IBM Data Science Professional Certificated

Capstone Project

Final Report

Zinan Lin

May 21st, 2019

Table of Contents

[Introduction 3](#_Toc9390407)

[Background 3](#_Toc9390408)

[Problem Scope 3](#_Toc9390409)

[Audience 3](#_Toc9390410)

[Datasets 4](#_Toc9390411)

[Data of Michelin Starred Restaurants 4](#_Toc9390412)

[Source 4](#_Toc9390413)

[Description 4](#_Toc9390414)

[Example 4](#_Toc9390415)

[Data of arrondissements in Paris 5](#_Toc9390416)

[Source 5](#_Toc9390417)

[Description 5](#_Toc9390418)

[Example 6](#_Toc9390419)

[Data of Restaurant Information 6](#_Toc9390420)

[Source 6](#_Toc9390421)

[Description 6](#_Toc9390422)

[Example 7](#_Toc9390423)

[Data from Foursquare API 7](#_Toc9390424)

[Source 7](#_Toc9390425)

[Description 7](#_Toc9390426)

[Example 8](#_Toc9390427)

[Methodology 8](#_Toc9390428)

[Data Collecting 8](#_Toc9390429)

[Data Wrangling 9](#_Toc9390430)

[Data Visualization 10](#_Toc9390431)

[K-means Clustering 12](#_Toc9390432)

[Results 14](#_Toc9390433)

[Discussion 15](#_Toc9390434)

[Conclusion 15](#_Toc9390435)

[Reference 15](#_Toc9390436)

# Introduction

## Background

Paris is the captain city of France, it is famous for its history, arts, and food. Paris has 20 arrondissements, which are regions that the city is separated into, like the picture [1] showed below. There are about 40,000 [2] restaurants in Paris and it is currently the city with the second most Michelin starred restaurants in the world [3]. I am a food lover and I am motivated to find out more about the fine-dining places in Paris.

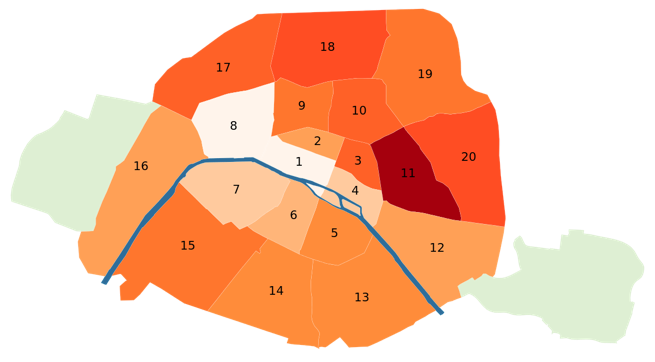


Fig. 1: The arrondissements of Paris

## Problem Scope

What are the similarities in the arrondissements of Paris in terms of restaurants?

Are the restaurants in the central (downtown) areas more favored by clients?

Optionally, how does the number of Michelin Starred restaurants affect the way that that restaurants are clustered?

We will need to find out about the above problems with the support of datasets.

## Audience

The audience of this problem are tourist, who planned to stay in Paris for vacation. Knowing which area/ areas are similar in terms of restaurants can help them deciding on where to live and where to eat.

This problem can also benefit restaurant owners/ businessmen who are interested in knowing the distribution of restaurants in Paris

# Datasets

## Data of Michelin Starred Restaurants

### Source

List of all Michelin starred restaurants:

<https://www.theupcoming.co.uk/2019/01/21/all-the-paris-michelin-star-restaurants-2019-on-a-map-and-full-list/>

### Description

Either using beautiful soup to scrape the web page or manually download the lines into an excel file. The data size is small (~100 rows, 2 columns) and the webpage is not very well organized, therefore it’s easy to manually create an excel file.

### Example

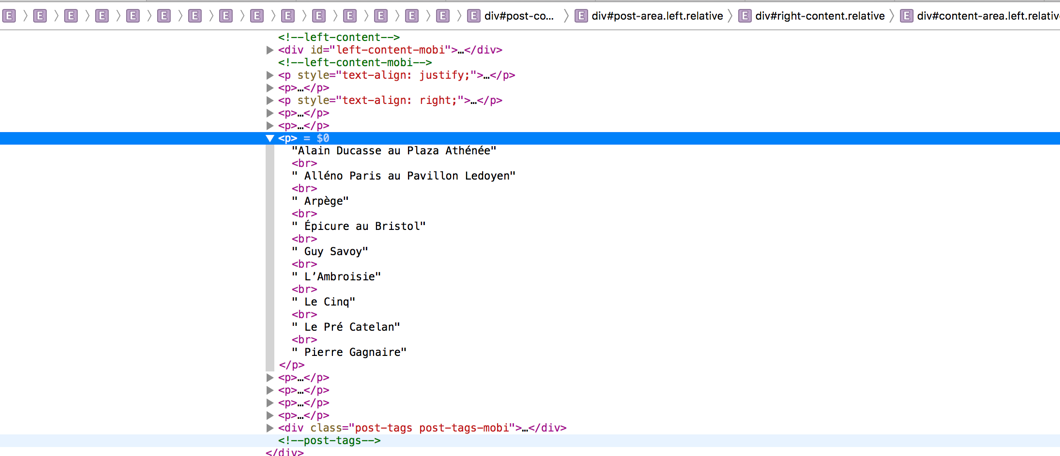


Fig 2. The website HTML, showing the labels of Michelin starred restaurants



Fig 3. Manually downloaded data

## Data of arrondissements in Paris

### Source

List of arrondissements in Paris

<https://opendata.paris.fr/explore/dataset/arrondissements/export/>

### Description

The data is nicely formatted from the official website of Paris open Data. It supports download, so I will import the data as a csv file using pandas.

