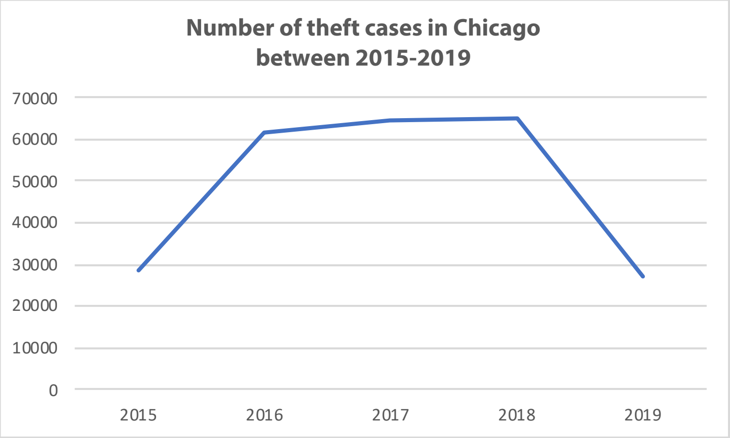
CAPSTONE PROJECT

BUSINESS PROBLEM

You want to open a new Ice Cream Shop in Chicago, but have a few questions in order to open this shop. You are hoping that data will help you solve them.

The Chicago Data Portal has released a dataset in which one can learn more about the crimes between 2001 and present. The crime that most often impacts a business, is theft. There have already been 27321 cases of theft in 2019 alone – moreover, in 2018, the crime with the most cases was also theft, with a total 65200.



Link: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data

You place a high importance on the safety of your customers, so you want to make sure you open up your ice cream shop in an area where you know your customers will be safe. You start to think about what data you would need to determine which communities are the safest. You do not want to limit your location possibilities, so you will focus on the top 5 safest communities in Chicago. Moreover, considering a high safety score will also minimize the risk of robbery that the business may face.

Moreover, you know your target market will be families with young children. However, you do not know where you will target your marketing efforts – after all, according to Chicago Health Atlas, there were ±630,000 people in between the ages of 0 – 19 in Chicago in 2017. Maybe you can segment this target group into a smaller age group – what data would we need for that?

Lastly, to determine of the five communities where you will locate your ice cream shop, the current level of supply in the community will be examined. In other words, the number of competitors who are currently present in that specific community will be analyzed. If this is high, the likelihood is lower to start the business in this neighbourhood.

Overall, this will ensure that we are able to set up our business in a community that is safe for your customers (especially considering the young age of your target market) and in a community where there is a market gap, i.e. no ice cream shop is currently present.

DATA

Different datasets will be used to solve this problem.

1. A governmental dataset called 'CHICAGO_PUBLIC_SCHOOLS', which includes all columns such as school name, community name, zipcode, geospatial coordinates, safety score, etc. The safety score will be an average of all safety scores within that zipcode. This dataset will be used to determine the 5 communities with the highest safety score of which we will focus on. *Link:* https://www.kaggle.com/chicago/chicago-public-schools-data

2. Geographical coordinates using Folium: Used in order to plot the zipcodes onto the map. This will help determine if any outliers exist, i.e. zipcodes that are extremely far from the rest. These will be removed from the potential locations.

3. Foursquare data: This data will be used to determine the most common venues of each zipcode. This will help to see whether there are any gaps in the zipcode area in which demand for an ice cream shop is not being met.

4. A governmental dataset called 'CENSURDATANEW' that includes six socioeconomic indicators of public health significance and a "hardship index" by community area in Chicago between 2008 and 2012. *Link: https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2*

5. A governmental dataset called 'TOTALPOPULATION' from Chicago Health Atlas, which outlines the number of individuals per age group over a time frame of 2010 – 2017. Each age group is split every four years. This will help segment our target market into groups of 0-4, 5-9, 10-14 and 15-10. From here, we can determine which of these segments would be most lucrative to focus our marketing efforts on. *Link: https://www.chicagohealthatlas.org/indicators/total-population*

