**Introduction to Machine Learning.**

**Tutorial 01: Loading and Viewing Data.**

**Requirements:**

* Basic understanding of Python 3
* Jupyter notebook
  1. Scikit Learn
  2. Numpy
  3. Pandas

R**esources:**

1. I prefer using Anaconda Distribution as it includes Python, Jupyter notebook and all the packages required.

Follow the link below to download and install Anaconda Distribution for Windows, Mac OS and Linux.

<https://www.anaconda.com/download>

A detailed video on how to download and install anaconda is linked below.

<https://www.youtube.com/watch?v=HW29067qVWk>

1. Titanic Dataset from Kaggle.com.

We will analyze the famous Titanic dataset. Follow the below link to download train.csv and test.csv

<https://www.kaggle.com/c/titanic/data>

After downloading the files save the files in the current working directory.

**Loading and Analyzing Data Using Pandas in Jupyter Notebook.**

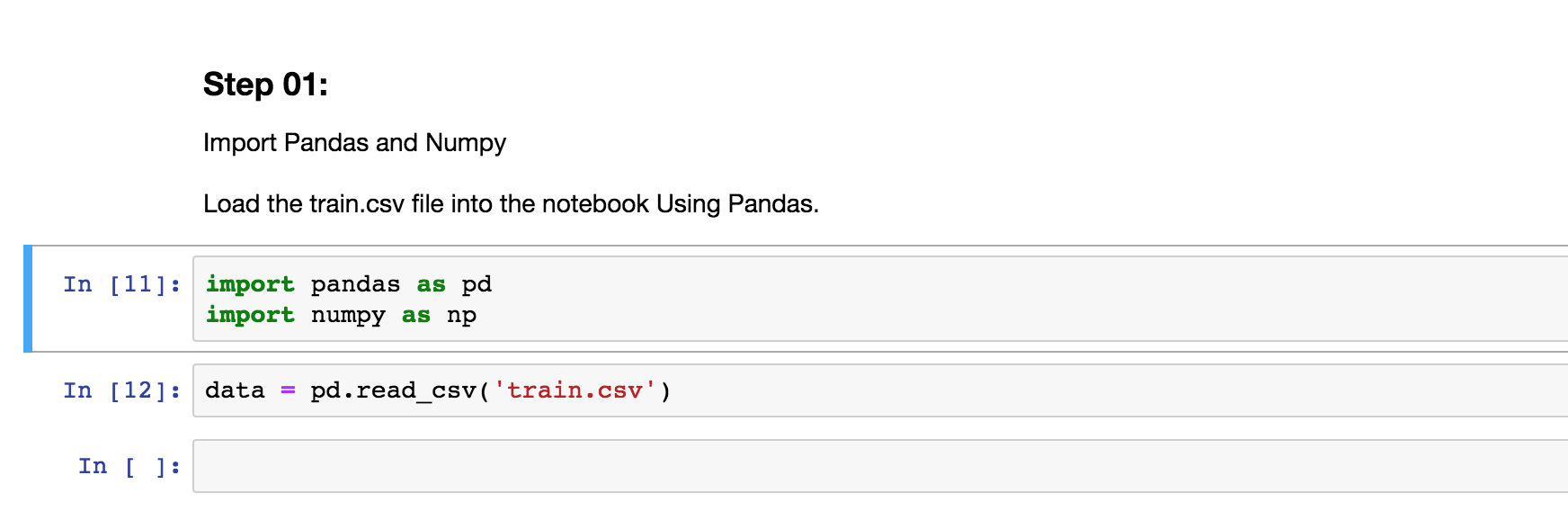
**Step 01:**

Import pandas as pd

Import numpy as np

Load the train.csv file into the Jupyter notebook using pandas, and save the file into a variable.

data = pd.read.csv(‘train.csv’)



