Were to Rent an Apartment in Dublin?

Introduction: Defining the Business Problem

- Mr. John Guinness, a software developer, found a new job in Dublin He is interested to rent an apartment there. His questions are:
  - Where are the best places for him to leave?
  - Which districts in Dublin best fit his preferences?
- Mr Guinness' list of his preferences: theaters, parks, beaches, gyms, public transportation, pubs, restaurants (especially Italian and Chinese), coffee shops, shopping malls, supermarkets, groceries and convenience shops.
- There are 22 districts in Dublin. In order to answer Mr. Guinness questions, we looked for the most relevant venues for each district, we grouped the districts regarding the categories of venues in his preferences list and we have to find the group of districts which best matches his preferences.

Data Acquisition and Preparation – Districts of Dublin: Building the Data Frame

- One can find a list of the districts and their corresponding postal codes on wikipedia: https://en.wikipedia.org/wiki/List_of_Dublin_postal_districts

Dublin's postal districts	
Northside, covering local government area	**Southside**, covering local government area
Dublin 1 (D1) Dublin	Dublin 2 (D2) Dublin
Dublin 3 (D3) Dublin	Dublin 4 (D4) Dublin, Dun Laoghaire Rathdown
Dublin 5 (D5) Dublin	Dublin 6 (D6) Dublin, Dún Laoghaire Rathdown
Dublin 7 (D7) Dublin	Dublin 6W (D6W) Dublin, South Dublin
Dublin 9 (D9) Dublin	Dublin 8 (D8) Dublin
Dublin 11 (D11) Dublin, Fingal	Dublin 10 (D10) Dublin
Dublin 13 (D13) Dublin, Fingal	Dublin 12 (D12) Dublin
Dublin 15 (D15) Fingal	Dublin 14 (D14) Dublin, Dún Laoghaire–Rathdown, South Dublin
Dublin 17 (D17) Dublin, Fingal	Dublin 16 (D16) Dún Laoghaire–Rathdown, South Dublin
	Dublin 18 (D18) Dún Laoghaire–Rathdown
	Dublin 20 (D20) Dublin, South Dublin
	Dublin 22 (D22) South Dublin
	Dublin 24 (D24) South Dublin
"County Dublin"; Fingal, South Dublin, Dún Laoghaire–Rathdown, and small pockets of Meath	

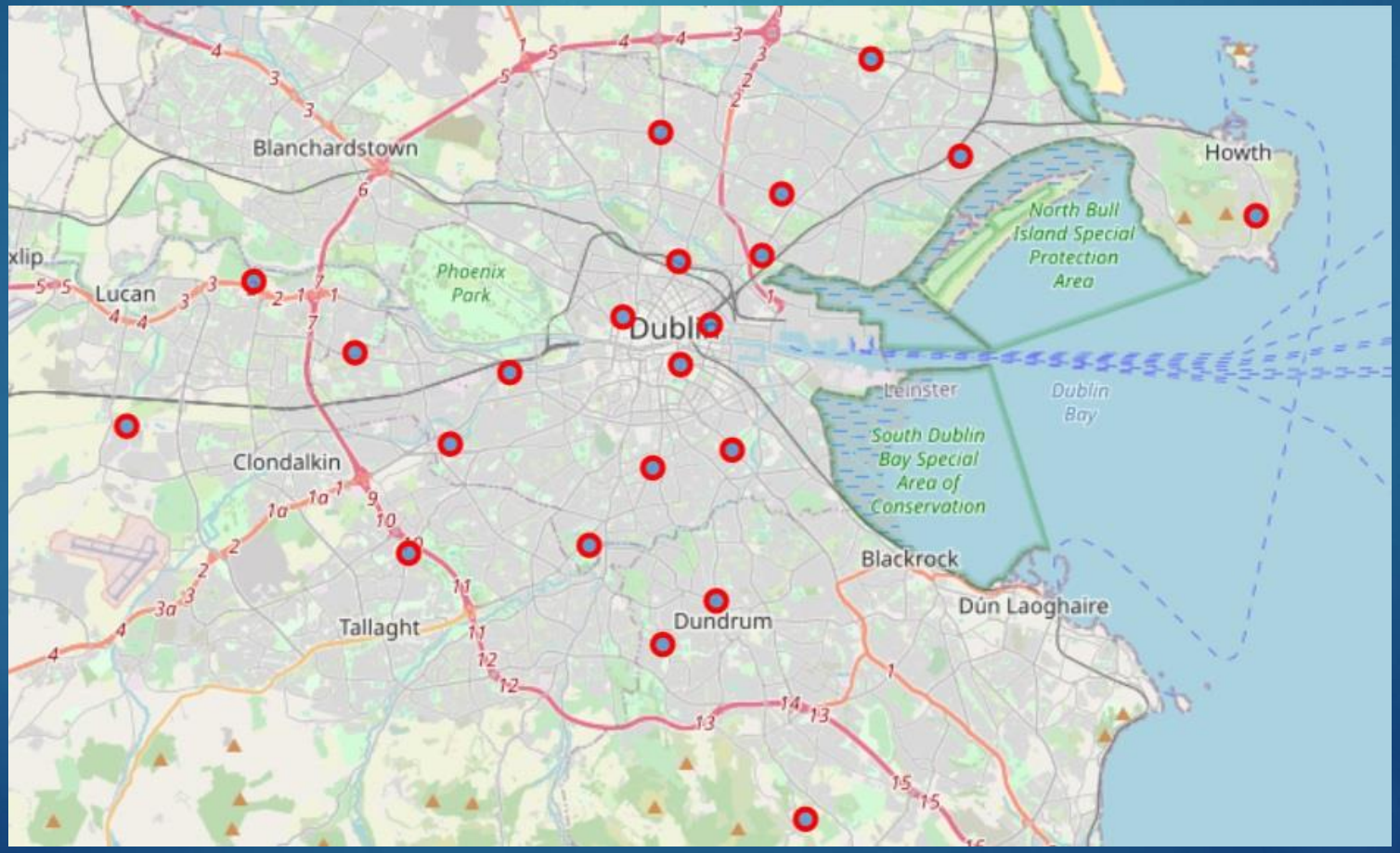
Data Acquisition and Preparation – Districts of Dublin: Building the Data Frame (2)

