Coursera IBM Data Science with Python Specialization – Course 9 Capstone Project

Contents:

1. Topic Description

For the Capstone, I decided to use the 4 Major cities NewYork, Delhi, Sydney, Toronto dataset available in the course and append it with data from geojson.

To get a general overview of eachcities venue categories, four Dataframes created for each city . Furthermore, a new dataset is created from the processed All the cities venue data (available from Foursquare) that contains data.

We will see that the venues in each city greatly.

Interested stakeholders might take recommendations from the data as to what businesses are at the top to invest on, where to open a shop or similar. Or they don't, depends on them.

2. Data Description

The four cities data for this topic was originally taken from the Capstone course of IBM data science, from Wikipedia and other course sources, the city of Toronto's homepage and from the Internet.

Starting from Wikipedia's data on, a data frame is created from HTML parsing and combined with geographical coordinates datsets. Note that the are not aggregated based on their boroughs since the geopgraphical data uses later. The geo coordinates are also taken from other sources and imported as CSV files.

Using the coordinate Data is extracted from Foursquare API for each Latitude and Longitude with the top 100 venues(as Foursquare has limitations for number of calls per day).

Due to the maximum amount of calls to Foursquare, this can't be repeated very often per day.

In further steps, the data regarding the amount of venues and categories is being grouped and save in their own excel file. The data frames/tables each give an overview of the most common venue for each city is and can quickly be analyzed within Python.

3. Methods used for data exploration

The methodologies during the Toronto data analysis can be summarized quickly:

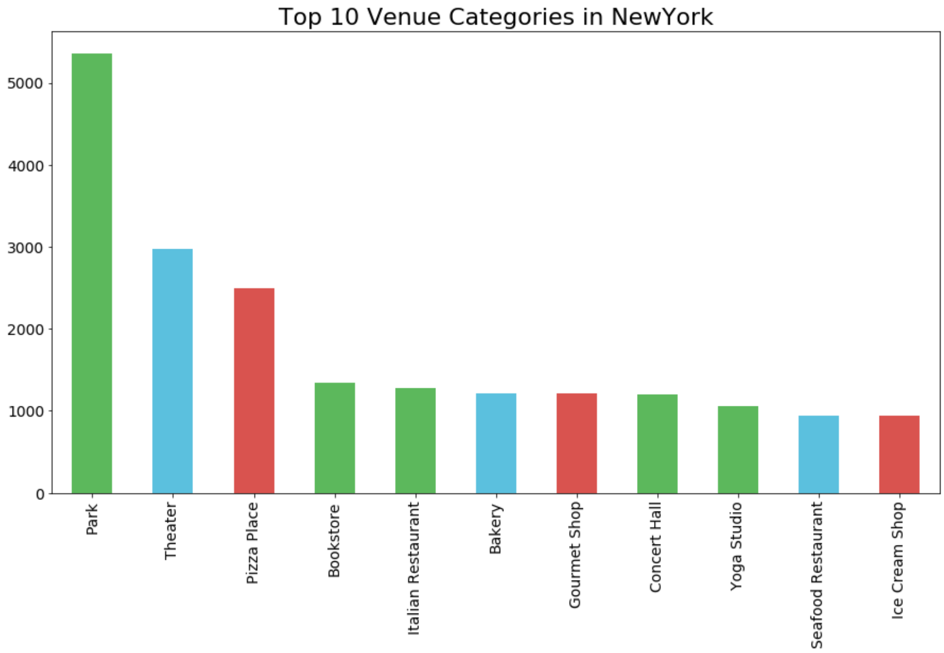
- Data wrangling with Python functions, loops, algorithms and regular expressions
- Pandas dataframe manipulation and HTML parsing
- Data visualization with Matplotlib and Jupyter Notebooks
- Basic clustering with sklearn and KMeans

The import of the data as well as the manipulation and formatting were done with Python “pandas” toolkit to create and manipulate data frames.

