PANDAS

The pandas package is the most important tool at the disposal of Data Scientists and Analysts working in Python today. The powerful machine learning and glamorous visualization tools may get all the attention, but pandas is the backbone of most data projects.

What's pandas for?

Through pandas, you get acquainted with your data by cleaning, transforming, and analyzing it.

For example, say you want to explore a dataset stored in a CSV on your computer. Pandas will extract the data from that CSV into a DataFrame — a table, basically — then let you do things like:

Calculate statistics and answer questions about the data, like

1.What's the average, median, max, or min of each column?

2.Does column A correlate with column B?

3.What does the distribution of data in column C look like?

3.Clean the data by doing things like removing missing values and filtering rows or columns by some criteria

4.Visualize the data with help from Matplotlib. Plot bars, lines, histograms, bubbles, and more.

5.Store the cleaned, transformed data back into a CSV, other file or database

INSTALL AND IMPORT

Install commands

```
pip install pandas
```

```
conda install pandas
```

To import Pandas

```
import pandas as pd
```

Core components of pandas:

1.Series

2.DataFrames

The primary two components of pandas are the Series and DataFrame.

A Series is essentially a column, and a DataFrame is a multi-dimensional table made up of a collection of Series.

Creating DataFrames From Scratch

```
data = {
    'apples': [3, 2, 0, 1],
    'oranges': [0, 3, 7, 2]
}
purchases = pd.DataFrame(data)

purchases
```

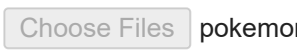
```
purchases = pd.DataFrame(data, index=['June', 'Robert', 'Lily', 'David'])

purchases
```

So now we could locate a customer's order by using their name:

```
purchases.loc['June']
```

```
from google.colab import files
uploaded = files.upload()
```



- **pokemon.csv**(application/vnd.ms-excel) - 44028 bytes, last modified: 9/21/2016 - 100% done

```
Saving pokemon.csv to pokemon.csv
```

LOADING DATA INTO PANDAS

```
import pandas as pd
dataset=pd.read_csv('pokemon.csv')
dataset
#print(dataset.head(3))
#dataset['HP']
```

READING DATA IN PANDAS

Read Headers

```
dataset.columns
```

Read each Column

```
#print(dataset.Name)
#print(dataset['Name'][0:5])
print(dataset[['Name', 'Type 1', 'HP']])
```

Read each Row

```
#print(dataset.head(4))
print(dataset.iloc[0:4])
#print(dataset.iloc[1])
#for index, row in dataset.iterrows():
#     print(index, row['Name'])
#dataset.loc[dataset['Type 1'] == "Grass"]
```

Read specific location(R,C)

```
print(dataset.iloc[2,1])
```

SORTING/DESCRIBING DATA

```
#dataset.describe()

#dataset.sort_values('Name',ascending=False)

dataset.sort_values(['Type 1', 'HP'], ascending=[1,0])
dataset
```

MAKING CHANGES TO THE DATA

```
#dataset['Total'] = dataset['HP'] + dataset['Attack'] + dataset['Defense'] + dataset['Sp. Atk

#dataset = dataset.drop(columns=['Total'])

#dataset['Total'] = dataset.iloc[:, 4:10].sum(axis=1)

cols = list(dataset.columns)
dataset=dataset[cols[0:4]+cols[7:-1]+cols[4:7]]
```

```
#dataset = dataset[cols[0:4] + [cols[-1]]+cols[4:12]]
#dataset

dataset.head(5)
```

SAVING OUR DATA (EXPORTING INTO DESIRED FORMAT)

```
#dataset.to_csv('modified.csv')

from google.colab import files
dataset.to_csv('modified.csv')
files.download('modified.csv')
```

FILTERING DATA

```
#new_dataset = dataset.loc[(dataset['Type 1'] == 'Grass') | (dataset['Type 2'] == 'Poison')]
#new_dataset = dataset.loc[(dataset['Type 1'] == 'Grass') & (dataset['Type 2'] == 'Poison') &


#new_dataset=new_dataset.reset_index()
#new_dataset.reset_index(drop=True, inplace=True)

#dataset.loc[~dataset['Name'].str.contains('Mega')]

import re
#dataset.loc[dataset['Type 1'].str.contains('fire|grass',flags=re.I, regex=True)]
dataset.loc[dataset['Name'].str.contains('^pi[a-z]*',flags=re.I, regex=True)]


#new_dataset
```

CONDITIONAL CHANGES

```
#dataset.loc[dataset['Type 1']=='Flamer','Type 1']='Fire'
#dataset.loc[dataset['Type 1']=='Fire','Legendary']=True


#dataset.loc[dataset['Total'] > 500, ['Generation','Legendary']] = ['Test 1', 'Test 2']

dataset = pd.read_csv('modified.csv')

dataset
```

AGGREGATE STATISTICS (GROUP BY)

```
#dataset = pd.read_csv('modified.csv')


#dataset.groupby(['Type 1']).mean().sort_values('Attack',ascending=False)


dataset['count'] = 1
dataset.groupby(['Type 1']).count()['count']

#dataset.groupby(['Type 1', 'Type 2']).count()['count']

#dataset
```