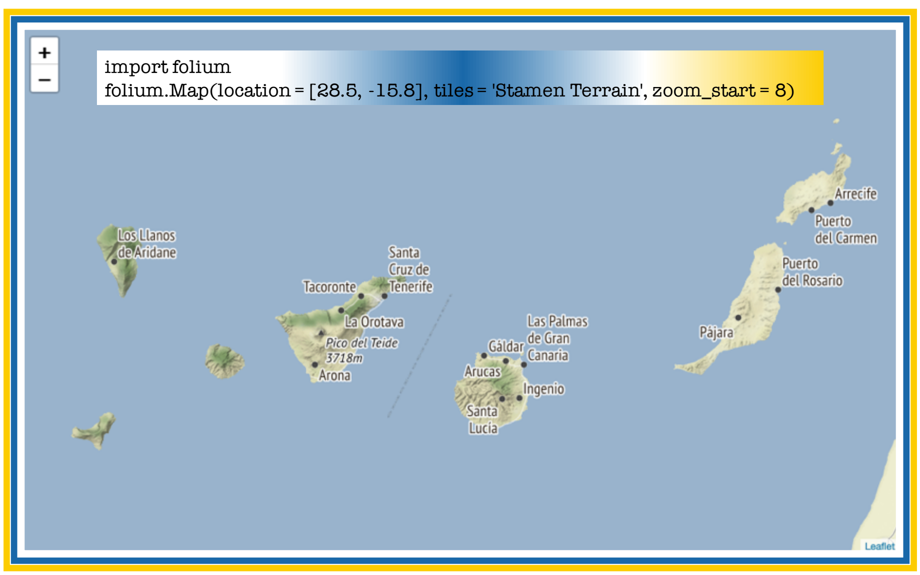
IBM Data Science Professional Certificate on Coursera

The Battle of the Neighbourhoods

Final assignment: Applied Data Science Capstone

Bij-Na Kim

January 23, 2021



1 Introduction

Gran Canaria is one of the group of volcanic islands that constitute Canary Islands. Located to the south of Spain and to the west cost of Morocco, the Canary Islands take their presence in the Atlantic Ocean. Given their location, the islands enjoy a very pleasant climate throughout the year thus attracting many tourists particularly during winter months. The tertiary sector accounts for 75 % of the economy in Canarias [1], where tourism is at the core of the islands activities. Construction has been booming as a result, as people are always seeking to invest in the housing market. Whether the intention is to use it solely as a touristic accommodation, or for a second home and explore the option of a short-let when the house is not in use, there is a need for a data-driven decision making.

Gran Canaria covers an area of $\sim 1500\,\text{km}^2$ and has $\sim$850,000 inhabitants. The island is composed of 21 municipalities [2], where the highest population density is in Las Palmas de Gran Canaria, which also serves as the capital city of the island. Each municipality is then made up of districts and neighbourhoods. Each municipality has its own unique characteristics. For instance, those located in the southern part of the islands are the top destinations for tourists due to the famous beaches, while those towards the centre of the island are ideal places for rural holidays and enjoying nature.

The choice of location is therefore important, and will largely be determined by what buyers are looking for, e.g. proximity to the beach, proximity to bars and restaurants, proximity to nature, quiet rural areas for resting, etc. Additionally, buyers will also need to have an indication on the return of investment. Historical data can be used to answer questions such as: when is the peak season for tourism in the island? How much does a tourist spend on accommodation on average? Therefore, the purpose of the current work is to assist potential buyers in making decisions on buying a holiday property in Gran Canaria.

2 About the dataset

As previously described in the Introduction, two broad data categories can be highlighted:

- What are the characteristics of the different areas within Gran Canaria? Understand what features make the areas distinct.
- What additional information can be gathered from public sources in terms of tourism trends in Gran Canaria?

Foursquare[1] will be used for discovering the venues within Gran Canaria. Taking advantage of the

[1]https://foursquare.com

free open API services Foursquare offers, the list of venues and their location data will be imported into Python.

