Introduction and Business requirement

A. Backgound of the problem:

Opening a new branch of the Bank in Almaty city, Kazakhstan

The majority of international cities like Toronto, New York or other cities have districts called "neighborhoods", these are actually geographical locations of different size areas. In the city where I live division of the city geographically done by seven administritive districts. Each of these districts have its own Governor, hospitals, Universities and almost similiar infustructure with typical houses of former USSR country, Kazakhstan.

Almaty is a previous capital of Kazakhstan with population around two millions of people, located under mountains with lots of facilities and financial organizations. Currently, Almaty is considered to be the financial capital of Kazakstan and the biggest financial hub of Central Asia. As on March 2020, there are 17 local second-tier Bank's officially registered in Kazakhstan and 10 international Banks of different nature: russian banks, chinese banks, korean banks and one american. The Central Bank is also located in Almaty.

A great number of Banks, especially local with a very developed infrastructure (ATM, Branches) create very serious competition to other International Banks. As an example the biggest by assets local Bank " Halyk Bank" has more than 70 branches in Almaty. Total number of Branches, Banks in the city is more than 200, which is a really a lot for the medium size city.

The main business requiement of this project is to identify the best district for opening of new Bank's Branch within same city as Head Office, Almaty. Main requirement of the business is to identify places with high potentials, good developed infrastructure and preferably in the center of city.

The definite location is critically important for the Bank as:

1. **Branch will be separate business unit with decentralized Management powers**
2. **Bank has no ATM's**
3. **New Branch will be providing financial services to SME and individuals**
4. **Retail Business is a priority**
5. **Parking availability for clients**
6. **Better location provides competetive advantage for Retail business**

Why it is important?

We live in the world where financial services bring same importance to our lives as food, internet, transportation. Financial institutions have to be very competetive to ensure best services to the customers in terms of location of their offices, quality of products.

B.Data description

I will try to solve all these problems in this research by using different sources of information (data). As the main part of these research is location data i will use FOURSQARE Api's to extract the coordinates of organizations, cafes, and etc. However, the parsing sources of Google maps is too required as Foursqare is not very popular web in Almaty and to solve real problem some relevant resourses such as OutScraper also will be used. In order to define the population of each district of the city, it's density and other data i will use wikipedia tables as its a really good a renewable sourse of information about city and it administrative districts.

Target audience :

1. **Business personnel, Top Management of the Bank, Shareholder and other Departments. This research will be a guide how to define a better location of organization and see competitors locations to avoid leakage of income.**
2. **Anyone who is interested in accumulating information about exploratory Data Analysis of their cities.**

C.Methodology

Intruments for this research are open sources which are available for everyone. I used Jupiter Notebook and Github repository for this research. My master data which has the main components district (address_borough), Latitude and Longitude off all of the Banks and their Branches in the city. It is pertinent to mention that I have used two main sources Foursqare to define Head Offices of the Bank, and Google Maps to find all the Branches.

