Venues of Cracow: Main square vs Kazimierz District

Piotr Waga, 15.05.2020r.

1. Introduction

Cracow is a city with a thousand-year history, the former seat of Polish kings and the capital of the country, and today one of the most important European metropolises. The Cracow Old Town is a unique treasury of works of art, historical souvenirs and monuments that represent almost all architectural styles - from the Middle Ages to modern times.

The heart of the city for hundreds of years is the Main Market Square - the largest urban square in medieval Europe, preserved unchanged since 1257 and inscribed on the first UNESCO World Heritage List in 1978. The bell tower of St. Mary's Basilica has been playing every hour for 600 years . On the other hand, the Sukiennice - located in the middle of the market square - a medieval market hall - are one of the most recognizable Polish monuments.

In this project an analysis will be carried out considering two very popular neighborhoods of Cracow, which are rich in venues:

- **Main Square**
- **Kazimierz District**

Conducted comparison will be aimed at finding features unique for each districts, similarities and leads for investors planning to open new venues in one of mentioned districts.

2. Data

The data needed for this analysis will be acquired through Foursquare developers API, and from Google Maps. The data set will consist of:

- Districts latitudes and longitudes
- 100 venues for each district, and for each venue:
    - Latitude and longitude
    - Name and Foursquare id
    - Category
    - Rating
    - Count of tips, likes, and photos

The above data will be used to compare the districts, in regard of features of it's venues. Below is presented the final Data Frame which will be used for analysis:

	District	id	Venues	Venue Latitude	Venue Longitude	Venue Category	Verified	Rating	Tips Count	Likes Count	Photos Count
0	Kazimierz	56193c32498e5b08a853b2e6	Youmiko Sushi	50.050314	19.943130	Sushi Restaurant	False	9.5	89	237	102
1	Kazimierz	55ce5da3498e5b80f82f9afc	BARaWINO	50.048434	19.944891	Wine Bar	False	9.2	13	67	21
2	Kazimierz	5639c271cd1086837a2fea42	Nolio	50.049356	19.942901	Pizza Place	False	9.0	77	251	50
3	Kazimierz	5283cd6411d296aa84e3d8d7	Bottiglieria 1881	50.048738	19.946163	Italian Restaurant	False	8.7	5	22	11
4	Kazimierz	57262e7c498e0c9c72558bb5	Good Lood	50.048777	19.944739	Ice Cream Shop	False	8.6	34	109	38
...	...	...	...	...	...	...	...	...	...	...	...
195	Main square	55448431498e731178b9e24b	Little Havana Hostel	50.061960	19.934831	Hostel	False	7.6	3	10	6
196	Main square	4fb68199e4b0204f4d42f2e0	Muzeum Farmacji	50.063268	19.940134	History Museum	False	7.5	1	8	7
197	Main square	4cdda698db1254810bab2bce	CK Dezerter	50.060336	19.936326	Eastern European Restaurant	False	7.4	6	26	28
198	Main square	51802019498ef82bbe68129f	Tiffany Ice cream	50.063795	19.935643	Ice Cream Shop	False	7.8	23	50	30
199	Main square	4db6deca5da3a76f446edef3	Szara Resto&Bar	50.061323	19.938078	Restaurant	False	7.4	31	54	123

(Fig 1. Final Data Frame for analysis)

This data set was created by merging two Data Frames:

1. Venues, which was created with data acquired through **explore** endpoint of Foursquare API. Data consisted of venues: id, name, latitude, longitude and category.
2. Venues Details, which was created with data acquired through **details** endpoint of Foursquare API. Data consisted of venues: id, ratings, tips count, likes count and photos count.

3. Methodology

First action carried out on data was finding the most popular categories of venues for each district. Preparation of data consisted of splitting Data Frame in two (one for each district). Next by grouping and counting each Data Frame by venue's category the new Data Frames was created which were used to plot data on bar graphs (Fig 2). Condition was introduced for ploting – that count of category must be above or equal to 4.

Second action carried out on data was descriptive statistics, which provided informations about means, standard deviation, min and max values for each features on splitted Data Frames (Fig 3 and Fig 4).

Third and last analysis carried out was regression. Aim of this action was attempt to find if there is a correlation between rating of venues (target, depended value), and other details of the venues – tips count, likes count and photo count (independent features).

