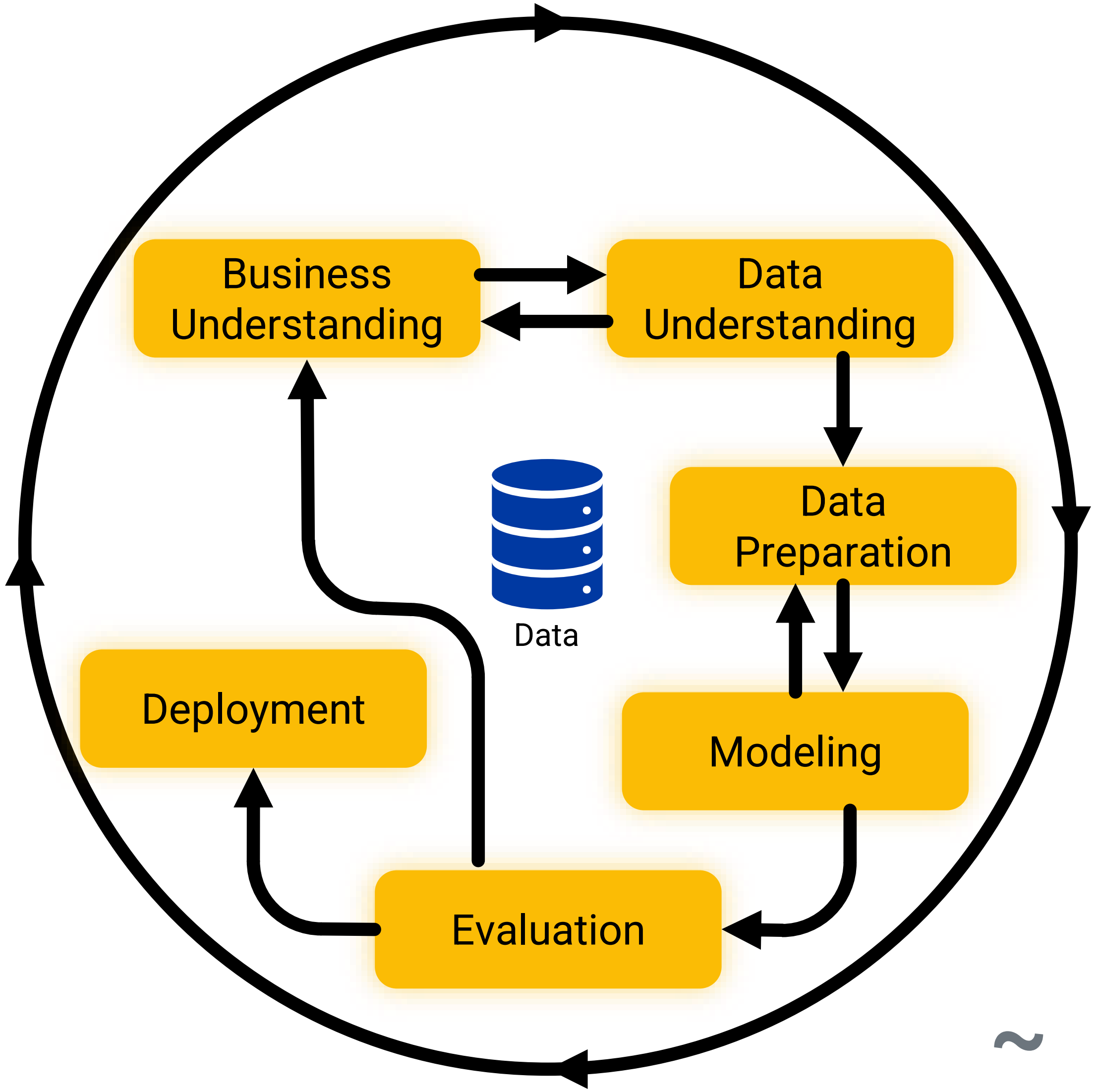


Source: EPI-USE Enterprise Machine Learning

Why ML Now?

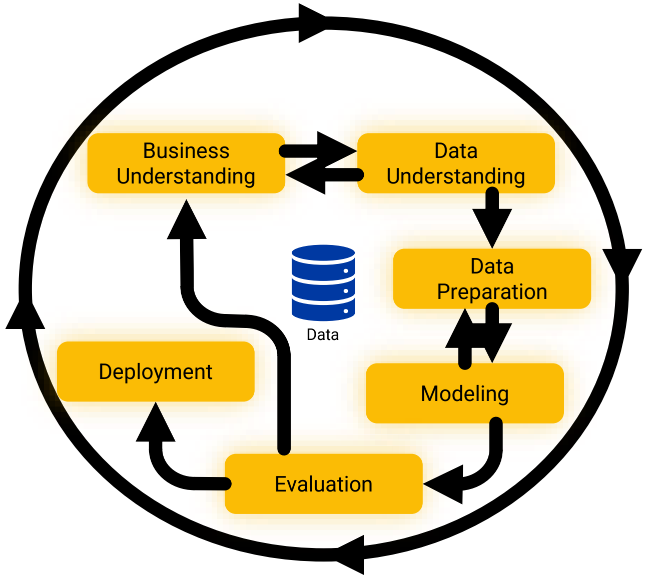
- ☑ **Data** explosion allows us to better train models
- ☑ **Low-Cost Computing** is more accesible
- ☑ **Bigger and Faster Algorithms** allow faster training and more robust models.

Data Science Workflow



~ CRISP-DM

SPAI

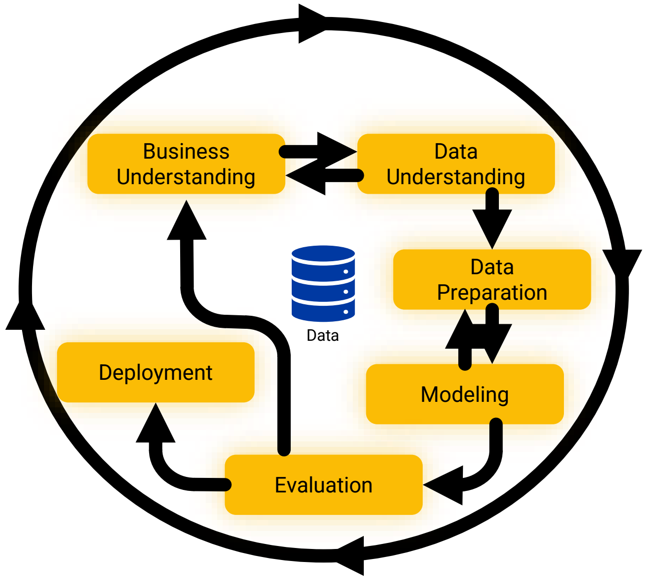


Business Understanding

Actual problem wanted to be solved

Definition of a successful project

Source of data



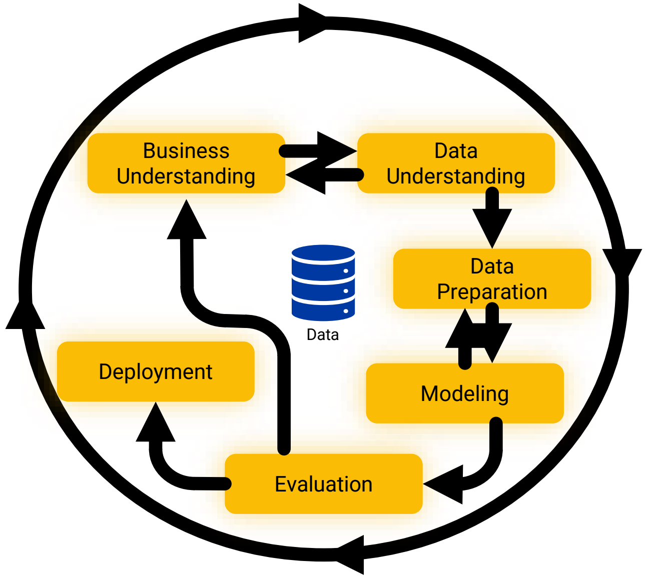
Data Understanding

Acquire data

Explore data

- ? Is the data of good quality
- ? Is the data able to help you achieve your goal

Make use of descriptive statistics and data visualization

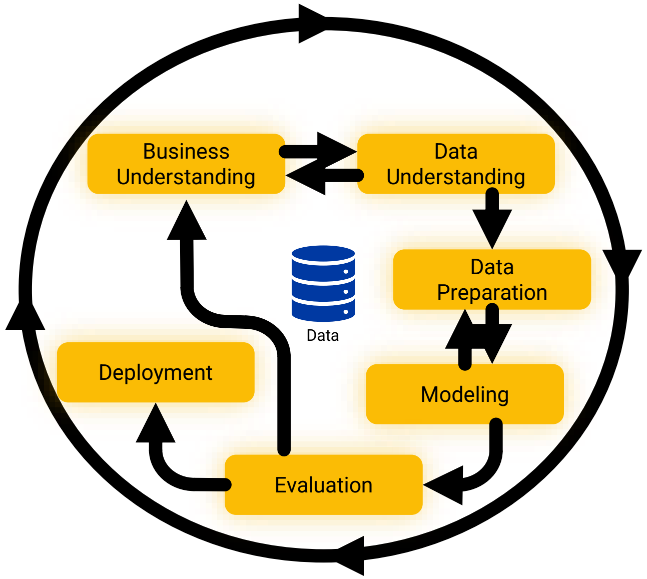


Data Preparation

“Garbage In, Garbage Out”

Clean up data

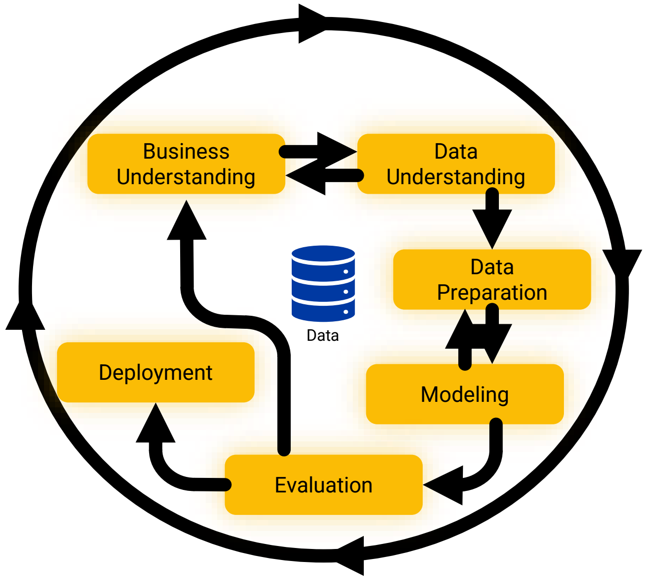
Model makes use of full possible information from data.



Modelling

Models are systems created to do tasks.