### Example

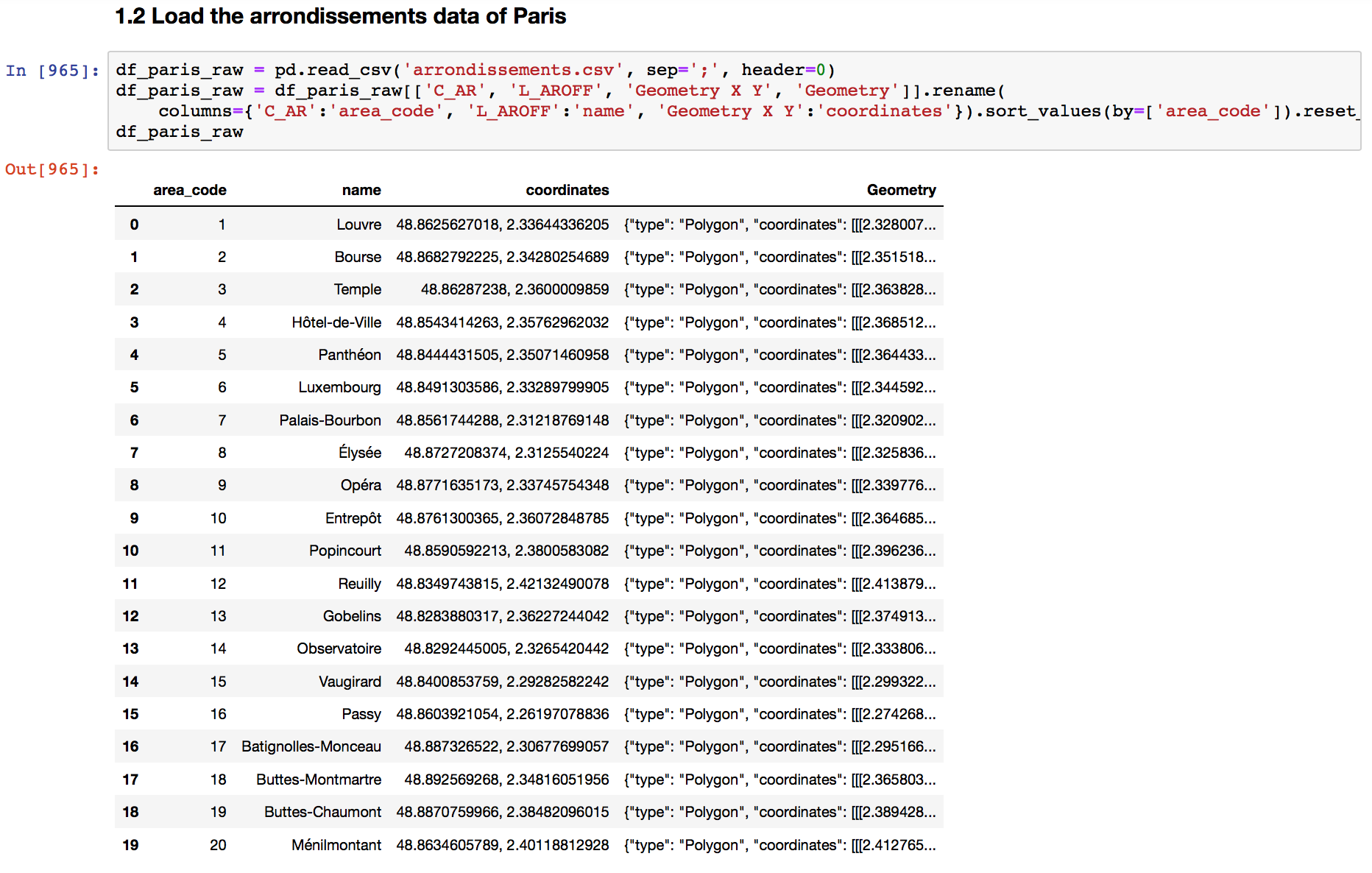


Fig 4. Code Snippet of importing arrondissement data as a data frame

## Data of Restaurant Information

### Source

Kaggle, TripAdvisor Restaurants Info for 31 Euro-Cities:

<https://www.kaggle.com/damienbeneschi/krakow-ta-restaurans-data-raw>

### Description

This is a dataset consisting Ratings and reviews for restaurants across 31 European cities, it’s 28.7MB. I will download it and import it as a CSV file. Later on, I will filter out only restaurants from Paris and use it to find out the details about restaurants I will be analyzing.