The data source used for the island-related statistics will be from Instituto Canario de Estadistica (ISTAC), managed by the regional Government of the Canaries[2]. ISTAC contains an exhaustive catalogue of statistical data, ranging from economy, society, demographics, primary/secondary/tertiary sectors, average appraised home value, etc. All this information will weigh in when making a decision. The table summarises the datasets used in this work and the source:

Table 1: Table summarising the datasets to be used for the study.

Data	**Source**
List of venues	Foursquare
Location data for municipalities and districts	ISTAC
Basic demographic data for the municipalities	ISTAC
Average appraised home value for Las Palmas *province*	ISTAC
Expected high and low seasons for tourism	ISTAC

NB: For the average home value, the statistics are shown for the *province* of Las Palmas. The province of Las Palmas includes the islands of Gran Canaria, Fuerteventura and Lanzarote. Nevertheless, it will be interesting to observe the general trends within the islands.

3 Methodology

There are two sections to be completed in this work:

- Part 1: districts segmentation by types of venues present using Foursquare data.
- Part 2: key tourism data.

The Juypter Notebook used throughout the work, with detailed steps, can be found on GitHub:

`https://github.com/bee-57/Coursera_Capstone/tree/submission`

[2]https://www.gobiernodecanarias.org & https://datos.canarias.es

3.1 Part 1: districts segmentation by types of venues present using Foursquare data

3.1.1 Municipalities and districts in Gran Canaria

Figure 1 shows an extract of the Jupyter Notebook listing all municipalities and their corresponding municipality ID.

Figure 1: List of municipalities in Gran Caanria. Extract from Jupyter Notebook.

```
In [4]: # Filter to only obtain Gran Canara, i.e. 'cd_isla' needs to have a value of 'ES705'.
        municipalityGC = municipalityCanaries.loc[municipalityCanaries['gcd_isla'] == 'ES705'].reset_index(drop=True)
        print('There are {} municipalities listed for Gran Canaria.'.format(len(municipalityGC)))

        There are 21 municipalities listed for Gran Canaria.

In [10]: # We are only interested in keeping columns 'geocode' and 'etiqueta' from this table.
         # 'Geocode' contains the municipality identifier code and 'etiqueta' is the label, i.e. the name of the municipality.
         # Let's rename the columns.
         municipalityGC = municipalityGC[['geocode', 'etiqueta']]
         municipalityGC = municipalityGC.rename(columns={'geocode':'municipalityId', 'etiqueta':'municipalityName'})
         municipalityGC

Out[10]:
```

	municipalityId	municipalityName
0	35001	Agaete
1	35002	Agüimes
2	35005	Artenara
3	35006	Arucas
4	35008	Firgas
5	35009	Gáldar
6	35011	Ingenio
7	35012	Mogán
8	35013	Moya
9	35016	Las Palmas de Gran Canaria
10	35019	San Bartolomé de Tirajana
11	35020	La Aldea de San Nicolás
12	35021	Santa Brígida
13	35022	Santa Lucía de Tirajana
14	35023	Santa María de Guía de Gran Canaria
15	35025	Tejeda
16	35026	Telde
17	35027	Teror
18	35031	Valsequillo de Gran Canaria
19	35032	Valleseco
20	35033	Vega de San Mateo

Figure 2 shows the list of all districts in Gran Canaria. On the `disrictCanaries` dataframe, the municipalities are listed as municipality ID (`gcd_municipio`). The municipality ID will be used to merge the two dataframes in Pandas.

The final table resulting from the data transformation is shown on the `geolocGC` data shown in Figure 3. On this table, there are 6 columns:

- municipalityId - ID to refer a municipality,
- municipalityName - name of the municipality,
- districtLabel - label given to the district,
- longitude - longitude of the district centre,
- latitude - latitude of the district centre,
- surface - surface area of the district in km^2.

