# **Data Science Project ITS 2022-23**

**Requests:**

The construction company Crack contacted us because they want to build a new research center and they requested an analysis of the areas in Milan, Paris and Berlin.

The main points of the analysis are:

* The physical network within the city(how it's connected)
* The best internet infrastructures
* The available services
* The amount of petty crimes in the area
* The cost of local labor
* The best investment by 2030

We started working on a POC for the job.

**Fields(Data scavenging):**

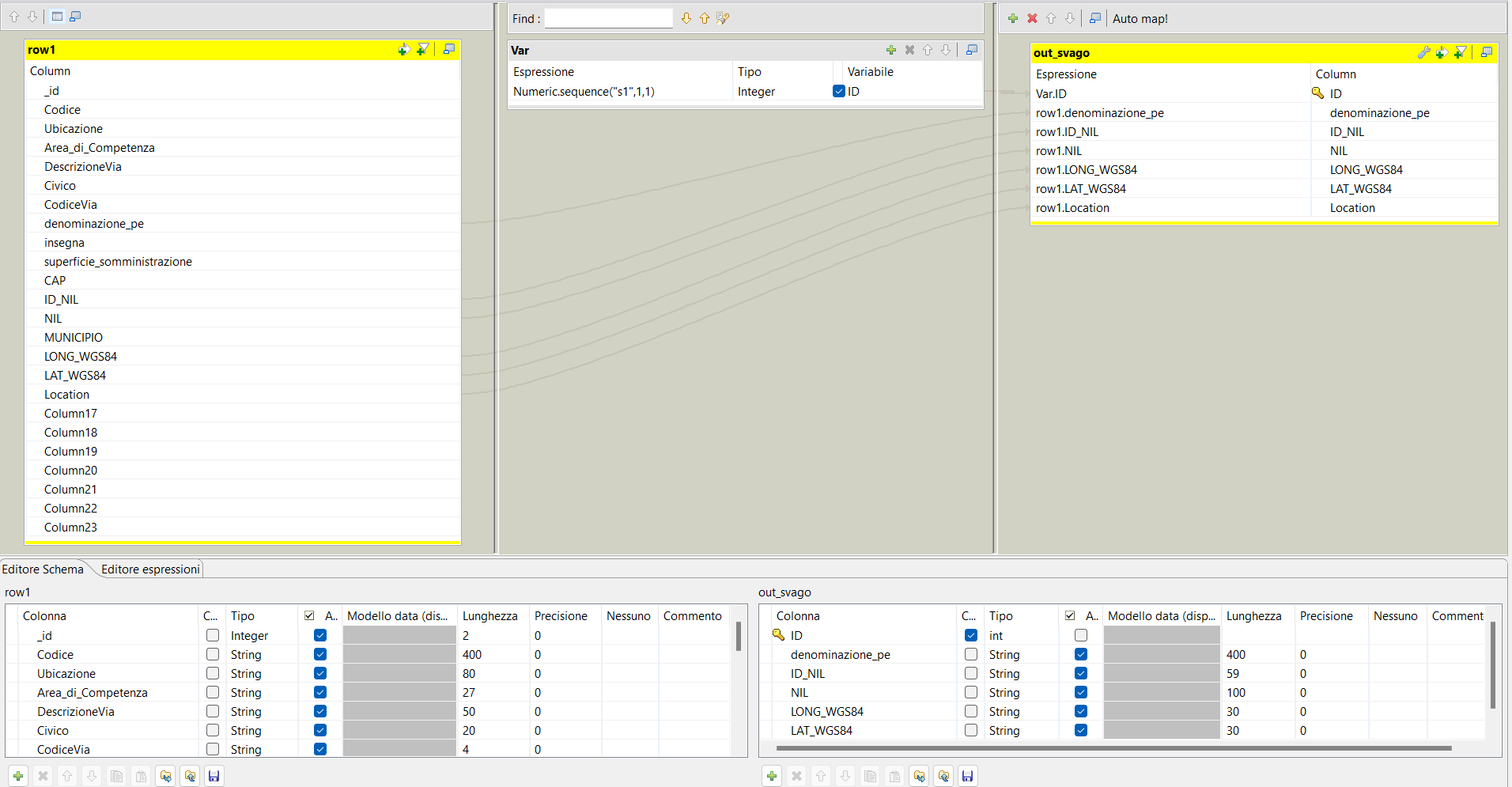
With a quick research we found the following open data form a few trusted resources:

* **Milan:**
  + Public transport: location of railway stations
    - Link: https://dati.comune.milano.it/dataset/ds80\_infogeo\_stazioni\_ferroviarie\_localizzazione\_
  + ATM - Urban surface line routes
    - Link: https://dati.comune.milano.it/dataset/ds538\_atm-percorsi-linee-di-superficie-urbane
  + BikeMi Stations:
    - Link: https://dati.comune.milano.it/dataset/ds65\_infogeo\_aree\_sosta\_bike\_sharing\_localizzazione\_
  + Public Station’s Parking locations:
    - Link: https://dati.comune.milano.it/dataset/ds45\_infogeo\_parcheggi\_interscambio\_localizzazione\_
  + Location of public car parks:
    - Link: https://dati.comune.milano.it/dataset/ds342-trafficotrasporti-parcheggi-pubblici-localizzazione
  + ATM - Metro routes:
    - Link: https://dati.comune.milano.it/dataset/ds539\_atm-percorsi-linee-metropolitane
  + ATM - Composition of metro lines:
    - Link: https://dati.comune.milano.it/dataset/ds533\_atm-composizione-percorsi-linee-metropolitane
  + ATM - Metro stops:
    - Link: https://dati.comune.milano.it/dataset/ds535\_atm-fermate-linee-metropolitane
  + Bicycle parking:
    - Link: https://dati.comune.milano.it/dataset/ds670-trasporto-pubblico-sosta-biciclette
  + Nuclei d'Identità Locale (NIL) (neighborhoods):
    - Link: https://dati.comune.milano.it/it/dataset/ds964-nil-vigenti-pgt-2030
  + Business for leisure:
    - Link: https://dati.comune.milano.it/it/dataset/ds58\_economia\_pubblici\_esercizi\_in\_piano
  + Public Hospitals:
    - Link: https://www.dati.lombardia.it/Sanit-/Elenco-strutture-di-ricovero-e-cura-pubbliche/jjbx-ccsh/data
  + Preschools:
    - Link: https://dati.comune.milano.it/dataset/ds671-infogeo-scuole-infanzia-localizzazione
  + Primary schools:
    - Link: https://dati.comune.milano.it/dataset/ds40-infogeo-scuole-primarie-localizzazione
  + Secondary schools:
    - Link: https://dati.comune.milano.it/dataset/ds46-infogeo-scuole-secondarie-i-grado-localizzazione
  + Construction costs (base 2015=100):
    - Link: http://dati.istat.it/Index.aspx?DataSetCode=DCSC\_FABBRESID\_1&Lang=it
  + Crimes in Milan:
    - Link: https://dati.comune.milano.it/dataset/ds564-reati-denunciati-all-autorita-giudiziaria-dalla-forze-di-polizia
  + Networking Milan:
    - Link: https://maps.agcom.it
* **Paris:**
  + Metro access and public parking:
    - Link: https://opendata.paris.fr/explore/dataset/plan-de-voirie-acces-pietons-metro-et-parkings/map/?disjunctive.num\_pave&disjunctive.lib\_level&location=10,48.81952,2.13203&basemap=jawg.streets
  + Metro access, Railway track, Bus stop post, Ticket dispenser, Tram:
    - Link: https://opendata.paris.fr/explore/dataset/plan-de-voirie-mobiliers-urbains-abris-voyageurs-points-darrets-bus/information/?disjunctive.num\_pave&disjunctive.lib\_level
  + Cycle paths:
    - Link: https://opendata.paris.fr/explore/dataset/plan-de-voirie-pistes-cyclables-et-couloirs-de-bus/information/?disjunctive.num\_pave&disjunctive.lib\_classe
  + Taxi call points:
    - Link: https://opendata.paris.fr/explore/dataset/bornes-dappel-taxi/information/?disjunctive.code\_postal
  + Parking on public roads:
    - Link: https://opendata.paris.fr/explore/dataset/stationnement-sur-voie-publique-emprises/information/?disjunctive.regpri&disjunctive.regpar&disjunctive.typsta&disjunctive.arrond&disjunctive.zoneres&disjunctive.locsta&disjunctive.parite&disjunctive.signhor&disjunctive.signvert&disjunctive.confsign&disjunctive.typemob
  + Mobiblib car parks, the car-sharing service in Paris:
    - Link: https://opendata.paris.fr/explore/dataset/liste-des-stations-de-services-de-vehicules/information/?disjunctive.code\_post&disjunctive.motorisation&disjunctive.fonctionnelle
  + Buildable Land:
    - Link: https://opendata.paris.fr/explore/dataset/emprise-batie-et-non-batie/information/
  + Preschools:
    - Link: https://opendata.paris.fr/explore/dataset/secteurs-scolaires-ecoles-elementaires/information/?disjunctive.id\_projet&disjunctive.zone\_commune&disjunctive.annee\_scol
  + Primary schools:
    - Link: https://opendata.paris.fr/explore/dataset/secteurs-scolaires-ecoles-elementaires/information/?disjunctive.id\_projet&disjunctive.zone\_commune&disjunctive.annee\_scol
  + Secondary schools:
    - Link: https://opendata.paris.fr/explore/dataset/secteurs-scolaires-colleges/information/?disjunctive.id\_projet&disjunctive.zone\_commune&disjunctive.annee\_scol
  + Crimes in Paris:
    - Link: https://www.data.gouv.fr/fr/datasets/bases-communale-et-departementale-des-principaux-indicateurs-des-crimes-et-delits-enregistres-par-la-police-et-la-gendarmerie-nationales/

