Student Name: Ganga Lingden

Mini Project 1: (Yelp web scrapping)

Project Goal: To suggest the top five restaurants(i.e breakfast and brunch) based on Helsinki area by using web scraping technique.

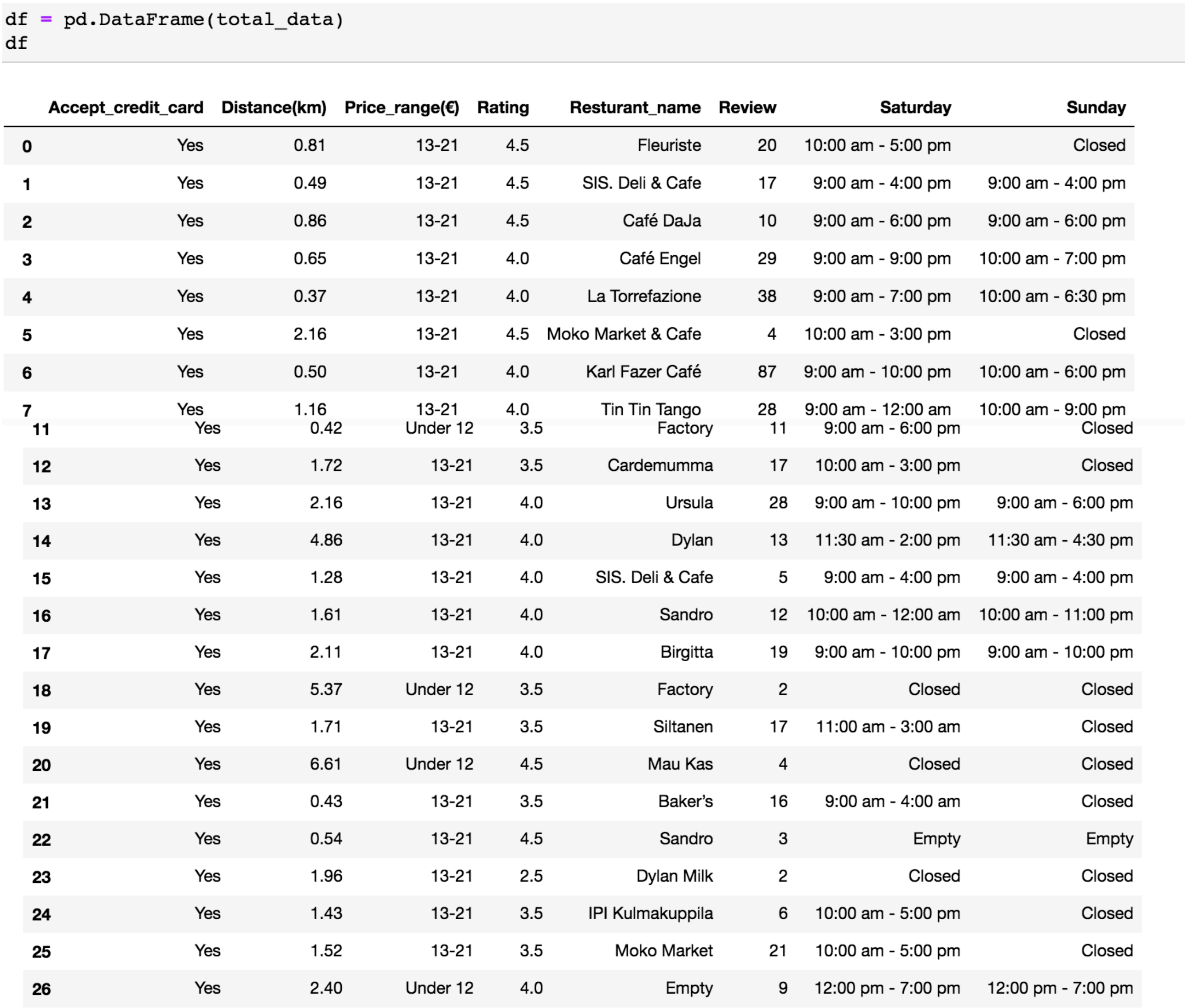
Procedure: In the section, the steps that were taken during the scraping of restaurant information are covered

1.Data scrapping: In the project, the Beautiful Soup was used to carried out the task. Firstly, manual filtering on yelp search bar were executed by selecting criteria, such as, location(Helsinki), category type (breakfast and brunch), payment method, price range(€ and €€). By filtering with these selection makes more convenient way to reach the project goal since these are some of the requirements to find the top five restaurants.

After manual filtering, it was shown 38 restaurants in two pages, but unfortunately the script only pulled 27 restaurants urls and collected all of them in a list (in my case, total_url). The reason behind to pull only 27 restaurants' urls is that all restaurants are not in the same format (for example, the code: **resturant_url = soup.find_all('a', {'class': 'nowrap'}**) doesn't find url of some restaurants since urls are not presented in the page.

Now the next one was to find the essential attributes for the analysis. In my case, the attributes such as resturant_name, price_range, accepted_credit_card, rating, review_count, days(saturday and sunday) and distance were scraped as dictionary data and append each of these data in one toatal_data as a list. Distance was calculated by taking longitude and latitude of each restaurant . The address, Kaivokatu 1, 00101 Helsinki was as central point for the distance calculation.

