Final Report

1) Introduction/ Business Background

Due to the spread of Covid-19, good health care has been shown to be an important factor in containing and fighting pandemics. Therefore, the possible location of a new pharmacy in Manhattan will be analysed in order to make a recommendation for the opening of a branch for a pharmacist. The turnover of a pharmacy and the supply of medicines to the population depends not only on the number of pharmacies but also on the population density of a neighbourhood.

And thus any new business venture or expansion needs to be reviewed carefully and strategically targeted so that the return on investment will be sustainably reasonable and more importantly the investment can be considered as less risky.

2) Data Description

To solve the problem identified, site data will be used and analysed to derive a recommendation. In particular, the Foursquare API is used for this purpose, which uses queries to output the venues in each neighbourhood within a given radius. This data is combined with public data on the demographics of each neighborhood. This data is publicly available from the City of New York and can be accessed at https://data.cityofnewyork.us/api/views/8m6s-esnp/rows.csv (https://data.cityofnewyork.us/api/views/8m6s-esnp/rows.csv). An extract of the data is shown below.

import pandas as pd link="https://data.cityofnewyork.us/api/views/8m6s-esnp/rows.csv?accessType=DOWNLOAD (https://data.cityofnewyork.us/api/views/8m6s-esnp/rows.csv?accessType=DOWNLOAD)" manhattan_pop= pd.read_csv(link) manhattan_pop.head()

As you can see the record contains the borough, the county code and NTA code, and the name of the neighborhood. As no further information is needed besides the name of the neighbourhood and the population, the other features are deleted. The dataset is from 2010, so it is a somewhat outdated dataset, but due to its easy availability it should serve as a basis for this analysis.

For this dataset dataset some neighborhoods are summarized within one row. We need to clean the data set, in order to have population data for all neighborhoods we do the following:

1) transform the NTA Names to a list containing the certain neighborhoods

2) add length of each neighborhood-list as new column

3) iterate over dataframe and append all neighborhoods as single to a new list. For Neigborhoods that have been summarized we take the mean (which is the population divided by number of neighborhoods)

3) Methodology

For a complete analysis of possible neighborhoods and locations, a data set was first required that included the geolocation of the respective city districts. For this purpose, the dataset in JSON format was used and transferred into a Pandas Dataframe.

This dataset was then enriched with the population of the respective neighbourhood. For this purpose, the data had to be cleaned and adapted to the format of the existing first dataset. These steps were already sufficiently described in chapter 2 (Data Description). From the population data, a so-called population score (Pop Score) was then derived, which represents the relative population of Manhattan in the respective neighborhood. This score is to serve as one of the bases for decision-making at a later stage of the analysis.

The next step was to use the Foursquare API to request and store specific information on the individual venues in the respective neighborhoods. For this purpose, all venues within a radius of 500 meters were used accordingly. With the help of this data, a relative frequency of pharmacies in the respective neighbourhood was calculated and a so-called Pharmacy Score was derived, which is used again in the later analysis. This score will be used as a further basis for decision-making at a later stage of the analysis.

Since all the necessary data was now available, the next step was to combine this data into a common data frame. Weights were introduced for the two assessment bases number of pharmacies in the neighbourhood (Pharmacy Score) and relative population of Manhattan in the respective neighbourhood (Pop Score). The Pharmacy Score represents the possible competition from competitors and is therefore the worse the higher the Pharmacy Score is. The Pop Score represents an estimator for the respective sales potential that is based on the relative population of the neighborhood and is therefore better the higher it is. To combine both scores, their relative weight is set at 50% each, as this weight is considered realistic by the authors of the study. An overall score is then calculated from the difference between the Pop Score times relative weight minus the Pharmacy Score times relative weight.

The higher the respective overall score, the more promising the neighborhood is for a possible location.

In the last step, a K-Means cluster analysis was performed to identify possible similarities between individual neighborhoods. For this purpose the number of initial clusters was set to 5.

4) Results

As you can see, the Upper West Side is the best neighborhood for a new pharmacy. Two factors play a major positive role here. On the one hand, the Upper West Side is the most populous neighborhood in Manhattan and therefore has an immense sales potential and a great demand for drugs and medical supplies. On the other hand, the competition with other pharmacies here is very low, as the relative number of other pharmacies is very small.

In [9]:
```
pd.read_csv("manhattan_fin.csv").head()
```

Out[9]:

	Unnamed: 0	Neighborhood	Latitude	Longitude	Population	Pop Score	Pharmacy Score	Ov S
0	6	Upper West Side	40.787658	-73.977059	132378.0	12.487813	0.000000	6.24
1	4	Yorkville	40.775930	-73.947118	77942.0	7.352620	1.000000	3.17
2	16	West Village	40.734434	-74.006180	66880.0	6.309092	0.000000	3.15
3	15	Lincoln Square	40.773529	-73.985338	61489.0	5.800534	1.041667	2.37
4	23	Hamilton Heights	40.823604	-73.949688	48520.0	4.577110	0.000000	2.28

As you can see, the Upper West Side is the best neighborhood for a new pharmacy.In both neighborhoods the relative population is very high and competition from other pharmacies is very low. For the sake of completeness it should be noted that the highest competition from other pharmacies is on the Lower East Side and in Marble Hill. Marble Hill closed as the worst neighbourhood for opening, not least because the relatively lower population was not able to compensate for this either.

