20 september 2019

Finding optimal salad bar location

IBM Applied Data Science course
Capstone Project

Xander Mol

Table of Contents

Cover page photo credit:

Jakub Kapusnak

https://www.foodiesfeed.com/free-food-photo/healthy-vegetarian-salad-buffet/

Executive Summary

- Problem to be solved: What is the optimal neighborhood to start a new salad bar, based on a comparison between neighborhoods in the four cities considered using demographic and income data combined with data on other venues in the area, based on the following requirements and background:
    - Only locations in the four biggest cities in the Netherlands are considered, being Amsterdam, Rotterdam, Utrecht and The Hague;
    - From market research a projected ideal target group is defined: single household females;
    - Areas with abundant availability of other restaurants are preferred as experience indicates that this attracts more potential customers to new restaurants;
    - Preferably none or few competing salad bars in the area;
- Analys done using demographic and income data from the Dutch Central Bureau of Statistics and Foursquare data on venues at selected locations, including presence of salad bars;
- Methodology used was first clustering all postal codes in the selected cities in 5 clusters based on common characterisctics on demographics, income and venues present.
  These 5 clusters were analysed given our selection criteria above, resulting in concluding on a preferred cluster.
  Within this preferred cluster the postal code with the highest percentage of females was selected.
- Selected postal code is 3572 in the city of Utrecht. This postal code meets all the stated criteria.

Introduction

Background

This exercise has its origin in the IBM Applied Data Science Capstone course on Coursera and the Peer Graded Assignment concluding this course.

Assignment: 'The Battle of Neighborhoods'
'Come up with an idea to leverage the Foursquare location data to explore or compare neighborhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data to solve'.

The chosen problem is a fictional research assignment to research the optimal place to start a specified type of new business given the specified requirements.

Problem

What is the optimal neighborhood to start a new salad bar, based on a comparison between neighborhoods in the four cities considered using demographic and income data combined with data on other venues in the area, based on the following requirements and background:

- Only locations in the four biggest cities in the Netherlands are considered, being Amsterdam, Rotterdam, Utrecht and The Hague;
- From market research a projected ideal target group is defined: single household females;
- Areas with abundant availability of other restaurants are preferred as experience indicates that this attracts more potential customers to new restaurants;
- Preferably none or few competing salad bars in the area.

Target group for research results

Results of this research is targeted on the aspiring new founders of this salad bar to be established. There interest is to choose a location for their new businesses that has the highest chance of being successful given market research on target group they already did perform.

Their interest in using a data science approach for this problem is to leverage data to make the most objective decision possible.

Data

Demographic data

Given the target group requirement of single household females, demographic data is necessary for the different areas we want to compare.

As basis for the area boundaries, in the Netherlands the postal code system is a very good proxy of the different neighborhoods as it is more consistent and also much more granular than for example borough ('wijken') and neighbourhood ('buurten') names.

Dutch postal code have a four digit two letter format (1234 AB). Together with the house number it uniquely identifies an individual postal address. Good proxy for neighbourhoods are the four digits, the full postal code is on street level so very low level.

For demographic data, the Dutch Central Bureau of Statistics (CBS) is by far the most logical source. It derives it's data from the government own systems and from data delivery that companies are legally obliged to deliver to the government. Therefore, CBS data is considered to be the most accurate, complete and comprehensive.
Additionally, the CBS provides almost all data for free and open for public use in several formats and views.

After browsing available datasets, this set was selected as best fitting the requirements for this project, being demographic data on at least sex and household size per four digit postal code, where the dataset as of 1 January 2016 was the most recent to be found:

Population by four digit postal code as of 1 January 2016

https://www.cbs.nl/nl-nl/maatwerk/2016/51/bevolking-per-viercijferige-postcode-op-1-januari-2016

Format is an Excel file, fields in this file include:

- Four digit Dutch postal codes;
- Total population;
- Population split per sex and five-year age cohort;
- Migration background;
- Household size and nature of households (kids/no kids)

Geolocation data

To be able to map our postal codes, and also to be able to retrieve venue data for this postal code, geolocation data (latitude/longitude) is required for all four digit postal codes.

Amongst others on this location a file can be found exactly doing this:

4pp -Cities, postal codes and coordinates

https://git.tuxm.nl/tuxmachine/postcodes/src/4329c858db24b79523fd3fbbaf2df138ccaf16cd

Format is a Comma Separated Values (CSV) file.