Figure 2: List of districts in Gran Canaria. Extract from Jupyter Notebook.

```
districtCanaries = pd.read_csv('https://datos.canarias.es/catalogos/estadisticas/dataset/1f8a16
districtCanaries
```

Out[5]:

	geocode	etiqueta	granularidad	gcd_isla	gcd_municipio	superficie	utm_x	utm_y	longitud
0	20170101_35001_D01	Distrito 01 - Agaete	DISTRITOS	ES705	35001	4452.5952	432237.32	3105495.86	-15.689645
1	20170101_35002_D01	Distrito 01 - Agüimes	DISTRITOS	ES705	35002	7877.5930	455417.28	3085973.99	-15.453004
2	20170101_35003_D01	Distrito 01 - Antigua	DISTRITOS	ES704	35003	24956.1021	603313.77	3136753.89	-13.945814
3	20170101_35004_D01	Distrito 01 - Arrecife	DISTRITOS	ES708	35004	407.7228	641771.25	3205171.30	-13.544991
4	20170101_35004_D02	Distrito 02 - Arrecife	DISTRITOS	ES708	35004	2001.6834	639719.60	3206522.95	-13.565874
...	...	...	...	...	...	...	...	...	...
169	20170101_38050_D03	Distrito 03 - Vallehermoso	DISTRITOS	ES706	38050	4261.0030	274764.89	3109110.56	-17.292308
170	20170101_38051_D01	Distrito 01 - La Victoria de Acentejo	DISTRITOS	ES709	38051	1823.4949	359072.69	3144204.96	-16.438806
171	20170101_38052_D01	Distrito 01 - Vilaflor de Chasna	DISTRITOS	ES709	38052	5642.2377	338263.87	3115921.95	-16.647254
172	20170101_38053_D01	Distrito 01 - Villa de Mazo	DISTRITOS	ES707	38053	7033.2659	226221.30	3164212.52	-17.798915
173	20170101_38901_D01	Distrito 01 - El Pinar de El Hierro	DISTRITOS	ES703	38901	8277.9702	203948.07	3065907.92	-18.001556

174 rows × 10 columns

In [6]:
```
# Again, we need to set a filter so that we only get those districts within Gran Canaria.
districtGC = districtCanaries.loc[districtCanaries['gcd_isla'] == 'ES705'].reset_index(drop=Tru
print('There are {} districts listed for Gran Canaria.'.format(len(districtGC)))
```

```
There are 53 districts listed for Gran Canaria.
```

Figure 3: Dataframe `geolocGC` listing all districts and geolocation data.

In [37]:
```
# Before running function 'getNearbyVenues' the dataframe 'venuesGC' needs some further manipulation.
# The API call requires a radius value. Given that districts are different in size, an arbitrary value will cause
# inaccuracies: smaller neighbouring districts will have an overlap, whereas larger districts may not be completely
# covered. Given that we have the surface area of each district, we can estimate the shape of the circle.
# The radius we extract (equivRadius, units meters) can be used for the API call.
geolocGC['equivRadius'] = (((geolocGC['surface']/math.pi)**0.5)*1000).round(decimals=0)
geolocGC['equivRadius'] = geolocGC['equivRadius'].astype('int64')
geolocGC.head()
```

Out[37]:

	municipalityId	municipalityName	districtLabel	longitude	latitude	surface	equivRadius
0	35001	Agaete	Distrito 01 - Agaete	-15.689645	28.073135	4452.5952	37647
1	35002	Agüimes	Distrito 01 - Agüimes	-15.453004	27.897893	7877.5930	50075
2	35005	Artenara	Distrito 01 - Artenara	-15.682969	28.018081	6641.6517	45979
3	35006	Arucas	Distrito 01 - Arucas	-15.520911	28.119561	228.0795	8521
4	35006	Arucas	Distrito 02 - Arucas	-15.540187	28.115284	332.6582	10290

The module `folium` was used for visualising all districts on the map shown in Figure 8(a).

