KU LEUVEN

LEUVEN STATISTICS RESEARCH CENTRE

# *Modern Data Analytics*

*Exam Project GOZ39a*

## Summary

There is no exam for this course, in stead a project has to be handed in. The project involves a data analytics assignment. The projects are assigned (randomly) to the different teams :

- Water Security
- Uncovering the network
- Politicians and Climate Change
- Heat Waves Impacts
- ECB Bond Purchases
- Cities and Greenhouse Gas Emissions

Each of the projects are described in detail in the appendices of this document.

## Open Research Questions

The research questions are really "open". In other words: **these are just a start**. The same holds for the initial dataset. By no means are you restricted by the suggested data. On the contrary, a good data-scientist is capable to "think out of the box" and to retrieve data from other sources and blend these together.

The teaching team would be sad, if the only thing you achieved in our course, is producing a vanilla Jupyter Notebook where you make us some standard charts. With this course, you should have advanced much further on the learning curve.

## What do we expect from each Team ?

- In a nutshell, **we want you to think as a data-scientist throughout the whole production pipeline:** retrieving & pre-processing data, exception handling, building a model, hyperparameter optimization, etc...
- We expect that you bring the topics explained during the course into practice. Your team should be able to bring value to the data. You can use techniques that were not covered during the course and can bring other python packages into the project.
- Make sure you start from the same python environment, used in the course. Of course you can update packages, install new ones,...
- Make sure that you understand the underlying mathematics in the approach that you use (supervised, unsupervised, nlp, AI,..). A data-scientist is much more than an expert in Sklearn, NLTK, Pytorch,etc...

## What do you need to hand in ?

The deliverables shall consist of:

- a program (python only !)
- a report
- a presentation

## Deliverable

The project is a group-effort which has to result in **Three** deliverables

1. Your Python Code shared on a GitHub account
   This should be either Jupyter Notebook(s) or an App
2. Report (pdf) of maximum 3 pages
3. Each team will be invited for a presentation on-campus or on-line (teams are free to choose).

## Presentation

I have asked the examination committee to schedule 4 days during which the presentations can take place (June). During this presentation the team will present their findings (15 min) and will answer questions (5 minutes).

## Assigned Project

You will receive a mail before March 31st, with the project that has been assigned to your team.

## Delivery Date

**Before** June 1st, 2021
Failure to deliver the report in time, results in a no-pass grade for this course

## Delivery Mechanism

Each team will be assigned an S3-bucket on AWS. There is no need to have an AWS account yourself. The team-captains will receive an invite with access to the S3 buckets.
This bucket has three folders "data","code" and "report".

### Data

In this folder you save all the data you have used (or links to the data) to solve your project

### Code

If you are not able to deliver your python code in a GitHub folder, you can use this folder in the AWS S3-bucket.

### Report

In this S3-bucket you can drop your report (pdf)

## Grading

- 15% on the presentation
- 20% on the written report
- 60% on the work done (project)
- 5% peer evaluation

## Grading Criteria

Below is in bullet-point format a non-exhaustive list of the criteria that we will take into account when we evaluate your work.

### Modeling

- Are you able to reach out to different data-sets outside the assigned dataset ?
- Visualisation
- Code: Style & Organisation of your Python Code.
- Does the code actually work ?
  We should be able to clone your code on github and run it on our computer. Make sure that you use a requirements file to specify the python packages you require.
- Delivery App
  If you deliver an App, the code should be on S3. The app should be deployed.
  Note that after the Easter break, we plan a course on app-building in Python.

### Content of the report

- Your pipeline : from retrieving data to the actual model
- Introduction and problem statement
- Research method & scientific character of the work done
- Argumentation
- Results: discussion / interpretation
- General conclusion
- Coherence / logical composition
- Originality & creativity
- References

### Presentation

- Presentation: used language
- Presentation: content / accessibility
- Presentation: form / composition / timing
- Understanding underlying mathematics
- Answering questions on Python and ML Code

## Failure

There are two ways to obtain a "no-pass" result:

- Your project did not receive a pass grade
- Your team did not hand in a report in time.