Data is fed from ML models to infer relationships in data and make predictions



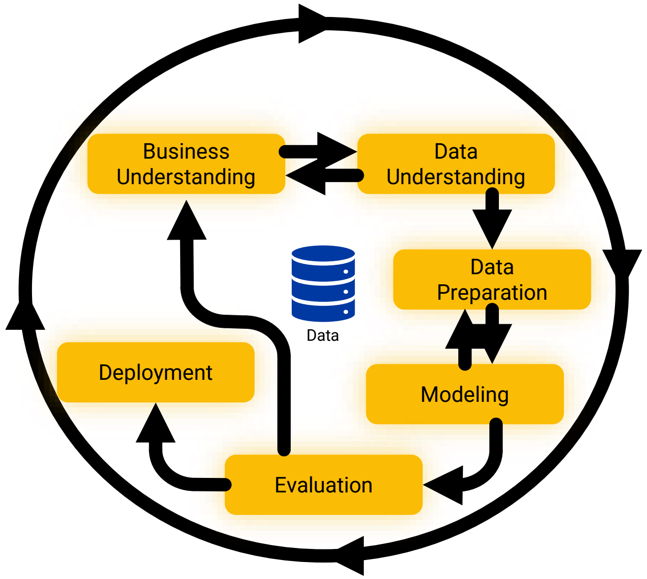
Evaluation

Not all models are good

Ensure model succeeds in solving task

Ensure model is not memorizing answers

If model performs poorly, repeat previous stages.



Deployment

When model is good enough

Operationalize model

Python for Data Science

☑ Currently most popular coding language for Data Science

☑ Many toolboxes/libraries for Python.

- NumPy – Multidimensional arrays, Mathematical functions
- SciPy – Scientific computing
- Pandas – Data manipulation and Analysis
- Sci-kit Learn – Machine Learning

☑ Visualization libraries

- Matplotlib
- Seaborn

2

Exploratory Data Analysis

Approach to analyze dataset to summarize their main characteristics using Graphical & Non-Graphical Methods

What is EDA

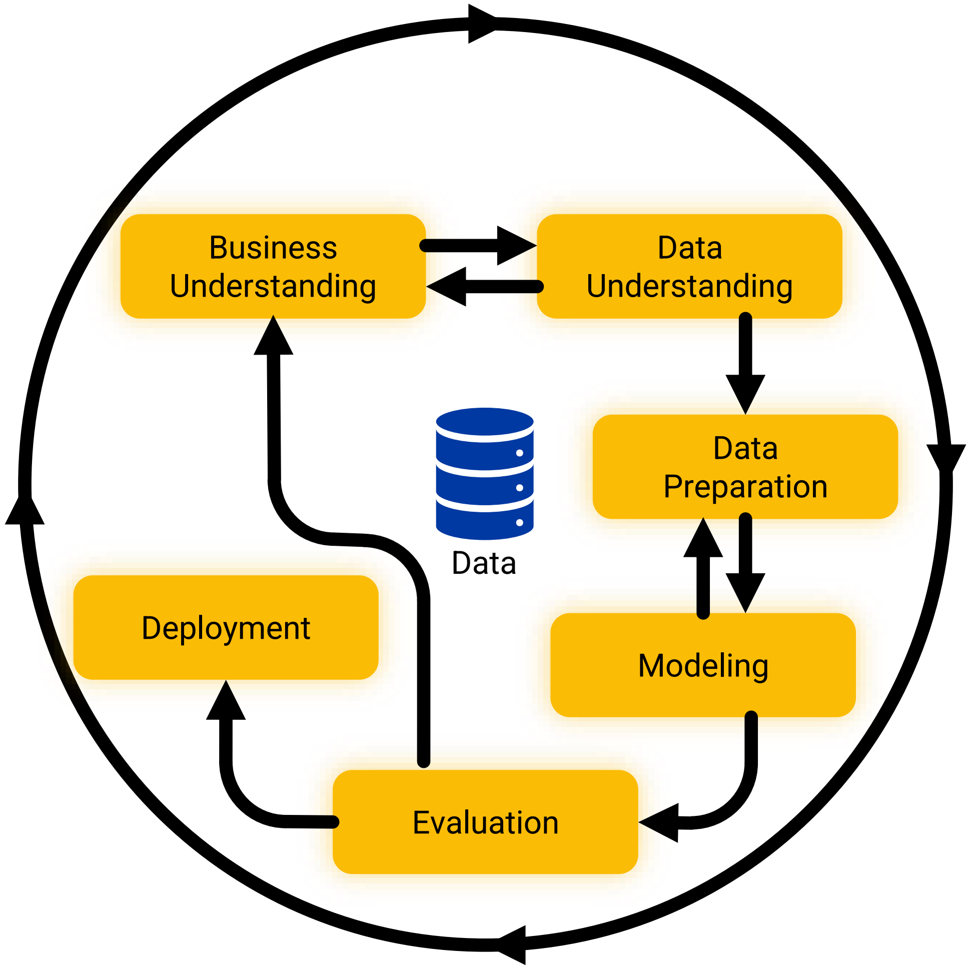
Goals For EDA

- ☑ Check for mistakes in the data. (E.g. Missing Values, Outliers,)
- ☑ Understand the context of the datasets. (E.g. Meaning of Each Columns)
- ☑ Identify trends and relationship in the dataset for feature engineering(creating more feature) or feature elimination

Knowledge Check

Which part of the Data Science Workflow would today's topic come under?

A. Business Understanding
B. Data Understanding
C. Modelling
D. Data Preparation



Approach to EDA

- ☑ Load in data
- ☑ Describe data
- ☑ Check for errors in data
- ☑ Analyze each feature with graphical & non-graphical methods
- ☑ Analyze relationships between features.

Python Libraries (EDA)

- **Pandas** – Data Manipulation and Analysis
- **Numpy** – Multidimensional Array, Mathematical Functions
- **Matplotlib** – Data Visualization Libraries
- **Seaborn** – Data Visualization Libraries

Libraries Installations

☑ Install via Terminal

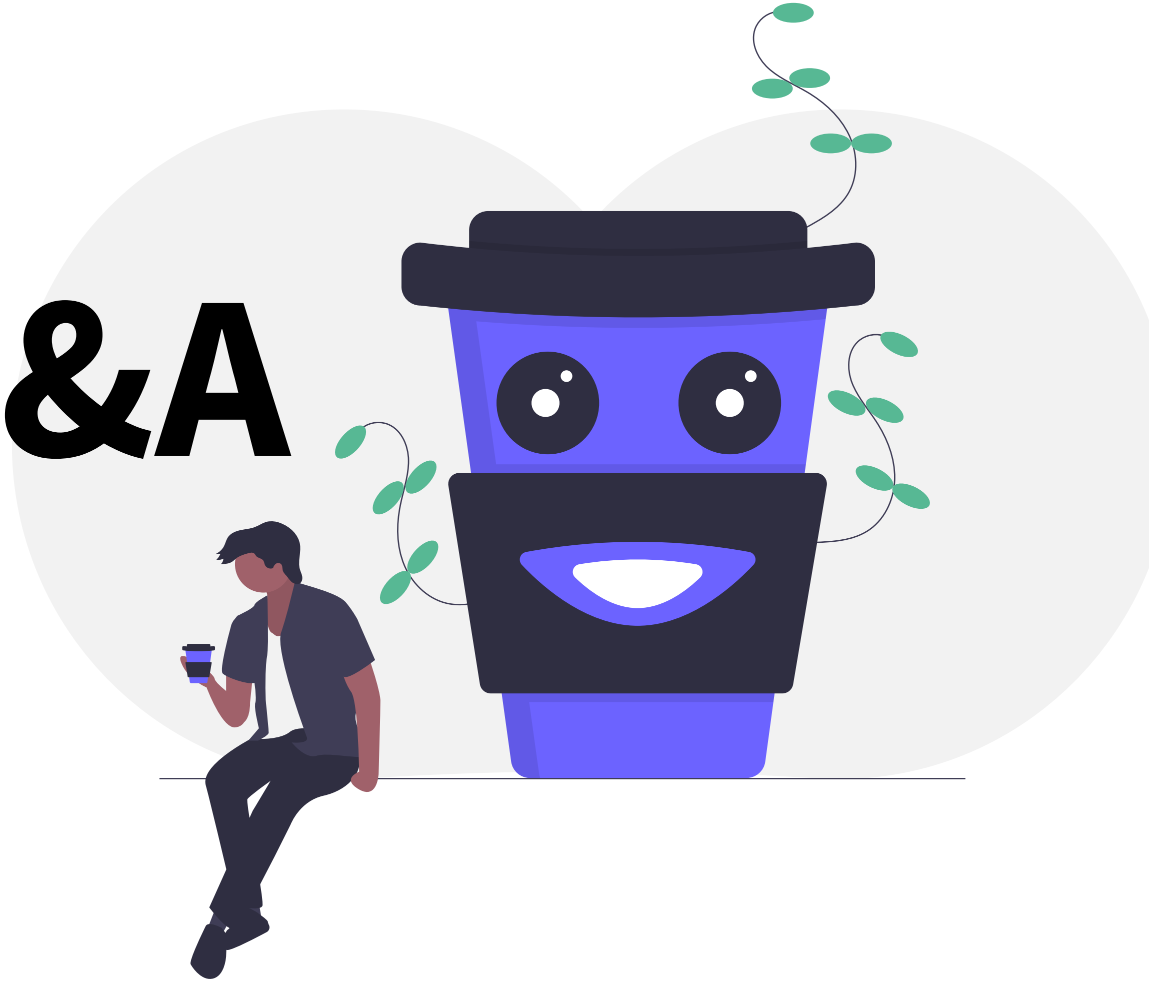
```
pip install pandas numpy matplotlib #for Windows

pip3 install pandas numpy matplotlib #for MacOS
```

☑ Import package

```
import pandas as pd #pandas
import numpy as np #numpy
import matplotlib.pyplot as plt #matplotlib
```

Break and Q&A

10 Minutes



3.1

Pandas

Introduction

Pandas is a Python package providing fast and flexible data structures.

