# Applied Data Science Capstone Project
# The Battle of the Neighborhoods

*”Manhattan eats Finance”*

*By Fabian Schmidt*

*12/30/2020*

# 1. Introduction/Business Problem

'Restaurbank' is a small New York City-based finance company with a niche business model. They are specialized in loan-funding small and medium sized restaurant businesses within Manhattan. Restaurants are considered being a riskier industry sector than others due to higher default rates and less fixed assets that could be sold to repay the loan. Despite this fact, usually 'Restaurbank's business model works quite well as they claim to perform a more rigorous due diligence before handing out the loan. Also, they argue their collaborative approach to credit lending with advice and networking among the restaurant community makes their clients more successful.

However, due to the current events around the COVID-19 pandemic new clients who would like to start a restaurant in Manhattan are rare. At the same time, a higher client default rate is putting risk management back onto the firm's schedule. So, the company's problems are both acquiring new, solid clients as well as protecting their existing business from further defaults.

Therefore, the risk management department set up a task force with a few statisticians to assess if this new hype topic called 'Data Science' might help the firm with their current challenges. They stumbled upon the geospatial data provider Foursquare and came up with an idea on how help both their department with managing risk as well as the front office staff in their client acquisition. They scheduled a meeting with the managers of the front office and the risk management department to present their results.

In the meeting they introduced their 'Manhattan eats Finance' initiative: By using restaurant ratings on Foursquare, they want to find hot- and blackspots of great restaurants. The front office can leverage these findings to guide new clients asking for a loan to areas where there are fewer great restaurants and therefore competition is less intense. Also, they can approach the owners of top-rated restaurants to restructure their debt with 'Restaurbank', providing lower interest rates than they currently pay for their loans. The risk management department could include the restaurant rating from Foursquare in their credit risk assessment and management, assuming lower rated restaurants might not be able to withstand competition and default.

# 2. Data

To achieve this, different types of data from Foursquare will be used. At first, a list of restaurants across Manhattan is retrieved that contains the venue ID, name, its coordinates given by longitude and latitude as well as the venue's category. To do so, Foursquare is queried with the search term restaurant. Below is a screenshot of the first few rows of the resulting Pandas DataFrame. Note that the venue category also has values like 'bakery' or 'snack place'.

| Venue ID | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|
| 5894c9a15e56b417cf79e553 | Xi'an Famous Foods | 40.715232 | -73.997263 | Chinese Restaurant |
| 3fd66200f964a520bce61ee3 | La Bella Ferrara | 40.717450 | -73.998373 | Bakery |
| 5c965dad5455b2002c058659 | Yi Ji Shi Mo Noodle Corp | 40.718254 | -73.995930 | Chinese Restaurant |
| 4bcf9774a8b3a5939497625f | Shanghai 21 | 40.714423 | -73.998904 | Shanghai Restaurant |
| 4a00e0a7f964a520bc701fe3 | Singapore Malaysia Beef Jerky | 40.718527 | -73.995824 | Snack Place |

In a second step the scores for these venues are queried. To retrieve these a premium call to the Foursquare API is needed querying the details of every single venue in the list. Therefore, the amount of data is limited, as only 500 such premium calls per day are allowed with the free Foursquare developer account. Below you can find a screenshot of the resulting DataFrame including the venue ID as well as the venue's score.

| Venue ID | Venue Score |
|---|---|
| **5894c9a15e56b417cf79e553** | 8.9 |
| **3fd66200f964a520bce61ee3** | 8.9 |
| **5c965dad5455b2002c058659** | 8.9 |
| **4bcf9774a8b3a5939497625f** | 8.9 |
| **4a00e0a7f964a520bc701fe3** | 8.8 |

In the next step both both DataFrames are joined on the venue ID index resulting the data set that will be used for the further project.

| | Venue | Venue Latitude | Venue Longitude | Venue Category | Venue Score |
|---|---|---|---|---|---|
| **Venue ID** | | | | | |
| **5894c9a15e56b417cf79e553** | Xi'an Famous Foods | 40.715232 | -73.997263 | Chinese Restaurant | 8.9 |
| **3fd66200f964a520bce61ee3** | La Bella Ferrara | 40.717450 | -73.998373 | Bakery | 8.9 |
| **5c965dad5455b2002c058659** | Yi Ji Shi Mo Noodle Corp | 40.718254 | -73.995930 | Chinese Restaurant | 8.9 |
| **4bcf9774a8b3a5939497625f** | Shanghai 21 | 40.714423 | -73.998904 | Shanghai Restaurant | 8.9 |
| **4a00e0a7f964a520bc701fe3** | Singapore Malaysia Beef Jerky | 40.718527 | -73.995824 | Snack Place | 8.8 |

