Opening and Reading with Files

Opening & Reading TXT Files

with

```
with open([file_address], [mode('r'/'w')], encoding=[encoding]) as file:
    content = file.read()
    print(content)
```

Get Rainbow CSV Extension

What it is? An extension that works on CSV files.

Why do I need it? CSV files come as lines with comma-separated values. This extension automatically colours each 'column' differently for better visibility.

Where do I get it? VSCode → ctrl + shift + X → search & install 'Rainbow CSV'

Opening & Reading CSV Files

with

```
import csv

with open([file_address], [mode], [encoding]) as file:
    reader = csv.reader(file)

    for row in reader:
        print(row)
```

Analysing Tables with pandas

Turn CSV Files into pandas dataframes

read_csv

```
import pandas as pd

df = pd.read_csv([file_address])
```

Displaying Table Info with pandas

Syntax	What it Does
`df.head(k)`	shows first k rows of the table, default is 5
`df.tail(k)`	shows last k rows of the table, default is 5
`df.shape`	returns a tuple of (num_of_rows, num_of_cols)

Manipulating Columns with pandas

Syntax	What it Does
`df[column_name]`	returns all values in a given column as an iterable series
`df[df[col_name] == val]`	compares all values in a given column to value, returns only rows that pass
`df[new_col_name] = series`	assigns elements of a series one-by-one to the lines of the df as new col
`df.sort_values(col_name)`	sorts col values in ascending/alphabetical order

Looking up Specific Cells pandas

Syntax	What it Does
`df.loc[k, col_name]`	returns a cell in the k-th row of the specified col
`df.iloc[m, n]`	returns a cell in the m-th row of the n-th col

Grouping Values with pandas

`df.groupby(col1).func()`	gets unique elements (groups) from col1 and applies aggregation func
`df.groupby(col1)[col2].func()`	groups unique values in col1, applies aggregation func to col2

Other Stats with pandas

`df.describe()`	gets basic stats for all numeric cols
`df[col_name].median()`	prints median of a given col
`df[col_name].mean()`	prints mean of a given col
`df[col_name].std()`	prints standard deviation of a given col
`df[col_name].sum()`	prints sum of a given col
`df[col_name].min()`	prints the minimal value in a given col
`df[col_name].max()`	prints the maximal value in a given col

Learn more about all the exciting stuff you can do with pandas dataframes here