

# Restaurants Analysis —



## Second Project

**Big Data** Course - Prof.ssa Simona Colucci - 2020/2021

**Group 6:** Dario Di Palma - Marco Servadio

# Overview



# Big Data on restaurants


---

In the last years, the application of data science on restaurants are increasing:

- Evaluation of people to understand more about it
- Suggestion of food
- Suggestion of restaurant
- Recommendation system



# Outline

- With the growth of review sites such as TripAdvisor, a site for publishing reviews on hotels, bed and breakfasts and restaurants.
  - We can have a lot of data that can be used to create a recommendation systems.
- Analysis on Data
  - Collaborative Filtering

Results and Analysis



# Dataset and Tools

# The dataset

- It contains 9 data files
  - The collaborative filter technique use only one the rating\_final.csv
  - Other 8 files was used to analysis and extrapolation all useful information to understand specifically things in the dataset
- 
- Example of the structure in rating\_final

▲ userID	🔍 placeID	# rating	# food_rating	# service_ra...
U1077	135085	2	2	2
U1077	135038	2	2	1
U1077	132825	2	2	2

kaggle



# Methods and libraries

- Collaborative Filtering (MLlib)
  - Map Transformation and Reduce Action
  - Pandas
  - Matplotlib
  - Seaborn
  - SciPy
- } data visualization and analysis tool



# Jobs, Results and Comments

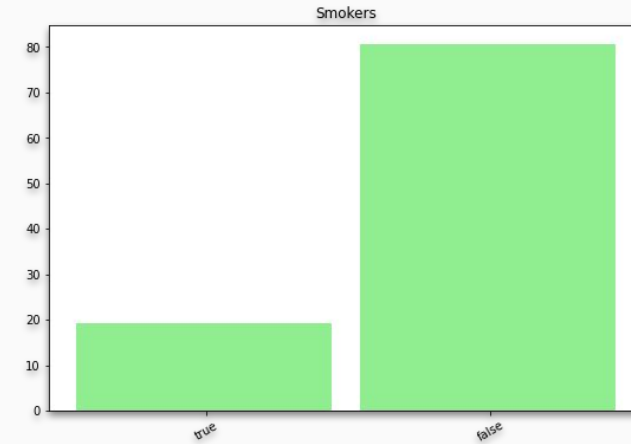
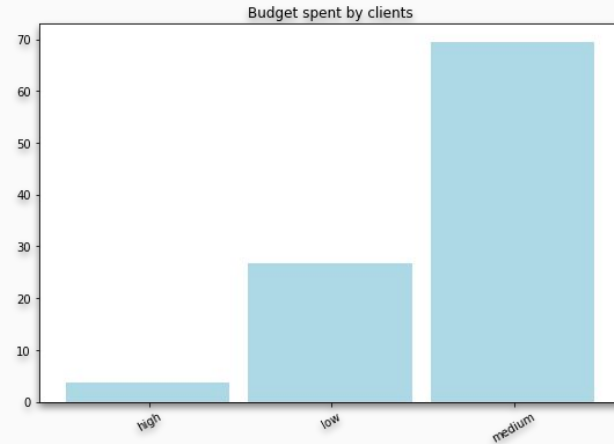
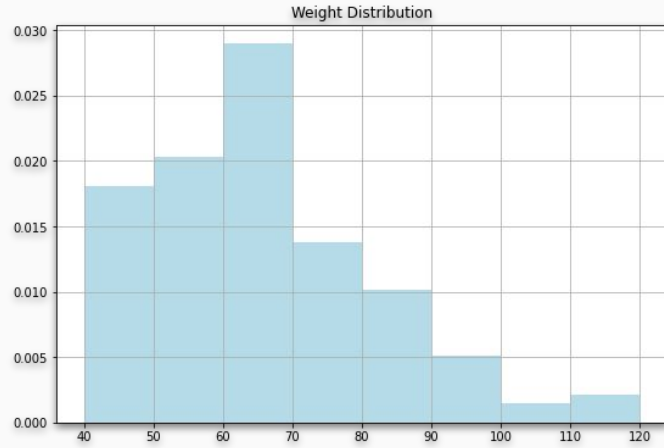


# 1. Preliminary analysis on restaurants and consumers.

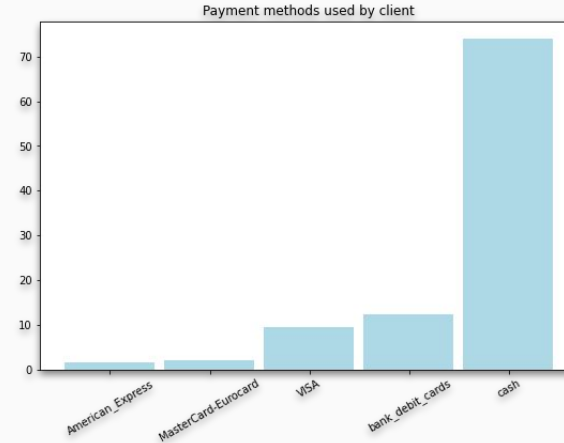
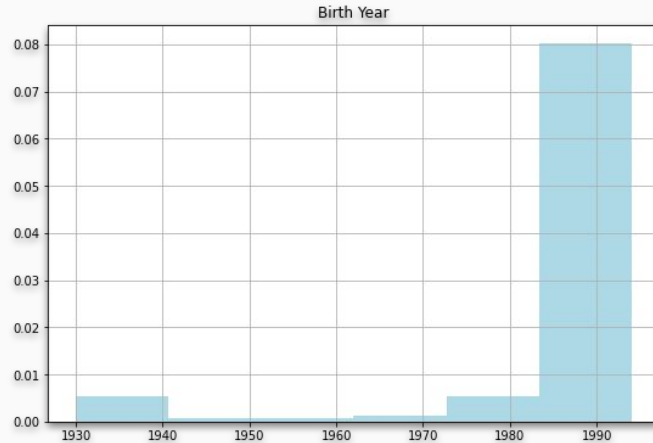
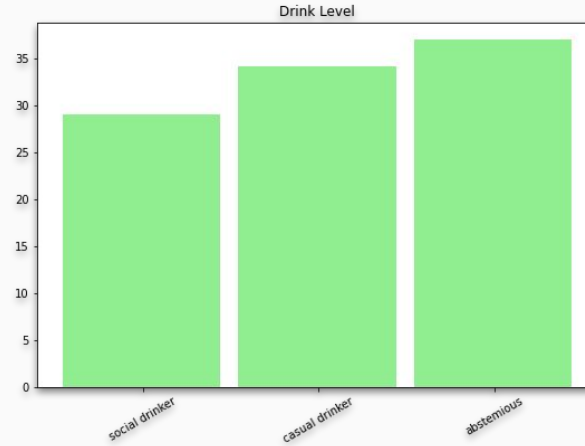
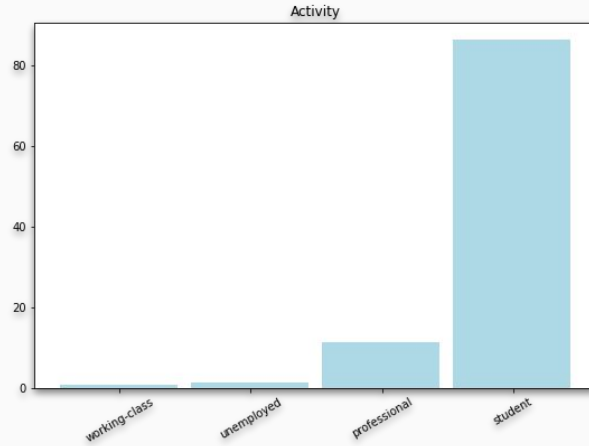
- 
- **Problem:** Understand basic info about our customers and restaurants to learn more about the dataset, analyzing the average user and restaurant features in our data.
  - For the customers we considered:
    - Weight, Height, Budget, Smoke, Activity, Drink level, Payment methods, Birth year, Location and Type of cuisine
  - For the restaurants we considered:
    - Payment methods, Drink served, Smoking rules, Price level, Location and Type of cuisine
- 