After extracting the data in one list (total_data), then this was converted in to pandas data frame for further analysis of the information. The screen-shot of total _data in pandas data frame is shown in the below :



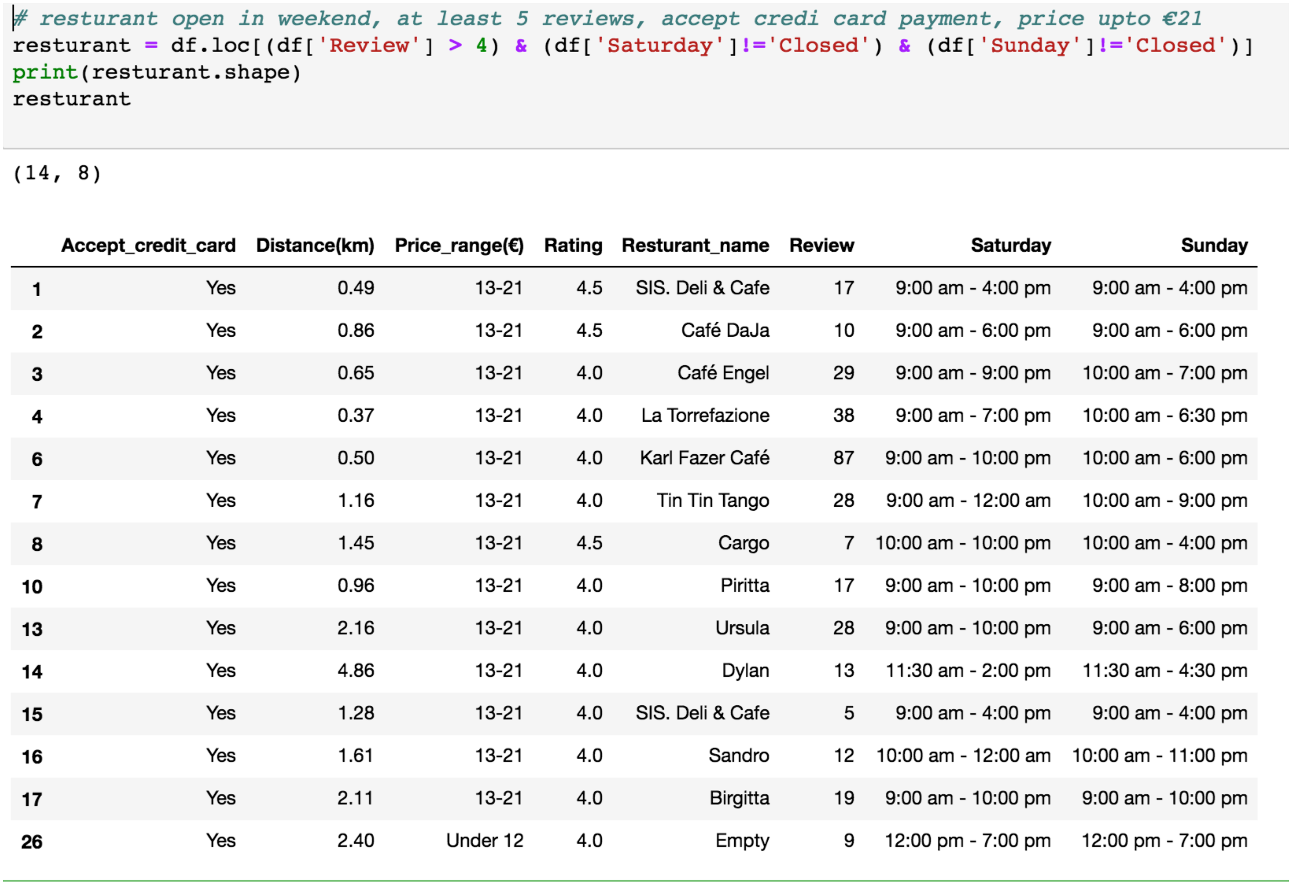
```
df = pd.DataFrame(total_data)
df
```

