IBM Data Science Capstone Project

INTRODUCTION

FINDING A NEW HOME

- In many cases when a person must move to another city or even country, she knows nothing about the new location, and it becomes extremely difficult to choose optimal place to search for a new home/flat.
- At the same time, this person has some preferences about what she does and doesn't want to have near her future home, alongside with some restrictions and limitations, for example – budget limitations.
- The system that was developed during this capstone projects is designed to help solve the stated problem based on such parameters.

DATA

ACUISITION
CLEANING
TRANSFORMATION

DATA SOURCES

- BetaNYC portal (data.beta.nyc):

  Information about New York Census Tracts;

  Geoshapes of each Tract;

- Foursquare API:

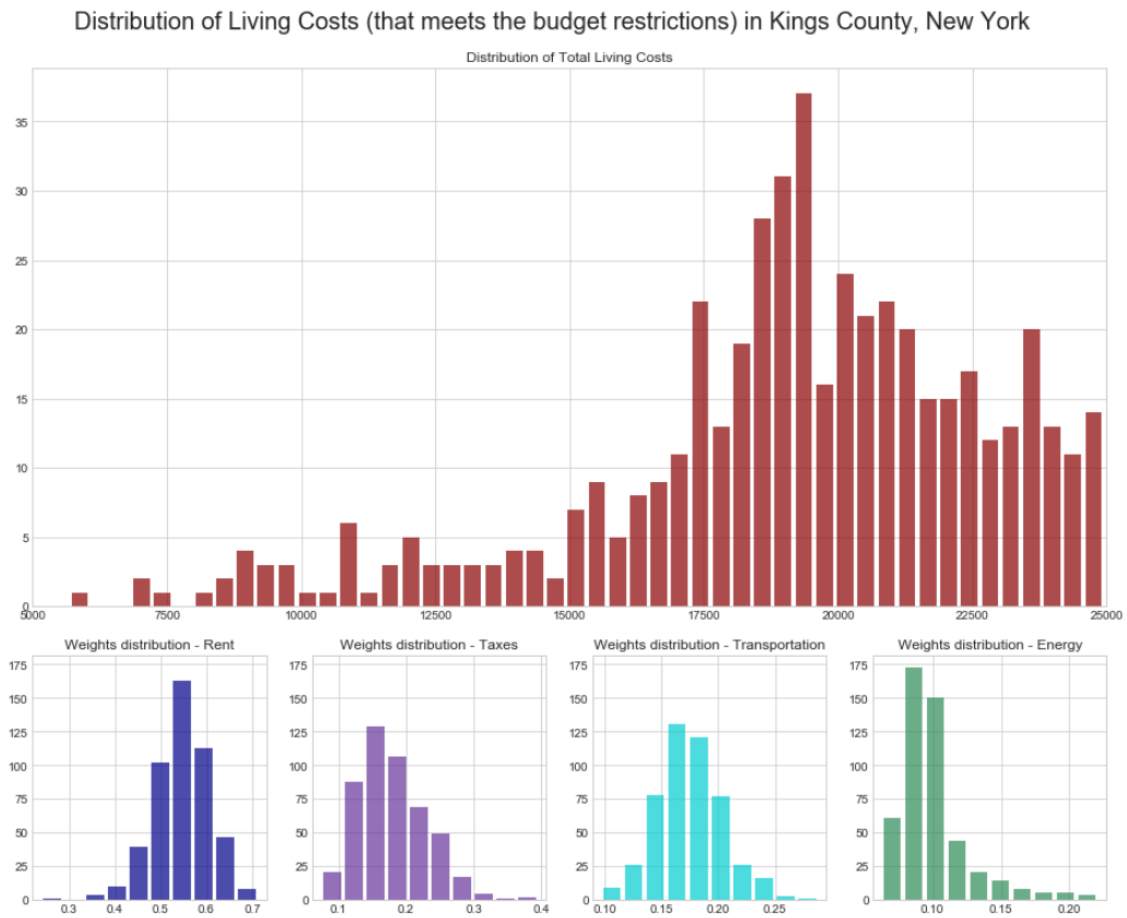
  Information about venues in each Census Tract

- Venue preferences of a made-up family that must relocate from Torrington to New York

FINDING A NEW HOME

DATA CLEANING and TRANSFORMATION

- Removed location records that had zero costs;
- Removed records that didn't satisfy budget limitations;
- Combined data from different sources into a single dataframe.



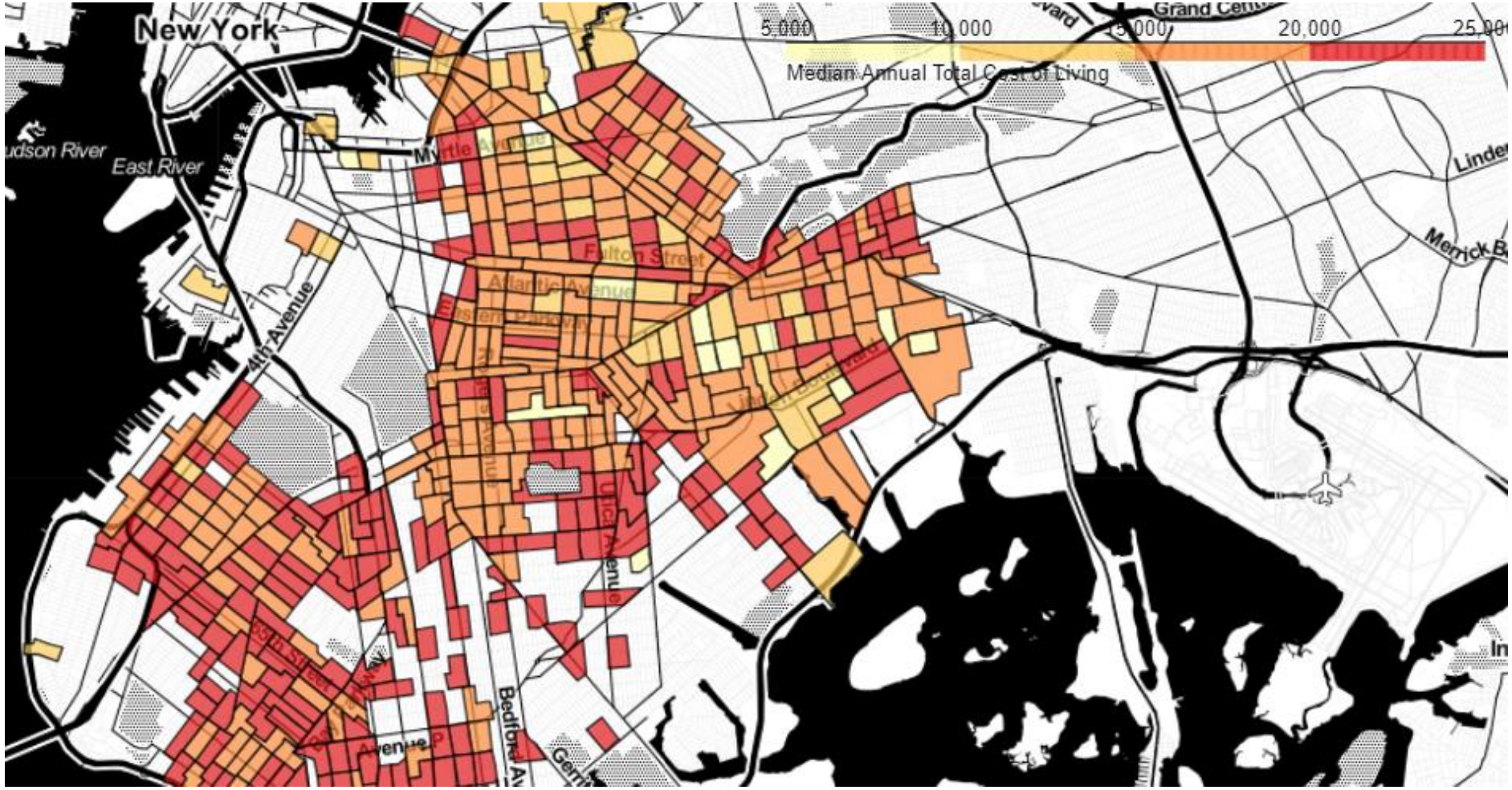
LOCATIONS DISTRIBUTION

By medium Annual Total Cost of Living and by Costs structure

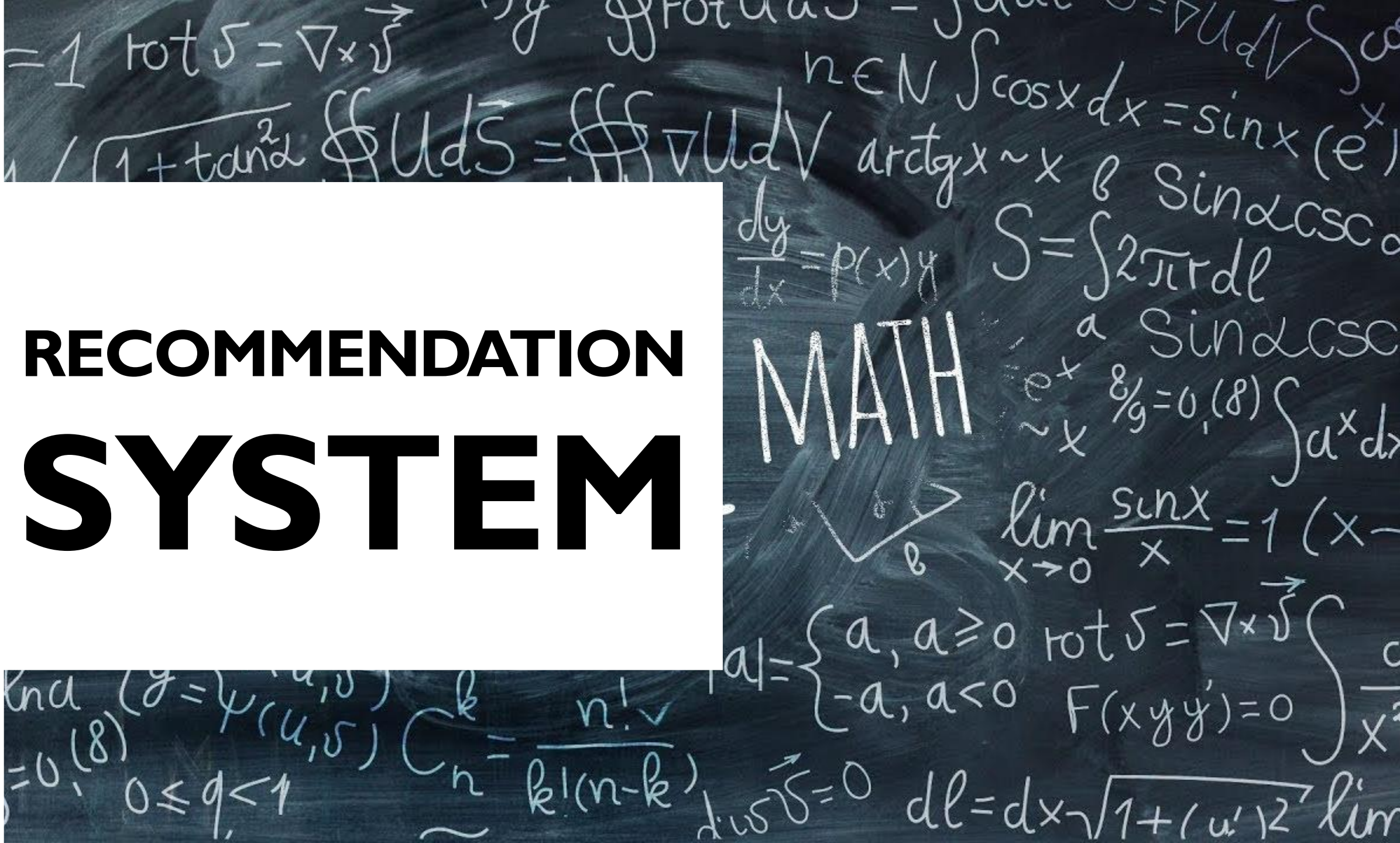
- Most of the Tracts that meet budget limitations have annual cost of living from $15,00 to $25,000
- In majority of Tracts more than 50% of total expenses fall on Rent payments

KINGS COUNTY MAP

COLOR: MEDIUM ANNUAL TOTAL COST OF LIVING



RECOMMENDATION SYSTEM



FINDING A NEW HOME

SYSTEM MECHANICS

VECTOR MULTIPLICATION BETWEEN USER PREFERENCES AND LOCATION CHARACTERISTICS

	Score
Café & Restaurant	6
Infrastructure	8
Parks, landscape and POV	6
Pet Care	5
Health, Sport and Beauty	8
Bars and Clubs	-8
Tourism	-7

*

Category ID	Café & Restaurant	Infrastructure	Parks, landscape and POV	Pet Care	Health, Sport and Beauty	Bars and Clubs	Tourism
36047000200	14	0	0	1	1	3	1
36047002000	4	2	0	0	2	3	0
36047002200	3	1	1	0	0	1	0
36047002300	12	1	8	0	9	6	1
36047002901	4	2	4	0	8	2	0

ID	Score	Café & Restaurant	Infrastructure	Parks, landscape and POV	Pet Care	Health, Sport and Beauty	Bars and Clubs	Tourism	Total Costs	Latitude	Longitude
36047010600	214	33	0	0	0	3	1	0	18123.882368	40.638875	-74.006040
36047079801	199	26	1	1	1	7	4	0	18803.643250	40.660011	-73.958719
36047029000	180	26	1	0	0	2	0	0	20452.762195	40.603174	-73.994348
36047032700	171	23	1	2	1	5	4	0	17782.949692	40.662647	-73.958796