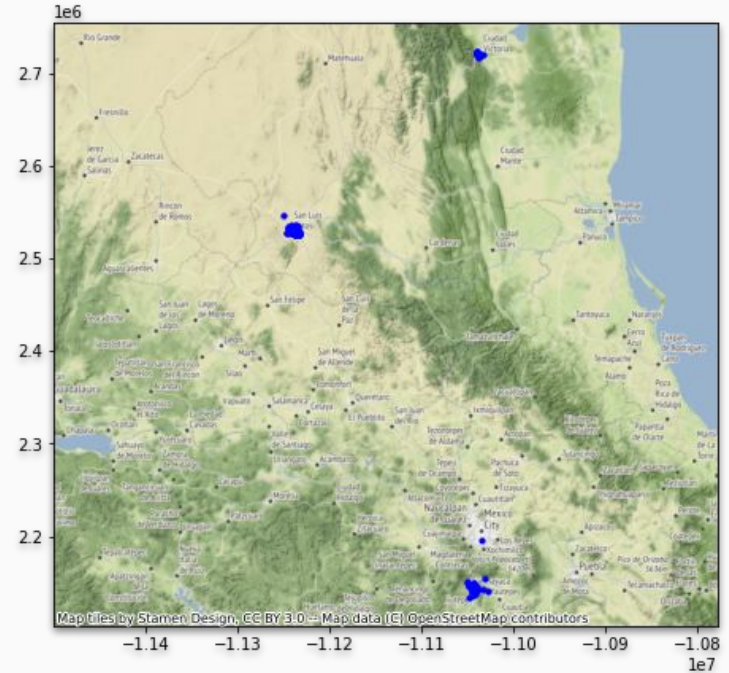
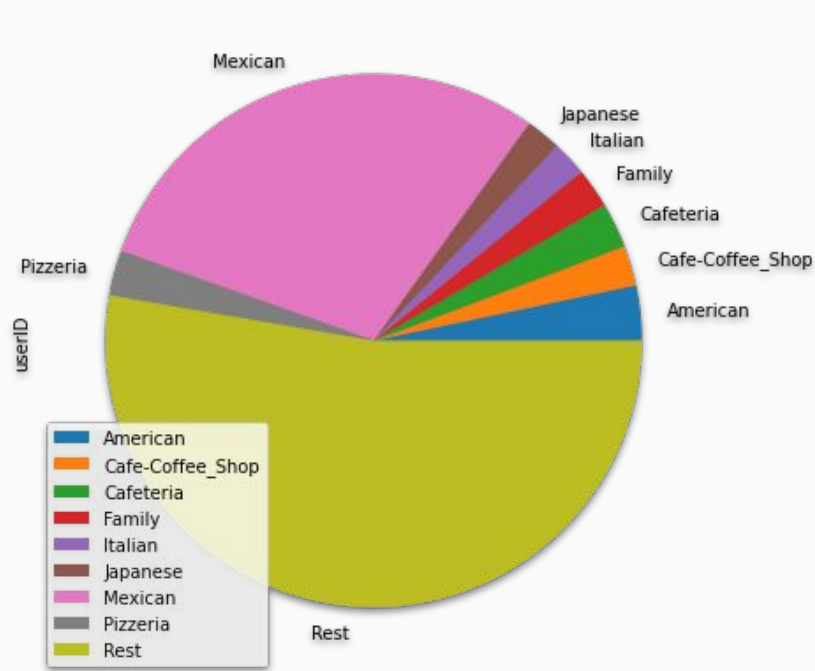
# Preliminary analysis on the customer base



# Preliminary analysis on the customer base



# Preliminary analysis on the customer base

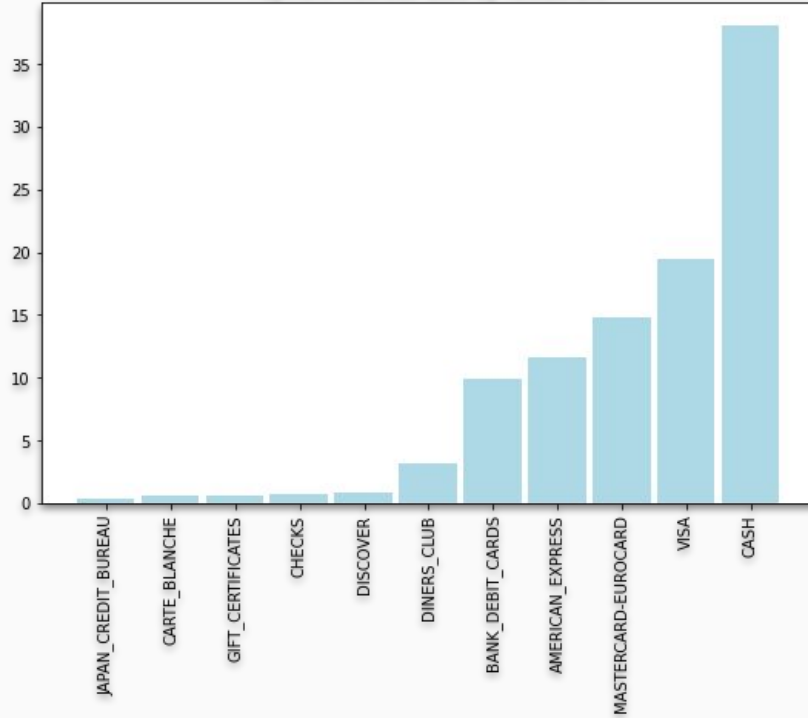


- We have an high probability that our average user is a Mexican, tall between 1.60-1.70m, weights between 60-70kg, is a non-smoking student and with an high probability of drinking, that he was born in around 1990 and that he pay mostly in cash.