Income data

To assess how affluent our prospected customers are, income data per postal code is handy as well to enrich the data.

For this, another dataset of the CBS is used:

Spendable income per postal code, 2004-2014

https://www.cbs.nl/nl-nl/maatwerk/2017/15/besteedbaar-inkomen-per-postcodegebied-2004-2014

Format is an Excel file, fields in this file include:

- Four digit Dutch postal codes;
- Number of households;
- Average spendable income;
- Spendable income standardised for householdsize.

Only the 2014, the most recent year available in this format, will be used as we will not research changes in time.

Names of areas and neighbourhoods based on postal code

As names are more easy to relate to as numbers, it is handy to enrich the four digit postal code to the area/neighbourhood/city names it relates to.

For this, again a CBS dataset is used:

Neighborhood, district and municipality 2018 for postcode house number

https://www.cbs.nl/nl-nl/maatwerk/2018/36/buurt-wijk-en-gemeente-2018-voor-postcode-huisnummer

Format is a zip file with four Comma Separated Values (CSV) files:

- pc6hnr20180801_gwb-vs2.csv with link per postal code/housenumber to neighborhoodm district and municpality codes;
- buurtnaam2018.csv with all names of the neighborhoods per neighborhood code;
- gemeentenaam2018.csv with all names of the municipality per municipality;
- wijknaam2018code.cvs with all names of the districts per district code.

Foursquare location data

To obtain information of surrounding venues for a given longitude/latitude, the Foursquare API is used.

See https://developer.foursquare.com/docs for API documentation.

This API will be used to obtain a list of the (max 100) venues surrounding the center of each postal code area and to obtain specific data on presence of salad bars.

Methodology

Also see explanation and code used on:

https://github.com/xahmol/Coursera_Capstone/raw/master/Applied%20Datascience%20Capstone%20week%204-5%20(1.ipynb

Data preparation

Dutch demographic data based on postal code

Importing of Excel file

The source file is an Excel file with a multi row header, therefore this can not directly be used as header for a dataframe. Imported therefore without a header, while adding own header colums that are adapted from the source.

Skipped header and footer rows and specified the column numbers to import as the last columns contain data we are not interested in as we will obtain them already from a different source.

Also converted all data to float apart from the postal code column to easier allow for mathematical operations.

Changing columns to percentage of total in row instead of absolute numbers

In order to make numbers comparable, all columns containing inhabitant numbers have been converted to a percentage of total inhabitants of the postal code. All household data are converted to percentages of total households.

Geolocation data

CSV while from source imported without any modification needed.

In order to reduce filesize and computing times, all postal codes we will not need are dropped:

- All postal codes with classification P.O. Box (as not people are living there) or Unknown (also no inhabitants) are dropped;
- Everything but the postal codes in the four selected cities (Amsterdam, Rotterdam, Utrecht and The Hague) is dropped.
  After doing so, investigation of the resulting dataframe shows that one postal code outlier remains, namely a city called Nieuw-Amsterdam. Logical it is not dropped as it contains Amsterdam, but we won't need it, so also dropped.

Income data per postal code

The source file is an Excel file with a multi row header, therefore this can not directly be used as header for a dataframe. Imported therefore without a header, while adding own header colums that are adapted from the source.

Skipped header and footer rows and specified the column numbers to import: only income data for the most recent year available (2014). Household data is not imported as we have that already from the demographic data source.

Names of areas and neighbourhoods based on postal code

First, the ZIP file from the source is selected and extracted. This results in four CSV files. Only three are used as we already have obtained municipality names from the geolocation data.

These files are combined by first importing the complete list of all Dutch postal codes and housenumbers linked to area, neighbourhood and municipality code. Next the three files containing the names linked to these codes are imported. Finally, the four dataframes are merged based on the postal code field. All unneeded columns are dropped: housenumber (too low level) and the codes. T

Next, a new column is made for only the four digits in the postal code as this is the level we want to work on. A dataframe is created by grouping the data on this new column. After this the full postal code column can be dropped.

Analysis on data types shows that this dataframe has the postal code as strings, while the other files have this column as integers. Therefore the column is recasted as integer to be able to merge all dataframes into one.

After merging all dataframes all unneeded columns are dropped.

