Pieter du Toit
12 October 2019

IBM Datascience

Capstone Project

IBM Datascience 1
Capstone Project 1
Problem: 3
Who would be interested: 3
Description: 4
Methodology: 5
Results: 6
Discussion and observations: 18
Conclusion: 19

Problem:

A company is looking at setting up a pyrolysis plant to turn plastic into fuel, they want to set it up Cape Town and want to know which areas will be best suitable. They want neighbourhoods with Warehouses, Factories, Grocery stores, industrial parks, storage facilities and Liquor Stores, as they can place plastic recycle bins for collection at these locations. They also want to know where would be the best location to for the plant to run it efficiently and cost effective.

Who would be interested:

Any company that works in Waste Management and recycling, as well as Municipalities, as it will give them a better idea which areas may require more recycle bins, due to a higher cluster of certain types of stores in the area.

Description:

Data was obtained from various sources, I did a scrape of https://en.wikipedia.org/wiki/List_of_Cape_Town_suburbs and grouped all neighbourhoods with the same street code together. Some neighbourhoods that did not have street codes had to be dropped, these are informal settlements in the city. I then obtained the geolocation for each neighbourhood, using nominatum from geopandas. I exported these results to a csv file so it could be imported into the final project. I also did a scrape of https://www.expatistan.com/cost-of-living/toronto?currency=USD , https://www.expatistan.com/cost-of-living/cape-town?currency=USD and https://www.expatistan.com/cost-of-living/new-york to get the cost of living for these 3 cities, so a comparison could be done between costs in developed countries vs a developing country. The data from the site had to be cleaned up a bit, and was left with the normal expense types an person will have to a live in a City.

I used the geolocation data from the saved csv file with the Four Square api, to collect the locations of Convenience Stores, Department Stores, Discount Stores, Factories, Flea Markets, Fruit & Vegetable Stores, Grocery Stores, Liquor Stores, Markets, Recycling Facilities, Shopping Malls, Storage Facilities, Supermarkets, Warehouses and Warehouse Stores. The Warehouses, Factories and Recycling Facilities was included so we can see which neighbourhoods already has industrial areas.

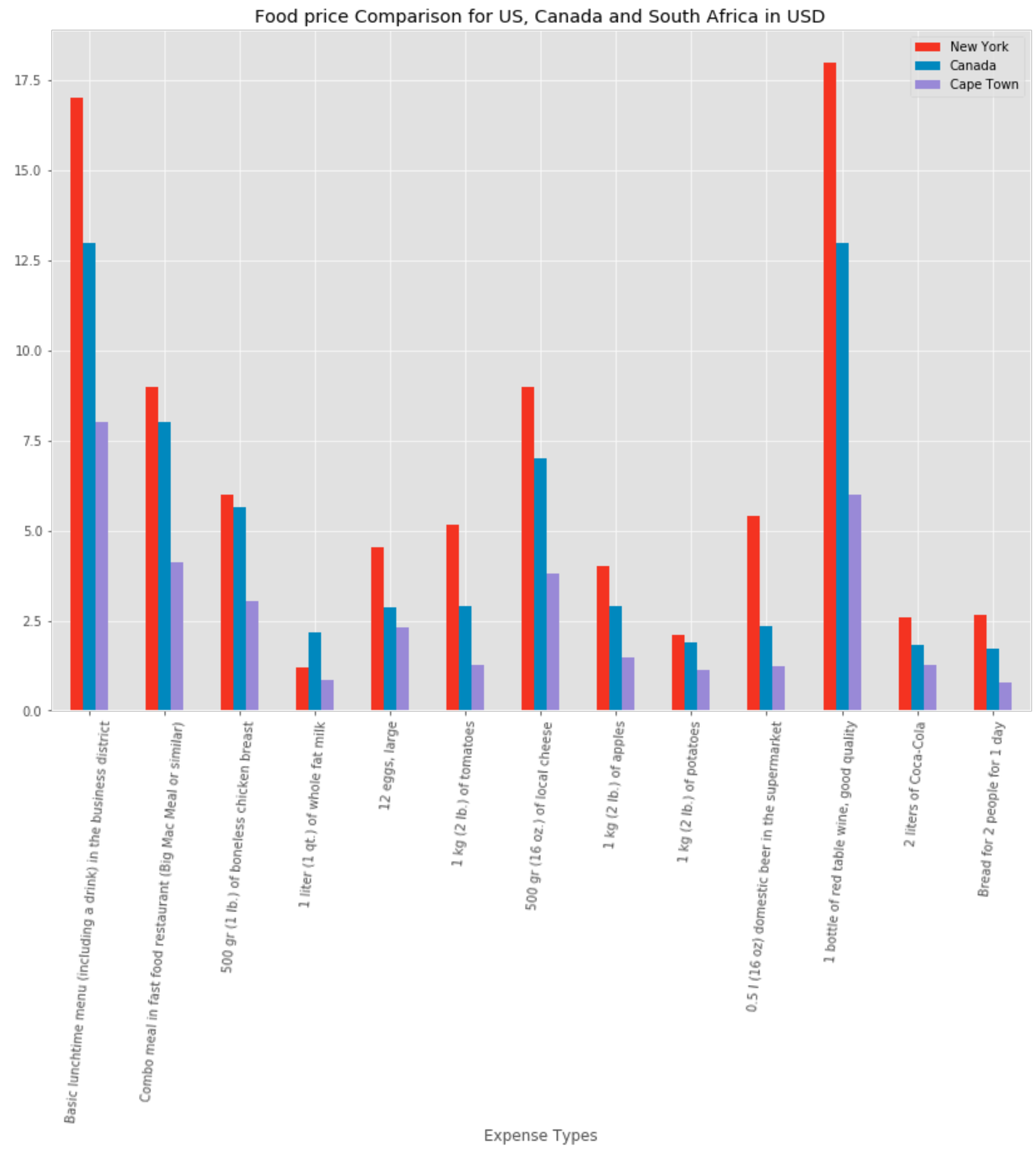
Methodology:

First I made to graphs to compare the cost of living for the 3 cities New York, Toronto and Cape Town. As you will see there was some interesting result here.

