IBM

Applied Data Science

Capstone Project

Opening a retail-entertainment center in Ukraine

Author: Filjuk Valentin

Dnepr 2020

TABLE OF CONTENTS

Executive Summary

This report provides an analysis and evaluation of the potential place of retail-entertainment center for investors (relevant stakeholders). Methods of analysis include k-Means Clustering. All calculations can be found in jupyter notebook. Results of analyzed data show a possible places to build a new retail-entertainment center. The resulted data was divided in 3 cluster:

- Cluster 0. High number of retail-entertainment centers (with 2 or more retail-entertainment centers per district);
- Cluster 1. Moderate number of retail-entertainment centers (with 1 retail-entertainment center per district);
- Cluster 2. Without retail-entertainment centers (with 0 retail-entertainment center per district)

The recommendations are:

- build a new retail-entertainment center in Cluster 0 or Cluster 1;
- Cluster 0 would a good choice to avoid competition;
- Cluster 1 would provide a competition with stable revenue.
- build in Cluster 2 is bad idea, because you just drop an income from retail-entertainment center.

The report has limitations such as size of population, income of residents, shape of districts.

I have gone through whole process of submitting capstone project and how would look like the real project from data scientist

Introduction

Finding and visiting scattered shopping and entertainment venues takes extra time. A retail-entertainment center saves people’s time, allow them to relax and make purchases. This way of organizing involves working of shops, restaurants, cafes, service centers, cinemas and other places of leisure in a single place. Those factor leads to a stable profit for each participant in the work process in the retail-entertainment center, especially the lessor.

The aim of the project is to help investors to choose a place for the construction of a new retail-entertainment center.

1. Analysis

1.1. Characteristics of the subject area and object of study

Shopping center is a group of architecturally united shops managed as a whole and built in a special area. A shopping center also provides a parking zone.

The most important characteristic of a shopping center is the anchor.

Anchor is the shops that attract major customers flow.

A retail entertainment center has the anchor consisting from shops (supermarket, shoes shop, pharmacy, etc.) and entertainment (cinemas, entertainment centers, restaurants, etc.). Gift shops, accessories, audio and video products and service shops are act as secondary tenants.

The location of a retail entertainment center is mainly affected by the size of the economically active population, average per capita income, demographic structure of the population, the number of people employed in the economy, investment volumes.

1.2. The Data

1.2.1. Brief overview

I limited the data set due to the large number of variables that need to be considered. This type of research goes beyond the time allocated for the project.

For this project I used information about districts of top 8 big cities of Ukraine, namely: list of districts, latitude and

longitude. The I would search for venue data related to retail entertainment center.

Altogether gathered info would lead us to a potential place for construction of a new retail-entertainment center.

1.2.2. Cities and districts

I created a list of cities and districts presented in research.

City	*District*
Kyiv	Holossijiw, Sviatoshynskyi, Solomianskyi, Obolonskyi, Podilskyi, Pecherskyi, Shevchenkivskyi, Darnytskyi, Dniprovskyi, Desnianskyi
Kharkiv	Shevchenkivskyi, Kyivskyi, Slobidskyi, Osnovianskyi, Kholodnohirskyi, Moskovskyi, Novobavarskyi, Industrialnyi, Nemyshlyanskyi
Odessa	Suvorovsky, Prymorsky, Malynovsky, Kyivsky
Dnipro	Amur-Nyzhnodniprovskyi, Shevchenkivskyi, Sobornyi, Industrialnyi, Tsentralnyi, Chechelivskyi, Novokodatskyi Samarskyi
Zaporizhia	Oleksandrivskyi, Dneprovsky, Voznesenskyi, Khortytskyi,

City	*District*
Zaporizhia	Shevchenkivskyi
Lviv	Halytskyi, Zaliznychnyi, Lychakivskyi, Frankivskyi, Shevchenkivskyi, Sykhivskyi
Kryvyi Rih	Dolgintsevskiy, Inhuletskyi, Metalurhiynyy, Pokrovsky, Saksahanskyi, Ternivskyi, Tsentral'no-Gorodskoy
Kherson	Suvorovskiy, Korabel'nyy, Dneprovsky

All data about cities and districts contains in file ‘Ukraine_cities_n_districts.csv’. Also, latitude and longitude of districts are in the same file. I prepared it beforehand.

2. Methodology

2.1. Brief overview of k-Means Clustering

The k-means algorithm searches for a pre-determined number of clusters within an unlabeled multidimensional data set.