To test this data, a map is created showing all postal codes on the map. For this first the latitude/longitude of the center of the Dutch Randstad area is obtained.

Resulting map:

Foursquare data on venues near to postal code center point

Next, per postal code in this dataframe, nearby venues (with a maximum of 100) are obtained using the Foursquare API. Results is stored in a new dataframe.

From this dataframe several new dataframes/stratictics are created for analysis and further use:

- Number of venues per postal code;
- Number of unique venues per category;
- One-hot encoding and, following, grouping creating mean of the venue categories per postal code to use for statistical analysis and clustering;
- Dataframe with top 10 venue categories per postal code.

Cluster postal code areas based on venue type, demographic data and income data

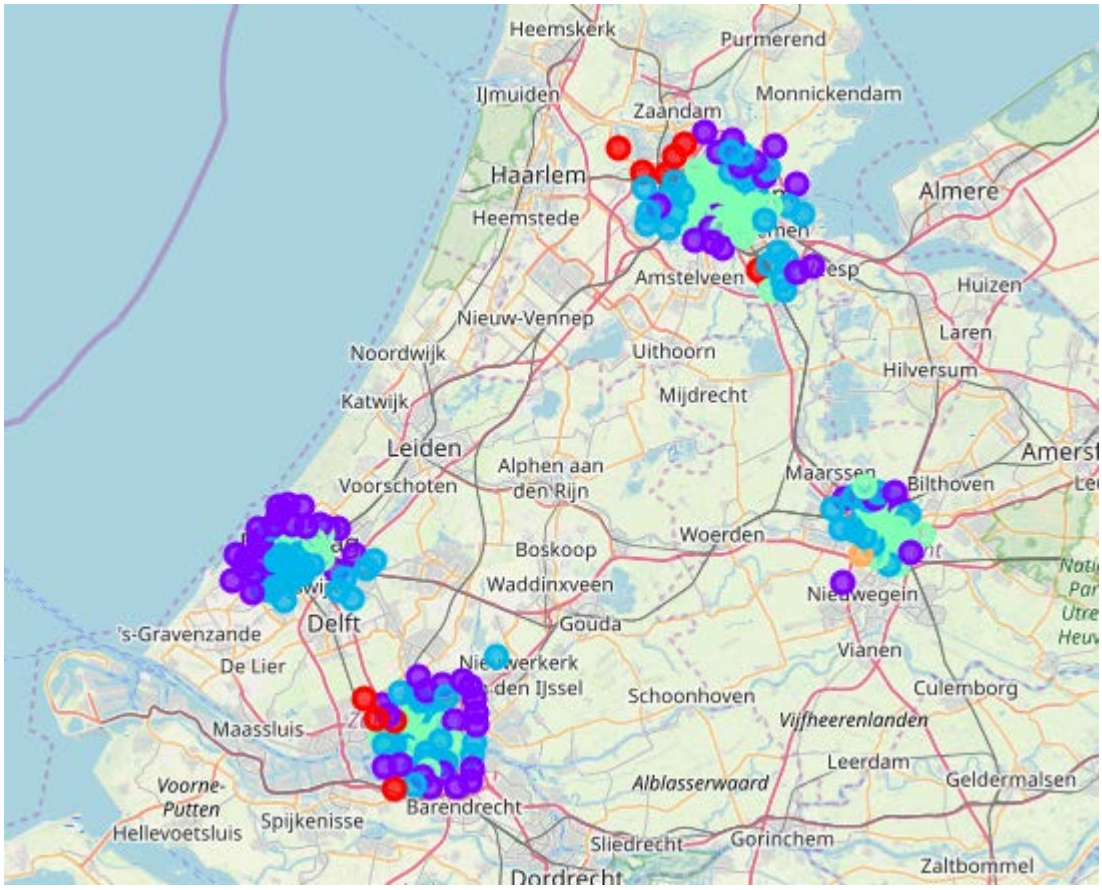
Next part is to cluster the postal codes in different clusters based on venue types present, demographic data and income data. Reason for doing so is that we want to narrow down our search for the optimal founding spot to similar postal codes having the same characteristics that we deem to be given higher chance of success.

To do this, a dataset is created with only the numerical values on venue, demographics and income. This dataset is normalized to ensure that no single characteristic is overadjusted for due to the fact it has a larger range than others. The standard scaler in the Scikit package is used.