What is Pandas

SPAI

Pandas

Provides expressive data structures

Makes working with relational or labeled data easy and intuitive

Fundamental high-level building block for practical and real-world data analysis

3.2

Pandas

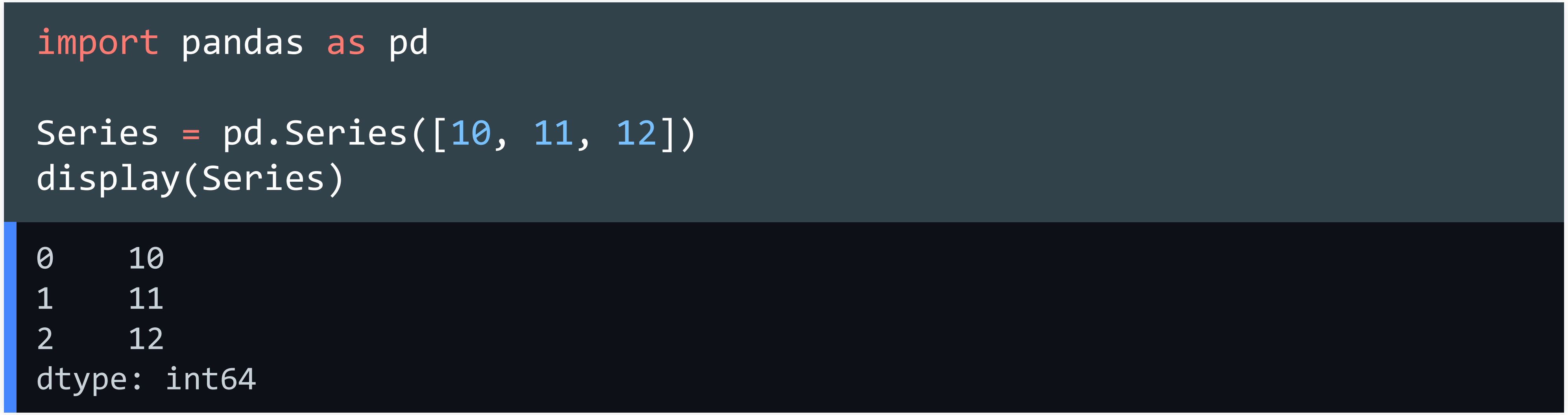
DataFrame & Series

Data Structures

✓ Series – 1D Array

✓ DataFrame – Table like 2D Array

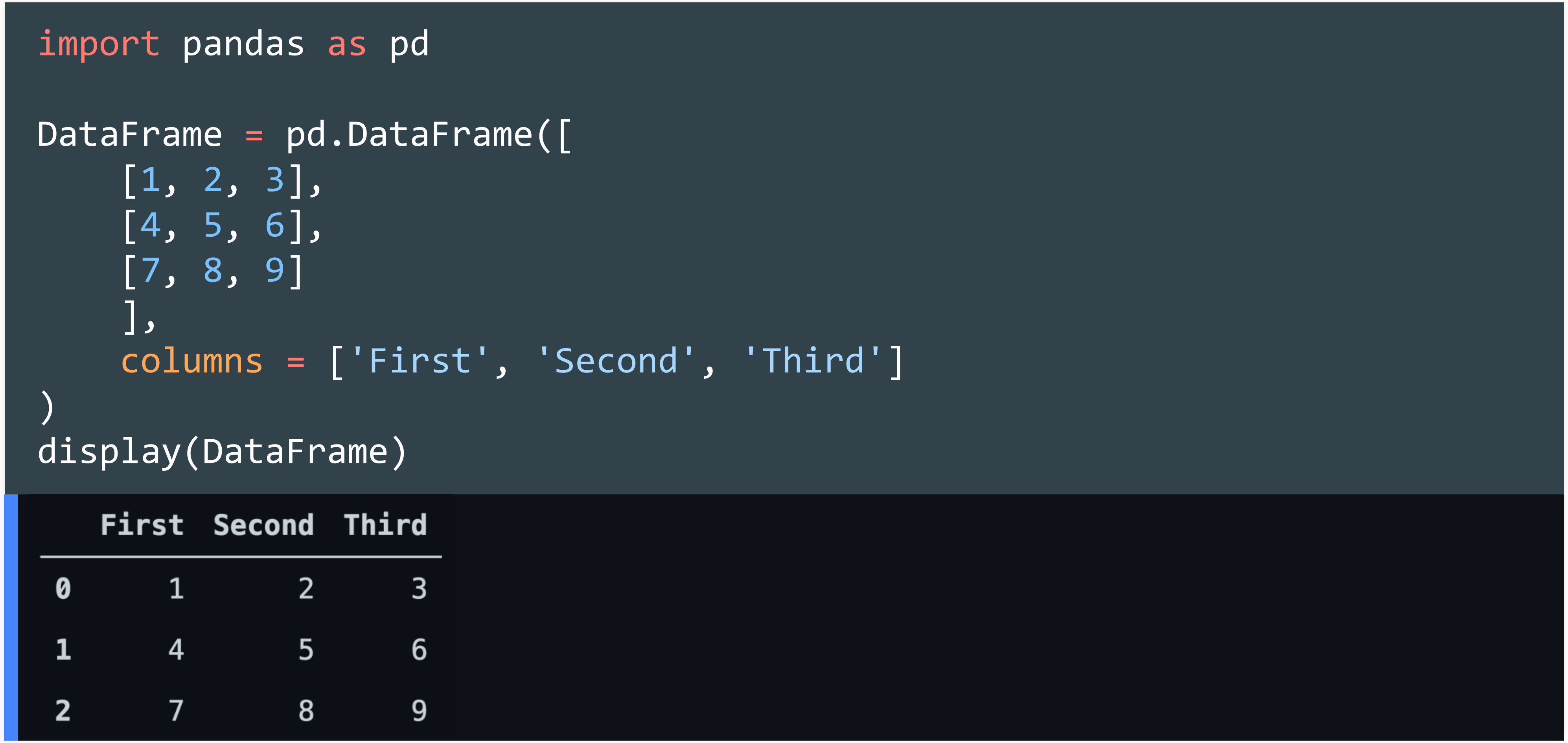
Series – 1D Array



```python
import pandas as pd

Series = pd.Series([10, 11, 12])
display(Series)
```

```
0    10
1    11
2    12
dtype: int64
```

DataFrame – 2D Array



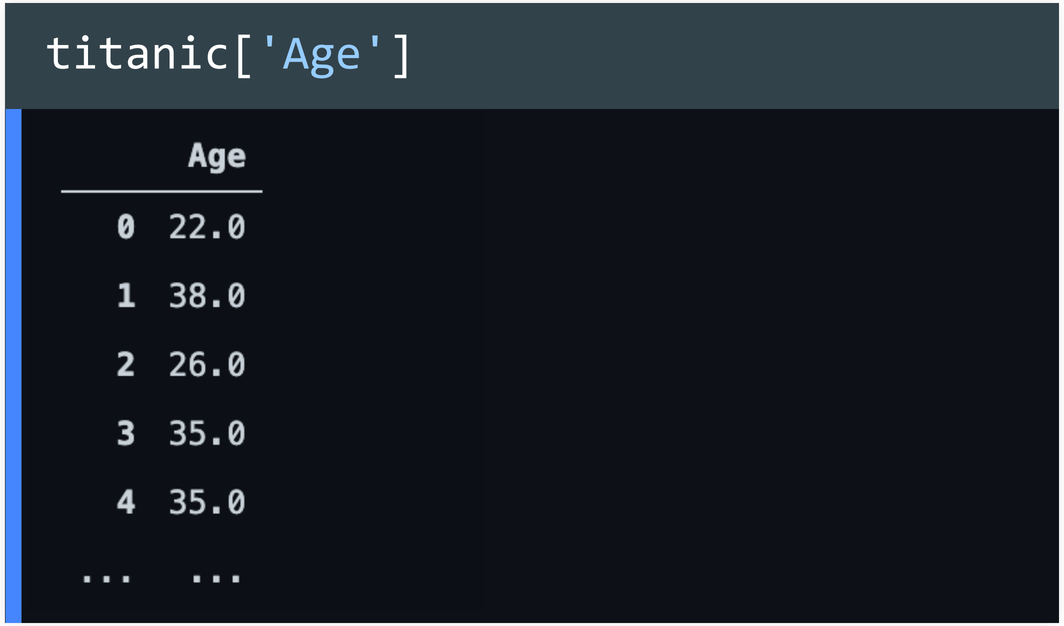
```
import pandas as pd

DataFrame = pd.DataFrame([
    [1, 2, 3],
    [4, 5, 6],
    [7, 8, 9]
    ],
    columns = ['First', 'Second', 'Third']
)
display(DataFrame)
```

```
   First  Second  Third
0      1       2      3
1      4       5      6
2      7       8      9
```

Column Slicing

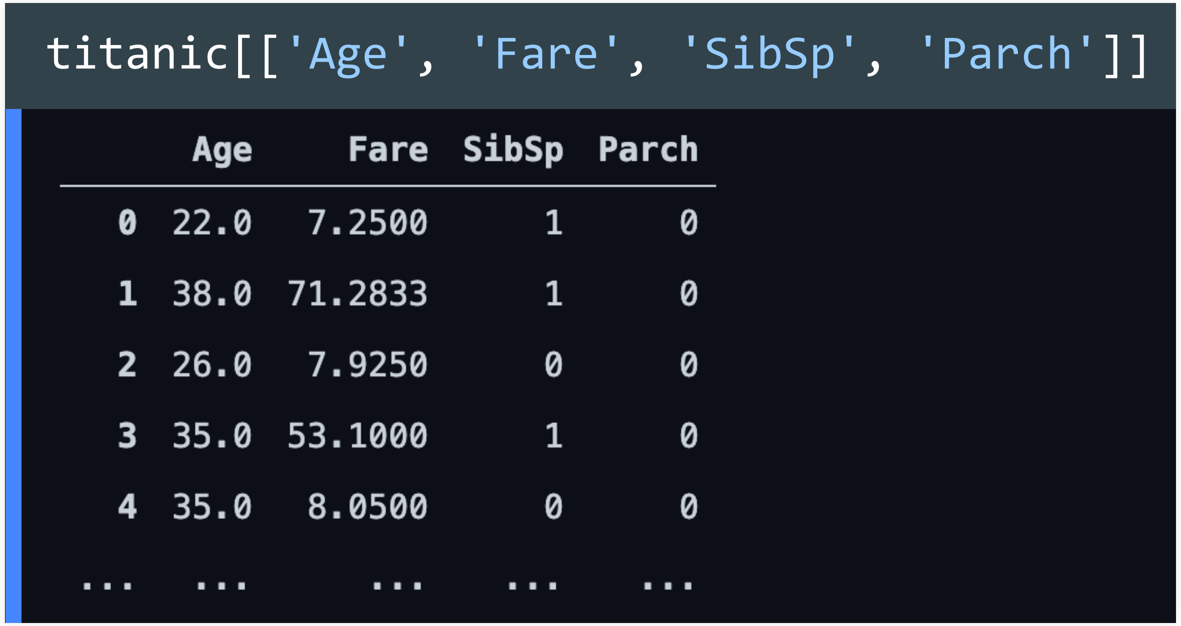
☑ Select one column

☑ Select multiple columns



```
titanic['Age']
```

```
      Age
0    22.0
1    38.0
2    26.0
3    35.0
4    35.0
...   ...
```



```
titanic[['Age', 'Fare', 'SibSp', 'Parch']]
```

```
      Age     Fare  SibSp  Parch
0    22.0   7.2500      1      0
1    38.0  71.2833      1      0
2    26.0   7.9250      0      0
3    35.0  53.1000      1      0
4    35.0   8.0500      0      0
...   ...      ...    ...    ...
```

Approach to EDA

- ☑ **Load in data**
- ☑ Describe data
- ☑ Check for errors in data
- ☑ Analyze each variable with graphical & non-graphical methods
- ☑ Analyze relationships between variables.

Loading Data

- ☑ Ensure that data file is in same folder as python file

- ☑ Run the following code

```
titanic = pd.read_csv('titanic.csv')
```

Approach to EDA

☑ Load in data

☑ **Describe data**

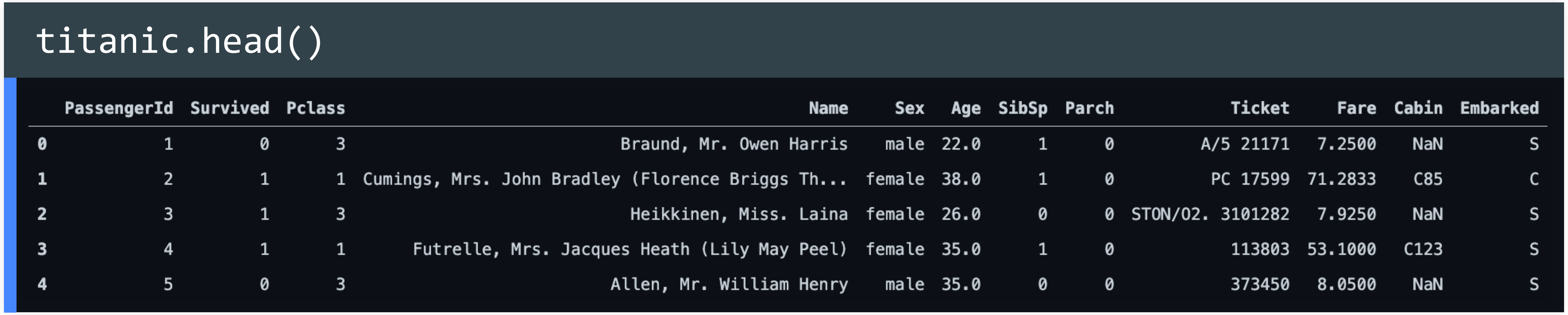
☑ Check for errors in data

☑ Analyze each variable with graphical & non-graphical methods

☑ Analyze relationships between variables.