3.1.2 Retrieving list of venues from Foursquare

Using the `geolocGC` dataframe a series of API calls were made to Foursquare. Aside from the user credentials, there are five other properties that need to be passed when requesting data from Foursquare:

- version: API version, '20180604',
- latitude: district latitude,
- longitude: district longitude,
- radius: radius of search from coordinates supplied,
- limit: limit on number of venues, '100'.

Since the districts are all different in size, having an arbitrary value for radius will introduce inaccuracies in the data. For smaller districts, the radius may extend to neighbouring districts, whereas for larger districts, the specified radius may partially cover the area.

Building up from the course tutorial (case study: Manhattan), in order to address this issue, the equivalent radius, `equivRadius`, was calculated using the surface area values provided on `geolocGC`, assuming a circular shape of the district:

$$\text{equivRadius} = \sqrt{\frac{\text{surface}}{\pi}} \tag{1}$$

Figure 4: Dataframe `venuesGC` listing the venues retrieved from Foursquare for each district.

```
In [38]: # For example, let's print the top rows from District 02 in Las Palmas.
         # This is where I personally like to hang out! :))
         (venuesGC.loc[venuesGC['district'] == 'Distrito 02 - Las Palmas de Gran Canaria']).head()
```

Out[38]:

	district	districtLatitude	districtLongitude	venue	venueLatitude	venueLongitude	venueCategory
2030	Distrito 02 - Las Palmas de Gran Canaria	28.11232	-15.421078	Pastelería Colomar	28.115799	-15.421875	Cupcake Shop
2031	Distrito 02 - Las Palmas de Gran Canaria	28.11232	-15.421078	Restaurante Allende	28.107074	-15.417960	Spanish Restaurant
2032	Distrito 02 - Las Palmas de Gran Canaria	28.11232	-15.421078	Regaliz Funwear	28.105516	-15.417707	Men's Store
2033	Distrito 02 - Las Palmas de Gran Canaria	28.11232	-15.421078	Teatro Pérez Galdós	28.103382	-15.414024	Theater
2034	Distrito 02 - Las Palmas de Gran Canaria	28.11232	-15.421078	La Azotea De Benito	28.102523	-15.415288	Beer Garden

Repeating the API call for all districts in Gran Canaria, the following were captured from Foursquare: venue name, venue coordinates and venue category. The dataframe `venuesGC` stored all venues retrieved from Foursquare. Figure 4 shows an extract of the venues within District 2 of Las Palmas de Gran Canaria.

Based on the ranked type of venues frequently present, Figure5, districts can be grouped based on their similarities. k-means clustering was applied using the `KMeans` module from `scikit-learn` was used, where the number of clusters used, k, was 4.

Figure 5: Dataframe `districtsVenuesSorted` listing the most frequent venues retrieved from Foursquare for each district.

```
districtsVenuesSorted.head()
```

Out[41]:

	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	Distrito 01 - Agaete	Beach	Hotel	Restaurant	Scenic Lookout	Spanish Restaurant	Tapas Restaurant	Ice Cream Shop	Italian Restaurant	Surf Spot	Plaza
1	Distrito 01 - Agüimes	Beach	Hotel	Italian Restaurant	Bar	Café	Scenic Lookout	Spanish Restaurant	Restaurant	Tapas Restaurant	Ice Cream Shop
2	Distrito 01 - Artenara	Hotel	Beach	Spanish Restaurant	Scenic Lookout	Restaurant	Italian Restaurant	Resort	Tapas Restaurant	Ice Cream Shop	Surf Spot
3	Distrito 01 - Arucas	Restaurant	Tapas Restaurant	Spanish Restaurant	Beach	Scenic Lookout	Plaza	Shopping Mall	BBQ Joint	Italian Restaurant	Café
4	Distrito 01 - Firgas	Restaurant	Spanish Restaurant	Plaza	Scenic Lookout	Hotel	BBQ Joint	Beach	History Museum	Italian Restaurant	Tapas Restaurant

3.2 Part 2: key tourism data

In this section, the data we are aiming to get are basic demograhic data, average appraised home value for Las Palmas province, and expected high and low seasons for tourism.