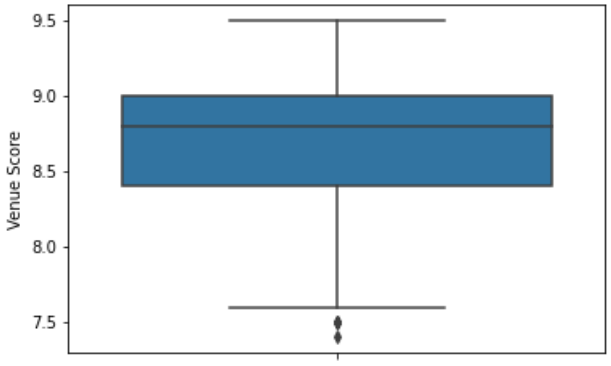
# 3. Methodology

An initial explorative analysis of the 'raw' Foursquare data was performed. An initial finding was, that there is a brought range of venue types including restaurants, banks, and shops. As we are interested only in restaurants, we should filter the results. One approach would be to retrieve all venues and then filter for restaurants. However, as you can see in the screenshot below, restaurant-like food places like bagel shops, bakeries or even steak houses would then be missed. A manual exception for all of these venue types does not seem feasible. Therefore, a different approach was chosen where we would include the search term 'restaurant' in the Foursquare API query. Foursquare's built-in logic recognizes food places that are not explicitly a restaurant venue type and returns them. In the screenshot above you can see, that Foursquare returned results that are of type bakery and snack place when searching for restaurants. Additionally, this approach helps mitigating one of the limitations of the free Foursquare account. As it limits explore results to 100 venue results, we make sure that these 100 results are all relevant for our problem. Instead we would have to filter on restaurants and would end up with lesser data.

| Art Gallery | Art Museum | Arts & Crafts Store | Asian Restaurant | Athletics & Sports | Auditorium | Australian Restaurant | Austrian Restaurant | Auto Workshop | BBQ Joint | Baby Store | Bagel Shop | Bakery | Bank | Bar | Baseball Field | Basketball Court | Bed & Breakfast | Beer Bar | Beer Garden | Beer Store | Big Box Store |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Further data analysis was done with the most important value for our problem: The venue's rating score. As we know that people are biased to giving more positive feedback but at the same time are more eager to publish their opinion when not satisfied, we should assume the score might have statistical anomalies.

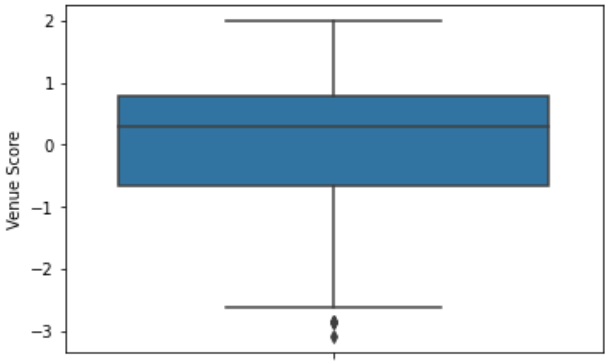
Indeed, checking the key statistics of the score yields interesting results. The score on Foursquare can range from 0 to 10. However, the received restaurant scores mean in Manhattan is 8.67. Below you can see a boxplot of the score, showing its distribution, as well as the descriptive statistics of the column. Note how they only range between 7.4 and 9.5.




```
count    401.000000
mean       8.674065
std        0.412524
min        7.400000
25%        8.400000
50%        8.800000
75%        9.000000
max        9.500000
Name: Venue Score, dtype: float64
```

Boxplot and statistics before normalization

Due to this unequal distribution of the score, we decided to normalize the score prior to processing. The Z-score was chosen as normalization approach. Below you can see the boxplot and the descriptive statistics after normalization. Observe how they the median's value is ~0, while the mean is negative. With the normalized values, above-average rated restaurants are now weighted positive, while below average restaurants have a negative weight.