	Accept_credit_card	Distance(km)	Price_range(€)	Rating	Resturant_name	Review	Saturday	Sunday
0	Yes	0.81	13-21	4.5	Fleuriste	20	10:00 am - 5:00 pm	Closed
1	Yes	0.49	13-21	4.5	SIS. Deli & Cafe	17	9:00 am - 4:00 pm	9:00 am - 4:00 pm
2	Yes	0.86	13-21	4.5	Café DaJa	10	9:00 am - 6:00 pm	9:00 am - 6:00 pm
3	Yes	0.65	13-21	4.0	Café Engel	29	9:00 am - 9:00 pm	10:00 am - 7:00 pm
4	Yes	0.37	13-21	4.0	La Torrefazione	38	9:00 am - 7:00 pm	10:00 am - 6:30 pm
5	Yes	2.16	13-21	4.5	Moko Market & Cafe	4	10:00 am - 3:00 pm	Closed
6	Yes	0.50	13-21	4.0	Karl Fazer Café	87	9:00 am - 10:00 pm	10:00 am - 6:00 pm
7	Yes	1.16	13-21	4.0	Tin Tin Tango	28	9:00 am - 12:00 am	10:00 am - 9:00 pm
11	Yes	0.42	Under 12	3.5	Factory	11	9:00 am - 6:00 pm	Closed
12	Yes	1.72	13-21	3.5	Cardemumma	17	10:00 am - 3:00 pm	Closed
13	Yes	2.16	13-21	4.0	Ursula	28	9:00 am - 10:00 pm	9:00 am - 6:00 pm
14	Yes	4.86	13-21	4.0	Dylan	13	11:30 am - 2:00 pm	11:30 am - 4:30 pm
15	Yes	1.28	13-21	4.0	SIS. Deli & Cafe	5	9:00 am - 4:00 pm	9:00 am - 4:00 pm
16	Yes	1.61	13-21	4.0	Sandro	12	10:00 am - 12:00 am	10:00 am - 11:00 pm
17	Yes	2.11	13-21	4.0	Birgitta	19	9:00 am - 10:00 pm	9:00 am - 10:00 pm
18	Yes	5.37	Under 12	3.5	Factory	2	Closed	Closed
19	Yes	1.71	13-21	3.5	Siltanen	17	11:00 am - 3:00 am	Closed
20	Yes	6.61	Under 12	4.5	Mau Kas	4	Closed	Closed
21	Yes	0.43	13-21	3.5	Baker's	16	9:00 am - 4:00 am	Closed
22	Yes	0.54	13-21	4.5	Sandro	3	Empty	Empty
23	Yes	1.96	13-21	2.5	Dylan Milk	2	Closed	Closed
24	Yes	1.43	13-21	3.5	IPI Kulmakuppila	6	10:00 am - 5:00 pm	Closed
25	Yes	1.52	13-21	3.5	Moko Market	21	10:00 am - 5:00 pm	Closed
26	Yes	2.40	Under 12	4.0	Empty	9	12:00 pm - 7:00 pm	12:00 pm - 7:00 pm

2. Clean up data:

In this section, all the data which are not opened in weekend and have review less than 5 were removed since weekend opening, review at lest 5 are the requirements. And, payment method and price look fine since filtering was done (taking account these two requirements) before scraping the data. So, no need to worry about these two.

After the cleaning of data, the only 14 instances(restaurants) were remained. The below is the screen shoot:



```
# resturant open in weekend, at least 5 reviews, accept credi card payment, price upto €21
resturant = df.loc[(df['Review'] > 4) & (df['Saturday']!='Closed') & (df['Sunday']!='Closed')]
print(resturant.shape)
resturant
```

```
(14, 8)
```

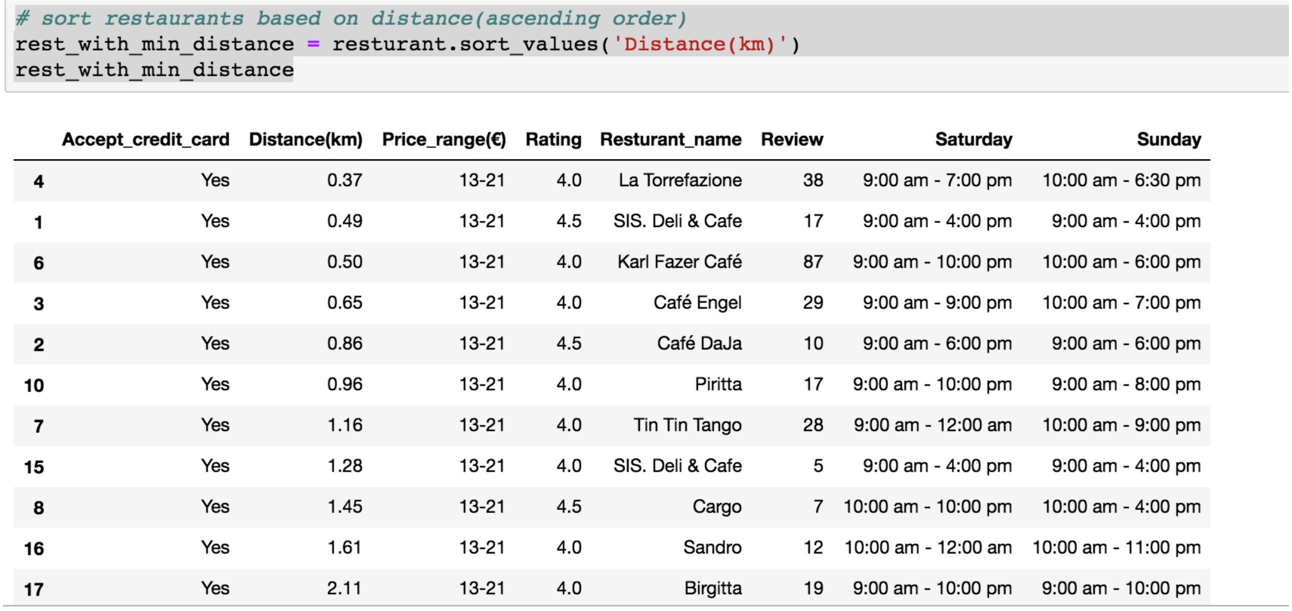
	Accept_credit_card	Distance(km)	Price_range(€)	Rating	Resturant_name	Review	Saturday	Sunday
1	Yes	0.49	13-21	4.5	SIS. Deli & Cafe	17	9:00 am - 4:00 pm	9:00 am - 4:00 pm
2	Yes	0.86	13-21	4.5	Café DaJa	10	9:00 am - 6:00 pm	9:00 am - 6:00 pm
3	Yes	0.65	13-21	4.0	Café Engel	29	9:00 am - 9:00 pm	10:00 am - 7:00 pm
4	Yes	0.37	13-21	4.0	La Torrefazione	38	9:00 am - 7:00 pm	10:00 am - 6:30 pm
6	Yes	0.50	13-21	4.0	Karl Fazer Café	87	9:00 am - 10:00 pm	10:00 am - 6:00 pm
7	Yes	1.16	13-21	4.0	Tin Tin Tango	28	9:00 am - 12:00 am	10:00 am - 9:00 pm
8	Yes	1.45	13-21	4.5	Cargo	7	10:00 am - 10:00 pm	10:00 am - 4:00 pm
10	Yes	0.96	13-21	4.0	Piritta	17	9:00 am - 10:00 pm	9:00 am - 8:00 pm
13	Yes	2.16	13-21	4.0	Ursula	28	9:00 am - 10:00 pm	9:00 am - 6:00 pm
14	Yes	4.86	13-21	4.0	Dylan	13	11:30 am - 2:00 pm	11:30 am - 4:30 pm
15	Yes	1.28	13-21	4.0	SIS. Deli & Cafe	5	9:00 am - 4:00 pm	9:00 am - 4:00 pm
16	Yes	1.61	13-21	4.0	Sandro	12	10:00 am - 12:00 am	10:00 am - 11:00 pm
17	Yes	2.11	13-21	4.0	Birgitta	19	9:00 am - 10:00 pm	9:00 am - 10:00 pm
26	Yes	2.40	Under 12	4.0	Empty	9	12:00 pm - 7:00 pm	12:00 pm - 7:00 pm

