# Selection of a Location for New Tourist-Oriented Bakery in Yekaterinburg, Russia

BY EDUARD MEILAKH

# 1. Introduction

## 1.1. Background

Yekaterinburg is the largest city of Ural Federal District, Russia and is the administrative center of Sverdlovsk Region. The city is located on the Iset River between the Volga-Ural region and Siberia.

According to Wikipedia, the population of Yekaterinburg is roughly 1.5 million residents. Yekaterinburg is the fourth-largest city in Russia.

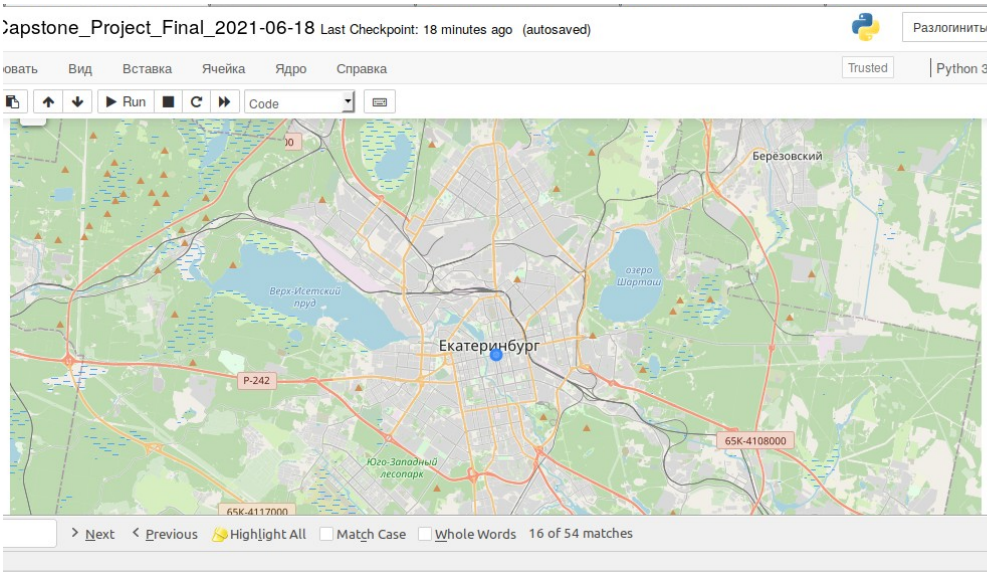


Figure 1. Map of Yekateriburg (blue point on the map is the city center).

## 1.2. Business Problem

There are a lot of hotels in Yekaterinburg. Meanwhile, market niches for venues, including bakeries, near hotels are overflowing and highly competitive. That's why a task to choose a location for new tourist-oriented bakery in Yekaterinburg is complex and comprehensive.

Yekaterinburg is business city, so the most tourists are located near hotels and business centers. One of the possible approaches to solve the task of choosing the location for new bakery is to explore territories near hotels. On the one hand, the number of venues within 300 meters of a hotel will give an understanding of capacity of a market niche and number of potential customers, and on the other hand, it is good idea to look for hotels where there is no bakeries nearby.

# 2. Data Acquisition and Cleaning

## 2.1. Data Sources

There are a lot of sources of information about hotels in Yekaterinburg.

We need the following data for our project:

- list of hotels in Yekaterinburg, their names, coordinates, stars;

- venues around each hotel, their number and types;
- the city center coordinates.

We will obtain data about hotels's name, location (coordinates), class (number of stars from 1 to 5) by scraping https://www.gogototour.com/ru/city/yekaterinburg/, https://tophotels.ru/ and via API of Foursquare. We will use the list of hotels for which all information is available (location, stars, names).

Venues nearby each hotel will be obtained via Foursqaure API.

The city center coordinates will get on Wikipedia page.

```
CENTER_LOCATION = (56.8383, 60.6036)
```

Also, we will calculate a distance between each hotel and the city center (in meters). We will use python module `vincenty` for this purpose:

```
from vincenty import vincenty
```

## 2.2. Data Cleaning

Data are obtained from multiple sources, that is why one sub-task of data cleaning is to unify hotel names when possible and delete duplicate entries. Hotel names are written with use of both Cyrillic alphabet, and Latin (English) alphabet. Also, hotel names may contain words like 'hotel', 'Yekaterinburg' (both English, and Russian versions of these words), or may not.