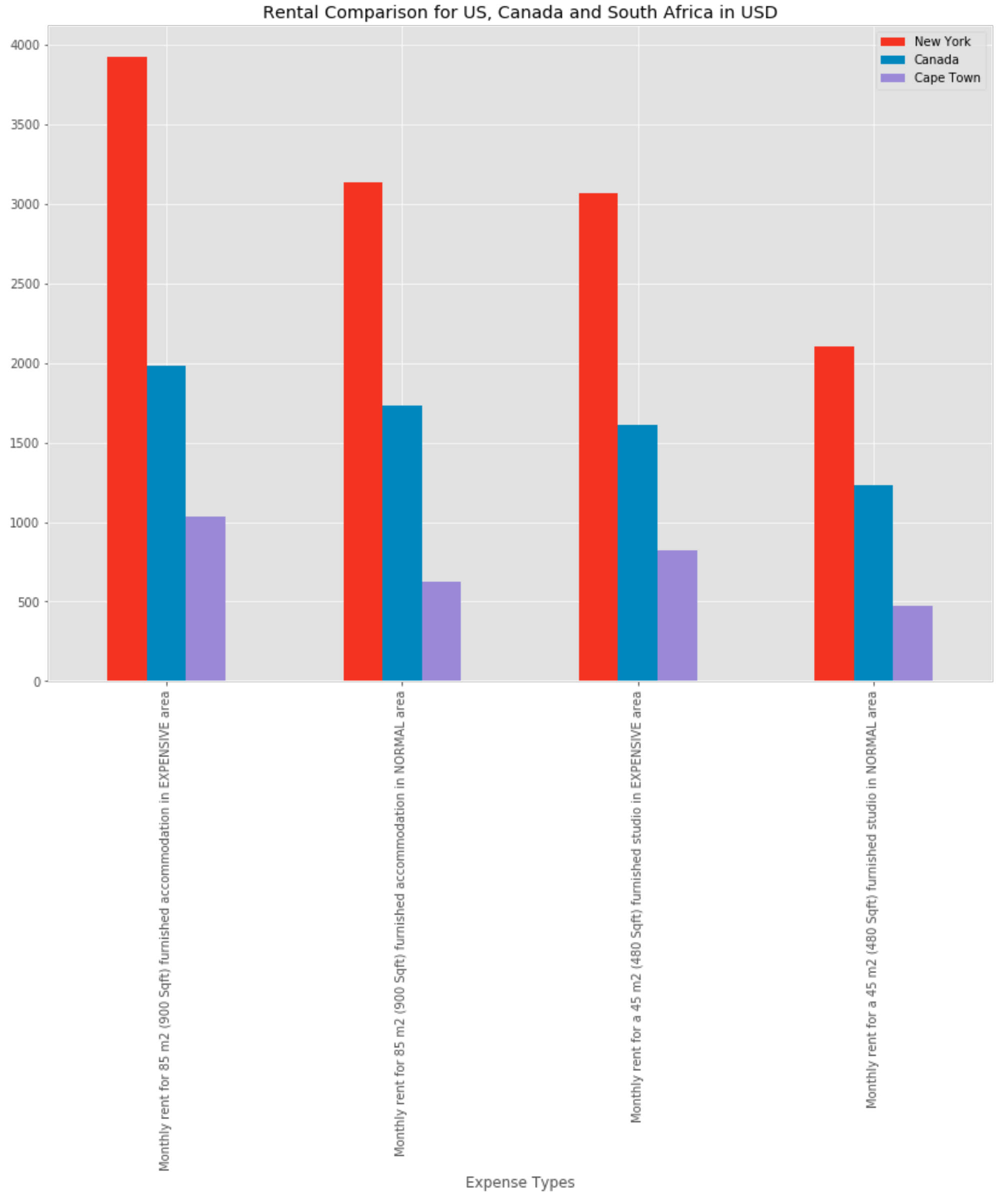
I used the venue data from square and got the 10 most popular places for each neighbourhood. K-means clustering of 5 was used to cluster neighbourhoods that had a similar venues together. I then did a count on the cluster values to determine which cluster had the most similar neighbourhoods. To confirm this I mapped the data with folium using different dot markers for each cluster. K means was used as it as a type of unsupervised machine learning that can be used on data that doesn't seem to have a pattern, it will cluster the data together using feature similarity.

Once I had the cluster I selected only the venues for those neighbourhoods. I calculated the mean of the coordinates for all the different venues to select an area where the company can look to place their pyrolysis plant. I then mapped all the venues out using a different colour for shopping malls and recycling venues, and used a larger radius for the point the company should start looking at placing the plant.

