Iceland Day-in-Life Excursion

Sally Chan

May-5-2021

1. Introduction
   1. Background

LoadsOfFun is a travel company aim to deliver unforgettable short excursions to their customers. Their customers are usually people travel for business, they are in town for a short time either working or attending a conference, who usually have a day or half a day of free time to explore the country they are visiting, they want to make the most out of their limited free time.

* 1. Problem

The people travel for work usually have very busy schedules, they do not have time to do research or planning. They will not join a regular tour with fixed schedule. They want to see the major sights but after taking a memorable picture, they do not want to spend a full day there. They want to experience what locals enjoy and emerge in their culture briefly in their leisure time hoping to take back with them a memorable day-in-life of being a local.

* 1. Interest

LoadsOfFun decided to fill the gap by offering short excursions that delivers local cultural experience. The product development team will form a small team to work on a pilot tour package for customer trial. This team is composed of young IT professionals equipped with data science skills who are enthusiastic travelers themselves. The place they decided to work on is Iceland, it is a country filled with adventurous sites such as volcano, hot spring, and rugged mountains with extraordinary nature scenery.

* 1. Methodology

The key tools for the project are Jyputer Notebook and Python. Jyputer Notebook is a web-based application for software programming using a browser. Python is a programming language; it has a rich library for data analytics. The team will be using Python library such as geopy to locate the latitude and longitude of an address; folium for mapping and geospatial analysis and visualization; matplotlib for plotting; and scikit-learn for machine learning to understand data similarity.

The steps to do the project are outlined in Figure 1. Data is gathered for the project, it is cleaned and formatted to the stage that can be processed by Python. Data integration and transformation will take place if we need more information for analysis or build data models.

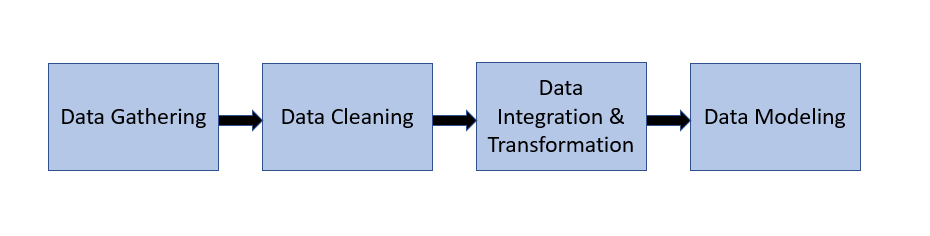


Figure 1 shows major steps in processing data for analysis

In each steps of data handling we gain a deeper understanding of the data used for the problem. Data modeling helps visualization of the data in addition to textual format.

This is not a waterfall approach, for example retrieved data will always be cleaned and visualized for problem solving. If more data is needed at a later stage, new data will be added and it will go through the cleaning and visualization before integrated into the final data set for analysis.

2. Data acquisition and cleaning

2.1 Data sources

The team has not visited Iceland so they rely on the internet to find all the information they need for the developing the excursion project. They browse the web for postal code with the understanding that from the postal code they will be able to find information on cities and towns.

The team decided to use the Iceland postal code link from Wikipedia <https://en.m.wikipedia.org/wiki/List_of_postal_codes_in_Iceland> , after reviewing the website they believed it contains trusted data. The postal code is divided into regions, It is contained in a table showing postal code, area using the postal code, and post office address. Figure 2 shows a snap shot of the postal code data from the web link.



Figure 2. Iceland postal code in the Capital Region

2.2 Data cleaning

The data from the web was bring into Jyputer Notebook using Python library Beautiful Soup to read in the html table. It is put into Pandas dataframe. Figure 3 shows the dataframe containing the postal code data from the web link.