Steps of hotel names processing:

- write all words in lower case;

- delete punctuation and common words like 'hotel' and 'Yekaterinburg';

- transliterate Cyrillic letters into English ones;

- capitalize words (each word will begin with a capital letter).

Functions for these purposes have been developed:

```
def capwords(words): # capitalize each word
    return ' '.join(word.capitalize() for word in words.split())

punctuation = '"«»/,()'
COMMON_WORDS = ['hotel', 'hotels', 'отель', 'гостиница', 'екатеринбург', 'ekaterinburg',
'yekaterinburg']
RUS_ABC = ('а', 'б', 'в', 'г', 'д', 'е', 'ё', 'ж', 'з', 'и', 'й', 'к', 'л', 'м', 'н',
           'о', 'п', 'р', 'с', 'т', 'у', 'ф', 'х', 'ц', 'ч', 'ш', 'щ', 'ъ', 'ы', 'ь',
           'э', 'ю', 'я')
TRANSLIT = ('a', 'b', 'v', 'g', 'd', 'e', 'yo','zh','z', 'i', 'j', 'k', 'l', 'm', 'n',
           'o', 'p', 'r', 's', 't', 'u', 'f', 'h', 'c', 'ch','sh','sch','\'','y', '\'',
           'e', 'yu','ya')
LAT_ABC = tuple("abcdefghijklmnopqrstuvwxyz-")

def clean_name(name):
    if "/" in name:
        eng_name = name.split("/")[1].strip()
        rus_name = name.split("/")[0].strip()
        if len(set(eng_name) - set(RUS_ABC)-set(punctuation)) < 3:
            eng_name, rus_name = rus_name, eng_name
        return clean_name(eng_name)
    cleaned_name = ''.join(ch for ch in name.lower() if ch not in
                           punctuation
                           )
    cleaned_name = ' '.join(word for word in cleaned_name.split() if word not in
COMMON_WORDS).strip()
    if len(set(cleaned_name) - set(RUS_ABC)) < 3:
        additional_symbols = tuple(set(cleaned_name)-set(RUS_ABC))
        ABC1 = RUS_ABC + additional_symbols
        ABC2 = TRANSLIT + additional_symbols
        cleaned_name = ''.join(ABC2[ABC1.index(ch)] for ch in cleaned_name)
    return capwords(cleaned_name)
```

# 3. Methodology

## 3.1. Descriptive Statistics and Data Visualization

The main data I used have the following features: *Hotel Name* (names), *Latitude* (lats), *Longitude* (longs), *Stars*, also *the Distance to the city center* (to_center) has been calculated for each hotel.

Out[113]:

| | names | lats | longs | stars | to_center |
|---|---|---|---|---|---|
| 0 | Viz'avi | 56.837612 | 60.548379 | 4 | 3.371344 |
| 1 | Park Inn | 56.836491 | 60.618449 | 4 | 0.928455 |
| 2 | Best Eastern Uralsky Dvor | 56.827579 | 60.616588 | 4 | 1.433141 |
| 3 | Atlaza City Residence | 56.823979 | 60.637920 | 4 | 2.633022 |
| 4 | Grand Hall | 56.828640 | 60.559570 | 4 | 2.894994 |
| 5 | Senator Business | 56.841709 | 60.576708 | 4 | 1.684621 |
| 6 | Ramada | 56.775280 | 60.717830 | 5 | 9.896437 |
| 7 | Ural | 56.841729 | 60.572492 | 4 | 1.936623 |
| 8 | Onegin | 56.828814 | 60.613890 | 4 | 1.228979 |
| 9 | Soldi | 56.903450 | 60.578182 | 4 | 7.418683 |
| 10 | Hyatt Regency | 56.843521 | 60.591788 | 5 | 0.926131 |
| 11 | Moskovskaya Gorka | 56.822496 | 60.587947 | 4 | 2.002584 |

Also, information about venues nearby each hotel has been obtained:

In [123]:
```
venues.head(10)
```

Out[123]:

| | Hotel | Hotel Latitude | Hotel Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Angelo Airport | 56.751256 | 60.798917 | Зал прибытия | 56.750443 | 60.800361 | Airport Service |
| 1 | Angelo Airport | 56.751256 | 60.798917 | Koltsovo International Airport (SVX) (Междунар... | 56.750094 | 60.801070 | Airport |
| 2 | Angelo Airport | 56.751256 | 60.798917 | Angelo Airport Hotel 4* | 56.750881 | 60.799312 | Hotel |
| 3 | Angelo Airport | 56.751256 | 60.798917 | Ресторан Sunlight | 56.750981 | 60.798772 | Eastern European Restaurant |
| 4 | Angelo Airport | 56.751256 | 60.798917 | Cherry Berry | 56.750432 | 60.802674 | Juice Bar |
| 5 | Angelo Airport | 56.751256 | 60.798917 | Angelo Jazz bar | 56.750825 | 60.799158 | Café |
| 6 | Angelo Airport | 56.751256 | 60.798917 | Паспортный контроль / Passport Control | 56.750122 | 60.803202 | Airport Service |
| 7 | Angelo Airport | 56.751256 | 60.798917 | Charter's Pub | 56.750101 | 60.802049 | Pub |
| 8 | Angelo Airport | 56.751256 | 60.798917 | Мамуля | 56.749940 | 60.800945 | Comfort Food Restaurant |
| 9 | Angelo Airport | 56.751256 | 60.798917 | Coffeeshop Company | 56.750167 | 60.801731 | Coffee Shop |

I used python **folium** library to visualize geographic details of hotels location in Yekaterinburg.

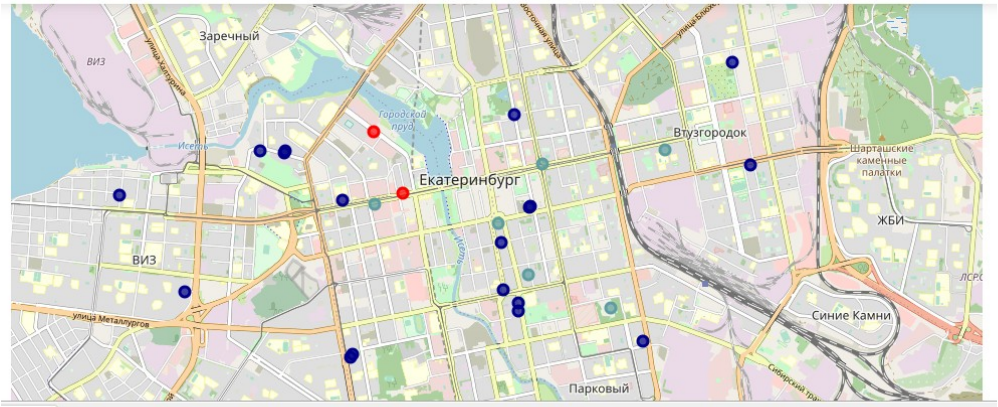


Figure 2. Map of Yekaterinburg with hotels on it. Colors designate stars of the hotels.

I used scatter plot visualization in order to explore if there is a correlation between distance to the city center and number of venues nearby a hotel, and between hotel's stars and number of venues nearby a hotel.