In the above data frame, all the instances fulfil the requirements of the project goal (i.e, must be open in weekend, price range up to €21, reviews counts at least 5 and accept read card payment method). But now the top five listed in the table are the really top five restaurants or not? This is the important question to ask. Thus, I want to further analysis the data so that get top five using the intersection of top 8 restaurants (you can use any number of restaurant but not all of them) based on 'rating' and 'review' but I did not select distance. This is because I believe that the restaurant that has high rating with some great number of reviews give good indication of top restaurant. No matter how far is the distance. This assumption might be differ with other analyst.

3. Finding the Top Five Restaurants:

In the section, I tried to sort the restaurant based on different attributes and try to find the common restaurant between top 8 restaurants based on rating and review counts.

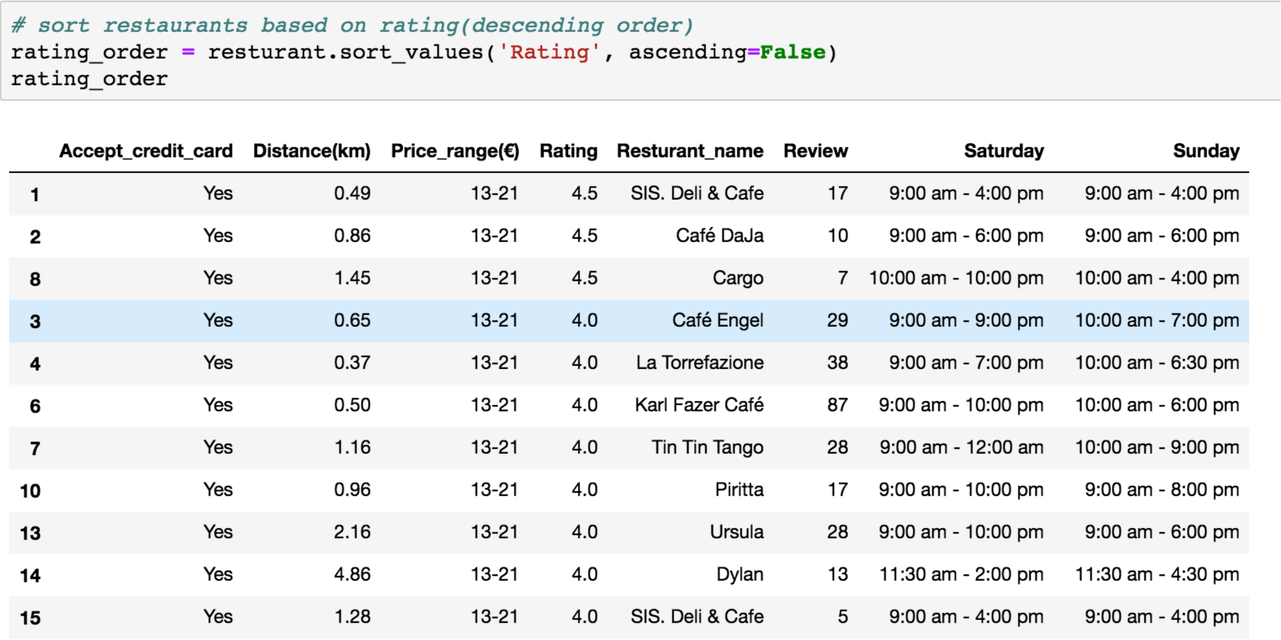
Restaurants with minimum distance at first:



```
# sort restaurants based on distance(ascending order)
rest_with_min_distance = resturant.sort_values('Distance(km)')
rest_with_min_distance
```

