Railway stations in Lombardy

In this project I will work with two datasets:

- `railwaystats` : data about railway stations in Lombardy, downloaded from the Open Data portal of regione Lombardia.
- `population` : population by municipality of Lombardy, downloaded from the Geoportal of regione Lombardia

I will then join them geospatially and study them.

Data cleaning and preparation

Firstly I import the two shape files as layers in `QGIS` . Then I create a connection between QGIS and `PgAdmin` to import the relative tables in the second one.

Now in PgAdmin I clean and pre-process the data.

Railway data

First let's visualize the main variables of the imported data

```
# Check the imported data of railwaystats
SELECT codestaz, reg, prov_, stazione, linea_fisi, saliti7_9, discesi7_9, corse7_9, sa
liti24h, corse24h,saliti_s, saliti_r, saliti_re, corse_s, corse_r, corse_re, anno
FROM railwaystats
ORDER BY stazione;
```

codstaz character	prov_ charac	stazione character varying(254)	linea_fisi character varying(254)	saliti7_9 numeric	discesi7_9 numeric	corse7_9 numeric	saliti24h numeric	corse24h numeric	saliti_s numeric	saliti_r numeric	saliti_re numeric	corse_s numeric	corse_r numeric	corse_re numeric	anno numeric
215	BS	COLOGNE	LECCO-BRESCIA	63	22	3	206	31	0	206	0	0	31	0	2016
291	PV	LAMBRINIA	PAVIA-CASALPUST.	32	3	5	77	30	0	77	0	0	30	0	2015
120	CO	ALBATE-CAMERLATA	CHIASSO-MILANO	274	103	16	654	117	607	47	0	94	23	0	2017
196	LC	OSNAGO	TIRANO-MILANO Mz-Lc	406	51	7	785	62	785	0	0	62	0	0	2015
266	LO	S.STEFANO LODIGIANO	MILANO-BOLOGNA	41	5	4	86	36	0	86	0	0	36	0	2017
335	MN	MANTOVA	CODOGNO-MANTOVA	110	236	4	966	28	0	99	867	0	7	21	2019
2797	CO	Pontelambro-Castelmarte	Seveso - Asso	99	20	6	238	40	0	238	0	0	40	0	2018
246	BS	OSPITALETTO TRAVAGLIATO	MILANO-VENEZIA	131	41	8	385	60	0	385	0	0	60	0	2015
309	CR	PIADENA	CODOGNO-MANTOVA	269	106	11	1011	66	0	402	609	0	45	21	2019
265	LO	CODOGNO	MILANO-BOLOGNA	921	400	15	2370	106	0	1434	936	0	85	21	2016

We observe that there are some problems with some variables, so we fix them

```
# data cleaning
# Delete the rows which refer to a region different than Lombnardy or the station is
"-"
DELETE FROM railwaystats
WHERE stazione = '\"-\"' OR reg <> 'LO';
```

```
# Notice that all instances of the attribute "calendario" are "inverno".
SELECT calendario
FROM railwaystats
GROUP BY calendario;
```

	calendario
▶	Invernale

```
# Therefore we can drop the variable since it is not bringing additional information
(as well as the variable "reg" wich now only contains LO for Lombardy)
ALTER TABLE railwaystats
DROP COLUMN calendario, reg;
```

Population data

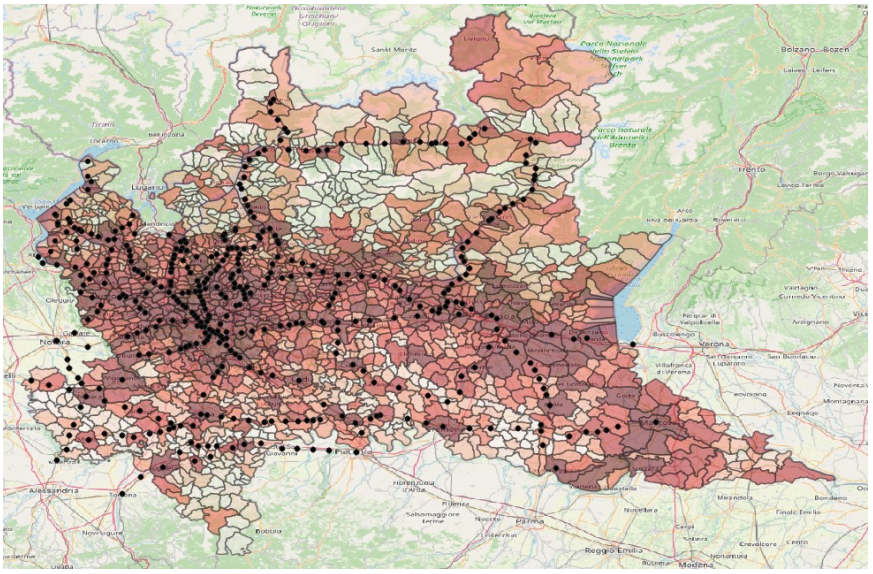
Check the imported data. It has as a geometry a multi-polygon representing the area of each province, and then the population and surface of the area.

```
SELECT *
FROM population;
```