In [122]:
```
df.head(10)
```

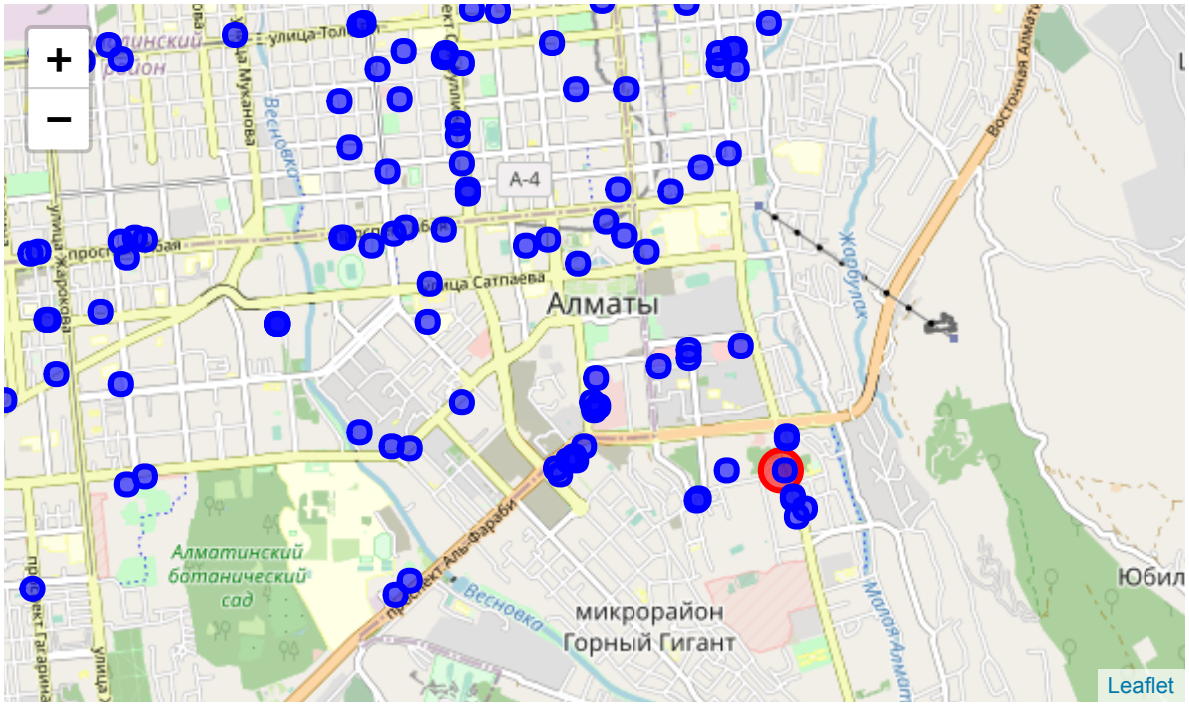
Out[122]:

	name	type	address_borough	lat	lng
0	Citibank Kazakhstan	Bank	Medeu District	43.257438	76.956957
1	Tengri Bank	Bank	Bostandyk District	43.240714	76.927797
2	KZI BANK (Kazakhstan Ziraat International Bank)	Bank	Bostandyk District	43.235870	76.902687
3	Center Credit Bank	Bank	Medeu District	43.261060	76.960007
4	Center Credit Bank	Bank	Medeu District	43.226612	76.942683
5	National Bank of Pakistan	Bank	Medeu District	43.225966	76.961241
6	Capital Bank	Bank	Bostandyk District	43.240537	76.947502
7	MUFG Bank Almaty Representative Office	Bank	Medeu District	43.227407	76.944123
8	National Bank of Kazakhstan	Central bank	Bostandyk District	43.235131	76.917940
9	Eurasian Development Bank HQ	Bank	Medeu District	43.228007	76.961421

During the research I was using folium library to vizualize all geographical areas of where Banks are located and labelled them.

In [123]:
```
alm_venues_map
```

Out[123]:



After vizualization we can come to conclusion that city has a great number of financial servises located in every part, therefore I used count() method to see what are the numbers of Banks located in each district and parsed wikitable with administrative information like population. All this data was merged to define population/banks offices ratio to understand what is real picture in each area.

