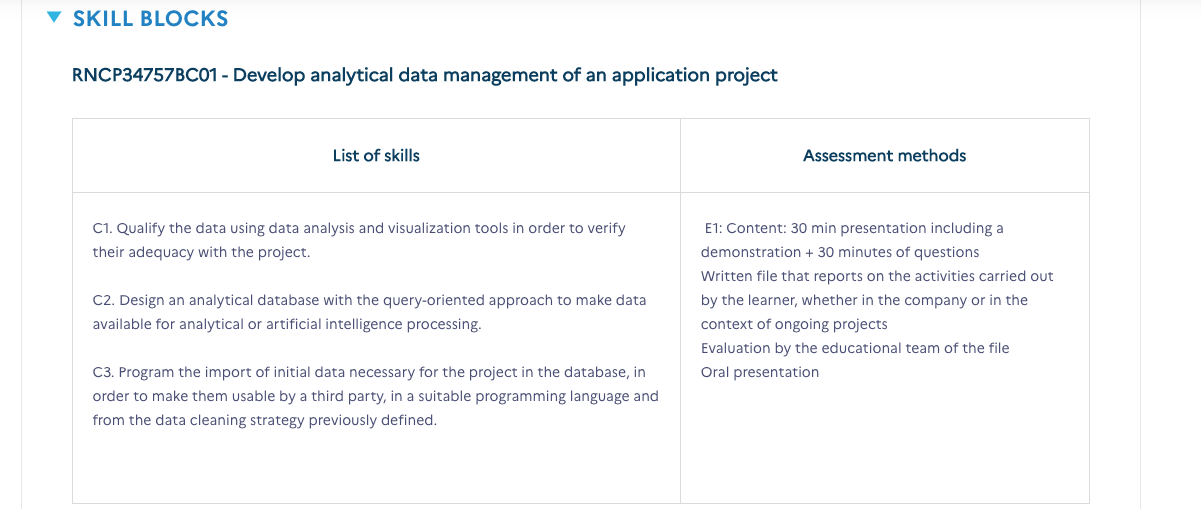
***Certifications professionnelles link to description:*** [**https://www.francecompetences.fr/recherche/rncp/34757/**](https://www.francecompetences.fr/recherche/rncp/34757/)

1. **Explain the use case**
2. **Collect data. Explain why you choose this data, from these data sources**
3. **Clean data and execute exploratory data analysis (+vizualisation)**
4. **Choose the database type (compare several types and explain why)**
5. Create an entity-relationship diagram (at least 4 entities)
6. Create a database (database, tables)
7. Add data to the database
8. Create 5 scripts showing the insights
9. Prepare the 10 pages report
10. Prepare presentation



## **Requirements & Deliverables**

Each student should upload all the project materials to Github.

You should deliver:

* **Planning of your project in Trello/Jira,**
* Code in Python for data collection and cleaning,
* ER model,
* data sources and metadata,
* database script,
* **report (10 pages)**
* slides.

The link to the Github repository and the report in pdf format should be sent on Friday, **06/06/2023 till 13 PM.**

## **Presentation**

The presentation time limit is 30 minutes. You will have 5 minutes for Q&A.

The slides of your presentation must include the content listed below and a demo of your project:

* Title of the project + Name
* Description of your Project (context and objective)
* Planning
* Data sources, data collection
* Data cleaning
* EDA/Visualization
* ERD, Database Schema, Queries
* Highlights
* Model/Machine learning
* Main results
* Demo
* Challenges
* Conclusions

Possible Data sources:

<https://opendata.paris.fr/pages/home/>

<https://data.gov.uk/>

<https://data.iledefrance.fr/pages/home-open-data/>

|  |  |
| --- | --- |
|  | Data Analytics |

ARIF

Acute Respiratory Infections Forecast

IronHacker : Romain Courtois

June, 2023

**Table of content**

**Introduction**

Business Use Case

Acute respiratory infections (ARI)

Acute respiratory infection is a serious infection that prevents normal breathing function. It usually begins as a viral infection in the nose, trachea (windpipe), or lungs. If the infection is not treated, it can spread to the entire respiratory system. Acute respiratory infection prevents the body from getting oxygen and can result in death.

