|  |
| --- |
|  |
| Data Science Capstone Project |
| New Metro – New Opportunity |

|  |
| --- |
| Kacper Madej  2020-01-11 |

# 1. Introduction

## 1.1 Background

September 2016 Wola, Warsaw; Gülermak started setting up construction sites for three new metro stations: Płocka, Moczydło and Księcia Janusza. First and third ones got their names from streets they are on. Moczydło got its name from famous park near planned station. In later stage it was renamed Młynów after suburban railway station which has been built in the neighborhood.

Metro seems to be the most convenient way to commute. It’s main key aspects are: high frequency, high average speed and high capacity. It’s popularity brings masses of people to each station every day.

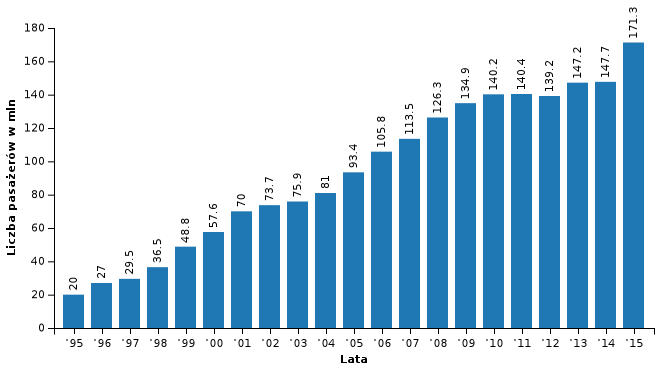


Figure 1 Number of passengers (millions) per year

## 1.2 Problem

Due to its popularity, fresh new metro stations areas are excellent opportunity to start a business. Flat rents are significantly higher in metro areas. More and more investors try to buy a flat as soon as possible, before metro is done just to get paid higher rent after it will be finished. More people also means more customers. Convenient stores, like Żabka (popular polish convenient store franchise), spring up like mushrooms. The same way restaurant industry could use this mechanism. While people go to and from work or school, they get hungry. If the restaurant is near station, it could also be reached from further districts since travelling by metro is fast and comfortable.

## 1.3 Interest

This report is targeted for investors interested in setting up a restaurant. Although the report will be focused mainly on restaurant density, it can be helpful to get general insight about new business possibilities.

# 2. Data

## 2.1 Data sources

All information about new stations are from official ZTM site (Public Transport Authority in Warsaw). To scrap data about each restaurant in designated areas, Google Maps API can be easily used. It require unique API key, which is obtained with free trial account. Google provide basic json format with restaurant name, location (latitude and longitude), rating, price level and total number of ratings.



Figure Dataframe

## 2.2 Data cleaning

Some restaurants does not have i.e. price level information in their json file. Each entry should consist of name, location, rating price level and total number of ratings. If an entry lack of any needed information, it has to be removed.

## 2.3 Data selection

In this report restaurants will be filtered by their rating and total number of ratings, but this two features are often in correlation with price level. Cheaper restaurants tend to be more popular, thus expectations towards them are lower. There need to be a check if price level has significant impact on key features. Any difference should be visible on two following boxplots on the next page.