3.2.1 Working with geodata

The goal is to be able to create thematic maps to represent key demographics of the island. In order to do so, it is first necessary to define the district boundaries. Using the resources available at ISTAC, the geodata delimiting the districts were loaded on to Python as `geodata` as shown in Figure 6. As highlighted in the figure, `geocode` is the property that will be used to load the geodata on the map in Figure 9.

Figure 6: `geodata` for Gran Canaria listing the district boundaries.

```
In [3]: # Let's display the geodata to understand the structure and labels used.
        geodata
```

```
Out[3]: {'type': 'FeatureCollection',
         'name': 'distritos_20170101',
         'crs': {'type': 'name',
          'properties': {'name': 'urn:ogc:def:crs:OGC:1.3:CRS84'}},
         'features': [{'type': 'Feature',
           'properties': {'geocode': '20170101_35001_D01',
            'etiqueta': 'Distrito 01 - Agaete',
            'granularidad': 'DISTRITOS',
            'gcd_isla': 'ES705',
            'gcd_municipio': '35001',
            'superficie': 4452.5952,
            'utm_x': 432237.32,
            'utm_y': 3105495.86,
            'longitud': -15.689645,
            'latitud': 28.073135},
           'geometry': {'type': 'MultiPolygon',
            'coordinates': [[[[-15.745333825523996, 28.050992249135742],
               [-15.745215454351655, 28.05101315861234],
               [-15.744777324653075, 28.051401864503973],
```

3.2.2 Basic demographic data

Using the demography data available at ISTAC (refer to Notebook for the link), the basic demograhic data was imported, transformed and stored as `demogDistrShort`. The original table contains a very comprehensive list of measurements, but the `demogDistrShort` dataframe summarises the population, mean age of inhabitants and the ratio of foreign to national population. The reason being that these demographic information will weigh in when choosing the area for property investment.

Figure 7: Summary of demographics for Gran Canaria broken down by districts.

```
In [5]: # 'demographyDistricts' contains many many columns, but we are only interested in seeing
        # the average age of inhabitants per municipality, and the ratio of foreign population.
        demogDistrShort = pd.DataFrame(columns = ['district', 'population', 'meanAge', 'ratioForeignPop
        demogDistrShort['district'] = demographyDistricts['geocode']
        demogDistrShort['population'] = demographyDistricts['poblacion']
        demogDistrShort['meanAge'] = demographyDistricts['poblacion_edad_media']
        demogDistrShort['ratioForeignPop'] = demographyDistricts['poblacion_extranjera_pc']
        demogDistrShort.head()
```

Out[5]:

	district	population	meanAge	ratioForeignPop
0	20170101_35001_D01	5526	44.9	4.7
1	20170101_35002_D01	30882	39.1	5.4
2	20170101_35005_D01	1096	50.4	0.5
3	20170101_35006_D01	7542	42.4	2.1
4	20170101_35006_D02	2922	45.4	1.2

3.2.3 Average home values and peak tourism season

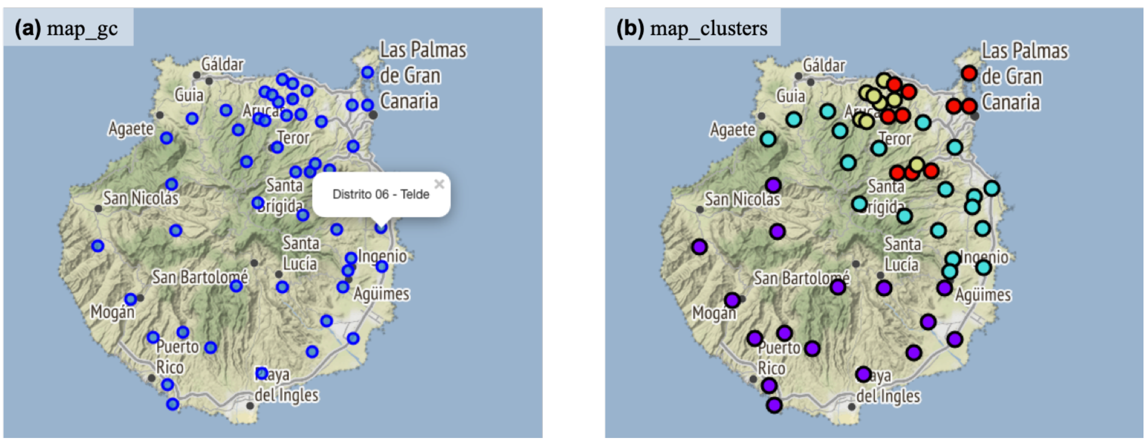
These two datasets were also obtained from ISTAC, but were manually prepared given their size and (language) translations requirements. These datasets come from:

- ISTAC - "*Valor tasado medio de vivienda libre según antigüedad de la vivienda de hasta 5 años. Provincias por comunidades autónomas y periodos*". Dataframe `df_homevalue`.

- ISTAC - "*Tarifa media diaria (ADR), ingresos por habitación disponible (RevPAR) e ingresos totales por municipios de alojamiento de Canarias y periodos*". Dataframe `tourismIncome`.

4 Results

Figure 8: Map of Gran Canaria showing (a) all districts labelled in blue, (b) results from clustering the distics based on their similarities.



Following from section 3.1, Figure 8(a) shows the location of all districts, where `Folium` was used for rendering the map of Gran Canaria. Figure 8(b) shows the results from clustering. As previously mentioned, the number of clusters used was 4. The table below summarises the most common venue categories identified after clustering the districts.

Table 2: Table summarising the main findings from clustering the districts based on venues similarities.

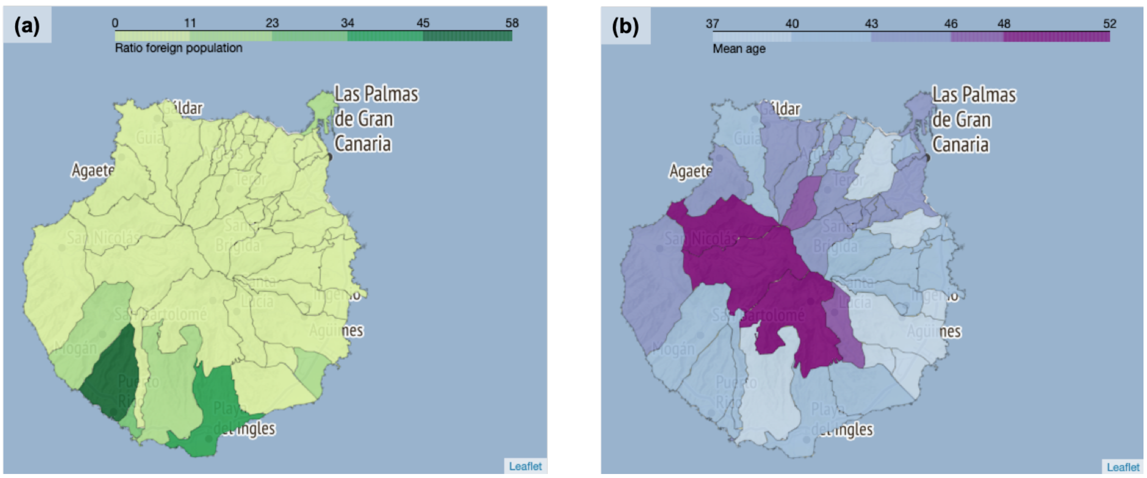
Cluster	Colour on map (Figure 8)	Top venues categories
First (index 0)	red	Restaurant, Spanish restaurant and beach
Second (index 1)	violet	Hotel, beach and Italian restaurant
Third (index 2)	blue	Beach, restaurant and coffee shop
Fourth (index 3)	green	Restaurant, Spanish restaurant and plaza

The accuracy of the data shown during clustering ultimately relies on the available venues listed on Foursquare. It will be a fair assumption that not all venues get to be featured on Foursquare.