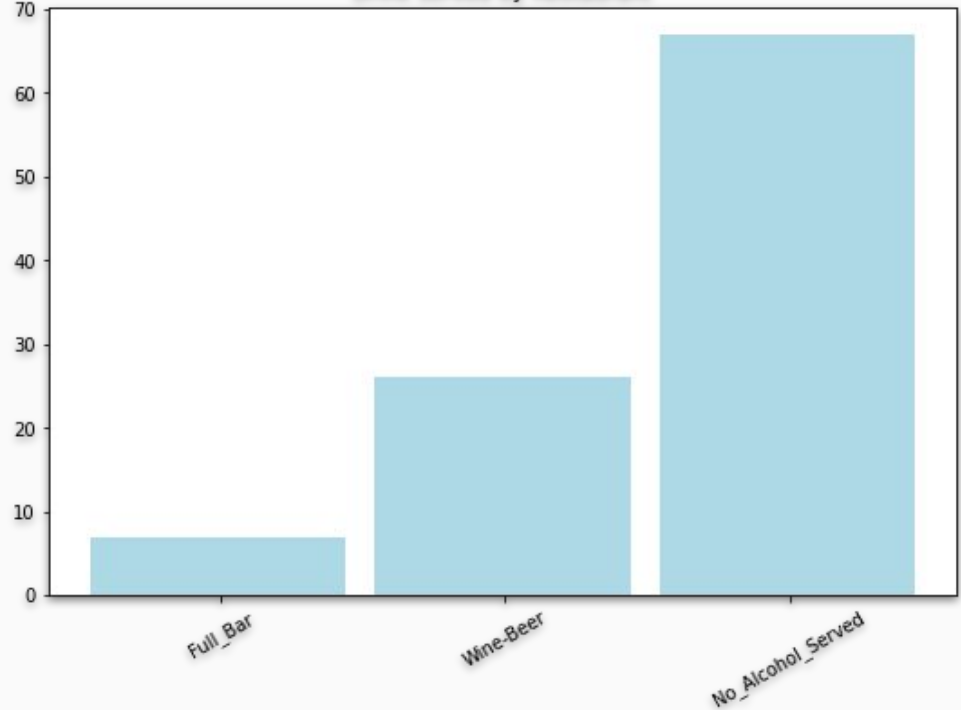


# Preliminary analysis on the restaurants

Payment methods used by Restaurant

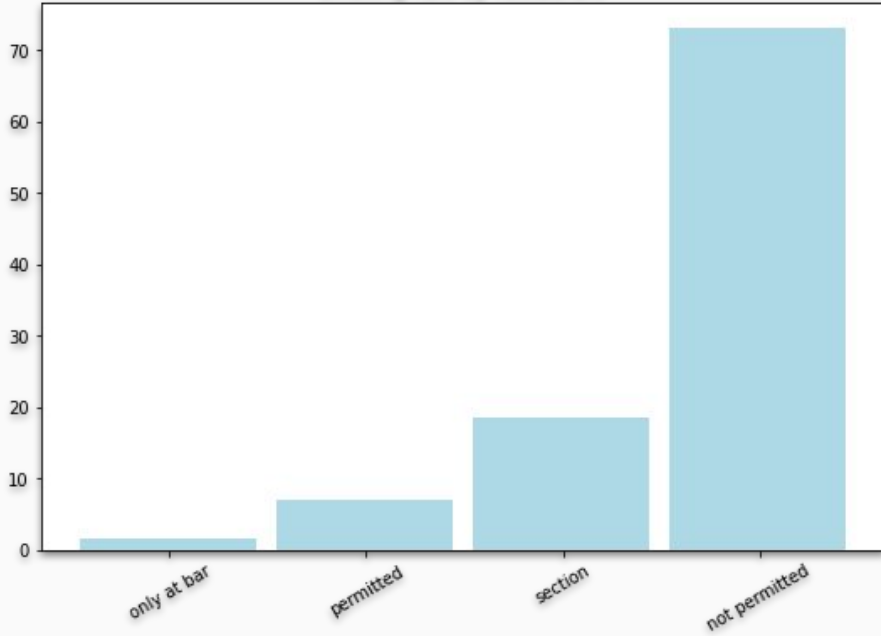


Drink served by Restaurant

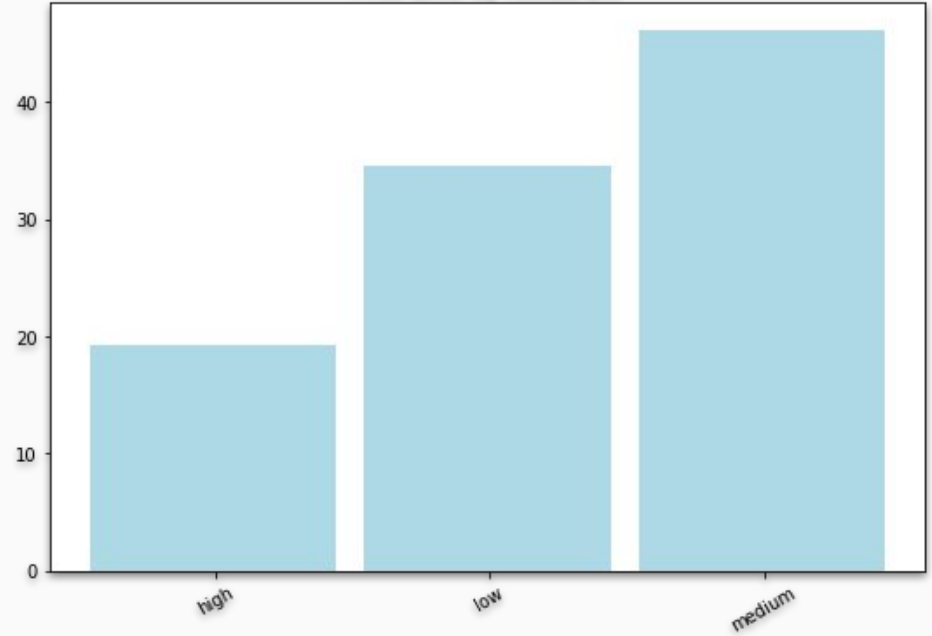


# Preliminary analysis on the restaurants

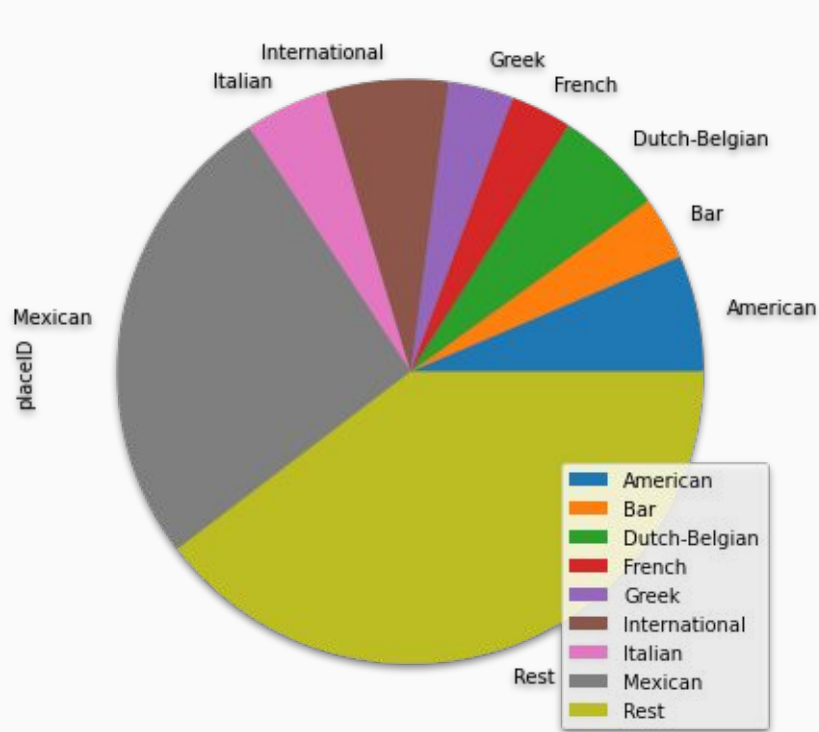
Smoking Rule by Restaurant



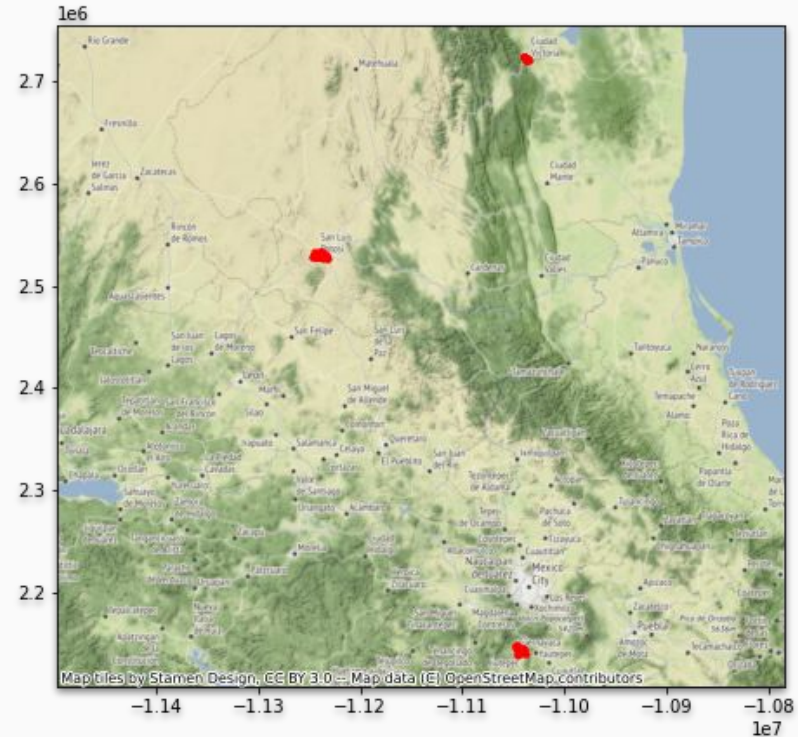
Price Level by Restaurant