The amount of cluster centroids (cluster centers) determined by amount of clusters (k-value) you want to split up the data.

The first step of the algorithm is that it randomly initialize position of cluster centroids. Then all data points assignment to the cluster centroids. The data points assigns to the closest cluster centroid. Afterward, cluster centroids are moved to location of mean of data points assignment to the cluster in previous step. Then we again assign all data points to the cluster centroids. Some data points points may change cluster centroids. We run the process until no more data points change their cluster centroids. We get clusters in the end of the process.

2.2. Implementation

I will use data from 'Ukraine_cities_n_districts.csv' to get list of:

- cities;
- districts;
- latitude and longitude of districts.

After that, I will use Foursquare API (explore API) to get top 100 venues that are withing 2500 meters radius (avg. size of district). You need to have developer's account to get access to

Foursquare API, so I registered there beforehand. I will use latitude and longitude of districts to get venues. Foursquare would return JSON. I need venues name, latitude and longitude of venues, category list of venues types from JSON data. Then I would analyze each district by retail-entertainment center occurrence. There is no retail-entertainment center category. So, I searched thorough Shopping Maul category and manually removed shopping centers that can't be categorized as retail-entertainment center. I would group up data by districts to get amount of retail-entertainment center. The data would be prepared for K-mean clustering.

I clustered data on 3 clusters based of the amount of retail-entertainment center in districts. It would allows to identify which districts have high concentration of retail-entertainment centers, moderate concentration of retail-entertainment centers and districts without of retail-entertainment centers.

2.3. Limitations

In this project, I would omit factors like population, income of residents, other shops that could influence the location of a new construction place of retail-entertainment center.

The real size of districts are presented as complex figure and can't be described as by a radius.

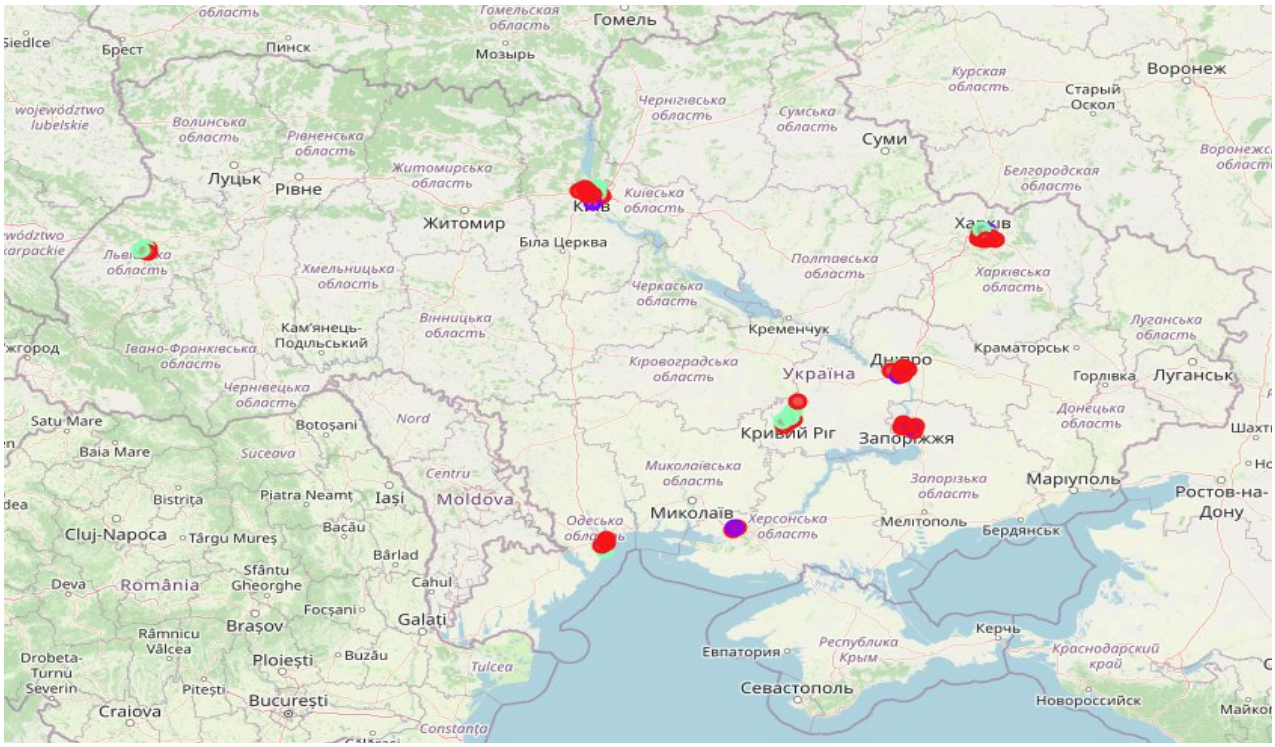
The project can't be implemented in real construction project and requires addition modeling and calculations.