### Example



Fig 5, High level Overview of the data

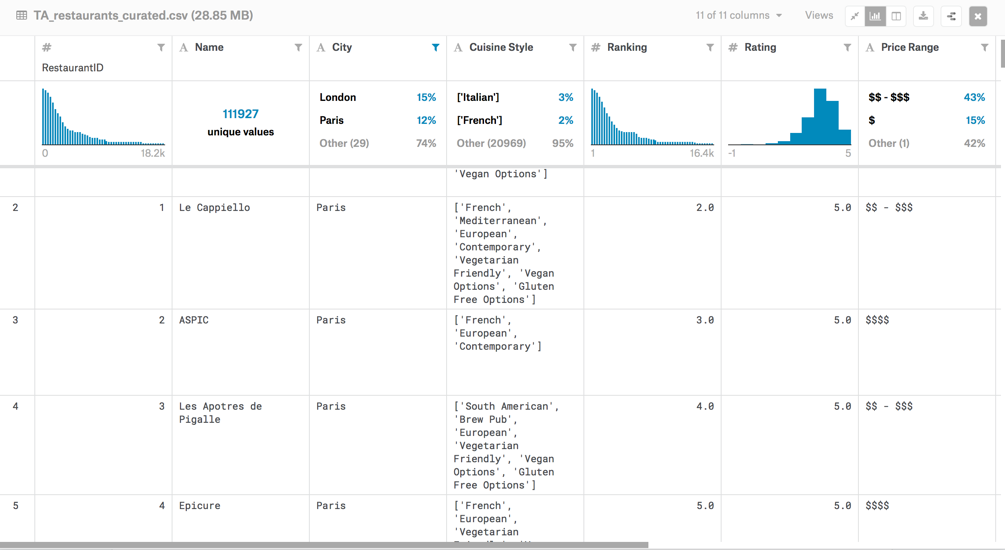


Fig 6, Example of the data

## Data from Foursquare API

### Source

Foursquare API:

<https://foursquare.com/developers/apps>

### Description

Using Foursquare API and it’s search endpoint, I will be able to get data on which arrondissement are restaurants from. As my current subscription, I can make 99,500 Regular Calls / Day, which should be more than enough for me as the total number of restaurants in Paris is about 40,000 [2].

### Example

Documentation for search endpoint: <https://developer.foursquare.com/docs/api/venues/search>



Fig 7, Code snippet of the function that makes call to the API

# Methodology

## Data Collecting

I am working with four data frames, three of which are directly imported from files: they are df\_michelin, df\_paris and df\_restaurants.

df\_venues were from Foursqaure API, by specifying food = '4d4b7105d754a06374d81259', I was able to only get food category from the venues near the input coordinates --- the input coordinates are the coordinates of the arrondissements from df\_paris.

Please look at the link to the notebook [4] for more about the code of data collection.

## Data Wrangling

I started by cleaning the data, cut off unwanted columns in each data frame and drop NaN values where I need the label for future analysis.

Other than eliminating anomalies, data normalization is important as well. In df\_restaurant, there are “ratings” and “number of reviews”. Using “rating” alone will not be good enough because, for example, a restaurant is rated 5 by 3 users is not necessarily better than a restaurant that is rated 4.5 by 30 users. Therefore, I will need to normalize the rating based on the number of reviewers.

I used Bayesian estimate of the weighted review as following:

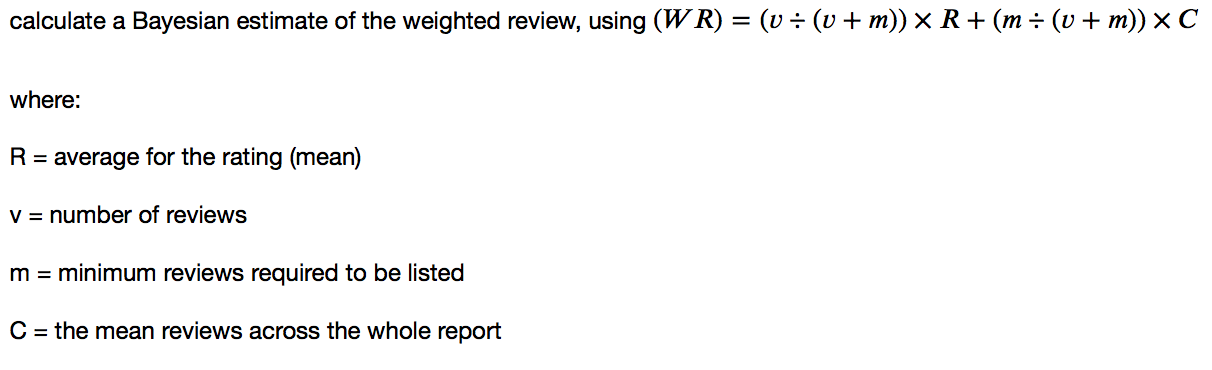


Fig 8, Normalization equation [5]