**Ingestion and ETL(Talend):**

We decided to structure and store the data we found on a MySQL database hosted on AWS cluster.

In order to load the data (from csv files) we used Talend, an open source application for ETL, where we mapped every column from each file and loaded only the ones we previously deemed useful.

In addition we also added a variable column, in the tMap object, using the function “Numeric.sequence(‘s1’, 1, 1)”, wich creates a sequence of integers starting with 1 and incremented by 1 each row. We used this ID columns as primary keys for each table in the DB.

**Data mining(Jupiter Notebooks):**

For the POC we decided base our work on 2 Jupiter Notebooks, one for Milan and one for Paris, using Google Collab free package, we went thought similar processes for both. We started with:

* + **Importing Data:** We used a python api to import the tables we needed from our database into pandas dataframes
  + **Data Wrangling:** The goal of this step is to have all of the dataframes with the same 3 columns: latitude, longitude and the type of point they’re indicating. So for this we used only the dataset that contained coordinates, added the type manually if it wasn’t already present.
  + **Encoding:** In order to feed the data to the clustering algorithm it as to be all numerical, so firstly we dropped any row with missing coordinates and then converted those to floats, they were treated as strings up until this point, than we used the pandas function getDummies() to encode the Type column.
  + **Modelling:** We used a hierarchical clustering model, AgglomerativeClustering from the sklearn library, to decide on the number of cluster we used the elbow method, for Paris we had to trim the datasets because the free machine in Google Collab couldn’t handle the data all at once, could also be improve using different machine learning library.
  + **Exporting:** We then exported the clusterd points ready to be visualized with power BI.