Impact of Apartment`s storey on it`s Price, Prediction of the Price for Apartment & Analysis of Venues in the Boroughs of the city Nur-Sultan.

Asset Alkhanov

June 5, 2020

1. A description of the problem and Discussion of the background

Astana(Nur-Sultan) is a capital of Kazakhstan. It is one of the biggest cities in Kazakhstan with population over than a million people. I decided to make this analysis due to request of my friend, who is real estate seller and lives in this city. Also the most interesting thing for him was an impact of floor number to the price of the flat. People who want to sell their flats quite often ask him to estimate the flat and suggest a reasonable price for the flat in order to publish the flat info and price in the advertisement. The price for the flat should be calculated very carefully, usually it shouldn't be too high and not too low. My friend has to consider lot's of similar supplies in the market and by taking into account all the characteristics and other factors related to the flat, he has to produce a sensible price for it. The process of giving a price to the flat is very time consuming. Therefore I came up with an idea to help to my fiend to predict a price for the flat using some Data Science tools. Imported all necessary python libraries. Scraping data from a flat advertising agency's website using Beautiful soup

2. A description of the data and how it will be used to solve the problem

To get data about flats in market, I did the following following research:
I have analyzed all the websites and online systems for selling and advertising flats. I have chosen the most popular web service for selling flats among customers in Kazakhstan. The website wasn't having any API, therefore I decided to scrape the data from it. I cleaned the data and retrieved all necessary information about each flat in the market of Nur-Sultan(Astana), though couldn't retrieve location data of each flat from the website. The data I scraped from the website was about building built year, borough, floor number, price, address and square meters. I made decision to analyze flats with very similar characteristics and chose only 2 - bedroom flats starting from 57 to 62 square meters. I picked flats built from 1990 year, I wanted to retrieve more data on tall buildings as much as possible, since houses built until 1990 weren't tall.

I used Foursquare API in order to retrieve number of venues in radius of 3000 metres from the center of each borough. I took the center coordinates of each borough from the 2GIS website. There are 4 main boroughs in Astana (Nur-Sultan), Esil is the biggest, Baikonyr, Almaty and Saryarka is the smallest borough in the city.

I couldn't find border coordinates for each borough, therefore I decided to draw a map of boroughs myself in the www.geojson.io, once I finished drawing the borders for the boroughs, the website automatically generated coordinates of the borders. I used the coordinates to draw a Choropleth map, which displayed number of venues in each borough. I Cleaned all the scraped data for a better processing and readability. Finally I designed a dictionary out of the cleaned data to create a dataframe.

	Title	Borough	Address	Price	Built year	Sqrms	Flat Floor	Building max floor
0	2-комнатная квартира	Алматы р-н	Нажимеденова	27400000	2018	60.0	12	13
1	2-комнатная квартира	Есиль р-н	Е16 2	19000000	2016	58.0	9	9
2	2-комнатная квартира	Есиль р-н	Керей-жанибек хана 9	19000000	2006	60.0	2	9
3	2-комнатная квартира	Есиль р-н	Мәңгілік Ел 48	27000000	2015	61.0	8	8
4	2-комнатная квартира	Есиль р-н	Акмешит 7 — Ханов Керея и Жанибека	22500000	2010	59.0	1	9
...	...	...	...	...	...	...	...	...
598	2-комнатная квартира	NA	Бараева — Иманбаевой	25000000	2002	59.0	3	4
599	2-комнатная квартира	Есиль р-н	Керей и Жанибек хандар 9	21000000	2008	58.9	5	9
600	2-комнатная квартира	Есиль р-н	К. Мухамедханова 12	20500000	2018	59.0	4	10
601	2-комнатная квартира	Алматы р-н	Кенена Азербаева 12	18000000	2017	58.0	4	5
602	2-комнатная квартира	Сарыарка р-н	Акан серы 16 — Тлендиева	16500000	2016	59.0	9	13

