Assignment: Applied Data Science Capstone project

Marc Jellema

December 17, 2020

Abstract

This report is a deliverable for my capstone project *Battle of the neighbourhoods*, part of my graduation for my **IBM Data Science Professional certificate**. First leg of the capstone project is the introduction and data part of this report that you can find below. The other sections will be filled in the second and final leg of the capstone project. Feel free to leave comments or send me an email: mjellema@omnia.nl

Introduction

I tasked myself with the assignment to find additional ways to enhance automatic estimation of house prices in the Netherlands. The use of the **FourSquare** API is mandatory for this assignment, part of the IBM Data Science professional certificate [1]. As a tech savvy person myself, I'd heard of FourSquare and I do remember seeing the FourSquare logo on venues in the past, but never came across its footprint in the Netherlands. Wondering if I could use the FourSquare API in finding an answer for a challenge in the Netherlands, I did some research and as it turns out, the FourSquare database is more than comprehensive enough to use for data science challenges in the Netherlands as well. Not knowing a single person actively using FourSquare to let their friends know where they check-in, it puzzled me how FourSquare actually accumulates their data. James D. Walsh wrote a nice article [2] on how FourSquare accomplishes this. Enough about FourSquare for now, let's focus on introducing the question I want to answer with help of the FourSquare API and the data it can provide:

> *Based on **FourSquare** data, can we find a correlation between the average house price, the number of restaurants and the number of inhabitants in the provinces of the Netherlands?*

Although the COVID pandemie is having an impact on the overall economy world wide, it seems house prices have not yet taken a hit due to the pandemic. Startups like **Promodomo** [3] use algorithms to calculate house prices, even when they are not on the market yet. Algorithms like Promodomo's always benefit from additional sources to enrich their calculation. Searching for a correlation based on FourSquare data is just one off the many possibilities to augment the estimation of house prices and this is why I choose to take this challenge and come up with answers in this capstone project of my IBM Data Science Professional certificate.

1 Data

To answer the challenge I need the following data points:

1. Number of inhabitants per province in the Netherlands
2. Average house price per province in the Netherlands
3. Number of restaurants per province in the Netherlands

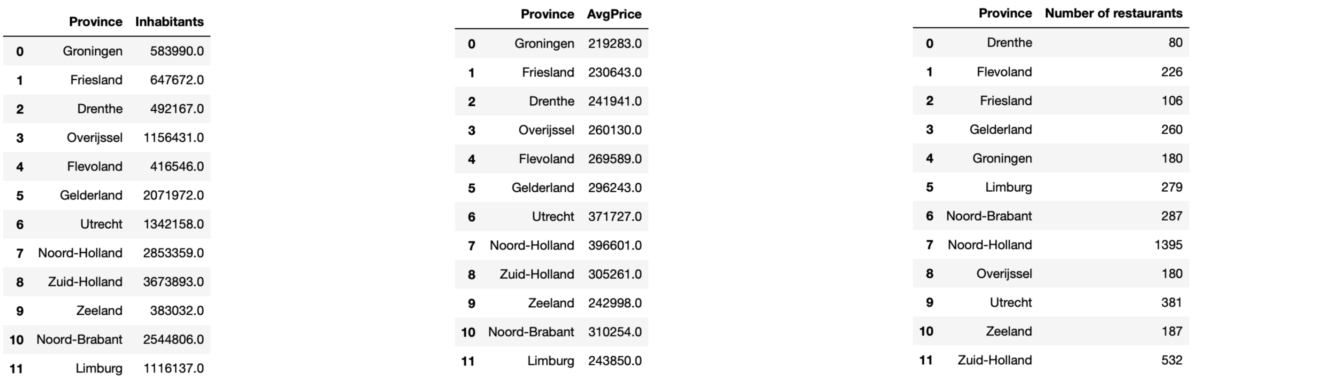
Ad 1. The number of inhabitants per province in the Netherlands can be downloaded from the Centraal Bureau voor Statistiek (**CBS**): a government owned, freely accessible web site with tons of statistical datasets (over 4K of them!) to use in all kind of analysis. I did some research on which table would best provide my need for the number of inhabitants per province and its title turns out to be `[Regionale kerncijfers Nederland]`. The resulting dataframe, including some data cleaning as removing substring `[(PV)]` from the province name.