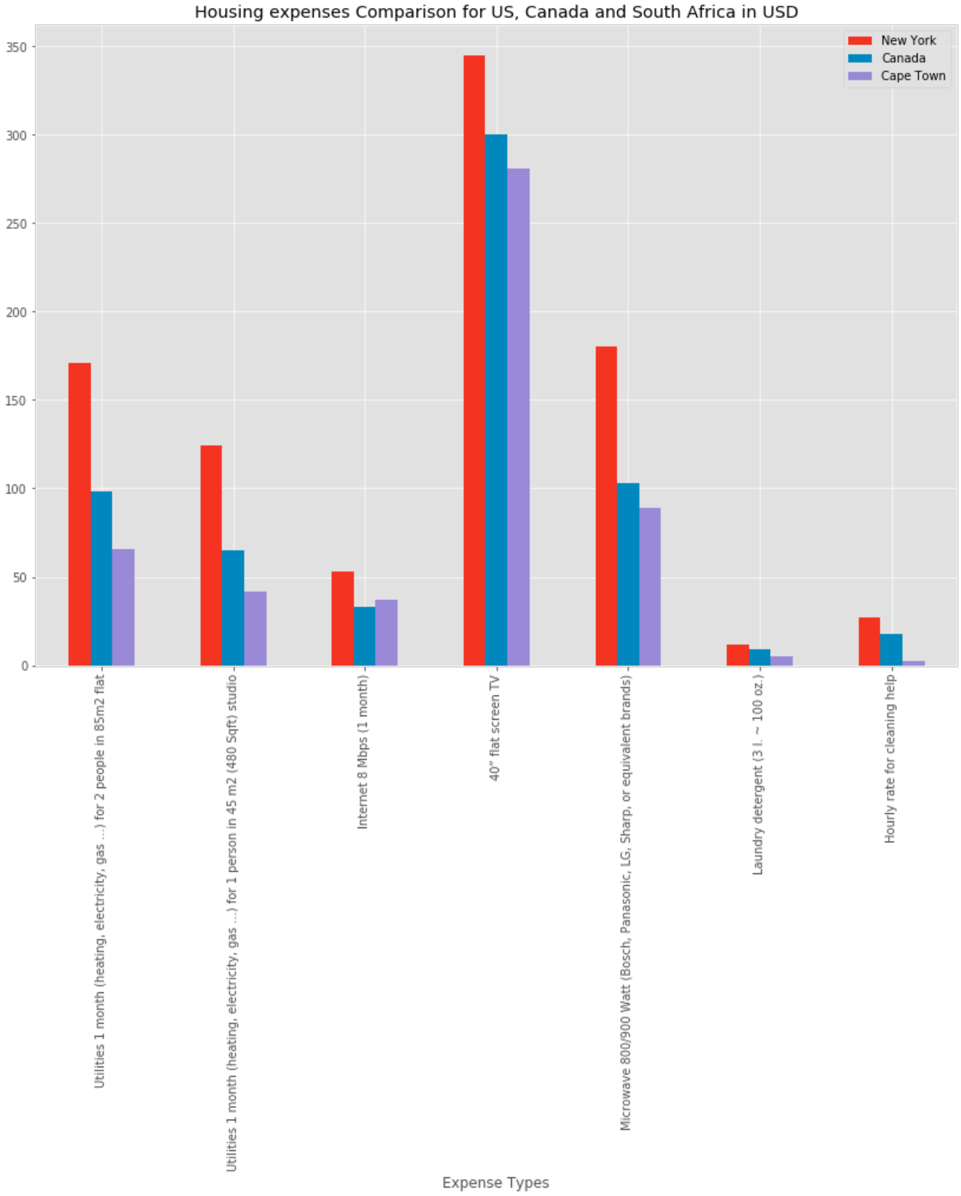
Results:



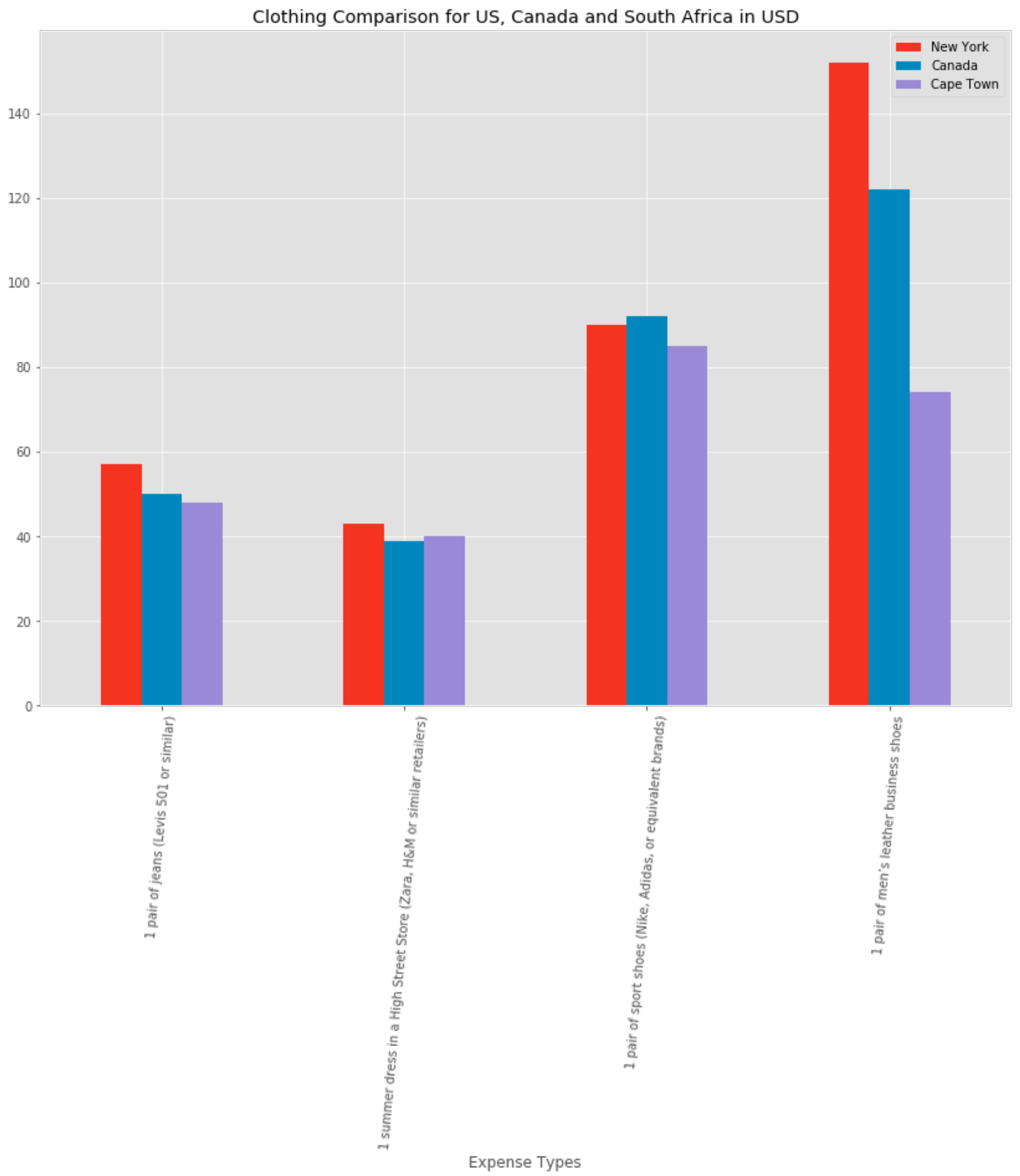
From the graph comparing food prices in the 3 cities, we can see that the cost of food in Cape Town when converted to USD is much lower than that of the two other cities.



