**Educational Migration:**

**Studying the Neighbourhood Amenities of Ljubljana (Slovenia) and Villach (Austria)**

Nic Whitehead  
  
22nd August 2021

**1. Introduction**

Tertiary education in Australia can be an expensive and sometimes cost-prohibitive exercise for many residents. Ex-pats (non-Australian permanent residents) living in Australia can be subject to international student rates which are even more highly priced. With no access to government assisted loans, many ex-pats and children of ex-pats are excluded from the quality tertiary education programs available in Australia.

So, what possibilities do these ex-pats have to educate either themselves or their children without expending vast sums of money or taking on massive personal debt?

Well, for those are lucky enough to have ties to the European Union, or can only afford lower-cost tuition fees, then Slovenian universities may be the answer. Slovenia provides free and affordable courses to both EU citizens and international students. Many of the courses are in English and can be purchased at a fraction of the cost in Australia.

Some parents may want to consider this as an option for their families where finances allow. Then at the same time, take the opportunity to live and work in Europe while their children study at an international university.

The Universities of Ljubljana and Maribor are the 2 top-ranked universities in Slovenia. Ljubljana is the capital of Slovenia and is the site of the main university campus. Villach is a sizeable community just across the Austrian border and is a short 90-minute train ride away.

Maribor is the largest city in Eastern Slovenia. It is the economic, administrative, educational, and cultural centre of the area. Graz is the second largest city in Austria (pop. 633,168 @ 2015) and is less than 90-minutes away from Maribor by train. (<https://en.wikipedia.org/wiki/Graz>)

Both pairs of cities offer many amenities, employment and lifestyle opportunities. However, which city would be best to live in while the children participate in further education?

**2. Business Problem**

The intended stakeholders for this analysis are Australians who are trying to decide whether to take their families on an educational migration to Europe.

The analysis will assist in choosing which areas in each city will provide the most relevant amenities and thus provide guidance on where to live for a period of time.

There are 4 cities to analyse, all of which provide a variety of lifestyle options. However, depending on where the children are studying (Ljubljana or Maribor), the challenge is deciding which city of each pair will be most appropriate to meet the needs of the family lifestyle.

For the purposes of this analysis the assumption is that a family will be looking for an Urban/Suburban lifestyle with certain amenities within a relatively close range of the chosen location.

Some of these amenities are:

1. Gym or fitness centre
2. Cafes
3. Restaurants
4. Supermarkets
5. Parks
6. Cinemas
7. Train Station
8. Theatre
9. Shopping facilities
10. Entertainment Venues

The report will highlight venues of interest in each city, within a 500m radius of each district.

Thus, illustrating the most suitable location to consider staying while a member/s of the family are studying in Slovenia.

**3. Data Description**

Geographical location data is required for Ljubljana, Villach, Maribor and Graz.  
Postal codes in each city serve as a starting point.

Postal codes will be used to find the neighbourhoods, districts, venues and their most popular venue categories.

**3.1 Postcode & GPS Data**

To derive the solution, postcode and GPS data is obtained from a variety of web sources and compiled as a CSV. This is due to scarcity and inconsistent formatting of the information for each location. Each CSV contains the following:

1. *district*: Name of the district within the city location
2. *post\_code:* Postal codes for each district
3. *latitude:* Latitude for district
4. *longitude:* Longitude for district

Ljubljana, Villach, Maribor and Graz have few distinct districts, so compiling this list manually is relatively straight forward.

The CSV was stored in a github repository, then imported into a pandas dataframe to be used as a starting point for the analysis.

**3.2 Foursquare API Data**

This analysis will require data about different venues in different districts of each location. In order to gain that information, "Foursquare" locational information will be used. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

After finding the list of districts, a connection to the Foursquare API is established to gather information about venues inside each district, for each city. For each district, we have chosen the radius to be 500 meters.

The data retrieved from Foursquare contains information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1. *Neighbourhood*: Name of the District
2. *Neighbourhood Latitude*: Latitude of the District
3. *Neighbourhood Longitude*: Longitude of the District
4. *Venue*: Name of the Venue
5. *Venue Latitude*: Latitude of Venue
6. *Venue Longitude*: Longitude of Venue
7. *Venue Category*: Category of Venue

Based on all the information collected for Ljubljana, Villach, Maribor and Graz, there will be sufficient data to build the model. Districts will be clustered together based on similar venue categories. Using this data, stakeholders can then compare city areas against the required venue categories and decide which location is most suited to their requirements.

