Certification CDSD
Block 1 & 3

August 28th, 2025 – Louis Le Pogam

Agenda

- Block 1 - Build & Manage a Data Infrastructure – Kayak Project
- Block 3 – Unsupervised Machine Learning – Uber Pickups Project

Project Reminder

Project

- Kayak Marketing Team would like to create a holiday recommendation application based on :
  - Weather
  - Hotels in the area
  - Based on real-time data

Goal

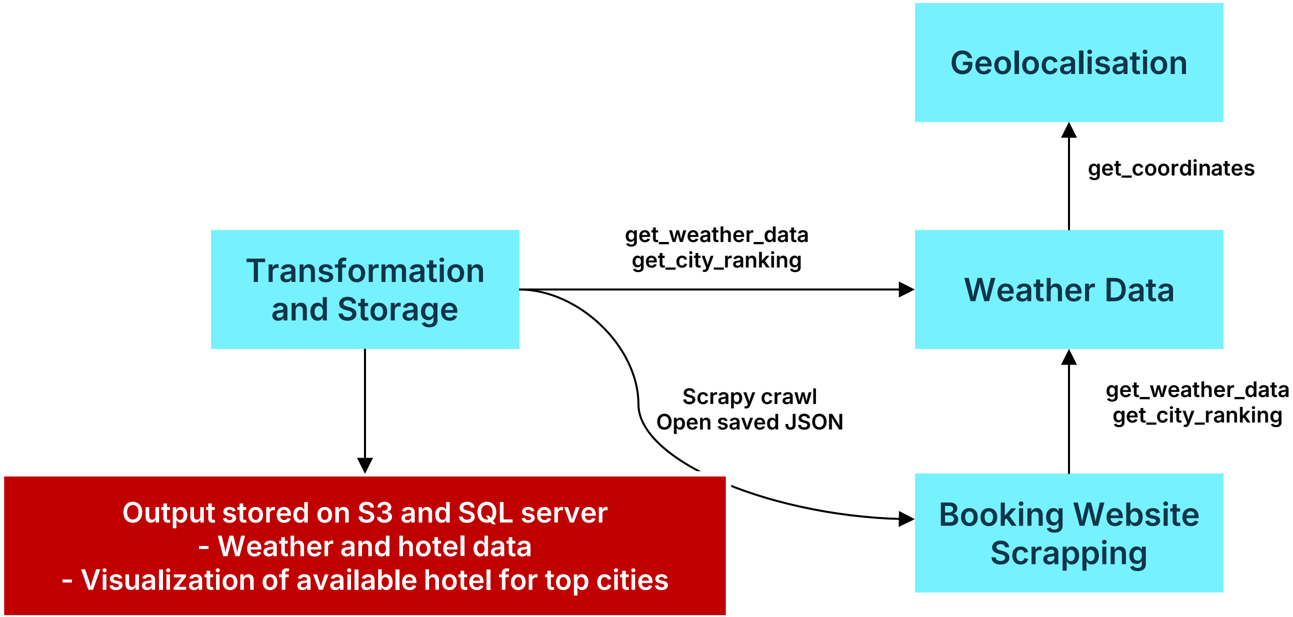
- The data are not available and the goal is to get the needed data as following:
  - Scrape data from destinations.
  - Get weather data from each destination.
  - Get hotels' info about each destination.
  - Store all the information above in a data lake.
  - Extract, transform and load cleaned data from your datalake to a data warehouse.

4 building blocks for the data scrapping model

	Description
1 Geolocalisation	• Processing of list of cities by obtaining GPS coordinates and INSEE codes with API • Saving all data to CSV for weather queries
2 Weather Data	• Retrieving 7-day weather forecasts for cities based on INSEE codes • Ranking cities based on customizable criteria and creates aggregated rankings • Saving ranked list based on number of favorable days to CSV
3 Booking Website Scrapping	• Taking the top 5 cities based on previous analysis • Searching for hotels in each city available on booking.com • Saving in into a JSON file
4 Transformation and Storage	• Processing previous data by removing low-quality hotels and identifying consecutive night availability • Creating interactive visualizations and uploads all results to AWS S3 and SQL server

KAYAK

All 4 blocks are used to get the final output



The weather module scraps the weather by city and ranks them based on defined ideal conditions

get_weather_data

- Reads a list of French cities from a file based on the INSEE code
- Connects to a weather service (Meteo Concept API) for each city
- Gathers 7-day forecasts including:
  - Daily rainfall predictions
  - Temperature highs and lows
  - Wind speed information
- Organizes everything in a DataFrame where each row represents one day's weather forecast for a specific city