6. A governmental dataset called 'CHILDOBESITY' from Chicago Health Atlas, which outlines the percentage of children whose BMI is equal to or greater than the 95th percentile on the CDC BMI-for-age growth charts, indicating obesity. This data is split per community, thus we can determine health-conscious behavior per community. This will help us to determine whether the chosen community is health-conscious, and thus whether we will consider this as a threat or opportunity to opening up in that area. *Link: https://www.chicagohealthatlas.org/indicators/child-obesity*

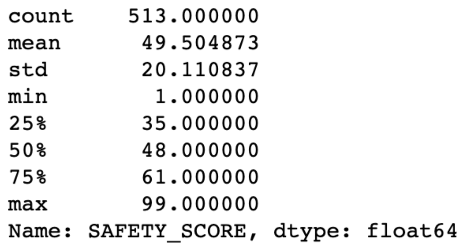
METHODOLOGY

DATA EXTRACTION

The different datasets were imported into the IBM DB2 database in order to be extracted using the sql load extension. This required entering database credentials and creating the dsn connection string, subsequently creating the connection to the database. Multiple libraries also needed to be imported.

STATISTICAL TESTS

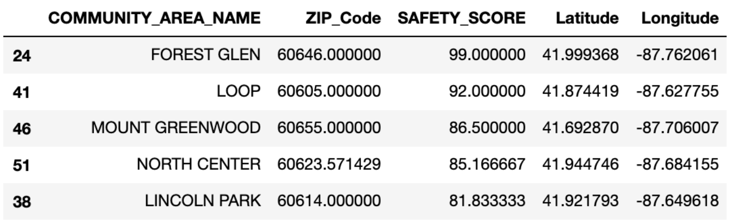
Statistical tests were run in order to understand the data. For example, the following describes the statistics of the SAFETY_SCORE, a variable of the CHICAGO_PUBLIC_SCHOOLS dataset. We can see above that the minimum safety score is 1 and the max is 99. We can also see other statistics, like the mean score being 49.5 and that 75% of all safety scores are below 61.



```
count    513.000000
mean      49.504873
std       20.110837
min        1.000000
25%       35.000000
50%       48.000000
75%       61.000000
max       99.000000
Name: SAFETY_SCORE, dtype: float64
```

DATA CLEANING

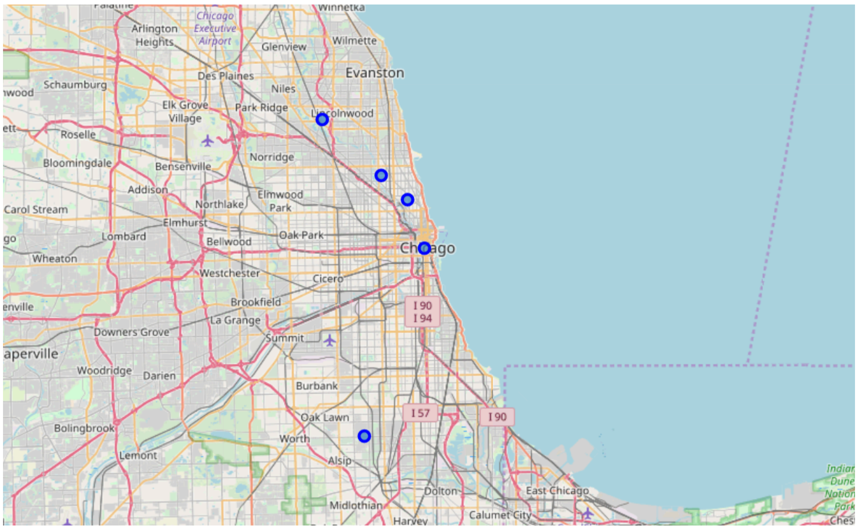
Data needed to be cleaned in order to retrieve the dataframe in the form we want. This included accessing only specific columns, sorting, setting and resetting indexes and creating new dataframes. For example, as we wanted to focus on those with the highest safety score, scores are sorted in descending order in order to determine the top 5 safest communities. A new dataframe was created with the data below.