# Preliminary analysis on the restaurants



- Being in Mexico it is quite obvious that the prevalence of restaurants is focused on Mexican cuisine.



- Restaurants are located mostly in the three city near customer analyzed before.



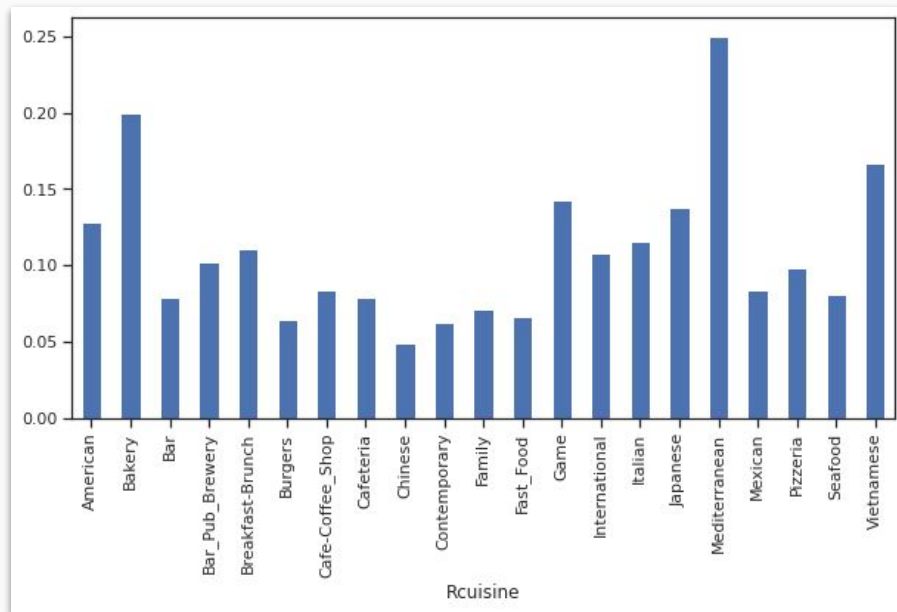
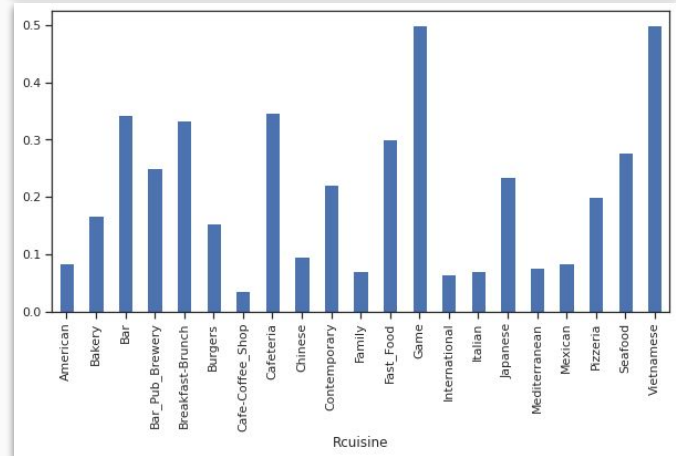
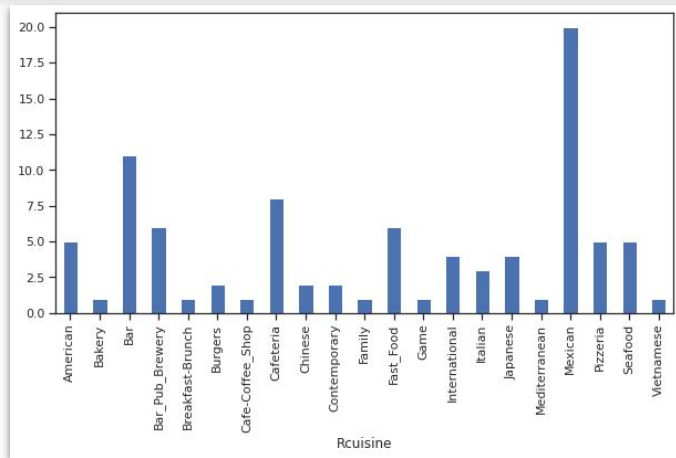
## 2. Relation between certain features of a restaurant and the received rating