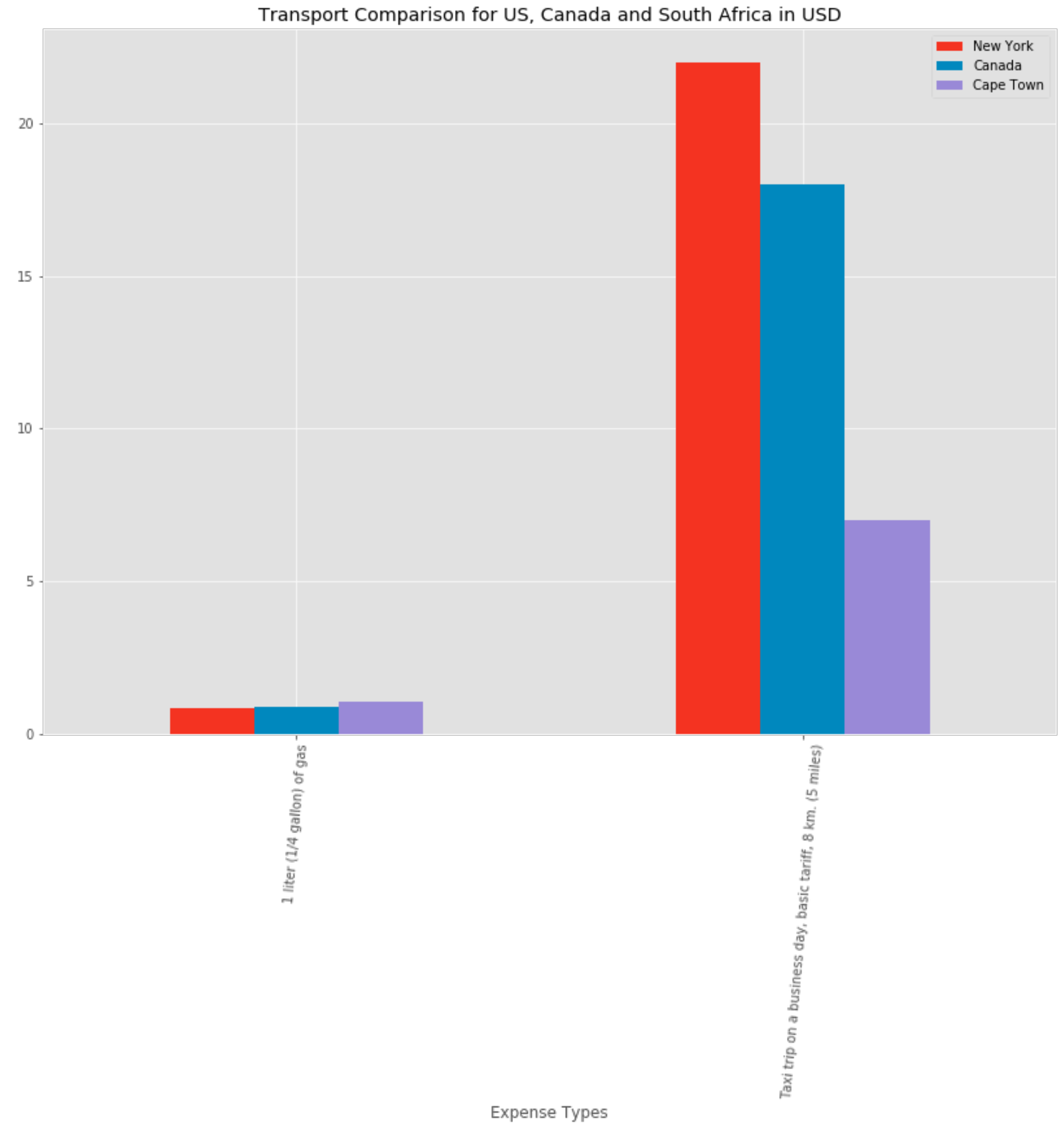
With Apartment rental, the same trend is there, the cost is much lower in Cape Town, where in New York people are spending almost 4x more on rental than in Cape Town.