3. Methodology

3.1 Data cleaning and filtering

In order to avoid errors related with non numeric values, I removed all 'NA' values from the fields 'Flat Floor' and 'Borough', it left me with 533 records. Also, I picked the flats with price less than 100 million, since it's very unreal to find a flat that might cost more than 100 million tenge(Kazakhstan currency) for 2 bedroom flat with 57-62 squared meters, however sometimes you might find unique cases, though it might be just incorrectly entered data. All in all, I sorted the data frame by 'Price' field in descending order. When printing first 10 records of the data frame, it becomes obvious that the 10 out of 10 most expensive flats are in the 'Есиль р-н'('Esil' in latin) borough. The sorted data frame also shows that, in the top ten expensive flats there is only one flat located on the floor higher than five and it is floor 18.

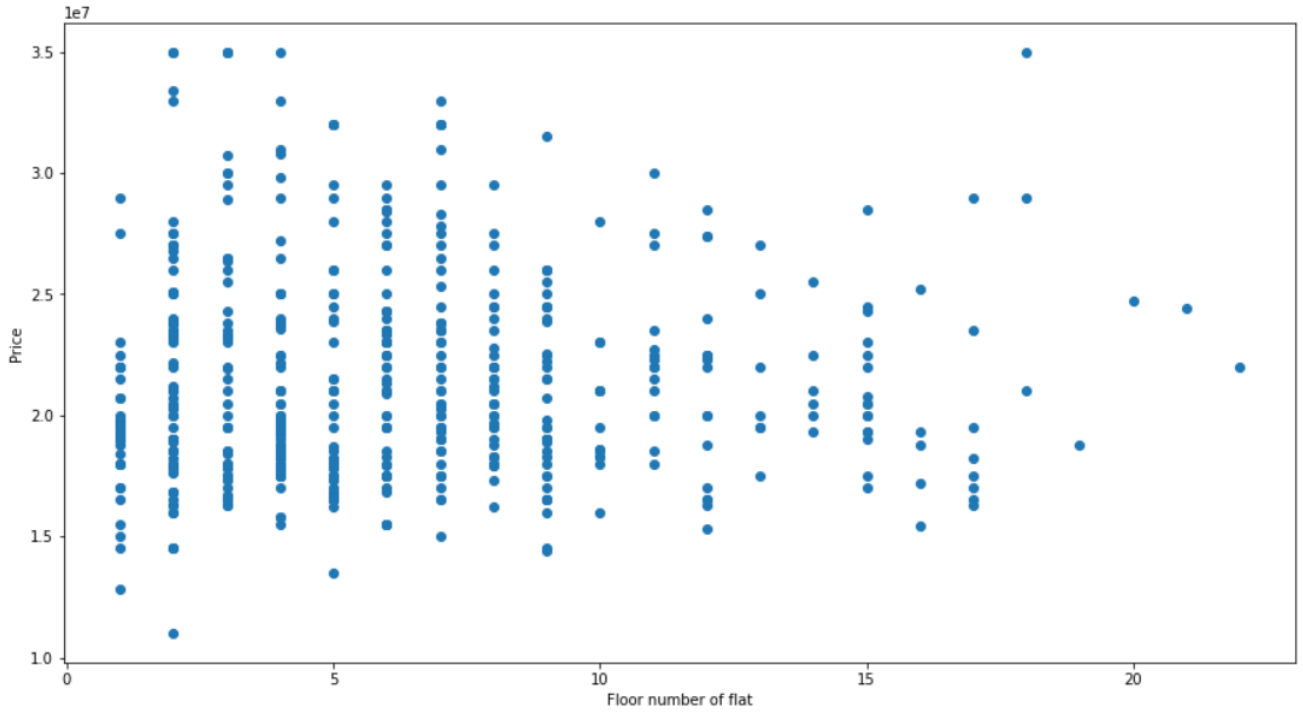
	Title	Borough	Address	Price	Built year	Sqrms	Flat Floor	Building max floor	Borough ID
0	2-комнатная квартира	Есиль р-н	Орынбор — проспект Мангилик Ел	35000000	2016	62.0	4	7	4
1	2-комнатная квартира	Есиль р-н	Сауран 10Б	35000000	2018	60.0	18	18	4
2	2-комнатная квартира	Есиль р-н	Туркестан 18	35000000	2018	61.0	3	12	4
3	2-комнатная квартира	Есиль р-н	Алматы 11 — Туркестан	35000000	2017	60.3	2	10	4
4	2-комнатная квартира	Есиль р-н	Алматы 11 — Туркестан	35000000	2017	60.3	2	10	4
5	2-комнатная квартира	Есиль р-н	Орынбор — Акниет	35000000	2017	61.1	3	14	4
6	2-комнатная квартира	Есиль р-н	проспект Кабанбай Батыра 13/5	35000000	2016	60.0	3	8	4
7	2-комнатная квартира	Есиль р-н	Кабанбай Батыра	33400000	2017	58.0	2	8	4
8	2-комнатная квартира	Есиль р-н	Кабанбай Батыра 58Б — Улы Дала - Сауран	33000000	2017	59.0	2	8	4
9	2-комнатная квартира	Есиль р-н	38-я улица 23	33000000	2018	62.0	4	7	4