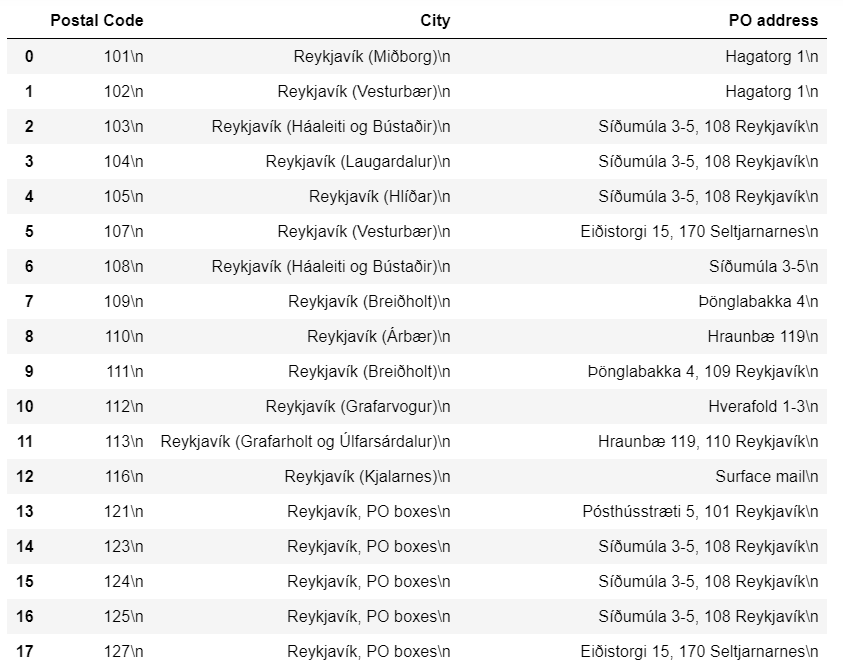


Figure 3. Postal code data in pandas dataframe before cleaning

As we can see, the data has unwanted characters and format. The team needs to clean the data to a stage that they can use for analysis. They need the city name and the neighborhood names only.

The City column contains the city and district in (), sometimes district also means PO Box. The team selected index 0 to 12, taking only the lines with unique district names and removing duplicate lines.

The City column is then separated into two columns City and Neighborhood. The City column is for the city names and the Neighborhood column is for the district names. The characters ‘(‘ and ‘)’ are removed.

The post code and the PO address column is not useful for analysis, so they are deleted. Figure 4 shows the new dataframe after cleaning.



Figure 4 shows the dataframe after data cleaning

To know more about the city and its neighborhoods the team wants to see a map. They use Python geopy to get the latitude and longitude, this data is necessary for folium to create the map for visualization. Figure 5 shows the latitude and longitude of Reykjavik city and its neighborhoods.

The new data, latitude and longitude, give an idea to the team where the neighborhoods are located. They also noticed that Hlíðar’s longitude is out of normal range and may be an issue.

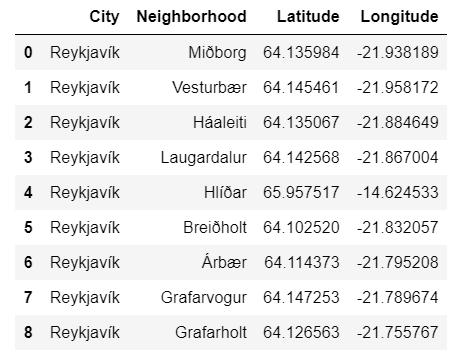


Figure 5. Shows latitude and longitude for every neighborhood in the city of Reykjavik

Figure 6 shows a map of Reykjavik and it’s neighborhood. The map only shows eight dots, this reassures the team that Hlíðar is out of range and it will be a good idea to omit it for their project. The team regenerated the dataset. Figure 7 shows the final dataset.