	Accept_credit_card	Distance(km)	Price_range(€)	Rating	Resturant_name	Review	Saturday	Sunday
4	Yes	0.37	13-21	4.0	La Torrefazione	38	9:00 am - 7:00 pm	10:00 am - 6:30 pm
1	Yes	0.49	13-21	4.5	SIS. Deli & Cafe	17	9:00 am - 4:00 pm	9:00 am - 4:00 pm
6	Yes	0.50	13-21	4.0	Karl Fazer Café	87	9:00 am - 10:00 pm	10:00 am - 6:00 pm
3	Yes	0.65	13-21	4.0	Café Engel	29	9:00 am - 9:00 pm	10:00 am - 7:00 pm
2	Yes	0.86	13-21	4.5	Café DaJa	10	9:00 am - 6:00 pm	9:00 am - 6:00 pm
10	Yes	0.96	13-21	4.0	Piritta	17	9:00 am - 10:00 pm	9:00 am - 8:00 pm
7	Yes	1.16	13-21	4.0	Tin Tin Tango	28	9:00 am - 12:00 am	10:00 am - 9:00 pm
15	Yes	1.28	13-21	4.0	SIS. Deli & Cafe	5	9:00 am - 4:00 pm	9:00 am - 4:00 pm
8	Yes	1.45	13-21	4.5	Cargo	7	10:00 am - 10:00 pm	10:00 am - 4:00 pm
16	Yes	1.61	13-21	4.0	Sandro	12	10:00 am - 12:00 am	10:00 am - 11:00 pm
17	Yes	2.11	13-21	4.0	Birgitta	19	9:00 am - 10:00 pm	9:00 am - 10:00 pm

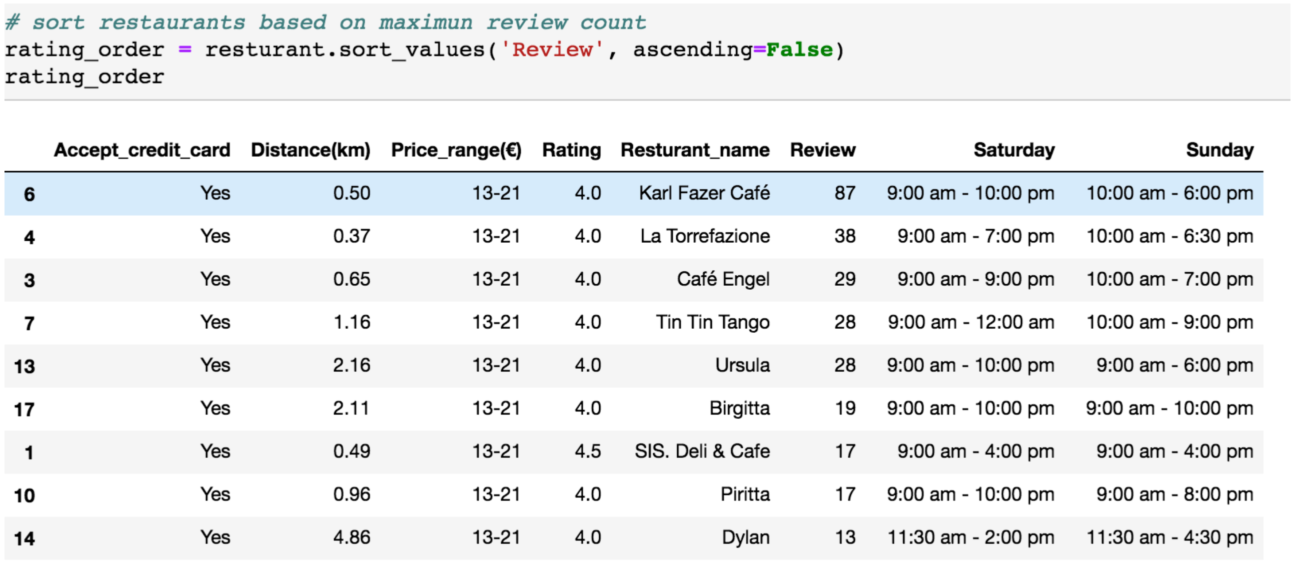
Restaurants with maximum rating:



```
# sort restaurants based on rating(descending order)
rating_order = resturant.sort_values('Rating', ascending=False)
rating_order
```