In case of a "no-pass" result, students can participate in the August exam. Here new projects will be made available. The August session will be individual, not in team work.

## Q&A

There will be extra Q&As where you can ask questions. The schedules of these Q&A's will be made available on Toledo. If you think that your question is too specific to be covered on the Q&A session, you can drop me an email.

Appendix A

# Water Security

## Introduction

The current climate change scenario predicts that almost half of the world's population will live in areas of high water stress by 2050 with limited access to fresh clean water. Governments, national, and international institutions, as well as water management companies, are looking for solutions that can address this growing global water demand. Cities are encouraged to take action on water security, to build resilience to water scarcity and manage this finite resource for the future.

## Proposed Research Question

Are all cities reporting consistent data ? Are there data gaps in some regions ? Can you predict the likelihood of a water shortage ? What about the population densities in those areas ?

## Initial Data Set

The place to start your enquiry is (data.cdp.net and www.cdp.net). The climate disclosure panel (CDP) is a not-for-profit charity that runs the global disclosure system for investors, companies, cities, states and regions to manage their environmental impacts. The data for cities and regions is free for access & downloading.

Appendix B

# Uncovering the network

Working your way through 13 F filings

## Introduction

An investment manager that has investment discretion over $100 million or more in financial instruments, must report its holdings quarterly on a so-called "Form 13F" with the Securities and Exchange Commission (SEC) of the United States.

Such a 13F report requires disclosure of the name of the institutional investment manager that files the report and a list of the shares held in position as of the end of the calendar quarter for which the report is filed, and the total market value of each such position.

More information can be found on the following link. There are two types of 13 F filings; 13F-HR and 13F-NT. Our interest is in 13F-HR, this report contains the different holdings of an investment manager.

## Proposed Research Question(s)

To what extent can you apply network analysis on the holdings of different investment companies ? How are investment firms connected ? Has the connectivity increased ? Is there a change in the network metrics ?

Note that you can limit the size of the network. By no means are you required to investigate the whole US fund management industry !

## DataSet

The SEC has an on-line portal EDGAR, where investment companies can upload the filings and where investors can download all the uploaded information.

A python package exists :

```
$ pip install sec-edgar-downloader
```

## Information

A list of all the funds (and the corresponding CIK-codes) can be downloaded via a link on the SEC's website. Just note that the CIK has 10 characters. To make this easy, we pre-processed this file and have posted this on S3 (public): https://goz39a.s3.eu-central-1.amazonaws.com/13F/CIK_list.csv

### Example:

The vanguard group is an investment management company. It has a Central Index Key or "CIK" number (0000102909). This is a number given to an individual or a company by the United States Securities and Exchange Commission.

```
from sec_edgar_downloader import Downloader
```

```
loader = Downloader()
loader.get("13F-HR", "0000102909")
```

The code snipped above downloads, the full history of the 13F-HR filings of the vanguard group. When you run this code, you will retrieve for this investment manager the different holdings in a text-file.

In one of the first lines in the report you find the date for which the report is run (YYYYMMDD-format)
"CONFORMED PERIOD OF REPORT:  20190630"

Using an XML-style-formatting, you find all the information grouped by holding. The text-fragment below, gives the typical output for a holding of 214,190 shares in the "Performance Food Group"

```
<infoTable>
    <nameOfIssuer>PERFORMANCE FOOD GROUP CO</nameOfIssuer>
    <titleOfClass>COM</titleOfClass>
    <cusip>71377A103</cusip>
    <value>8574</value>
    <shrsOrPrnAmt>
      <sshPrnamt>214190</sshPrnamt>
      <sshPrnamtType>SH</sshPrnamtType>
    </shrsOrPrnAmt>
    <investmentDiscretion>DFND</investmentDiscretion>
    <otherManager>01</otherManager>
    <votingAuthority>
      <Sole>214190</Sole>
      <Shared>0</Shared>
      <None>0</None>
    </votingAuthority>
  </infoTable>
```

The most important fields are

- **nameOfIssuer** = name of the holding
- **cusip** = a unique identifier for each security (important because the same security can appear several times in a file). The investment manager can either tell the number of shares they have (**sshPrnamtType**=SH) or the market value in USD
- The quantity (number of shares or market value) is stored in the field **sshPrnamt**

Appendix C

# Politicians and Climate Change

## Introduction

Every year since 1947, representatives of UN member states gather at the annual sessions of the United Nations General Assembly. The centrepiece of each session is the General Debate. This is a forum at which leaders and other senior officials deliver statements that present their government's perspective on the major issues in world politics.