**4. Methodology**

The model will be created with Python so the following packages will need to be imported to process the data.:

**import** **pandas** **as** **pd**

**import** **requests**

**import** **numpy** **as** **np**

**import** **matplotlib.cm** **as** **cm**

**import matplotlib.pyplot as plt**

**import** **matplotlib.colors** **as** **colors**

**import** **folium**

**from** **sklearn.cluster** **import** **Kmeans**

**import** **json**

**from** **geopy.geocoders** **import** **Nominatim**

**from** **pandas.io.json** **import** **json\_normalize**

Package breakdown:

* *Pandas*: To collect and manipulate data in CSV and conduct data analysis
* *requests*: Handle http requests
* *matplotlib*: Detailing the generated maps and creating plots for analysis
* *folium*: Generating maps of Ljubljana, Villach, Maribor & Graz
* *sklearn*: To import Kmeans which is the machine learning model that we are using.
* *Json*: library to handle JSON files
* *geopy.geocoders/Nominatim*: convert an address into latitude and longitude values
* *pandas.io.json/json\_normalize:* transform JSON file into pandas dataframe

The approach taken here is to explore each of the cities individually, plot the map to show the neighbourhoods being considered and then build the model by clustering all the similar neighbourhoods together.

Finally, a new map is plotted with the clustered neighbourhoods. Then insights and observations can be made from the resulting visuals.

**4.1 Data Collection**

In the data collection stage, we begin obtaining the required information for the cities of Ljubljana, Villach, Maribor and Graz. The data requires postal codes, districts and latitude/longitude specific to each of the cities.

Sourcing this information directly and cleanly from web locations was challenging. Most pages relating to each of these cities had partial data and were very inconsistent in their formatting. This made web-scraping difficult.

Fortunately, there are few districts in each area, compared to other larger European or American locations. Therefore, the decision was taken to research and manually collect the relevant data.

A CSV was created and loaded into github, then read into pandas to begin the location analysis.

Table

Description automatically generated

\* The table above contains the postcodes, districts and Lat/Long for all four cities.

**4.2 City by City Analysis**

At this point, analysis using the following processes can begin:

* Map visuals
* FourSquare
* K-Means Clustering with Silhouette analysis

For each city location, the same analysis process has been conducted. After all four locations have been completed, the locations can be reviewed as a group to assist stakeholders in determining which will be the most desirable location to live.

**4.2.1 Ljubljana Analysis**

Below is the location map of Ljubljana showing all the locations in the analysis for the city.

Details of the full analysis process also follow.

Map

Description automatically generated

The districts of Ljubljana are spread fairly evenly across the city and surrounding areas. They are also not too far from the centre of the city. It is likely that some of these areas will have many venues that are preferred by an expat family from Australia. However, others may be commercial or industrial districts with little in the way of lifestyle options.

A connection to FourSquare was established using a previously created API. It returns the top 100 venues within 500m of each District’s Longitude and Latitude. The resulting Json file is cleaned and made ready for additional analysis.

One Hot Encoding was used to sort and calculate the resulting dataframe. It grouped the districts and calculated the mean of the frequency of occurrence of each category (up to a maximum of 100 venues per district).

Graphical user interface

Description automatically generated with medium confidence

The above table is further sorted into the Top 10 venues per district.

The resulting view is a more easily readable and understandable data set for stakeholders to start making decisions about where the most desirable place would be to reside.

A screenshot of a computer

Description automatically generated with low confidence

**4.2.1.1 Ljubljana Analysis: Clustering of Venues**

To assist with the decision-making process for potential ex-pats moving to Ljubljana, the districts will be clustered using K-Means clustering. This will group the districts according to the main type of venues found nearby the Lat/Long for the centre of each district. Venues of similar type are clustered together in this un-supervised algorithm.

Once the clusters have been calculated, they are plotted on the city map for easier consumption.

However, K-Means requires a manual entry for the number of clusters. The question is how many clusters will be the optimal number of districts and venues in the sample?

In this case the silhouette method has been chosen to predict the optimal number of clusters for the dataset. The silhouette method allows for a rating of how well separated each cluster is from the other clusters. This is indicated by a value no greater than 1. The closer a value is to 1, the denser and well separated the cluster is from other clusters in the sample. For each city, a simple line plot has been chosen to visually indicate how many clusters will be optimal for the K-Means analysis.