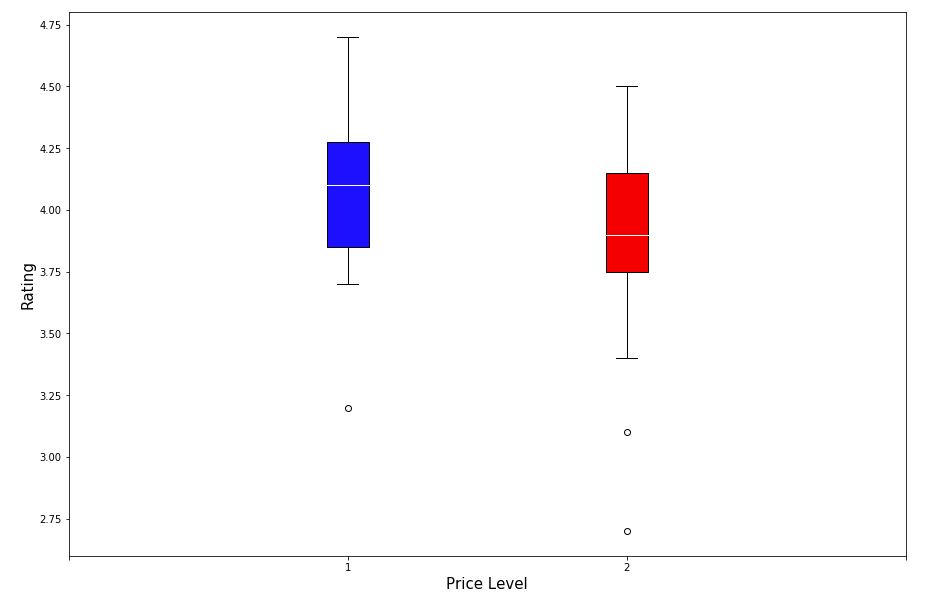
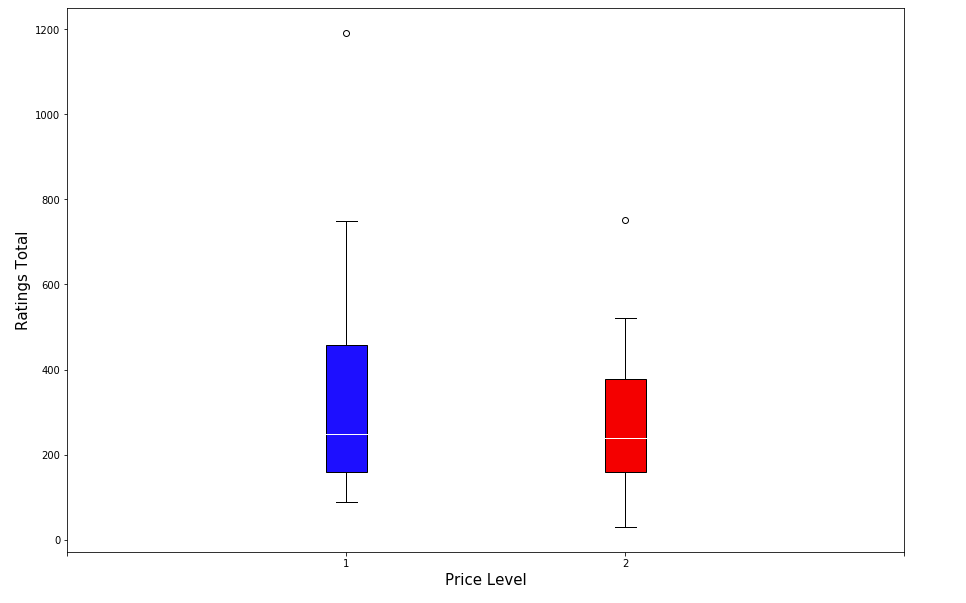
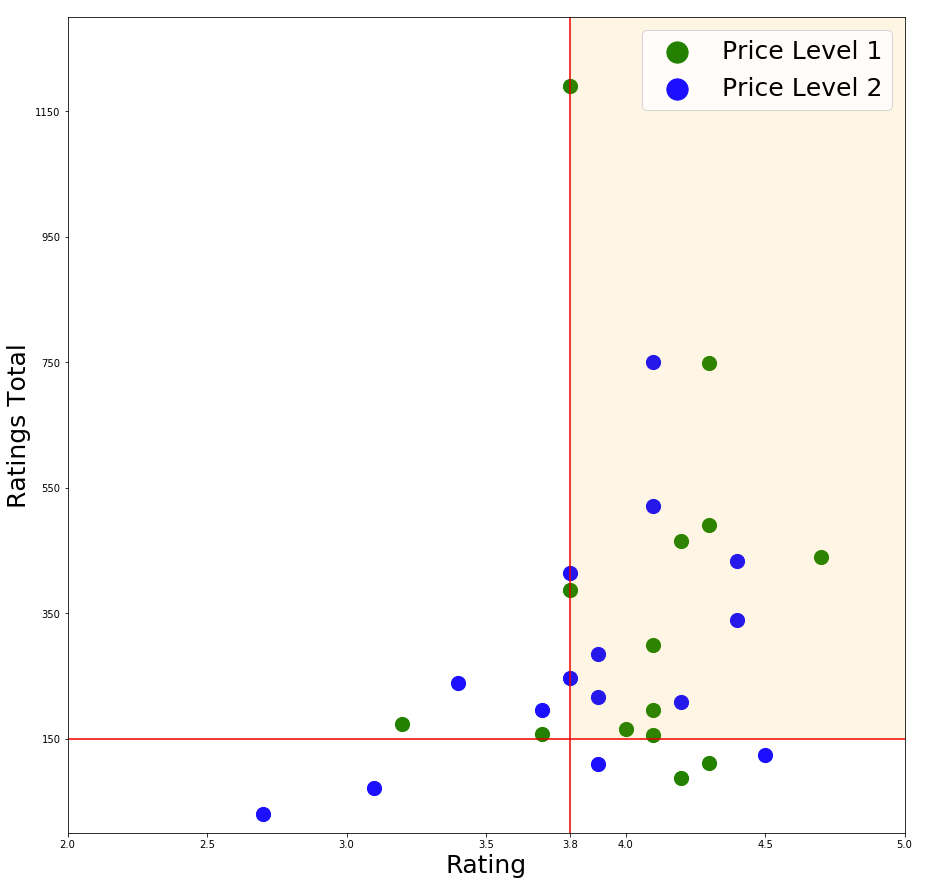
****