In [124]:
```
df_merged.head()
```

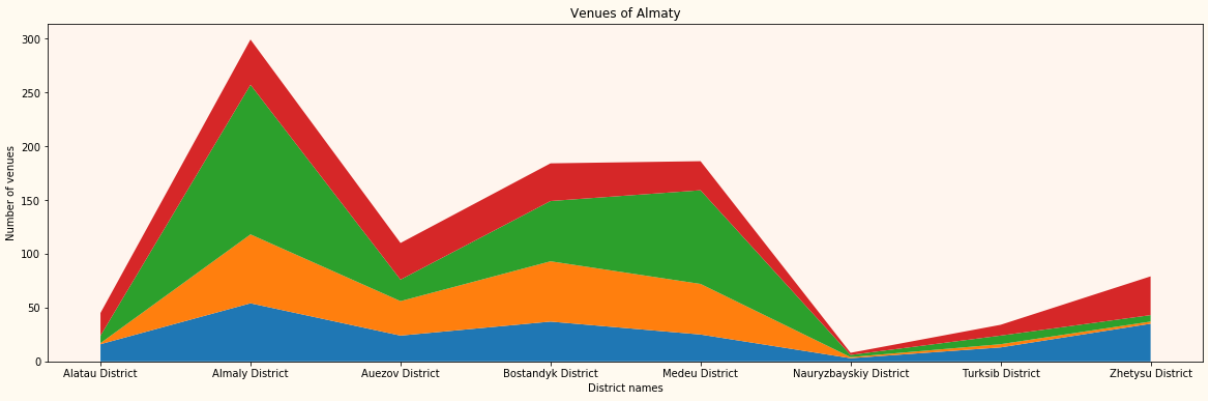
Out[124]:

	address_borough	populations	n_of_banks	pop_to_bank_ratio
1	Almaly District	215768.0	64.0	3371
5	Medeu District	209836.0	47.0	4464

	address_borough	populations	n_of_banks	pop_to_bank_ratio
3	Bostandyk District	343541.0	56.0	6134
2	Auezov District	295543.0	32.0	9235
7	Turksib District	235357.0	3.0	78452

Now we can see the different districts of Almaty and the number of Banks in each district, but having only quantative data we cannot do any research conclusion
Lets see how the data is displayed and what additional information it can bring to us!

In [125]:
```
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt

x = fgrs_for_plot['address_borough']

y1 = fgrs_for_plot['n_companies']
y2= fgrs_for_plot['n_of_banks']
y3 = fgrs_for_plot['n_of_pastimes']
y4 =  fgrs_for_plot['n_of_molls']

fig, ax = plt.subplots()

ax.stackplot(x, [y1, y2, y3, y4])

ax.set_facecolor('seashell')

fig.set_figwidth(20)
fig.set_figheight(6)
fig.set_facecolor('floralwhite')

plt.title('Venues of Almaty')
plt.ylabel('Number of venues')
plt.xlabel('District names')

plt.show()
```



Due to the fact that Almaty city has 848 venues its if complicated to vizualize all them and I decided to choose top 10 venues and accumulated them in the table below. Most of them are the Banks, cafes, wholesale trade and different shops.

In [126]:
```
neighborhoods_venues_sorted.head(10)
```

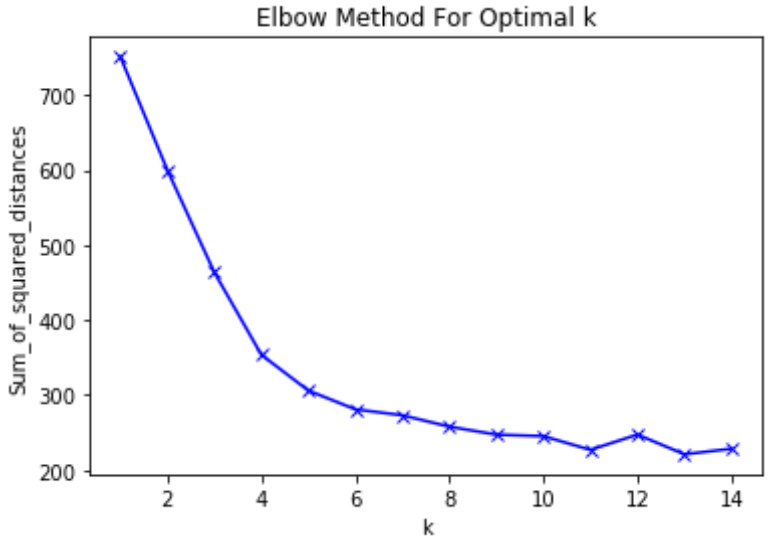
Out[126]:

	Cluster_Labels	address_borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4t Co
0	4	Alatau District	Торговый центр	Кафе	Офис компании	Ма
1	3	Almaly District	Кафе	Bank	Торговый центр	Рес
2	1	Auezov District	Торговый центр	Bank	Кафе	ком
3	3	Bostandyk District	Bank	Торговый центр	Кафе	Рес
4	3	Medeu District	Bank	Кафе	Торговый центр	Рес

	Cluster_Labels	address_borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4t Co
5	0	Nauryzbayskiy Disctrict	Bank	Ювелирный магазин	Ломбард	Магазин
6	2	Nauryzbayskiy District	Производитель	Торговый центр	Фармацевтическая компания	Ювели ма
7	1	Turksib District	Торговый центр	Кафе	Производитель	
8	4	Zhetysu District	Торговый центр	Офис компании	Кафе	Строител ма

We have a lot of similiar categories of venues in different districts. For research purposes it is critically important to understand how similiar or dissimiliar these venues and the best way to solve this problem is clustering the data. I have used method of Kmeans of unsupervised learning mechanism.

In [127]:
```
plt.plot(K, Sum_of_squared_distances, 'bx-')
plt.xlabel('k')
plt.ylabel('Sum_of_squared_distances')
plt.title('Elbow Method For Optimal k')
plt.show()
```



Before starting clustering the districts i had to understand the best optimal number of clusers which I can use to get a better grouping of the objects by Kmeans method. The best for this purpose is to use elbow method and it was found as 5. So I can start grouping and labeling each of the venue with a specific cluster. After merging we have a result below:

In [127]:
```
joined_table.head(5)
```

Out[127]:

	address_borough	lat	lng	address_street	type	Cluster_Labels	1st M Comm Ve
2	Zhetysu District	43.289880	76.926703	Ryskulov Ave 57/B	Продажа автомобилей	4	Торго це
3	Zhetysu District	43.284758	76.948181	просп. Суюнбая 89,Алматы	Офис компании	4	Торго це

	address_borough	lat	lng	address_street	type	Cluster_Labels	1st M Comm Ve
4	Zhetysu District	43.288218	76.940784	Seyfullin St 288	Офис компании	4	Торго це
7	Almaly District	43.249144	76.932853	Seyfullin St 563	Офис компании	3	К
8	Almaly District	43.263537	76.945241	Nazarbayev Ave уг	Продажа оборудования	3	К

When we analyze above table we can label each cluster as follows:
Cluster 0 : “Naurizbayskiy district” - region with private houses, mostly local pastimes for families, cafes and shops

Cluster 1 : “Auezov and Turksib districts" - mixed with private houses, car services stations, ATM's and medium size living neighboorhoods

Cluster 2 : “Naurizbayskiy district” with venues closer to the city - private houses, a couple of Banks' Branches, more shops and trading centers

Cluster 3 : "Almaty, Medeu, Bostandyk"- most developed districts in the city with multiple infastructure items

Cluster 4 : "Alatau district" - district with mixed private houses, petrol stations, most car sales companies and producers are located in this part of city

Clustering of different districts and grouping them can give us indication of similiarity or dissimiliarity of the city, but unfortunately it doesn't allow to indicate descriptive analitics of each district of the city. For this purpose I have gathered data in the table with the most important infustructure items such as trading centeres, big commercial and trading companies, producers, and other companies from absolutely different sectors of economy. Table is really good indicator for this purpose as we can compare different