Acute respiratory infections are infectious, which means they can spread from one person to another.

According to the World Health Organization (WHO), acute respiratory infections kill an estimated 2.6 million children annually every year worldwide.

Santé publique France : 40,000 deaths per year and nearly 8 months of life expectancy lost due to exposure to fine particles

ARIs are caused by various respiratory viruses including SARS-CoV-2 (Covid-19), influenza viruses and other respiratory viruses such as RSV, rhinovirus, or metapneumovirus. The purpose of ARI surveillance is to monitor epidemics caused by these viruses.

Certains facteurs semblent favoriser la survenue de telles pathologies :

Le sexe masculin ;

L’âge (le risque de décès est plus important chez les nourrissons âgés de 1 à 3 mois) ;

La prématurité ;

Le climat et la saison (les infections se développent essentiellement par temps froid et pluvieux) ;

La pollution ;

La promiscuité ;

L’état nutritionnel ;

L’état immunologique ;

Un bas niveau d’éducation ;

Un bas niveau socio-économique du pays.

Goals:

Main goal : Create a tool to predict forecast of ARI number cases using weather forecast

Alternate goal: Perform an analysis of correlation between ARIs, weather and pollution.

Plan

**Data and data source**

# ARI\_Dataset : [French Weekly epidemiological surveillance](https://public.opendatasoft.com/explore/dataset/healthref-france-sentinelles-weekly/information/)

The Sentinelles network (INSERM/Sorbonne University, <https://www.sentiweb.fr>) collects a set of data allowing the epidemiological progress of certain diseases to be monitored with a weekly frequency.

A picture containing text, map, atlas

Description automatically generated 

# POLU\_Dataset : ["Real-time" data from measurements of concentrations of regulated air pollutants](https://www.data.gouv.fr/fr/datasets/donnees-temps-reel-de-mesure-des-concentrations-de-polluants-atmospheriques-reglementes-1/)

A close-up of a logo

Description automatically generated with low confidence

A picture containing screenshot, text, map

Description automatically generated

The data available through these processes is hourly data from automatic analyzers. The concentrations of the following atmospheric pollutants are measured:

Ozone (O3)

Nitrogen dioxide (NO2)

Sulfur dioxide (SO2)

Particles with a diameter of less than 10 µm (PM10)

Particles with a diameter of less than 2.5 µm (PM2.5)

Carbon monoxide (CO)

# SYNOP\_Dataset : [Historical meteorological observation France](https://public.opendatasoft.com/explore/dataset/donnees-synop-essentielles-omm/information/)

Observation data from international surface observation reports (SYNOP) circulating on the Global Telecommunication System (GTS) of the World Meteorological Organization (WMO). Atmospheric parameters measured (temperature, humidity, wind direction and force, atmospheric pressure, amount of precipitation) or observed (weather sensitive, description of clouds, visibility) from the earth's surface.

A picture containing map

Description automatically generated

**Data collection**

ARI\_Dataset Collection

healthref-france-sentinelles-weekly.csv 5.4 Mo 72306 rows 12 columns



For ARI : from 23/03/2020 to 08/05/2023 2145 rows

1 line of data per region per week : This set my geographical granularity to French regions, and my time granularity to weeks.

During integration, other dataset will be aggregated by region and weeks to reduce the size manipulated.

POLU\_Dataset Collection

A screenshot of a computer

Description automatically generated with low confidence

From 01/01/2021 to now

1 test file downloaded : FR\_E2\_2021-01-01.csv

10,9 Mo 50257 rows, 24 columns



We need a python script to download 1 file per day and aggregate it per pollutant per station.

SYNOP\_Dataset Collection

donnees-synop-essentielles-omm.csv 1.1 Go

2 247 960 rows 57 columns

From 05/01/2010 to 26/04/2023

**Data cleaning and Exploratory data analysis**

**Data base type selection**

**Entities. ERD**

**Challenges**

Big size of data

Mapping geographical data

Take account of seasonality, COVID impact