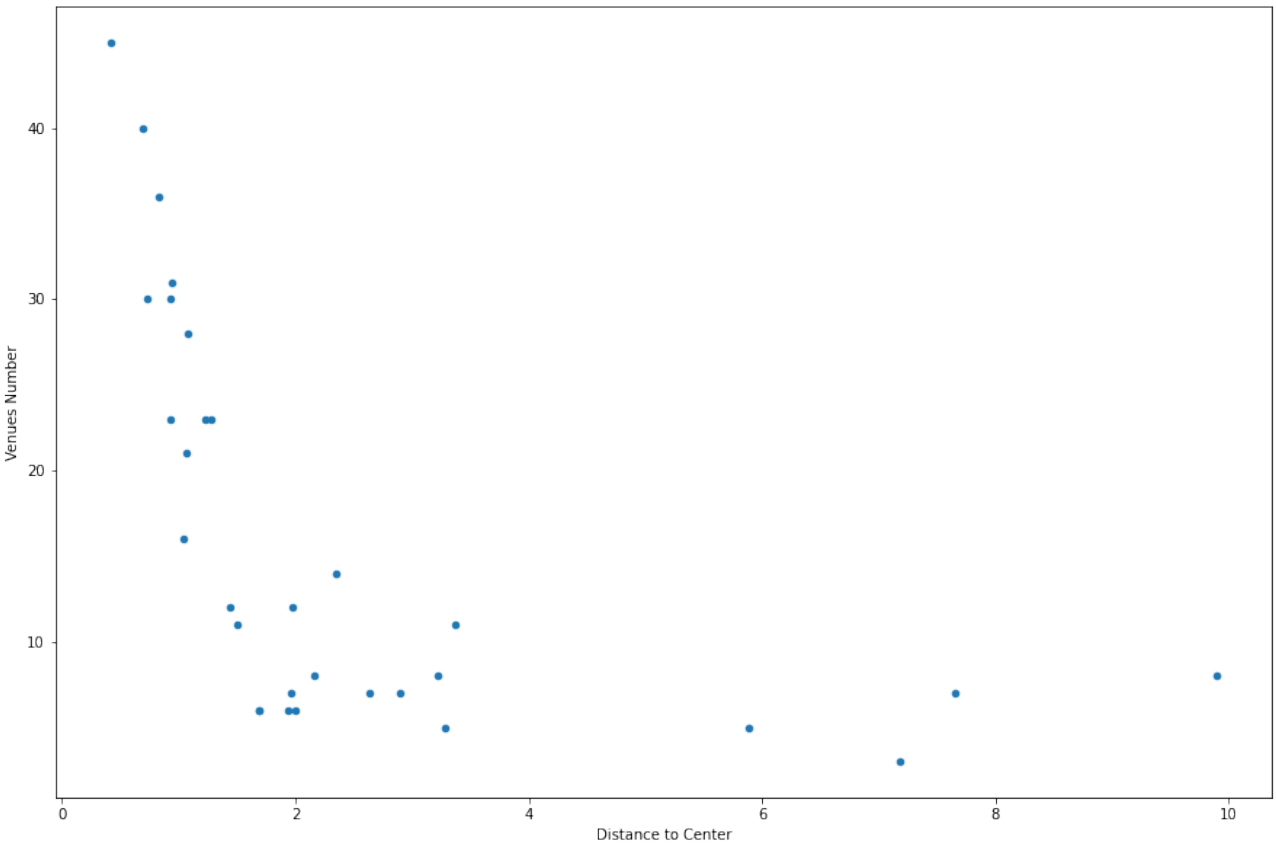


Figure 3. Correlation between a distance from a hotel to the city center, m (X-axis) and the number of venues nearby a hotel (Y-axis).

As expected, the closer to the city center, the more venues are located.