- We downloaded the file from wikipedia, we parsed it using the *html5lib* parser and we scraped it using the *BeautifulSoup* scraper.
- We looked for the table containing the information about the districts and we built a data frame containing the districts of Dublin, their corresponding postal codes and their descriptions.
- Using the *Nominatim* geocoder from the *geopy* library, we converted addresses into latitudes and longitudes and added them to the data frame.

	Postal code	District	Description	Latitude	Longitude
0	D1	Dublin 1	Dublin	54.654106	-6.234020
1	D2	Dublin 2	Dublin	54.656109	-6.231326
2	D3	Dublin 3	Dublin	52.653599	-7.243743
3	D4	Dublin 4	Dublin, Dun Laoghaire Rathdown	53.324171	-6.240135
4	D5	Dublin 5	Dublin	53.385074	-6.161141

Data Acquisition and Preparation – Districts of Dublin: Rendering the Map

- Using the *folium* library, we created a map of Dublin and we added to the map a marker for each district:



Data Acquisition and Preparation – Venues in the Districts of Dublin

- In order to find the venues in the districts of Dublin we used the *Foursquare* API.
- For each district in Dublin, we got the venues situated in a circle with a radius of 2 kilometers, centered in the latitude and longitude coordinates of the district.
- We created a data frame which contains all the venues in the districts of Dublin.
- Each venue is a row in our data frame and is characterized by the following features: name of district, latitude of district, longitude of district, name of venue, latitude of venue, longitude of venue and category of venue.

Methodology – Exploratory Data Analysis: Analyzing a District in Dublin

- We explored a central district in Dublin, the one with the D1 postal code.
- Using the *Foursquare* API, we got the venues situated in a circle with a radius of 2 kilometers, centered in the latitude and longitude coordinates of the district.
- We transformed the *json* format list into a *pandas* data frame and filtered the columns regarding to venue name, venue category, venue latitude and venue longitude.
- We got a data frame containing 100 venues in the D1 district of Dublin.

	name	categories	lat	lng
0	EPIC The Irish Emigration Museum	Museum	53.348323	-6.248131
1	The Famine Memorial	Sculpture Garden	53.348059	-6.250108
2	Seven Wonders	Café	53.348313	-6.243677
3	Bread 41	Bakery	53.344812	-6.251619
4	Shoe Lane Coffee	Café	53.347147	-6.255075

Methodology – Inferential Statistical Testing: Analyzing all the Districts in Dublin

- Using the paradigm described in the previous slide, we created a data frame which contains all the venues in the districts of Dublin.
- Each venue is a row in our data frame and is characterized by the following features: name of district, district latitude, district longitude, name of venue, venue latitude, venue longitude and venue category.