The top 10 cheapest flats are illustrated in the dataframe below. All the cheapest flats are mainly located in two boroughs, 'Алматы р-н'('Almaty' in Latin) and 'Сарыарка р-н'('Saryarka' in Latin). There are five flats in 'Almaty' and two flats in each 'Saryarka' and 'Baikonyr' boroughs.

	Title	Borough	Address	Price	Built year	Sqrms	Flat Floor	Building max floor	Borough ID
545	2-комнатная квартира	р-н Байконур	Асан Кайгы — Гумар Караша	15000000	2013	60.0	7	25	3
546	2-комнатная квартира	Алматы р-н	Кордай	14550000	2012	57.0	2	14	1
547	2-комнатная квартира	Алматы р-н	Кордай	14550000	2012	57.0	2	14	1
548	2-комнатная квартира	Алматы р-н	Кордай	14550000	2012	57.0	2	14	1
549	2-комнатная квартира	Сарыарка р-н	189-ая улица	14500000	2013	60.0	9	9	2
550	2-комнатная квартира	Алматы р-н	Жабаева 12/2	14500000	2012	57.1	1	6	1
551	2-комнатная квартира	Сарыарка р-н	189 улица	14400000	2013	61.0	9	9	2
552	2-комнатная квартира	Алматы р-н	Жабаева 12/2 — Тасты	13500000	2010	60.0	5	5	1
553	2-комнатная квартира	Есиль р-н	Лесная поляна 16	12800000	2012	60.6	1	5	4
554	2-комнатная квартира	р-н Байконур	Кусжолы 8	11000000	2099	61.0	2	2	3

3.2 Plotting relationship between 'Price' and 'Floor number of flat'

In order to see the whole picture and relation of the data I collected, I created a scatter plot from the data frame using fields: 'Flat Floor' and 'Price'. I used these fields, cause the 'Flat floor number' was the most interesting characteristic, for my client. The scatter plot below doesn't illustrate neither positive nor negative correlation.



3.3 Applying Linear and Polynomial models

I decided to check whether my data fits to Linear or Polynomial regressions. The two lines below showing that there is only 1% of fitting for Linear Regression and only 3% for Polynomial Regression. It means that I can't predict prices for other flats, cause percentage value in both cases is very low. Hence, my data doesn't fit to both Linear and Polynomial regressions.