	Accept_credit_card	Distance(km)	Price_range(€)	Rating	Resturant_name	Review	Saturday	Sunday
1	Yes	0.49	13-21	4.5	SIS. Deli & Cafe	17	9:00 am - 4:00 pm	9:00 am - 4:00 pm
2	Yes	0.86	13-21	4.5	Café DaJa	10	9:00 am - 6:00 pm	9:00 am - 6:00 pm
8	Yes	1.45	13-21	4.5	Cargo	7	10:00 am - 10:00 pm	10:00 am - 4:00 pm
3	Yes	0.65	13-21	4.0	Café Engel	29	9:00 am - 9:00 pm	10:00 am - 7:00 pm
4	Yes	0.37	13-21	4.0	La Torrefazione	38	9:00 am - 7:00 pm	10:00 am - 6:30 pm
6	Yes	0.50	13-21	4.0	Karl Fazer Café	87	9:00 am - 10:00 pm	10:00 am - 6:00 pm
7	Yes	1.16	13-21	4.0	Tin Tin Tango	28	9:00 am - 12:00 am	10:00 am - 9:00 pm
10	Yes	0.96	13-21	4.0	Piritta	17	9:00 am - 10:00 pm	9:00 am - 8:00 pm
13	Yes	2.16	13-21	4.0	Ursula	28	9:00 am - 10:00 pm	9:00 am - 6:00 pm
14	Yes	4.86	13-21	4.0	Dylan	13	11:30 am - 2:00 pm	11:30 am - 4:30 pm
15	Yes	1.28	13-21	4.0	SIS. Deli & Cafe	5	9:00 am - 4:00 pm	9:00 am - 4:00 pm

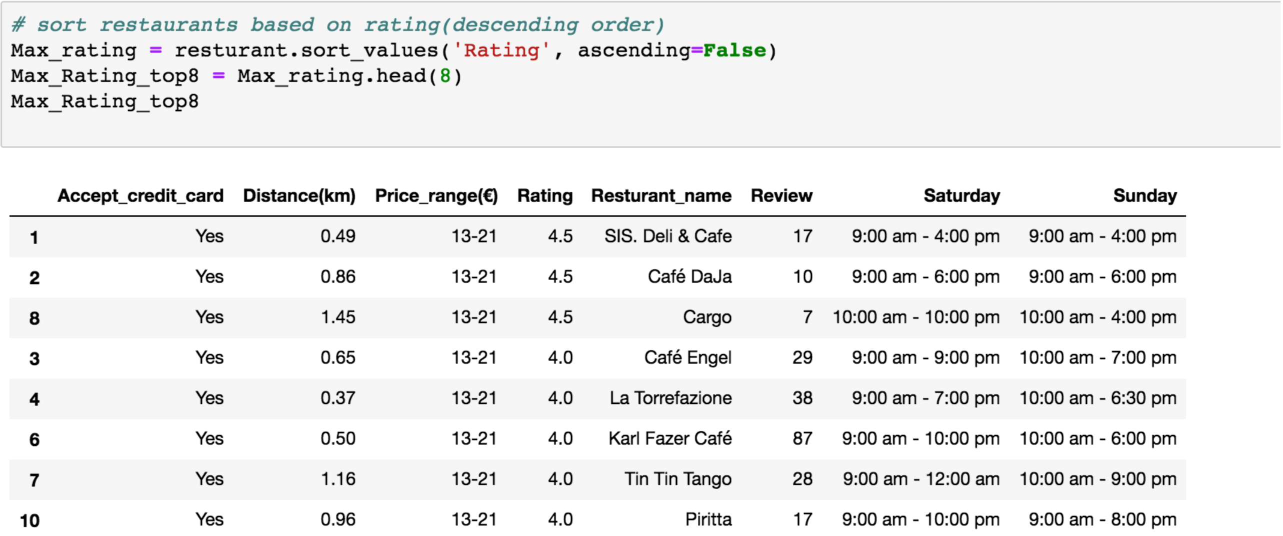
Restaurants with maximum reviews count:



```
# sort restaurants based on maximun review count
rating_order = resturant.sort_values('Review', ascending=False)
rating_order
```

	Accept_credit_card	Distance(km)	Price_range(€)	Rating	Resturant_name	Review	Saturday	Sunday
6	Yes	0.50	13-21	4.0	Karl Fazer Café	87	9:00 am - 10:00 pm	10:00 am - 6:00 pm
4	Yes	0.37	13-21	4.0	La Torrefazione	38	9:00 am - 7:00 pm	10:00 am - 6:30 pm
3	Yes	0.65	13-21	4.0	Café Engel	29	9:00 am - 9:00 pm	10:00 am - 7:00 pm
7	Yes	1.16	13-21	4.0	Tin Tin Tango	28	9:00 am - 12:00 am	10:00 am - 9:00 pm
13	Yes	2.16	13-21	4.0	Ursula	28	9:00 am - 10:00 pm	9:00 am - 6:00 pm
17	Yes	2.11	13-21	4.0	Birgitta	19	9:00 am - 10:00 pm	9:00 am - 10:00 pm
1	Yes	0.49	13-21	4.5	SIS. Deli & Cafe	17	9:00 am - 4:00 pm	9:00 am - 4:00 pm
10	Yes	0.96	13-21	4.0	Piritta	17	9:00 am - 10:00 pm	9:00 am - 8:00 pm
14	Yes	4.86	13-21	4.0	Dylan	13	11:30 am - 2:00 pm	11:30 am - 4:30 pm