Next, k-means clustering is performed on this dataset with k set to 5 clusters.

Cluster numbers are inserted back to the full dataframe.

The resulting dataframe is visualized using a map showing the different clusters in different colors:



Analysis and examination of the clusters is performed by comparing characteristics of each cluster.

Given our criteria for the location these aspects are reviewed:

- Nature of venues already present: a large presence of existing restaurants attracts potential customers to our salad bar;
- Habitation: larger population means more customers;
- Percentage of females and percentage of one person households: our target demographic group based on market research are single household females;
- Standarised income: higher income is favourable on spending amount.

Results are shown in the Results section, see next chapter.

Potential competing salad bars already present

For the selected cluster, a new dataframe is created of all salad bars already present. As we do not want to found close to competing salad bars, all postal codes are dropped in which salad bars are already present.

Selecting postal code with highest percentage of females

As final selection criterium, we will select the postal code with the highest percentage of females.

Final check on selected postal code

For the selected postal code, a final Foursquare query is performed to double check if really no salad bars are present in the selected postal code.

Results

Analysis of the different clusters

For full source data, statistics and code used, see:

https://github.com/xahmol/Coursera_Capstone/raw/master/Applied%20Datascience%20Capstone%20week%204-5%20(1.ipynb

Full dataset as comparison

- Average inhabitants: 9,068.463115
- Average percentage females: 49.9288%
- Average with migration background non-western: 31.2550%
- Average single person households: 49.9532%
- Average standardised income: 21.982787k

Cluster 0

Key stats:

- Average inhabitants: 63.500000
- Average percentage females: 32.5661%
- Average with migration background non-western: 22.0694%
- Average single person households: 59.8779%
- Average standardised income: 0.0

Analysis:

- Almost no inhabitants, almost no income.
- Venues include mostly harbor/marina, snack places and leasure zones
- See also map and area names: industrial and harbour area

Conclusion:
Cluster is harbour area. Not suited for settling our salad bar

Cluster 1

Key stats:

- Average inhabitants: 7884.453125
- Average percentage females: 51,4327%
- Average with migration background non-western: 19,4456%
- Average single person households: 45,2300%
- Average standardised income: 25.329687

Analysis:

- High population density
- Lower than average non-western migration
- Higher than average income
- Venues mostly leasure, sport and shopping

Conclusion:
Cluster is residential zone, few restaurants. Not ideal for a salad bar, although it has a high average income and relatively many females.

Cluster 2

Key stats:

- Average inhabitants: 10401.506024
- Average percentage females: 50.0095%
- Average with migration background non-western: 45,9970%
- Average single person households: 41,8285%
- Average standardised income: 21.854217

Analysis:

- Very high population density
- Relatively high migration background, higher than average income, but lower than cluster 1.
- Venues inludes many restaurants

Conclusion:
Cluster is densily populated city area with many restaurants. Very good candidate for opening salad bar.

Cluster 3

Key stats:

- Average inhabitants: 9815.465116
- Average percentage females: 50,1683%
- Average with migration background non-western: 26,6659%
- Average single person households: 60.7361%
- Average standardised income: 22.427907

Analysis:

- Polulation density higher than average, but slightly lower than cluster 2
- Relatively low migration background, higher than average income and higher than cluster 2, but lower than cluster 1.
- Venues inludes shops and many restaurants
- Very many single person households
- City centers

Conclusion:
Busy area with many restaurants. Very high on single person households. Very good candidate for opening salad bar.

Cluster 4

Key stats:

- Stats do not say much given very low total population of 10

Analysis:

- Very similar to cluster 0, but no harbours.
- Also almost no inhabitants, almost no income.
- Venues include mostly snack places and leasure zones See also map and area names: industrial area

Conclusion:
Cluster is industrial area. Not suited for settling our salad bar

Conclusion on cluster analysis

Best candidates for opening our salad bar are clusters 2 and 3. Both have presence of many other restaurants attract people to our restaurant. We chose cluster 3 as the percentage of single person households is much higher, also average income is slightly higher.

Potential competing salad bars already present

In the selected cluster of 86 postal codes, salad bars are already present in 6 postal codes of these. We therefore exclude those 6 from selection.

Selecting postal code with highest percentage of females

Within those 80 remaining postal codes, the postal code with the highest percentage of females is selected.

