Coursera Capstone Report

Introduction

In this project I will be looking at the largest cities in europe with a view to determining the most common venue in each. This information will be useful to new businesses opening in large european cities as their strategists will be able to weigh the likelihood of there being a gap in the market against the pressure put on a business by proximity to competing businesses.

Data

In order to discover the most common venues in the largest european cities, first it is necessary to scrape a list of the largest european cities from wikipedia. This can be done with a package such as beautiful soup however in the instance that the data is already tabulated in a desirable manner then it is possible to simply use the read_html function present in the pandas library. Helpfully the list on wikipedia already includes the latitude and longitude values needed, otherwise it would be necessary to use a geocoder to obtain coordinates.
Once this data is obtained it can be used as an input for a function built to use the Foursquare™ api to provide a list of venues. By building a function it is possible to loop the api request in order to collect information from all the cities in the dataframe with a single line of code.

Methodology

After using the read_html function on the appropriate wikipedia page it was necessary to drop the columns that weren't needed using the df.drop function. After this the data needed cleaning so that it would pass freely through the later functions, this involved the use of a loop to work through all rows of the data frame, splitting longitude and latitude and removing excess text that was present in the data frame such as reference hyperlinks, for this the str.split, list.append, and str.replace functions were used. Once the longitude and latitude were separated into their appropriate columns and cleaned they were type cast as float values using the df.astype function. For ease of understanding in the presentation these were then mapped using the folium library.
At this stage the foursquare function was built to request a list of venues in each city centre. The above described data frame was then passed to the foursquare function, providing a list of venues in each city centre.
After using one hot encoding a new function was built to display a table of the venues in each city listed by their frequency.

Results and Discussion

The most common venue in most cities was a hotel, appearing as the most common venue in 4 out of the 15 cities. Ideally a larger search radius would have been used to measure the entire city rather than just the center however due to the limit put in place by foursquare of 100 results per API call, this was not possible and a smaller section of the city was chosen in order to differentiate the cities and provide reproducible data for discussion.

Conclusion

This project was a success though more interesting findings could have been developed if api calls were made for each borough in each city and this may have circumvented the 100 result limit.