Figure Ratings Total boxplot

Figure Rating boxplot

There is slight difference two price levels, but it will not affect further analysis.

I arbitrary chose two restriction on which the restaurants will be filtered. For me, good restaurant has been rated for 3.8 or better. In addition, we should seek for restaurants to have more than 150 ratings, to make sure that it’s rating score is meaningful. To be sure that these filters are adequate to the dataset, I put the entries in the scatter plot below. Red lines limit the area where restaurants have rating greater than 3.8 and ratings number above 150. Hopefully, most of the entries passed those criteria, so only relatively good restaurants are taken into account.

Figure Rating/Ratings Total scatterplot

# 3. Methodology

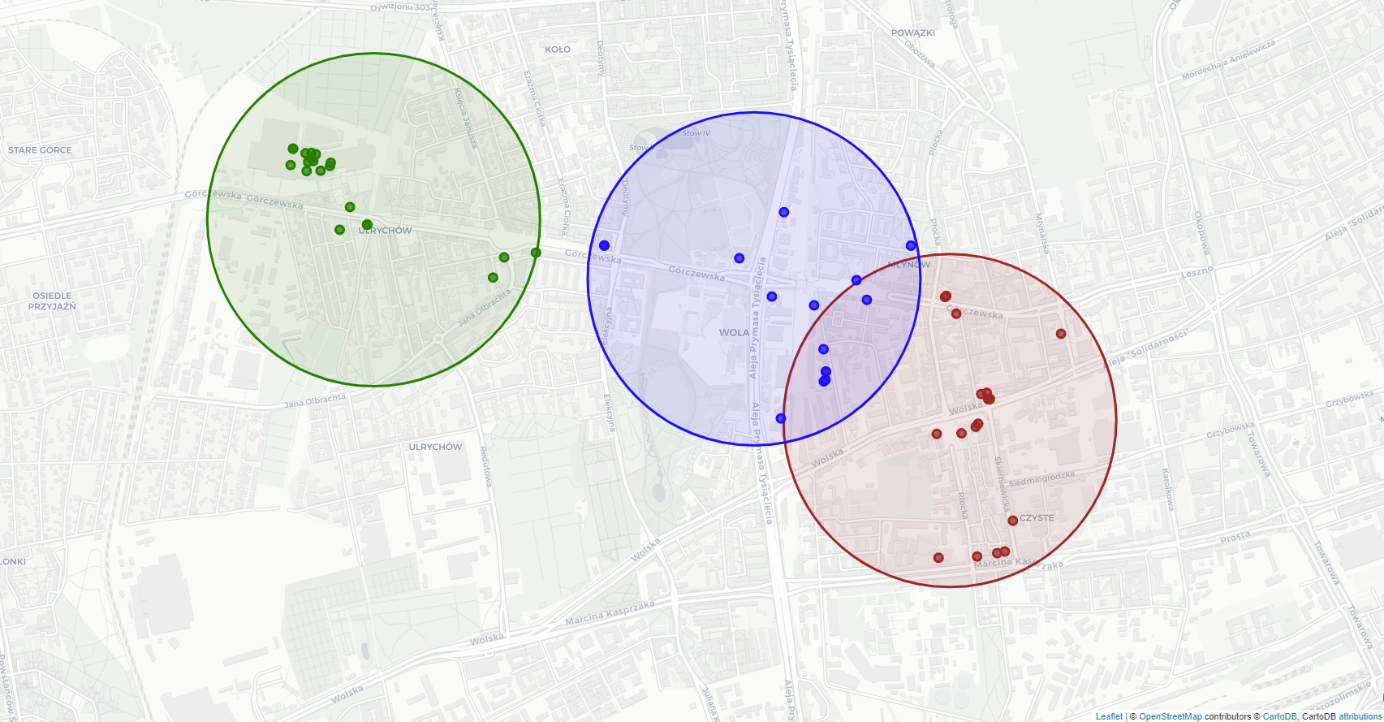
## 3.1 General idea

The main idea of this report is to analyze each station area separately and combine the final results.

A plan for each area is to:

1. Gather and filter nearby restaurants data
2. Create a dataframe
3. Display restaurants map

Figure 6 All restaurants grouped by each station



In the next section I present created dataframes and maps, ready to compare stations by restaurant density.

## 3.2 Station data

|  |  |  |  |
| --- | --- | --- | --- |
| Station name | Station colour | Number of restaurants: | |
| before filtering | after filtering |
| Płocka | Brown | 21 | 10 |
| Młynów | Blue | 9 | 2 |
| Księcia Janusza | Green | 20 | 7 |