RESULTS

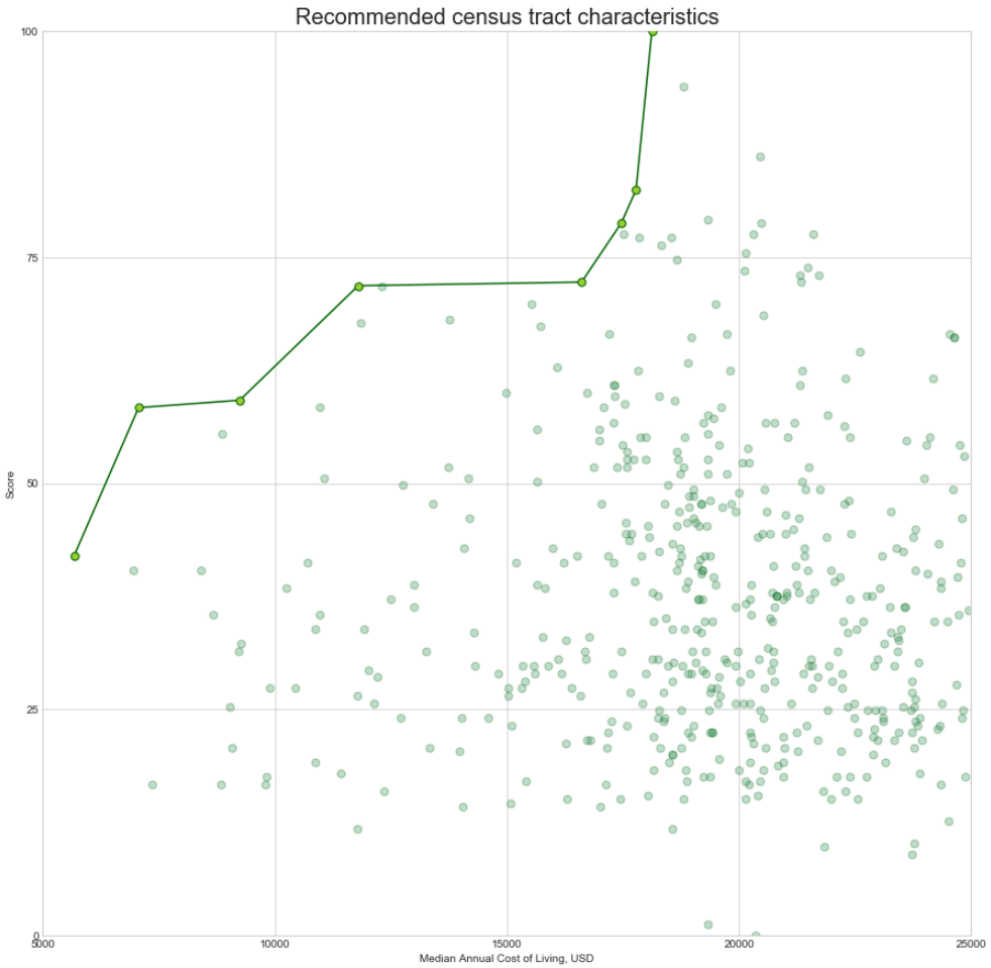
SCORING OPTIONS

1 ABSOLUTE SCORE

If we want to suggest them locations with the highest Score, the only thing that we need to do is to take top N rows of the result dataframe

2 OPTIMAL SCORE

Choose locations that provide best score for a given sum of money



OPTIMAL LOCATIONS

SUBTITLE HERE

- Points located bellow the line are possibly less attractive
- Why? Because for each of these options exists another one that provides higher score or/and lower annual cost.
- Why 'possibly'? Because this is only score representation of computed 'attractiveness' of each recommended location, in the real life it could be the situation that a location with a lower score would be much more interesting to the final Customer.

KINGS COUNTY MAP

COLOR: TRACT SCORE, YELLOW DOTS: OPTIMAL LOCATIONS, BLUE DOTS: TOP 20 BY SCORE

CONCLUSION AND FUTURE DIRECTIONS

FINDING A NEW HOME

- Although the implemented model is very simple, it already helped us to narrow down list of possible locations for a new home from almost 500 to 20 or even 8 options;
- Recommendation system could be further improved, for example, by extracting the category scores from the list of neighborhoods/tracts with their ratings provided by the customer/user;
- We made an assumption that the number of desirable/unwanted venues equally important as their scores. The system could be improved by introducing some kind of weight coefficients for each additional venue of the same type;

SERGEI MERSON

sergei..merson@gmail.com

THANK YOU