Capstone project - Study of the correlation between obesity and restaurant density in England

Fabio Vaianella

February 2019

1. Introduction: Business Problem

Obesity is the second cause of premature death in Europe, preceded only by smoking. A research has demonstrated that 1 in 7 premature deaths could be avoided if people had a healthy weight. What if the risk of obesity was linked to the density of restaurants in the surrounding of the person at risk? What if there were correlations between the number of temptations at which a person is subjected to and the risk that is person will suffer from obesity?

The following project aims to provide answer to the following question:

"Is there any link between obesity and the density of restaurants in the surrounding of where a person lives in England?"

The results of this study could be useful for the Department of Health and Social Care of United Kingdom but could be expanded to any other region that suffer from high mortality due to obesity such as the rest of Europe or the United States.

2. Data

Based on our problem, we needed as data:

- We first needed obesity statistics in each of the neighborhood. These data are available on Kaggle but require to create an account. We will thus locally download these data and upload them for the project. These data consist of a number of admission as obesity case in each neighborhood in 2015/2016. The dataset contains information for each neighborhood of England, but we will focus only on the Greater London for this project. Each row of the dataset (after some cleaning) consists of various attributes, but we will only use some of them, namely the name of the neighborhood and the number of admissions per 100 000 population for male, female and both. (Kagglle)
- We needed also the locations of the neighborhood of the Greater London. We will use the package Geopy to obtain the geographical coordinates of each neighborhood. We will look for the coordinates by searching them with Geopy.geocoder using the names of the neighborhood obtained in the previous step. (Geopy)
- We finally needed Foursquare data to study the surrounding of the center of each neighborhood and evaluate their density of restaurants. We will search the venues for each neighborhood in a certain radius and then filter the results by searching the venues that contain the word "restaurant". (Foursquare)

We started by downloading locally the data for obesity. The file contains a 14 tables. First thing to do was to select the correct table and read its rows. After removing useless rows, we had a dataset with 152 rows (152 districts) as seen in Figure 1.

	index	E12000001	Unnamed: 1	A	North East	20188	7039	13148	Unnamed: 7	781	565	991
0	0	E06000047	NaN	116	County Durham	3373.0	1283.0	2089.0	NaN	647.0	502.0	787.0
1	1	E06000005	NaN	117	Darlington	649.0	274.0	375.0	NaN	623.0	541.0	698.0
2	2	E08000037	NaN	106	Gateshead	1428.0	405.0	1023.0	NaN	716.0	427.0	1003.0
3	3	E06000001	NaN	111	Hartlepool	554.0	208.0	346.0	NaN	616.0	473.0	751.0
4	4	E06000002	NaN	112	Middlesbrough	829.0	298.0	531.0	NaN	657.0	496.0	812.0

Figure 1: dataset after reading the file

First column of the dataset is just used for indexing the rows, second column is the ONS code, third column is an empty column, fourth column is LA code, fifth column is the name of the district, sixth column the total number of admission in the district, seventh column the number of male admission, eighth column

the number female admission, ninth column an empty column, and the remaining columns are the admission per 100,000 population. We then removed the unnecessary column and rewrite the column names to make sense as in Figure 2.

	Regions	All	Male	Female
0	County Durham	647.0	502.0	787.0
1	Darlington	623.0	541.0	698.0
2	Gateshead	716.0	427.0	1003.0
3	Hartlepool	616.0	473.0	751.0
4	Middlesbrough	657.0	496.0	812.0

Figure 2:Obesity dataset after cleaning

Being on possession of the names of the districts, we then used Geopy to get the location of every district. Note that geopy did not return all districts coordinates, so we lost 14 district in this process on 152, which is quite satisfying.

With the locations data, we then explored the venues on each district and then select all the venues related to food (restaurant, pub, lounge, food, tavern, diner, steakhouse). In this way, we had our final cleaned dataset with obesity statistics and restaurant densities (Figure 3)

	District	All	Male	Female	Latitude	Longitude	Restaurants
0	County Durham	647.0	502.0	787.0	53.872616	-1.705963	2
1	Darlington	623.0	541.0	698.0	51.536151	-0.134964	4
2	Gateshead	716.0	427.0	1003.0	54.958554	-1.605700	2
3	Middlesbrough	657.0	496.0	812.0	54.576042	-1.234405	4
4	Newcastle upon Tyne	737.0	451.0	1038.0	54.973847	-1.613157	18

Figure 3: Final dataset

3. Methodology

We firsted start our analysis by visualization of the data by looking at the choropleth map of the obesity admission and compared it to the choropeth map of the density of restaurants. The GeoJson file for the choropleth map of England was found on a dedicated Github. We then extended our analysis to quantitative analysis by searching correlation between the density of restaurants and the obesity admissions, for male, female and both. An linear regression model was developed in this sense.