3. Results

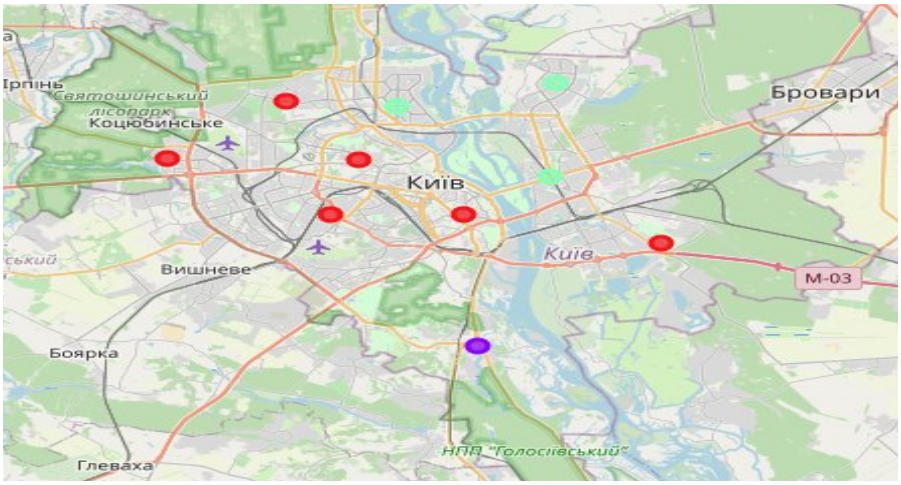
The results are 3 clusters with:

- Cluster 0. High number of retail-entertainment centers (with 2 or more retail-entertainment centers per district);
- Cluster 1. Moderate number of retail-entertainment centers (with 1 retail-entertainment center per district);
- Cluster 2. Without retail-entertainment centers (with 0 retail-entertainment center per district).

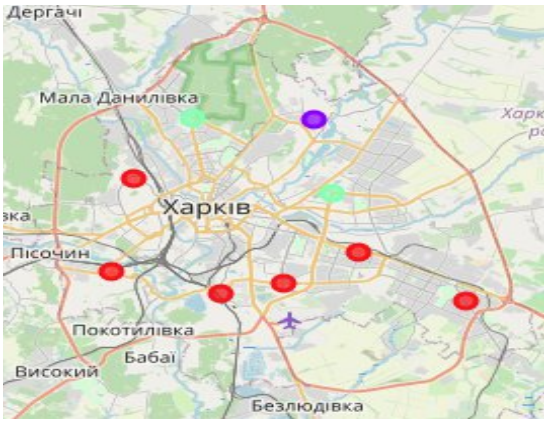
The results of clustering are visualized in the maps below. The Cluster 0 is in the red color, the cluster 1 is in the purple color and cluster 2 is in the light green color.



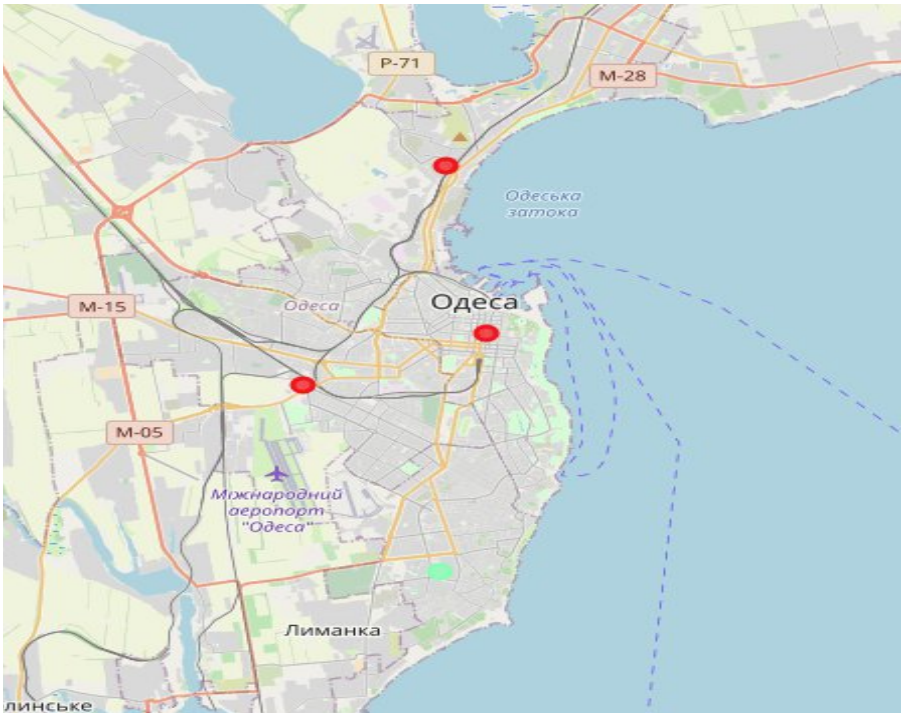
Picture 1. The map of Ukraine with clusters