Rating before the Normalization:

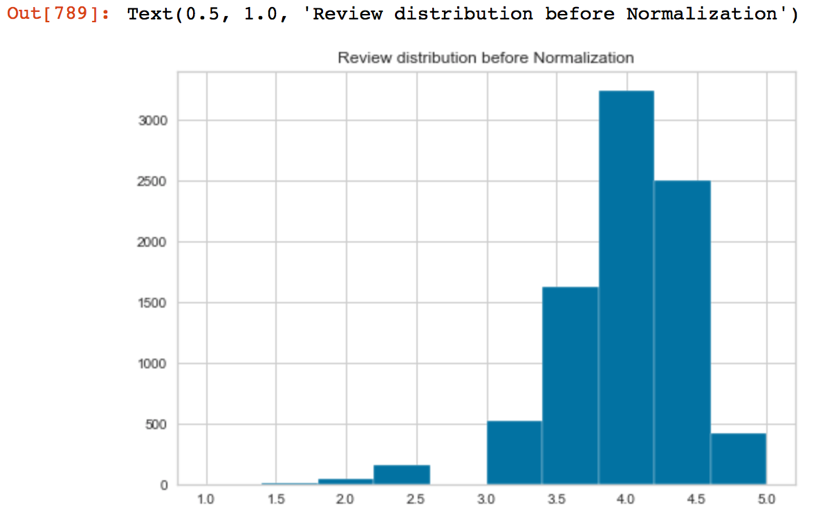


Fig 9, Rating Histogram before the Normalization

Rating after the Normalization:

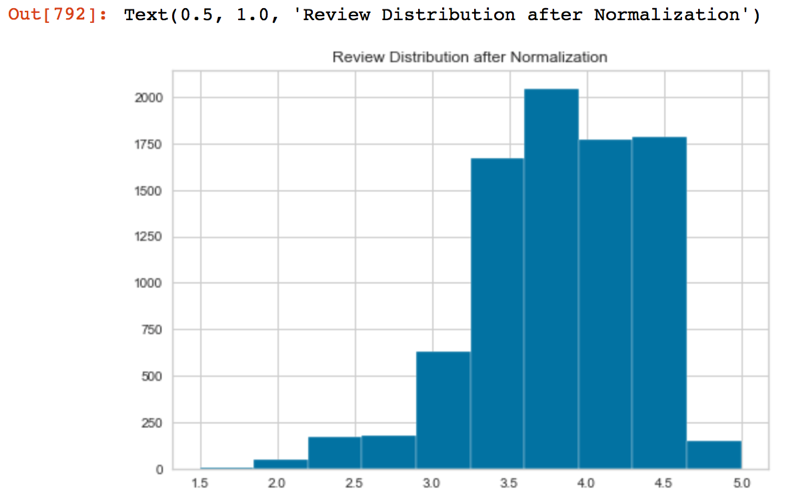


Fig 10, Rating Histogram after the Normalization

Distribution of the Number of Reviews:

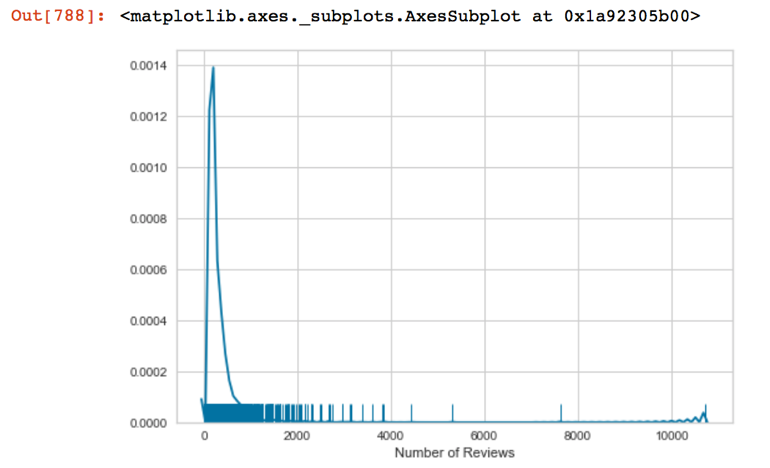


Fig 11, Distribution of the Number of Reviews

## Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. [6]

In order to get a better understanding of the datasets, I employed some data visualization in the data exploring process.

