LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

Carlos J. Costa

EXPLORATORY DATA ANALYSIS

```
import pandas as pd
import matplotlib.pyplot as plt
data = pd.read_csv('WorldBankPort.csv',sep=';', index_col='year’)
data
```

Content identification

- Analysing the data starts with the characterization of the dataset, namely:
- number of variables
- the number of records

```
data.shape
```

Single Variable Analysis

- ***Variables Type and Domain***
- ***Missing values***
- ***Variables Distribution***
- ***Granularity***

Variables Type and Domain

```
data.dtypes


catVariables = data.select_dtypes(include='object')
for i in catVariables:
        print(i, data[i].unique())
```

Missing values

- It is possible analyse the numbers by plotting them through a bar chart.

Variables Distribution

- Several metrics may be used to understand the distribution

```
data.describe()
```

- It is possible to graph the distribution

```
data.boxplot(figsize=(10,6))
plt.show()
```

Granularity

- histogram for each variable

```
columns = data.select_dtypes(include='number').columns
rows = len(columns)
cols = 5
plt.figure()
fig, axs = plt.subplots(rows, cols, figsize=(cols*4, rows*4), squeeze=False)
bins = range(5, 100, 20)
for i in range(len(columns)):
    for j in range(len(bins)):
        axs[i, j].set_title('Histogram for %s'%columns[i])
        axs[i, j].set_xlabel(columns[i])
        axs[i, j].set_ylabel("probability")
        axs[i, j].hist(data[columns[i]].dropna().values, bins[j])
fig.tight_layout()
plt.show()
```

Multi-Variate Analysis

- ***Sparsity***
- ***Correlation analysis***

Correlation analysis

- Correlation analysis

```
import seaborn as sns
import matplotlib.pyplot as plt
fig = plt.figure(figsize=[12, 12])
corr_mtx = data.corr()
sns.heatmap(corr_mtx, xticklabels=corr_mtx.columns,
yticklabels=corr_mtx.columns, annot=True, cmap='Blues')
plt.title('Correlation analysis')
plt.show()
```