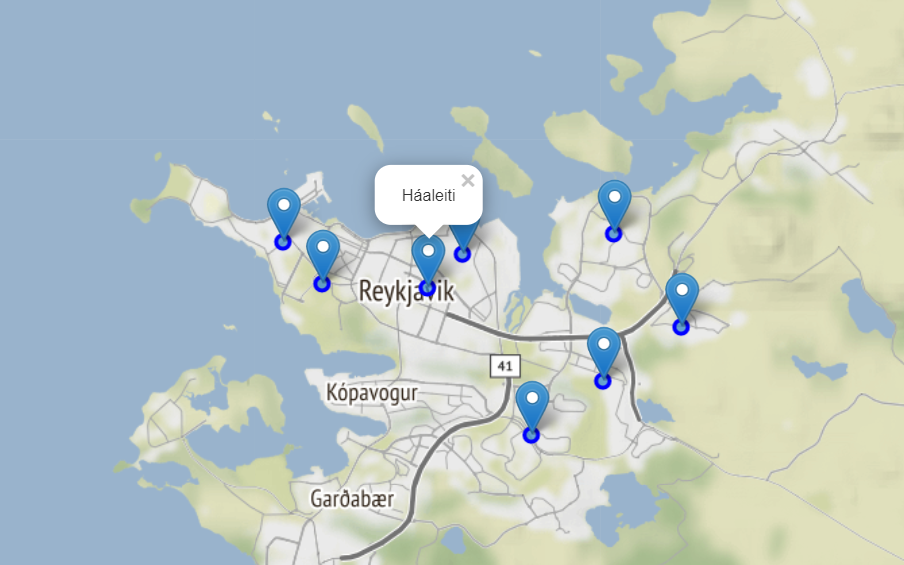


Figure 6 shows a map of Reykjavik and it’s neighborhood



Figure 7 shows the final dataset

2.2 Adding New Data

The map helps the team to visualize where the neighborhoods are location, which two neighborhoods are in close in proximity. To build an excursion package the development team needs to know in-depth what each neighborhood has to offer to the visitor. The best place for this information is using Foursquare API to get data of popular places people visited and recommanded. Figure 8 shows the Json file from Foursquare and the data in it.

  
Figure 8 shows the Json file from Foursquare and the data in it

The team extracts the data out of the json file and stores it in a dataframe for analysis similar to reading in the html file for postal code. Figure 9 shows the datafame containing venues for neighborhood Miðborg.

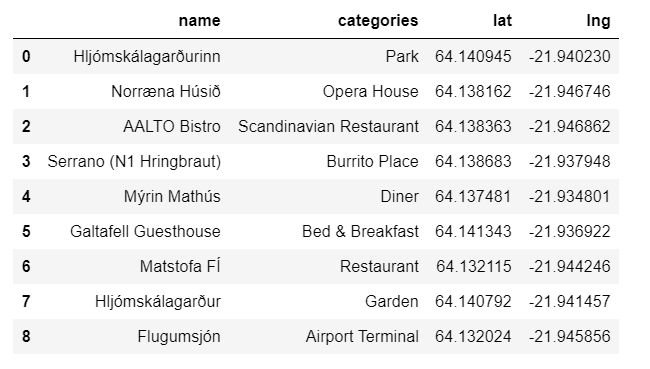


Figure 9 shows the datafame containing popular places for Miðborg

2.3 Explore Data in other neighborhoods

This is useful data for building excursion so the team decided to get similar information for all the neighborhoods in Reykjavik. Figure 10 shows the popular venues in Reykjavik neighborhoods.



Figure 10 shows the popular venues in Reykjavik neighborhoods

The eight (8) neighborhood has a total of fifty-seven (57) venues, and twenty-nine (29) unique venue categories. The Venue Category column in Figure 11 shows Háaleiti and Vesturbær has the most variety venues to attracts visitors.

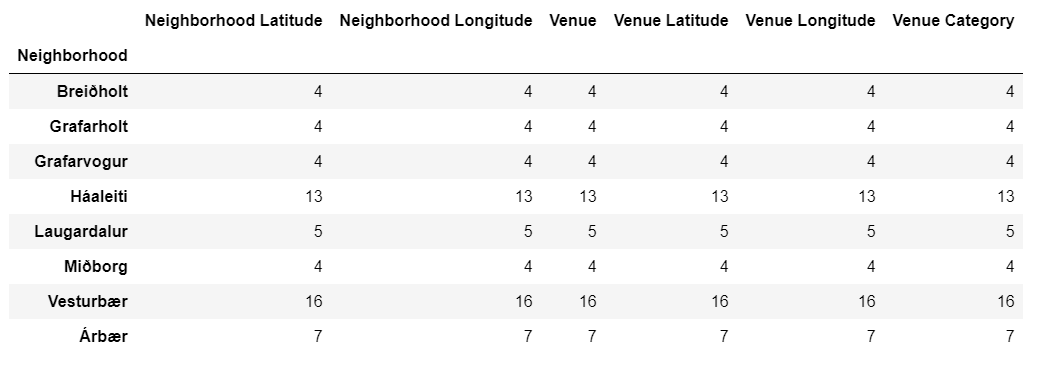


Figure 11 Háaleiti and Vesturbær has the most variety venues

3. Data Model

3.1 Data Discovery

To visualize the data, the team decided to build a machine leaning model to show cluster of data. To do that they use one hot encoding to represent each venues in the 9 neighborhood using 0 or 1. 0 means not available and 1 means it exists. Figure 12 shows neighborhood venues represented in one hot encoding. The venues are then grouped by neighborhood with the frequency of occurrence shown in each category in Figure 13.