For df\_paris, I used folium to visualize the map with each arrondissement outlined:

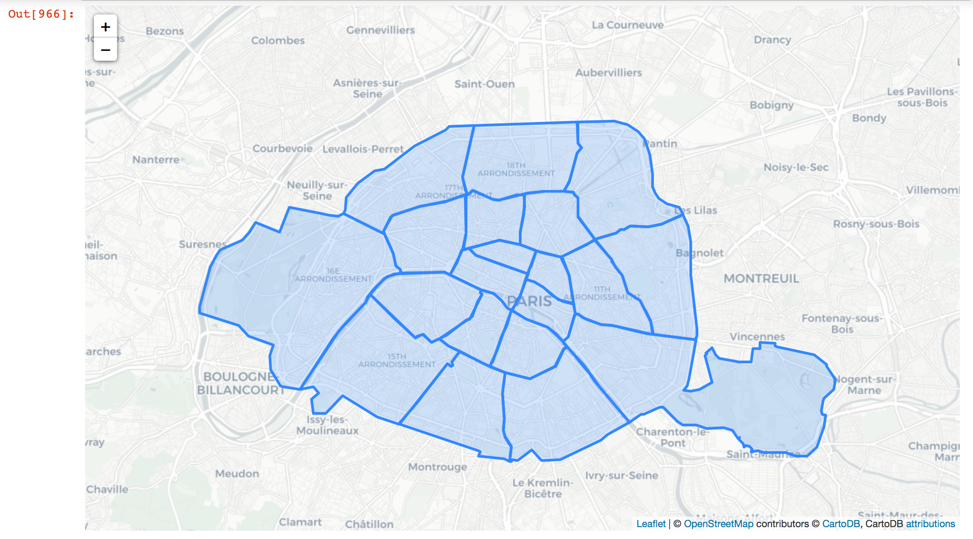


Fig 12, Map of Paris

For ratings of the restaurants, I plot rating with box plot in ascending order against each arrondissement to have an intuitive view of the data.

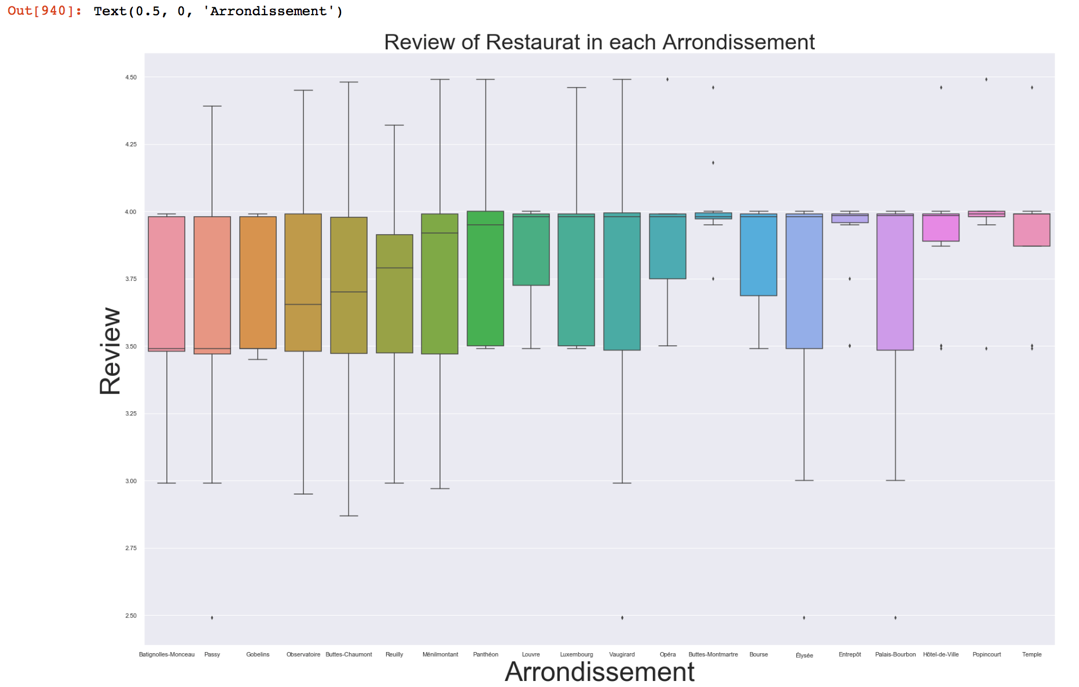


Fig 13, Box plot

For the percentage of the distribution of Michelin starred restaurants, I chose to use a pie plot. The result wasn’t satisfying but I will cover the reason in the discussion section.