- 
- **Problem:** Understand if certain features have a weight in the liking of a restaurant
    - Research of a possible relation between the best rated restaurants and the kind of cuisine they serve
    - Research of a possible relation between availability of parking and service rating
- 





## 2.1 Research of a possible relation between the best rated restaurants and the kind of cuisine they serve



**First attempt:**

absolute data

**Second attempt:**

normalization by number of restaurant

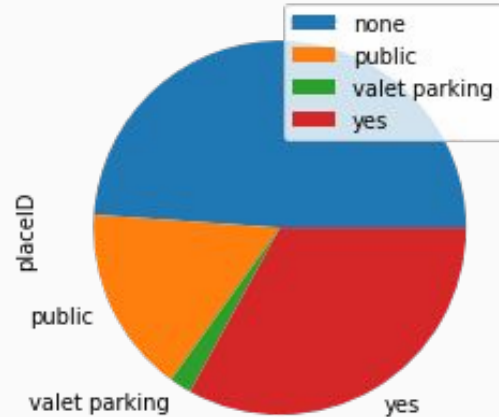
**Third attempt:**

normalization by number of reviews

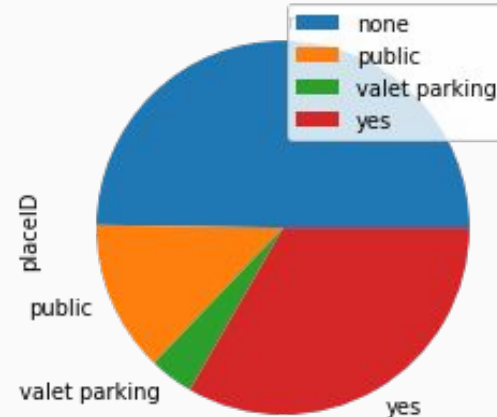


## 2.2 Research of a possible relation between availability of parking and service rating

Parking in restaurants with minimum score



Parking in restaurants with maximum score



- **Conclusion:** parking and service ratings are mostly incorrellated.
- The only small change we were able to find was in the valet parking service percentage. However this service is often offered by high profile restaurants, so it's natural that the service score is generally higher in these kinds of places.

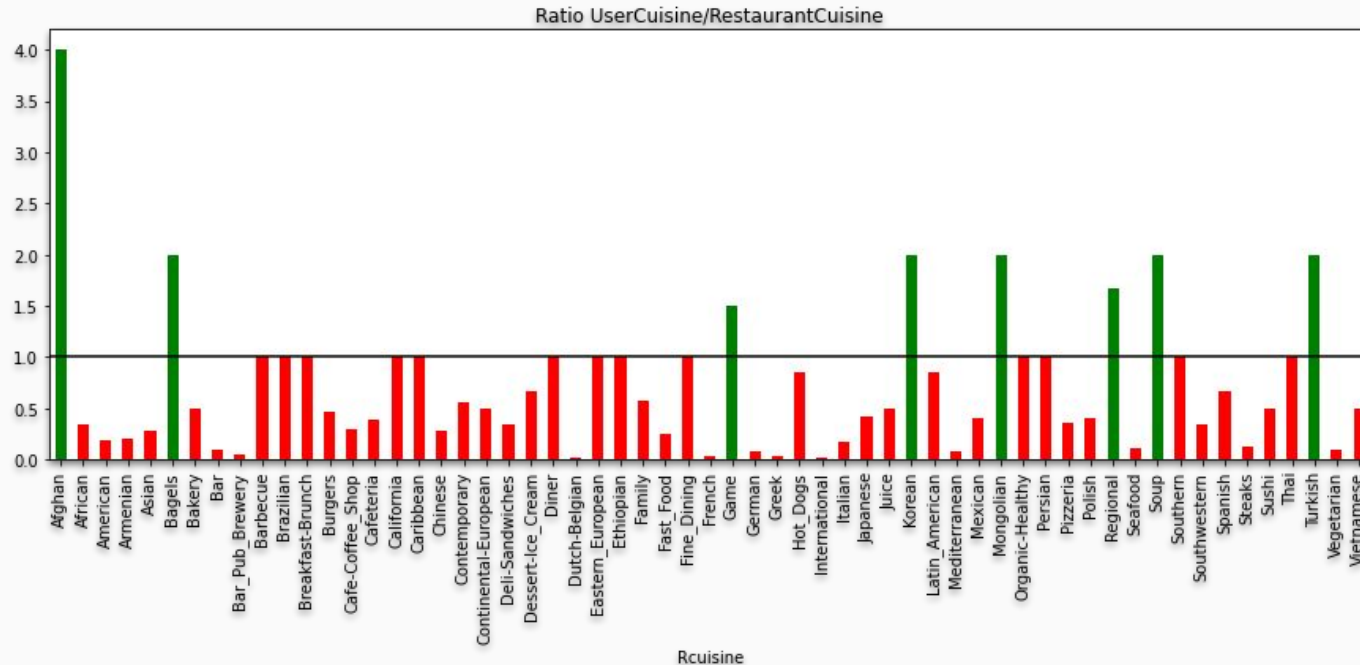


### 3. Supply and demand analysis for each kind of cuisine

- 
- **Problem:** Assuming the basic economic rule of supply and demand, look for which cuisines there is more demand than supply, so that it is possible to have a good return on investment.
- 



### 3. Supply and demand analysis for each kind of cuisine



- The green bars in the graph represent the kinds of cuisine for which the demand is not metted quite enough.
- In a purely business-economic analysis these are the types of restaurants which have less competitors.
- We set threshold ratio at 1.0, which seemed quite right to exclude a large part of the already saturated market.