Therefore, complementary data are required to enable decision-making.

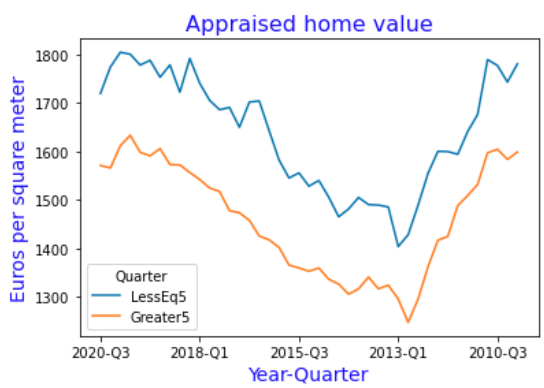
Based on the work on Part 2, Figure 9 summarises the demographic data for all districts. Figure 9(a) represents the ratio of foreign to national population. As expected, the sourthern coast of the island shows the highest proportion of foreign population. This is then followed by the district within Las Palmas that predominantly features the beach.

Figure 9: Choropleth maps showing the ratio of foreign population and inhabitants mean age for the districts in Gran Canaria.



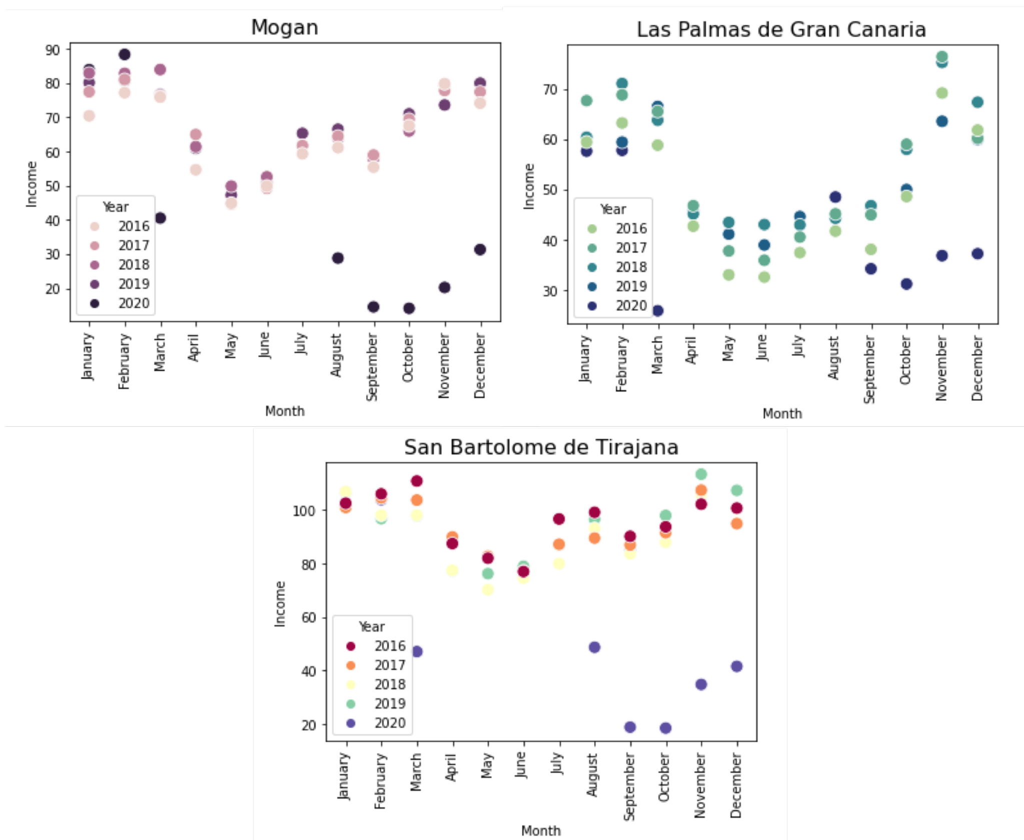
Shown in Figure 9(b), the mean age of districts' inhabitants are displayed. The rural parts towards the centre of the island show signs of ageing population. On coastal areas there are signs of a larger active population, perhaps due to tourism-related job opportunities.