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Dublin 1	53.350062	-6.247473	EPIC The Irish Emigration Museum	53.348323	-6.248131	Museum
1	Dublin 1	53.350062	-6.247473	The Famine Memorial	53.348059	-6.250108	Sculpture Garden
2	Dublin 1	53.350062	-6.247473	Seven Wonders	53.348313	-6.243677	Café
3	Dublin 1	53.350062	-6.247473	Bread 41	53.344812	-6.251619	Bakery
4	Dublin 1	53.350062	-6.247473	Shoe Lane Coffee	53.347147	-6.255075	Café

Methodology – Inferential Statistical Testing: Analyzing all the Districts in Dublin (2)

- Applying the *get_dummies* function from the *pandas* library on the venue category column of the data frame, we transformed it into a new one with numerical values.
- We grouped rows by district and by taking the mean of the frequency of occurrence of each category of venues and we named the new data frame *dublin_grouped*.
- In order to better understand our data, we sorted the categories of venues in descending order regarding the frequency of occurrence of each category and built another data frame containing each district along with the top most common venues.

	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Dublin 1	Coffee Shop	Café	Pub	Park	Italian Restaurant	Plaza	Hotel	Pizza Place	Theater	Ice Cream Shop
1	Dublin 10	Coffee Shop	Supermarket	Fast Food Restaurant	Gym	Bistro	Furniture / Home Store	Pub	Grocery Store	Burger Joint	Café
2	Dublin 11	Coffee Shop	Supermarket	Park	Grocery Store	Fast Food Restaurant	Soccer Field	Café	Sandwich Place	Restaurant	Pub
3	Dublin 12	Supermarket	Park	Grocery Store	Fast Food Restaurant	Coffee Shop	Motorcycle Shop	Tram Station	Diner	Go Kart Track	Convenience Store
4	Dublin 13	Café	Pub	Trail	Deli / Bodega	Golf Course	Ice Cream Shop	Seafood Restaurant	Coffee Shop	Playground	Museum

Methodology – Inferential Statistical Testing: Analyzing all the Districts in Dublin (3)

- Based on the information in the data frame presented in the previous slide and considering Mr. Guinness' preferences, we extracted from the *dublin_grouped* data frame the relevant columns :*'District', 'Pub', 'Café', 'Coffee Shop', 'Park', 'Beach', 'Gym', 'Theater', 'Supermarket', 'Grocery Store', 'Convenience Store', 'Shopping Mall', 'Chinese Restaurant', 'Italian Restaurant', 'Light Rail Station', 'Bus Stop', and built a new data frame.*
- Its subset containing only numerical data (the frequencies of occurrence of each venue category) was used for the clustering process.