Linear Regression fitting %: 0.028512105701604187
Polynomial Regression fitting %: 0.0077226270716241885

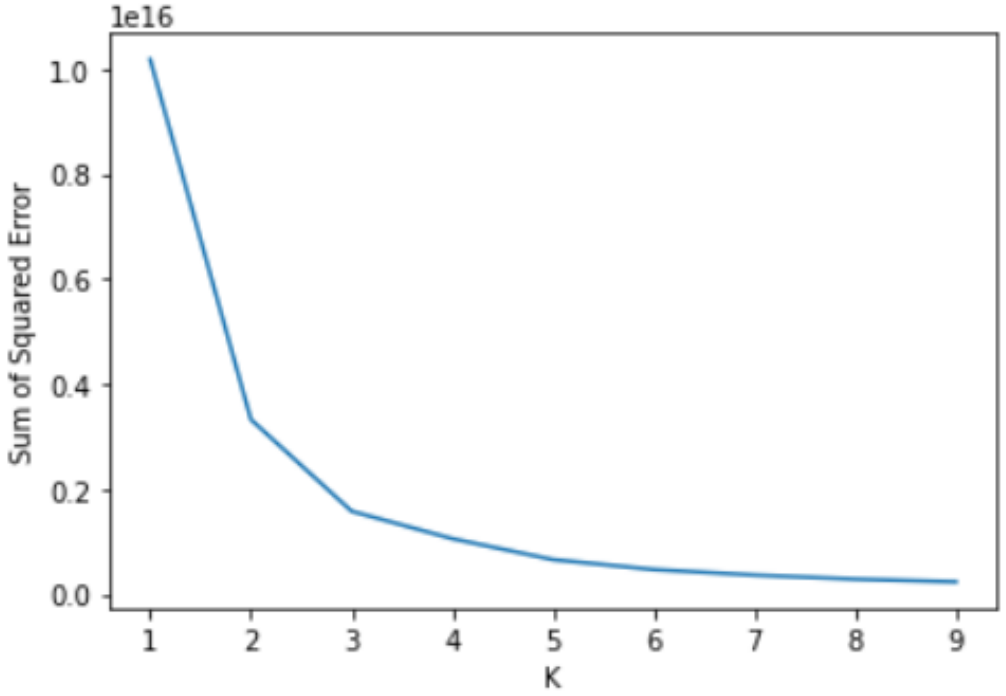
3.4 Predicting a price for a new flat, applying Multiple regression

Since the Linear and Polynomial regressions didn't fit to my dataset, I made decision to apply Multiple regression for predicting price of a flat. Although, I used not only 'Flat floor', but also 'Borough ID', 'Square meters' and 'Building built year' characteristics to predict the price. To test the prediction model, I passed the following values, 'Flat floor' = 6, 'Borough ID' = 2 (Saryarka), 'Square meters' = 61 and 'Building built year' = 2016. The predicted price seemed quite reasonable.

Predicted price of a flat is: 21092984.48722668

3.5 KMeans clustering

In order to divide the data into different clusters I decided to use KMeans clustering. In order to decide number of clusters needed to cluster my data, I applied 'Elbow' method. The first step was calculation of Sum of Squared Error in order to implement the 'Elbow' method. By means of plotting relation of K to SSE we can see an 'elbow', the line breaks when K is 2 and 3, we now can see that number of clusters should be 2 or 3. I decided to stick with 3 for my data.



The next step was applying KMeans object with 3 clusters. Afterwards, I did fitting and predicting clusters for 'Flat floor' and 'Price' vaues in the dataframe. The result of clustering the dataframe can be seen below.

I created a new column 'cluster' and added all the predicted clusters in the data frame. All the flats become clustered now. The dataframe below demonstrates the result.