Figure 10: Average appraised home value for the province of Las Palmas between years 2010 and 2020



An additional aspect that will be of interest to any investors looking for real estate opportunities, is the historical average of home values around the area. The values shown in Figure 10 are the average appraised home value, in euros per m^2, over the past two decades. A distinction is made based on the age of the house. Those that have been constructed less than 5 years ago, are valued $\sim$500 €/m^2 higher than those above 5 years of age.

Figure 11: Daily income per available guestrooms for municipalities Mogán, Las Palmas de Gran Canaria and San Bartolomé de Tirajana.



Lastly, knowing when the high and low tourism seasons are, as well as the potential revenue, will help in the decision-making process. Figure 11 shows the average daily income per available guestrooms for the most popular touristic destinations: Mogán, Las Palmas de Gran Canaria and San Bartolomé de Tirajana.

The plots cover the last 5 years. It is worth noting that the Canary Islands were hit hard due to the global pandemic in 2020. The metrics show the average value per available guestrooms. The magnitude of this number can give an indication of the occupancy rate too. Empty available rooms would not generate any income, thus bringing the overall average income.

Nonetheless, focusing on years 2016 $\sim$ 2029, it is clear that the high season for tourism comprises of November, December, January, February and March. Furthermore, the district of San Bartolomé de Tirajana consistently shows better performance over the other districts.

4.1 Discussion

Various datasets have been explored in order to compare and contrast the different districts in Gran Canaria.

As a first approach, the venues publicly listed on Foursquare were utilised in order to group the districts based on their similarities using k-means clustering as a technique. Although some improvements can be made in future studies, one aspect that is worth nothing is the second cluster. The second cluster covers most of the coastline towards the south of the island, which indeed corresponds to the highest density of hotels. The municipalities of San Bartolomé de Tirajana and Mogán are very popular touristic destinations, and there are many hotels, bungalows, apartments and villas in these regions.

This is further supported by Figure 11. For example, the average income per guestroom during low season in San Bartolomé de Tirajana exceeds the high season in Las Palmas de Gran Canaria. One would therefore expect a steadier revenue stream from a holiday accommodation in the southern coast of Gran Canaria.

In addition, as observed in Figure 9(a), the highest ratio of foreign population is also found in the southern part of the island. Where there is a higher resident population of foreigners, there are likely to be more services available to tourists, such as international restaurants, interpreters, health clinics offered in various languages, etc.

Las Palmas is still a good candidate. It has the advantages of being a big city, with beaches, a slightly higher than average foreign population index, and a large working age population.

Unfortunately, no real estate information at such local levels could be obtained for the current study. Comapring the average house values at the district level would have been ideal. Nevertheless Figure 10 still provides some insight. Older houses are cheaper than newer houses, and overall, house prices have been steadily rising since early 2013 until recently. As the market fluctuates, close monitoring of the real estate market will therefore be required in choosing the right time to invest.

4.2 Conclusion

The purpose of the current work was to assist potential buyers in making decisions on buying a holiday property in Gran Canaria. By using k-means clustering and data visualisation tools, many aspects were covered, ranging from demographics, real estate, tourism and venues offered by the different regions of the island.

5 Acknowledgements

- Open data from ISTAC.

- Open data from Foursquare.

References

[1] Gobierno de Canarias, http://www3.gobiernodecanarias.org/medusa/ecoblog/casilher/la-economia-en-espana/la-economia-canaria, last accessed on 12 January 2021.

[2] Cabildo de Gran Canaria, https://cabildo.grancanaria.com/en/los-municipios/, last accessed on 13 January 2021.