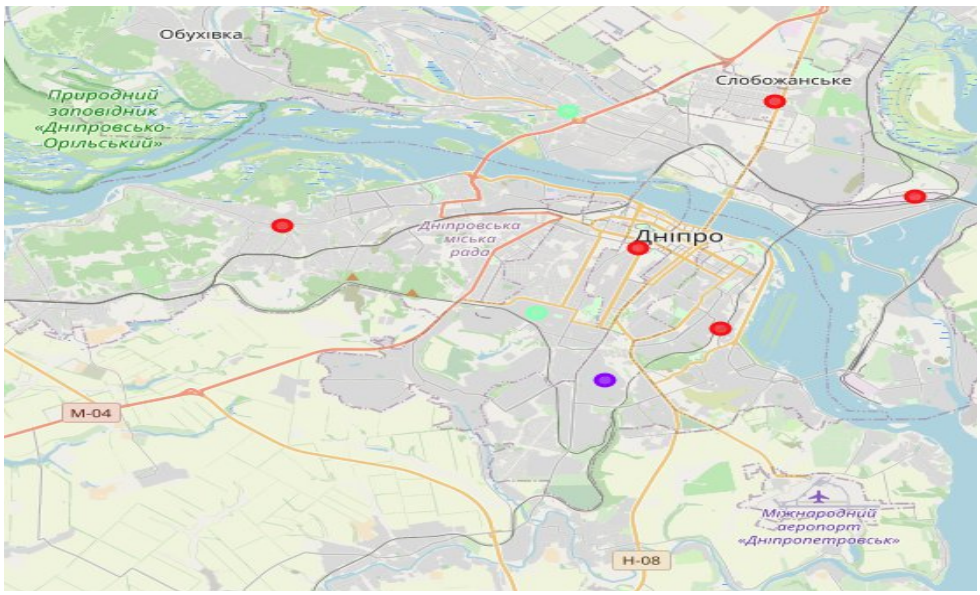
Picture 2. The map of Kyiv, Ukraine with clusters



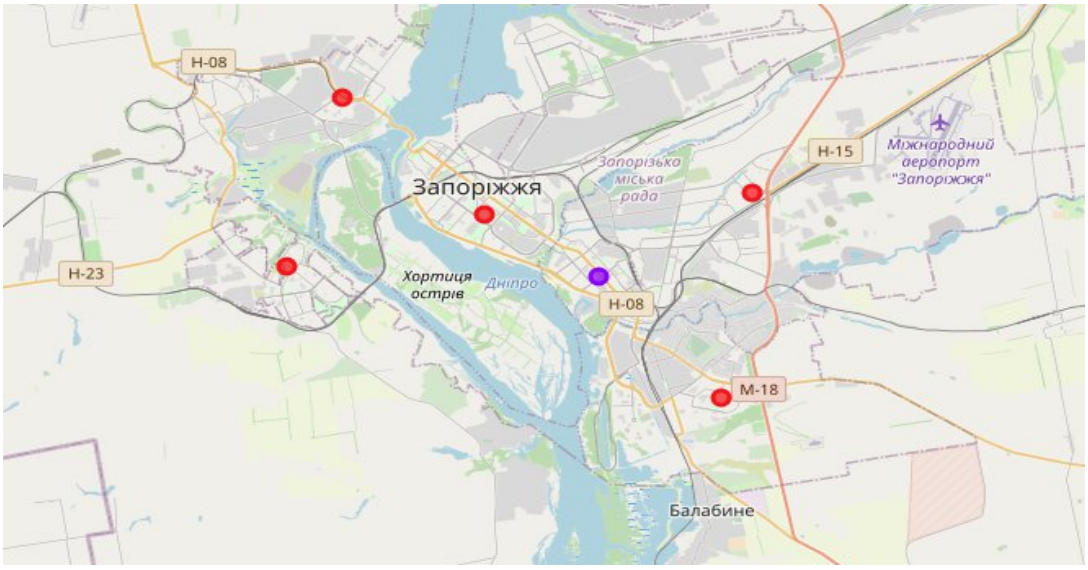
Picture 3. The map of Kharkiv, Ukraine with clusters