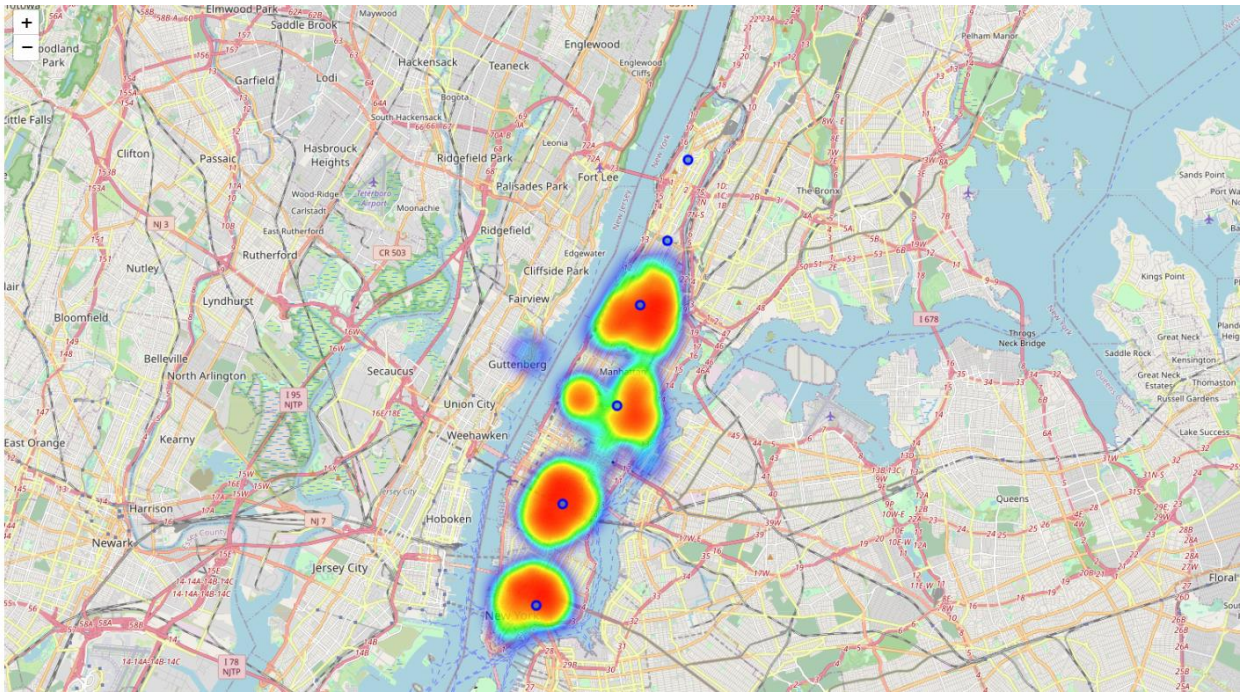



```
count    4.010000e+02
mean    -2.905960e-15
std      1.000000e+00
min     -3.088466e+00
25%     -6.643617e-01
50%      3.052799e-01
75%      7.901007e-01
max      2.002153e+00
Name: Venue Score, dtype: float64
```

Boxplot and statistics after normalization

# 4. Results

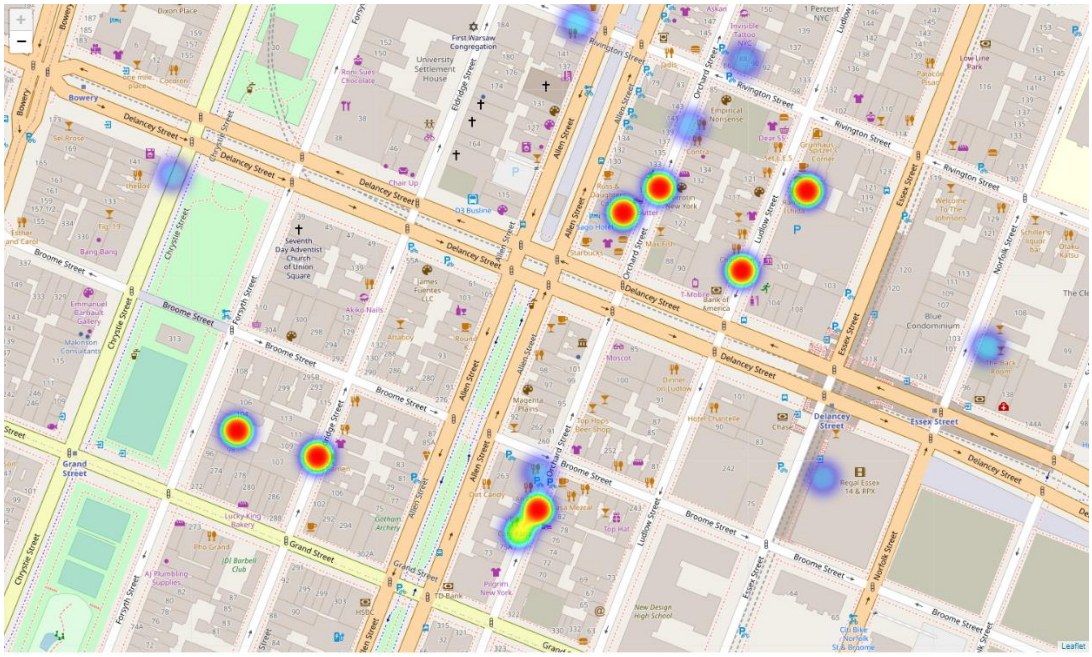
Below you can see the main result of the ‘Manhattan eats Finance’ project. A heat map of Manhattan where individual blobs represent a restaurant and their intensity depicts the restaurants score. Additional markers were set for the coordinates that were used as centres of the search areas. With this, it should be possible to spot and explain any anomalies related to the search settings like blank spots at the edges or accumulations around the centres.