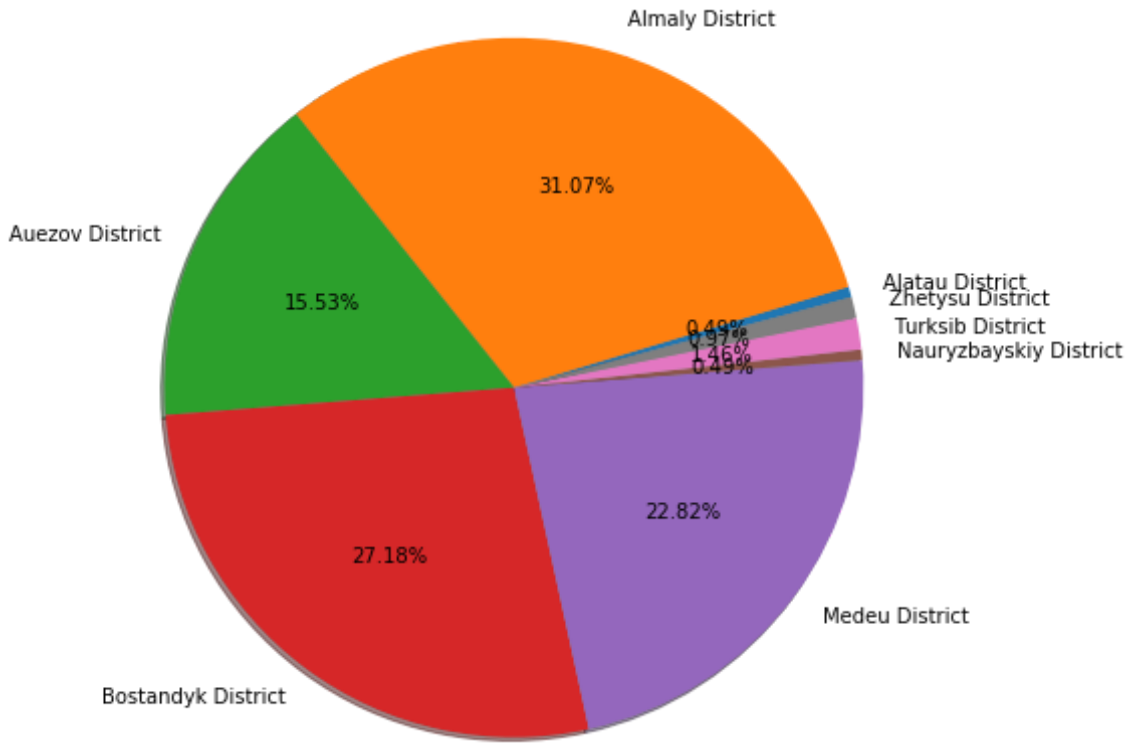
parts of the city and see how financial services are developed in the city. As well we can indicate population to Bank ratio which is a leading indicator in addition to clusters.

In [128]:
```
aggregated_table.head(10)
```

Out[128]:

	address_borough	n_companies	populations	n_of_banks	pop_to_bank_ratio	n_of_molls	n_of
0	Alatau District	16	260441.0	1.0	260441	21	
1	Almaly District	54	215768.0	64.0	3371	42	
2	Auezov District	24	295543.0	32.0	9235	34	
3	Bostandyk District	37	343541.0	56.0	6134	35	
4	Medeu District	25	209836.0	47.0	4464	27	
5	Nauryzbayskiy District	3	128169.0	1.0	128169	2	
6	Turksib District	13	235357.0	3.0	78452	10	
7	Zhetysu District	35	166001.0	2.0	83000	36	

In [129]:
```
plt.pie(aggregated_table['n_of_banks'],  labels= aggregated_table['addr
ess_borough'],
        startangle=15,
        shadow = True,
       radius = 2, autopct = '%0.2f%%')
plt.show()
```