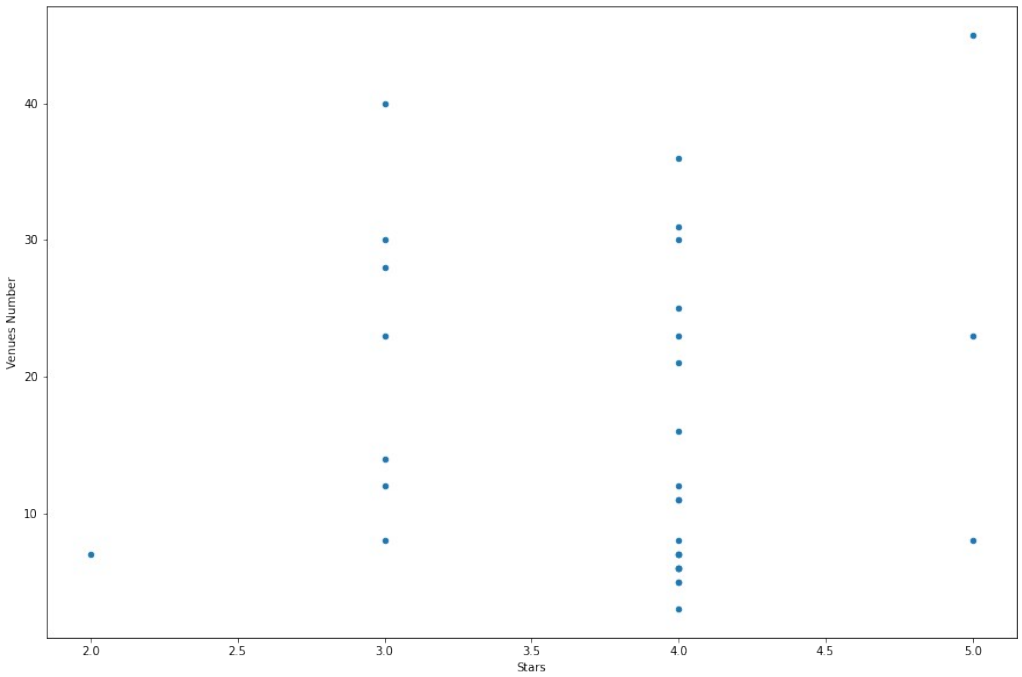


Figure 4. No correlation between hotel's stars (X-axis) and the number of venues nearby a hotel (Y-axis) is observed

Using Foursquare API I get venues for each hotel and their categories. There are 130 unique categories of venues in our study.

I used GitHub repository in my study to store my jupyter notebook.

## 3.2. Cluster Analysis

I used K-Means method for clustering hotels according to most common categories of venues near each hotel. Calculations have been performed with `sklearn` module.

Data preparation for clustering had the following steps:

1. The data about venues categories have been transformed into categorical variables by using `get_dummies` function of `pandas` package for Python:

```
import numpy as np
onehot = pd.get_dummies(venues[['Venue Category']], prefix="", prefix_sep="")

# add Hotel column back to dataframe
onehot['Hotel'] = venues['Hotel']

# move Hotel column to the first column
fixed_columns = [onehot.columns[-1]] + list(onehot.columns[:-1])
onehot = onehot[fixed_columns]

onehot.head()
```

Out[127]:

| | Yoga Studio | ATM | Airport | Airport Lounge | Airport Service | Airport Terminal | American Restaurant | Arcade | Art Gallery | Art Museum | ... | Theater | Theme Park | Theme Park Ride / Attraction | Toy / Game Store | Ukrainian Restaurant | Vegetarian / Vegan Restaurant |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 |

5 rows × 130 columns

2. The next step was to group rows by hotel and calculate the mean of the frequency of occurrence of each category. For data understanding, venues categories have been sorted by their frequencies in descending order for each hotel (from the 1st most common venues to 10th most common venues).

Out[131]:

| | Hotel | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue | Venues Number |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Angelo Airport | Coffee Shop | Airport Service | Café | Airport Lounge | Pie Shop | Pub | Eastern European Restaurant | Pelmeni House | Juice Bar | Russian Restaurant | 25 |
| 1 | Atlaza City Residence | Hunting Supply | Café | Shopping Mall | Liquor Store | Tennis Stadium | Steakhouse | Intersection | Convenience Store | Cosmetics Shop | Dance Studio | 7 |
| 2 | Atrium Palace | Coffee Shop | Pizza Place | Bookstore | Fast Food Restaurant | Café | Gastropub | Clothing Store | Restaurant | Gym / Fitness Center | Pet Store | 45 |
| 3 | Avs | Bath House | Gym | Playground | Park | Karaoke Bar | Volleyball Court | Concert Hall | Comfort Food Restaurant | Convenience Store | Cosmetics Shop | 7 |
| 4 | Best Eastern | Gym / Fitness | Street Art | Nightclub | Café | Restaurant | Park | Coffee Shop | Gift Shop | Grocery | Comfort Food | 12 |

3. The third step is to prepare data for clustering (a column contained hotel names has been dropped).

4. Clustering analysis has been perform to divide all hotels into three clusters. Cluster labels for hotels have been defined:

```
In [138]: kclusters = 3
          # run k-means clustering
          kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(clustering)

          # check cluster labels generated for each row in the dataframe
          kmeans.labels_

Out[138]: array([2, 2, 2, 2, 2, 2, 0, 2, 2, 2, 2, 2, 0, 0, 2, 0, 2, 1, 2, 2, 2, 2,
                 0, 2, 2, 2, 2, 2, 0, 2, 2, 2], dtype=int32)
```

We can explore each cluster in order to define its characteristics.