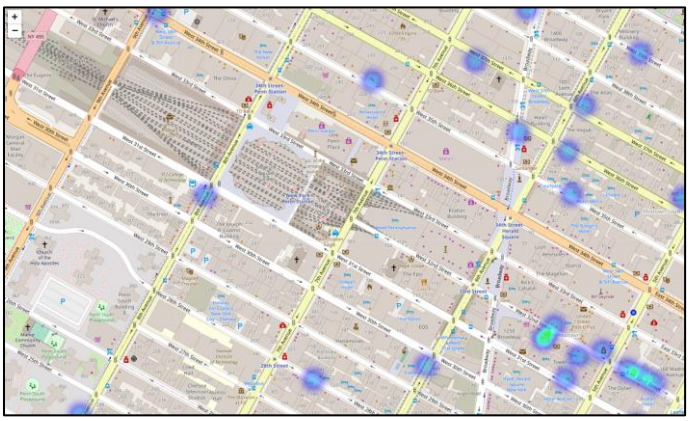

This map already provides first insights, e.g. we can see the gap at the third search area that is Central Park. We can also see that data for the Upper Manhattan search areas could not be retrieved to the API call limitations of Foursquare. Now, lets zoom in a bit.:

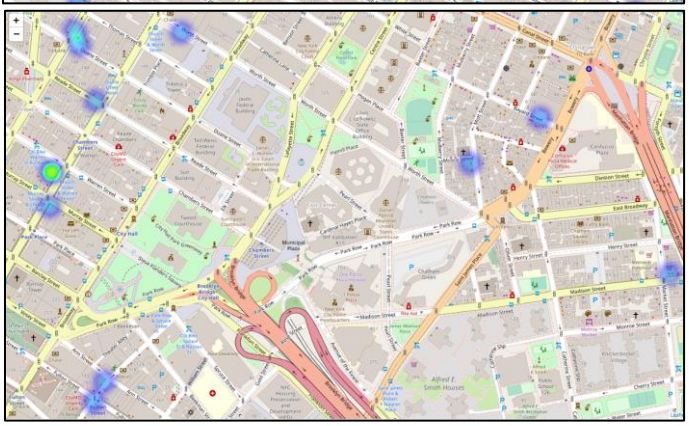
Further investigation provides detailed insights:

As you can see on the below maps, there is a heavy cumulation of highly rated restaurants around a certain part of Delancey St. New clients of 'Restaurbank' should be advised to pick different locations for their soon-to-open restaurants.