**4.2.1.2 Ljubljana Silhouette Score Plot**

Chart, line chart

Description automatically generated

According to the results, the optimal number of clusters for Ljubljana is 3. It is worth noting that at 3 clusters, the silhouette score is only 0.36, which is not a particularly high score by itself. However, when viewed in the context of the other number of clusters, it is proportionally much higher than the majority of clusters tested.

Below is the map plot of Ljubljana with a 3-cluster plot:

**4.2.1.3 Ljubljana Clustered Districts (K-Means)**

Map

Description automatically generated

Rudnik

Vic

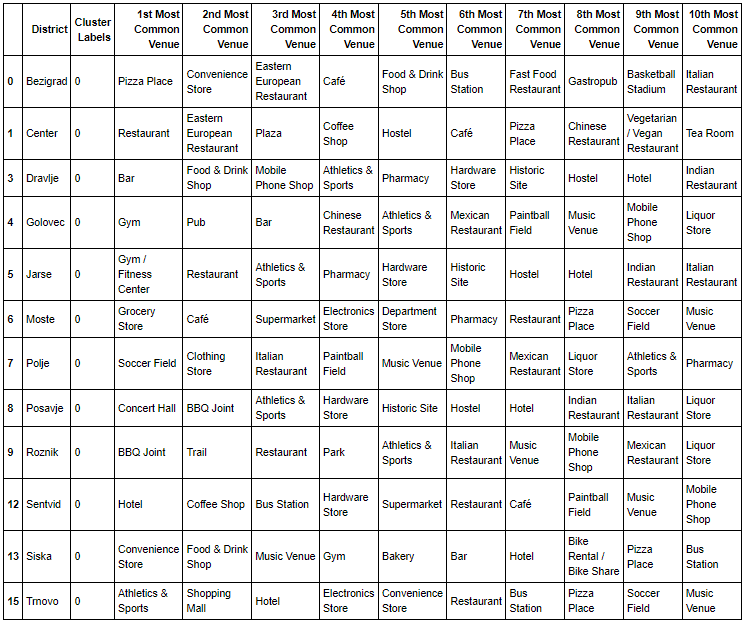
The K-Means algorithm has grouped the majority of districts into one cluster. With two other districts isolated in their own cluster each (Vic – purple/Rudnik – teal).

Once the clusters had been identified and sorted, dataframes were created for each group including their Top 10 venues of interest. These tables can be used by stakeholders to evaluate the suitability of an area for living in.

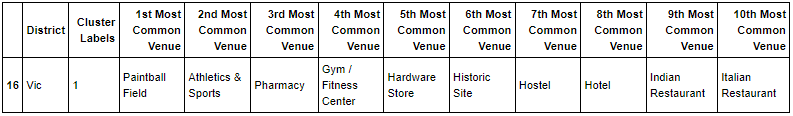
Each table below represents a colour/group of districts.

**4.2.1.4 Ljubljana Venues by Clustered Districts**

**Cluster 1 – Red Circles**



**Cluster 2 – Purple Circle (Vic)**



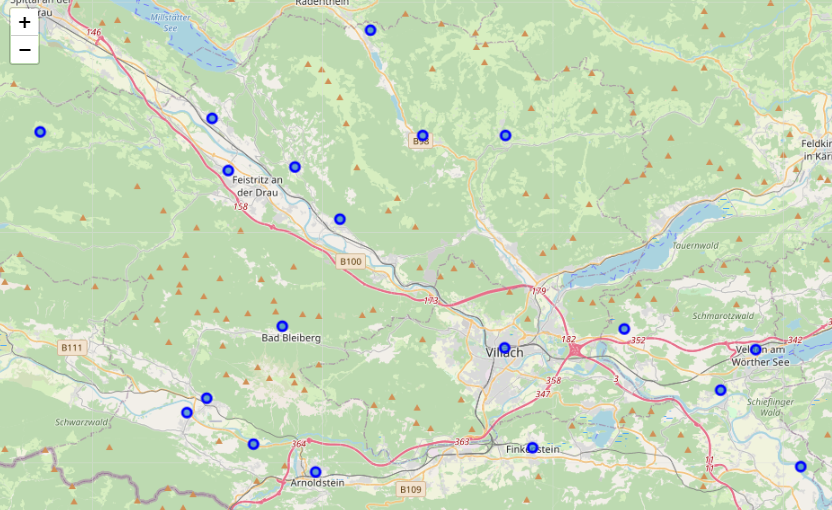
**Cluster 3 – Teal Circle (Rudnik)**



The above tables will be reviewed later in this report, against the suitability criteria listed in the Business Problem (section 2). All remaining cities analysed in this project are shown below.