## 4. Analysis on the ratings

- 
- **Problem:** Analyze using statistical and graphical operations, data about the ratings released by customers.
- 



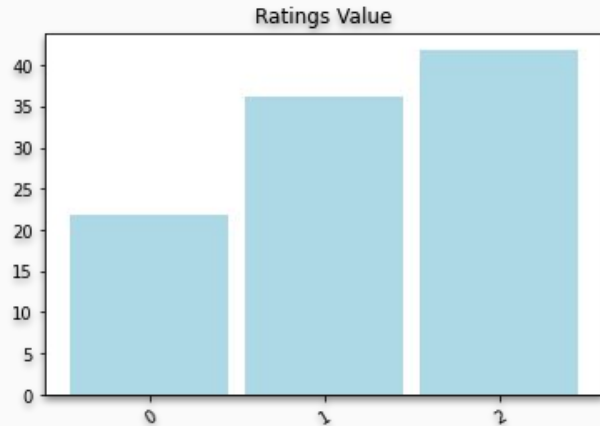
## 4. Analysis on the ratings

- Number of ratings and other statistical parameters
- Rate of usage for each score



Total Ratings = 1161  
Total Users = 138  
Total Places = 130

Min rating: 0  
Max rating: 2  
Average rating: 1.20  
Median rating: 1  
Average of rating released by user: 8.41  
Average of ratings receipts per places: 8.93

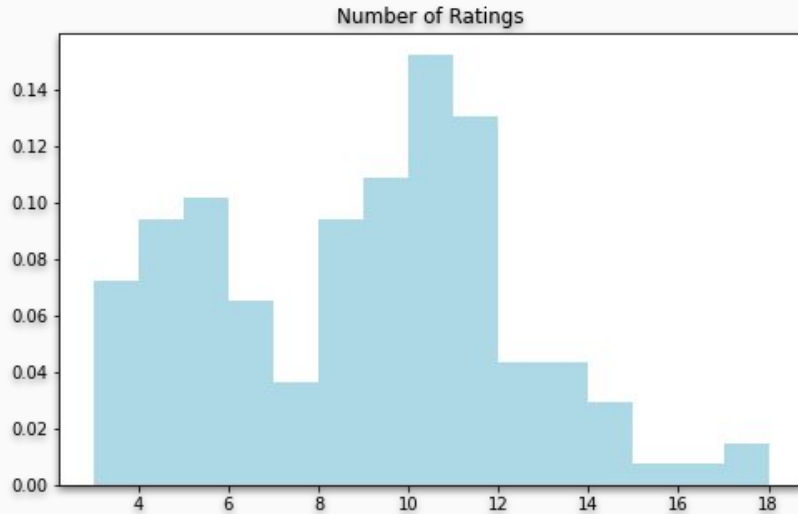


Here we can see that the ratings are well distributed, with a trend towards higher ratings.

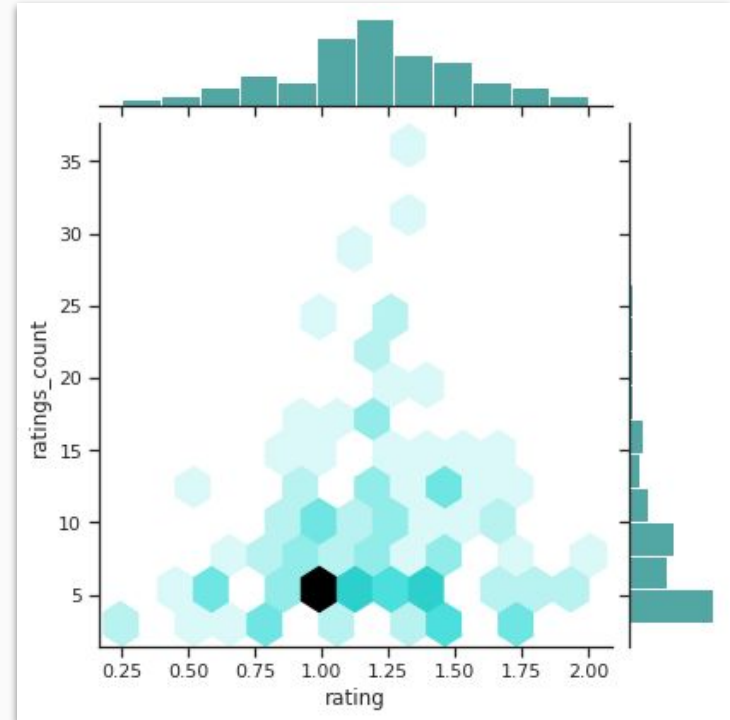


## 4. Analysis on the ratings

- Number of reviews for each customer



- Distribution of average ratings for each restaurant and number of ratings



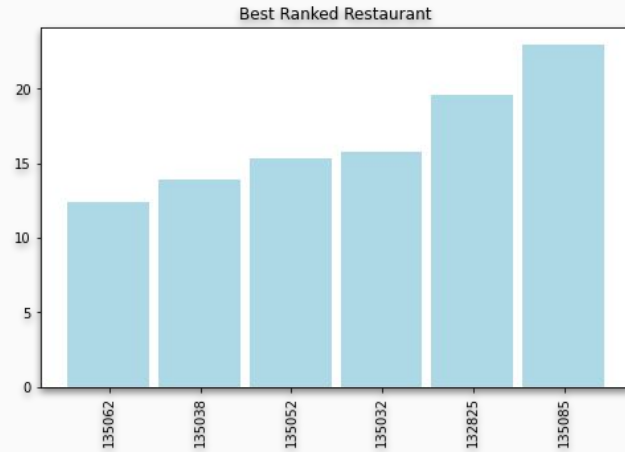
## 5. Recommendation system based on the most liked restaurant and type of cuisine

- 
- **Problem:** Give some basic recommendation to users based on the popularity (positive reviews) of restaurants.
    - Popularity based on absolute ratings among all restaurants
    - Popularity based on ratings for each kind of cuisine
- 





## 5. Recommendation system based on the most liked restaurant and type of cuisine



- Most liked restaurants
- Search for the most popular restaurants by type of cuisine