Describing Data

☑ Exploring data through top few rows

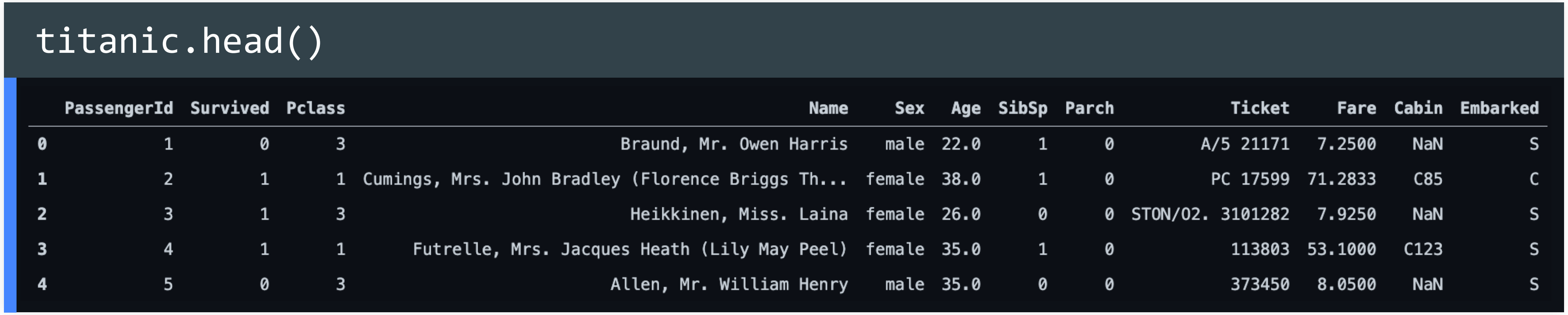


```
titanic.head()
```

```
   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
```

Describing Data

☑ Exploring data through top few rows

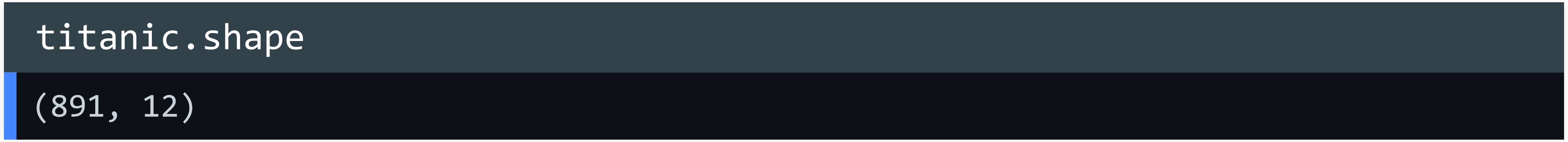


```
titanic.head()
```

```
   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
```

Describing Data

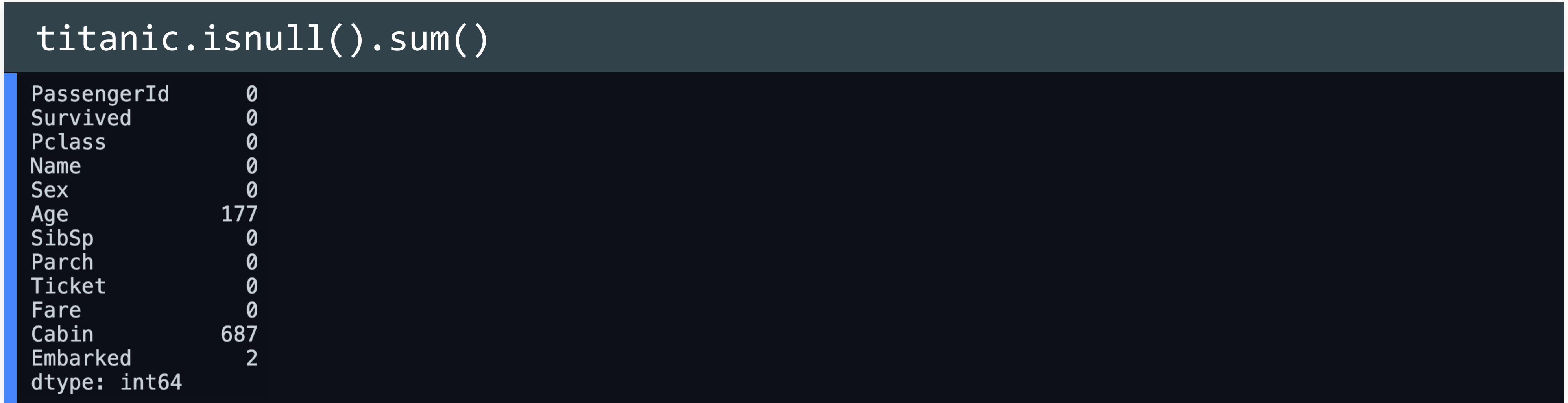
☑ Show total rows and columns.



```
titanic.shape
```

```
(891, 12)
```

☑ Check total ‘NaN’ values.

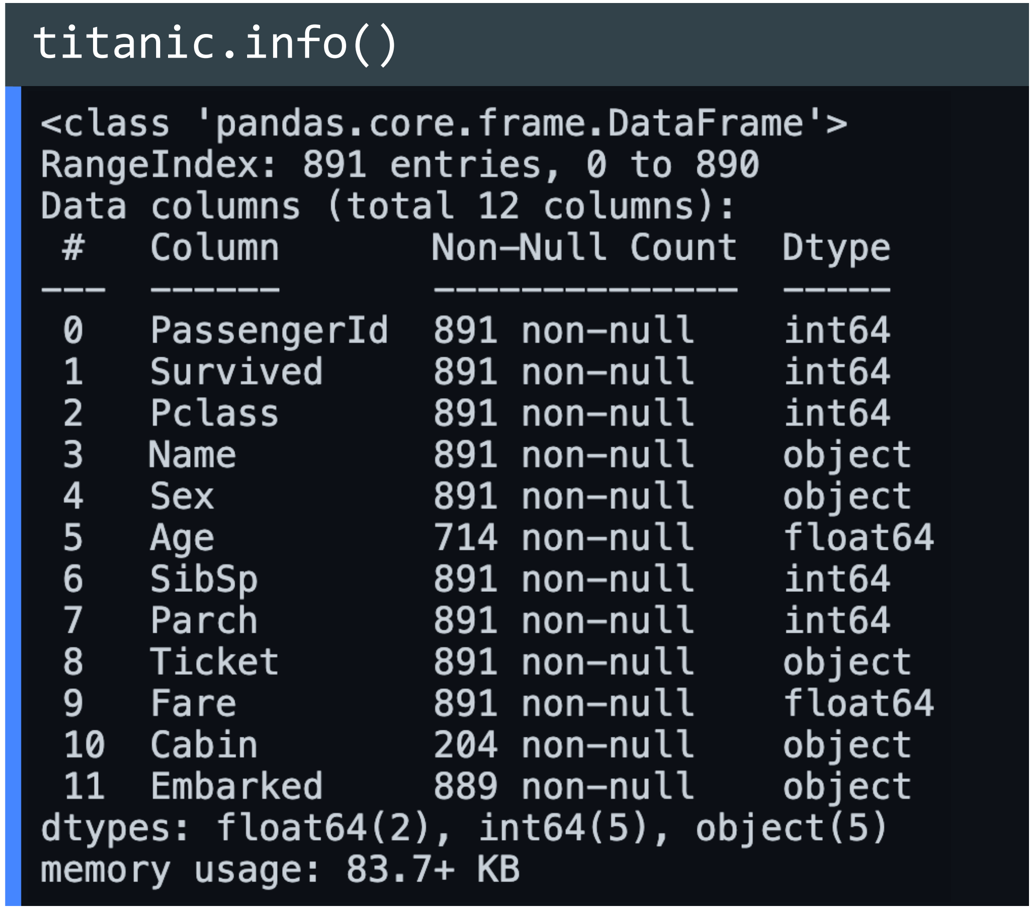


```
titanic.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Describing Data

Show quick summary



```
titanic.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

Dtype

```
object  #String values (‘SPAI’)
float64 #Decimal values (0.01)
int64   #Integer values (1)
```

Lunch Time

1 Hour

Approach to EDA

☑ Load in data

☑ Describe data

☑ **Check for errors in data**

☑ Analyze each variable with graphical & non-graphical methods

☑ Analyze relationships between variables.

Checking Data Types

⚠ Every data in a column must have the same data type

⚠ If a column contains multiple data types, pandas will assign a data type to accommodate all data types

⚠ This might cause issues in later processes

Checking Missing Values

- ⚠ Data can be missing in our dataset for various reasons.
- ⚠ Identify number of missing values in our dataset using isnull().sum()
- ⚠ Missing Values should be discarded or imputed before feeding into ML models.

Approach to EDA

- ☑ Load in data
- ☑ Describe data
- ☑ Check for errors in data
- ☑ **Analyze each variable with graphical & non-graphical methods**
- ☑ Analyze relationships between variables.

4

Types of Features

Types of Features

Numerical Feature

Categorical Feature

Numerical Feature

Represented with numbers

Continuous – Measured data
(e.g. Temperature, Height)

Discrete – Counted data
(e.g. Number of Rooms, Number of People)

Categorical Feature

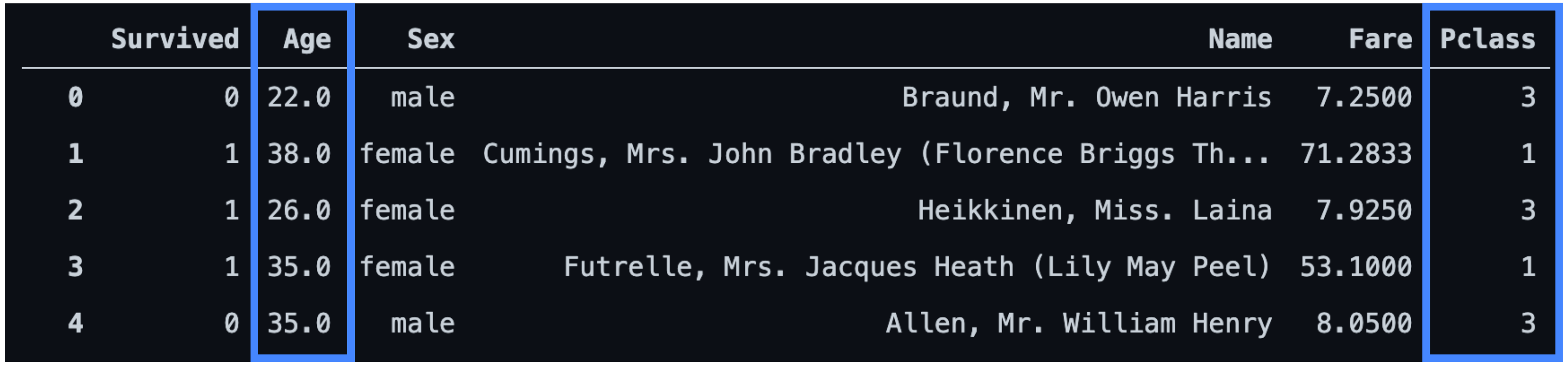
Descriptive by nature

Ordinal – Ordered/Ranked data
(e.g. Grades: A,B,C,D….)