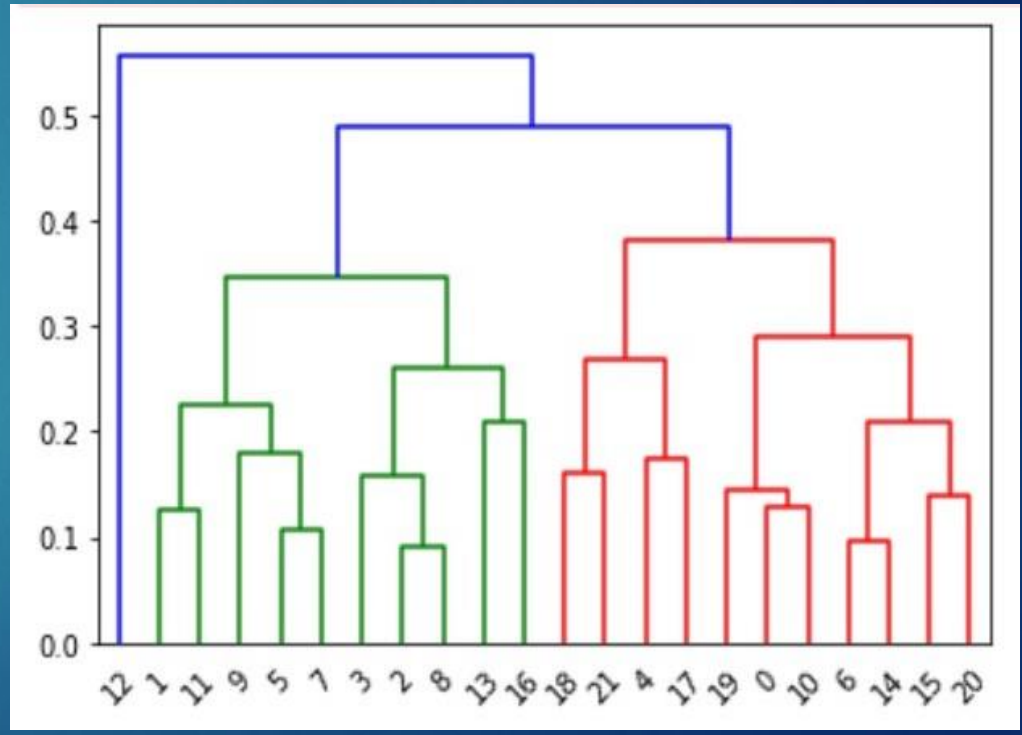
	District	Pub	Café	Coffee Shop	Park	Beach	Gym	Theater	Supermarket	Grocery Store	Convenience Store	Shopping Mall	Chinese Restaurant	Italian Restaurant	Light Rail Station	Bus Stop
0	Dublin 1	0.050000	0.090000	0.110000	0.040000	0.0	0.010000	0.030000	0.010000	0.010000	0.000000	0.000000	0.000000	0.030000	0.0	0.0
1	Dublin 10	0.036364	0.036364	0.072727	0.036364	0.0	0.036364	0.000000	0.072727	0.036364	0.036364	0.018182	0.000000	0.018182	0.0	0.0
2	Dublin 11	0.037736	0.037736	0.094340	0.075472	0.0	0.018868	0.018868	0.094340	0.075472	0.018868	0.018868	0.018868	0.000000	0.0	0.0
3	Dublin 12	0.000000	0.025000	0.050000	0.075000	0.0	0.000000	0.000000	0.125000	0.050000	0.025000	0.025000	0.025000	0.000000	0.0	0.0
4	Dublin 13	0.125000	0.156250	0.031250	0.000000	0.0	0.000000	0.000000	0.000000	0.031250	0.031250	0.000000	0.000000	0.000000	0.0	0.0

Methodology – Machine Learning Technics: Clustering the Districts in Dublin

- We used the agglomerative clustering algorithm which performs a hierarchical clustering using a bottom up approach: pairs of clusters are merged together as they move up the hierarchy.
- The steps of the algorithm are as follows:
  - create n clusters and assign one for each data point.
  - compute the distance matrix.
  - repeat:
    - merge the two closest clusters.
    - update the distance matrix.
  - until only a single cluster remains.

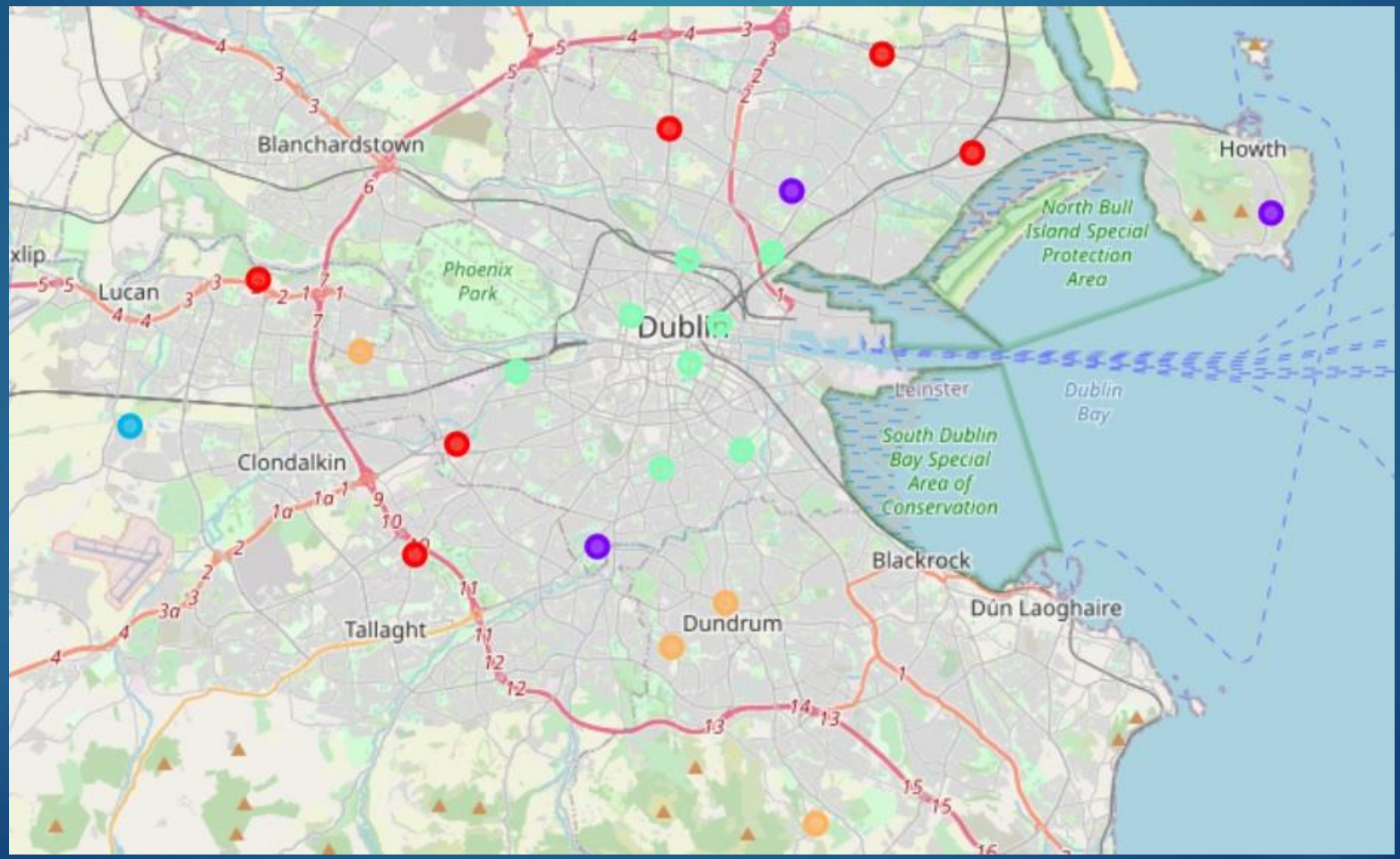
Methodology – Machine Learning Technics: Clustering the Districts in Dublin (2)