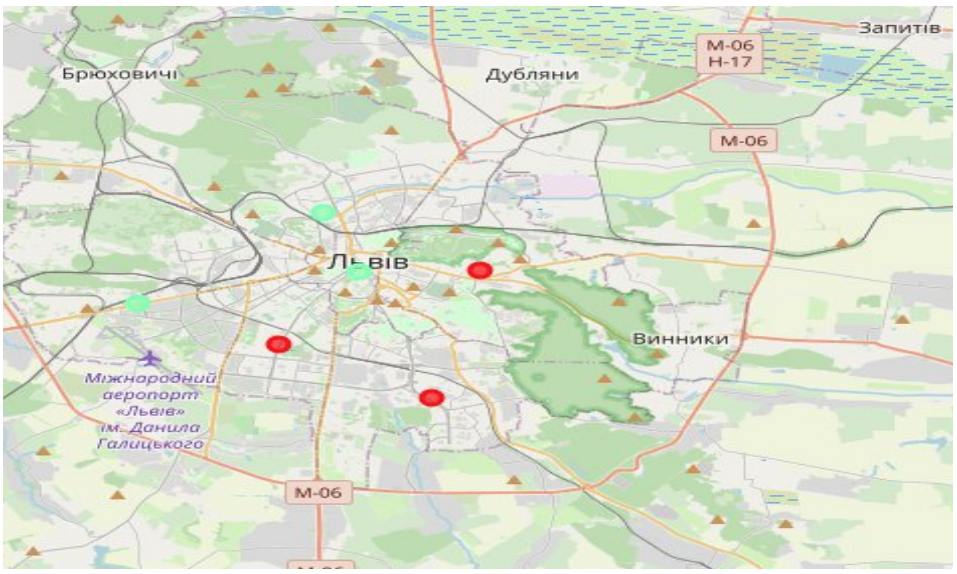
Picture 4. The map of Odessa, Ukraine with clusters



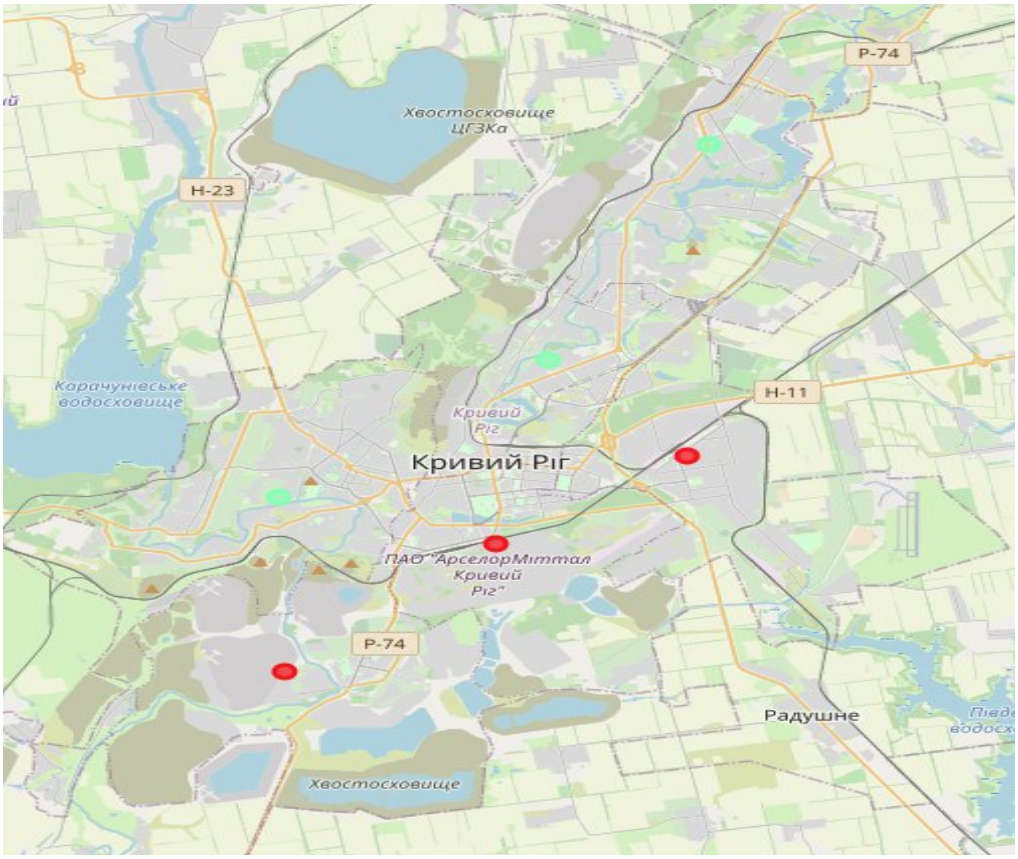
Picture 5. The map of Dnipro, Ukraine with clusters



