MSc - Data Mining

Topic 01 : Module Overview

Part 06 : Top X pandas commands

Dr Bernard Butler and Dr Kieran Murphy

Department of Computing and Mathematics, WIT.
(bbutler@tssg.org and kmurphy@wit.ie)

Spring Semester, 2021

Outline

- Reading data formats
- Computing descriptive statistics
- Processing data by filtering and grouping

Part I

Input and Output

Setup

Minimal

We begin every data mining project with importing the three core data science packages:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
plt.style.use('seaborn-darkgrid')
```

numpy — fast array operations
pandas — data manipulation
matplotlib — visualisation

- We give modules nicknames (`np`, `pd`, . . .) to simplify their later use, and we access properties/functions of a package using the dot notation (`np.max`, `pd.DataFrame`, . . .).

Extra

```
import seaborn as sns
import statsmodels.api as sm

pd.set_option('display.max_columns', 500)
pd.set_option('display.width', 1000)
```

seaborn — statistical visualisation
statsmodels — statistical data exploration
pandas options to show all columns for wider datasets

Reading data from a CSV file

Pandas supports a huge variety of input/output formats so best approach is to focus on what is needed to process the given data and verify input. Our marks dataset is in CSV format so we start with

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
plt.style.use('seaborn-darkgrid')
```

and input using