Icon

Description automatically generated

Events in Paris

Name: Mathieu JOMAIN

June, 2022

Table of Contents

[Introduction 3](#_Toc105180635)

[Data 3](#_Toc105180636)

[Intro, data sources and meta data 3](#_Toc105180637)

[Cleaning 5](#_Toc105180638)

[Data visualization 8](#_Toc105180639)

[Data exportation 10](#_Toc105180640)

[SQL 10](#_Toc105180641)

[ER Model 10](#_Toc105180642)

[Database and tables 11](#_Toc105180643)

[Importing data to database 12](#_Toc105180644)

[Queries 13](#_Toc105180645)

# Introduction

The project is the following: find and manipulate data on the subject of our choice.

The goal is to realize the different steps of the creation of a database, starting with the collection of data, the cleaning, the visualization and finally the creation of a database with several tables.

The first step is to plan the project using a project management tool. I chose to use Trello. You can find my board here : <https://trello.com/b/z8nmgMzG/rncp-project>

Graphical user interface, text, application, chat or text message

Description automatically generated

# Data

### Intro, data sources and meta data

In order to realize this project, I went to the Paris Data website, which provides a lot of free data. I chose data about events taking place in Paris.

This data set is available online on the Opend Data Paris website, or you can reach it using this link:  
<https://opendata.paris.fr/explore/dataset/que-faire-a-paris-/information/?disjunctive.tags&disjunctive.address_name&disjunctive.address_zipcode&disjunctive.address_city&disjunctive.pmr&disjunctive.blind&disjunctive.deaf&disjunctive.transport&disjunctive.price_type&disjunctive.access_type&disjunctive.programs>.

In addition, I also used another data source regarding postal codes / city name / insee code, from the Data Gouv website; available here:  
<https://www.data.gouv.fr/fr/datasets/base-officielle-des-codes-postaux/>

I chose this data because on the one hand it is a very tangible subject, so it is natural to understand the content of the columns.

On the other hand, it is a database with very little numerical data. This makes, in my opinion, the handling more complex, which makes it a real challenge for me.

Meta data 1: events

|  |  |
| --- | --- |
| Column name | Explanation |
| ID | Basically an unique integer number for each event |
| URL | Website link where we can find information about the event |
| Titre / chapeau | Name and some information about the event and what it is about |
| Date début / fin | Starting and ending date of the event |
| Occurences | At first list of dates of the event, then turned into the number of occurences |
| Mots cles | Keyword to indicate the nature / related field of the event |
| Nom / adresse du lieu | Name of the place and adress |
| Postal code / ville | Postal code and city of the adress |
| Access PMR / mal voyant / mal entendant | If the location suits people with different disabilities needs |
| URL / telephone / email de contact / facebook | Website link, phone number, email address and facebook page of the organisator |
| Type de prix | Field to know if the event as either a paid or a free entrance |

Meta data 2: insee

|  |  |
| --- | --- |
| Column name | Explanation |
| Code insee | INSEE unique code of the location (something like a postal code) |
| Nom de la commune | Name of the city |
| Code postal | Postal code of the city |
| Coordonnées GPS | Exact GPS coordinates |

### Cleaning

The first step we have to do once we have the data store on our computer is to import the data to python using the command read from pandas.

Then we start by exploring the data to try to better understand how the information is structured, what are their types....

Graphical user interface, text

Description automatically generated

According to me, the next step is to identify potential columns to be deleted for lack of relevance.

Text

Description automatically generated

Another factor that can lead to the removal of a column is its lack of information. We should therefore consider looking at the missing values of the whole dataframe.

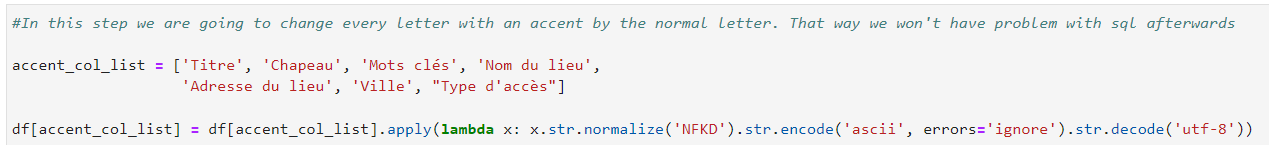
Graphical user interface, application

Description automatically generated

I then did some formatting on the columns containing text, to reduce the case of the characters and to remove the potential accents.

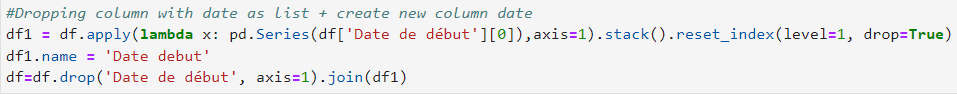
Graphical user interface, text, application, chat or text message

Description automatically generated



I then transformed the information contained in text in different ways. I extracted the information from the date columns in order to have only one date and delete the additional useless content.