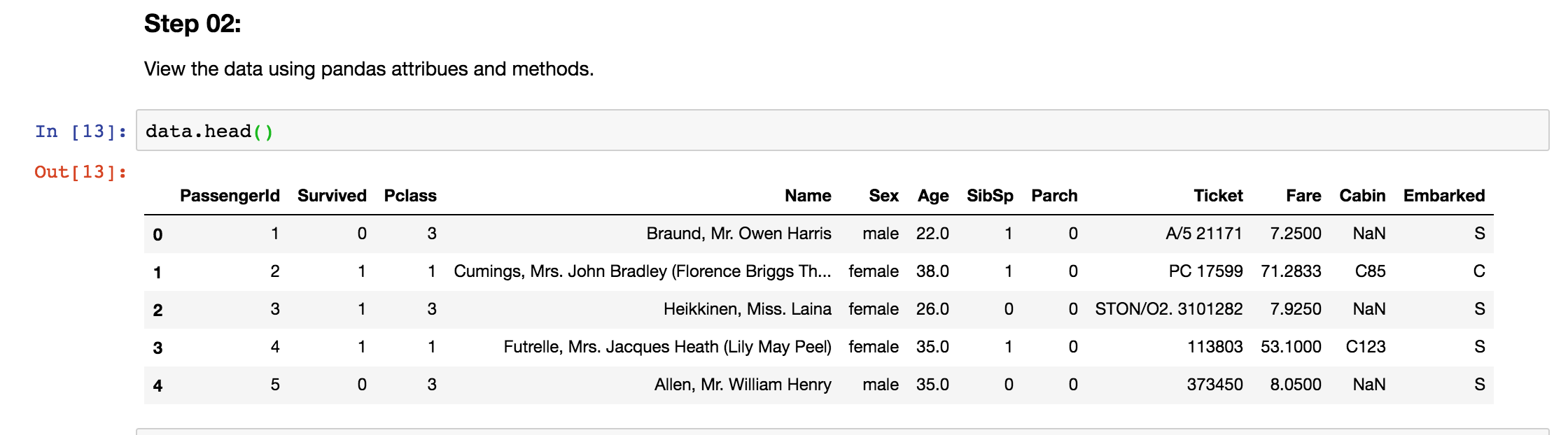
**Step 02:**

* Now we have successfully loaded the file for analysis. Next, we can look at the data, the rows columns etc.
* First, we will call the **head method** to look at the **first 5 rows** of the dataset.

**data.head()**

The columns are called **Features**

The rows are called **observations**

****

* Next, we can call the shape attribute to look at the number of features and the observations we have in the dataset.

**data.shape**



As we see we get a tuple (891,12).

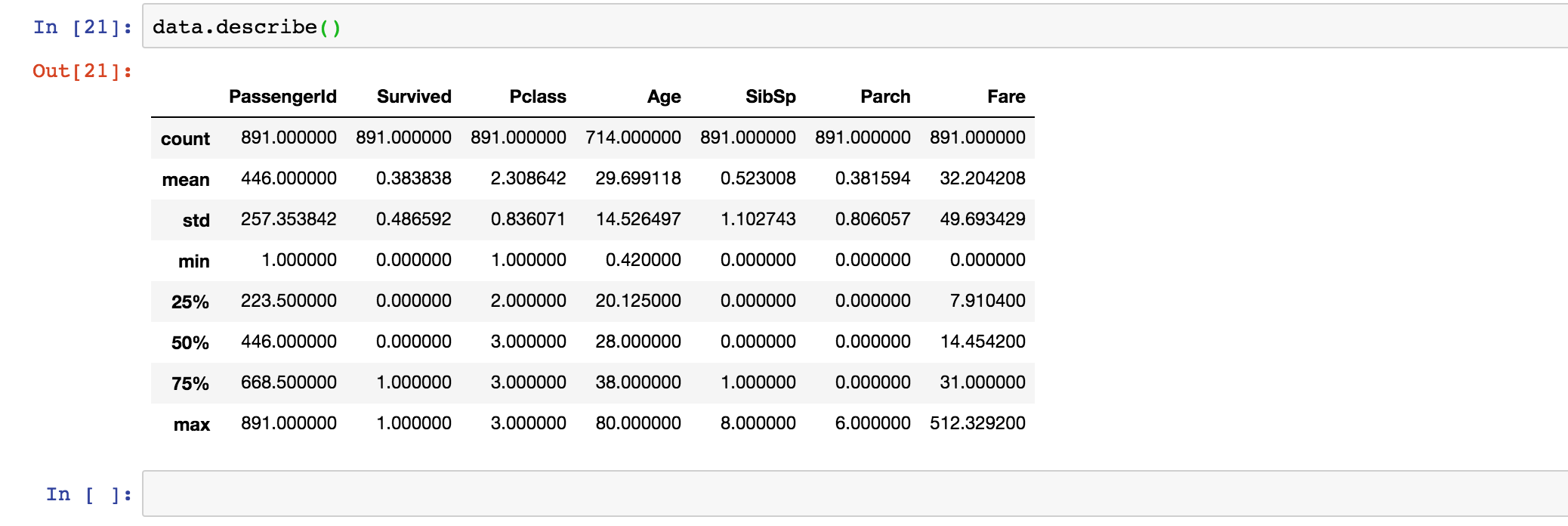
891 is the number of rows, **observations**  in the dataset.

