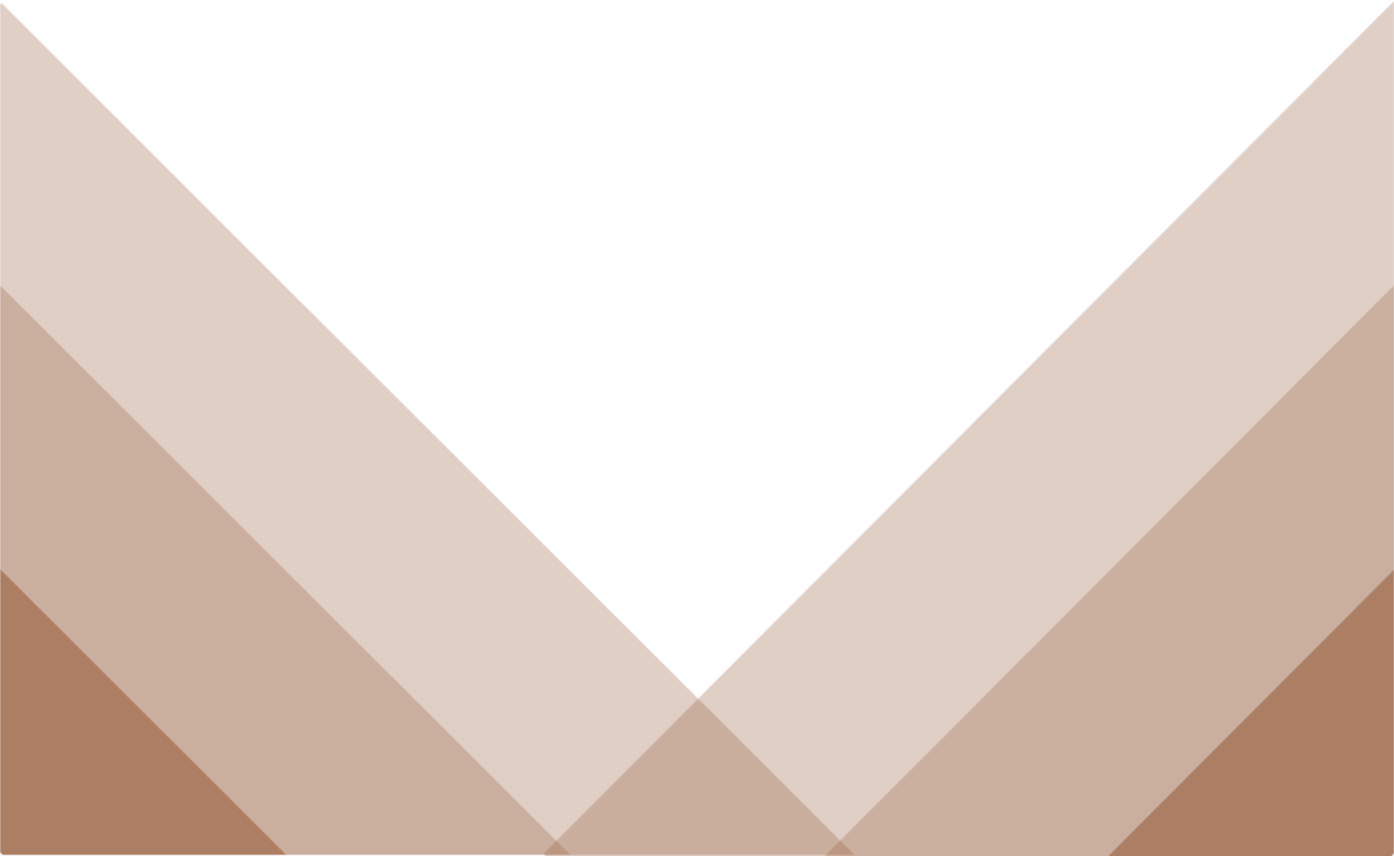


# SAÉ 2.04

## Working with a database

*Exploitation d'une base de données*



# Table des matières

<b>Introduction .....</b>	<b>2</b>
<b>Our group.....</b>	<b>2</b>
<b>Reading data .....</b>	<b>3</b>
Our data: migration statistics in Barcelona.....	3
What is it?.....	3
Why is it relevant? .....	3
Making a database from the aforementioned data .....	4
How is the data organised? .....	4
Entity-relationship model .....	5
Converting the existing data to the format of our database .....	6
<b>How is Barcelona organized? .....</b>	<b>7</b>
<b>Working with data.....</b>	<b>8</b>
Python functions to interact with our data .....	8
Creating diagrams to illustrate relevant information .....	10
Immigration and emigration per district .....	11
Immigrants' nationalities for each relevant district, each year .....	14
EIXAMPLE .....	14
LES CORTS .....	15
CIUTAT VELLA.....	16
SARRIA-SANT GERVASI .....	17
Immigration and emigration by age range for each relevant district, each year .....	18
EIXAMPLE .....	18
LES CORTS .....	19
CIUTAT VELLA.....	20
SARRIA-SANT GERVASI .....	21
Making estimates from existing data.....	22
<b>Conclusion.....</b>	<b>22</b>

## Introduction

This project consists in analyzing CSV spreadsheets in order to create a database that can hold all of their data with as little redundancy as possible.

Then, we can work on this database, and use its data to create diagrams that illustrate relevant information, like the evolution of specific values, or comparisons between different elements.

## Our group

We were two students on this project, and we split responsibilities as follows:

- Hugo: Researches, Diagrams
- Matthieu: Researches, CSV Conversion scripts, Report

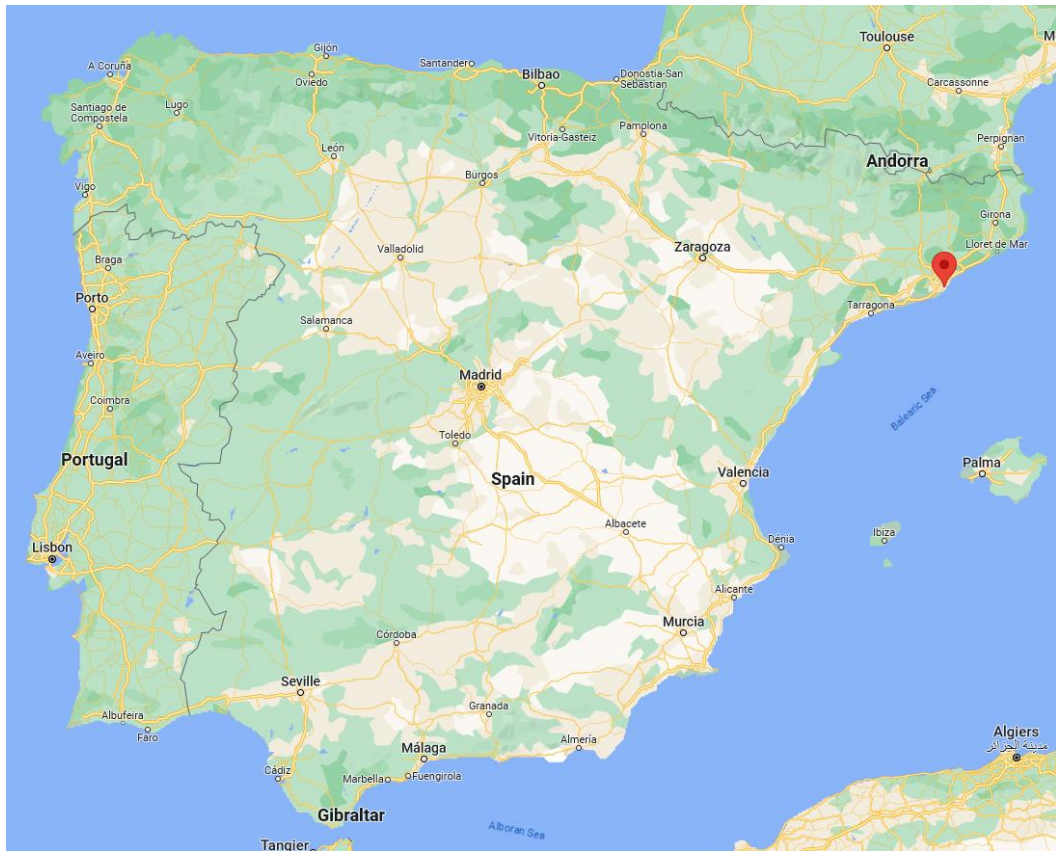
Do note that this task distribution is not 100% accurate: we both often helped each other. It's a group project, after all.