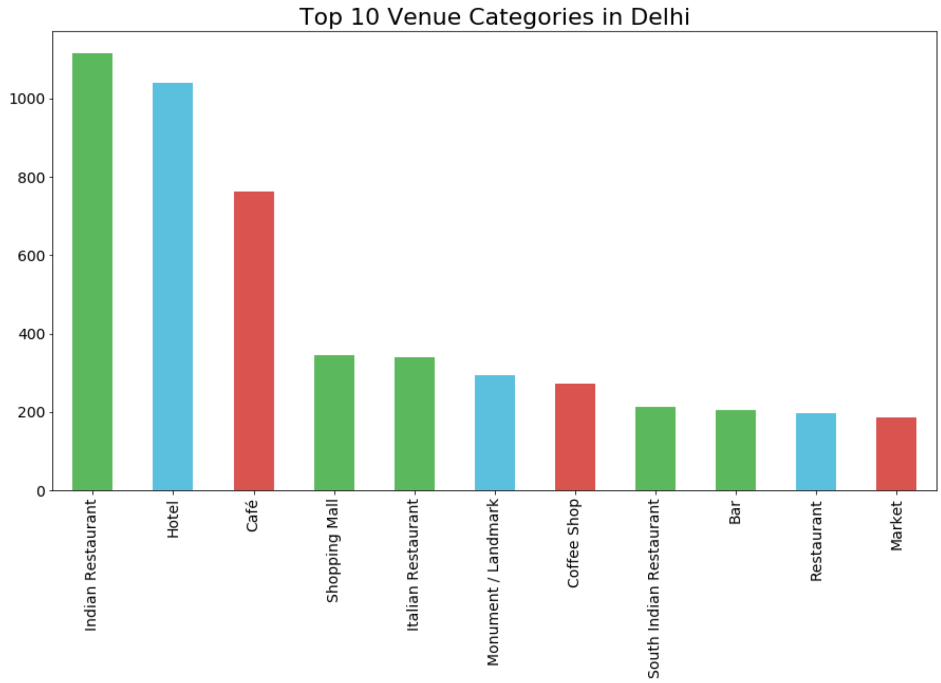
Functions of pandas such as groupby, count etc. were used together to format the data for later analysis during each step of the code.

With the Python module “Matplotlib”, a bar charts was created to show the most common venues in each city and the overall common venue in 4 It should (again) be noted that these data files and the corresponding maps might not be 100% coordinated with the datasets.