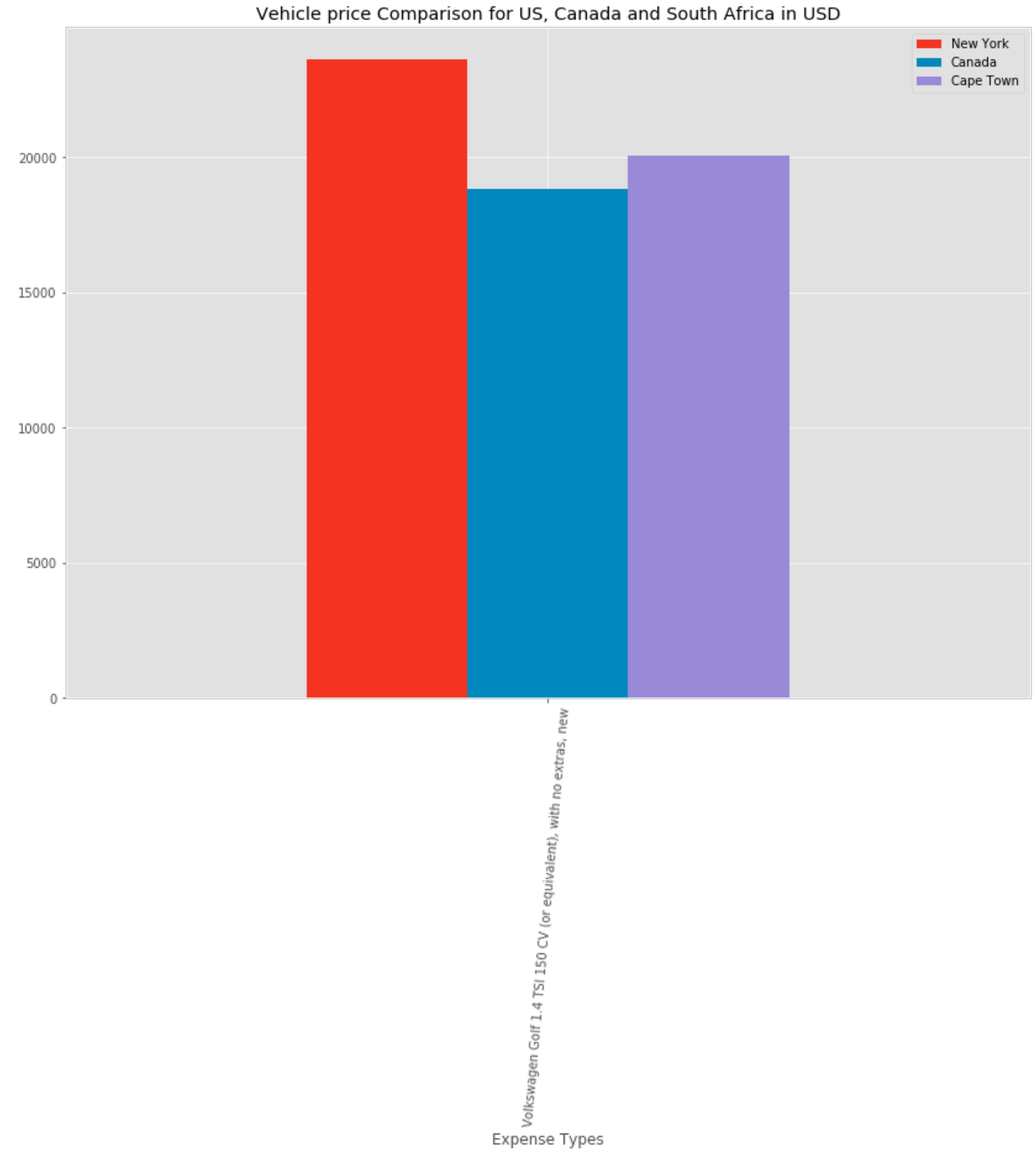
On normal housing expenses, we see that Toronto is the cheapest of the three cities when it comes to internet. But for all the rest Cape Town is much cheaper.



The clothing expenses for all the cities looks to be somewhat equal except for mens shoes that follows the same trend as all the other data. The primary reason why Cape Town might not be cheaper in this category is for the fact that some business found that it was cheaper to import clothing than to source local manufactured goods.



Comparing fuel prices we also see an anomaly, fuel in Cape Town is more expensive than in New York and Toronto, but the public transport is almost 3x cheaper than New York. The reason for this is that fuel is heavily taxed in South Africa, almost 60% of the fuel price is tax. The taxi industry in South-Africa makes up about 69% of the public transport system in South-Africa. The taxi industry spends around 2.6 billion USD a year on fuel, and about 135 million USD on insurance. The biggest problem for the government is that a large percentage of these taxi operators are not registered.



With vehicle price comparison we see that Toronto, is a little bit cheaper than South-Africa.

Analysing Cape Town Neighbourhoods:

Cape Town has 8 boroughs and 41 neighbourhoods.

The below shows the header values in the data frame.

	Borough	Neighbourhood	Latitude	Longitude
0	Cape Flats	Delft	-33.965556	18.644444
1	Cape Flats	Macassar	-34.066116	18.767495
2	Helderberg	Firgrove, Somerset West	-34.040539	18.455753
3	Helderberg	Gordon's Bay, Strand	-34.161125	18.868687
4	Northern Suburbs	Brooklyn, Kensington, Maitland, Rugby	-33.908889	18.479167

Below is the header for the venues that was collected using the Four Square API.