# Reading data

## Our data: migration statistics in Barcelona

### What is it?

Barcelona is a city on the coast of northeastern Spain. It is the capital and largest city of the autonomous community of Catalonia.



The city of Barcelona is home to over 1,620,000 people. As all big cities, a lot of people arrive and leave the city each year.

Therefore, they record the amount both the immigration and emigration in the city and publish statistics on it.

They sort these statistics by categories, like age range, or nationality.

### Why is it relevant?

The point of this study is to represent Barcelona's migration statistics in a way that makes it easy to understand, and easy to interpret.






Our data goes from 2013 up to 2017, and by looking at the evolution of the many figures it has, we can predict future data, or figure out missing past data.

For the city of Barcelona, it might be quite useful to be able to predict future immigration and emigration streams, so they can prepare the necessary infrastructure and resources to make sure all citizens live in decent conditions.

## Making a database from the aforementioned data

### How is the data organised?

The data is split across five CSV files :

-  immigrants\_by\_nationality.csv
-  immigrants\_emigrants\_by\_age.csv
-  immigrants\_emigrants\_by\_destination.csv
-  immigrants\_emigrants\_by\_destination2.csv
-  immigrants\_emigrants\_by\_sex.csv

They all use the following structure :

	A	B	C	D	E	F	G
1	Year	District Code	District Name	Neighborhood Code	Neighborhood Name	Nationality	Number
2	2017	1	Ciutat Vella	1	el Raval	Spain	1109
3	2017	1	Ciutat Vella	2	el Barri Gòtic	Spain	482
4	2017	1	Ciutat Vella	3	la Barceloneta	Spain	414
5	2017	1	Ciutat Vella	4	Sant Pere, Santa Caterina i la Ribera	Spain	537
6	2017	2	Eixample	5	el Fort Pienc	Spain	663
7	2017	2	Eixample	6	la Sagrada Família	Spain	1181
8	2017	2	Eixample	7	la Dreta de l'Eixample	Spain	1063
9	2017	2	Eixample	8	l'Antiga Esquerra de l'Eixample	Spain	1177
10	2017	2	Eixample	9	la Nova Esquerra de l'Eixample	Spain	1593
11	2017	2	Eixample	10	Sant Antoni	Spain	883
12	2017	3	Sants-Montjuïc	11	el Poble Sec	Spain	826
13	2017	3	Sants-Montjuïc	12	la Marina del Prat Vermell	Spain	23

*First few rows from immigrants\_by\_nationality.csv*

Each row represents the recording of one statistic for a specific neighborhood located in a given district.

For instance, the first row claims that in 2017, "el Raval", a neighborhood located in the district of Ciutat Vella, received 1109 spanish immigrants.

Most files include not only immigration data, but also emigration data :

	A	B	C	D	E	F	G	H
1	Year	District Code	District Name	Neighborhood Code	Neighborhood Name	Age	Immigrants	Emigrants
2	2017	1	Ciutat Vella	1	el Raval	0-4	154	108
3	2017	1	Ciutat Vella	2	el Barri Gòtic	0-4	58	33
4	2017	1	Ciutat Vella	3	la Barceloneta	0-4	38	37
5	2017	1	Ciutat Vella	4	Sant Pere, Santa Caterina i la Ribera	0-4	56	55
6	2017	2	Eixample	5	el Fort Pienc	0-4	79	60
7	2017	2	Eixample	6	la Sagrada Família	0-4	111	95
8	2017	2	Eixample	7	la Dreta de l'Eixample	0-4	121	78
9	2017	2	Eixample	8	l'Antiga Esquerra de l'Eixample	0-4	97	63
10	2017	2	Eixample	9	la Nova Esquerra de l'Eixample	0-4	123	130
11	2017	2	Eixample	10	Sant Antoni	0-4	57	80
12	2017	3	Sants-Montjuïc	11	el Poble Sec	0-4	80	99
13	2017	3	Sants-Montjuïc	12	la Marina del Prat Vermell	0-4	1	2

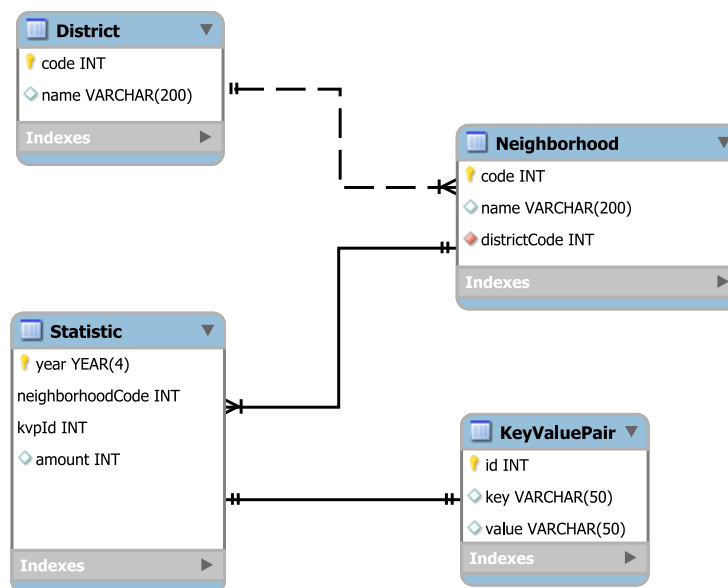
*First few rows from immigrants\_emigrants\_by\_age.csv*

The structure is the same, but the number column was split in half : one part for immigrants, and one part for emigrants.

## Entity-relationship model

In order to be added to a usable database, we needed to rearrange all this data in tables that avoid redundancy as much as possible.

Therefore, we thought about the following model :



First off, districts and neighborhoods are identified by their IDs in a table. This way, we don't have to repeat their names in each statistics, only the ID of the required neighborhood.

Then, we organise each statistic in a table that stores its year, neighborhood ID, and the statistic's type and amount.

The statistic's type is stored as a Key-Value pair.

It might not be very clear, so here is an example :

For the following line from `immigrants_by_nationality.csv`:

	A	B	C	D	E	F	G
1	Year	District Code	District Name	Neighborhood Code	Neighborhood Name	Nationality	Number
2	2017	1	Ciutat Vella	1	el Raval	Spain	1109

We'd have :

- In the District table :

	A	B
1	District Code	District Name
2	1	Ciutat Vella

- In the Neighborhood table :

	A	B	C
1	District Code	Neighborhood Code	Neighborhood Name
2	1	1	el Raval

- In the KeyValuePair table :

	A	B	C
1	ID	Key	Value
2	0	Nationality	Spain

- In the Statistic table :

	A	B	C	D
1	Year	Neighborhood Code	KVPID	Amount
2	2017	1	0	1109

It might look like we are uselessly complicating things ; and in a way, yes, we are. Gathering clear data from such a model won't be as straightforward as the original CSV.

But on a redundancy perspective, our model is much better : say you want to change the name of a district, you'd have to go through each row to edit it, while in our model it only takes one edit.

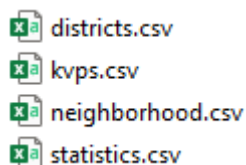
Same goes for statistics types.

### Converting the existing data to the format of our database

In order to import the existing data into a database that follows our entity-relationship model, we need to write a script that would parse all of the CSV files, and output new ones that follow the structure of our model.

Therefore, we have written a Python script that does exactly that. You will find it linked to this report, as `ConvertTables.py`. Make sure you put the source CSV files in the same folder, and have the pandas module installed on your Python setup.

Four new CSV files are generated :



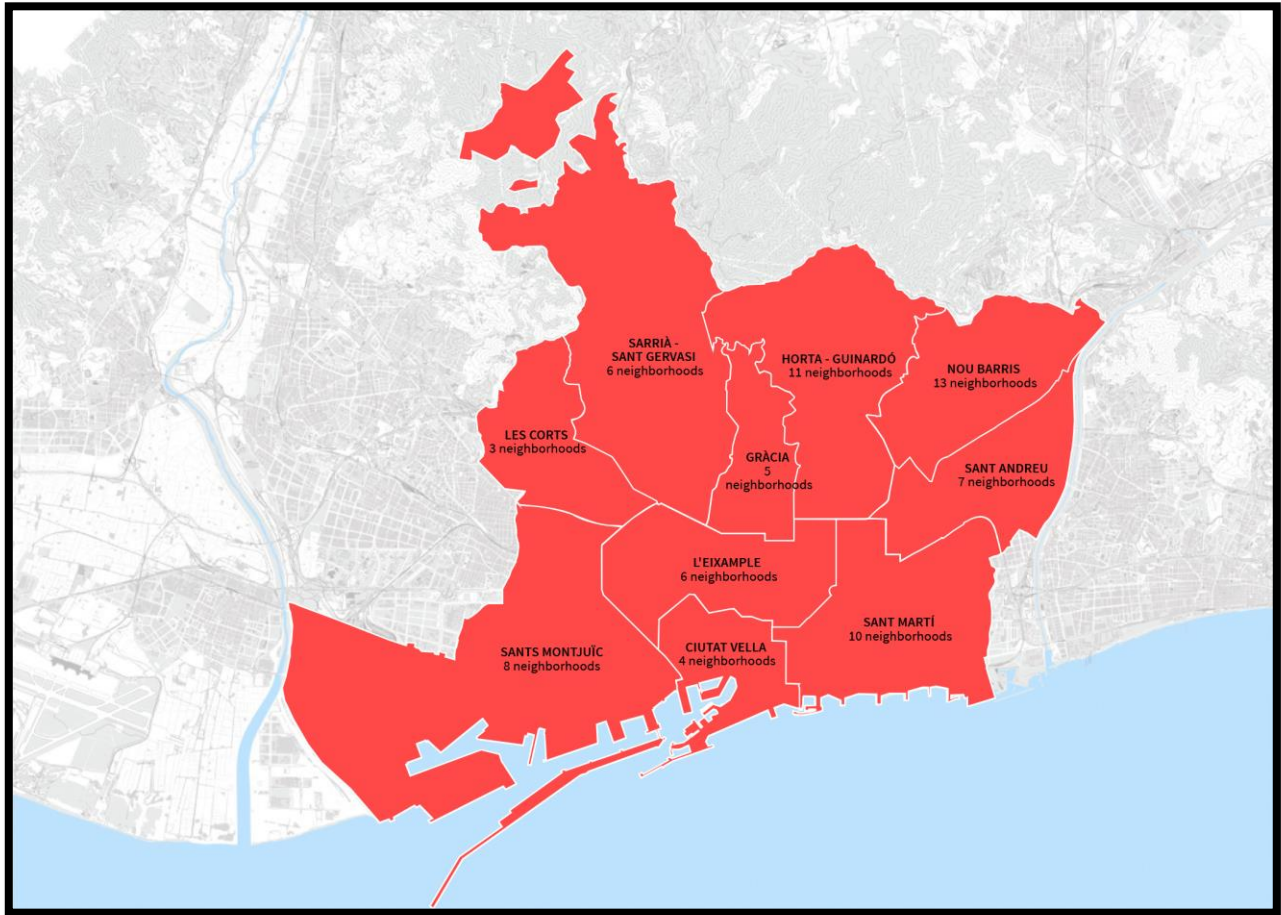
In terms of size, the five original CSV files all add up to 2.6 MB, whereas the four newly generate ones add up to 682.6 kB.

Now, we can import our CSV files into any MySQL-based database using the following syntax :

```
LOAD DATA INFILE '../bin/districts.csv' INTO TABLE District
FIELDS TERMINATED BY ";" LINES TERMINATED BY "\r\n" IGNORE 1 ROWS;
```

*Replace "District" and "districts.csv" respectively with the table name and CSV filename corresponding to the table to import.*

## How is Barcelona organized?



Barcelona is divided into ten districts, themselves divided into numerous neighborhoods. We are going to study both the immigration and emigration on each of these districts, from 2015 to 2017.



## Working with data

We have chosen Python to parse our data and to display it as diagrams easily.

We use the matplotlib, seaborn, pandas and numpy modules. The former two are for displaying diagrams from data, pandas is to parse CSV files as dataframes, and numpy is a math utility module.

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np
```

### Python functions to interact with our data

In order to simplify our work, we have written functions that interact with dataframes that hold our data and return relevant and usable data from input parameters that identify the exact kind of data that we want.

First, we parse our CSV files as dataframes:

```
# Reading all spreadsheets
districts = pd.read_csv('districts.csv')
neighborhood = pd.read_csv('neighborhood.csv')
kvps = pd.read_csv('kvps.csv')
statistics = pd.read_csv('statistics.csv')
```

Then, we write our many functions. Here is one of them:

```
# This function takes a district number as an input and returns its name
def GetDistrictNameFromNumber(numero):
    f = districts.loc[districts["District Code"] == numero, ["District Name"]]