	Name	Cuisine	Rating
0	tacos los volcanes	American	1.66667
5	little pizza Emilio Portes Gil	Armenian	1.25
6	Chaires	Bakery	1.4
7	Restaurant Bar Hacienda los Martinez	Bar	1.66667
20	emilianos	Bar_Pub_Brewery	2
26	la parroquia	Breakfast-Brunch	1
27	Log Yin	Burgers	1.75
32	Preambulo Wifi Zone Cafe	Cafe-Coffee_Shop	1.58333
33	cafe punta del cielo	Cafeteria	1.83333
42	Restaurant Wu Zhuo Yi	Chinese	1.25
45	Restaurante la Parroquia Potosina	Contemporary	1.75
47	Mariscos Tia Licha	Family	1.6
49	tortas hawai	Fast_Food	1.33333
57	KFC	Game	1.42857
58	Restaurant Las Mananitas	International	2
62	El Mundo de la Pasta	Italian	1.5
66	Michiko Restaurant Japonese	Japanese	2
71	Log Yin	Mediterranean	1.75
72	La Estrella de Dimas	Mexican	1.8
99	Little Cesarz	Pizzeria	1.3
104	puesto de gorditas	Regional	0.5
105	Mariscos El Pescador	Seafood	1.69231
110	Restaurant Familiar El Chino	Vietnamese	1.16667






## 6. Recommendation system using Collaborative Filtering with MLlib

- 
- **Problem:** Create a recommendation system based on Collaborative Filtering using Machine Learning Approach
    - Using explicit feedback released by customers in form of reviews
    - Using the Alternating Least Squares method
- 



# Collaborative Filtering



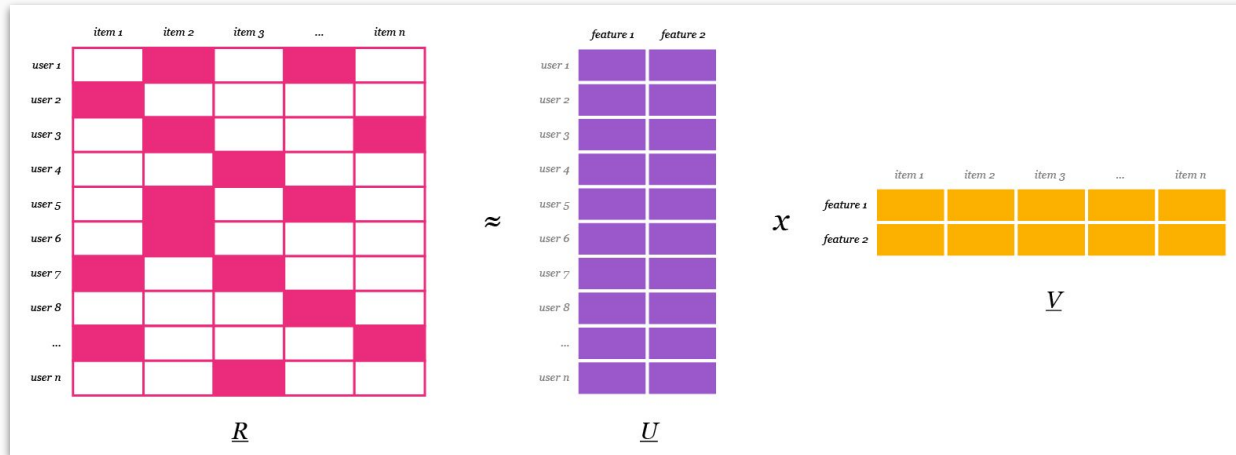
		Cuisines				
						
C u s t o m e r s	Marco	1	2	2	1	
	Dario		2		1	0
	Alex	1	0		2	
	Elettra	1		0	2	X
	Paola		2			2
	Oriana	1		2	1	

Prediction

- Based on the concept of “homophily” - similar people like similar things. The goal is to predict a user’s preferences based on the feedback of similar users.

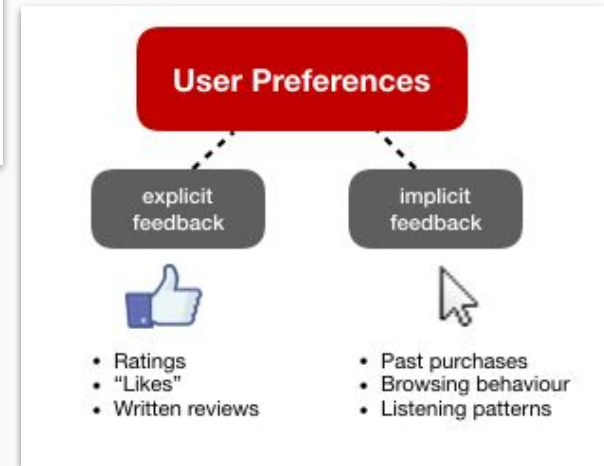


# Algorithm and User Preferences

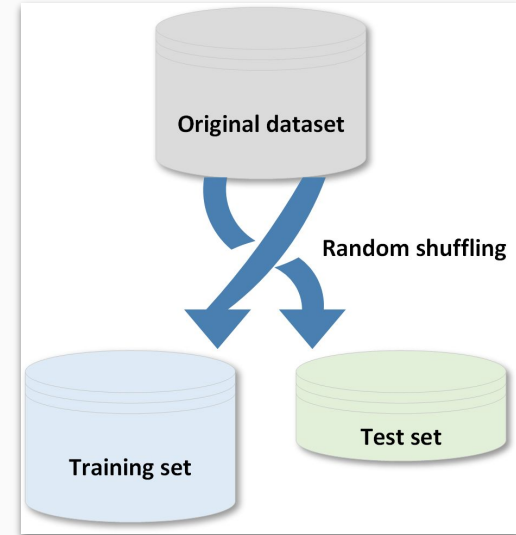
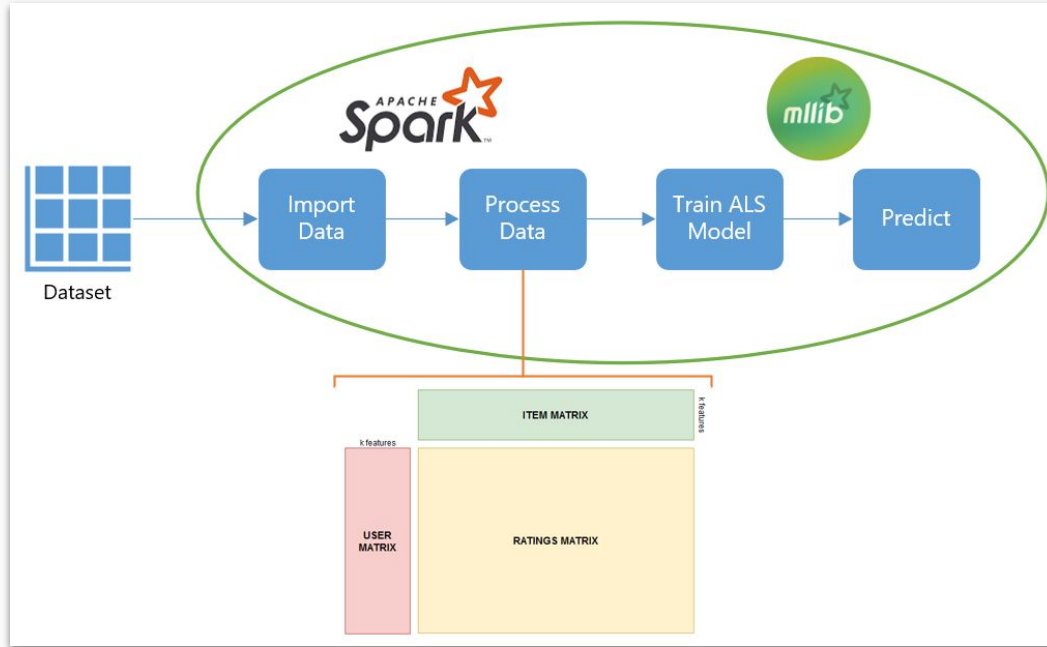