## Proposed Research Question

Is there a change , a trend you can observe in the different speeches delivered by our political leaders ? More important are these speeched exhibiting a changing view on climate change ? If not, what are the main topics ? (peace, financial stability, poverty,...)

## Initial Data Set

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/0TJX8Y

Appendix D

# Heat Wave Impacts

## Introduction

In the 1960s, Major cities experienced, on average, about two heat waves per year. In the 2010s, that number rose to more than six heat waves per year. These heat waves are also lasting longer, on average 47 days longer than in 1960. Even under different climate models and emission scenarios, results indicate that extreme heat events worsen.

Heatwaves, or heat and hot weather that can last for several days, can have a significant impact on society, including a rise in heat-related deaths. More than 70 000 people died during the 2003 heatwave in Europe. Workers who are exposed to extreme heat or work in hot environments may be at risk of heat stress. Exposure to extreme heat can result in occupational illnesses and injuries. Heat stress can result in heat stroke, heat exhaustion, heat cramps, or heat rashes. Humidity is an important factor in heat index assessment. When the humidity is high, water does not evaporate as easily and so it becomes difficult for the body to cool off through sweating.

More than 10 nuclear power plants were shut down in France because of cooling-water related issues ! Heatwaves also have an impact on productivity of workers and impact the food supply.

## Proposed Research Questions

- Can you predict the impact on mortality as a function of different parameters (Location, Age, …) caused by heath waves ?
- Can you spot trends in heat waves ?
- Can you predict the impact on the economy (GDP) resulting from heatwaves
- ...

## Initial Data Sets

- https://public.emdat.be
- Heat data https://earthdata.nasa.gov/learn/pathfinders/disasters/extreme-heat
- Temperature in different countries and cities https://sedac.ciesin.columbia.edu/data/set/sdei-global-uhi-2013/data-download
- Population Data https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-count-rev11/data-download
- Heat Stress https://www.cdc.gov/niosh/topics/heatstress
- Poverty Related Data https://sedac.ciesin.columbia.edu/data/sets/browse?facets=theme%3Apoverty
- Heathwaver & Mortality https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3324776/

Appendix E

# ECB Bond Purchases

## Introduction

The European Central Bank has been buying corporate bonds since 2015. This signifies an important cash injection in the European Economy. The ECB started buying assets from commercial banks as part of its non-standard monetary policy measures. These asset purchases, also known as quantitative easing or QE, support economic growth across the euro area and helps Europe to return to inflation levels below, but close to, 2%.

## Proposed Research Question

Has the European Central bank been supporting the green economy when purchasing corporate bonds ?

## Initial Dataset(s)

- https://www.ecb.europa.eu/mopo/implement/app/html/index.en.html#cspp
- https://sdw.ecb.europa.eu
- To find information of a sector / industry a company is belonging to, you might for example want to consult the Refinitiv database (API-available) https://permid.org/

Appendix F

# Cities and Greenhouse Gas Emissions

## Introduction

Cities house half the world's population but represent almost two-thirds of global energy demand and 70% of carbon emissions from the energy sector. Therefore, our cities have a vital role to play in the transition to a sustainable economy.

## Proposed Research Question

- Are all cities reporting consistent data ?
- Are there data gaps in some regions ?
- Are the emission improving ?
- Is there a link with the population density ?
- ...

## Initial Data Set

The place to start your enquiry is (data.cdp.net and www.cdp.net). The climate disclosure panel (CDP) is a not-for-profit charity that runs the global disclosure system for investors, companies, cities, states and regions to manage their environmental impacts. The data for cities and regions is free for access & downloading.

To understand the difference between the different greenhouse gas emissions (GHG), checkout the Greenhouse Gas Protocol (ghgprotocol.org)