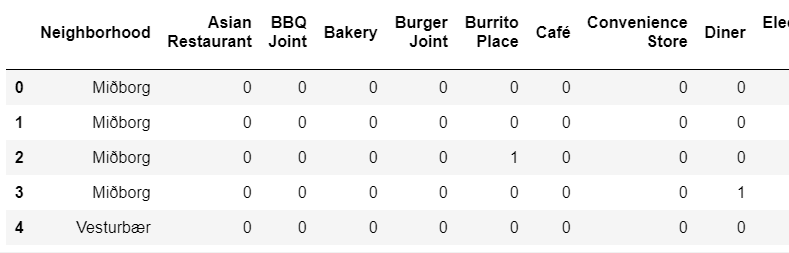


Figure 12. shows neighborhood venues represented in one hot encoding.

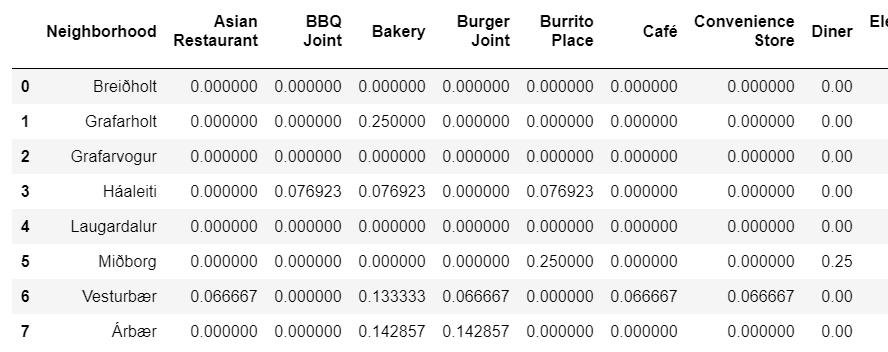


Figure 13. shows venues frequency grouped by neighborhood

Figure 14 shows a list of venues and frequency in Háaleiti and Vesturbær in a different format using the same data. This is good information for the development team to design the excursion trip based on one neighborhood.



Figure 14 shows a list of venues and frequency in Háaleiti and Vesturbær

By manipulating the data, one can visualize the top 10 most popular venues across the 9 neighborhoods in Figure 15. This data can make suggestion if one wants to stay in their neighborhood, what activities they can do there.



Figure 15 to shows the 10 most common venue in each neighborhood

The team uses K-means cluster data on the Reykjavik map to understand the similarities and difference venues offered in each neighborhood. Figure 16 shows the neighborhood similarities if we set K-mean cluster to 4. Figure 17 shows the neighborhood similarities if we set K-mean cluster to 2.