Output : DataFrame with weather prediction for eachs targeted cities

get_city_ranking

- Flags Less-Than-Ideal Weather Days when:
  - Too hot / Too cold / Too rainy / Too windy
  - Default value: 35°C / 20°C/ 10mm / 50 km/h
  - Can be changed depending on the season
- Calculates Problem Days for each city over the forecast period
- Ranks Cities by:
  - Fewest problematic weather days
  - Lowest average rainfall as tiebreaker

Output : DataFrame with a prioritized list of cities with optimal weather conditions for travelers

700 pages scrapped to get the available hotels of the next 7 days

Get Weather Data

- Open file of ranked cities and select top 5

Get URL for each city and date

- Loop on each cities and for the next 7 days
- Get the URL of the search for each city x date

Get the Details from the Search Results

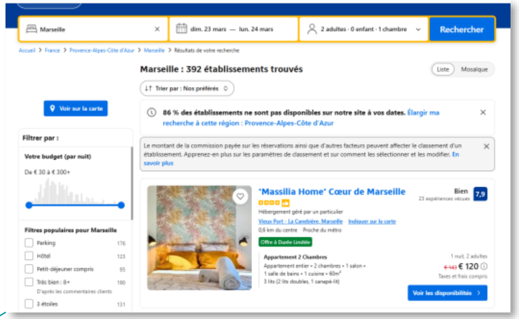
- For each results on the search page, get the main info : Name, ranking, price, distance...
- Keep only 20 first results to limit output size

Open each page and get detailed data

- For each hotel, open the hotel page to get detailed info : Adress, latitude, longitude, description, URL...

Store Data in a JSON file

- Gather all data in one JSON
- 700 pages scrapped (5 cities x 20 hotels x dates x 7 dates) in 6 minutes



Marseille | dim. 23 mars — lun. 24 mars | 2 adultes · 0 enfant · 1 chambre | Rechercher

Accueil > France > Provence-Alpes-Côte d'Azur > Marseille > Résultats de votre recherche

Marseille : 392 établissements trouvés

Liste | Mosaïque

Trier par : Nos préférés

Voir sur la carte

86 % des établissements ne sont pas disponibles sur notre site à vos dates. Élargir ma recherche à cette région : Provence-Alpes-Côte d'Azur

Le montant de la commission payée sur les réservations ainsi que d'autres facteurs peuvent affecter le classement d'un établissement. Apprenez-en plus sur les paramètres de classement et sur comment les sélectionner et les modifier. En savoir plus

Filtrer par :

Votre budget (par nuit)

De € 30 à € 300+

Filtres populaires pour Marseille

□ Parking	176
□ Hôtel	123
□ Petit-déjeuner compris	95
□ Très bien : 8+ D'après les commentaires clients	190
□ 3 étoiles	131

"Massilia Home" Cœur de Marseille

Bien 7,9
23 expériences vécues

Hébergement géré par un particulier

Vieux Port - La Canebière, Marseille · Indiquer sur la carte

0,6 km du centre · Proche du métro

Offre à Durée Limitée

Appartement 2 Chambres

Appartement entier • 2 chambres • 1 salon • 1 salle de bains • 1 cuisine • 60m²

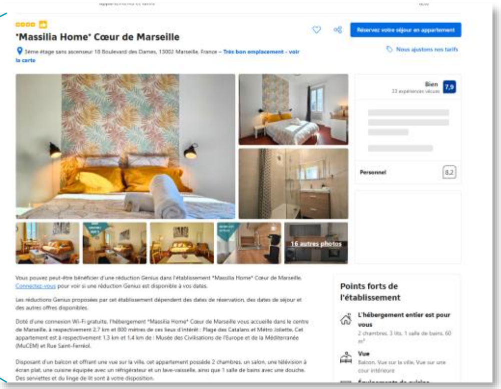
3 lits (2 lits doubles, 1 canapé-lit)

1 nuit, 2 adultes

~~€ 143~~ € 120

Taxes et frais compris

Voir les disponibilités >



"Massilia Home" Cœur de Marseille

Réservez votre séjour en appartement

3ème étage sans ascenseur 18 Boulevard des Dames, 13002 Marseille, France – Très bon emplacement - voir la carte

Nous ajustons nos tarifs

Bien 7,9
23 expériences vécues

Personnel 8,2

16 autres photos

Vous pouvez peut-être bénéficier d'une réduction Genius dans l'établissement "Massilia Home" Cœur de Marseille. Connectez-vous pour voir si une réduction Genius est disponible à vos dates.