	id integer	geom geometry(MultiPolygon,4326	nome_com_2 character varying(100)	istat_2017 bigint	nome_pro character varying(100)	pop_2017 double pre	pop_2014 double pre	variaz_17_ double preci	shape_area double precis	shape_len double preci
1	1	0106000020E6100000010000	GERMIGNAGA	12076	VARESE	3886	3857	29	79303.86534	13955.90115
2	2	0106000020E6100000010000	USMATE VELATE	108044	MONZA E DELLA BRIANZA	10211	10259	-48	71983.44342	20458.75157
3	3	0106000020E6100000010000	VAREDO	108045	MONZA E DELLA BRIANZA	13335	13160	175	93902.74165	13123.71569
4	1055	0106000020E6100000010000	CHIURO	14020	SONDRIO	2553	2518	35	761295.6905	44729.49619
5	4	0106000020E6100000010000	VEDANO AL LAMBRO	108046	MONZA E DELLA BRIANZA	7609	7535	74	79747.50305	7409.711881
6	5	0106000020E6100000010000	VEDUGGIO CON COLZANO	108047	MONZA E DELLA BRIANZA	4356	4443	-87	63683.05197	13085.77848
7	6	0106000020E6100000010000	VIMERCATE	108050	MONZA E DELLA BRIANZA	26062	25839	223	0618891.934	37459.08412
8	7	0106000020E6100000010000	GEMONIO	12074	VARESE	2871	2880	-9	81499.23713	10308.15600
9	14	0106000020E6100000010000	LAZZATE	108025	MONZA E DELLA BRIANZA	7803	7787	16	37227.74137	11002.03695
10	1080	0106000020E6100000010000	ROBECCHETTO CON INDUNO	15183	MILANO	4885	4886	-1	969426.1284	23392.98265

PostGIS visualizations

Going back in QGIS where we have initially imported the data, we can visualize individually the two geospatial layers: the red areas are given by the `population` shape file, and they represent the areas of the different municipalities, colored with increasing red according to the variable `pop_2017`. The black dots on the other hand are all the railway stations of Lombardy taken from the `railwaystats` file.



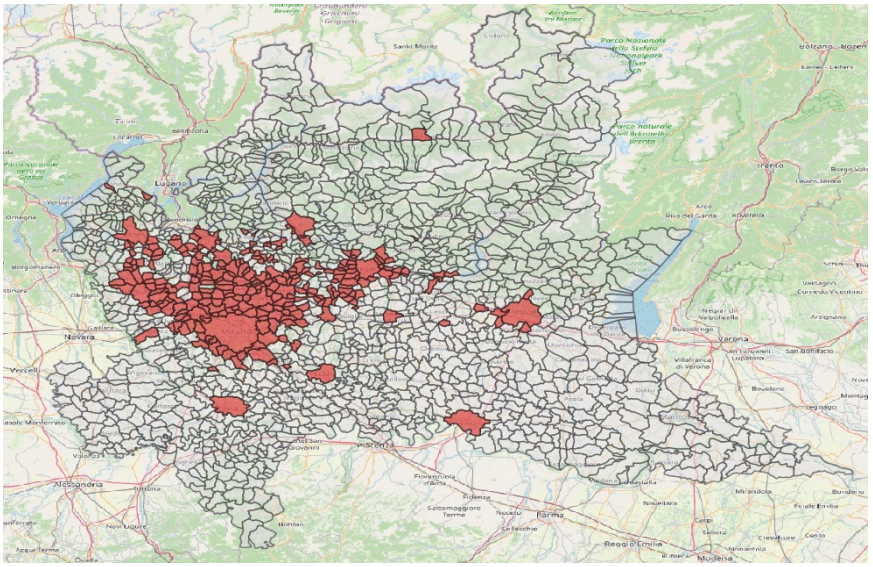
Heatmap of Lombardy's population of 2017 by municipality and representation of the railway stations.

Do high population density areas have a station inside of them?

To answer this question, we firstly create a table `pop_density` which selects only the municipalities of `population` which have a population density in 2017 bigger than 1000 inhabitants per square km.

```
CREATE TABLE pop_density AS
SELECT nome_com_2 AS comune, geom, pop_2017/(St_Area(ST_Transform(geom,32632))/100000
0) AS popdensity
FROM population
WHERE pop_2017/(St_Area(ST_Transform(geom,32632))/1000000) > 1000;
```

In the following visualization we can appreciate how the most highly populated areas are all the province capitals of the region (the spot areas) except for Mantova, but the main highly densely populated area is the one around the city of Milan, where the majority of the municipalities are all together red in the map.