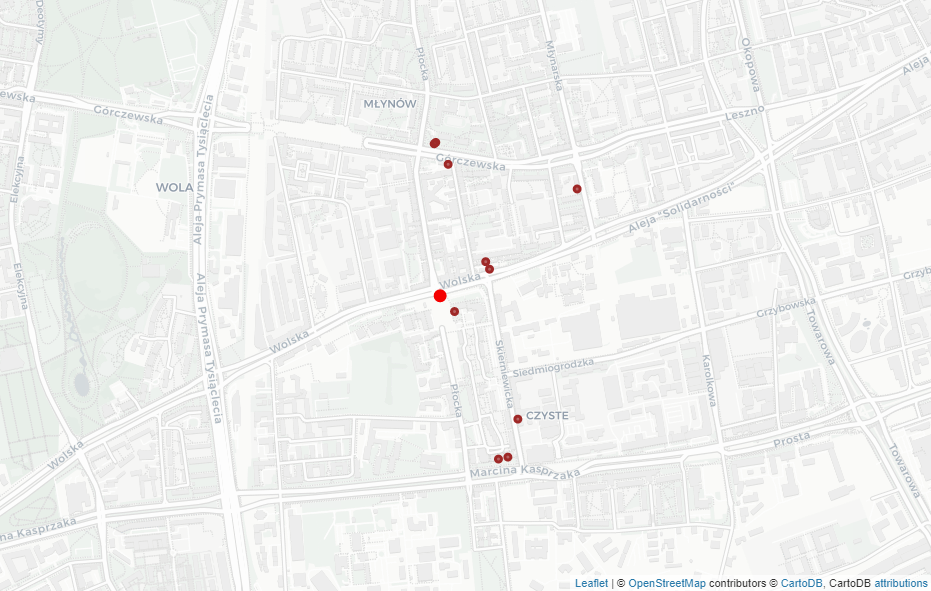
****

Figure Płocka

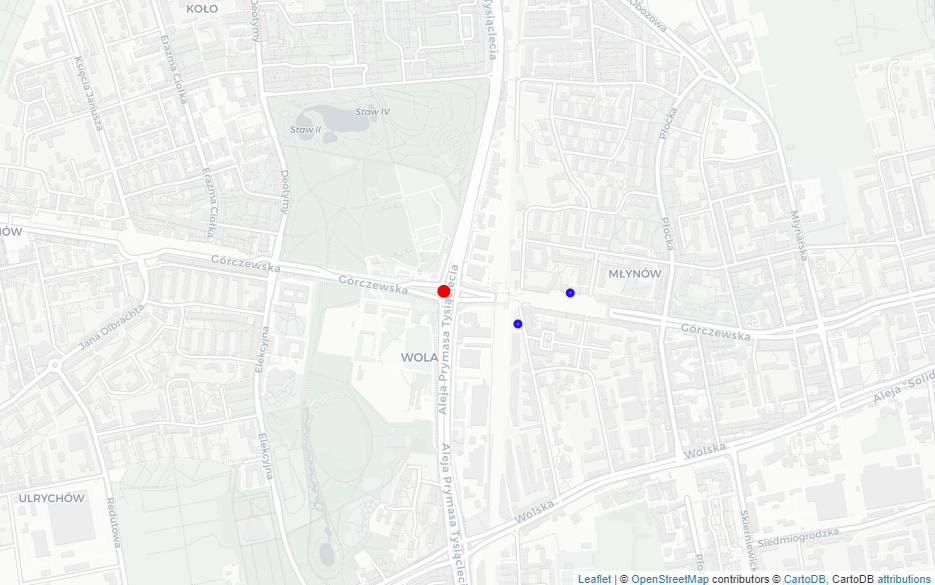
****

Figure Młynów

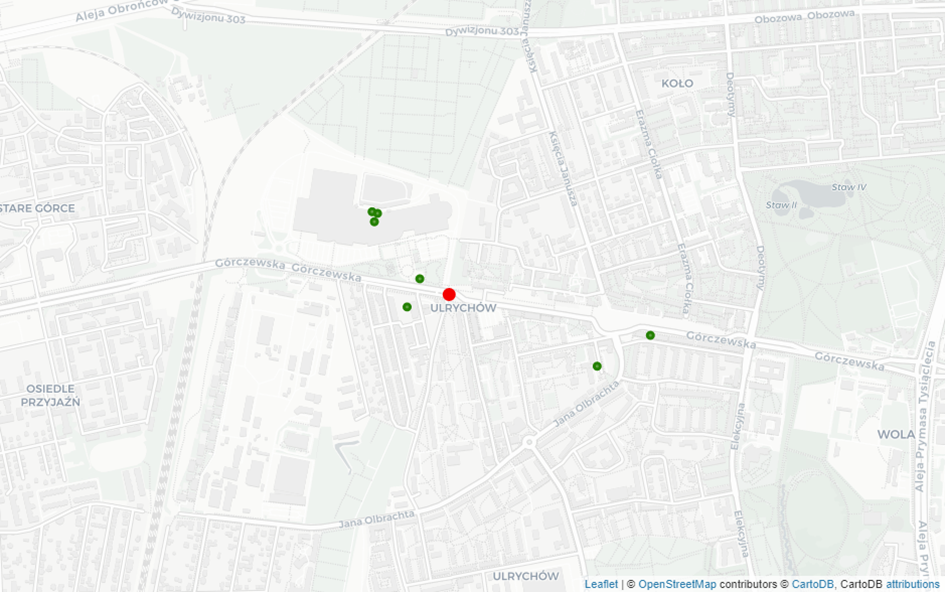
****

Figure Księcia Janusza

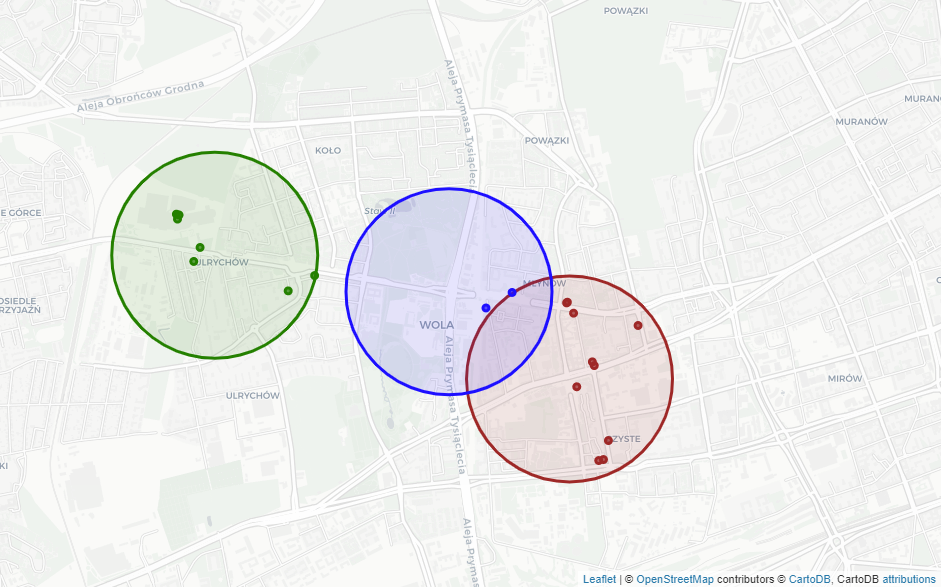


Figure Final map

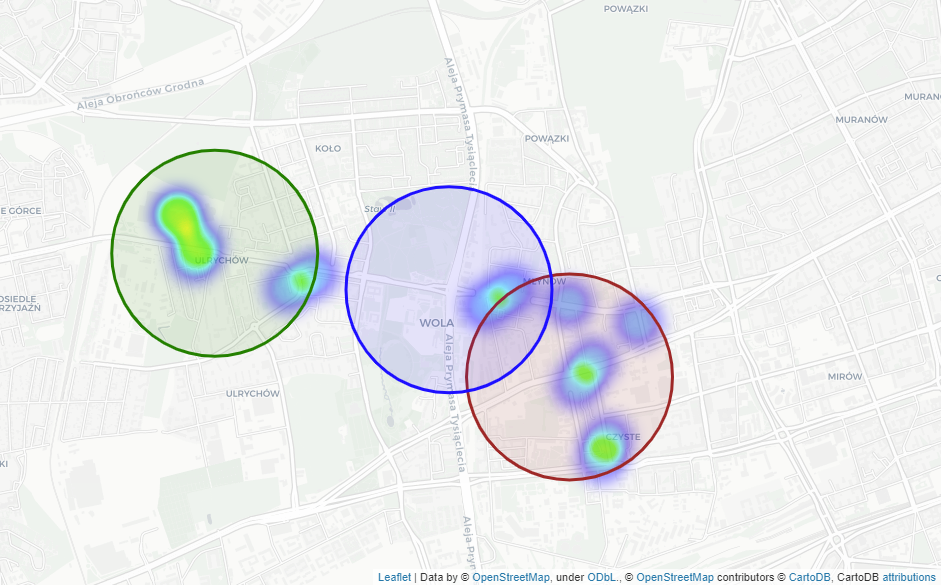
First three maps represents filtered restaurants’ locations. Fourth map combine the result. The fifth shows density of restaurants in each station area. (blue – low density, green – medium density, yellow – high density)

Figure Restaurant density heatmap

## 3.3 Machine learning – K-means clustering

On the heatmap there can be seen five green areas. To check if this observation is correct, restaurants’ locations will be processed through k-means clustering algorithm. Indeed, algorithms clusters are