    if len(f) == 0:
        return None

    return f.iloc[0, 0]
```

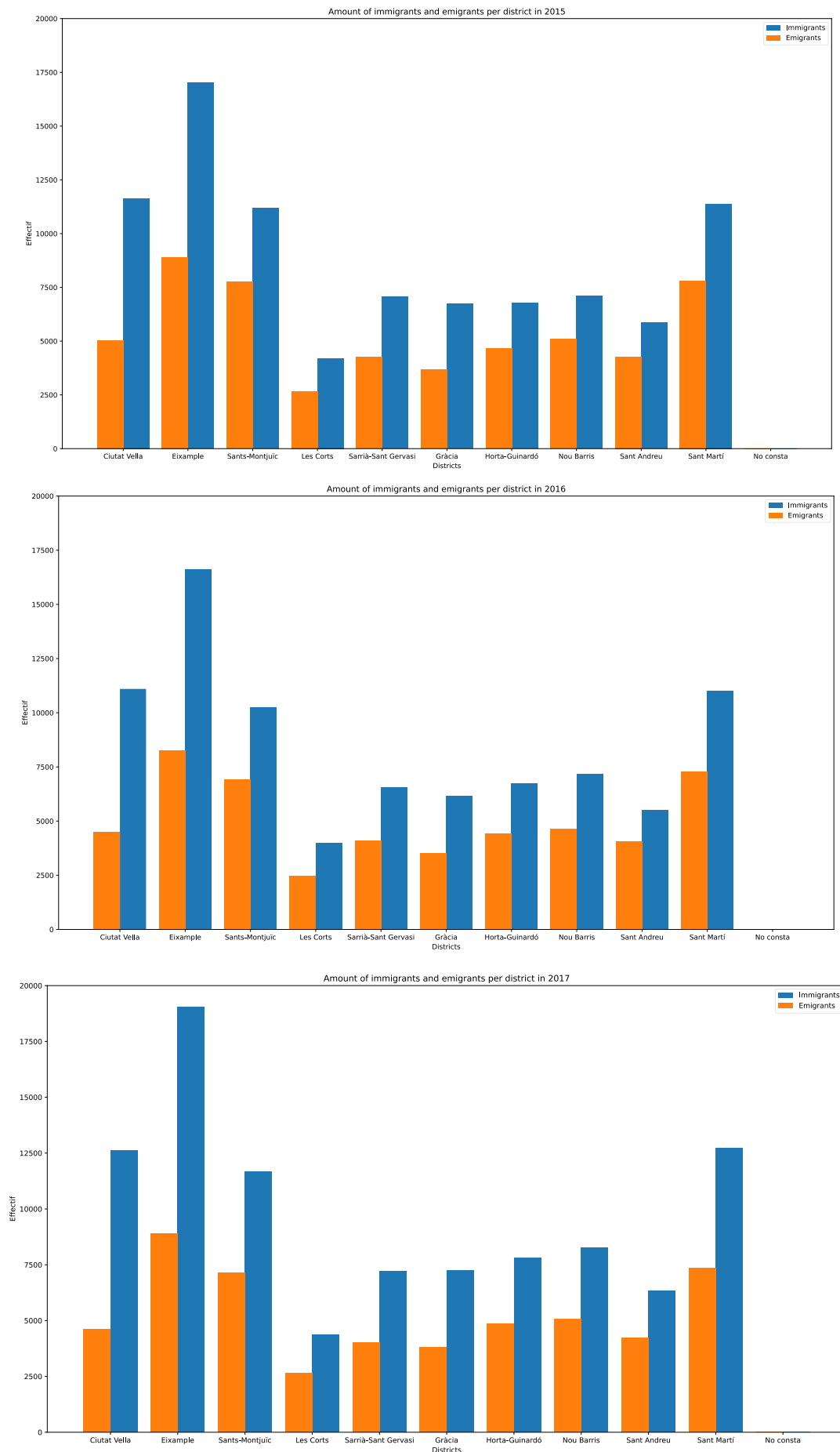
We won't go in details through each of them, so here is the list of functions we have made with similar purposes:

- GetDistrictNameFromNumber
- GetDistrictNumberFromName
- GetNeighborhoodNameFromNumber
- GetNeighborhoodNumberFromName
- GetEachNeighborhoodFromDistrictByNumber
- GetEachNeighborhoodFromDistrictByName
- GetImmigrantsAmountFromNeighborhoodByNumber
- GetImmigrantsAmountFromNeighborhoodByName
- GetImmigrantsAmountFromDistrictByNumber
- GetImmigrantsAmountFromDistrictByName
- GetTotalImmigrantsAmount
- ObtenirNombreImmigrantsNatioDeNeighborhoodParNumero
- ObtenirNombreImmigrantsNatioDeNeighborhoodParNom
- ObtenirNombreImmigrantsNatioDeDistrictParNumero
- ObtenirNombreImmigrantsNatioDeDistrictParNom
- ObtenirNombreImmigrantsNatioTotal
- ObtenirNombreImmigrantsParAge
- GetEmigrantsAmountFromNeighborhoodByNumber
- GetEmigrantsAmountFromNeighborhoodByName
- GetEmigrantsAmountFromDistrictByNumber
- GetEmigrantsAmountFromDistrictByName
- GetTotalEmigrantsAmount
- ObtenirNombreEmigrantsParAge

## Creating diagrams to illustrate relevant information

To display our data in an intuitive and explicit manner, we decided to write a python script that relies on seaborn, matplotlib and pandas to create diagrams that illustrate the information we thought to be relevant enough to be studied.

## Immigration and emigration per district



As can be seen, both immigration and emigration slightly lowered between 2015 and 2016. However, it raised to a never-before-seen point between 2016 and 2017, surpassing the 2015 statistics.

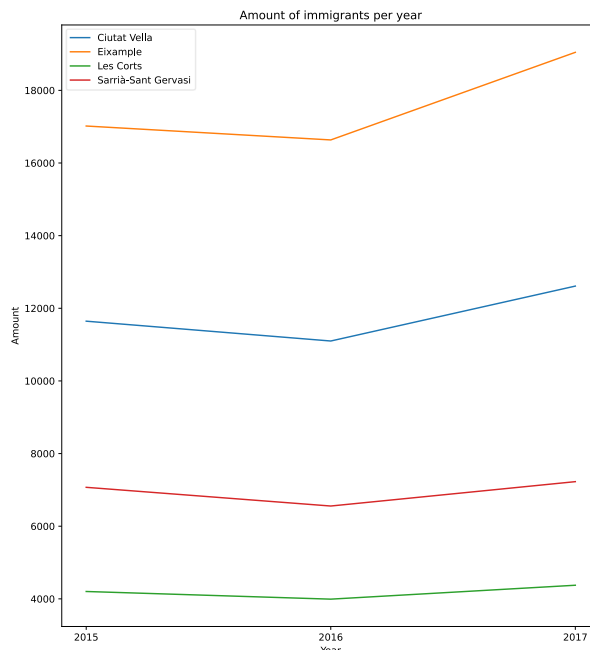
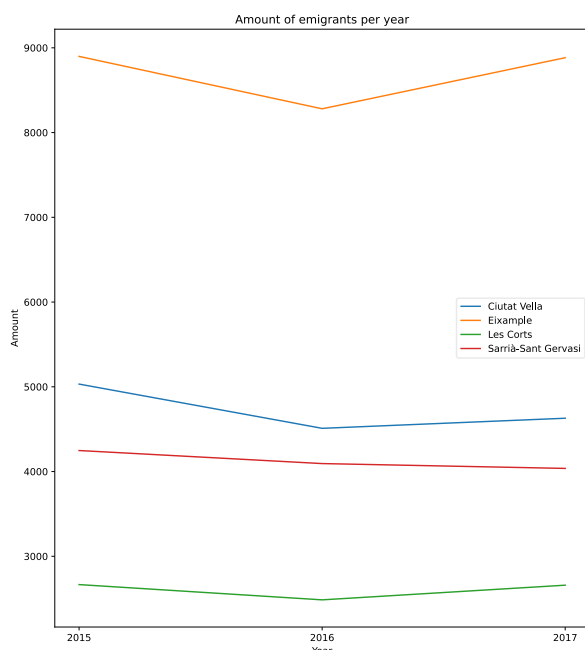
Besides that, some districts' figures vary more than others. This is the case for Eixample, located at the very center of Barcelona, holding the place of the most populated district in the entire city, which explains these significant changes.