Les réductions Genius proposées par cet établissement dépendent des dates de réservation, des dates de séjour et des autres offres disponibles.

Doté d'une connexion Wi-Fi gratuite, l'hébergement "Massilia Home" Cœur de Marseille vous accueille dans le centre de Marseille, à respectivement 2,7 km et 800 mètres de ces lieux d'intérêt : Plage des Catalans et Métro Joliette. Cet appartement est à respectivement 1,3 km et 1,4 km de : Musée des Civilisations de l'Europe et de la Méditerranée (MuCEM) et Rue Saint-Ferréol.

Disposant d'un balcon et offrant une vue sur la ville, cet appartement possède 2 chambres, un salon, une télévision à écran plat, une cuisine équipée avec un réfrigérateur et un lave-vaisselle, ainsi que 1 salle de bains avec une douche. Des serviettes et du linge de lit sont à votre disposition.

Points forts de l'établissement

L'hébergement entier est pour vous
2 chambres, 3 lits, 1 salle de bains, 60 m²

Vue
Balcon, Vue sur la ville, Vue sur une cour intérieure

Équipements de cuisine

KAYAK

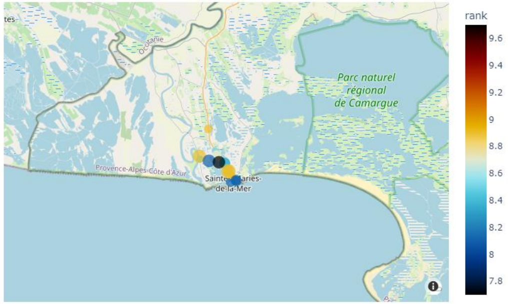
All outputs are stored on a AWS S3 and SQL server and ready to be used

Output 1 : All database

- Dataframe with weather forecast
- Dataframe with City Ranking
- JSON with all booking.com data scrapped

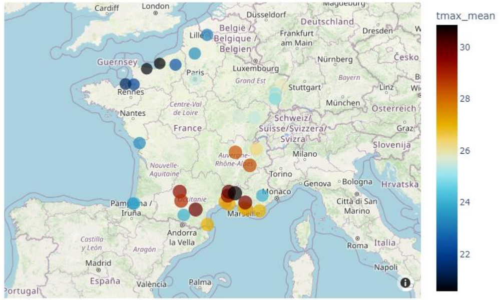
Output 2 : Visualization of available hotels for each top 5 cities

Hotels available in Saintes Maries de la mer



Output 3 : Overview of average temperature for the next 7 days of all targeted cities

Average max temperature for the next 7 days per city



All outputs are stored in a S3 and SQL server and ready to be extracted and used

Q&A

Agenda

- Block 1 - Build & Manage a Data Infrastructure – Kayak Project
- Block 3 – Unsupervised Machine Learning – Uber Pickups Project

Project Reminder

Uber

Project

- One of the main pain point that Uber's team found is that sometimes drivers are not around when users need them.
- Therefore, Uber's data team would like to work on a project where their app would recommend hot-zones in major cities to be in at any given time of day.

Goal

The target of the project is to

- Develop an algorithm to identify "hot zones" where drivers should position themselves
- Create time-based recommendations that adapt to changing demand patterns
- Visualize results for easy implementation by drivers

The dataset represents latitude and longitude of 564k pickups in April 2014

Dataset description

- Dataset of April 2014 used with 564k lines

- 4 columns:
  - Date
  - Latitude
  - Longitude
  - Base : Internal code, not used in the analysis

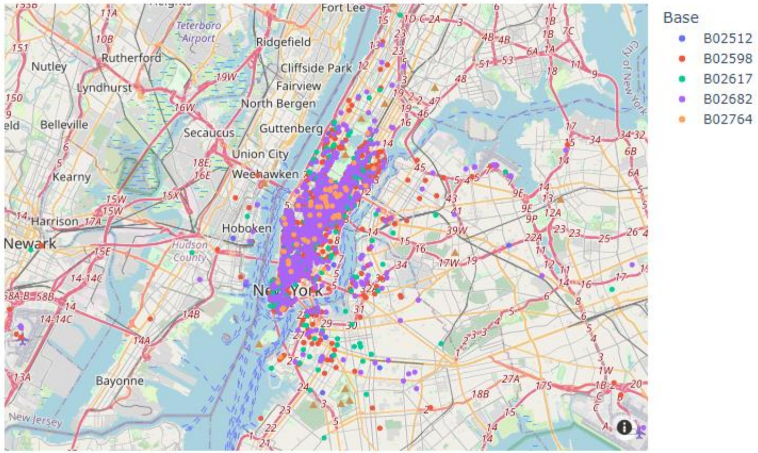
- Preprocessing limited to converting the date column into several sub-columns