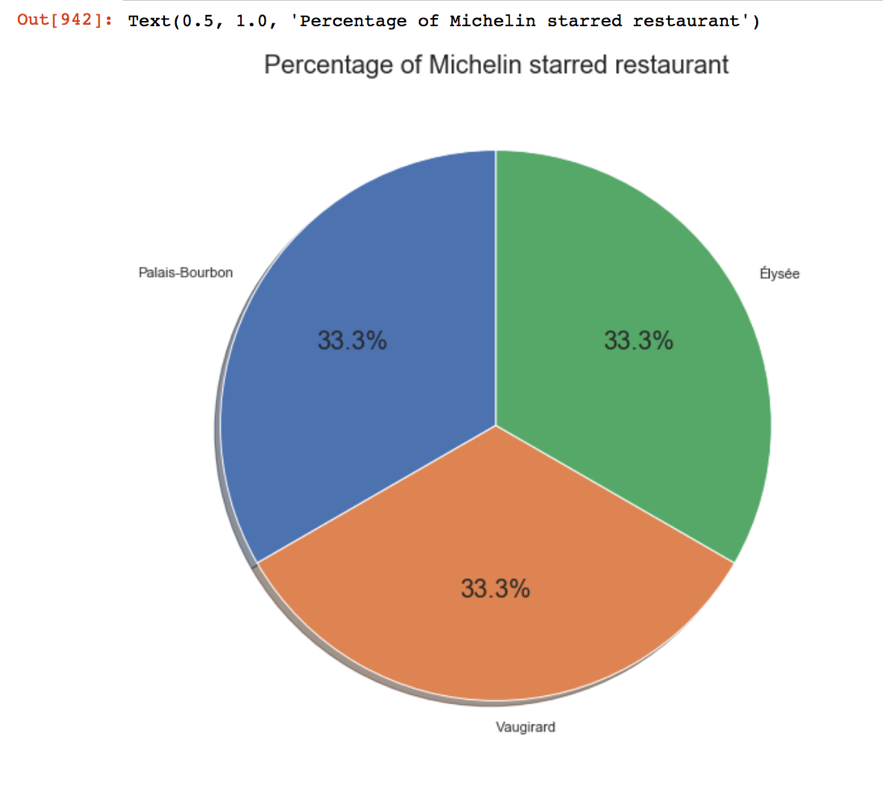


Fig 14, pie plot

Another Example of data visualization would be the plot for finding elbow point and visualize the clustered label on the map, which will be covered in the next section.

## K-means Clustering

k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. [6]

In this project, I aim to utilize k mean method to cluster restaurant data geometrically to find similar groups in terms of their location.

I first use one-hot encoding to separate categorized fields in binary, such as price range and normalized rating. Then I use the elbow method to find the optimal k, the Elbow method is a method of interpretation and validation of consistency within cluster analysis designed to help finding the appropriate number of clusters in a dataset. [7]

The code I used for plotting the elbow point graph comes from KElbowVisualizer library. Details can be found on <https://www.scikit-yb.org/en/latest/api/cluster/elbow.html>.

Here is the plot:

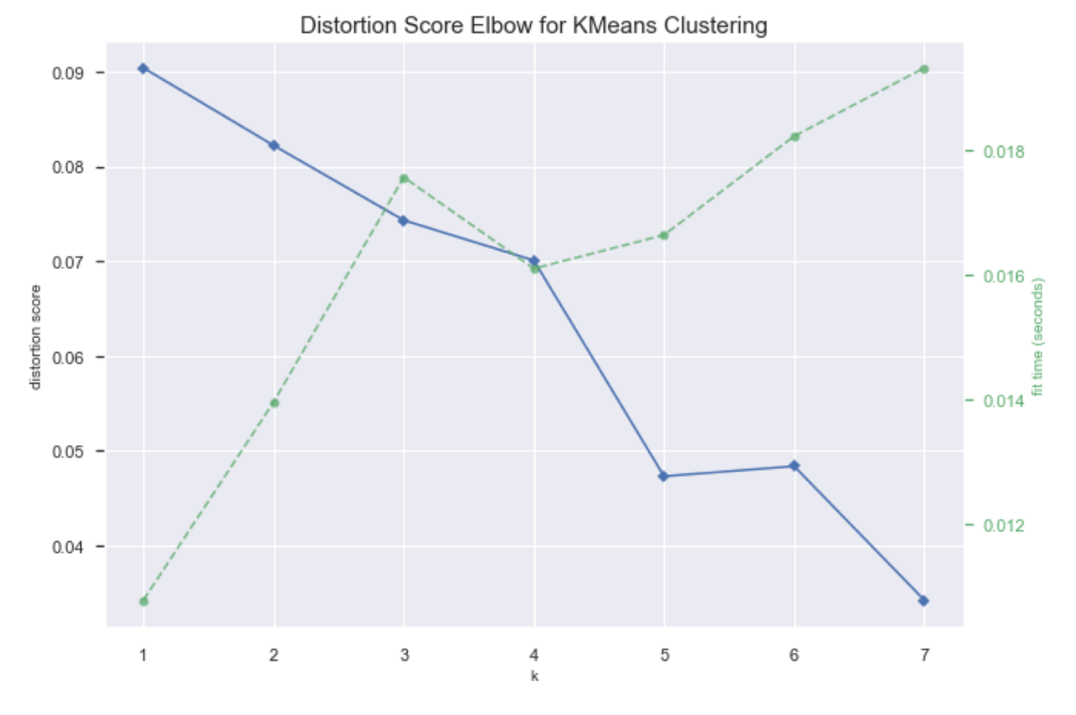


Fig 15, elbow point plot

It is not a pretty plot and the distribution score does not provide an obvious K, I chose k=5 to move on.