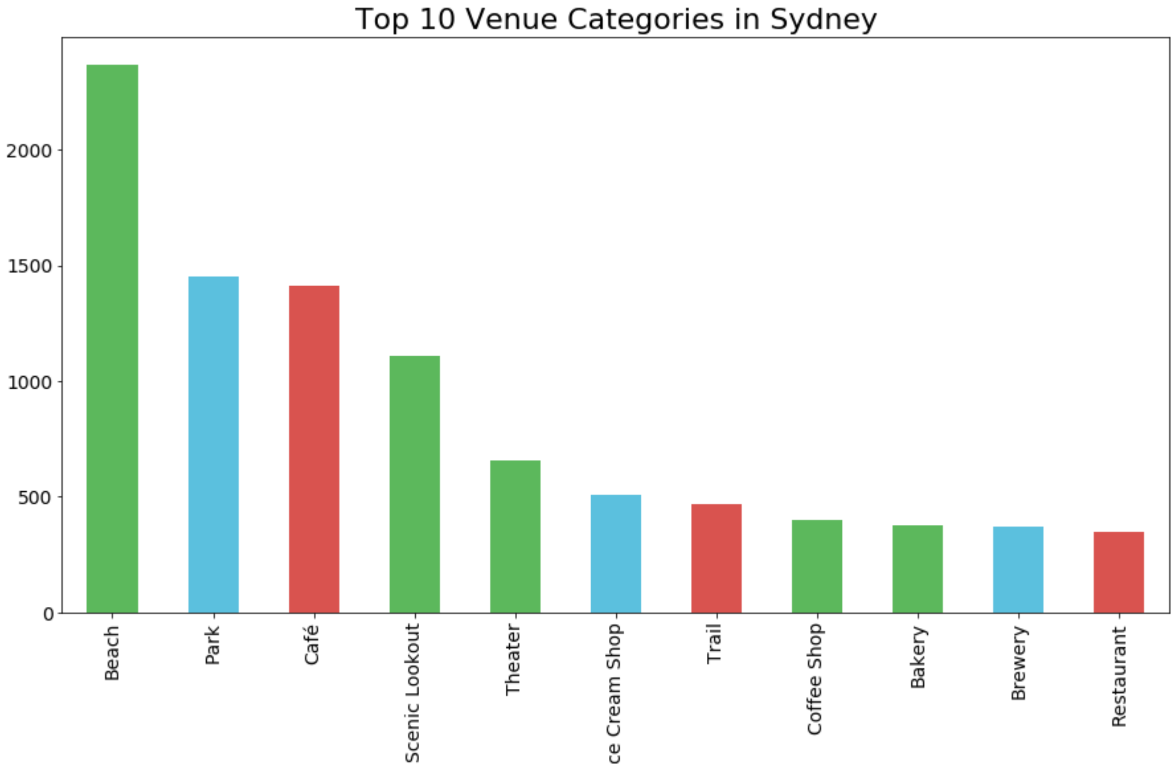
The resulting bar chart for NewYork:



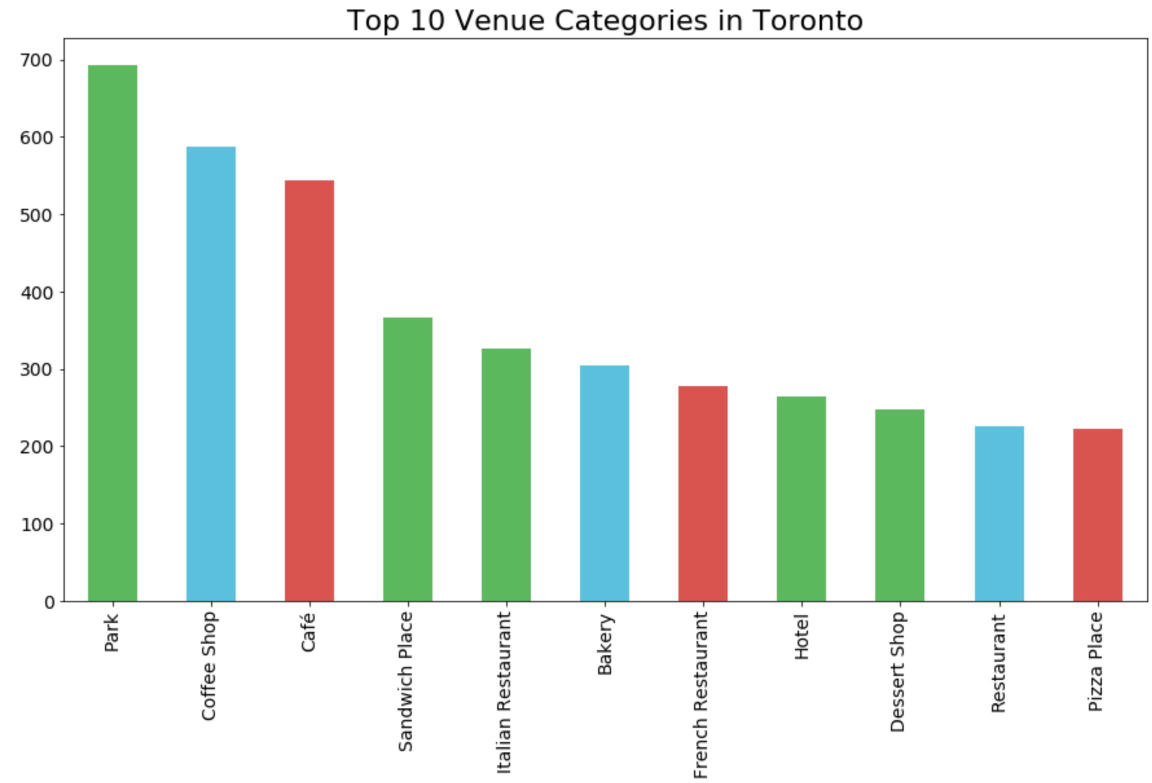
The resulting bar chart for Delhi:



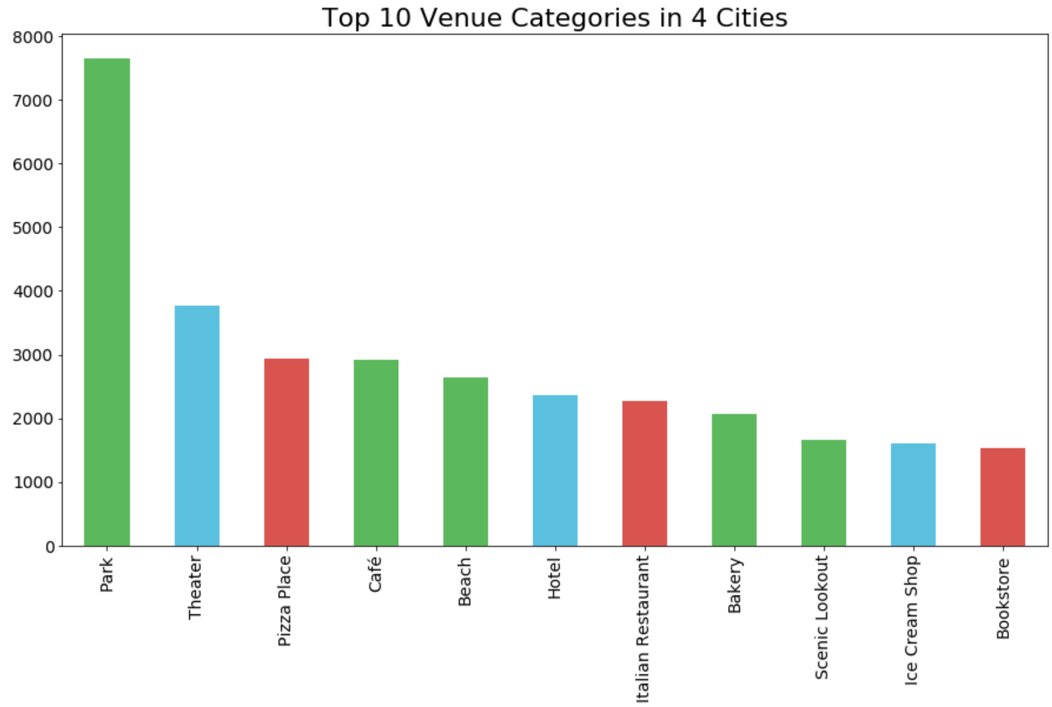
The resulting bar chart for Sydney:



The resulting bar chart for Toronto:



The resulting bar chart for 4 Cities:



4. Discussion

The results so far show that 4 cities are diverse and are different from each other.

Looking into detail, there is a large focus on food related places especially Café’s and Restaurants in each city. since their more common than other venue types. This might be due to Foursquare users being focused on visiting restaurants and giving their respective ratings.

From the perspective of a shop or restaurant owner, the competition among coffee shops and cafes seems rather high. Coffee shops are the most common venue types,

An interesting point is also the large spread of venue amounts between different Cities.

5. Conclusion

All in all, the venue distribution in 4 cities seems rather unequal due to the diversity. Food places take the top priority.

A future investor might be inclined to open his restaurant or a café bar in these 4 cities to be successful.