The selected postal code is 3572.

Complete data of that postcode:

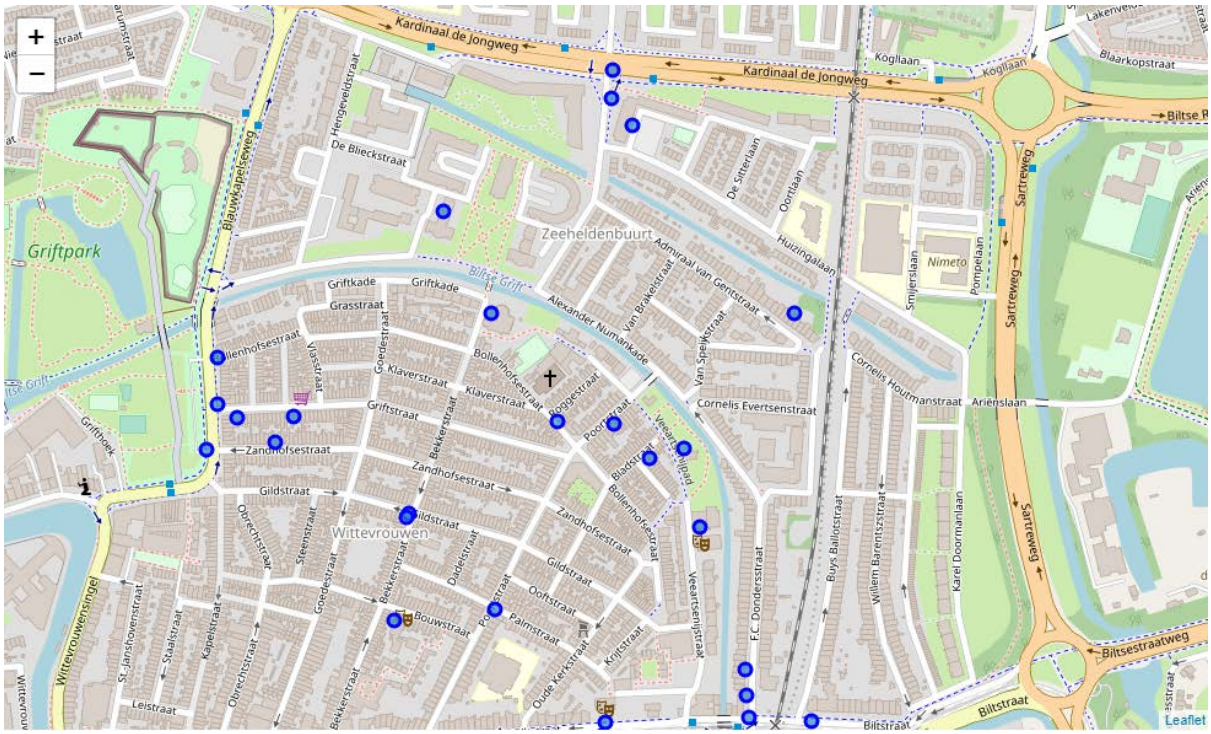
Postal code	3572
Cluster	3
District	Wijk 04 Noordoost
Neighbourhood	Tuinwijk-Oost
City	Utrecht
Municipality	Utrecht
Province	Utrecht
Latitude	52.0995
Longitude	5.13603
Total inhabitants	11,760
Percentage males	46.09%
Percentage females	53.87%
Migration background	20,75%
- western	13,48%
- non-western	7,31%
Households	6,990
- one-person	62,52%
- no kids	18,96%
- with kids	18,60%
Average household size	1.67
Average income	EUR 33,1k
Standardized income	EUR 24.9k

Age cohorts (inhabitants as percentage of total inhabitants):

Age cohort	Total	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50
Total	100,00%	5,91%	4,25%	4,17%	4,93%	13,48%	13,27%	9,95%	7,14%	6,34%	6,16%
Male	46,09%	3,06%	2,13%	2,00%	2,04%	5,10%	5,78%	5,06%	3,57%	3,10%	2,89%
Female	53,87%	2,85%	2,13%	2,13%	2,89%	8,33%	7,44%	4,93%	3,53%	3,23%	3,36%