in similar spots as observed green dots. The fact that most restaurants are within clusters means two things. First, maybe these locations attract more customer due to their neighborhood factors like number of bus/tram stops, density of residential flats or proximity of popular places like shopping center or park. Second, there is lot of space left which doesn’t belong to any cluster. With no restaurants nearby, new restaurant could fill the gap between more dense areas.

# 4. Result & Discussion

Figure 12 Final map processed through K-means algorithm

## 4.1 Map result

The heatmap clearly shows that Młynów area has lowest density than other stations. In fact, there are only two restaurants which meet the criteria, leaving a huge gap on the western side of this region. Płocka and Księcia Janusza areas include much more restaurants grouped in four clusters, two for each station. It should be pointed out that one of the clusters in Księcia Janusza area lies on shopping center called Wola Park and that is the reason why it is the most dense region from all clusters.

Comparing overall restaurants map before and after filtering (figure x and figure y), each area lost significant number of entries, but in a different way:

1. Płocka is the only area which remained mostly untouched by filters. Only few restaurants were removed, but it didn’t change the shape of clusters.
2. In Młynów only two restaurants left, leaving huge gap on the eastern side of Moczydło Park.
3. Księcia Janusza general shape was a bit altered, depicting that most well rated restaurants are in or nearby Wola Park.