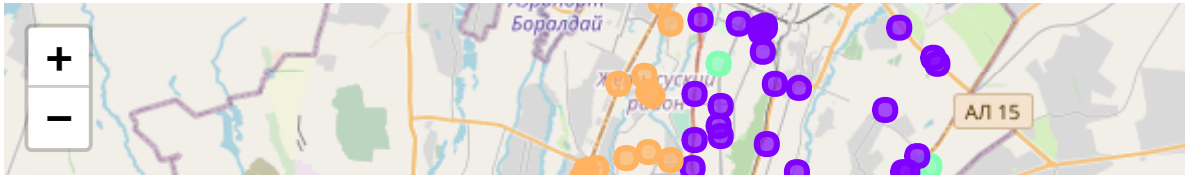


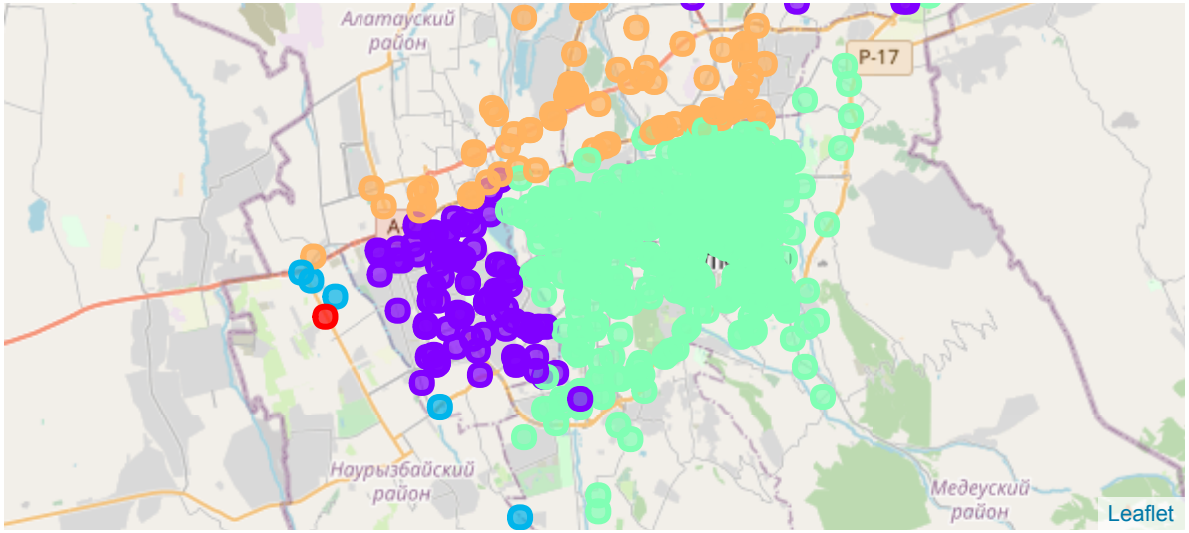
C.Result

As you can see in the abovementioned pie chart and table there is a very great difference in the descriptive data of districts. The districts are very different, but lets see them on the map to have a better understanding.

In [130]:
```
district_clusters
```

Out[130]:





We obviously see a great area in the center of the city under cluster 3. It means that according to administrative division of the city the districts Almaly, Medeu, Bostandyk were grouped with Kmeans method as similiar and some areas of other districts have also been included in cluster 3. I have labelled all items included in cluster 3 with:

1. **Name of district**
2. **Relation to the cluster**

Almost five hundred of venues are available in this cluster (3) of the city crossing by main roads, public transportation, with around 20 ATMs. But what district we can advise as the best to establish a Branch there?

D. Discussion

As I mentioned before, Almaty is financial hub of Central Asia with medium sized population on rather large area. The population of each district vary from 100 thousands to more than 300 thousands, but the density of population of all 8 districts bring three

leaders to the Top: Medeu, Bostandyk, Almaly districts. The main difficulty of Kmeans approach applied to this city is that the districts have rather great distances as compared to typical neighborhoods in big capitals like NY or LA. Using other types of segmentation like data on postal codes may devide Almaty on more clusters if we use a lower radius as well, but the difference between three leaders will be almost the same.

F. Conclusion

Now I can give the answer to the Bank what is a better location to open a Branch. I recommend to open a Branch in Almaly district of the city as it has been clustered in same category as more expensive by rent Medeu district with Head Offices of International Banks. A greater infrastructure is a second merit as here there are 139 places for families entertainment, 42 trading centers, 54 businesses. The location of the Branch in this district decrease uncertanity factor for Retail Business focus of the Bank.

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: