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

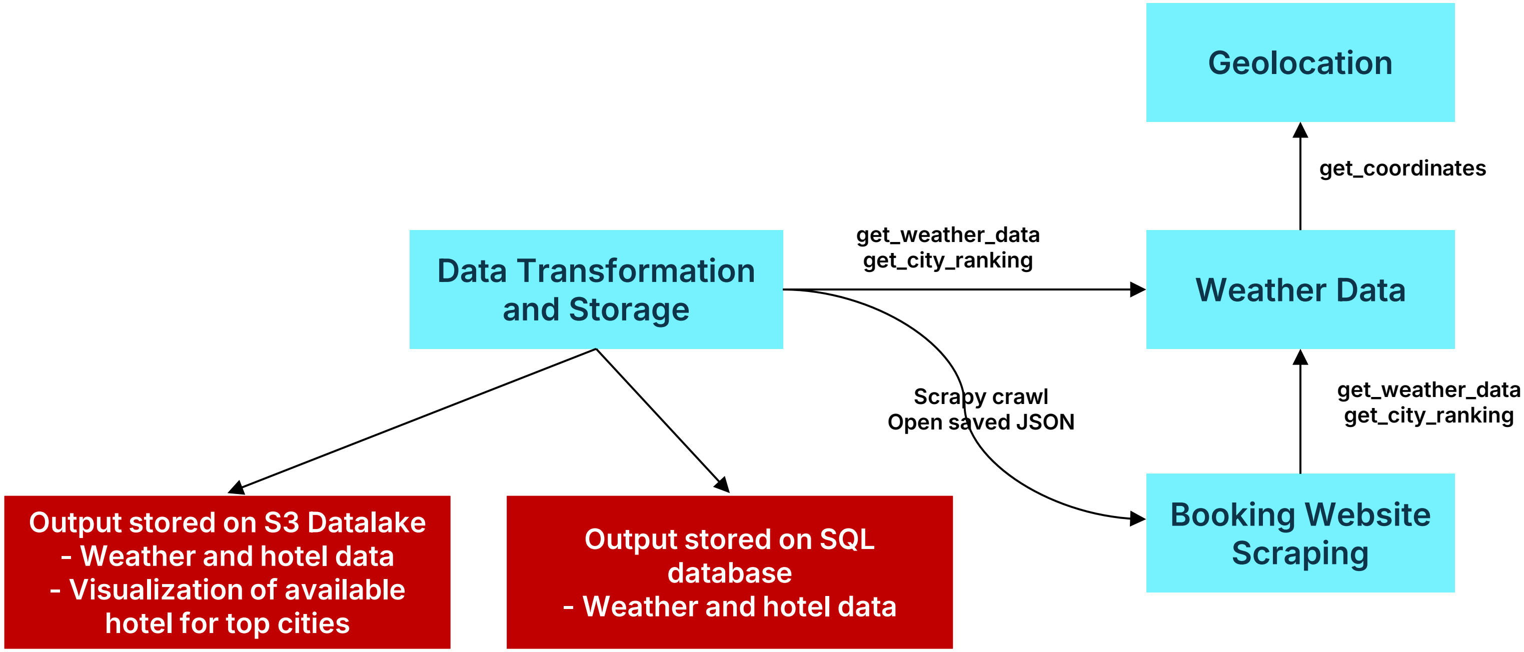
Project Reminder

Project	Kayak would like to create a holiday recommendation application based on : ▪ Weather ▪ Hotels in the area ▪ Based on real-time data
Goal	Get the needed data as following: ▪ Scrape data from destinations ▪ Get weather data and hotels' info from 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 scraping model

	Description
1 Geolocation	• 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 Scraping	• 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 Data Transformation & Storage	• Removing low rated hotels and identifying consecutive night availability • Creating interactive visualizations and storing results in AWS S3 (data lake) and an AWS RDS-hosted SQL relational database

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

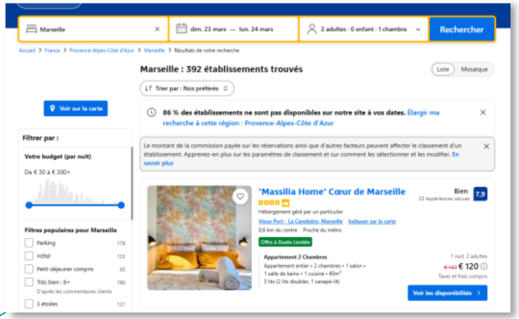
- For each results on the search page, get the main info : Name, ranking, price, distance...
- Keep only 20 first results to limit output size