In the next few parts, we will focus on four districts that we have chosen using these diagrams, following these criterias:

- Eixample: it is the city's lungs: the most populated district of all, but also the one with the highest development rate. Studying this district is key to understand Barcelona's evolution.
- Les Corts: this district has lots of buildings to host middle class people.
- Ciutat Vella: this is the poorest district of Barcelona, with the most affordable housings and the lowest life cost.
- Sarrià-Sant Gervasi: opposed to Ciutat Vella, this is the richest district of the city, with the highest life cost and life conditions.

Thanks to these four districts, we will be able to analyze migration depending on people's characteristics and hopefully figure out new information and make conclusions.

The two following diagrams focus more on these four districts, and measure migration per year to see the evolution more clearly:



Here, we can see how Eixample is demographically superior to the four other districts, which makes a lot of sense since it's the most populated one.

Generally, it is pretty clear that the population in each district raises. Even though the shapes on these diagrams may look similar, the scale is different: there is much more immigration (people going in) than emigration (people going out).

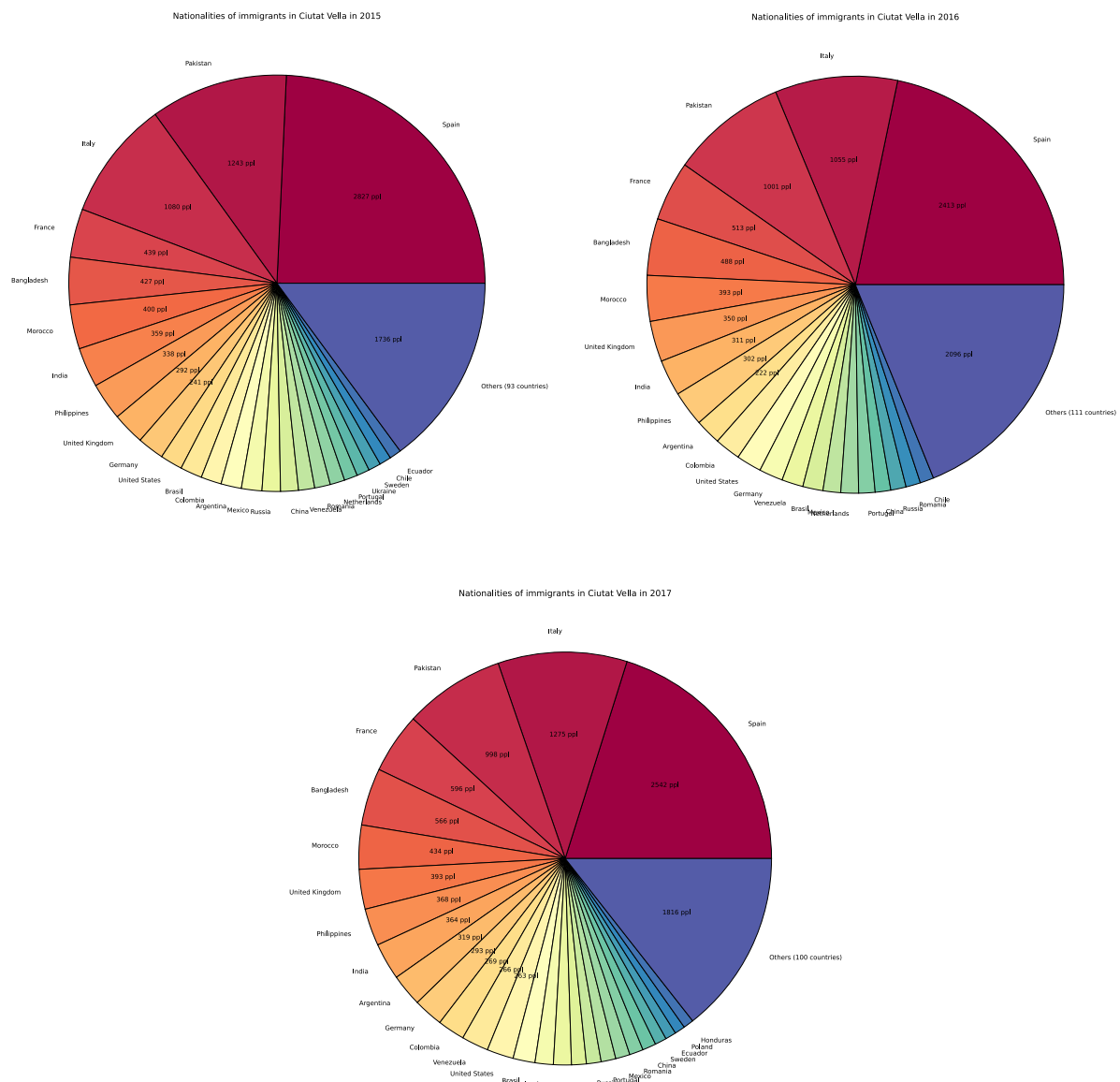
Eixample and Ciutat Vella being the two most affordable districts, they are obviously the ones to face the biggest population increases each year. Meanwhile, Les Corts and Sarrià-Sant Gervasi have their inhabitants count raise much slower.

Each year, all districts have leaving inhabitants. But while the emigration amount seems to lower down only in 2016 for all districts, Sarrià-Sant Gervasi's keeps lowering down even in 2017, likely due to the high life conditions in there.

## Immigrants' nationalities for each relevant district, each year

Across this part, we will represent immigrant' nationalities for each district, each year as pie charts. Countries with less than 100 people immigrating into Barcelona were merged into a single "Other" category for readability.