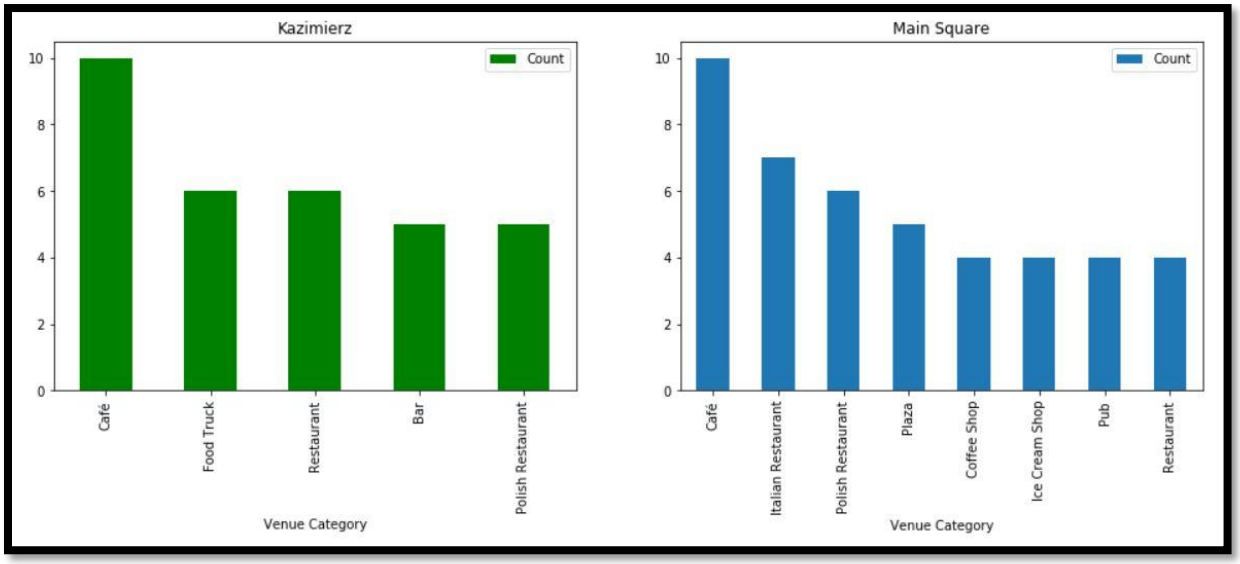
Regression plot was created for each of the combination of target and independent feature, to see if there is a visible linear correlations (Fig 4 - 9).

Data was splitted into training set (80%) and test set (20%).

Simple Linear Model was fitted with one of the combination of target and independent feature. Evaluation of the model with R2 score metrics was also conducted to check accuracy of the model. Also, for the same combination data was transformed and Polynomial Regression was carried out on it. Model was also evaluated with R2 score metric.

4. Results

4.1. Most popular categories of venues



(Fig 2. Most popular categories of venues)

Result of counting most popular categories of venues of each district is presented on Figure 2.

Most popular category of venues for both districts is Café, with count of 10 for each. Second most popular category for Kazimierz districts are Food Trucks and Restaurants (without specified category) with count of 6 each, and Main Square: Italian Restaurant (7 counts) and Polish Restaurant (6 counts).

4.2. Descriptive statistics

The results of descriptive statistics are two tables, gathering information on datasets for each district, respectively Fig 3 for Kazimierz District, and Fig 4 for Main Square.

Mean rating is very similar for both districts, 8.037 for Kazimierz and 8.068 for Main Square.

Mean counts for tips, likes and photos are bigger for Main Square venues. Biggest difference is noticeable in photos - mean count for Kazimierz is 53,21 photos per venue while for Main Square it is 158,40 photos per venue, which is 3 times more. Also worth noticing is maximum number of photos for one venue in Main Square which is 5546 compared to 404 for Kazimierz – ratio over 13 in favor of Main Square.

	Venue Latitude	Venue Longitude	Rating	Tips Count	Likes Count	Photos Count
count	100.000000	100.000000	100.000000	100.000000	100.000000	100.00000
mean	50.050685	19.945580	8.037000	27.730000	81.630000	53.21000
std	0.001646	0.002124	0.554441	37.447298	111.967136	68.33892
min	50.047327	19.940641	6.900000	0.000000	6.000000	0.00000
25%	50.049296	19.944371	7.600000	5.750000	16.750000	13.00000
50%	50.050824	19.945645	8.100000	14.500000	44.500000	27.50000
75%	50.051822	19.947206	8.400000	32.250000	78.500000	63.25000
max	50.053440	19.950407	9.500000	213.000000	700.000000	404.00000

(Fig 3. Descriptive statistics for Kazimierz District)