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

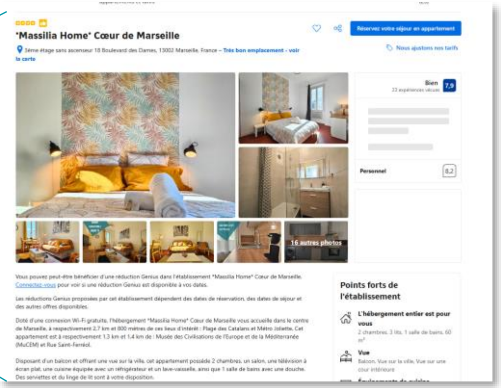
0,6 km du centre · Proche du métro

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

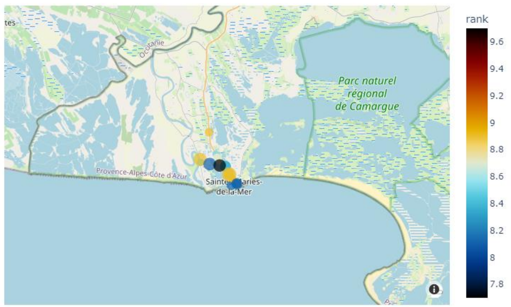
Vue
Balcon, Vue sur la ville, Vue sur une cour intérieure

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

Output 1 : All database

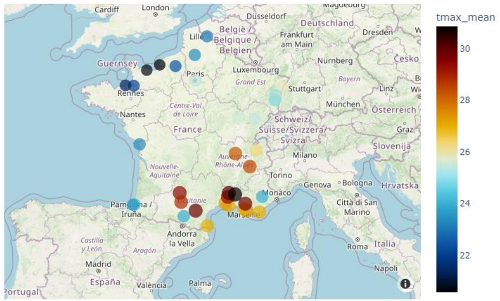
- Dataframe with weather forecast
- Dataframe with City Ranking
- JSON with all booking.com data scrapped

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 database 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

Uber's Challenge

- Drivers are not always located where and when riders need them
- This mismatch leads to longer wait times and reduced efficiency

Project Goal

- Detect real-time "hot zones" of high demand
- Provide actionable recommendations for driver reallocation
- Enable intuitive visualization for rapid decision-making

3 steps to define the hot zone

1. Exploratory Data Analysis
2. Model Selection
3. Real Time “Hot-Zone” Recommendation

3 steps to define the hot zone

Uber

1. Exploratory Data Analysis
2. Model Selection
3. Real Time “Hot-Zone” Recommendation

Uber has provided a cleaned database ready to be used

Uber

Dataset description

- Dataset of April 2014 used with 564k lines
- 4 columns:
  - Date
  - Latitude
  - Longitude
  - Base : Internal code, not used in the analysis