In [12]:
```
df= pd.read_csv("manhattan_fin_merge.csv")
df
```

Out[12]:

	Unnamed: 0	Neighborhood	Latitude	Longitude	Population	Pop Score	Pharmacy Score	
0	27	Upper West Side	40.787658	-73.977059	132378.0	12.487813	0.000000	6.
1	21	East Village	40.727847	-73.982226	44136.0	4.163548	0.000000	2.
2	18	Chinatown	40.715618	-73.994279	47844.0	4.513340	1.000000	1.

3	17	Lower East Side	40.717807	-73.980890	72957.0	6.882363	3.921569	1.
4	26	Yorkville	40.775930	-73.947118	77942.0	7.352620	1.000000	3.
5	9	Gramercy	40.737210	-73.981376	27988.0	2.640234	1.234568	0.
6	7	Midtown South	40.748510	-73.988713	14315.0	1.350398	0.000000	0.
7	12	Battery Park City	40.711932	-74.016869	19849.5	1.872493	0.000000	0.
8	0	Marble Hill	40.876551	-73.910660	23373.0	2.204880	4.000000	-0.
9	2	Midtown	40.754691	-73.981669	14315.0	1.350398	1.000000	0.
10	8	Murray Hill	40.748303	-73.978332	25371.0	2.393361	1.000000	0.
11	25	West Village	40.734434	-74.006180	66880.0	6.309092	0.000000	3.
12	24	Lincoln Square	40.773529	-73.985338	61489.0	5.800534	1.041667	2.
13	22	Clinton	40.759101	-73.996119	45884.0	4.328444	0.000000	2.
14	19	Lenox Hill	40.768113	-73.958860	40385.5	3.809746	0.000000	1.
15	3	Turtle Bay	40.752042	-73.967708	25615.5	2.416426	2.000000	0.
16	6	Little Italy	40.719324	-73.997305	10685.5	1.008011	0.000000	0.
17	15	Carnegie Hill	40.782683	-73.953256	30603.5	2.886966	0.000000	1.
18	5	Civic Center	40.715229	-74.005415	10685.5	1.008011	0.000000	0.
19	11	Chelsea	40.744035	-74.003116	17537.5	1.654391	0.000000	0.
20	10	Hudson Yards	40.756658	-74.000111	17537.5	1.654391	0.000000	0.
21	16	Upper East Side	40.775639	-73.960508	30603.5	2.886966	0.000000	1.
22	14	Morningside Heights	40.808000	-73.963896	55929.0	5.276035	2.631579	1.
23	20	Roosevelt Island	40.762160	-73.949168	40385.5	3.809746	0.000000	1.
24	1	Inwood	40.867684	-73.921210	23373.0	2.204880	3.448276	-0.
25	23	Hamilton Heights	40.823604	-73.949688	48520.0	4.577110	0.000000	2.
26	13	Manhattanville	40.816934	-73.957385	22950.0	2.164977	0.000000	1.
27	4	Stuyvesant Town	40.731000	-73.974052	10524.5	0.992824	0.000000	0.

If the average overall score is calculated for each cluster, clusters 0, 2 and 3 are particularly attractive for a particular branch concept (see below). Cluster 0 represents the densely populated lively neighborhoods such as Upper West Side, Lower East Side and Yorkville. Cluster 3 represents the affluent neighborhoods such as Roosevelt Island, Hamilton Heights and Manhattanvile, where a different branch concept that focuses on more affluent customers, for example, could be promising.

In [15]:
```
df.groupby(["Cluster"])[["Overall Score"]].mean()
```

Out[15]:

Cluster	Overall Score
0	2.302441
1	0.227642
2	1.396423
3	1.195289
4	0.496412

5) Discussion

Further analyses with the help of additional data are certainly useful to include, for example, costs for a branch, such as rent, personnel costs and incidental expenses, as these can vary from quarter to quarter. The turnover potential also depends not only on the population size but also on the average income in the respective neighbourhoods. These data could also be included in order to increase the significance of the model.

6) Conclusion

In conclusion, the top 3 neighborhoods Upper West Side, Yorkville and Westvillage can be recommended as the best neighborhoods for a new pharmacy. According to this analysis, however, opening a pharmacy in Midtown, Inwood or Marble Hill is not recommended. The cluster analysis has shown that different neighborhoods are similar and that different store concepts may offer promising business opportunities. The model can be further improved with the help of other factors and data, but it does provide a good initial assessment for the location search.