	Title	Borough	Address	Price	Built year	Sqrms	Flat Floor	Building max floor	Borough ID	cluster
0	2-комнатная квартира	Есиль р-н	Орынбор — проспект Мангилик Ел	35000000	2016	62.0	4	7	4	0
1	2-комнатная квартира	Есиль р-н	Сауран 10Б	35000000	2018	60.0	18	18	4	0
2	2-комнатная квартира	Есиль р-н	Туркестан 18	35000000	2018	61.0	3	12	4	0
3	2-комнатная квартира	Есиль р-н	Алматы 11 — Туркестан	35000000	2017	60.3	2	10	4	0
4	2-комнатная квартира	Есиль р-н	Алматы 11 — Туркестан	35000000	2017	60.3	2	10	4	0
...	...	...	...	...	...	...	...	...	...	...
550	2-комнатная квартира	Алматы р-н	Жабаева 12/2	14500000	2012	57.1	1	6	1	1
551	2-комнатная квартира	Сарыарка р-н	189 улица	14400000	2013	61.0	9	9	2	1
552	2-комнатная квартира	Алматы р-н	Жабаева 12/2 — Тасты	13500000	2010	60.0	5	5	1	1
553	2-комнатная квартира	Есиль р-н	Лесная поляна 16	12800000	2012	60.6	1	5	4	1
554	2-комнатная квартира	р-н Байконур	Кусжолы 8	11000000	2099	61.0	2	2	3	1

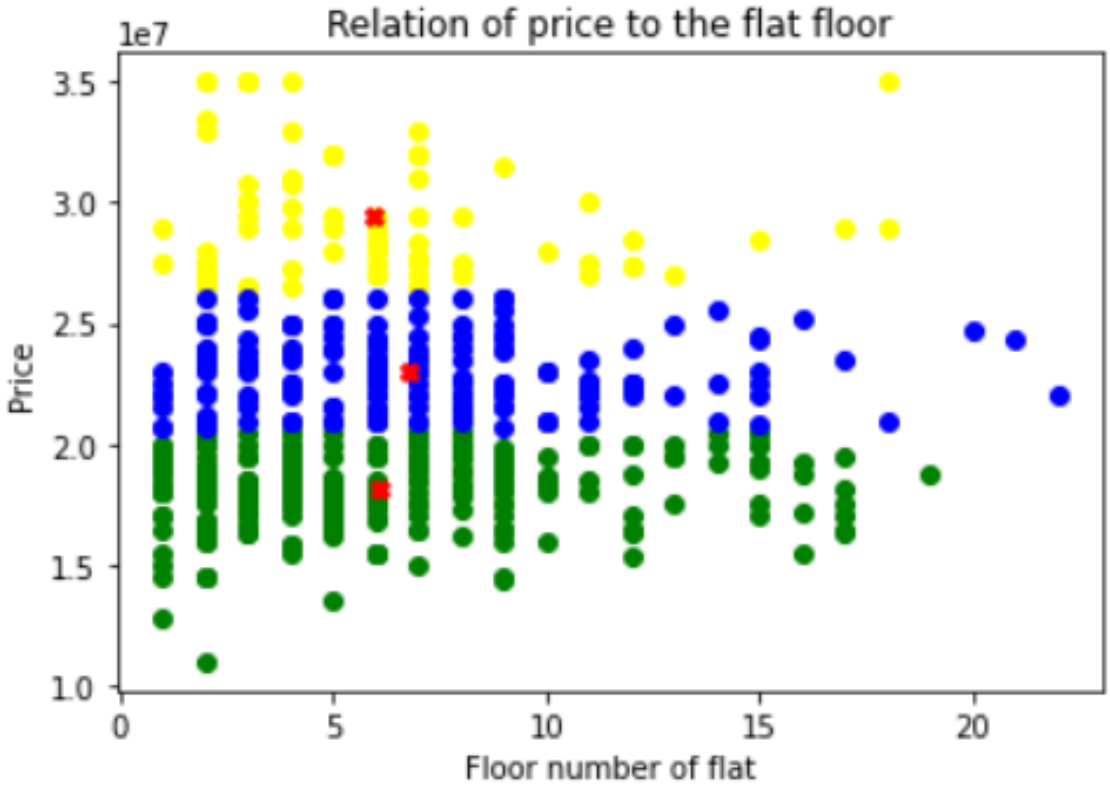
4. Results

4.1 Plotting three clusters on the Scatter plot

I created three data frames and assigned them three clusters from 0 to 2 inclusively. Afterwards, I added the fields 'Flat floor' and 'Price' of all the three data frames in the scatter plot, also I assigned three different colors for each cluster, namely yellow, green and blue. The three red crosses in the scatter plot denote centroids of the clusters. Finally, I added X,Y labels with appropriate text and gave a title for the scatter plot.