Age cohort	50-55	55-60	60-65	65-70	70-75	75-80	80-85	85-90	90-95	95-
Total	5,57%	5,48%	4,89%	3,19%	1,70%	1,53%	1,19%	0,64%	0,21%	0,09%
Male	2,72%	2,68%	2,34%	1,57%	0,72%	0,68%	0,47%	0,13%	0,00%	0,00%
Female	2,85%	2,76%	2,51%	1,62%	1,02%	0,85%	0,72%	0,47%	0,21%	0,04%

Selected postal code mapped with other Foursquare venues plotted, no competing salad bars present:



Discussion

First thing to point out is that from working with the data delivered by Foursquare it appeared that Foursquare data is not ideal for analysing locations in the Netherlands. Given own knowledge of the areas the data is outdated (restaurants listed that are no longer there) and incomplete (much more restaurants actually present than listed by Foursquare). This analysis could therefore be improved to use a more reliable source, such as Google Places or Tripadvisor. Both are much more used in the Dutch market and as a result more accurate and up to date.
But as the assignment was to use Foursquare data, this was not investigated.

Next thing that I struggled with in making this research is that it proofed difficult to find a usable geolocation shape JSON file (or an easy source to make it) for the Dutch four digit postal codes.
I would have liked to plot Choropleth maps heatmapped on income and percentage of females, but without a GeoJSON file for the postal codes used this proved not possible.

GeoJSON files could be found for the Dutch districts and municipalities, but choosing this would have as downside that districts cover a much larger area than the four digit postal code, and therefore much less granular
Also, the datasets found were coded on postal code. Moving to districts would have meant searching for new datasets, or converting this data to districts. This would have cost much additional time, while for the actual research question it would give a less granular and so less precise result.

For clustering, only the k sizing of 5 has been tried. Further investigation could have been done in other values of k. I have chosen not to do so as this number gave a clear and recognizable distinction between clusters, while already at this size a cluster with only 1 postal code resulted. Given more time this could be further substantiated though.
Also Density Based Clustering could have been investigated using the same data. Chosen not to do so in this timeframe, also because k-means clustering already provided meaningful results.
Finally, different selection of statistics to cluster on could have been tried. Now plainly all available statistics have been selected.

Conclusion

Using readily available data it proved possible to combine, cluster and analyse data to answer the problem asked by the identified target audience.

The optimal postal code to search for a location for a new to be established salad bar has been found: postal code 3572 in the city of Utrecht.

Revisiting the original selection criteria:

Only locations in the four biggest cities in the Netherlands are considered, being Amsterdam, Rotterdam, Utrecht and The Hague

✓ Met. Selected location is in the city of Utrecht.

From market research a projected ideal target group is defined: single household females

✓ Met. The selected postal code has a very high number of females meeting this target group.

Areas with abundant availability of other restaurants are preferred as experience indicates that this attracts more potential customers to new restaurants

✓ Met. The selected postal code has numerous other restaurants.

Preferably none or few competing salad bars in the area.

✓ Met. No other salad bars present according to Foursquare.

Datasets used

Population by four digit postal code as of 1 January 2016
https://www.cbs.nl/nl-nl/maatwerk/2016/51/bevolking-per-viercijferige-postcode-op-1-januari-2016

4pp -Cities, postal codes and coordinates
https://git.tuxm.nl/tuxmachine/postcodes/src/4329c858db24b79523fd3fbbaf2df138ccaf16cd

Spendable income per postal code, 2004-2014
https://www.cbs.nl/nl-nl/maatwerk/2017/15/besteedbaar-inkomen-per-postcodegebied-2004-2014

Neighborhood, district and municipality 2018 for postcode house number
https://www.cbs.nl/nl-nl/maatwerk/2018/36/buurt-wijk-en-gemeente-2018-voor-postcode-huisnummer

Foursquare location data
https://developer.foursquare.com/docs

Jupyter Notebook

Notebook with complete analysis, explanation, all data and all code:

Github:
https://github.com/xahmol/Coursera_Capstone/raw/master/Applied%20Datascience%20Capstone%20week%204-5%20(1.ipynb

IBM Watson Studio:
https://eu-de.dataplatform.cloud.ibm.com/analytics/notebooks/v2/cce71c0f-4580-4656-9e03-7078a87b0789/view?access_token=81020cd86e14b752141b21aabc80a6e87a7974bede5ef684e9328b98cac9aaa6