12 is the number of **features** in the dataset.

**Each feature is also called a pandas series. When we combine the 12 series we get a pandas data frame.**

* We could also look at the describe() method to get information about the numerical data.

**data.describe()**

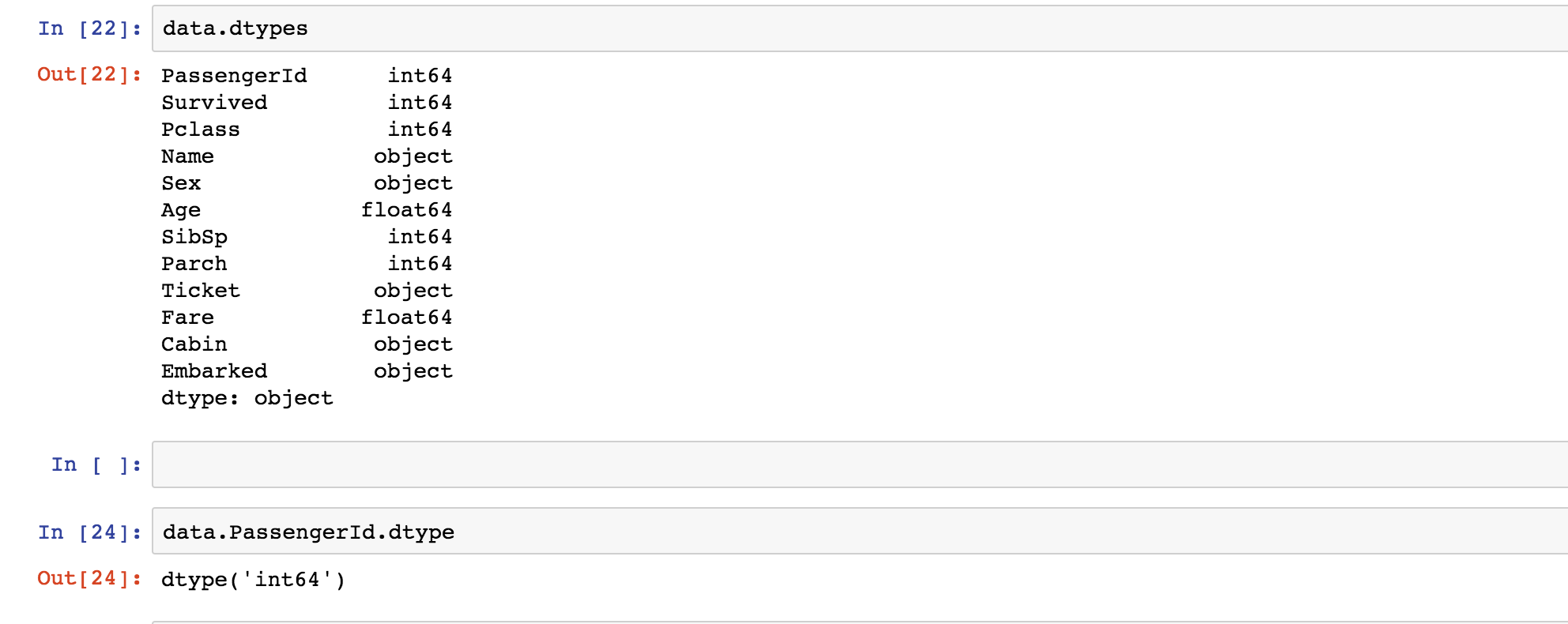


* Then we can check the dtypes attribute to check the data types of all the features ora single feature.