Nominal – Unordered/Unranked data
(e.g. Color: Blue, Red, Black)

Knowledge Check

What type of data would “Age” and “Pclass” be classified in as?



```
   Survived   Age     Sex                                               Name     Fare  Pclass
0         0  22.0    male                            Braund, Mr. Owen Harris   7.2500       3
1         1  38.0  female  Cumings, Mrs. John Bradley (Florence Briggs Th...  71.2833       1
2         1  26.0  female                             Heikkinen, Miss. Laina   7.9250       3
3         1  35.0  female       Futrelle, Mrs. Jacques Heath (Lily May Peel)  53.1000       1
4         0  35.0    male                           Allen, Mr. William Henry   8.0500       3
```

5

Statistical Primer

Measures of Central Tenancy

☑ Single number that best represents data

☑ Mean – Average numerical value

Useful when data has no outliers

☑ Median – Middle Value in sorted data

Better represents data with outliers

☑ Mode – Most common value

Useful when dealing with categorical data

Measures of Spread

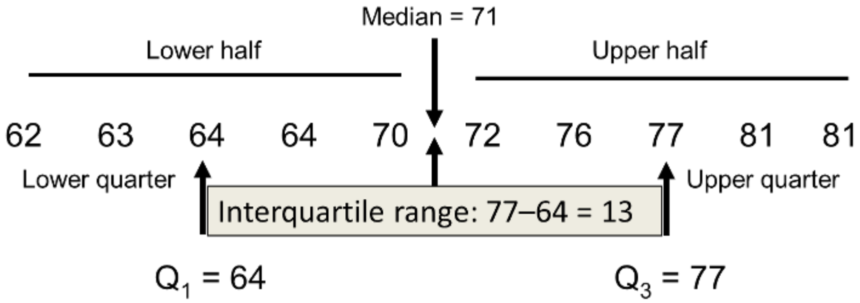
☑ Describes distance between values and the center

☑ Standard Deviation – average variation of data values from mean

- Lower standard deviation will result in values more “bunched together”
- Can be affected by outliers

☑ Interquartile Range – difference between two values

- Useful when data has outliers



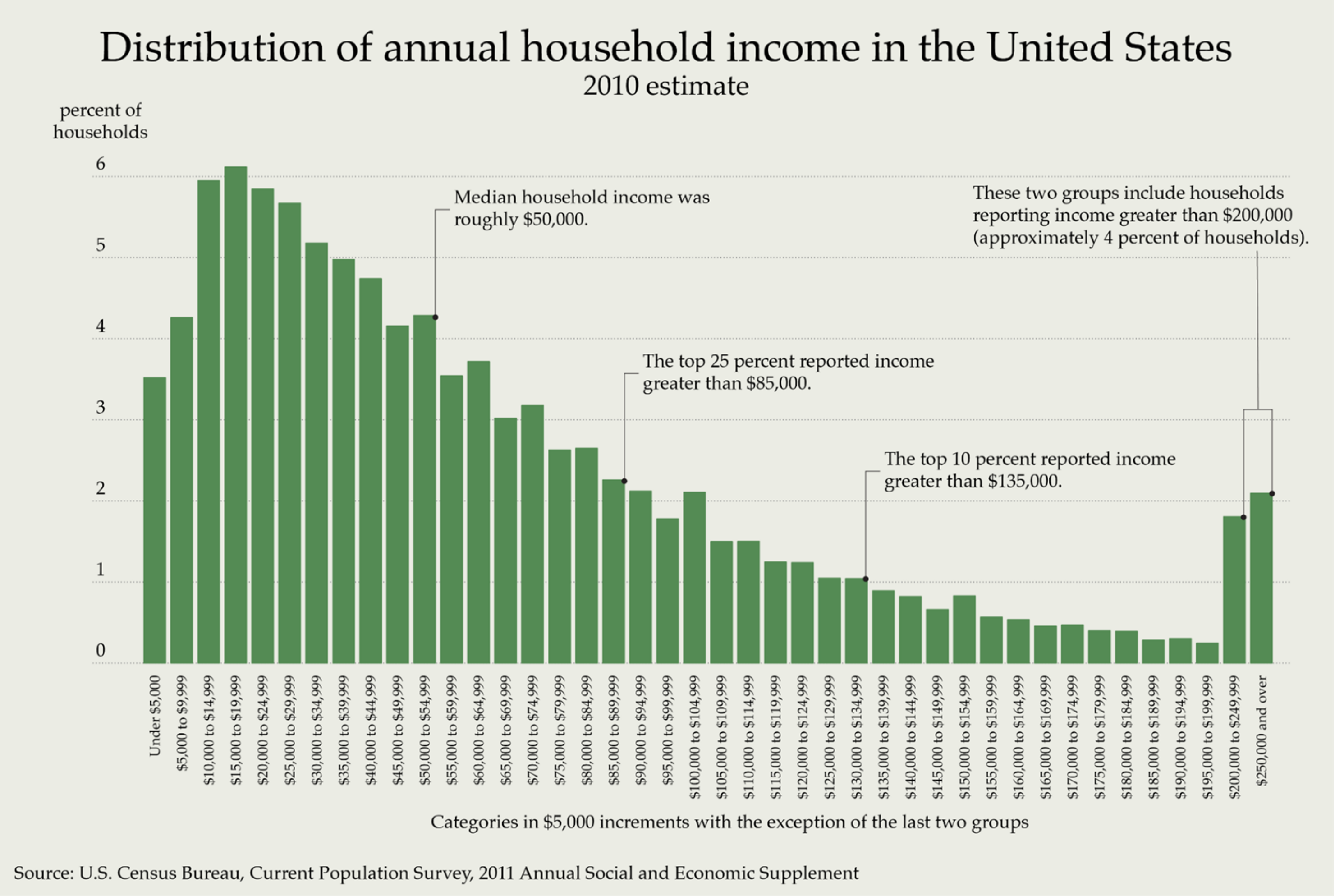
Outliers

- ☑ Abnormal numerical data with extreme values.
- ☑ Some Machine learning algorithms might be sensitive to them
- ☑ Therefore, handling outliers is important

Knowledge Check

Which statistical measure are more useful in describing this distribution?

A. Mean, Standard Deviation
B. Median, Standard Deviation
C. Mean, IQR
D. Median, IQR



Distribution of annual household income in the United States

2010 estimate

Median household income was roughly $50,000.

These two groups include households reporting income greater than $200,000 (approximately 4 percent of households).

The top 25 percent reported income greater than $85,000.

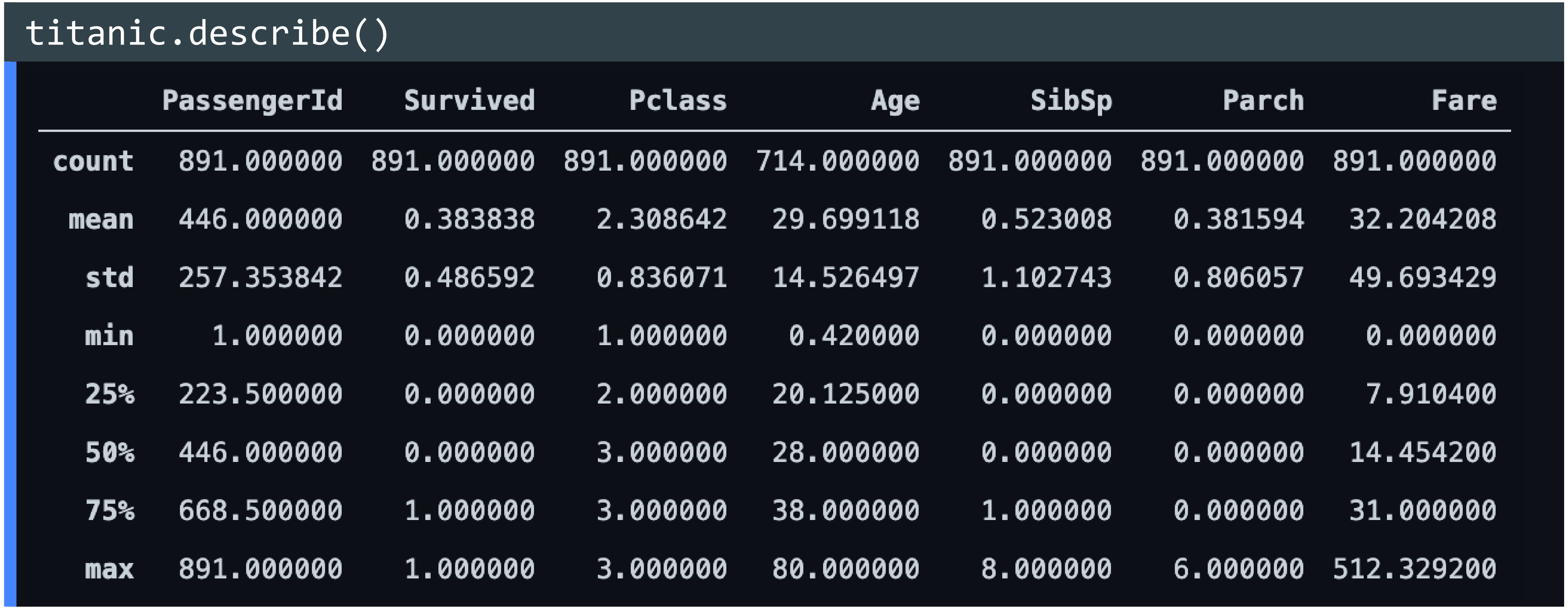
The top 10 percent reported income greater than $135,000.

Source: U.S. Census Bureau, Current Population Survey, 2011 Annual Social and Economic Supplement

SPAI

Descriptive Statistics

☑ Generating Descriptive Statistics



```python
titanic.describe()
```

```
       PassengerId    Survived      Pclass         Age       SibSp       Parch        Fare
count   891.000000  891.000000  891.000000  714.000000  891.000000  891.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008    0.381594   32.204208
std     257.353842    0.486592    0.836071   14.526497    1.102743    0.806057   49.693429
min       1.000000    0.000000    1.000000    0.420000    0.000000    0.000000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000    0.000000    7.910400
50%     446.000000    0.000000    3.000000   28.000000    0.000000    0.000000   14.454200
75%     668.500000    1.000000    3.000000   38.000000    1.000000    0.000000   31.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000    6.000000  512.329200
```

Describing Data

- ☑ Understand anomalies in data
- ☑ Count – Number of values in column
- ☑ Mean – Value of column
- ☑ STD – Spread of data
- ☑ Min, Max – Minimum and maximum value of column.
- ☑ 25%, 50%, 75% - Q1, Median, Q3

Data Visualization

- ☑ A form of visual art to interest us to keep us on the message
- ☑ Helps us visualise trends and outliers
- ☑ Visualization helps us internalise quickly.