	Venue Latitude	Venue Longitude	Rating	Tips Count	Likes Count	Photos Count
count	100.000000	100.000000	100.000000	100.00000	100.000000	100.000000
mean	50.061892	19.938012	8.068000	40.09000	134.140000	158.400000
std	0.001325	0.002168	0.434818	57.52612	231.843252	564.601782
min	50.059121	19.933834	7.400000	0.00000	3.000000	2.000000
25%	50.060873	19.936209	7.700000	6.00000	21.250000	20.750000
50%	50.061893	19.938046	8.000000	17.00000	52.000000	49.000000
75%	50.063055	19.939716	8.300000	50.25000	159.250000	117.000000
max	50.064213	19.942660	9.200000	365.00000	1742.000000	5546.000000

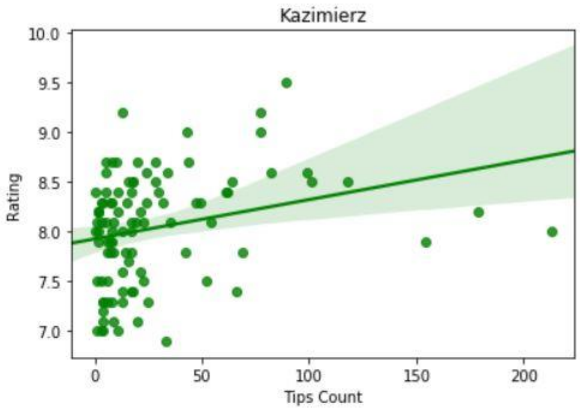
(Fig 3. Descriptive statistics for Main Square)

4.3. Linear and polynomial regression.

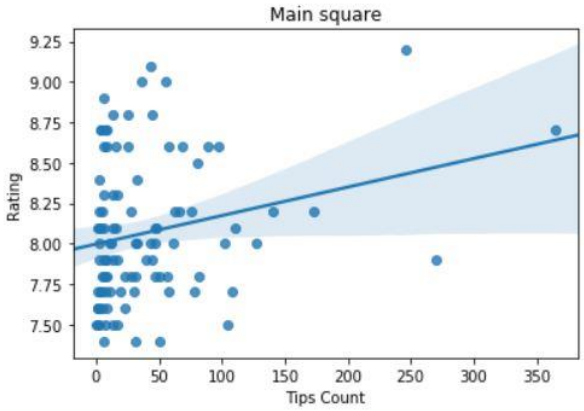
Results of ploting data on regplots is shown in Figures 4 – 9. While plots looks similar to each other, there no visible linear correlation between any combination of target and independent features.

R2 score for Simple Linear Model, trained on data where independent value was Tips Count from Kazimierz district was equal to R2 = -6,8212363524512.

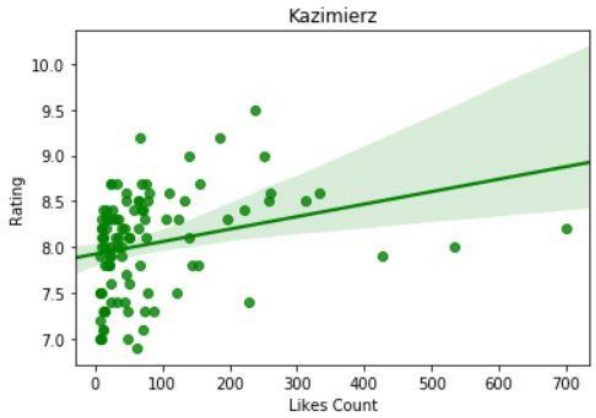
R2 score for Polynomial Regression, trained on same data and with 6 different degrees is presented on Figure 10.



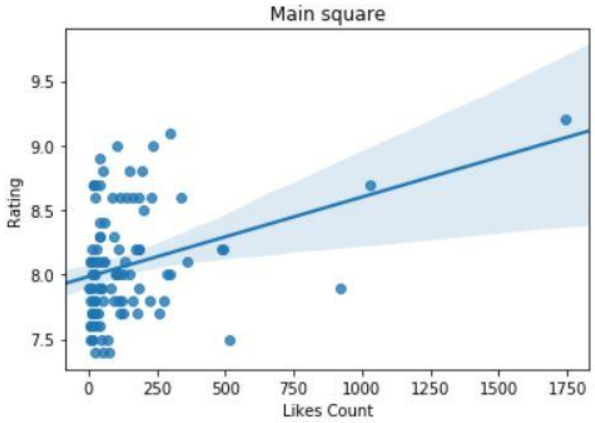
(Fig 4. Regplot for Tips Count, Kazimierz)