	COMMUNITY_AREA_NAME	ZIP_Code	SAFETY_SCORE	Latitude	Longitude
24	FOREST GLEN	60646.000000	99.000000	41.999368	-87.762061
41	LOOP	60605.000000	92.000000	41.874419	-87.627755
46	MOUNT GREENWOOD	60655.000000	86.500000	41.692870	-87.706007
51	NORTH CENTER	60623.571429	85.166667	41.944746	-87.684155
38	LINCOLN PARK	60614.000000	81.833333	41.921793	-87.649618

GEOGRAPHICAL ANALYSIS

In order to determine whether there are any geographical outliers, Folium will be used in order to plot the 5 communities with their coordinates onto the map. One can clearly see in the map that there is 1 outlier, namely Mount Greenwood with zipcode 60655. This will be discarded from the options, due to its large distance from the other possible communities.

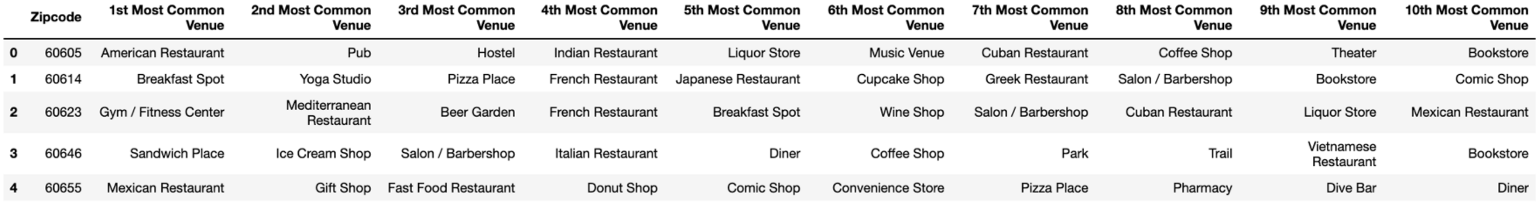


The remaining communities and zipcodes are:

Forest Glen	60646
Lincoln Park	60614
Loop	60605
North Center	60623

FOURSQUARE ANALYSIS

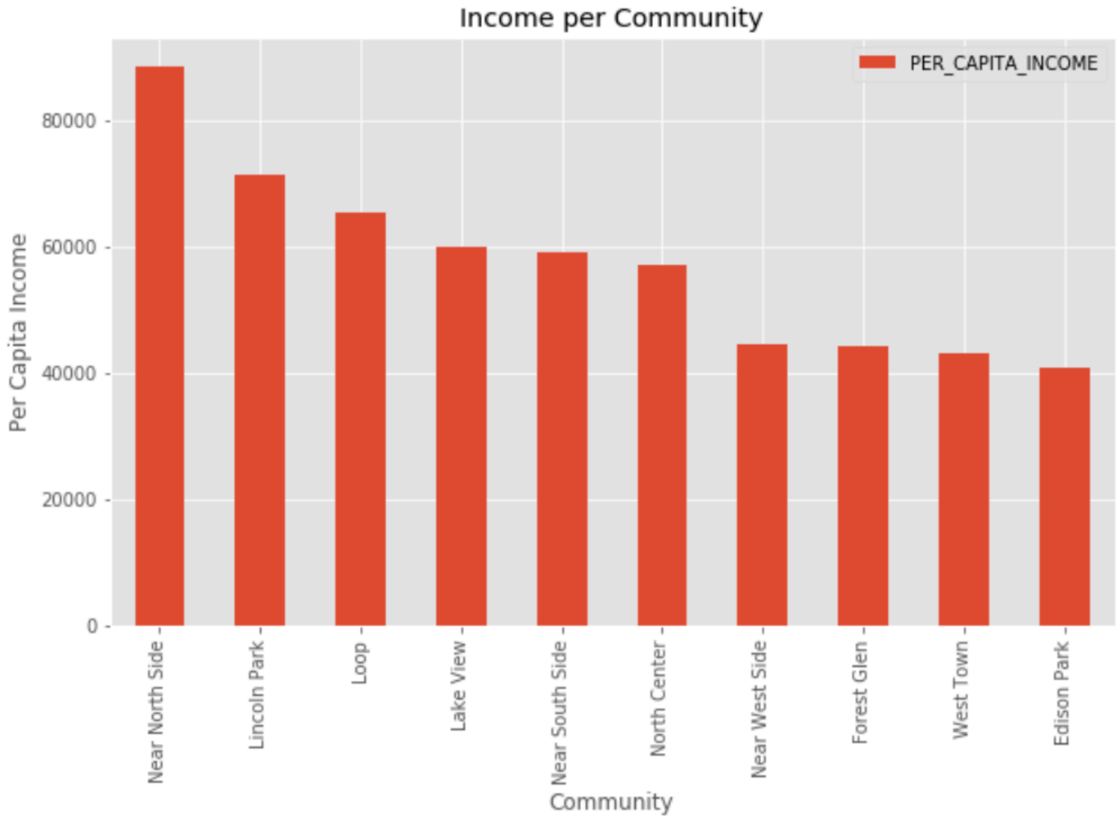
The following analysis focused on whether there already is a supply of ice cream shops in each zipcode. This will focus on the number of competitors in the surroundings. This step required creating a function that extracts the category of each venue, finds the common venues in each neighborhood and then converts this into a readable dataframe file. The output was the following dataframe. The results from this analysis will be discussed in the following section.