Pandas Data Visualization

☑ Provides built-in visualization tools

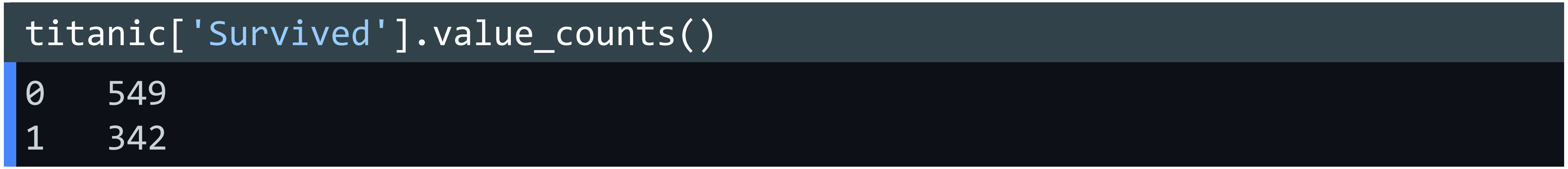
☑ May not be as detailed as Seaborn and Plotly

☑ Basic syntax to generate a plot

```
data_frame['column_name'].plot(kind='type_of_plot')
```

Bar Chart

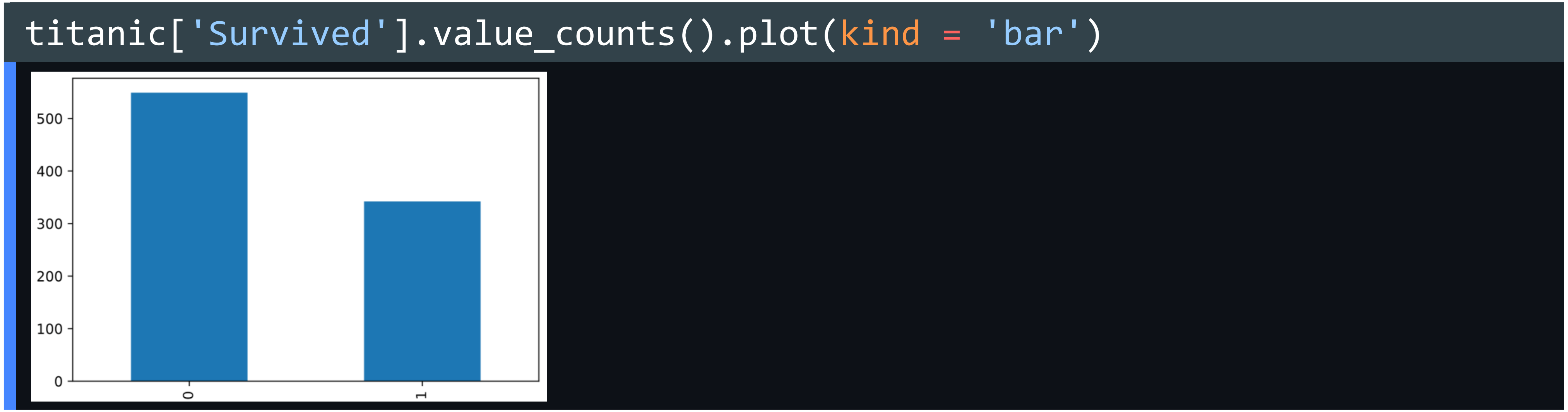
☑ Find total number of males and females



```
titanic['Survived'].value_counts()
```

```
0    549
1    342
```

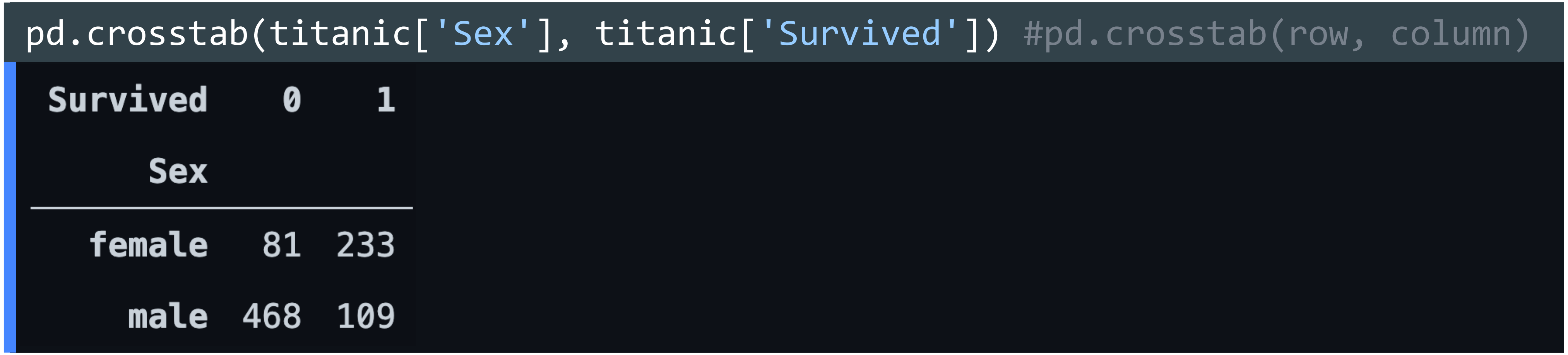
☑ Plotting graph



```
titanic['Survived'].value_counts().plot(kind = 'bar')
```

Cross-tabulation

☑ Analyse frequency of occurrence for two categorical variables.



```
pd.crosstab(titanic['Sex'], titanic['Survived']) #pd.crosstab(row, column)
```

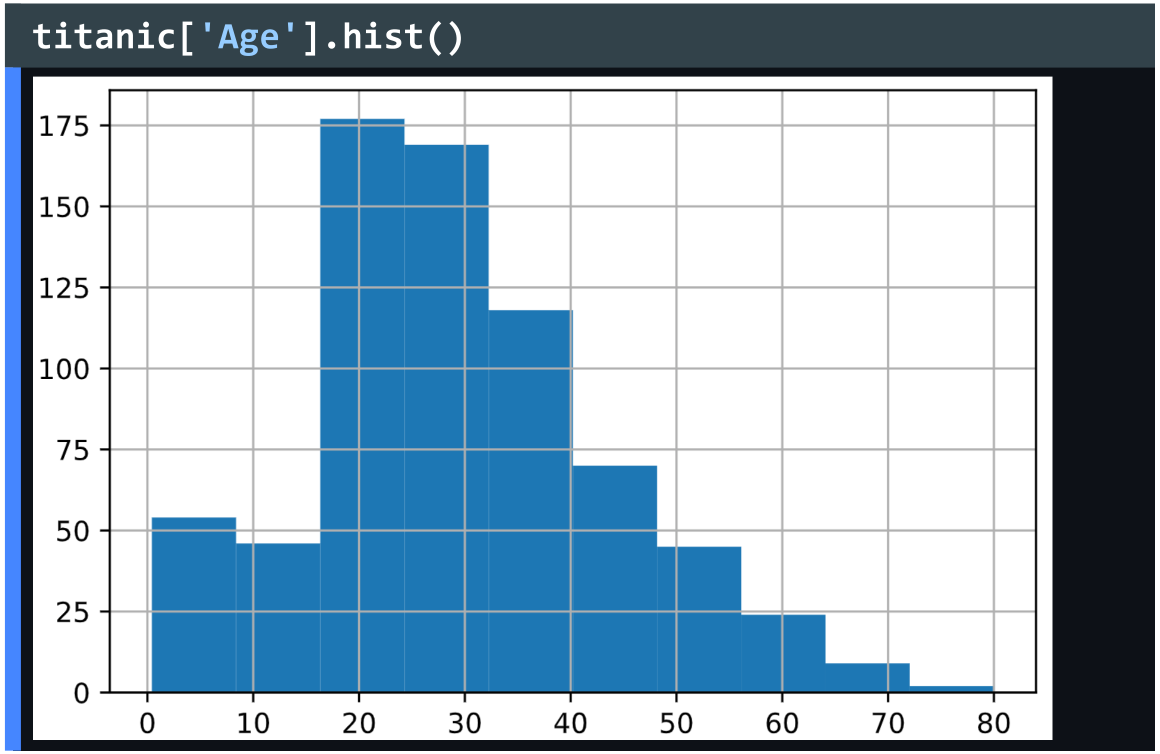
```
Survived    0    1
     Sex
------------------
  female   81  233
    male  468  109
```

Histogram

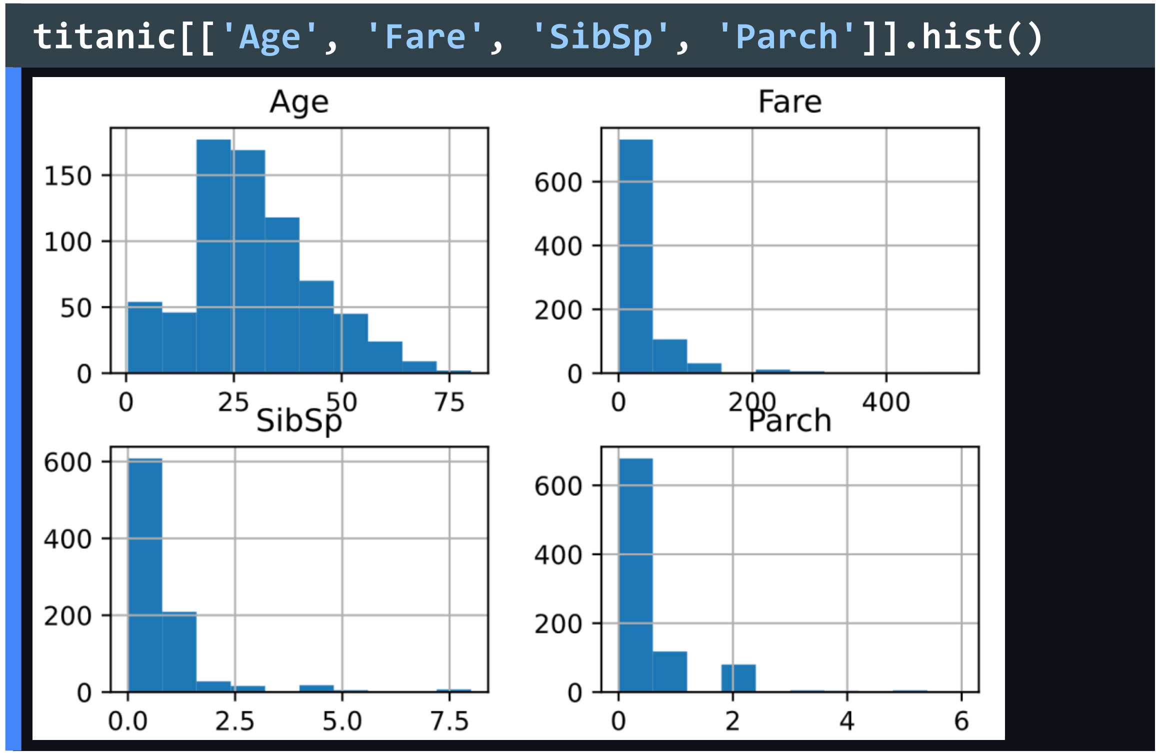
Show general distribution of numerical features.