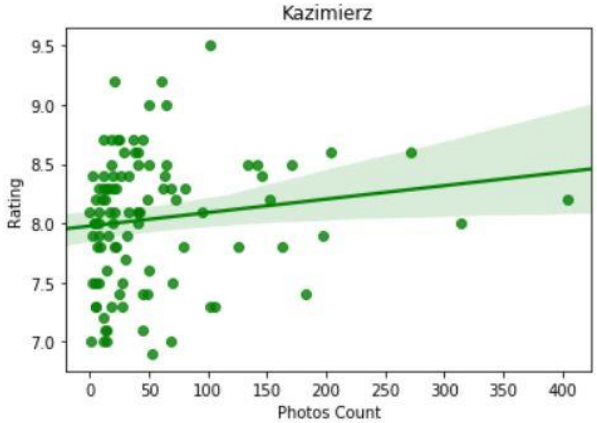
(Fig 5. Regplot for Tips Count, Main square)



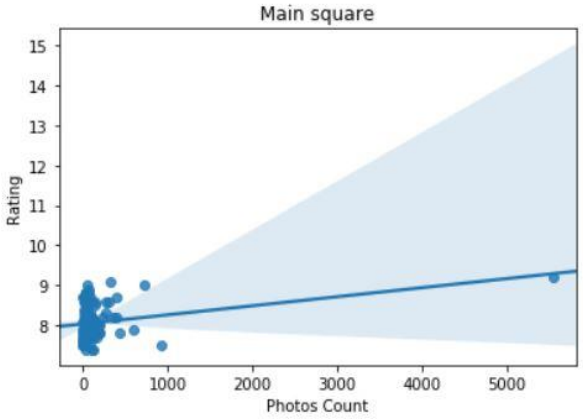
(Fig 6. Regplot for Likes Count, Kazimierz)



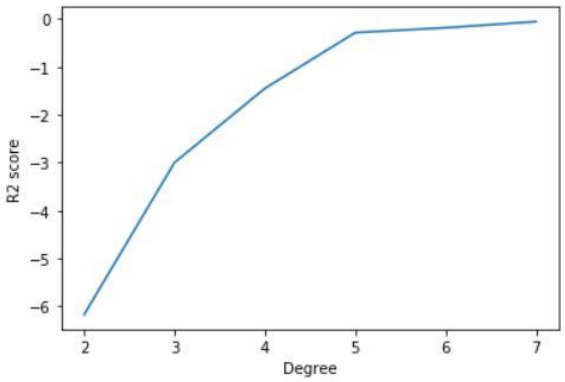
(Fig 7. Regplot for Likes Count, Main square)



(Fig 8. Regplot for Photos Count, Kazimierz)



(Fig 9. Regplot for Photos Count, Main square)



(Fig 10. Degrees of Polynomial Regression with related R2 score, for Photos Count, Kazimierz)

5. Discussion and conclusion.

The report showed similarities as well as differences between Main Square and Kazimierz District in Cracow, in terms of venues.

First similarity is the most popular category for venues in both districts which is Cafe. Next in order of popularity are Food Trucks and Restaurants for Kazimierz district, and Italian Restaurants with Polish Restaurants for Main Square. This difference can be explained by the fact that they Main Square is more, and probably the most, popular place for tourists visiting Cracow. Tourists might like to eat in place that they know or like (Italian pizza e.g.) or try something local hence popularity of Polish Restaurant on Main Square. On the other hand, Kazimierz District, while also relatively popular among tourist, might be better choice for local people to hang out and trying to avoid crowds of tourists, so that might be connected to more „standard" restaurants and food trucks.

Second similarity is mean of ratings of venues for both districts, which is around 8,0 for both. This indicates that venues in both districts are on similar (and high) level.

Count of tips, likes and photos per venues differs between districts. Biggest difference is noticeable in photos - average count for Kazimierz is 53,21 photos per venue while for Main Square it is 158,40 photos per venue, which is 3 times more. Also worth noticing is maximum number of photos for one venue in Main Square which is 5546 compared to 404 for Kazimierz – ratio over 13 in favor of Main Square. This difference might also be connected with fact that Main Square is more popular among tourists who takes a lot of photos, while Kazimierz District is also often visited by residents of Cracow, who might be less likes to make photos of their local venue.

Last analysis of report was linear and polynomial regression, which provided insight that there is no correlation between count of tips, likes and photos of venues, and theirs overall ratings in both districts. For chosen combination of targets and independent values, evaluation was conducted resulting in accuracy of the model in form on R2 score equals -6.8212363524512 which indicates absolutely no correlation between target and independent features (best accuracy is R2 score around 1,0 and slightly below).

Transforming features, in order to conduct polynomial regression, also resulted in negative values of R2 score, which indicates no correlation between target and independent features. While with increasing value for poly degrees, value of R2 score also seems to increase (Fig 10.), this only leads to overfitting of the model with higher values.