	Zipcode	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	60605	American Restaurant	Pub	Hostel	Indian Restaurant	Liquor Store	Music Venue	Cuban Restaurant	Coffee Shop	Theater	Bookstore
1	60614	Breakfast Spot	Yoga Studio	Pizza Place	French Restaurant	Japanese Restaurant	Cupcake Shop	Greek Restaurant	Salon / Barbershop	Bookstore	Comic Shop
2	60623	Gym / Fitness Center	Mediterranean Restaurant	Beer Garden	French Restaurant	Breakfast Spot	Wine Shop	Salon / Barbershop	Cuban Restaurant	Liquor Store	Mexican Restaurant
3	60646	Sandwich Place	Ice Cream Shop	Salon / Barbershop	Italian Restaurant	Diner	Coffee Shop	Park	Trail	Vietnamese Restaurant	Bookstore
4	60655	Mexican Restaurant	Gift Shop	Fast Food Restaurant	Donut Shop	Comic Shop	Convenience Store	Pizza Place	Pharmacy	Dive Bar	Diner

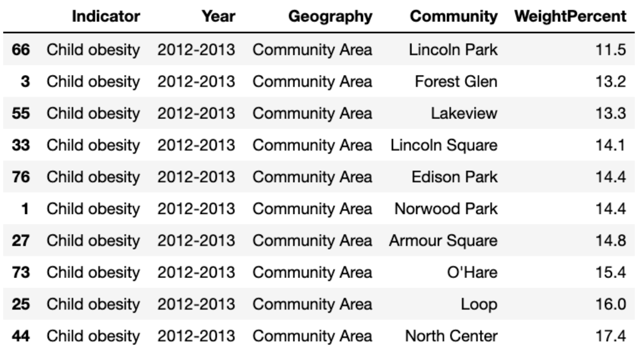
ZIPCODE ANALYSIS: INCOME

In the results, it will be discussed that the chosen community will be 60623, the community of North Center. In order to support this choice, additional analysis is performed. The first analysis looks at PER_CAPITA_INCOME vs ZIPCODE. As there is a correlation between PER_CAPITA_INCOME vs the safety of a community, a high income should be correlated with a higher community safety. Thus, if the zipcode appears in the following graph, we are able to state that it is indeed in a safer community and is appropriate for further consideration. This required the extraction of a new dataset and the plotting using matplotlib.pyplot. The results from this analysis will be discussed in the following section.



ZIPCODE ANALYSIS: CHILD OBESITY

In order to further support the choice of North Center, the relationship between child obesity and community (zipcode) will be analyzed. This required the extraction of another dataset. In this dataset, there are 77 communities in which WeightPercent refers to the percentage of children whose BMI is equal to or greater than the 95th percentile on the CDC BMI-for-age growth chart, and therefore considered in the obesity class. The following dataframe presents the 10 communities with the lowest child obesity score.