### EXAMPLE



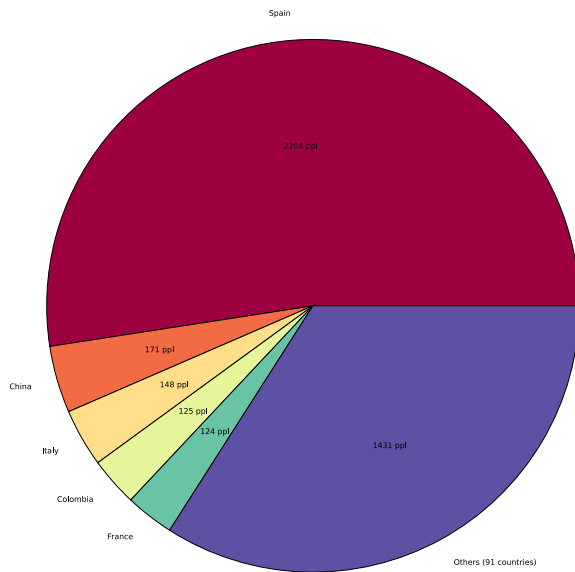
Eixample is definitely the district with the most widespread cultural differences in terms of immigrations. About are Spanish, but almost a quarter are from well developed countries. The rest are from various countries from all over the world.

About 10% of immigrants are represented by more or less 100 different countries on average, which shows that even though most people are from well known countries, there is also a good proportion that are from smaller, less present countries.

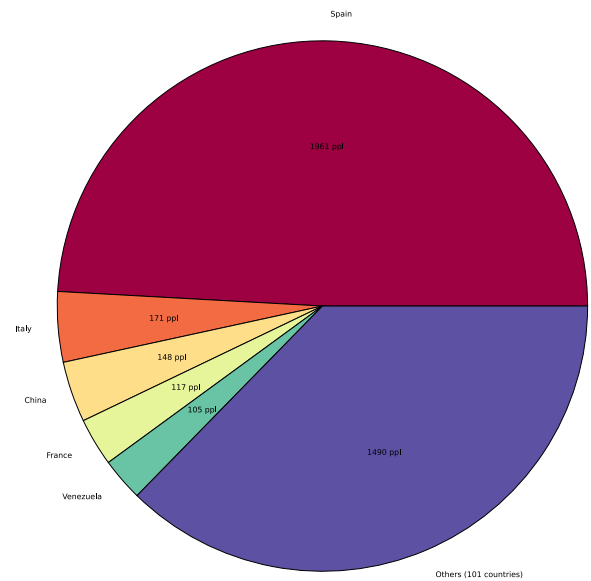
In terms of evolution, there are more and more Colombian and Venezuelan people immigrating into Barcelona, getting a bigger share than France since 2016.

On a more global note, the diversity met here show how socially diversified Eixample is, thanks to its excellent quality/life cost ratio.

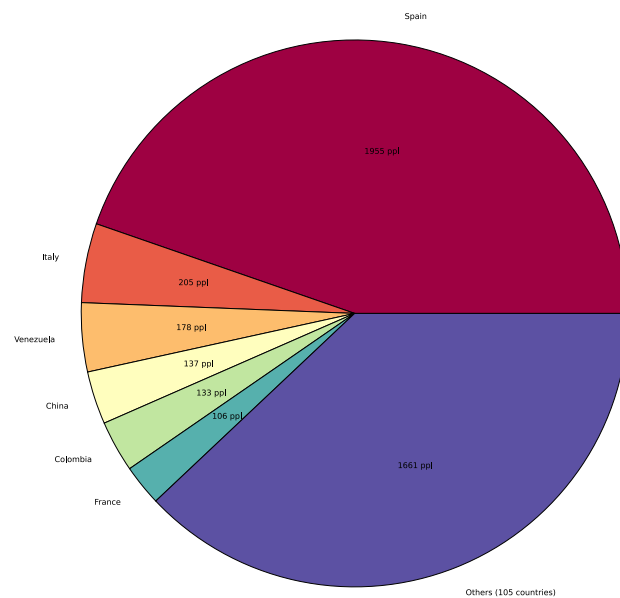
Nationalities of immigrants in Les Corts in 2015



Nationalities of immigrants in Les Corts in 2016



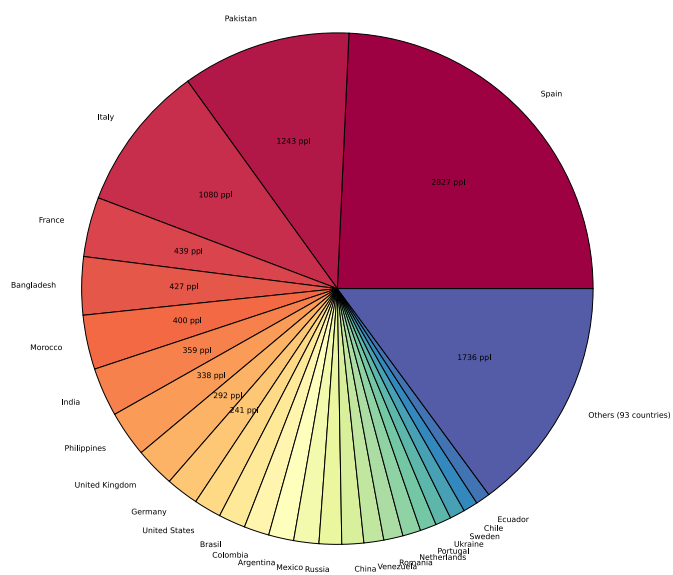
Nationalities of immigrants in Les Corts in 2017



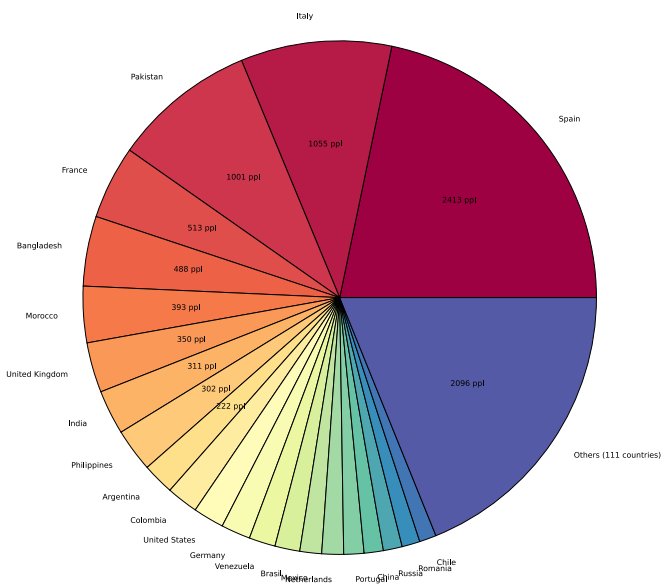
For Les Corts, diversity is way less present than for Exiample, although still quite existing: there is a clear Spanish majority. Italy, Venezuela, China, France and Colombia come next, slightly exchanging places from year to year. However, while Spanish keeps its place as the majority across the years, it still decreases, while the "other" countries' rises.



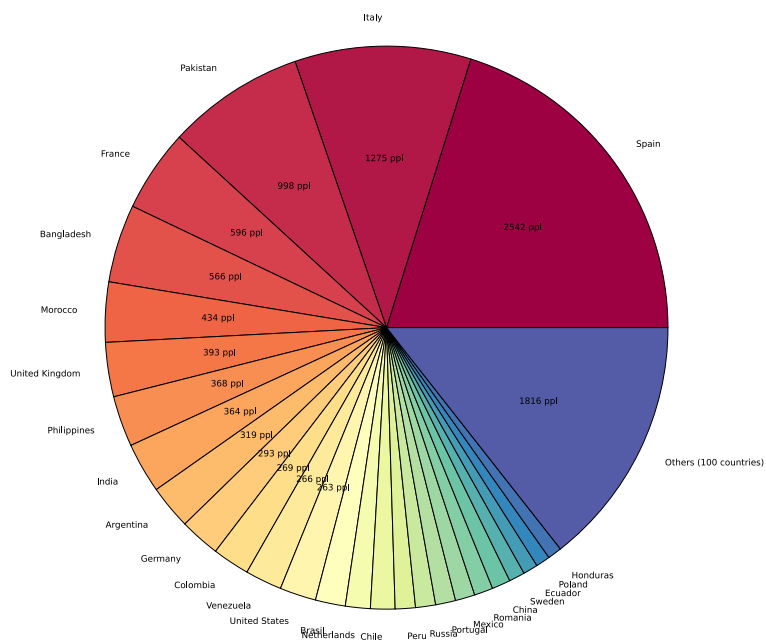
Nationalities of immigrants in Ciutat Vella in 2015



Nationalities of immigrants in Ciutat Vella in 2016

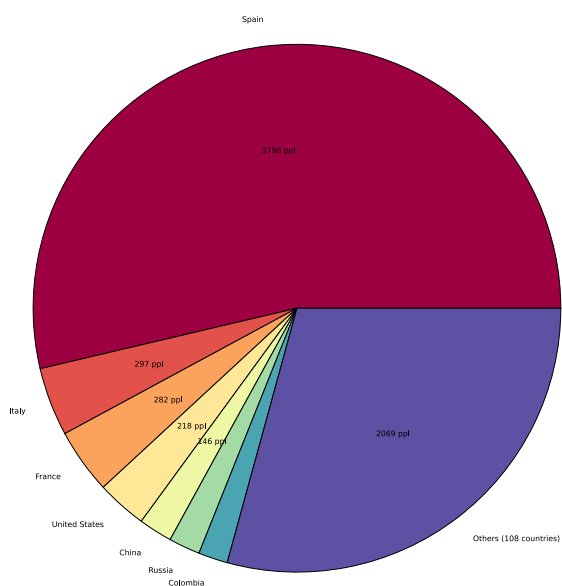


Nationalities of immigrants in Ciutat Vella in 2017

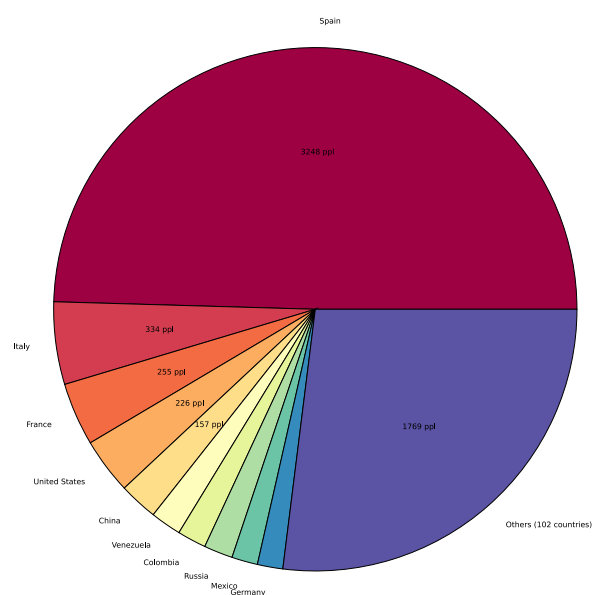