Cafes and coffee shops are located nearby hotels belonging to Cluster # 2 only.

Gym centers, spa, theme park ride / attractions are more characteristic for hotels belonging to Cluster # 0.

Cluster # 1 has only one hotel, grocery store, palace and garden are located nearby this hotel.

Clusters description is below:

**Cluster: 0**

| | |
|---|---|
| Gym / Fitness Center | 6.20% |
| Gym | 5.42% |
| Furniture / Home Store | 5.16% |
| Karaoke Bar | 5.16% |
| Theme Park Ride / Attraction | 5.16% |
| Spa | 4.17% |
| Pharmacy | 3.33% |

==========

**Cluster: 1**

| | |
|---|---|
| Grocery Store | 33.33% |
| Palace | 33.33% |
| Garden | 33.33% |

==========

**Cluster: 2**

| | |
|---|---|
| Café | 5.71% |
| Coffee Shop | 5.13% |
| Gym / Fitness Center | 4.59% |
| Restaurant | 3.79% |
| Park | 3.05% |
| Nightclub | 2.55% |
| Spa | 2.37% |

==========

Hotels are placed on the map, colors of the dots correspond to clusters (cluster 0 is purple, cluster 1 is green, cluster 2 is red).

Out[150]:

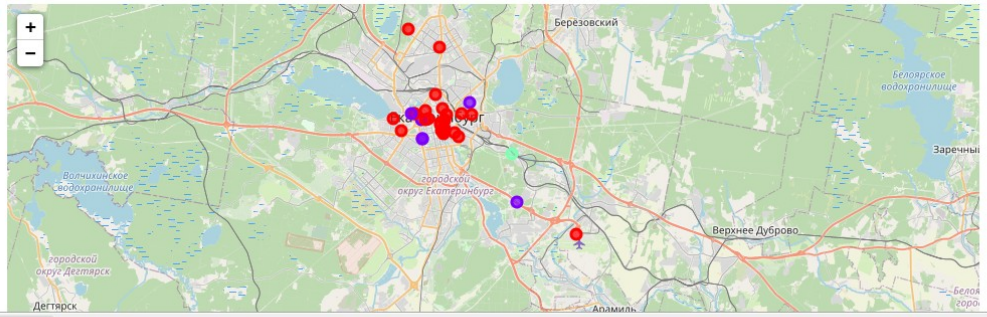


Figure 5. Clustering hotels (cluster 0 is purple, cluster 1 is green, cluster 2 is red)

## 4. Results

The results of the study:

1. Venues of 130 unique categories are located nearby hotels in Yekaterinburg, Russia.

2. There are more venues near hotels which are located closer to the city center. There is only one exception from this rule: the hotel located near airport (it is far from the city center, and there are a lot of venues of airport services near it).

3. The hotels in Yekaterinburg may be divided into three clusters. Cafes and coffee shops are located nearby hotels of one cluster; spa, sport venues are characteristic for another cluster; the third cluster includes only one hotel which has a specific location.

## 5. Discussion and Recommendations

The goal of this study is recommendations how to choose a location for new, tourist-oriented bakery in Yekaterinburg.

The fact that there are more venues closer to the city center gives us an insight that the closer a location to the center of Yekaterinburg, the higher capacity of market niches is. On the other hand locations that are further than 2,5 km from the city center have approximately equal density of venues. Taking into account that rental rate is higher in the center of the city, we can consider to place new bakery on a location 1,8-2,5 km of the city center.

The good idea is to select a location for the bakery near a hotel belonging to the cluster 2 ( the most common venues near hotels of this cluster are cafes and coffee shops), but to choose a hotel nearby which there is no bakery. There are few hotels which are met these conditions.

## 6. Conclusion

In this study I analyzed venues near hotels in Yekaterinburg, Russia using explanatory analysis, K-means for clustering, map visualization, scatter plots.