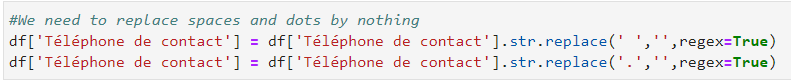


I have cleaned the list of keywords linked to each event by keeping only one value, in order to be able to process this information with much more ease.

Scatter chart

Description automatically generated

I have made changes to the phone number and postal code columns by removing potential spaces, letters and any other unwanted symbols.



I changed the nature of the content of the "occurrence" column to have a number instead of a list of dates.

Graphical user interface, text, application, email

Description automatically generated

In general, in data cleaning, there is a large part of the work that corresponds to the treatment of null values. Because I have mainly categorical values in my data, I had to change the NaN values by something else. Here I chose to replace it by “non renseigne”.

Text

Description automatically generated

### Data visualization

After finishing the data cleaning, I made some graphs in order to have a global idea of the data distribution. The first graph is the repartition of keyword. We can see that events related to music are dominant.

Chart, histogram

Description automatically generated

The next thing I wanted to know is the distribution of paid and free entrance. I was surprised to see that the repartition is almost even.

Chart

Description automatically generated

Then I used crosstabs to gather some more information. Furthermore, this way is more suitable to analyze categorical data.

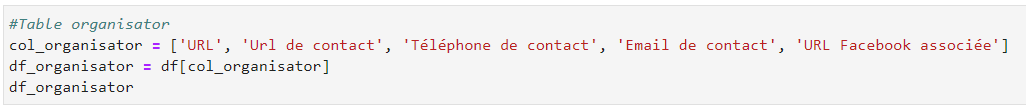
The first one is about the repartition of type of entrance for different location, here I used postal code. We can see that in the district 1 of Paris, there is almost 2 paying events for 1 free event.

A screenshot of a computer

Description automatically generated with medium confidence

### Data exportation

And finally, I decided to create what will be my future database tables. I used the same process for my four tables: create a list with the columns that I want and then create a new dataframe with only those elements.



Once the new dataframe is set, I exported it as csv file.



# SQL

### ER Model

Before diving into data cleaning and visualization in python, it is important to create an initial entity relationship diagram.

This allows us to have some guidelines. It is indeed better to handle the data when we have a first idea of the final result.

This ER model will then change according to the elements we obtain and the relationships we can set up between them.

Here is my final diagram:

Diagram

Description automatically generated

### Database and tables

The first step in working with SQL is to create the database. After that we need to let SQL know that we want to do some work on that particular database by writing the code “use”.

Now we can create the tables that we want by defining the columns names and their data type. We can select a column to be the primary key and also a foreign key from another table.

Graphical user interface, application

Description automatically generated

Example of a more complex table:

Graphical user interface, text, application

Description automatically generated

### Importing data to database

There are many ways to import data into SQL:

* create the table and import the data

Graphical user interface, text, application, email

Description automatically generated

* or with a code in SQL

Graphical user interface, text, application, chat or text message

Description automatically generated

* or even with a code in python, but it requires some extra steps (I used that way)

First, we need to initialize the connexion between python and sql.

Graphical user interface, text

Description automatically generated

Then we need to create a dataframe from the cleaned csv file and export it to SQL.

Graphical user interface, text, application

Description automatically generated

### Queries

Query 1:

Elements used: email, ID, titre (title of the event)

This first query shows how to reach a specific element form another table and add it to the input we want. Here we wanted to access the email addresses of organisators of events related to sports that occur more than 2 times.

Query 2:

Elements used: keyword + count

For the second query, I wanted to classify the events according to the keywords in order to know which ones works the best.

We can see that the events related to music are strongly dominant, more than 400, almost twice the number of events related to the second biggest category (which are the activities for children)

Query 3:

Graphical user interface, application

Description automatically generated

Here I wanted to know which district of Paris is the best for each type of budget. In one hand I want to know the district where the most of events (paid entrance or free entrance) take place.

We can see that in Paris 12e there are 172 paid events, followed closely by Paris 1e with 169 paid events.

As for the free events, in the 19th district of Paris there are 140 free events and 133 for Paris 18th.

With this analysis you can identify the places to go to participate in events according to your budget.

Query 4:

That query is very similar to query 3, but instead of displaying the type of price I wanted to know the repartition of events within the districts by the keyword.

With the output we can see that if you want to go to an event related to art for example, the best thing is to go in Paris 13e.

Query 5:

Graphical user interface

Description automatically generated

The last query shows the sum of occurrences in each district in order to know in which district takes place the most events. Thanks to the last 3 queries you can select the district that suits the best your desires in term of content, price and frequency.