Pandas cheatsheet 1

Import

```
import pandas as pd
```

Reading and Saving csv

```
pd -> pandas
df -> dataframe
```

to read a file into dataframe

```
df = pd.read_csv('filename')
```

to save a dataframe to a file

```
df.to_csv('filename')
```

Describing dataframe

look at the first 5 lines

```
df.head()
```

look at the last 5 lines

```
df.tail()
```

to get the dimensions of df

```
df.shape
```

to print all the column names

```
df.columns
```

to print data types

```
df.dtypes
```

to describe df

```
df.describe()
```

Describing columns

unique values in a column

```
df['column name'].unique()

# if no spaces we can omit ["]
df.column_name
```

describe (results depend on variable type)

```
df['column name'].describe()
```

value counts (frequencies)

```
df['col name'].value_counts()
```

Selection

return one column

```
df['column name']
```

return multiple columns

```
df[['column1', column2]]
```

return one line

```
df.loc[4]
```

return multiple lines

```
df.loc[[4,8]]
```

return a range of lines

```
df.loc[4:8]
```

return a cell

```
df.loc[4, 'col name']
df.loc[4]['col name']
```

return a range of lines and columns

```
df.iloc[0:4, 1:3]
```

Statistics (column with nums)

median (διάμεσος)

```
df['population'].median()
```

mean (μέση τιμή)

```
df['population'].mean()
```

sum

```
df['population'].sum()
```

minimum value

```
df['population'].min()
```

maximum value

```
df['population'].max()
```

standard deviation (τυπική απόκλιση)

```
df['population'].std()
```

correlation (συσχέτιση)

```
df['population'].corr()
```

Sorting

sort (by one column)

sorting can be done column wise – default is ascending

```
df.sort_values(by='col name')
```

use ascending=False for descending sort - "by=" may be omited

```
df.sort_values('col1', ascending=False)
```

sort by multiple columns

```
df.sort_values(['col1', 'col2'], ascending = [True,False])
```

Filtering

filtering with one condition

return rows where the condition is True

```
df[df['col1'] >= 5]
```

filtering with two conditions

with "&" both conditions must be True

```
df[(df['col1'] >= 5) & (df['col2'] != "Asia")]
```

with "|" either condition must be True

```
df[(df['col1'] >= 5) | (df['col2'] != "Asia")]
```

Grouping

get information about subsets

e.g. group rows with the same values in a column and get the mean of the corresponding values from another column

```
df.groupby("continent")["gdp_per_capita"].mean()
```

Df manipulation

create a new column

```
df['newCol'] = df['col3'] / 5
```

create a new dataframe

```
df2 = df[['c4','c1','c2']]
```