- In our model we used explicit preferences (given by the user), this fits well with our dataset where we have reviews.
- In real-world applications implicit feedback is more used (ratings gained from a data mining pipeline), but the results are not as good, because it adds another degree of uncertainty.

The ALS algorithm applies an approximate factorization to the sparse User-Item-Rating matrix, discovering a number (rank) of hidden features that can be used to make predictions.



# Pipeline and data splits generation



Starting from the dataset we do some processing to extract the needed features:

1. Import the dataset
2. Transform RDD using the rankings for training the ALS Model
3. Split the Original dataset to test the prediction



# Top 10 recommended

- Top 10 restaurants suggested for a user

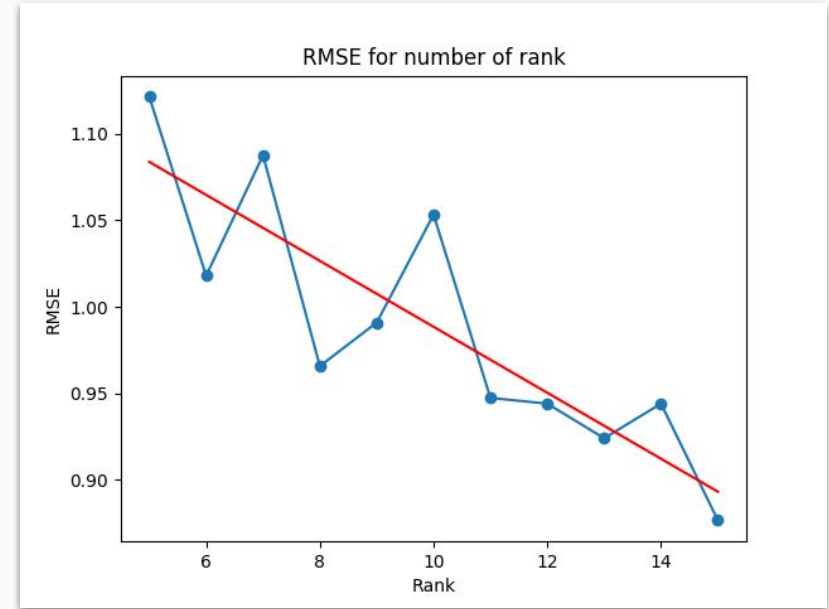
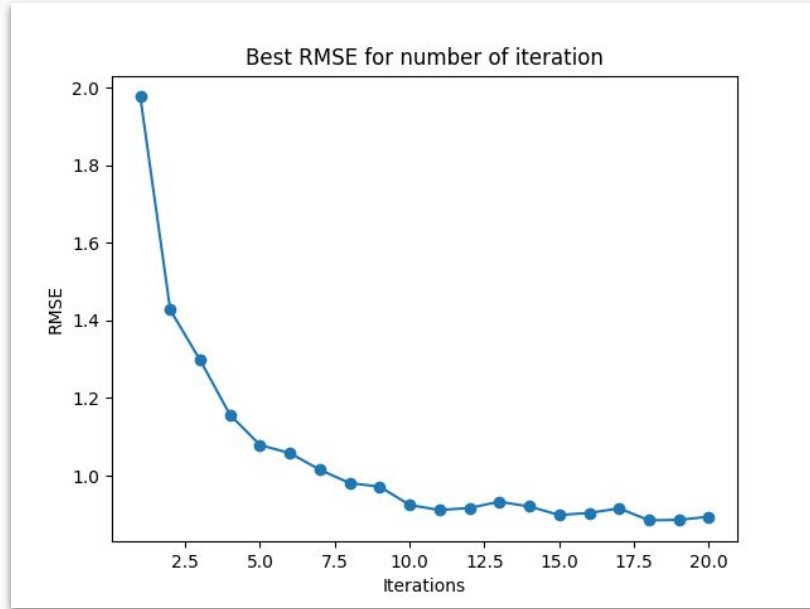
User	Restaurant	Rating
1077	135025	2.16364
1077	135042	2.03145
1077	132825	1.98648
1077	135075	1.93129
1077	135041	1.86017
1077	135057	1.81492
1077	132955	1.73593
1077	135032	1.73527
1077	132958	1.69776
1077	135072	1.69612

- Top 10 users suggested for a restaurant

User	Restaurant	Rating
1109	135108	2.16664
1053	135108	2.04773
1054	135108	2.0391
1111	135108	1.98165
1088	135108	1.98087
1126	135108	1.97576
1078	135108	1.96684
1037	135108	1.94618
1108	135108	1.77642
1071	135108	1.76059



# Evaluation of the model using the A/B testing



- **Root Mean Squared Error (RMSE)** was used to analyze the effect of changing Rank and Number of iteration on the system.
- Best value for not saturate the memory and obtain a good results it is 10 iterations and 11 rank



# Execution times

- 
- **Total Notebook execution time:**  $\approx 66$  sec
  - **Model Training rank = 11 and iterations = 10 :**
    - CPU times:** user 21.2 ms, sys: 87  $\mu$ s, total: 21.3 ms
    - Wall time:** 3.29 s
  - **Top-K Execution time:**
    - Top-10 restaurants for userID:**
      - CPU times:** user 5.62 ms, sys: 171  $\mu$ s, total: 5.79 ms
      - Wall time:** 222 ms
    - Top-10 customers for placeID:**
      - CPU times:** user 708  $\mu$ s, sys: 4.16 ms, total: 4.86 ms
      - Wall time:** 114 ms
- 





# Future developments

---

Use a recommendation algorithm based on Content-based, trying to deepen the recommendation systems more specifically, also by integrating state-of-the-art algorithms

# THANKS FOR WATCHING