Neighbourhood	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	
0	Delft	-33.965556	18.644444	FreshStop at Caltex Voorbrugh	-33.974548	18.642966	Convenience Store
1	Delft	-33.965556	18.644444	FreshStop at Caltex Delft	-33.955997	18.647024	Convenience Store
2	Delft	-33.965556	18.644444	Engen Bellstar Service Station	-33.951365	18.647774	Convenience Store
3	Delft	-33.965556	18.644444	SPAR	-33.942385	18.626650	Convenience Store
4	Delft	-33.965556	18.644444	Engen Hindle Road Service Station	-33.979934	18.657914	Convenience Store

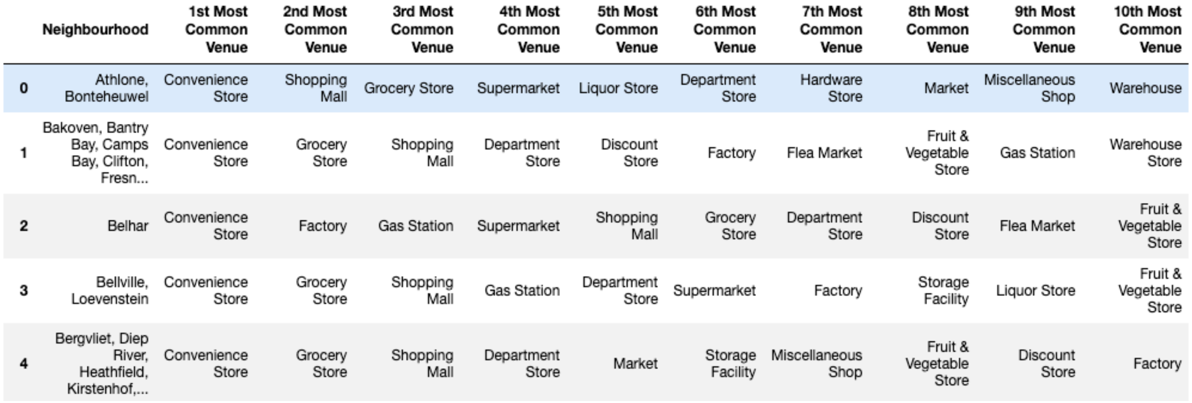
18 Unique categories was returned with a total of 1527 venues.

The following venues types was returned:

- Convenience Store
- Department Store
- Discount Store
- Factory
- Flea Market
- Fruit & Vegetable Store
- Gas Station
- Grocery Store
- Hardware Store
- Liquor Store
- Market
- Miscellaneous Shop
- Recycling Facility
- Shopping Mall
- Storage Facility
- Supermarket
- Warehouse
- Warehouse Store

All the category was converted with one hot encoder, to give the labels a value so we could get the most common venues for each neighbourhood.

Below is the header for the resulting data frame.



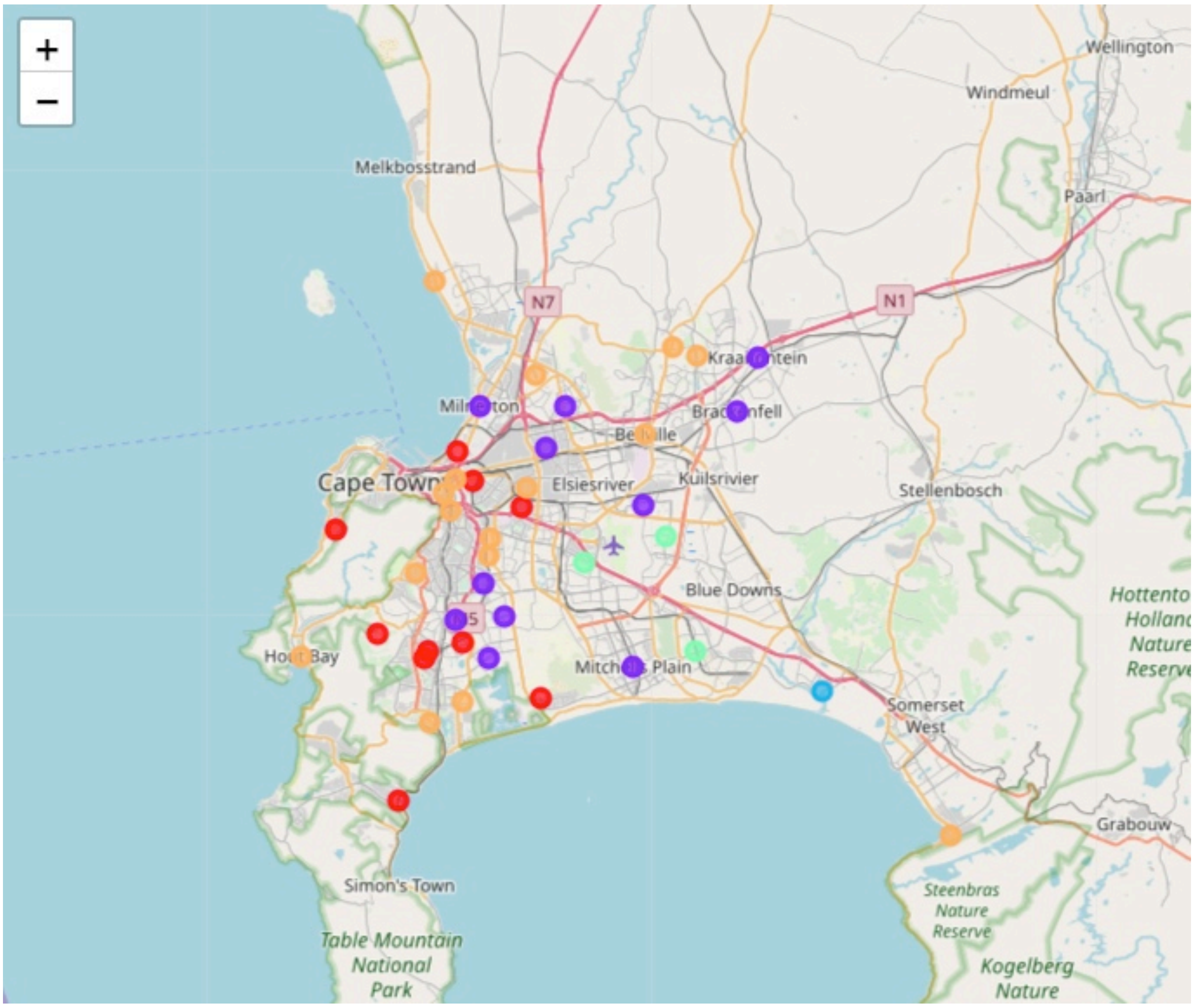
	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Athlone, Bonteheuwel	Convenience Store	Shopping Mall	Grocery Store	Supermarket	Liquor Store	Department Store	Hardware Store	Market	Miscellaneous Shop	Warehouse
1	Bakoven, Bantry Bay, Camps Bay, Clifton, Fresn...	Convenience Store	Grocery Store	Shopping Mall	Department Store	Discount Store	Factory	Flea Market	Fruit & Vegetable Store	Gas Station	Warehouse Store
2	Belhar	Convenience Store	Factory	Gas Station	Supermarket	Shopping Mall	Grocery Store	Department Store	Discount Store	Flea Market	Fruit & Vegetable Store
3	Bellville, Loevenstein	Convenience Store	Grocery Store	Shopping Mall	Gas Station	Department Store	Supermarket	Factory	Storage Facility	Liquor Store	Fruit & Vegetable Store
4	Bergvliet, Diep River, Heathfield, Kirstenhof,...	Convenience Store	Grocery Store	Shopping Mall	Department Store	Market	Storage Facility	Miscellaneous Shop	Fruit & Vegetable Store	Discount Store	Factory