- Focus on New York City inside this latitude and longitude line :
  - Latitude minimum = 40.4774
  - Latitude maximum = 40.9176
  - Longitude minimum = -74.2591
  - Longitude maximum = -73.7004

- Analysis done on the 30th of April 2014 at 5pm to limit the number of lines

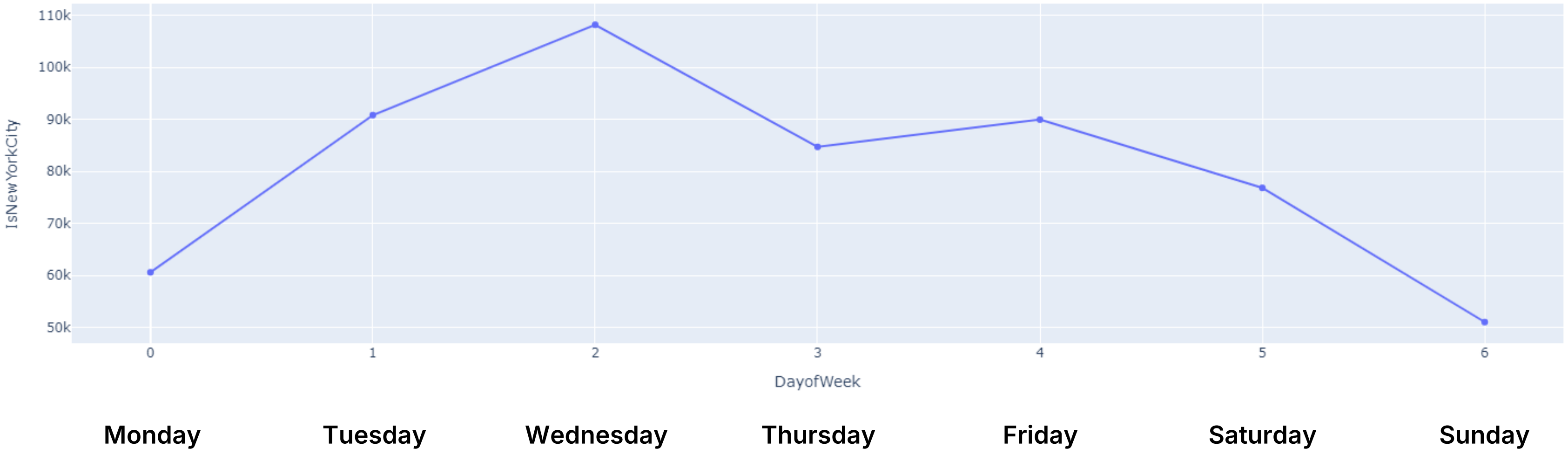
Overview of the pickup location for a given hour oof a given day

Pick up of the 30th at 5pm



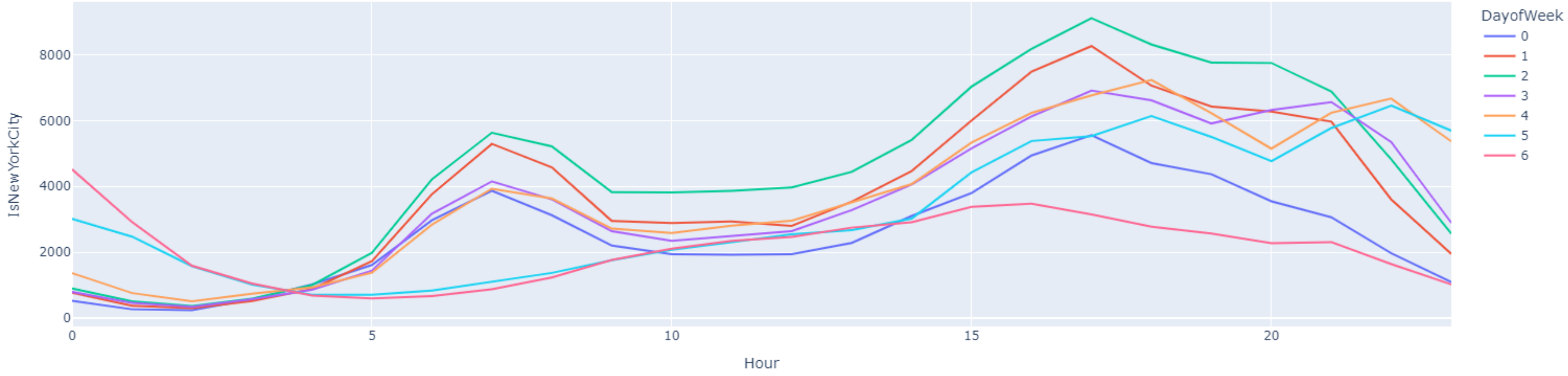
Monday and Sunday are the lowest day while the peak in on Wednesday