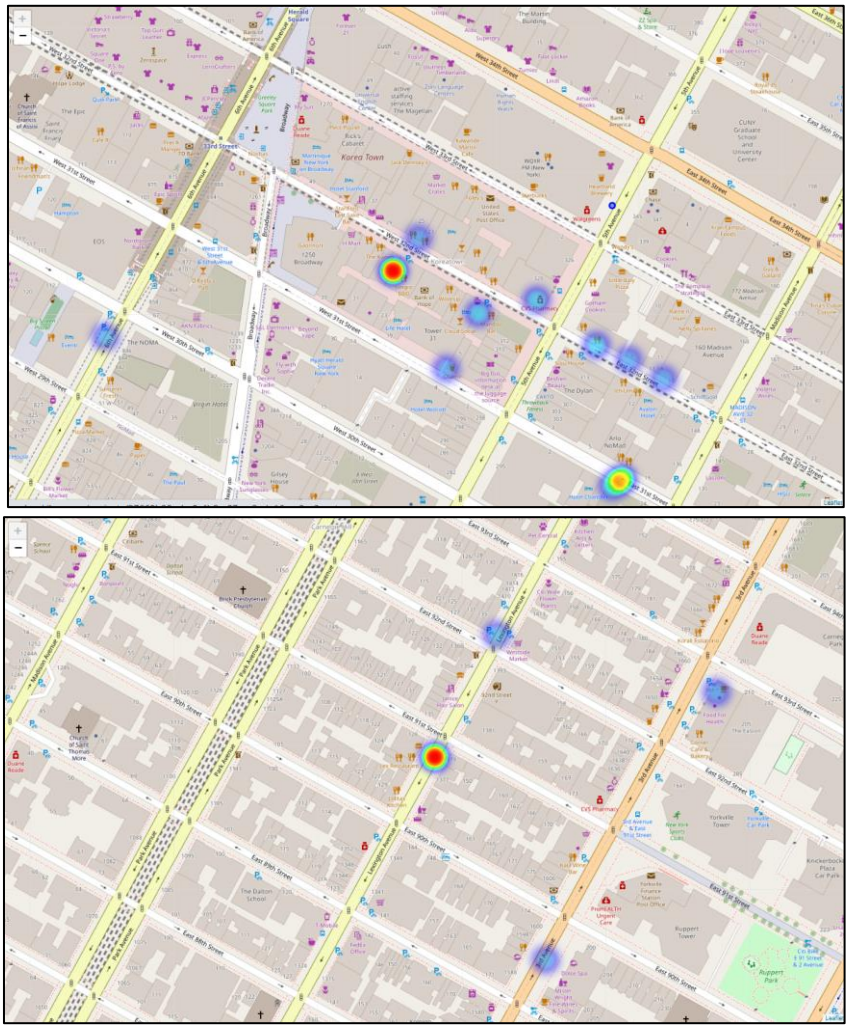



On the other hand, areas like the ones around PennStation or Civic Center look like blank spaces with only few high rated restaurants. Clients should be guided to check out some of these places in case they do not have a concrete location in mind for their restaurant. See below map sections for an illustration of these areas.





A third notable category are singular red points surrounded by less intense areas. These hint at top-rated restaurants in areas with few comparable competitors. ‘Restaurbanks’ front office staff should approach owners of these restaurants to check if they could become new clients, as we can assume their business is more solid in such an environment. Examples for this can be found at Delancey/Allen and Lexington/91st. Below are the map sections for these restaurants.



# 5. Discussion

It is obvious that above results are only a limited snapshot. Several limitations of Foursquare's free developer access do not allow a more thorough picture. Examples for this are the limitation of premium calls, which are the only way of retrieving venue scores, as well as the general result limit with explore API calls.

Also, comparison between all of the different venue types in data set might not be reasonable for 'Restaurbank'. A highly rated venue of type 'snack place' looks just the same on the map as a highly rated steak house. Just looking at the map, this might be perceived as a conflict but knowing the different venue types could lead to the conclusion that both 'restaurants' complement each other as they are visited on different types of occasions. This could be solved in a follow-up project to 'Manhattan eats Finance' by adding a layover with markers for the different restaurants that include their exact venue type.

Furthermore, with additional data the map could be more precise. Currently, the only insight on low-intensity areas is that there are few highly rated restaurants. However, this might have good reasons. For example this could be an remote, industrial area with low traffic of possible customers. This might not suit a certain type of restaurant or any restaurant at all. Here, more data is needed to provide more distinct information.

Nevertheless, the project can be considered a success as the project team was able to come up with a concise map that provides a good and easy to understand overview across Manhattan's restaurants.

The results explained in the previous chapter provide 'Restaurbank' with several approaches to secure their existing business and even expand it. However, follow-up from the different stakeholders is needed to actually implement these findings into actions.

# 6. Conclusion

Concluding this report, we can say that 'Restaurbank's risk department came up with a good idea to leverage Data Science for their business. A map of Manhattan showing restaurant scores as heatmap was produced that can be used to derive actions to solve their business problems. With high competition areas identified, credit risk assessment for clients in these areas can be adjusted and pre-emptive actions might be taken so these clients do not default. At the same time, new business opportunities were provided as highly rated restaurants can be identified as single high-intensity points in low-intensity areas. These assumedly solid business can be approach. Other new clients without existing restaurants can be guided to low-competition areas where these restaurant could possibly flourish better with less competition.

Initial ideas for further investigation were already discussed in the above section. Furthermore, the project served as knowledge building for the staff, enabling them to leverage the used data and technologies further. For example, if the Foursquare restaurant score proofs to be an indicator of client defaults it may be included into 'Restaurbank's credit risk assessment model. Another approach leveraging map views could be to plot defaulted customers on the map and see if there are anomalies observable.