4.Results and discussion

We started by comparing the choropleth maps of obesity admissions and restaurant density for both male and female, Figure 4.

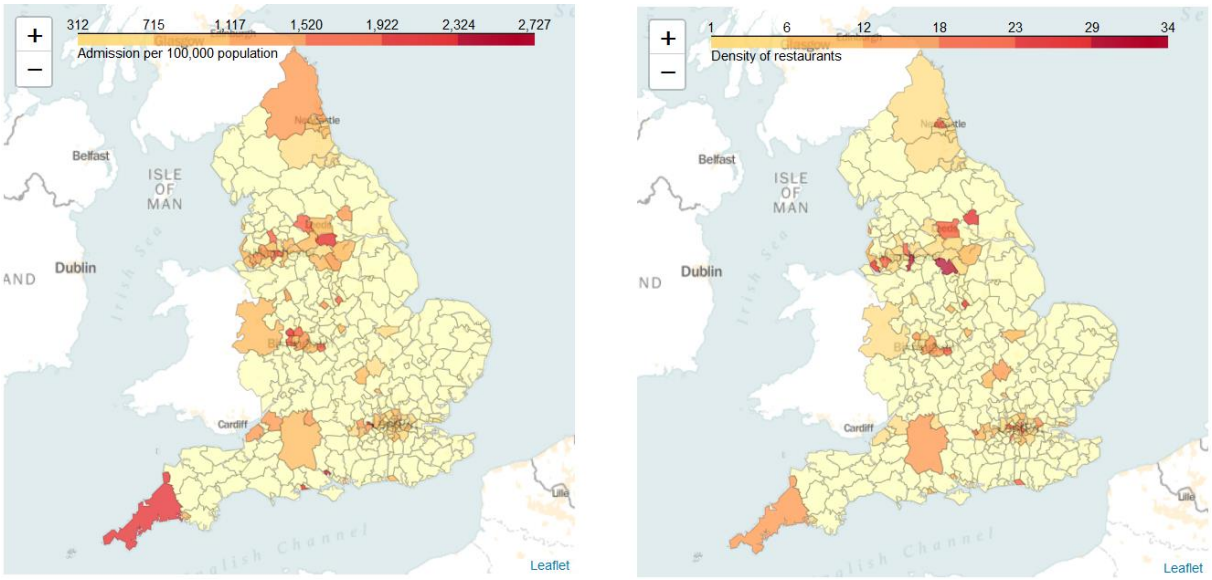

Figure 4: Choropleth map of the obesity admission per 100,000 population on the left and density of restaurants in each district on the right, in England

We can see from the two maps that they look similar. However, when looking closely we can see that the districts with larger obesity admissions are not the

ones with larger density of restaurants at all and our intuition could be wrong. Let's look at the scatter plot to have better insights of this (Figure 5).