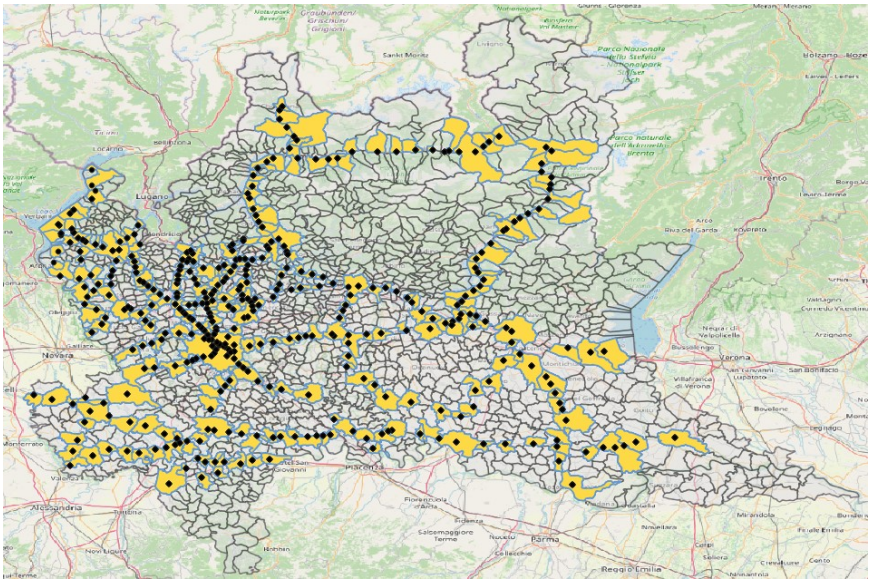


Municipalities with more than 1000 inhabitants per square km

We can perform a **geospatial join** between the two tables `railwaystats` and `population` where the geometries of the two tables intersects. What we get as an output the table `railway_pop` with a one-to-many relationship: for every municipality that contains at least a station, we get all the stations that it contains. In this way, we are also able to remove the stations that are not located in Lombardy.

```
CREATE TABLE railway_pop AS
SELECT stazione,codstaz, nome_com_2, prov_ , nome_pro, pop_2017, pop_2014, variaz_17_,
linea_fisi, saliti7_9, discesi7_9,
corse7_9, saliti24h, corse24h, saliti_s, saliti_r, saliti_re, corse_s, corse_r,
corse_re, anno, population.geom AS geom_p,  railwaystats.geom AS geom_r
FROM population, railwaystats
WHERE ST_Intersects(population.geom, railwaystats.geom);
```

We can see the matched municipalities (yellow) and railway stations (black dots) in the representation underneath.

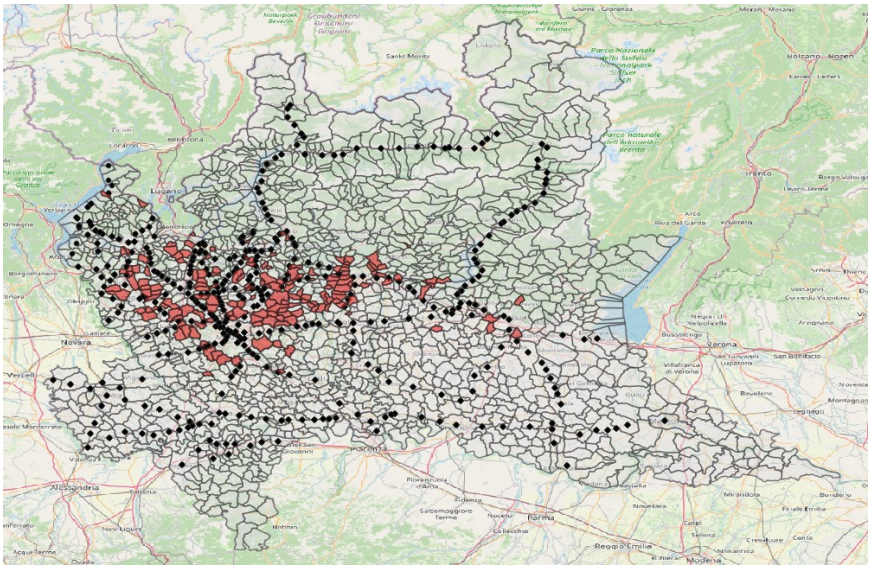


Municipalities which contain at least one station, and their relative stations.

Now we can exploit this new table and outer join it to the `pop_density` table, to get the highly populated municipalities that don't contain a railway station.

```
CREATE TABLE not_railway AS
SELECT comune, geom, popdensity
FROM pop_density
LEFT JOIN railway_pop ON comune = railway_pop.nome_com_2
WHERE railway_pop.nome_com_2 IS NULL;
```

Specifically we can now observe that the main area with high population and without train stations in the municipalities



Highly densely populated areas without a railway station inside