Pandas

- Pandas is a Python library used for working with **data sets**.
- It has functions for **analysing, cleaning, exploring, and manipulating data.**
- The name "Pandas" has a reference to both **"Panel Data"**, and **"Python Data Analysis"** and was created by Wes McKinney in 2008

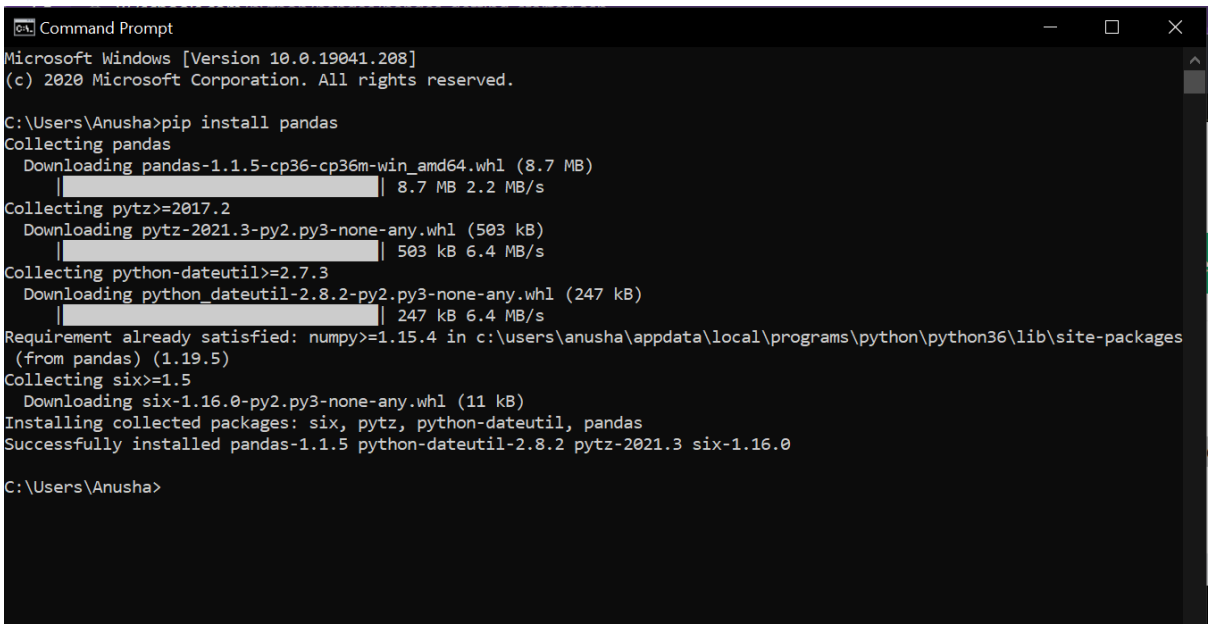
Uses of Pandas

- Pandas allows us to analyze big data and make conclusions based on statistical theories.
- Pandas can clean messy data sets, and make them readable and relevant.
- Relevant data is very important in data science.

Installation of Pandas

If you have Python and PIP already installed on a system, then installation of Pandas is very easy.

Step1: *Install it using this command:* ***pip install pandas***



Step2: ***import pandas***

Example

```
import pandas as pd

mydataset = {
  'cars': ["BMW", "Volvo", "Ford"],
  'passings': [3, 7, 2]
}

myvar = pd.DataFrame(mydataset)

print(myvar)
```

```
    cars  passings
0    BMW         3
1  Volvo         7
2   Ford         2
```

Pandas Series

*A **Pandas Series** is like a column in a table.* It is a one-dimensional array holding data of any type.

Example1:

```
import pandas as pd

a = [1, 7, 2]

myvar = pd.Series(a)

print(myvar)
```

Labels

If nothing else is specified, the values are labeled with *their index number*. First value has index 0, second value has index 1 etc.

This label can be used to *access a specified value*.

Example2:

```
import pandas as pd

a = [15, 78, 24,51,60,89]

m= pd.Series(a)

print(m[0])
```

Create Labels

With the `index` argument, you can name your own labels.

Example3:

```
import pandas as pd

a = [15, 78, 24,51,60,89]

m= pd.Series(a,index=["X","Y","Z","W","T","B"])

print(m)

print(m["W"])   #access an item by referring to the label
```

Key/Value Objects as Series

You can also use a key/value object, like a dictionary, when creating a Series.

```
import pandas as pd

c = {"day1": 420, "day2": 380, "day3": 390}

V = pd.Series(c)

print(V)
```

Note: The keys of the dictionary become the labels.

To select only some of the items in the dictionary, use the `index` argument and specify only the items you want to include in the Series.

```
import pandas as p

cl = {"day1": 420, "day2": 380, "day3": 390,"day4": 520, "day5": 480,"day6": 470}

Y = p.Series(cl, index = ["day1", "day2"])

print(Y)
```

DataFrames

Data sets in Pandas are usually **multi-dimensional tables**, called **DataFrames**.

Series is like a column, a DataFrame is the whole table.

Example1:

```
import pandas as pd
data = {
  "calories": [420, 380, 390],
  "duration": [50, 40, 45]
}
myvar = pd.DataFrame(data)
print(myvar)
```

Locate Row

As you can see from the result above, the DataFrame is like a table with *rows and columns.*

Pandas use the **loc** attribute to return one or more specified row(s)

Example2:

```
import pandas as pd
data = {
  "calories": [420, 380, 390],
  "duration": [50, 40, 45]
}
df = pd.DataFrame(data)
print(df)
print(df.loc[0])

print(df.loc[[0,1]])    #Return row 0 and 1
```