☑ Create a histogram with .hist()

☑ Create multiple histograms



```
titanic['Age'].hist()
```



```
titanic[['Age', 'Fare', 'SibSp', 'Parch']].hist()
```

Boxplot

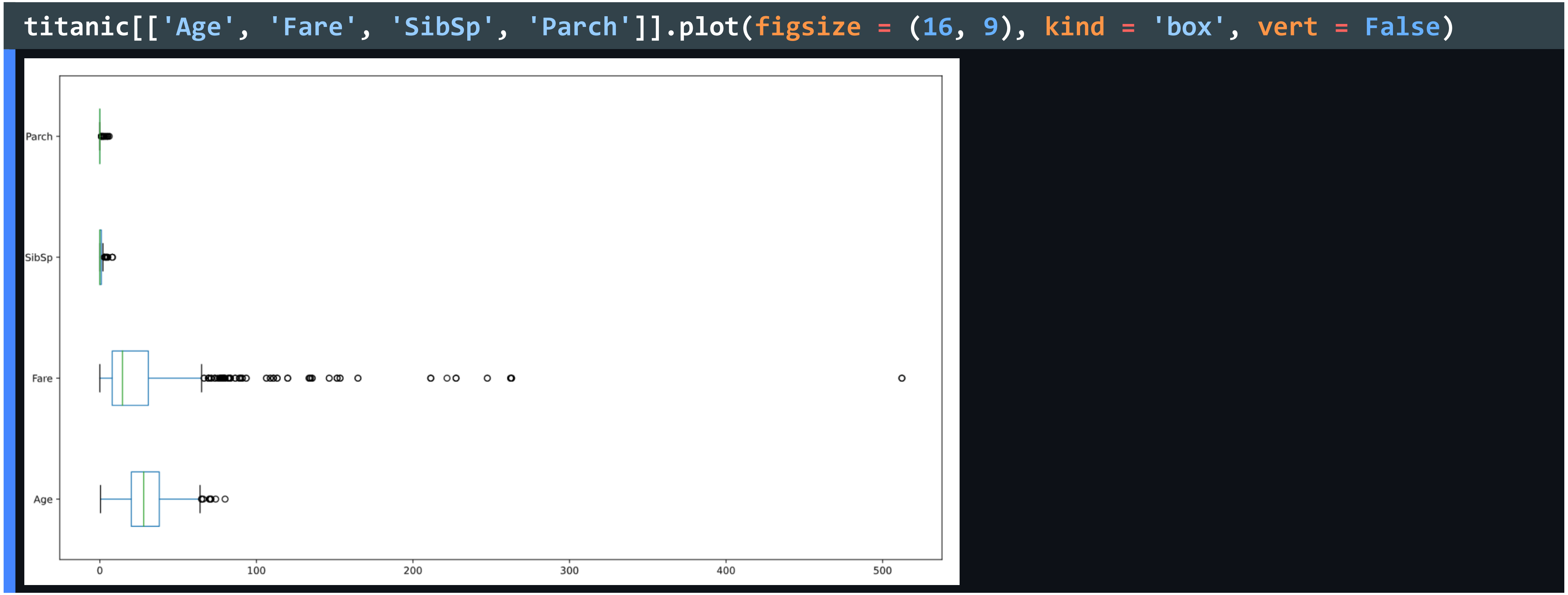
☑ Shows maximum, minimum and medium of data

☑ Used to check for outliers in data

☑ “figsize” argument used to expand size of figure

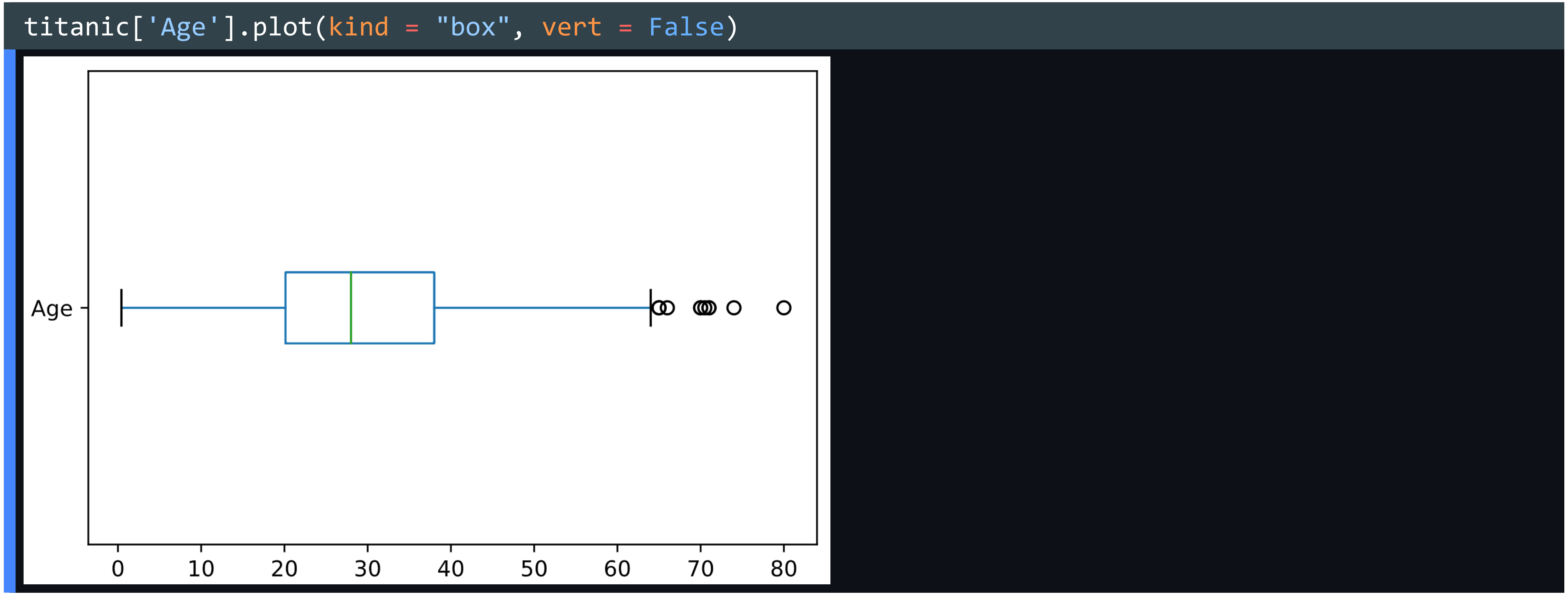
Boxplot

“vert = False” means setting boxplot horizontally



```
titanic[['Age', 'Fare', 'SibSp', 'Parch']].plot(figsize = (16, 9), kind = 'box', vert = False)
```

Boxplot



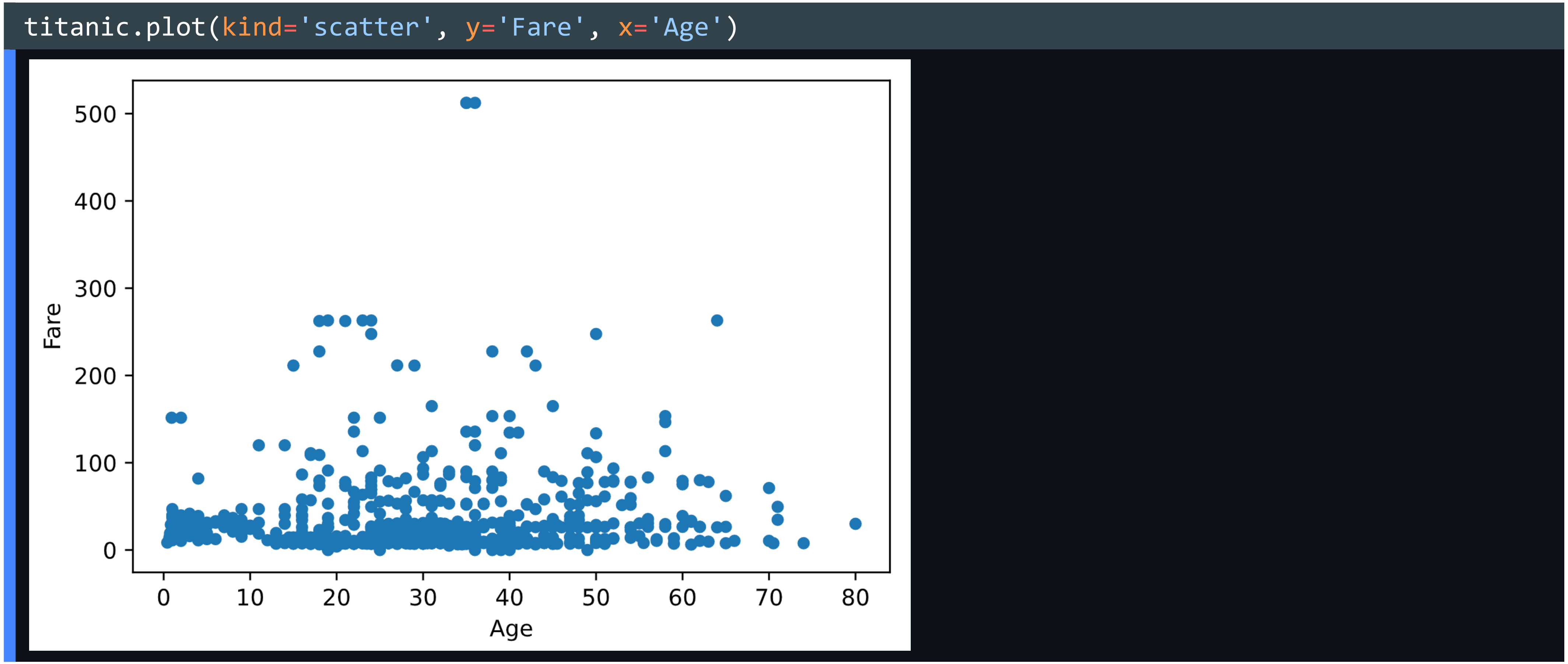
```python
titanic['Age'].plot(kind = "box", vert = False)
```

Approach to EDA

- ☑ Load in data
- ☑ Describe data
- ☑ Check for errors in data
- ☑ Analyze each variable with graphical & non-graphical methods
- ☑ **Analyze relationships between variables.**

Scatter Plot

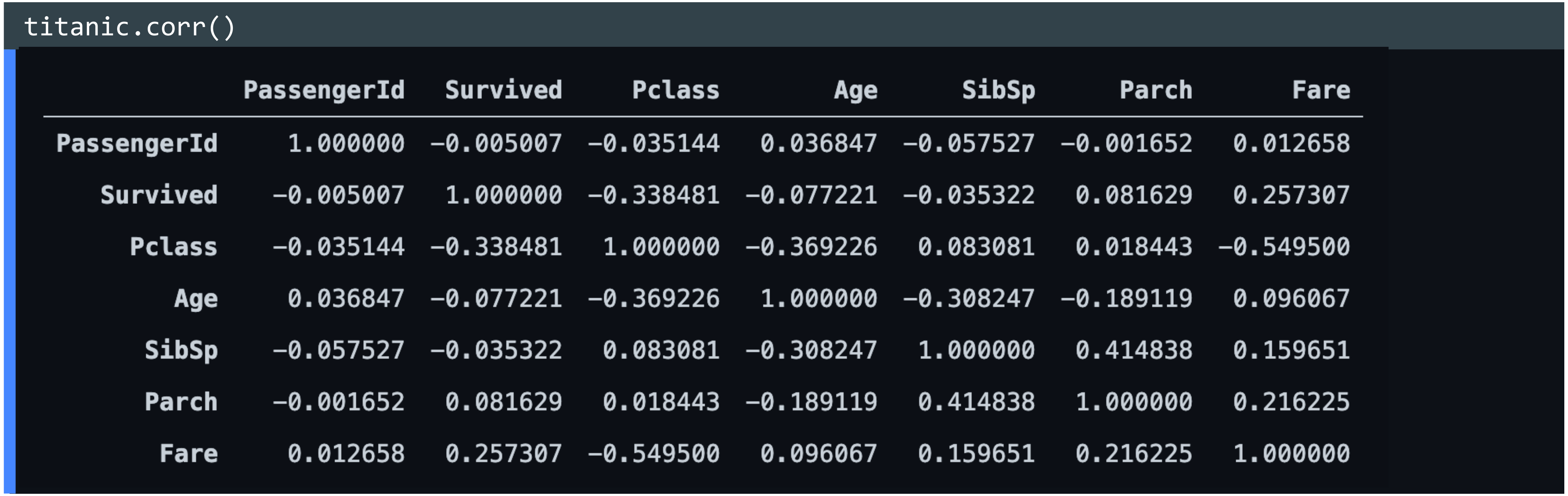
Show relationship between two numerical continuous features.



```
titanic.plot(kind='scatter', y='Fare', x='Age')
```