Ciutat Vella has a very important cultural diversity: the Spanish immigrants rate is way lower than for other districts. Unlike other districts, Ciutat Vella has a lot of immigrants from less developed and poor countries (Pakistan, Bangladesh, Morocco, Philippines, etc), which can be explained by the fact that it is Barcelona's poorest district, with the lowest life cost and most affordable housings.

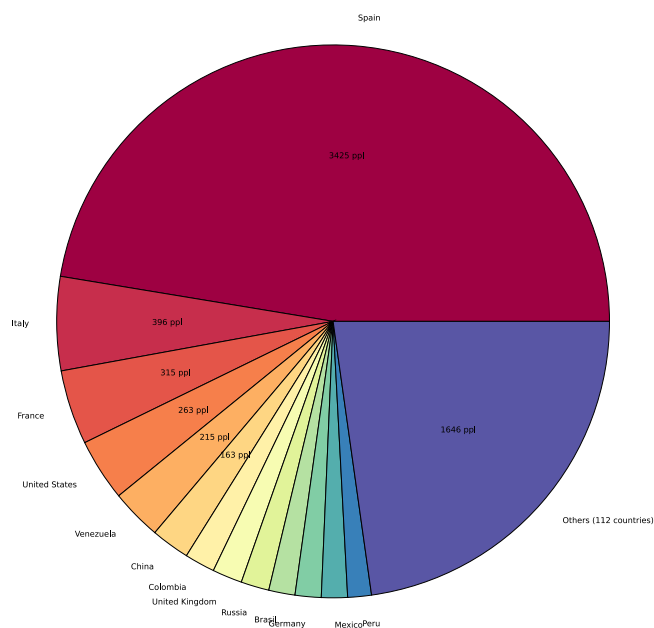
Nationalities of immigrants in Sarrià-Sant Gervasi in 2015



Nationalities of immigrants in Sarrià-Sant Gervasi in 2016



Nationalities of immigrants in Sarrià-Sant Gervasi in 2017

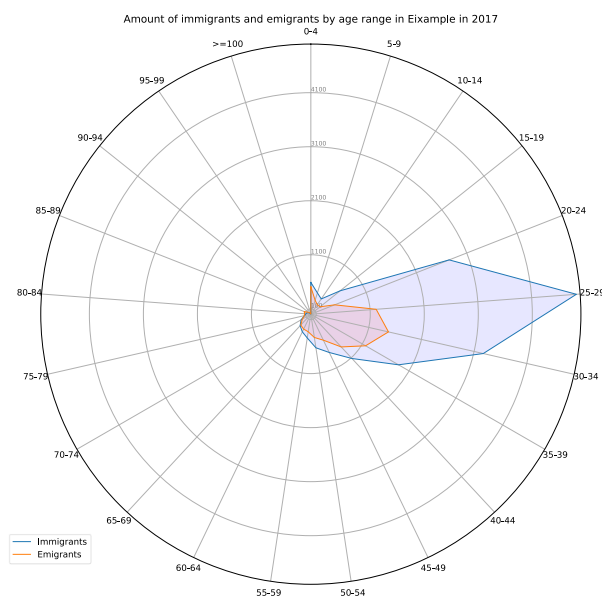
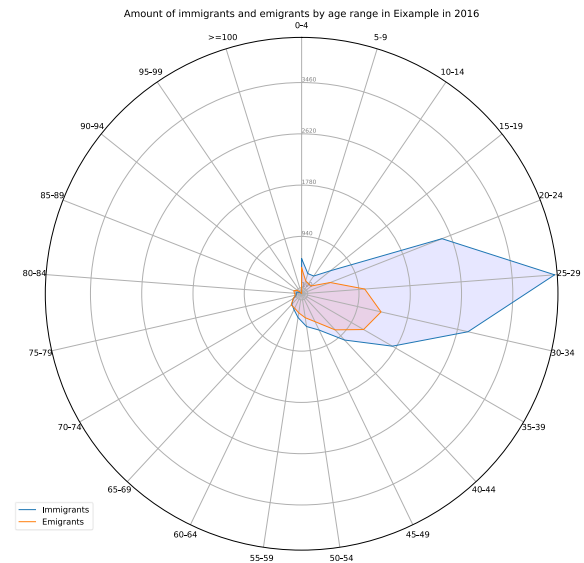
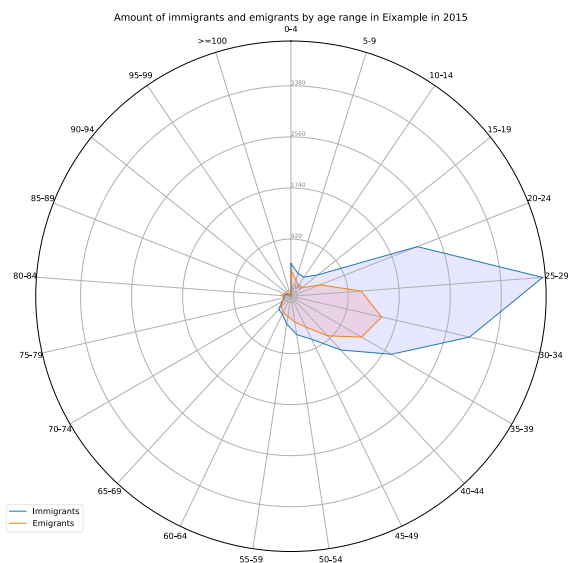


This district is the total opposite of Ciutat Vella: not only is its life cost very high, it does not welcome poor families very well. Most immigrants are Spanish, followed by Italy, France, and other developed countries. People who immigrated in here are doing so in the long run, which explains why there's less and less emigrants across the years in this district (Cf. previous diagrams).

## Immigration and emigration by age range for each relevant district, each year

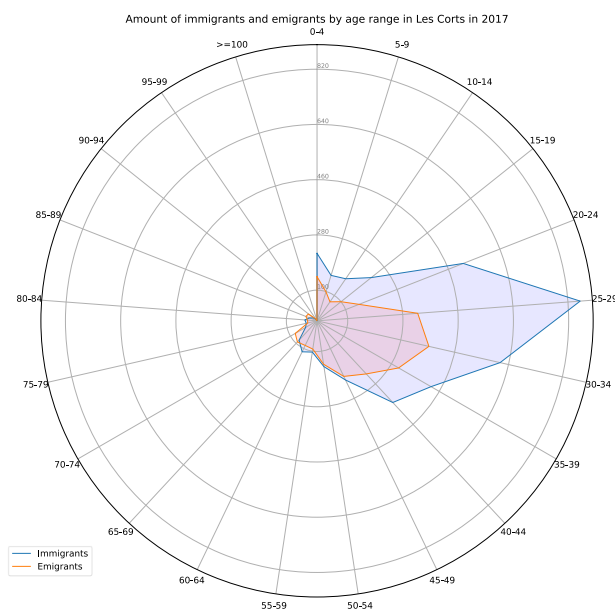
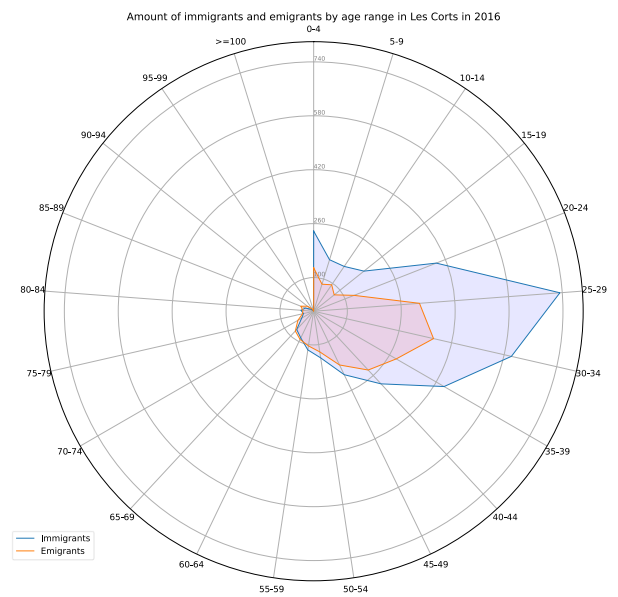
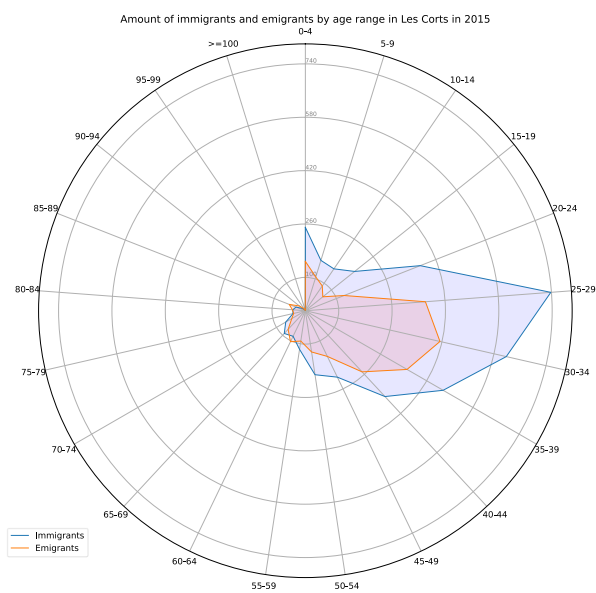
Let's now check the age range of both immigrants and emigrants. This is very important to figure out what kind of people join and leave specific districts, and to add to the facts determined by our previous analysis.

### *EIXAMPLE*



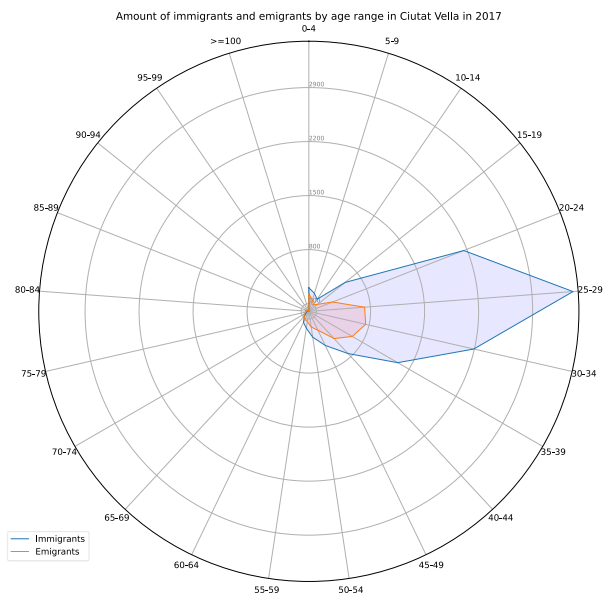
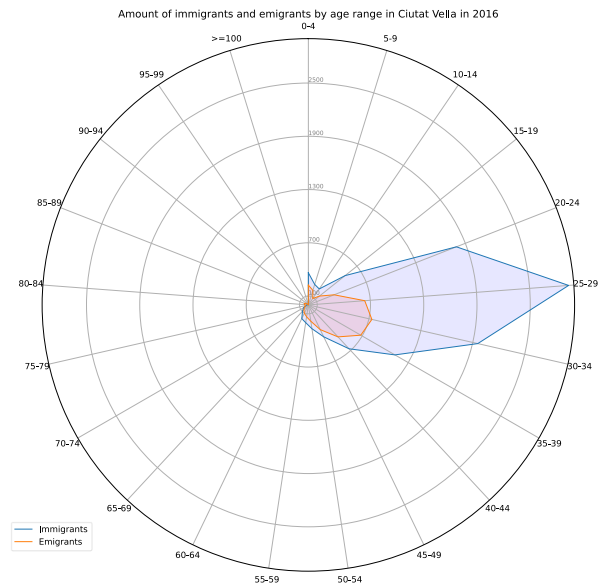
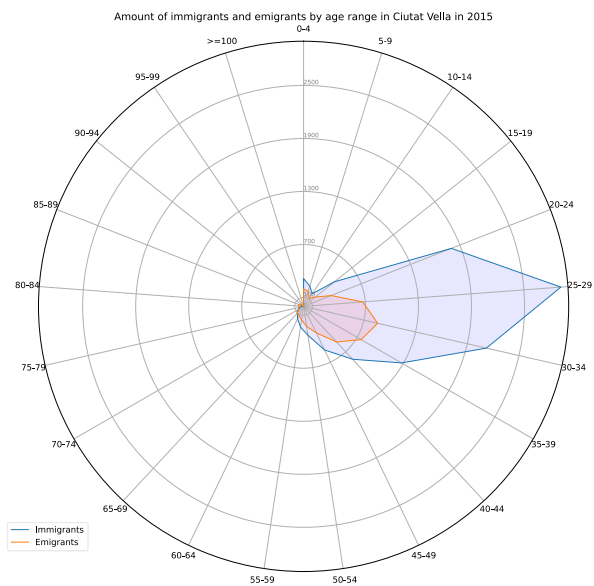
Emigration reduces across the years in Ciutat Vella. Immigration however is raising a lot and mostly targets people between 25 and 29 years old.

Our theory is that people coming in there are youngsters trying to find a job, and once they have got enough money, they can move to another district for better life conditions.

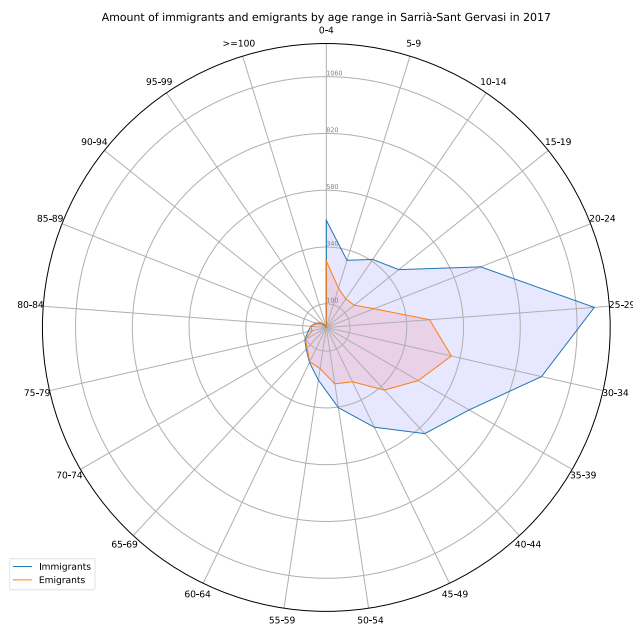
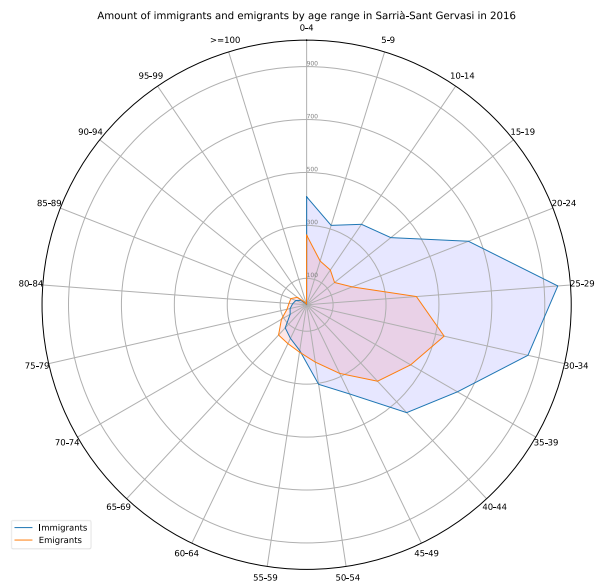