Ad 2. The average house price per province in the Netherlands in 2019 also can be downloaded from the CBS, this time based on the table with title `[Bestaande koopwoningen; gemiddelde verkoopprijzen, regio]`. After a bit of cleaning and tweaking (i.e. 'Friesland' ≠ 'Fryslân' and some column renaming had to be adjusted to make the dataset comprehensible for our international readers) the two dataframes could be joined.

Ad 3. The third dataset, the number of restaurants, came from FourSquare. I needed to find all restaurants within each province of the Netherlands. The FourSquare API has the following two drawbacks that I needed to overcome: the non commercial API limits the returned venues per call to max 100 and although the results have a `[state]` field, the API doesn't allow searching per state. The limit of max 100 venues per call I overcame with the help of the material I found of a fellow course student Guillermo (G.) Bareirro [4]. FourSquare returns max 100 venues per call, but if you make the call specific enough, the total results will not grow above the limit. Using the categories listed bij G. Bareirro, I was able to split querying all restaurants into their separate categories and thereafter grouped and summed them with pandas standard dataframe functionality.

To overcome the second drawback of not being able to query FourSquare by state, I tried to find a geo boundaries source online of all the provinces in the Netherlands. Turned out there is no such source readably available in the public domain. Knowing that an estimate of the restaurants per province would be sufficient for my analyses, I queried the restaurants (via their respective subcategories) in the capital city of each province.

The three data frames used in the analyses for this assignment are in the images below:



	Province	Inhabitants
0	Groningen	583990.0
1	Friesland	647672.0
2	Drenthe	492167.0
3	Overijssel	1156431.0
4	Flevoland	416546.0
5	Gelderland	2071972.0
6	Utrecht	1342158.0
7	Noord-Holland	2853359.0
8	Zuid-Holland	3673893.0
9	Zeeland	383032.0
10	Noord-Brabant	2544806.0
11	Limburg	1116137.0

(a) Inhabitants from StatLine [5]

	Province	AvgPrice
0	Groningen	219283.0
1	Friesland	230643.0
2	Drenthe	241941.0
3	Overijssel	260130.0
4	Flevoland	269589.0
5	Gelderland	296243.0
6	Utrecht	371727.0
7	Noord-Holland	396601.0
8	Zuid-Holland	305261.0
9	Zeeland	242998.0
10	Noord-Brabant	310254.0
11	Limburg	243850.0

(b) Average Price from StatLine [6]

	Province	Number of restaurants
0	Drenthe	80
1	Flevoland	226
2	Friesland	106
3	Gelderland	260
4	Groningen	180
5	Limburg	279
6	Noord-Brabant	287
7	Noord-Holland	1395
8	Overijssel	180
9	Utrecht	381
10	Zeeland	187
11	Zuid-Holland	532

(c) Restaurants from FourSquare [7]

Figure 1: Cleaned data frames as used in analysis

2 Methodology

3 Analysis

4 Results

5 Discussion

6 Conclusion

7 Acknowledgements

8 Appendices

References

[1] Coursera, *IBM Data Science Professional Certificate*. More info.

[2] J. D. Walsh, *Ten Years On, Foursquare Is Now Checking In to You*. Article link.

[3] Promodomo, *Promodomo website*. Visit website.

[4] G. Bareirro, *Best Cuisines*. Github repo link.

[5] CBS, *Number of inhabitants in 2019 in each province of the Netherlands*. Source on StatLine.

[6] CBS, *Average house price in 2019 in each province of the Netherlands*. Source on StatLine.

[7] FourSquare, *FourSquare developers API - explore endpoint*. Docs on FourSquare.