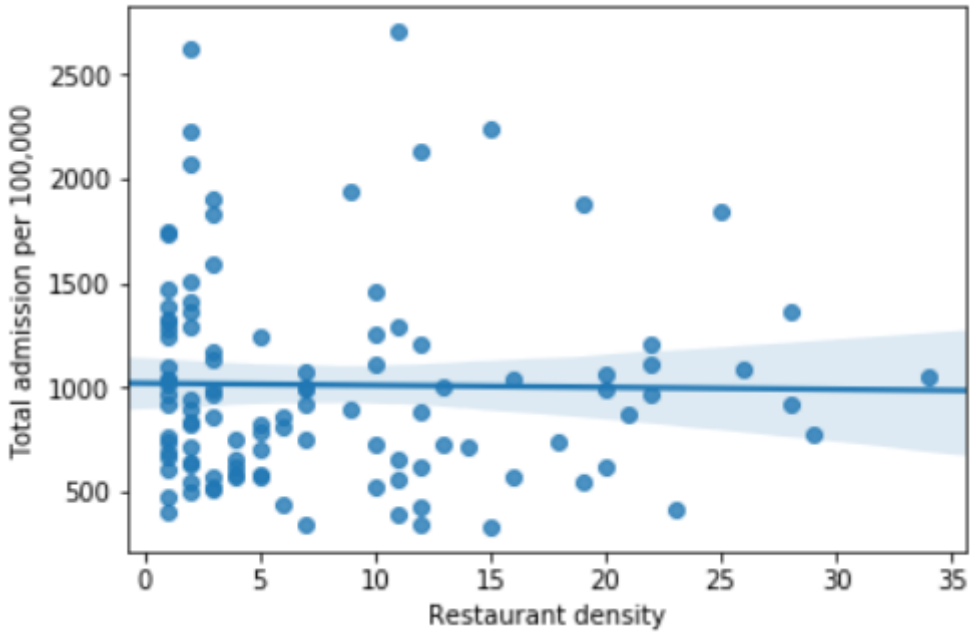

Figure 5: Regression plot of the total admission per 100,000 population versus the restaurant density.

As we can see, the data are spread out around the linear regression prevision. The maxima of admission do not correspond at all with the maxima of restaurant density. We can then predict that there should not be any correlation between these two quantities and can thus expect a bad R2 score. The same statement can be done analyzing the data for male and female only (Figure 6).

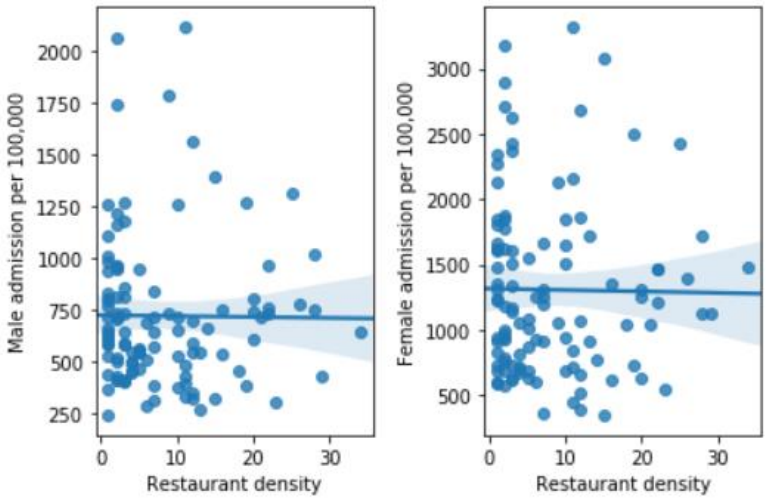

Figure 6: Regression plot of the admission per 100,000 population versus the restaurant density for male and female.

But let's analyze quantitatively this assumption by applying linear regression to our dataset. We divided our dataset in a training set (80% of the data) and a test set (20% of the data). After training our linear model, we calculated various accuracy metrics:

```
Mean absolute error: 381.55
Residual sum of squares (MSE): 262759.26
R2-score: 0.00
```

As we can see, the metrics for the linear regression are very bad, which means that the linear regression is not adapted at all to represent the data. Let's try the predictive accuracy by applying the model to the test set and calculate the accuracy metrics:

```
Mean absolute error: 332.31
Residual sum of squares (MSE): 160457.10
R2-score: -0.03
```

Again, the metrics for the linear regression are very bad, and the model is very wrong at predicting the obesity admission statistics through the density of restaurants. We can thus conclude from this that there are no link between obesity and the density of restaurants in England. Note that more complicated regression such as nonlinear, multiple linear or polynomial regression would not give better results since no tendency can be observed in the scatter plots.

5. Conclusion

From data on obesity and Foursquare data we have derived that there are no link between obesity and the number of restaurants in a district of England. Our model could however be refined by taking into account only restaurants that are more fat providers (such as fast-food, steakhouse,...) and maybe taking into account that the number of gyms or parks in a district can reduce the obesity rate. Moreover, one of the main problem encountered was that the number of venues we can get from Foursquare is very limited and forced us to restrict our analysis to a very small radius around the centre of each district. Increasing the radius could give better insights.