## 4.2 General observations

1. Płocka has some free space although it already consists two clusters with relatively high restaurant density. But both clusters are in some distance from the station so if Investor has an opportunity to rent a place just near the entrance, it may be profitable but risky.
2. Księcia Janusza is a difficult location due to Wola Park, a shopping center near the station. Wola Park attracts many citizens, but comes with bigger competition. Most of the restaurants in this area are famous brands like KFC or McDonald. The opportunity here is to seek for location near Górczewska, which is the main street. Probably a vegan restaurant serving healthy meals could fill the gap left by junky fast-foods.
3. Młynów seems to be the best choice. Low restaurant density with famous park nearby may bring many customers. This area also includes suburban railway station, Moczydło (an aquapark during summer, an ice rink during winter) and Moczydło park. Only two restaurants should be taken into account when choosing a type of future restaurant. One of them serves Italian cuisine while other serves sushi. Investor may be interested in Chinese takeaways or kebab.

# 5. Conclusion

## 5.1 Report conclusion

From this report it is clear that Młynów seems to be the best area to open a new restaurant. It has lowest restaurant density with both filtered and unfiltered data. Also the Młynów area has just one cluster comparing with two more dense clusters in each other station. Another advantage is that there is famous park called Moczydło nearby the station and with aquapark (ice rink in winter) next to it attracts lots of people. Last year, the suburban railway station was built in this region, so Młynów station could be potentially an interchange station between suburban railway and metro connecting city center with outskirts. Future investors should not worry about competition since there are only two good restaurants in the area.

## 5.2 Other aspects

This report focus only on density of good restaurants in specific areas. Although in my opinion it is one of major factors, the investor should be aware that it does not guarantee the profit. There are another aspects of choosing right place for investment such as restaurant type, investment size, target customers and much more. The goal of this report is to indicate in which area there could be an interesting spot to set up new business and therefore other steps should be taken.

Table of Contents

[1. Introduction 1](#_Toc29646619)

[1.1 Background 1](#_Toc29646620)

[1.2 Problem 1](#_Toc29646621)

[1.3 Interest 1](#_Toc29646622)

[2. Data 2](#_Toc29646623)

[2.1 Data sources 2](#_Toc29646624)

[2.2 Data cleaning 2](#_Toc29646625)

[2.3 Data selection 2](#_Toc29646626)

[3. Methodology 5](#_Toc29646627)

[3.1 General idea 5](#_Toc29646628)

[3.2 Station data 5](#_Toc29646629)

[3.3 Machine learning – K-means clustering 8](#_Toc29646630)

[4. Result & Discussion 9](#_Toc29646631)

[4.1 Map result 10](#_Toc29646632)

[4.2 General observations 10](#_Toc29646633)

[5. Conclusion 11](#_Toc29646634)

[5.1 Report conclusion 11](#_Toc29646635)

[5.2 Other aspects 11](#_Toc29646636)

Table of Figures

[Figure 1 Number of passengers (millions) per year 1](file:///C:\Users\Kacper\Desktop\Capstone%20Data%20Science.docx#_Toc29646637)

[Figure 2 Dataframe 2](file:///C:\Users\Kacper\Desktop\Capstone%20Data%20Science.docx#_Toc29646638)

[Figure 3 Ratings Total boxplot 3](file:///C:\Users\Kacper\Desktop\Capstone%20Data%20Science.docx#_Toc29646639)

[Figure 4 Rating boxplot 3](file:///C:\Users\Kacper\Desktop\Capstone%20Data%20Science.docx#_Toc29646640)

[Figure 5 Rating/Ratings Total scatterplot 4](file:///C:\Users\Kacper\Desktop\Capstone%20Data%20Science.docx#_Toc29646641)

[Figure 6 All restaurants grouped by each station 5](file:///C:\Users\Kacper\Desktop\Capstone%20Data%20Science.docx#_Toc29646642)

[Figure 7 Płocka 6](#_Toc29646643)

[Figure 8 Młynów 6](#_Toc29646644)

[Figure 9 Księcia Janusza 7](#_Toc29646645)

[Figure 10 Final map 7](#_Toc29646646)

[Figure 11 Restaurant density heatmap 8](file:///C:\Users\Kacper\Desktop\Capstone%20Data%20Science.docx#_Toc29646647)

[Figure 12 Final map processed through K-means algorithm 9](file:///C:\Users\Kacper\Desktop\Capstone%20Data%20Science.docx#_Toc29646648)