**4.2.2 Villach Analysis**

Villach location map



Villach shows a significant spread of districts across the countryside. Depending on the nature of the venues within each district and cluster, this may not be the best option for a Suburban/Urban lifestyle.

Villach has 19 districts. The below table shows a sample of the First 5 districts with the Top 10 venues per district:

A picture containing text, wall, cabinet, several

Description automatically generated

**4.2.2.1 Villach Analysis: Clustering of Venues**

Venues of similar type are clustered together using K-Means. Once the clusters have been calculated, they are plotted on the city map for easier analysis.

Silhouette method has been used to predict the optimal number of clusters for the dataset.

**4.2.2.2 Villach Silhouette Score Plot**

Chart, line chart

Description automatically generated

According to the results, the optimal number of clusters for Villach is 4. It is worth noting that at 4 clusters, the silhouette score is only 0.226, which is not a particularly high score by itself. However, when viewed in the context of the other number of clusters, it is proportionally much higher than the majority of clusters tested.

Below is the map plot of Villach with 4 clusters:

**4.2.2.3 Villach Clustered Districts (K-Means)**

Map

Description automatically generated

Rosegg

Finkenstein am Faaker See

Ferndorf

The K-Means algorithm has grouped the majority of districts into one cluster. With three other districts isolated in their own cluster each (Ferndorf – purple/Finkenstein am Faaker See – yellow/Rosegg - teal).

Once the clusters had been identified and sorted, dataframes were created for each group including their Top 10 venues of interest. These tables can be used by stakeholders to evaluate the suitability of an area for living in.

Each table below represents a colour/group of districts.

**4.2.2.4 Villach Venues by Clustered Districts**

**Cluster 1 – Red Circles**

Graphical user interface

Description automatically generated with medium confidence

Once again, most venues are concentrated in the largest cluster. These venues will need to be reviewed against the key criteria of the stakeholders.

**Cluster 2 – Purple Circle (Ferndorf)**

Graphical user interface, diagram

Description automatically generated with medium confidence

**Cluster 3 – Teal Circle (Rosegg)**

Table

Description automatically generated

**Cluster 4 – Yellow Circle (Finkenstein am Faaker See)**

A picture containing table

Description automatically generated

The above tables will be reviewed later in this report, against the suitability criteria listed in the Business Problem (section 2).

**4.2.3 Maribor Analysis**

Maribor location map

Map

Description automatically generated

The districts of Maribor are not too far from the city centre. It is likely that some of these areas will have many venues that are preferred by an expat family from Australia. It could also be true that no matter where a family chooses to live, they will not be far from a wide assortment of venues.

Maribor has 11 districts. The below table shows a sample of the First 5 districts with the Top 10 venues per district:Table

Description automatically generated

**4.2.3.1 Maribor Analysis: Clustering of Venues**

Venues of similar type are clustered together using K-Means. Once the clusters have been calculated, they are plotted on the city map for easier analysis.

Silhouette method has been used to predict the optimal number of clusters for the dataset.

**4.2.3.2 Maribor Silhouette Score Plot**

Chart, line chart

Description automatically generated

According to the results, the optimal number of clusters for Maribor is 3. It is worth noting that at 3 clusters, the silhouette score is only 0.325, which is not a particularly high score by itself. However, when viewed in the context of the other number of clusters, it is proportionally much higher than the majority of clusters tested.

Below is the map plot of Maribor with 3 clusters:

Map

Description automatically generated

Brezje Dogose Zrkovci

Center

As in previous examples, the K-Means algorithm has grouped the majority of districts into one cluster. With two other districts isolated in their own cluster each (Brezje Dogose Zrkovci – purple/Center - teal).

Once the clusters had been identified and sorted, dataframes were created for each group including their Top 10 venues of interest. These tables can be used by stakeholders to evaluate the suitability of an area for living in.

Each table below represents a colour/group of districts.