I could then cluster this data using the k-means cluster function, I used 5 clusters and created a cluster label for each each group that corresponds to a number.

The clustered data gave the below result.

Cluster label	Number of Neighbourhoods
4	16
1	11
0	10
3	3
2	1

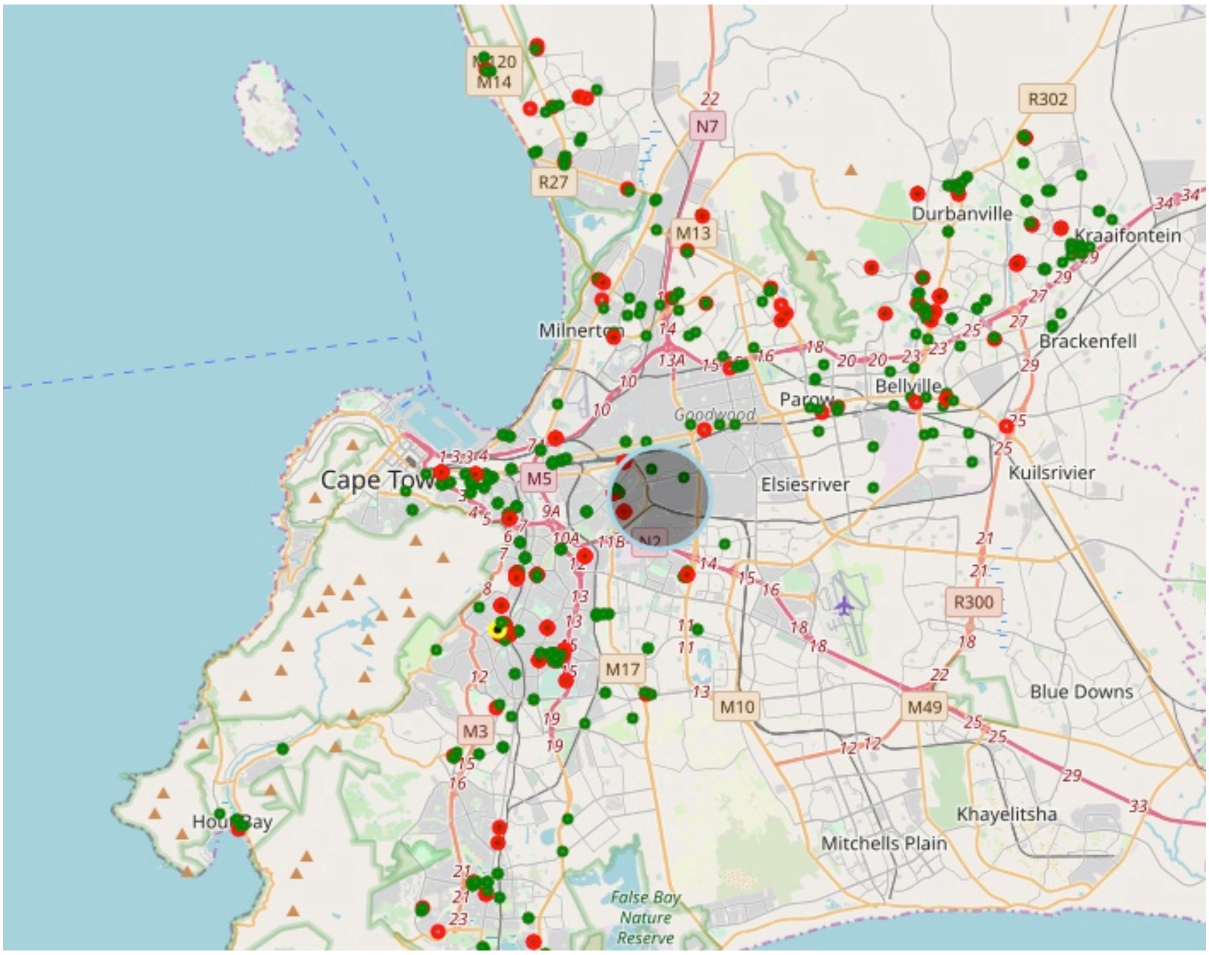
As per the below map of the clusters, we can see that cluster 4 which is in gold has the most dots on the map