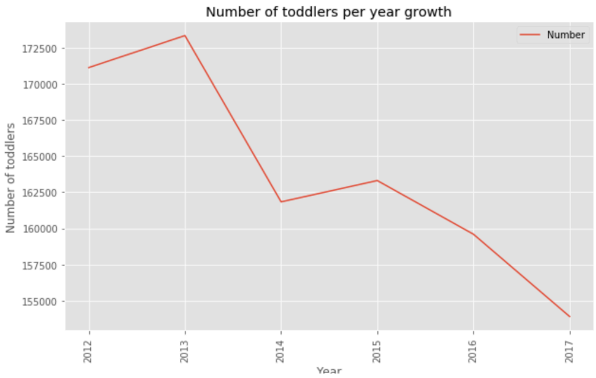
	Indicator	Year	Geography	Community	WeightPercent
66	Child obesity	2012-2013	Community Area	Lincoln Park	11.5
3	Child obesity	2012-2013	Community Area	Forest Glen	13.2
55	Child obesity	2012-2013	Community Area	Lakeview	13.3
33	Child obesity	2012-2013	Community Area	Lincoln Square	14.1
76	Child obesity	2012-2013	Community Area	Edison Park	14.4
1	Child obesity	2012-2013	Community Area	Norwood Park	14.4
27	Child obesity	2012-2013	Community Area	Armour Square	14.8
73	Child obesity	2012-2013	Community Area	O'Hare	15.4
25	Child obesity	2012-2013	Community Area	Loop	16.0
44	Child obesity	2012-2013	Community Area	North Center	17.4

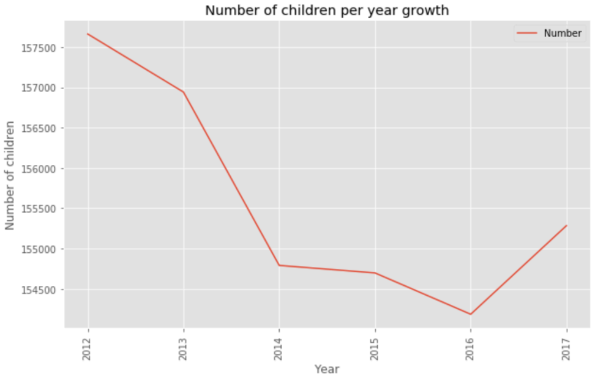
TARGET MARKET ANALYSIS

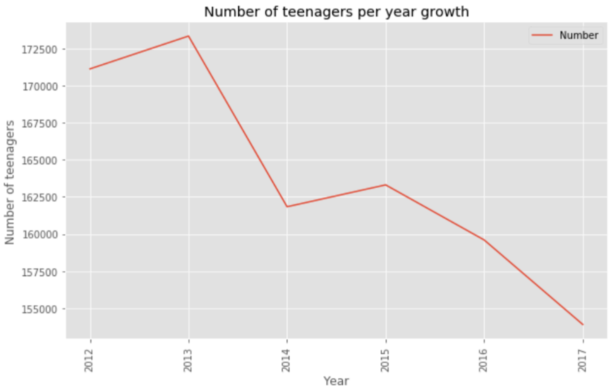
Another business decision to be determined was how to segment the target market into smaller groups that would be easier to target with marketing efforts. We would want to analyze the growth of the number of children, per segment, per year. First, data needed to be extracted and transformed into the appropriate dataframe. This required setting new indexes, creating categories for each child segment, and setting new indexes for each of these individual segments. The categories were:

Toddlers	5-9
Children	10-14
Teenagers	15-19

Note, the ‘baby’ category of 0-4 was removed. Afterwards, the data was plotted using matplotlib.pyplot in order to determine the growth trend of each category.





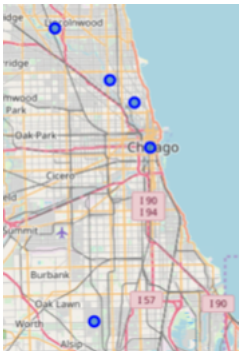


RESULTS & DISCUSSION

Firstly, it was determined that the communities with the highest safety score are:

Community Name	**Zipcode**	**Score**
Forest Glen	60646	99
Loop	60605	92
Mount Greenwood	60655	87
North Center	60623	85
Lincoln Park	60614	82

Then, after geographically analyzing the location of all communities, you immediately notice that the community with zipcode 60655 is an outlier from the others, which are all located north of Chicago. Therefore, we will remove this zipcode from the possibilities. The remaining possibilities are: 60646, 60614, 60605, 60623.