**4.2.3.4 Maribor Venues by Clustered Districts**

**Cluster 1 – Red Circles**

Graphical user interface, text, application

Description automatically generated

**Cluster 2 – Purple Circle (Brezje Dogose Zrkovci)**

Table

Description automatically generated

**Cluster 3 – Teal Circle (Center)**

Table

Description automatically generated

The above tables will be reviewed later in this report, against the suitability criteria listed in the Business Problem (section 2).

**4.2.4 Graz Analysis**

Graz location map

Map

Description automatically generated

The districts of Graz are quite centralised around the city centre, with few districts spread into outlying areas. It is likely that these areas will have many venues that are preferred by an expat family from Australia. It could also be true that no matter where a family chooses to live, they will not be far from a wide assortment of venues.

Graz has 17 districts. The below table shows a sample of the First 5 districts with the Top 10 venues per district:

A screenshot of a computer

Description automatically generated with low confidence

**4.2.4.1 Graz Analysis: Clustering of Venues**

Venues of similar type are clustered together using K-Means. Once the clusters have been calculated, they are plotted on the city map for easier analysis.

Silhouette method has been used to predict the optimal number of clusters for the dataset.

**4.2.4.2 Graz Silhouette Score Plot**

Chart, line chart

Description automatically generated

According to the results, the optimal number of clusters for Graz is 2, with a silhouette score is only 0.425, which is not a particularly high score by itself. However, when viewed in the context of the other number of clusters, it is proportionally much higher than the majority of clusters tested.

What is interesting, is that this dataset appears to contain venues that have greater similarity, and the clusters are better separated than all the other cities viewed so far.

For the Graz clustering exercise, silhouette scores of 3 or 4 could be used as these are similar scores to the other cities tested.

Below is the map plot of Graz with 2 clusters:

Map

Description automatically generated

As in previous examples, the K-Means algorithm has grouped the majority of districts into one cluster. With one other district isolated in its own cluster (Ries – red).

This example clearly shows how tightly grouped these districts are within the highly developed areas of the city. Graz could be a good option for an urban lifestyle with easy access to a wide variety of attractions and venues.

Once the clusters had been identified and sorted, dataframes were created for each group including their Top 10 venues of interest. These tables can be used by stakeholders to evaluate the suitability of an area for living in.

Each table below represents a colour/group of districts.

**4.2.4.4 Graz Venues by Clustered Districts**

**Cluster 1 – Red Circle (Ries)**

Table

Description automatically generated with medium confidence

**Cluster 2 – Purple Circles**

A screenshot of a computer

Description automatically generated with medium confidence

A computer screen capture

Description automatically generated with low confidence

**5. Results of Slovenia & Austria City Analysis**

With the analysis of all four cities complete, it is time to make comparisons against the key criteria of stakeholders.

However, the analysis criteria suggests that the choice of city is primarily going to be based on which location is chosen for study. If the preferred university is in Ljubljana, then the choice will be between Ljubljana and Villach. However, if Maribor University is preferred, then the cities to compare will be Maribor and Graz.

The best way to view the results and determine which will be the best location, is to review one pair of cities at a time. Each cluster within each city will be compared against the primary venue types preferred by the stakeholders.

In each city, a primary cluster will be identified as the optimal location for ex-pats to reside in. Then this will be compared against the top contender for the corresponding city. Each cluster will be ranked by the number of key venues identified in the group.

The key venues of interest are:

1. Gym or fitness centre
2. Cafes
3. Restaurants
4. Supermarkets
5. Parks
6. Cinemas
7. Train Station
8. Theatre
9. Shopping facilities
10. Entertainment Venues

A visual analysis of each cluster was conducted using the high-level categories shown above. Note that the categories above are not specific venue categories found in FourSquare. Rather they are broad groupings of amenities that the stakeholders are seeking in their preferred area or district.

The results are listed below in each table.

**5.1.1 Ljubljana Clusters**

|  |  |  |
| --- | --- | --- |
| CLUSTER | TYPE/DISTRICT | MATCHING VENUES |
| 1 | Red Circles | 87 |
| 2 | Purple Circle (Vic) | 6 |
| 3 | Teal Circle (Rudnik) | 6 |

It is clear that **cluster 1** would be the preferred area to settle in from a lifestyle point of view. A new resident seeking an area to live in, who is also looking for the venue types listed above, should look within these districts of Ljubljana.