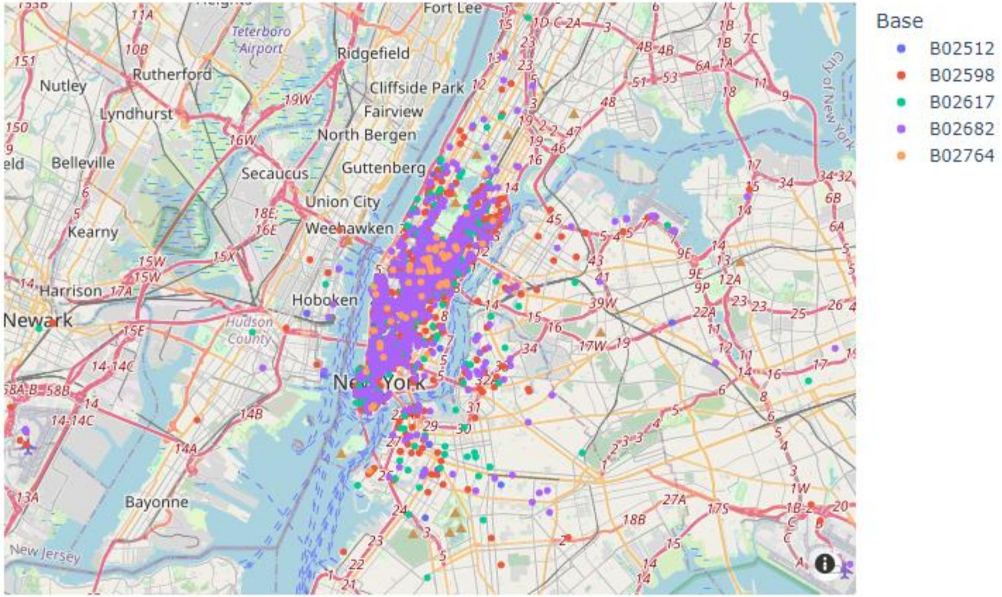
Preprocessing

- 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

Limited preprocessing performed on the database

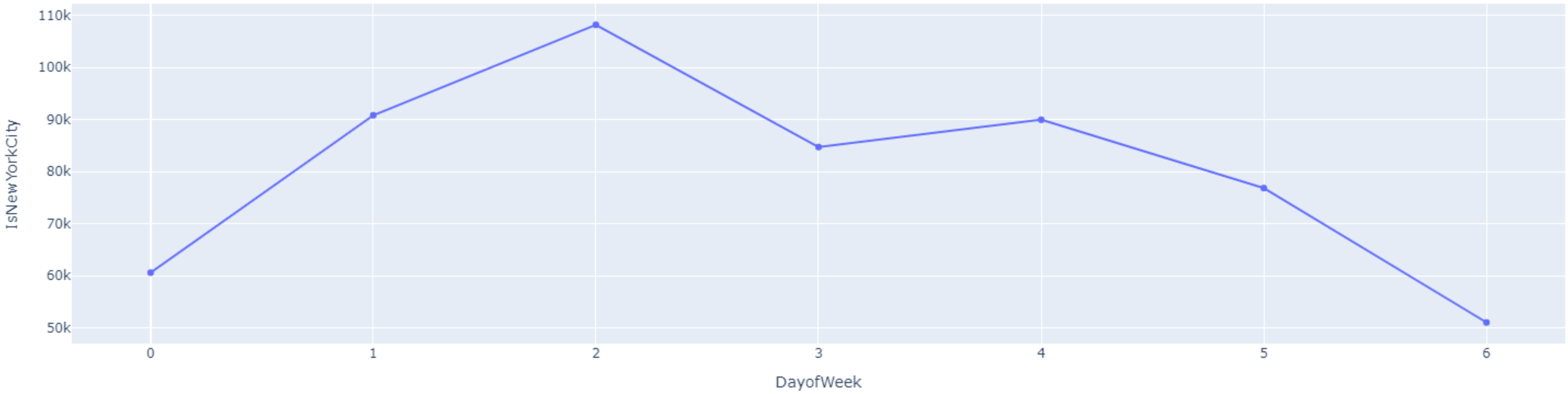
Database Overview: Each point corresponds to a single ride

Pick up of the 30th at 5pm



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

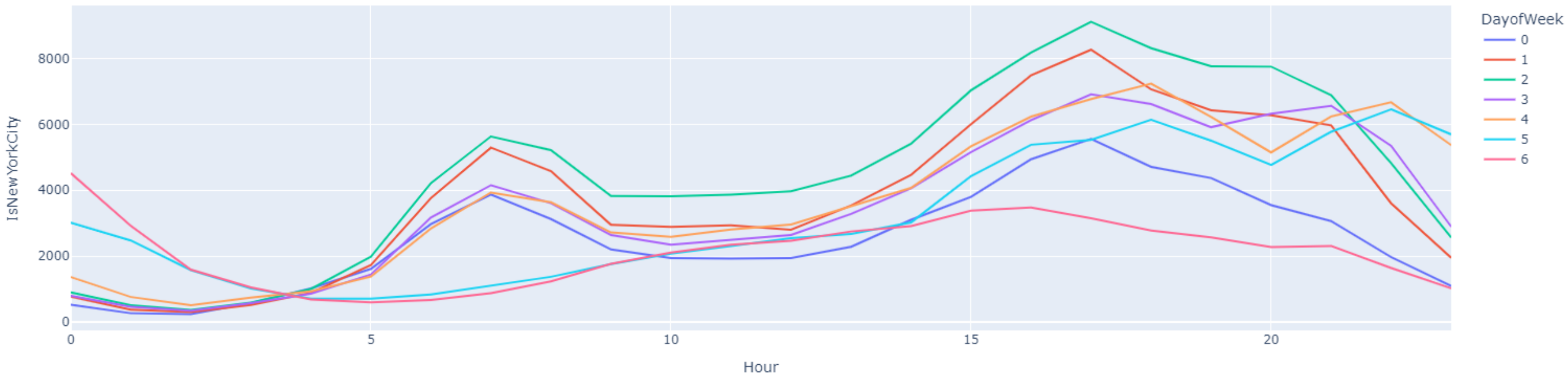
Number of pickups per Day | April 2014, NYC only, 0 = Monday



Monday Tuesday Wednesday Thursday Friday Saturday Sunday

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



3 steps to define the hot zone

Uber

1. Exploratory Data Analysis
2. Model Selection
3. Real Time “Hot-Zone” Recommendation

2 models can be chosen to define the hot zone

	Kmeans Clustering	DBScan
Principle	• Centroid-based clustering algorithm	• Density-based clustering algorithm
Shape	• Spherical only	• No shape
Strengths	• Simple and fast • Works well when clusters are spherical	• No need to know number of clusters • Handle complex shapes
Weaknesses	• Number of clusters must be known • Sensitive to initialization and outliers • Poor performance on irregularly shape	• Commutationnally heavier
Rationale for using it	• Speed and simplicity when number of clusters is known	• Complex shapes and number cluster not known