The scatter plot illustrates that the majority of the most expensive flats are located below 7th floor. In general, it can be inferenced that the lower the floor the higher the price for the flat. Nevertheless, this is not the only weighty factor impacting on the price of the flat, there are lot's of other reasons.

There is only one outlier flat on the 18th floor for 35 million tenge. It can be treated as a unique case, which is included in the top ten the most expensive flats.



4.2 Retrieving location data using Foursquare API

I retrieved venues data for each borough using Foursquare API. I used other GIS services to find locations for each borough. Afterwards, I applied 'explore' query with radius equal to 3000 meters from the center of each borough. The response was a JSON file, which required parsing the file and retrieving venue and location data. Overall, I created a data frame from a dictionary with borough names, venues, category, distance and locations.

The table below shows only 6 venues for 'Алматы р-н'('Almaty' in Latin), 8 for 'Сарыарка р-н'('Saryarka' in Latin), 30 for 'Есиль р-н'('Esil' in Latin) and 3 in 'р-н Байконур'('Baikonyr' in Latin) boroughs.

One of the reasons for location of the cheapest flats in the boroughs, such as 'Алматы р-н'('Almaty' in Latin) and 'Сарыарка р-н'('Saryarka' in Latin) might be in lack of venues nearby.

	Borough	Venue	Venue category	Distance from borough center(m)	Latitude	Longitude
0	Алматы р-н	KEN MART	Shopping Mall	1580	51.135385	71.573187
1	Алматы р-н	алматы ауданы	Park	1462	51.131819	71.590266
2	Алматы р-н	Хохлома	Bistro	1656	51.135506	71.571580
3	Алматы р-н	зангар	Restaurant	2179	51.148319	71.559158
…	……	……..	……..	…….	……	……
42	Есиль р-н	Massimo Dutti	Clothing Store	2020	51.088404	71.406917
43	Есиль р-н	Marwin	Toy / Game Store	2059	51.089011	71.407534
44	Есиль р-н	Лепим и варим	Pelmeni House	2218	51.089088	71.409821
45	Есиль р-н	Photobar	Photography Studio	2253	51.096841	71.408259
46	Есиль р-н	McDonald's	Burger Joint	2272	51.088626	71.410554