Correlation Plot

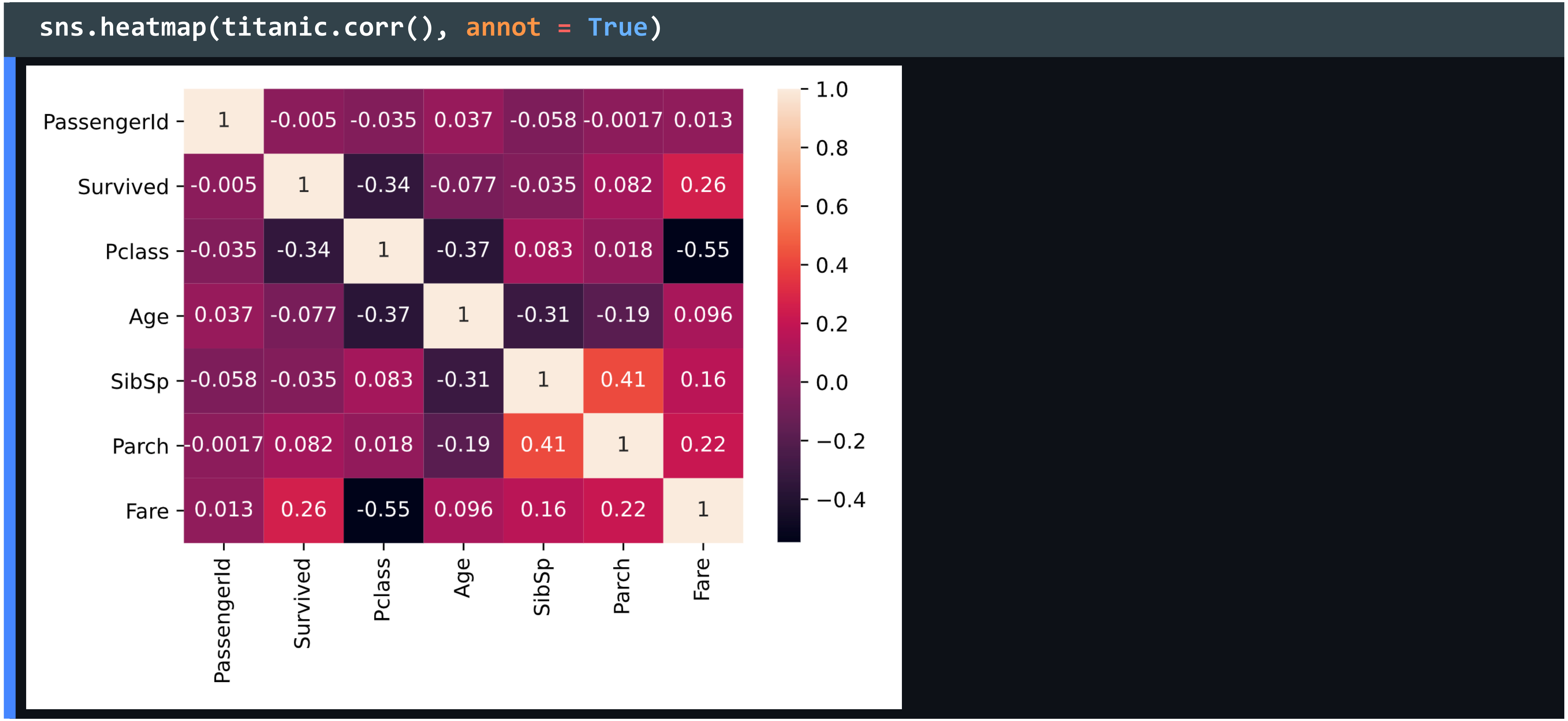
☑ Indicates strength and direction of linear relationship



```
titanic.corr()
```

```
             PassengerId   Survived    Pclass       Age     SibSp     Parch      Fare
PassengerId     1.000000  -0.005007 -0.035144  0.036847 -0.057527 -0.001652  0.012658
   Survived    -0.005007   1.000000 -0.338481 -0.077221 -0.035322  0.081629  0.257307
     Pclass    -0.035144  -0.338481  1.000000 -0.369226  0.083081  0.018443 -0.549500
        Age     0.036847  -0.077221 -0.369226  1.000000 -0.308247 -0.189119  0.096067
      SibSp    -0.057527  -0.035322  0.083081 -0.308247  1.000000  0.414838  0.159651
      Parch    -0.001652   0.081629  0.018443 -0.189119  0.414838  1.000000  0.216225
       Fare     0.012658   0.257307 -0.549500  0.096067  0.159651  0.216225  1.000000
```

Corroelation Plot



```
sns.heatmap(titanic.corr(), annot = True)
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1	-0.005	-0.035	0.037	-0.058	-0.0017	0.013
Survived	-0.005	1	-0.34	-0.077	-0.035	0.082	0.26
Pclass	-0.035	-0.34	1	-0.37	0.083	0.018	-0.55
Age	0.037	-0.077	-0.37	1	-0.31	-0.19	0.096
SibSp	-0.058	-0.035	0.083	-0.31	1	0.41	0.16
Parch	-0.0017	0.082	0.018	-0.19	0.41	1	0.22
Fare	0.013	0.26	-0.55	0.096	0.16	0.22	1

Knowledge Check

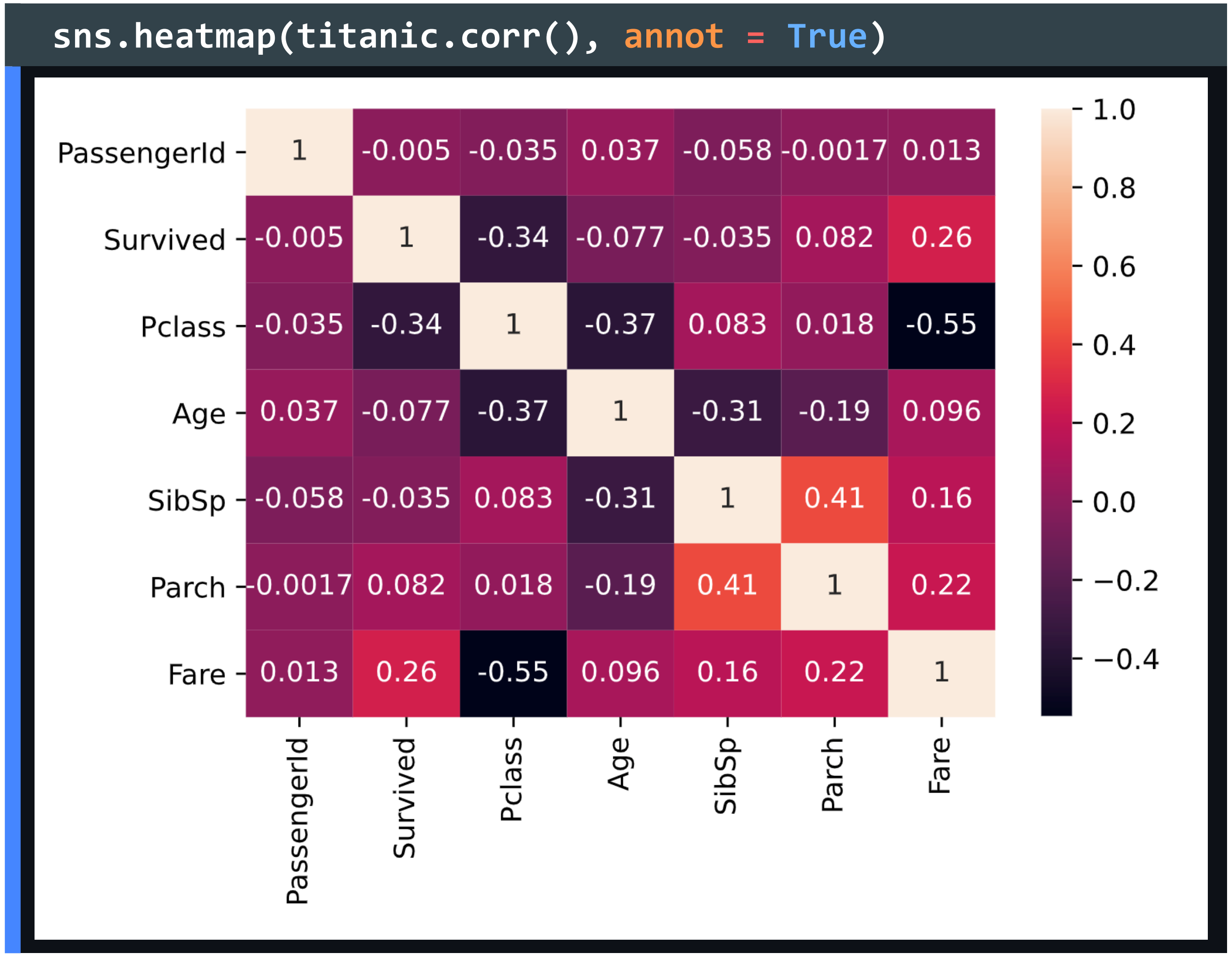
Out of the four options below, which feature has the greatest correlation with “survived”?

A. Fare

B. Age

C. Passenger Class

D. Sibsp (No. of Siblings and Spouses)



```
sns.heatmap(titanic.corr(), annot = True)
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1	-0.005	-0.035	0.037	-0.058	-0.0017	0.013
Survived	-0.005	1	-0.34	-0.077	-0.035	0.082	0.26
Pclass	-0.035	-0.34	1	-0.37	0.083	0.018	-0.55
Age	0.037	-0.077	-0.37	1	-0.31	-0.19	0.096
SibSp	-0.058	-0.035	0.083	-0.31	1	0.41	0.16
Parch	-0.0017	0.082	0.018	-0.19	0.41	1	0.22
Fare	0.013	0.26	-0.55	0.096	0.16	0.22	1

Pandas-Profiling as tool for EDA

✓ Pros

Robust

Can be converted into PDF or HTML file types.

✗ Cons

Computationally expensive especially for large datasets

Scan to mark attendance

Please scan the QR code to Sign Out and Tell Us Your Feedback

Feedback Form

SPAI

An AI Singapore Student Chapter

Thank You

@spai.sp

SPAI 21/22

SPAI