Seems more adapted to handle real-time data

DBScan was chosen for hot zone recommendation

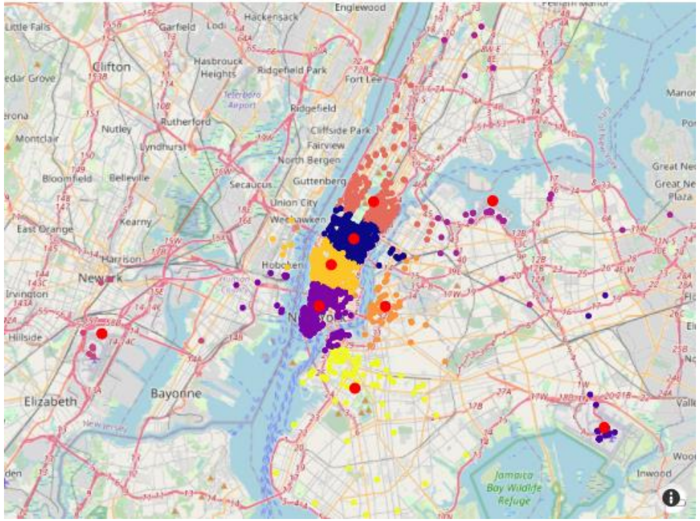
Uber

Kmeans Clustering

Details

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

Cluster Overview

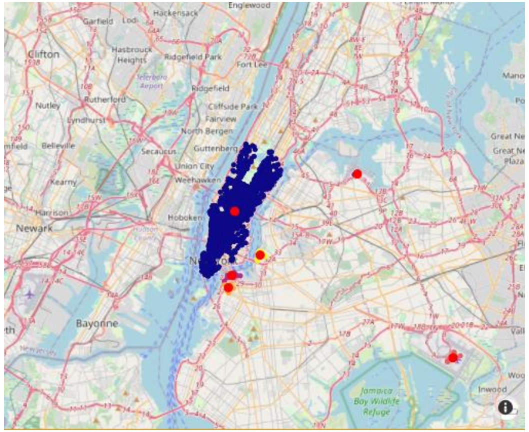


DBScan

Details

- Parameters : Epsilon = 0.15 / Min Sample = 10
- 6 clusters + outliers

Cluster Center Overview



Best model to handle variability in cluster location

3 steps to define the hot zone

Uber

1. Exploratory Data Analysis
2. Model Selection
3. Real Time “Hot-Zone” Recommendation

Hot zones are dynamically calculated with the DBScan algorithm

Uber

Plot_hot_zone function

Input	• Dataset, day of the week, hour, dbscan parameters
Calculation	• For any given hours, would calculate the clusters center and plot them
Output	• A map with the hot zones of this hour
Real-time visualization	• For a given day, the evolution of hot spots can be shown by looping over different hours

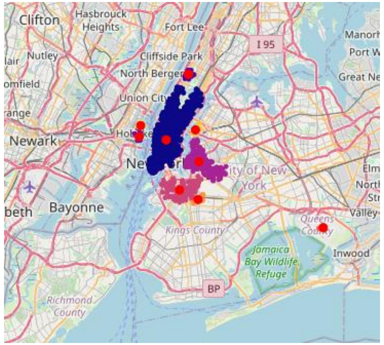
Hot zones can be dynamically calculated and provided to the drivers at any given day and time of the week

The hot zone are changing every hour of the day

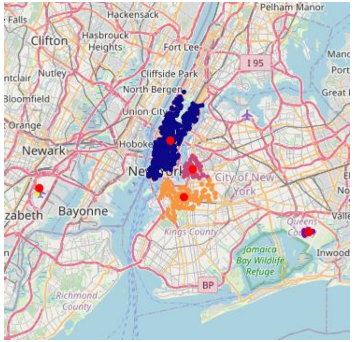
Uber

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

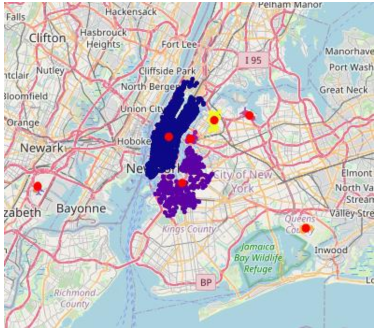
12am



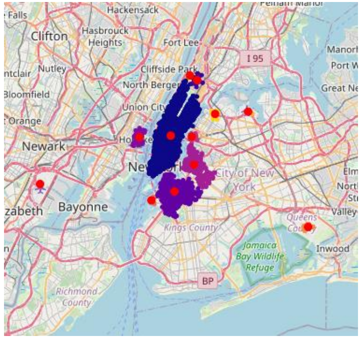
6am



12pm



6pm



Q&A

Thanks!