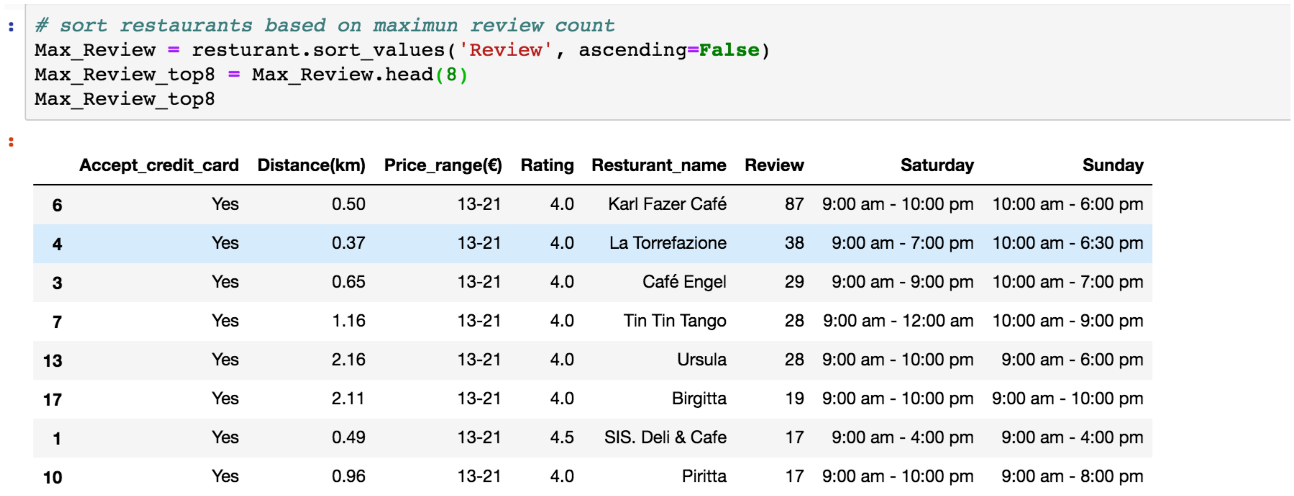
Top 8 restaurant based on Rating:



```
# sort restaurants based on rating(descending order)
Max_rating = resturant.sort_values('Rating', ascending=False)
Max_Rating_top8 = Max_rating.head(8)
Max_Rating_top8
```

	Accept_credit_card	Distance(km)	Price_range(€)	Rating	Resturant_name	Review	Saturday	Sunday
1	Yes	0.49	13-21	4.5	SIS. Deli & Cafe	17	9:00 am - 4:00 pm	9:00 am - 4:00 pm
2	Yes	0.86	13-21	4.5	Café DaJa	10	9:00 am - 6:00 pm	9:00 am - 6:00 pm
8	Yes	1.45	13-21	4.5	Cargo	7	10:00 am - 10:00 pm	10:00 am - 4:00 pm
3	Yes	0.65	13-21	4.0	Café Engel	29	9:00 am - 9:00 pm	10:00 am - 7:00 pm
4	Yes	0.37	13-21	4.0	La Torrefazione	38	9:00 am - 7:00 pm	10:00 am - 6:30 pm
6	Yes	0.50	13-21	4.0	Karl Fazer Café	87	9:00 am - 10:00 pm	10:00 am - 6:00 pm
7	Yes	1.16	13-21	4.0	Tin Tin Tango	28	9:00 am - 12:00 am	10:00 am - 9:00 pm
10	Yes	0.96	13-21	4.0	Piritta	17	9:00 am - 10:00 pm	9:00 am - 8:00 pm

Top 8 restaurant based on Review Count:

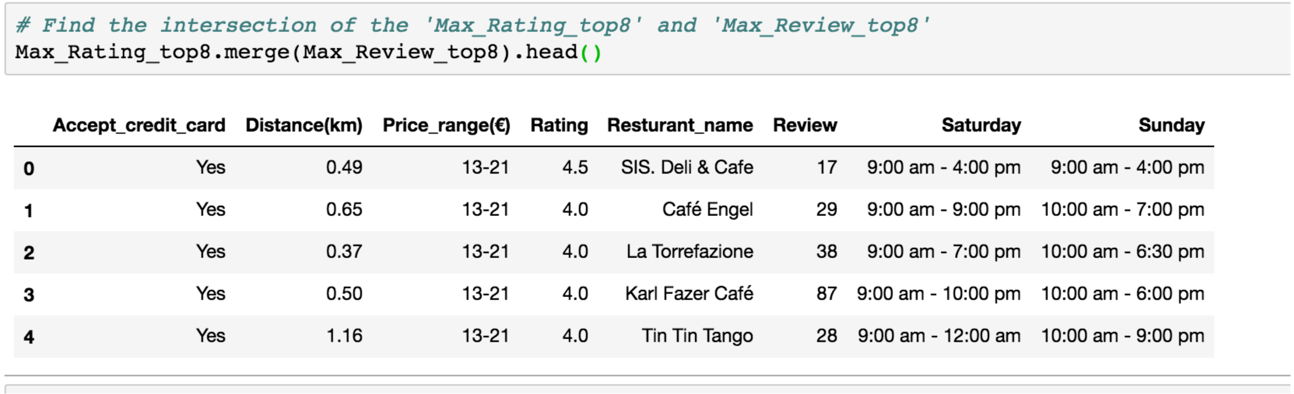


```
: # sort restaurants based on maximun review count
  Max_Review = resturant.sort_values('Review', ascending=False)
  Max_Review_top8 = Max_Review.head(8)
  Max_Review_top8
```

:

	Accept_credit_card	Distance(km)	Price_range(€)	Rating	Resturant_name	Review	Saturday	Sunday
6	Yes	0.50	13-21	4.0	Karl Fazer Café	87	9:00 am - 10:00 pm	10:00 am - 6:00 pm
4	Yes	0.37	13-21	4.0	La Torrefazione	38	9:00 am - 7:00 pm	10:00 am - 6:30 pm
3	Yes	0.65	13-21	4.0	Café Engel	29	9:00 am - 9:00 pm	10:00 am - 7:00 pm
7	Yes	1.16	13-21	4.0	Tin Tin Tango	28	9:00 am - 12:00 am	10:00 am - 9:00 pm
13	Yes	2.16	13-21	4.0	Ursula	28	9:00 am - 10:00 pm	9:00 am - 6:00 pm
17	Yes	2.11	13-21	4.0	Birgitta	19	9:00 am - 10:00 pm	9:00 am - 10:00 pm
1	Yes	0.49	13-21	4.5	SIS. Deli & Cafe	17	9:00 am - 4:00 pm	9:00 am - 4:00 pm
10	Yes	0.96	13-21	4.0	Piritta	17	9:00 am - 10:00 pm	9:00 am - 8:00 pm

Common restaurants between Top 8 restaurant based on Rating and Review Count (Top Five suggested Restaurant)

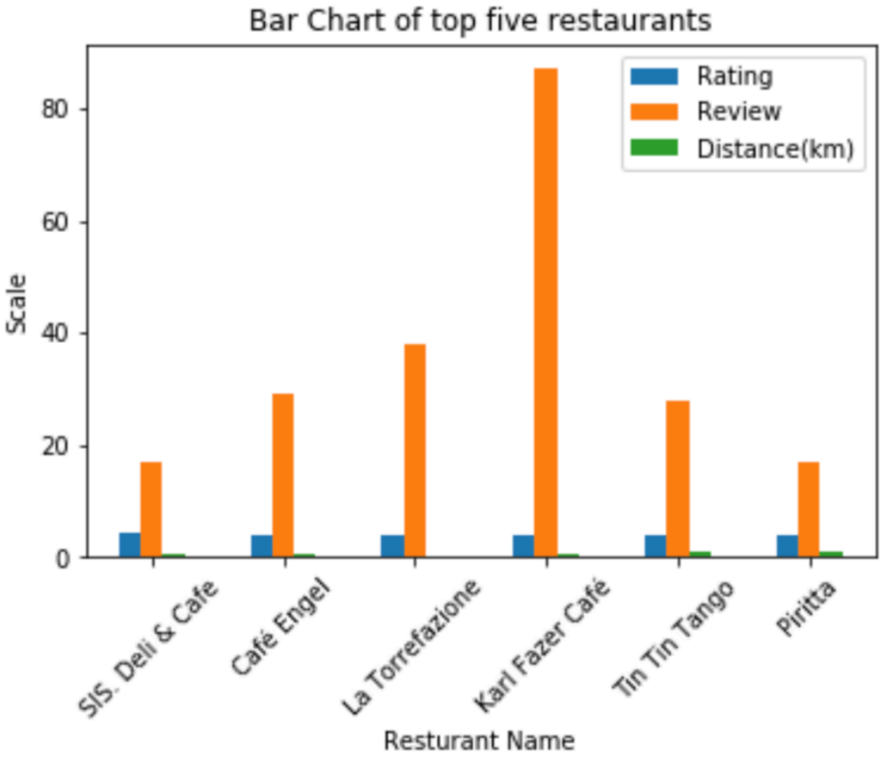


```
# Find the intersection of the 'Max_Rating_top8' and 'Max_Review_top8'
Max_Rating_top8.merge(Max_Review_top8).head()
```

	Accept_credit_card	Distance(km)	Price_range(€)	Rating	Resturant_name	Review	Saturday	Sunday
0	Yes	0.49	13-21	4.5	SIS. Deli & Cafe	17	9:00 am - 4:00 pm	9:00 am - 4:00 pm
1	Yes	0.65	13-21	4.0	Café Engel	29	9:00 am - 9:00 pm	10:00 am - 7:00 pm
2	Yes	0.37	13-21	4.0	La Torrefazione	38	9:00 am - 7:00 pm	10:00 am - 6:30 pm
3	Yes	0.50	13-21	4.0	Karl Fazer Café	87	9:00 am - 10:00 pm	10:00 am - 6:00 pm
4	Yes	1.16	13-21	4.0	Tin Tin Tango	28	9:00 am - 12:00 am	10:00 am - 9:00 pm

From the table above, surprisingly it shows that even though I performed the intersection between the two top 8 restaurants based on 'rating ' and 'review', the result is almost the same as the table after the data cleaning. The reason behind doing the intersection is to find the top restaurant that belong to top 8 restaurants in rating and review attributes. And, to avoid the situation that some restaurants have 5 stars rating with few review counts but appear as top one.

Bar char of top five restaurants:



Analysis: In the above bar graph among top five restaurants, 'Karl Fazer Cafe' has the highest review counts where as 'SIS. Deli and Cafe ' has lowest. 'Piritta' restaurant seams little far away from the centre (Kaivokatu 1, 00101 Helsinki). Rating is almost same to every restaurants i.e. equal to or higher than 4.