**5.1.2 Villach Clusters**

|  |  |  |
| --- | --- | --- |
| CLUSTER | TYPE/DISTRICT | MATCHING VENUES |
| 1 | Red Circles | 68 |
| 2 | Purple Circle (Ferndorf) | 6 |
| 3 | Teal Circle (Rosegg) | 6 |
| 4 | Teal Circle (Rudnik) | 8 |

It is clear that **cluster 1** would be the preferred area to settle in from a lifestyle point of view. A new resident seeking an area to live in, who is also looking for the venue types listed above, should look within these districts of Villach.

**5.1.3 Ljubljana Primary Cluster vs Villach Primary Cluster**

On the basis of the above data, it would appear that **cluster 1 in Ljubljana** offers significantly more options that match with the key criteria of our stakeholders.

Therefore, should the University of Ljubljana be chosen as the preferred university it would be better to reside within Ljubljana to gain access to a wider variety of the key amenities.

**5.2.1 Maribor Clusters**

|  |  |  |
| --- | --- | --- |
| CLUSTER | TYPE/DISTRICT | MATCHING VENUES |
| 1 | Red Circles | 62 |
| 2 | Purple Circle (Brezje Dogose Zrkovci) | 9 |
| 3 | Teal Circle (Center) | 8 |

It is clear that **cluster 1** would be the preferred area to settle in from a lifestyle point of view. A new resident seeking an area to live in, who is also looking for the venue types listed above, should look within these districts of Maribor.

**5.2.2 Graz Clusters**

|  |  |  |
| --- | --- | --- |
| CLUSTER | TYPE/DISTRICT | MATCHING VENUES |
| 1 | Red Circle (Ries) | 9 |
| 2 | Purple Circles | 136 |

It is clear that **cluster 2** would be the preferred area to settle in from a lifestyle point of view. A new resident seeking an area to live in, who is also looking for the venue types listed above, should look within these districts of Graz.

**5.2.3 Maribor Primary Cluster vs Graz Primary Cluster**

On the basis of the above data, it would appear that **cluster 2 in Graz** offers significantly more options that match with the key criteria of our stakeholders.

Therefore, should the University of Maribor be chosen as the preferred university it would be better to reside within Graz to gain access to a wider variety of the key amenities.

**6. Discussion**

The preceding analysis is a fairly simplistic method for selecting a location to settle in. While I appreciate that there are many other factors to choosing the best possible district in these four cities, it does provide some guidance to those who may want to engage in a migration for educational reasons.

Without a method of targeting the districts, moving to these cities would be a daunting prospect for Australian residents with little experience of Slovenia or Austria. That is where this analysis can be useful to direct the focus on a reduced number of locations that provide the lifestyle and local amenities that the stakeholders prefer.

It was interesting to note that the primary K-Means groupings were quite large, with a few outliers in each example. What this implies is that the most common venues were largely restricted to centralised areas. Which is what I did expect to some degree.

This does prove that the best place to access most of the required amenities is closer to the centre of each city.

Interestingly, the best location for families who have students in Ljubljana was to stay in Ljubljana. While the optimal location for those that have students in Maribor is Graz, Austria. Also noteworthy is that Graz provides the widest range out of all the cities of appropriate amenities for our stakeholders. This is likely due to Graz being the city with the largest population by a significant margin.

In future, this analysis would benefit from some additional refinements.

First, it would be useful to overlay the average price of rental/sale properties in each district as well. This would further target the districts for stakeholders. They would then be able to choose specific districts that are within their price range and that are nearby their preferred venue types.

Second, refining the FourSquare venues down to the types of amenities listed in the stakeholder’s primary list would also assist in providing a better result. These venue types could be developed into a table that groups the venue types from FourSquare into the broad preferred categories. A table like this can be merged with the venue data and used to limit the results specifically to stakeholder-required amenities. The resulting table could be used in the K-Means dataframe instead and may provide quite different clustering results.

**7. Conclusion**

With the ever-increasing cost of tertiary education in Australia, families seeking to find other means of educating either children (or themselves), may find this kind of analysis useful to assist in their relocation.

So, for Australians able to take advantage of cost-effective education in Slovenia, then choosing to live in either Ljubljana or Graz would appear to provide the best lifestyle options.

Although there are many other factors that may influence exactly where they would want to live, this kind of data can provide a good platform for reducing any potential for error in choosing an optimal location to stay for a considerable period of time.