After using scikit library for k-mean cluster and adding the label back into the data frame, I made the following plot overlaying with Fig. 12 to see how the restaurants are clustered:

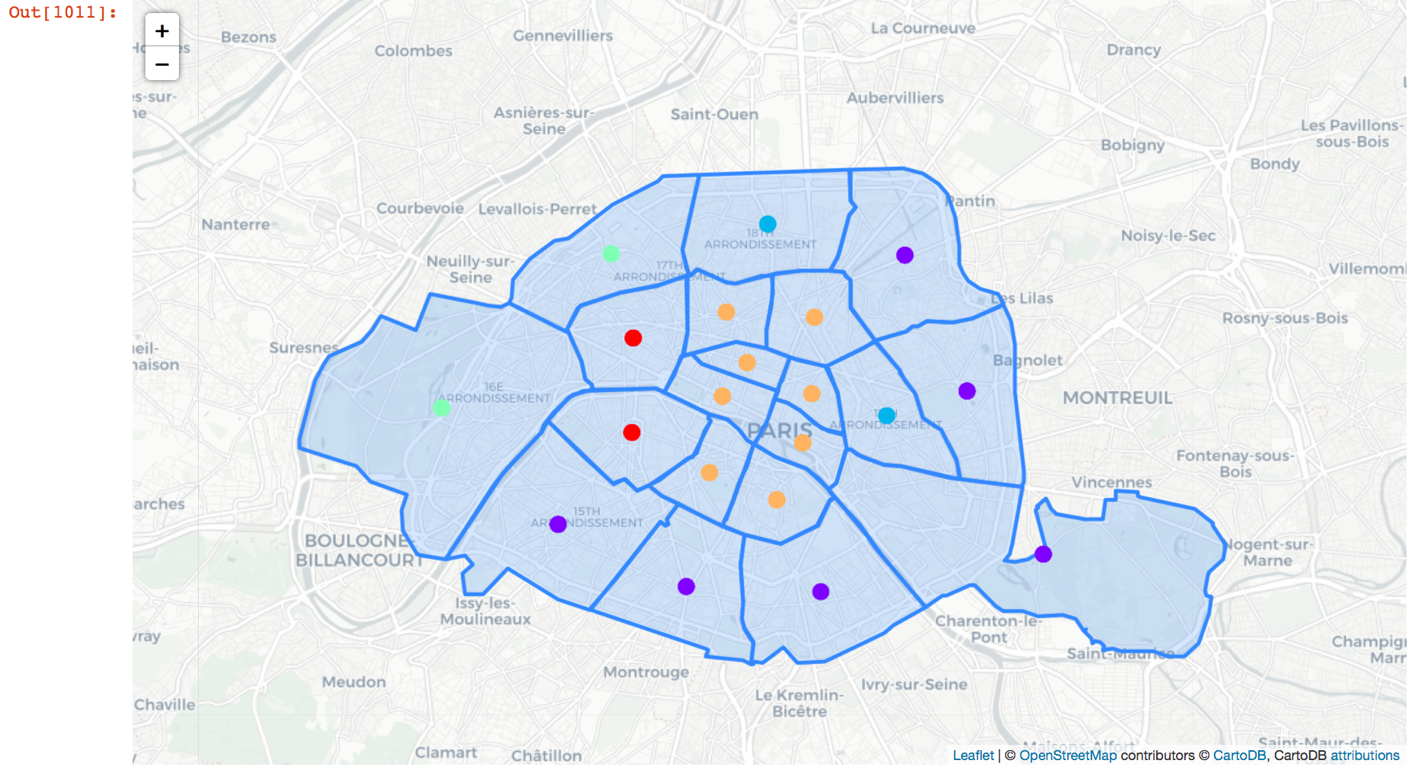


Fig 16, cluster label on Paris map

# Results

The results are quite intuitive with the plot above (Fig 16), the restaurants near the central city are clustered together and the ones further away from the city are clustered together.

If we take a closer look at the clustered results, we can see that the central city cluster tends to have more expensive and highly rated restaurants than the ones further away from downtown:

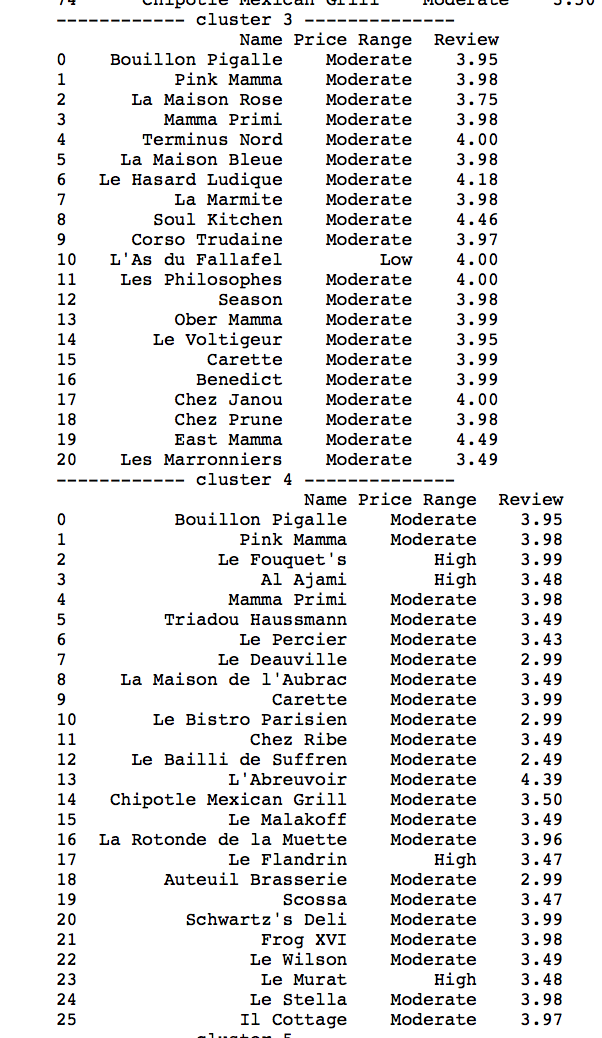
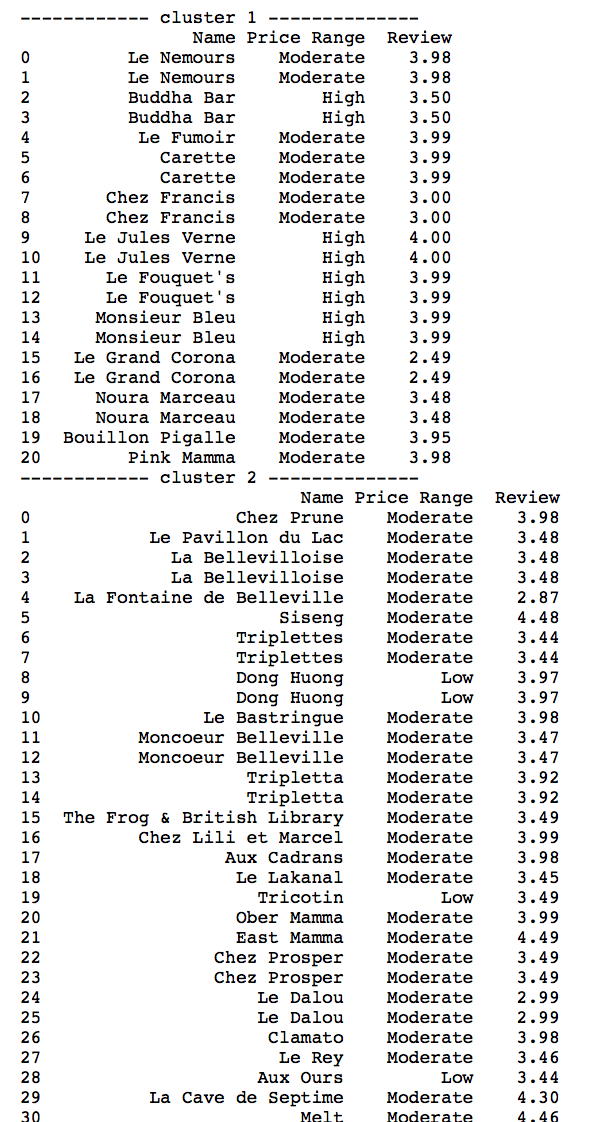


Fig 17, details of each cluster

To answer the following question:

What are the similarities in the arrondissements of Paris in terms of restaurants?

It seems like the distance to the centre/ downtown of the city has a linear relationship with regard to the characteristic of the restaurants.

Are the restaurants in the central (downtown) areas more favored by clients?

From our clustered ratings, yes.

Optionally, how does the number of Michelin Starred restaurants affect the way that that restaurants are clustered?

Non-conclusive, will need more data and analysis.

# Discussion

There are a couple points to discussion for this project

1, the data are limited --- as Foursquare get detailed info about a restaurant is a premium endpoint, I only have limited called per day therefore I had to use the Kaggle dataset as an alternative. Resulting in limited resources.

2, the radius of which I search for venues (in this case, restaurants) are not rigorous since the actual geographical radius are unknown to me. I specified 2500 as a magic number, which can surely be improved.

3, the choice of k was not careful enough as I did not spend more time in investigation.

4, the extend of how much the number of Michelin starred restaurant affects the clustering was non-conclusive due to the small number of Michelin starred restaurant details I was able to get.

# Conclusion

I am happy that I utilized the knowledge I learned throughout the course in this project. Although there are many imperfections, I am satisfied with the results and output of this project given the limited time frame. I look forward to expanding and carry on with this idea and hope to learn more about machine learning.

Good luck to you as well, my fellow classmates.

# Reference

[1] <https://en.wikipedia.org/wiki/Arrondissements_of_Paris>

[2] <https://www.quora.com/How-many-restaurants-are-in-Paris>

[3] <https://www.godsavethepoints.com/2019/02/15/13-most-michelin-starred-cities-in-the-world/>

[4] <https://github.com/linzinan/IBM-Data-Science-Certificate/blob/master/ParisCuisine.ipynb>

[5] <https://stackoverflow.com/questions/8542391/how-to-normalize-reviews-based-on-score>

[6] <https://www.tableau.com/learn/articles/data-visualization>