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

Topic: The Battle of Neighborhoods – Report

Contents:

1. **Topic Description**
2. **Data Description**
3. **Methods used for data exploration**
4. **Discussion**
5. **Conclusion**

Note: This report is kept short since most explanation is within the python code, either in Pycharm or the Jupyter Notebook. Also, I'm not sure how seriously people will look into this.

1. Topic Description

For the Capstone, I decided to use the Toronto dataset available in the course and append it with data from geojson.

To get a general overview of Toronto's venue density, several excel tables were created from python pandas dataframes. Furthermore, a new dataset is created from the processed Toronto venue data (available from Foursquare) that contains data regarding food related places in Toronto such as restaurants, bars, cafes etc.

We will see that the density of venues and food related venues greatly differs between Toronto's boroughs and neighborhoods.

Interested users might take recommendations from the data as to where to look for restaurants in Toronto, where to open a shop or similar. Or they don't, depends on them.

Link to the code in Jupyter notebook (without results, they can be seen in excel files):

https://github.com/StevenT1995/Cousera-IBM-Course-9---Capstone-Project/blob/master/Capstone%20Jupyter.ipynb

2. Data Description

The Toronto 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 Toronto's boroughs and neighborhoods, a data frame is created from HTML parsing and combined with geographical coordinates datsets. Note that the neighborhoods are not aggregated based on their boroughs since the geopgraphical data uses neighborhoods later.

The excel file for this dataset can be found here:

https://github.com/StevenT1995/Cousera-IBM-Course-9---Capstone-Project/blob/master/toronto_original.xlsx

Based on that table, Foursquare is used to generate a new data frame that includes the found venues and venue categories for each Toronto area plus their coordinates.

Due to the maximum amount of calls to Foursquare, this can't be repeated very often per day. The respective excel file is here:

https://github.com/StevenT1995/Cousera-IBM-Course-9---Capstone-Project/blob/master/toronto_venues.xlsx

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 Toronto's venue density and diversity and can quickly be analyzed within Python or Excel, e.g. to see maximum and minimum values, averages, variance etc.

File link:

https://github.com/StevenT1995/Cousera-IBM-Course-9---Capstone-Project/blob/master/toronto_groups.xlsx

Later, the Toronto content data on venues is used with a Toronto geojson to generate an overview of the venue densities.

The geojson was found on the internet, it delivers results without the guarantee of being 100% correct. The chart should be taken as an indicator, not as set in stone.

Based on the processed dataframes, four clusters are used to categorise Toronto's neighborhoods. This clustering only focuses on food related institutions (restaurants, bars etc.), therefore all other venue categories were dropped for this instance.

The neighborhoods are grouped into either high density of food places, rather high, rather low or low.

The clustered data frame was saved into an excel file that also contains a quick overview analysis of each cluster's numbers, KPIs etc.

Link here:

https://github.com/StevenT1995/Cousera-IBM-Course-9---Capstone-Project/blob/master/food%20cluster%20neighbor.xlsx

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 folium and Jupyter Notebooks
- Basic clustering with sklearn and KMeans
- Basic data analysis and sorting with MS Excel and indexing functions

The import of the data as well as the manipulation and formatting were done with Python “pandas” toolkit to create and manipulate data frames. The Python module “beautifulsoup4” was used to parse data on Wikipedia and import the results into data frames with pandas.

Functions of pandas such as groupby, aggregate, count etc. were used together to format the data for later analysis during each step of the code. When necessary, regular expression methods from Python module “regex” were used to filter data accordingly, e.g. for the creation of the food places data frame in the 4th section of the Python code.

With the Python module “folium”, a choropleth map was created to show the overall venue density in Toronto with the help of geojson data. It should (again) be noted that these data files and the corresponding maps might not be 100% coordinated with the datasets.