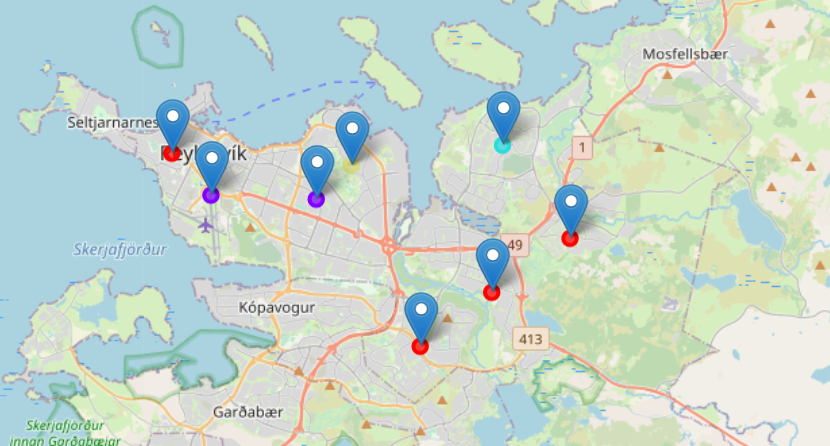


Figure 16 shows the neighborhood similarities if we set K-mean cluster to 4

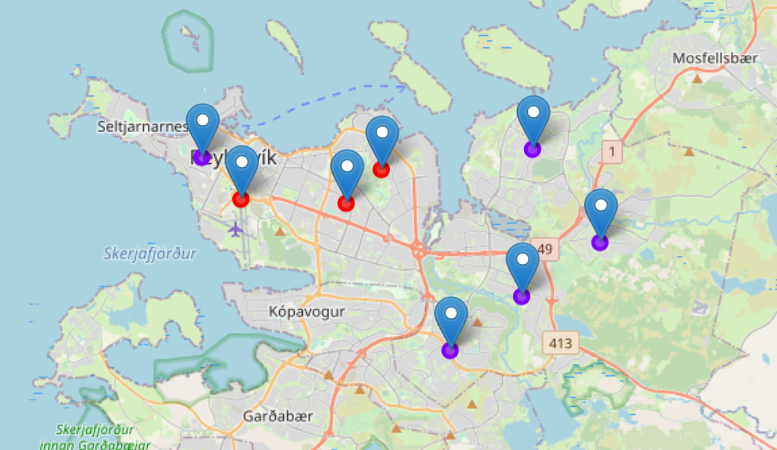


Figure 17 shows the neighborhood similarities if we set K-mean cluster to 2.

By changing the number of clusters in K-means, we learn the following:

When K is set to 2, two clusters are created from the data. The venues in the red neighborhoods, Miðborg, Háaleiti, and Laugardalur has similar venues and the purple neighborhoods, Vesturbær, Breiðholt, Árbær, Grafarvogur, and Grafarholt has similar venues. If compare the red and purple neighborhoods, they have different venues. The 2 cluster is a good indication that red neighborhoods are the tourist spot and purple neighborhood are residential area. This is further supported by the data such as opera house, hotels, and gardens are found in the red neighborhood.

When K is set to 4, we learned that Grafarvogur and Laugardalur may have something unique to offer while the Vesturbær, Árbær, Breiðholt, and Grafarholt remains in the same cluster as well as Miðborg and Háaleiti. It may be worth a trip for visitor to visit Grafarvogur and Laugardalur.

4. Development of excursion product

The excursion development team at LoadsOfFun had enough information to develop several excursion packages inviting anyone traveling to Iceland to try out

Excursions packages

|  |  |
| --- | --- |
| Modborg Half Day Trip | 1. Get a quick bite at AALTO Bistro in the morning 2. Take the bus to explore Modborg's surroundings 3. Have dinner at Myrin Mathus 4. Go to Opera House |
| Modborg Full Day Trip | 1. Stay at Galtafell Guesthouse 2. Eat at Serrano 3. Visit Hljómskálagarðurinn 4. Dinner at Matstofa FÍ 5. Go to Opera House 6. Go to airport |
| Háaleiti and Laugardalur Half Day Trip | 1. Get bakery at Mosfellsbakarí in the morning 2. Take a walk at the Laugardalurinn park 3. Lunch at Dirty Burger & Ribs 4. Have some icecream at Ísbúð Háaleitis |
| Háaleiti and Laugardalur Full Day Trip | 1. Get bakery at Mosfellsbakarí in the morning 2. Take a walk at the Grasagarðurinn garden 3. Go to Café Flóran for lunch 4. Try out skating at Skautahöllin Laugardal 5. Have some icecream at Ísbúð Háaleitis 6. Dinner at Mulakaffi |
| Breiðholt half day Trip | 1. Stay in vacation rental 2. Visit grocery store 3. Go to Indian Restaurant |

5. Conclusion

Data is crucial in understand everything around us. Using the right data and analytic tools can turn seemingly unimportant fact to useful information. In this project we use a simple postal code and latitude and longitude data to developed excursion package that delivers life time memorable experience for travelers are priceless. If the pilot trips are successful, one can build a business out of this concept and the possibilities to expand to other countries and cities is limitless.