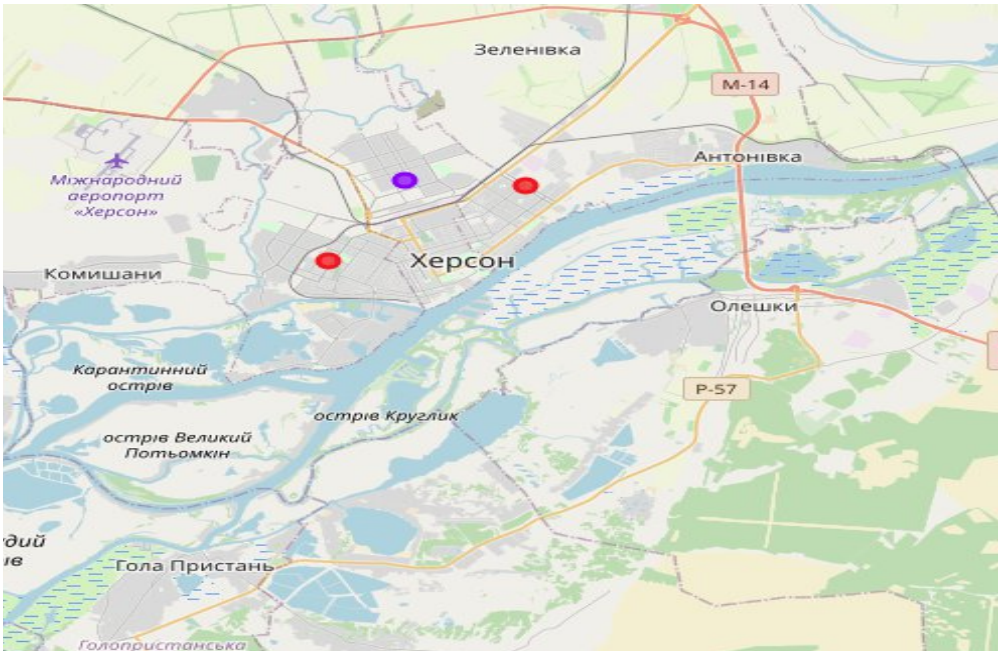
Picture 6. The map of Zaporizhia, Ukraine with clusters



Picture 7. The map of Lviv, Ukraine with clusters



Picture 8. The map of Kryvyi Rih, Ukraine with clusters



Picture 9. The map of Kherson, Ukraine with clusters

4. Discussion

The retail-entertainment centers are situated in the centers of districts.

Most of retail-entertainment centers are presented with one per district, with the highest number in cluster 2. Cluster 1 has a group with high density retail-entertainment centers. The Cluster 0 contains none retail-entertainment centers.

I would recommend to build a new retail-entertainment center in Cluster 0 or Cluster 1. Cluster 0 would a good choice to avoid competition, while Cluster 1 would provide a competition with stable revenue. The competition in Cluster 2 is bad idea, because you just drop an income from retail-entertainment center.

5. Conclusion

I have gone through whole process of submitting capstone project and how would look like the real project from data scientist.

I have gone through several steps, such as:

- identifying the business problem
- specifying the data required
- processing data and preparing the data
- performing machine learning by clustering
- reporting my recommendations to the relevant stakeholders (investors).