The resulting choropleth:

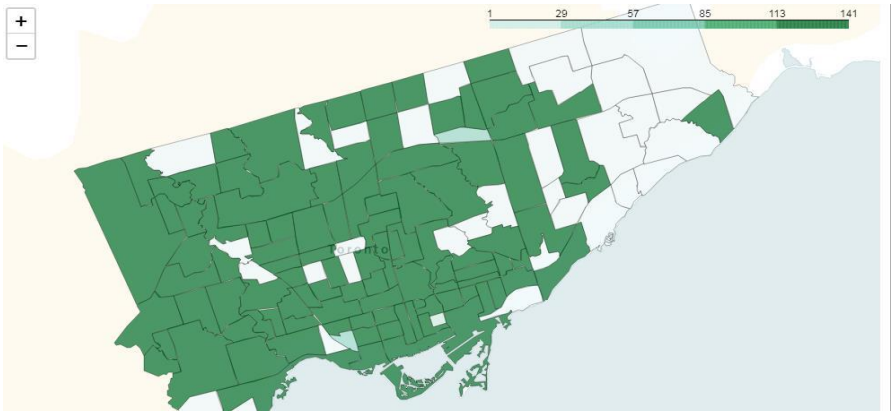

Image 1: Venue density in Toronto

Compare with a choropleth based on different key or geojson file:

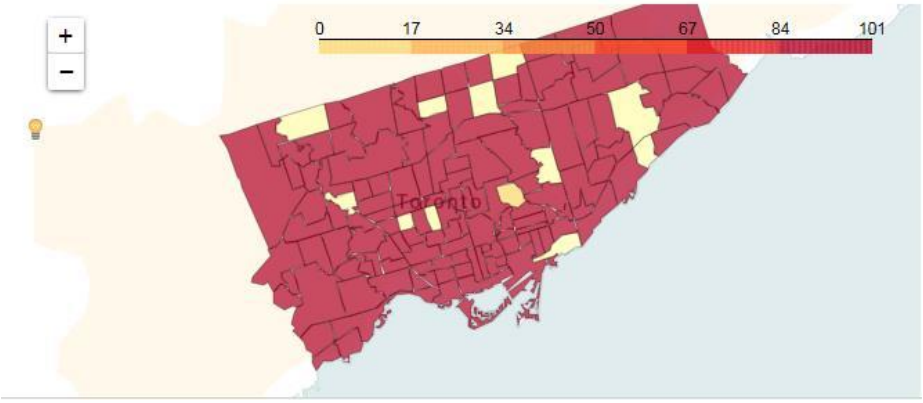

Image 2: Choropleth map of Toronto's venue density with very bad results

Problems here might occur due to the method "key_on" in the folium.choropleth attributes or due to broken geojson files/content data files.

See code snippet here:

```
# Create choropleth map
# Note: Due to differences in data from Foursquare and geojson, choropleth map might be prone to errors
to_map.choropleth(
    geo_data=geo,
    name='to_choropleth',
    data=df_to_choro,
    columns=['Neighborhood', 'Venue'],
    key_on='feature.properties.name',
    fill_color='BuGn',
    fill_opacity=0.7,
    line_opacity=0.3,
    threshold_scale=threshold_scale_1)

to_map
```

Image 3: Choropleth example code, see "key_on" attribute

In addition to the visualization, a quick and easy KMeans cluster was created with the Python module "sklearn". The Toronto neighborhoods were distributed into 4 clusters based on how many venues they each hold.

The statistics aspects here were kept very basic, my focus was on quick outputs and easy data differentiation.

Furthermore, all important data frames including the clustering results were saved into MS Excel files for possible further analysis or a quick overview.

4. Discussion

The results so far show that Toronto, as expected from a world metropole, contains a high number of neighborhoods and corresponding venues.

Looking into detail, there is a large focus on food related places in Toronto since their more common than other venue types. This might be due to Foursquare users being focused on visiting restaurants and giving their respective reatings.

The neighborhood with both the highest amount of single venues and he highest diversity of venue categories would be Jamestown with around 146 venues.

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, mostly due to the presence of Starbucks as can be seen from either the Python code or the excel files.

An interesting point is also the large spread of venue amounts between different neighborhoods. The spread tends to rather high overall, with a length of 50 venues on average between the highest density cluster and the least dense while the least dense cluster also inhabits the highest amount of neighborhoods overall.

Results can be seen in detail on the respective Github links above.

5. Conclusion

All in all, the venue distribution in Toronto seems rather unequal, but this could also be due to skewed data from Foursquare or the method application.

A future restaurant owner might be inclined to open his restaurant in a less density area of Toronto (see choropleth Image 1) to be protected from competition, given that there is also enough population.

Tourists should consider Jamestown because of its high venue diversity.