- Why did we chose the agglomerative hierarchical clustering algorithm?
  - The algorithm produces a dendrogram which can help to better understand the data Sand how the clusters were built.
  - Our data set is a small one, therefore we do not need a more efficient algorithm.
  - The algorithm always generates the same clusters.
- The algorithm can be visualized building a dendrogram. Each merge of clusters is a horizontal line in the dendrogram and the y coordinate of the horizontal line represents the similarity of the two clusters that were merged.



Results

- Five clusters have been generated by the agglomerative hierarchical clustering algorithm:



Results: Cluster 1 (Label 0) and Cluster 3 (Label 2)

District	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Dublin 5	0	Supermarket	Gym	Café	Pub	Chinese Restaurant	Shopping Mall	Convenience Store	Grocery Store	Beach	Park
Dublin 11	0	Supermarket	Coffee Shop	Grocery Store	Park	Café	Pub	Chinese Restaurant	Shopping Mall	Convenience Store	Theater
Dublin 12	0	Supermarket	Park	Grocery Store	Coffee Shop	Chinese Restaurant	Shopping Mall	Convenience Store	Café	Bus Stop	Light Rail Station
Dublin 17	0	Supermarket	Coffee Shop	Grocery Store	Park	Chinese Restaurant	Shopping Mall	Convenience Store	Café	Pub	Bus Stop
Dublin 20	0	Grocery Store	Supermarket	Coffee Shop	Pub	Park	Bus Stop	Italian Restaurant	Shopping Mall	Convenience Store	Light Rail Station
Dublin 24	0	Supermarket	Park	Pub	Light Rail Station	Shopping Mall	Coffee Shop	Bus Stop	Italian Restaurant	Chinese Restaurant	Convenience Store

District	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Dublin 22	2	Convenience Store	Bus Stop	Italian Restaurant	Grocery Store	Pub	Light Rail Station	Chinese Restaurant	Shopping Mall	Supermarket	Theater

Results: Cluster 4 (Label 3)