Number of pickups per Day | April 2014, NYC only, 0 = Monday



During the week, there is a peak at 7am then between 5pm and 8pm while the night is busy during the weekend

Number of pickups per Hour depending on the day | April 2014, NYC only, 0 = Monday



DBScan is used to calculate coordinates of hot zones at any given time

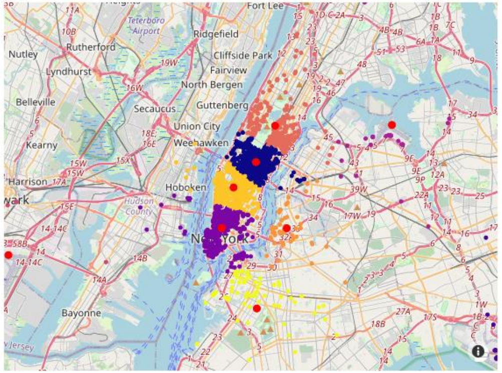
Uber

Kmeans Clustering

Description

- Elbow and silhouette methods to get the optimal number of clusters
- 9 clusters seems to be the best for April 30th

Cluster Overview

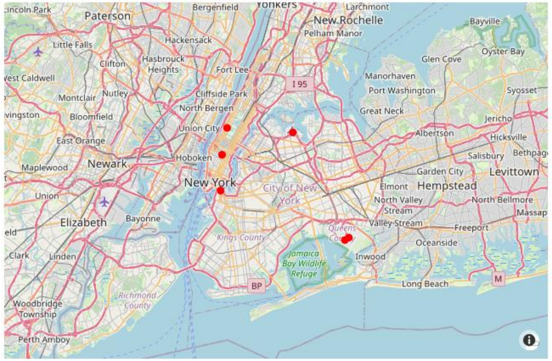


DBScan

Description

- DBScan used to handle different numbers of cluster depending on the time
- Parameters : Epsilon = 0.1 / Min Sample = 10
- 6 clusters + outliers

Cluster Center Overview



DBScan chosen for algorithm as the number of cluster adapts to the dataset

Hot zone are calculated and plotted for any given time with a DBScan clustering

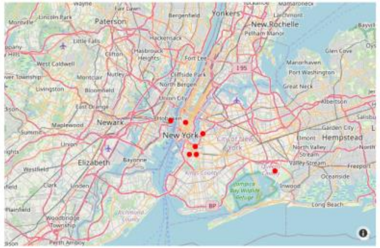
Uber

Plot_hot_zone function

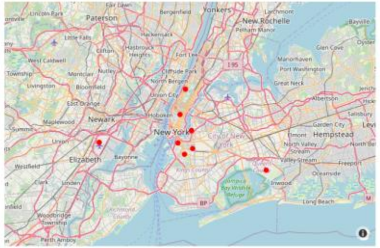
- Input : dataset, day of the week, hour, dbscan parameters
- For any given hours, would calculate the clusters center and plot them
- The output would be a map with the hot zones of this hour
- For a given day, the evolution of hot spots can be shown by looping over different hours

Evolution of hot zones for a Saturday at 12am, 6am, 12pm, 6pm

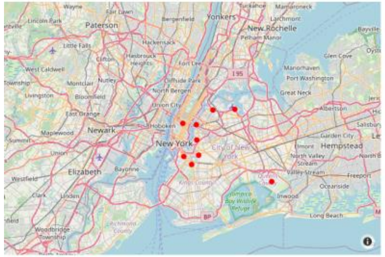
12am



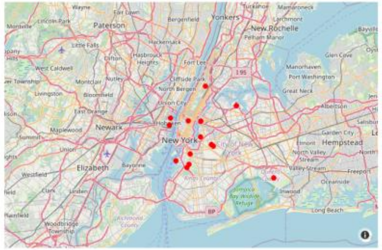
6am



12pm



6pm



Q&A

Thanks!