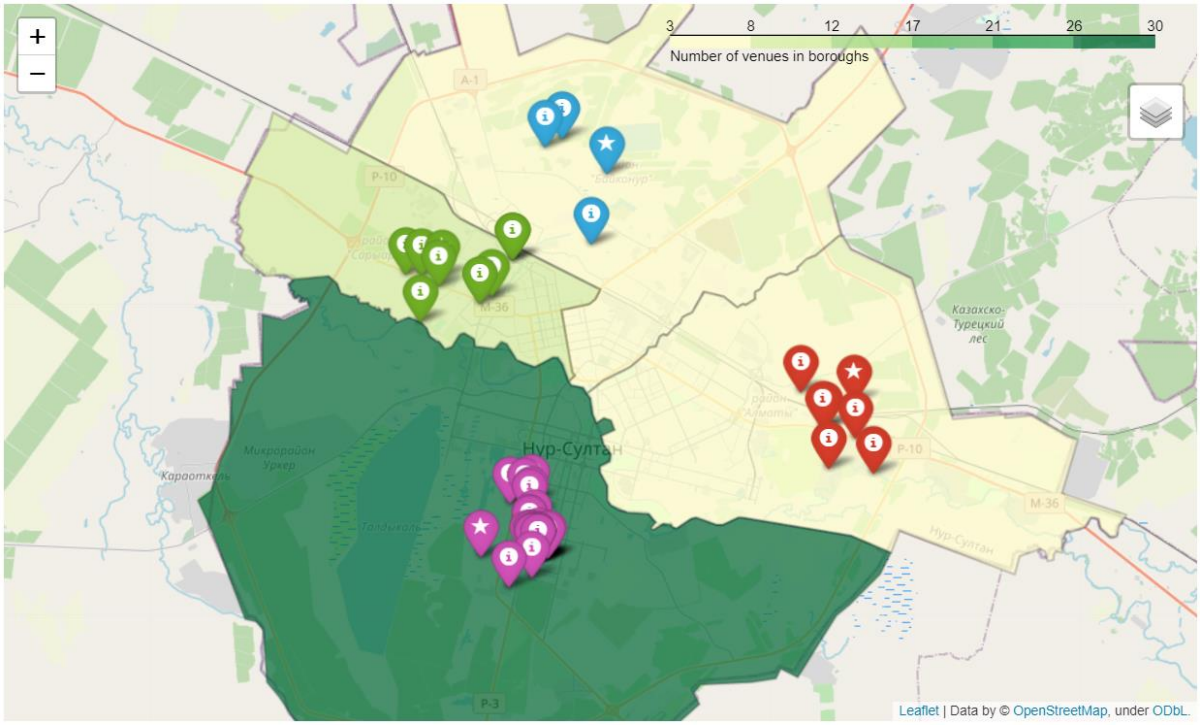
	Borough	Venues
0	Almaty	6
1	Saryarka	8
2	Baikonyr	3
3	Esil	30

4.3 Creating a Choropleth map

Creating a map of Astana(Nur-Sultan) city with it's boroughs and venues using Folium library required lot's of location data, such as latitude and longitude. The markers for boroughs centers and for venues in four different boroughs are colored with four different colors. The borough markers have 'star' icon, while venue markers have 'i' icon.

In order to create a Choropleth map I needed latitudes and longitudes of borders for each borough. I decided to draw the borders using www.geojson.io service, it has functions that can generate coordinates for the drawn polygons. I colored markers in four different colors, red, blue, green and purple.

The Choropleth map below shows, that the most number of close to the borough center venues are located in the 'Esil' borough. In the other three boroughs number of the venues is below ten. A high number of venues in the 'Esil' borough might be one of the reasons for a very high prices for flats in comparison with other boroughs.



5. Discussion

In process of the research I noted that the most expensive flats are located below 5th storey of the building and also that they are in the “Esil” borough. When trying to retrieve location data for each flat I come across with problems. Therefore, I decided to use only venue location data. Finding latitudes and longitudes of the boroughs was easier task; data was available in all GIS online services. Also, I noted that Foursquare API provides maximum 30 venues for a given coordinate when using either explore or search queries.

6. Conclusion

In this study, I analyzed the real estate market of the city Astana(Nur-Sultan) in Kazakhstan. For a better and thorough analysis I used data science tools and methods. My analysis covered relationship between price of the flat and floor number of the flat. In my analysis I collected data on flats with very similar characteristics, since my main aim was to find the flat floor numbers of the top most expensive flats in the city. The analysis revealed that the 90 % of the top ten the most expensive flats are located below 5th floor. Also I retrieved location data for each borough, in order to see if number of venues has an impact on the price of the flat. The hypothesis was true only for the biggest borough of the city, which is 'Esil'. The other boroughs have numbers close to each other, and they all are below 10. I also used predicting models, to predict prices for a new flats, the model that was applied in my research was a Multiple regression. This research can help for people, who want to sell flats or work in real estate market, but not sure about what price they want for their flat. Also it might be helpful for those who want to buy a flat.