Moreover, by looking at the most common venues of the remaining 4 communities, we were able to determine the following:

60605's 6th most common venue is an ice cream shop
60614's 7th most common venue is an ice cream shop
60623 does not have an ice cream shop
60646's 2nd most common venue is an ice cream shop

Thus, there is sufficient supply in the zipcodes 60605, 60614 and 60646. However, it would be interesting to take a further look into 60623, as there is currently no common ice cream shop.

After digging deeper into 60623, the following insights were found. When looking at the income_per_capita, each of the four communities is part of the top 6 highest income_per_capita communities:

60605: 3rd highest per capita income
60614: 2nd highest per capita income
60623: 6th highest per capita income

60646: 8th highest per capita income

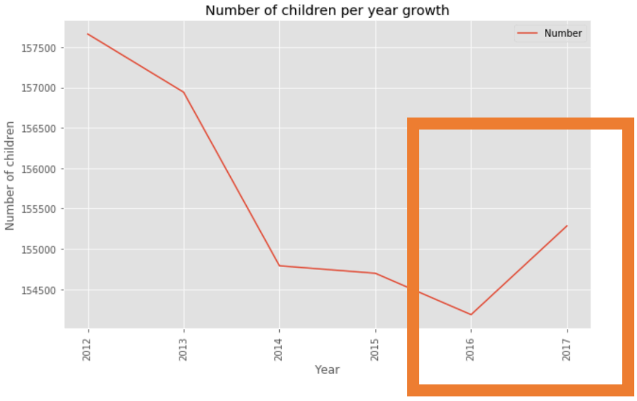
Therefore, we can safely assume that the community chosen, North Center with zipcode 60623, fulfills the safety criterion and thus is an appropriate choice.

We also found that each of the four communities is in the top 10 lowest child obesity rate communities, of the total 77 communities:

60605: 9th lowest in child obesity rate
60614: 1st lowest in child obesity rate
60623: 10th lowest in child obesity rate
60646: 2nd lowest in child obesity rate

As we can see, North Center is in the top 10 with lowest obesity rates. This data can both show us a threat, or opportunity with opening an ice cream shop in North Center. It may be a threat, as the community is very health-conscious with its resulting low child obesity score, meaning that parents do not often take their children to an "unhealthy place". On the other hand, it can be also seen as an opportunity, as we are now aware of their health-conscious behavior and thus can include "healthier" options.

Lastly, the target analysis graph demonstrated a general downward trend in the number of toddlers (5-9) and teenagers (15-19) between 2012 and 2017. On the other hand, although there was an initial downward trend of number of children over the years, it is slowly increasing since 2016. Therefore, the business conclusion we can make from this data is to focus on families with children between 10-14 years old in order to maximize our customer reach.



CONCLUSION

Overall, the prior analyses were able to determine multiple findings:

1. It determined the safest communities in Chicago, using their safety score. Safety is an important criteria considering the target audience of families with children. The top 5 were chosen to be 60646, 60614, 60605, 60623 and 60655.

2. Geographical outliers, i.e. any community which is extremely far from the rest, was not considered. This was determined using Folium. Zipcode 60655 was excluded.

3. There are many competitors (i.e. sufficient supply) in the zipcodes 60605, 60614 and 60646. However, it would be interesting to take a further look into 60623 (North Center), as there is currently no common ice cream shop that families go to in this community.

4. A general business analysis demonstrated that there is a general downward trend in number of toddlers (5-9) and teenagers (15-19), but a slow upward trend of children (10-14). Therefore, the targe audience is refined to focus on families with children between 10-14 years old in order to maximize customer reach.

5. Deeper analysis into zipcode 60623 showed that they are one of the top 6 communities in Chicago with the highest income per capita and thus, this confirms our criterion of being a safe zipcode.

6. Deeper analysis into zipcode 60623 (North Center) showed that it is amongst the top 10 'healthiest' communities due to its low child obesity score. This can present both a threat in terms of preferences of target market, but it can also be an opportunity in terms of changing your product portfolio to include healthier options.