District	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Dublin 1	3	Coffee Shop	Café	Pub	Park	Italian Restaurant	Theater	Grocery Store	Supermarket	Gym	Bus Stop
Dublin 2	3	Café	Coffee Shop	Pub	Park	Theater	Italian Restaurant	Chinese Restaurant	Grocery Store	Bus Stop	Light Rail Station
Dublin 3	3	Café	Pub	Italian Restaurant	Coffee Shop	Supermarket	Grocery Store	Park	Bus Stop	Chinese Restaurant	Gym
Dublin 4	3	Café	Park	Pub	Coffee Shop	Grocery Store	Italian Restaurant	Supermarket	Gym	Bus Stop	Light Rail Station
Dublin 6	3	Café	Pub	Park	Coffee Shop	Supermarket	Italian Restaurant	Chinese Restaurant	Grocery Store	Gym	Bus Stop
Dublin 7	3	Pub	Coffee Shop	Café	Theater	Park	Italian Restaurant	Grocery Store	Supermarket	Bus Stop	Light Rail Station
Dublin 8	3	Park	Café	Coffee Shop	Pub	Supermarket	Grocery Store	Italian Restaurant	Chinese Restaurant	Shopping Mall	Convenience Store
Dublin 15	3	Café	Pub	Coffee Shop	Supermarket	Park	Italian Restaurant	Grocery Store	Theater	Bus Stop	Light Rail Station

Results: Cluster 2 (Label 1) and Cluster 5 (Label 4)

District	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Dublin 6W	1	Pub	Supermarket	Café	Grocery Store	Italian Restaurant	Gym	Park	Coffee Shop	Bus Stop	Shopping Mall
Dublin 9	1	Pub	Café	Bus Stop	Supermarket	Grocery Store	Convenience Store	Gym	Park	Italian Restaurant	Chinese Restaurant
Dublin 13	1	Café	Pub	Convenience Store	Grocery Store	Coffee Shop	Bus Stop	Light Rail Station	Italian Restaurant	Chinese Restaurant	Shopping Mall

District	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Dublin 10	4	Supermarket	Coffee Shop	Convenience Store	Grocery Store	Gym	Park	Café	Pub	Italian Restaurant	Shopping Mall
Dublin 14	4	Supermarket	Coffee Shop	Café	Pub	Convenience Store	Italian Restaurant	Shopping Mall	Gym	Park	Light Rail Station
Dublin 16	4	Supermarket	Pub	Coffee Shop	Café	Shopping Mall	Park	Italian Restaurant	Gym	Light Rail Station	Theater
Dublin 18	4	Pub	Convenience Store	Coffee Shop	Café	Light Rail Station	Shopping Mall	Supermarket	Gym	Park	Bus Stop

Discussion: Observations

- For each district, the ten most relevant categories of venues were considered in order to describe it.
- Each cluster was described by the six most relevant venue categories among all its member districts.
- The table below summarize the information presented above and was presented to Mr. Guinness:

cluster	no. of districts	first cat.	second cat.	third cat.	fourth cat.	fifth cat.	sixth cat.
1	6	supermarkets	parks	shopping malls	grocery stores	coffee shops	pubs
2	3	pubs	cafés	grocery stores	bus stops	Italian restaurants	supermarkets
3	1	convenience stores	bus stops	Italian restaurants	grocery stores	pubs	light rail stations
4	8	cafés	coffee shops	pubs	Italian restaurants	parks	grocery stores
5	4	supermarkets	coffee shops	pubs	cafés	gyms	parks

Discussion: Recommendations

- If Mr. Guinness prefers a central district with cafés, coffee shops, pubs, Italian restaurants and parks, one of the eight districts in ***cluster 4*** will be the best choice.
- If he prefers supermarkets, parks, shopping malls, coffee shops and pubs, one of the six districts in ***cluster 1*** could be a good choice.
- If his first preferences are supermarkets, coffee shops, pubs, cafés, gyms and parks, then one of the four districts in ***cluster 5*** will be a good choice.
- One of the three districts in ***cluster 2*** could be a good choice if he prefers pubs, cafés, Italian restaurants, supermarkets and is interested in public transportation.
- The single district in ***cluster 5*** seems to be a poor choice. It contains convenience stores, bus stops, Italian restaurants, grocery stores, pubs and light rail stations.

Conclusion ¶

- The target of this project was to identify the districts in Dublin which best fit Mr. Guinness' preferences.
- A data frame, containing the districts of Dublin, their corresponding postal codes, their descriptions and their latitude and longitude values, was build.
- A second data frame containing all the venues in the districts of Dublin was built using the *Foursquare* API. The venues were grouped by district and by taking the mean of the frequency of occurrence of each category of venues.
- Considering Mr. Guinness' preferences, the relevant columns were extracted from this data frame and a new data frame was built.
- The agglomerative clustering algorithm, which performs a hierarchical clustering using a bottom up approach, was used in order to group the districts in Dublin.
- A summary of the results was presented to Mr. Guinness along with our recommendations.
- As future work, we consider to develop the presented model by adding information regarding rent levels for apartments in different districts in Dublin.