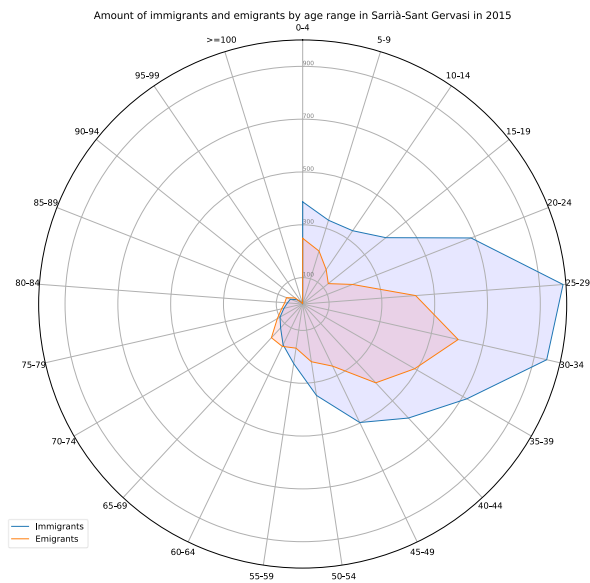


Emigration lowers across the years, although way less than for Eixample. Immigration mostly target 25 to 29 years old people as well, and emigration targets mostly 25 to 39 years old people.

There is a spike on 0 to 4 years old people though, who are likely the children of the immigrating young couples.



Just like the previous two, immigration is mostly targeted at 25 to 29 years old people, and emigration, which mostly target 30 to 34 years old people, decreases.



This diagram is by far the most interesting one on this part: immigration raises and emigration lowers as always, but here we can see that the age range is completely different.

While immigration is still targeted mainly at 25 to 29 years old people, there is also a decent part of 0 to 10 years old people immigrating. Considering this is the richest district, it is very likely that young 25 to 29 years old couples immigrate with their very young children, resulting in these figures.

But then, why are there also so many 0 to 4 years old emigrants? Well, considering that most emigrants are still young as well -30 to 34 years old), they likely emigrate with their children too, likely because the district had become too expensive for them to host not only themselves, but also their children.

Interestingly enough, 2015 and 2016 saw people between 65 and 90 years old emigrate more than immigrate, unlike other years and other districts.

## Making estimates from existing data

From the data we have seen, we can clearly say that the population of Barcelona as a whole is going to raise a lot. More and more immigrants come it while there are less and less emigrants leaving their districts.

## Conclusion

In conclusion, the analysis of this database enabled us to gain a better understanding of population movements in the city of Barcelona and the impact on its main districts.

We were able to learn how to sort and organize a large amount of data and then draw up a complete analysis.

Hugo & Matthieu