**data.dtypes**

Object = Strings

Other data types are integers and floats.

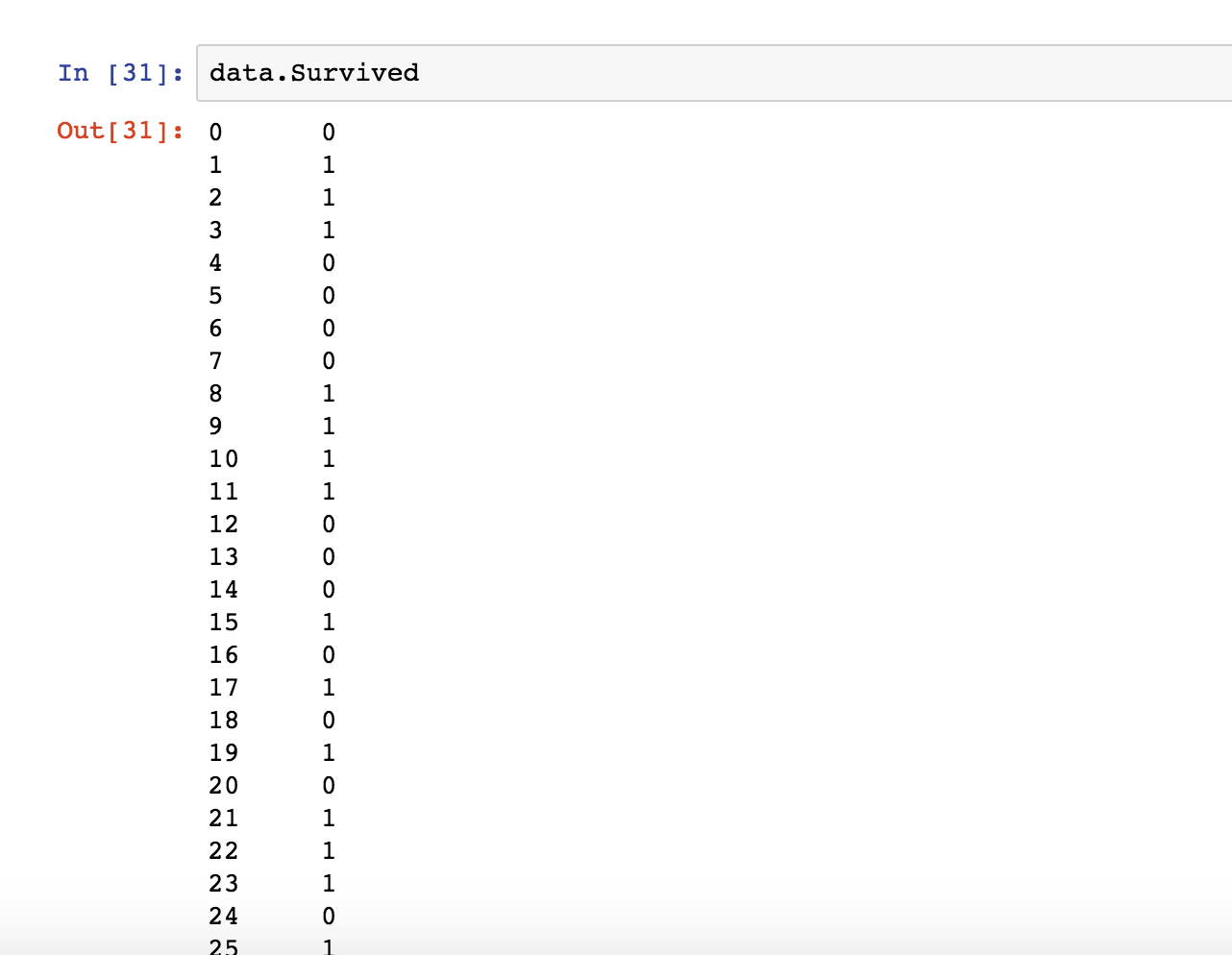
****

**Dot notation vs Bracket Notation.**

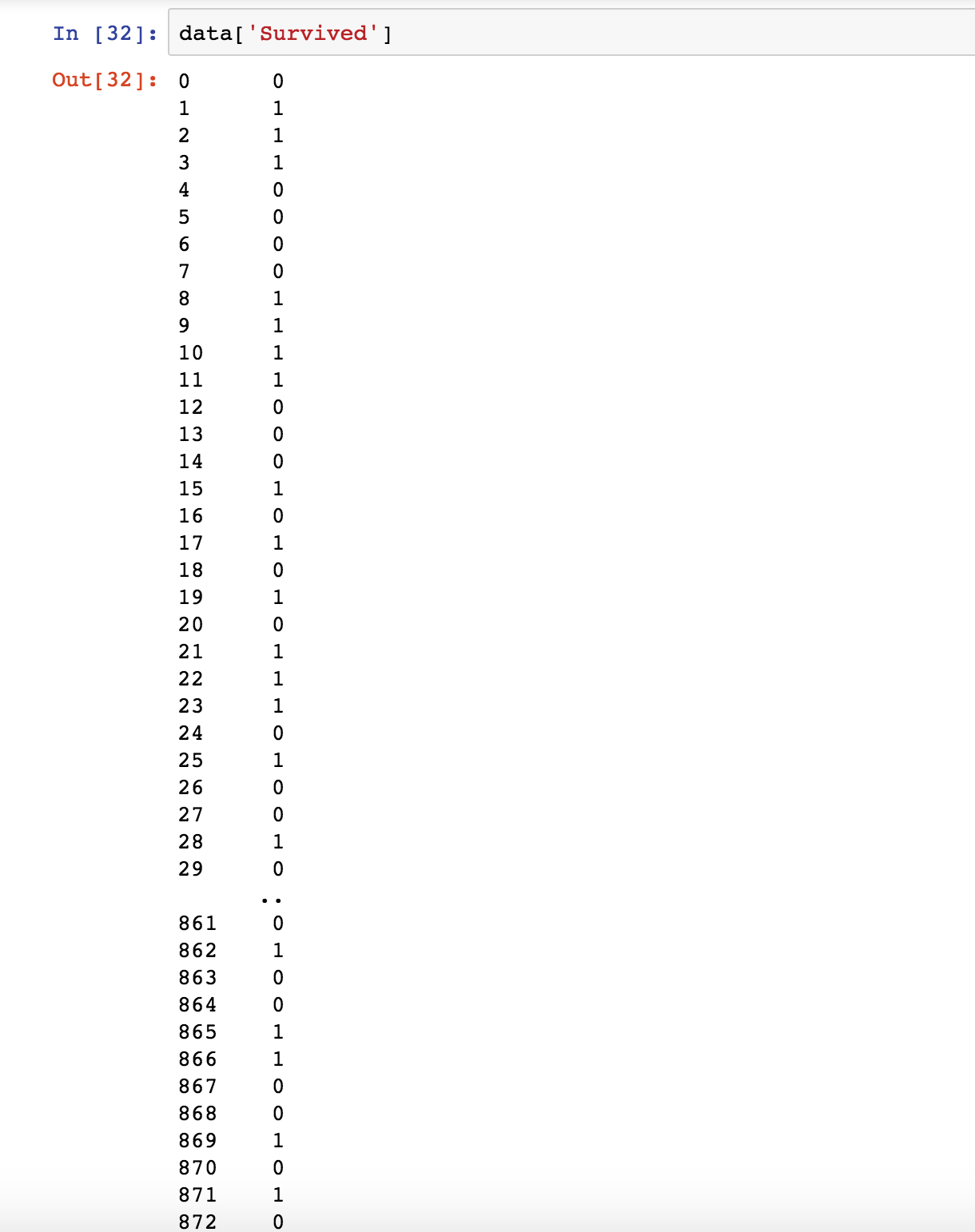
We can look at individual features using Pandas. This can be achieved by using dot notation and bracket notation.

Assume we need to look at the Survived feature in the dataset.

**data.Survived** ---- Dot notation.



**data[‘Survived’]** --- Bracket Notation.



* Assume we have a series **Survival Rate**. In this case, we need to use bracket notation as not notation would give us an error.

Our goal using the Titanic dataset is to predict the **Survival** of the passengers listed in the dataset. And to do that we need to use the dataset and the features given in the dataset. Therefore, the **predictor(target) vector is called y**, and the **feature matrix** used to predict y is denoted as **X**. Note how X is upper case and y is lower case.

