What it Can Do

- Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.
- Built on top of the Python programming language.
- Pandas can clean messy data sets, and make them readable and relevant.
- Pandas give you answers about the data. Like:
  - Is there a correlation between two or more columns?
  - What is the average value?
  - Max value?
  - Min value?
- Pandas are also able to delete rows that are not relevant or contain wrong values, like empty or NULL values. This is called cleaning the data.
- **Plot points**: You can plot histograms using the **plt.hist** function.
- If you have Python and PIP already installed on a system, then installation of Pandas is very easy.
  - *Install it using*: C:\Users\Your Name>pip install pandas
- *Once Pandas is installed, import it in your applications by adding the import keyword*: import pandas

Basic Functions (All in Python)

- Reading excel files (both XLS and XLSX) is as easy as the read_excel() function, using the file path as an input.

```
df = pd.read_excel('diabetes.xlsx')
```

POWERED BY DATACAMP WORKSPACE COPY CODE

- One of the most used methods for getting a quick overview of the data frame is the head() method. The head() method returns the headers and a specified number of rows, starting from the top.

Get a quick overview by printing the first 10 rows of the DataFrame:

```
import pandas as pd

df = pd.read_csv('data.csv')

print(df.head(10))
```

- There is also a tail() method for viewing the last rows of the data frame. The tail() method returns the headers and a specified number of rows, starting from the bottom.

Print the last 5 rows of the DataFrame:

```
print(df.tail())
```

- The .describe() method prints the summary statistics of all numeric columns, such as count, mean, standard deviation, range, and quartiles of numeric columns.

```
df.describe()
```

POWERED BY DATACAMP WORKSPACE COPY CODE

➢ Similarly, you might want to exclude certain data types using exclude argument.

```
df.describe(exclude=[int])
```

POWERED BY DATACAMP WORKSPACE COPY CODE

➢ Calling the .columns attribute of a DataFrame object returns the column names in the form of an Index object. As a reminder, a pandas index is the address/label of the row or column.

```
df.columns
```

POWERED BY DATACAMP WORKSPACE COPY CODE

➢ It can be converted to a list using a list() function.

```
list(df.columns)
```

POWERED BY DATACAMP WORKSPACE COPY CODE

➢ Pandas' apply function enables you to perform mathematical operations on data. This is quite useful because the dataset you have may not always be in the correct order, and using a mathematical procedure on the dataset will solve this issue. This is one of the Pandas' most appealing characteristics.

➢ You must consistently merge and join multiple datasets to generate a final dataset to assess it while analyzing data accurately. This is significant because if the datasets aren't properly combined or linked, the results will be compromised, which you don't want. Pandas can help you integrate multiple datasets quickly and efficiently, ensuring that you don't run into any issues while analyzing the data.