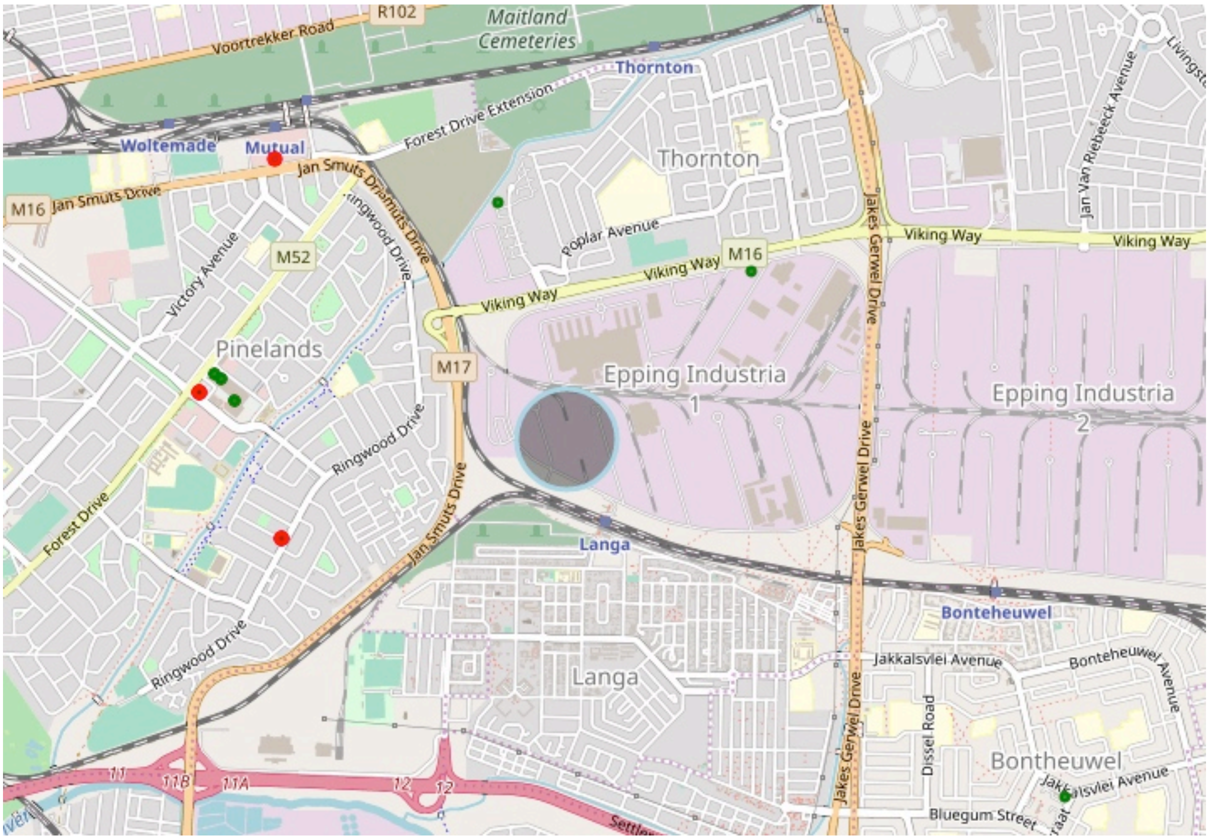
I then got all the venues for only the neighbourhoods in cluster 4 and calculated the mid point of the longitude and latitude.

```
(-33.93553910108302, 18.528002089769064)
```

I then created a new map and plotted all the points for the different venues, showing the Malls in red, and recycle facilities in yellow. I also placed a circle marker on the map for the midpoint calculated above, as that would be the most cost effective area for the company to look at setting up there plant, if will also allow for the trucks that need to to collect the recycle bins at the venues. The problem is, did I get an area that is in an industrial part of the city?



If I zoom in on the map to the suggested area, I can see the name of the neighbourhood.



We see the area is called Epping Industria 1, if we do a google search on Epping Industry we get the wikipedia page https://en.wikipedia.org/wiki/Epping,_Cape_Town , we see that Epping is indeed an industrial area in Cape Town.

Discussion and observations:

Most of the data could be found online, but due to layout of Cape Town with so many informal settlements, I had to push the search radius for the venue up to 4000 meters, to get results for some of the neighbourhoods. What was also interesting is that you could get data for air pollution and water pollution, but not really data for pollution of inland estuaries and ground pollution due to plastic. Water pollution data is mostly focused around contaminants in the water and does not contain much data about of how many tons of plastic is discarded and how much is recycled in a city. What I could find is that the average South African uses about 38kg of plastic a year.

What would be an interesting project for later is to see if a drone can be used with AI image recognition to classify and map areas with high plastic pollution on land, so that cities or recycling companies can get to it before it gets into the rivers and oceans. As access to clean drinking water is quickly becoming one of the scarcest resources in on our planet.

Conclusion:

The company has chosen well in deciding to use to Cape Town to setup an pyrolysis plant for plastic to fuel, as it is a developing country and most people won't be able to move away from fossil fuel power vehicles. Labour costs will also be cheaper than for other countries, but there is the risk of labour unions going on strike and halting operations a couple of times a year.

On the data end, I will be testing this model against some other cities as well, to see if all of the end results are close to industrial areas, for only then will I know if that the model is truly refined.