Web Scraping Report

MIDFLORIDA Branch Site Selection

By Justine Saylor

4-23-24

TABLE OF CONTENTS

Contents

TABLE OF CONTENTS 2
TABLE OF FIGURES 2
TABLE OF FILES 3
I. Introduction 5
II. What is Web Scraping? 5
III. Selecting Bank Locations 5
IV. Solution Paths 6
V. Manual Approach 6
VI. Automated Approach 7
VII. Analyzing Results 8
VIII. Selected Businesses 8
IX. Further Recommendations 9
X. Conclusion 10
XI. References 10

TABLE OF FIGURES

Figure 1: Adding all branches to Google Maps for visualization. 6
Figure 2: Demonstrating the manual method using Google Map Developers. 7
Figure 3: Demonstrating the automated method using G Maps Extractor browser extension. 8

TABLE OF FILES

1) Bank Locations for Scraping.docx
   a) Lists the branch names and addresses of all 45 bank locations selected for the dataset.
2) Manual Folder
   a) Manual Method Other Banks.xlsx
      i) Lists the branch name, branch address, along with the business name found, the business category, and number of locations for that business for the five Wells Fargo, five Truist, and five Suncoast locations selected.
   b) Manual Method Successful Midflorida Branches.xlsx
      i) Lists the branch name, branch address, along with the business name found, the business category, and number of locations for that business for the fifteen successful Midflorida locations selected.
   c) Manual Method Unsuccessful Midflorida Branches.xlsx
      i) Lists the branch name, branch address, along with the business name found, the business category, and number of locations for that business for the fifteen less successful Midflorida locations selected.
   d) Analyzed Results Manual Less Successful MidFlorida.xlsx
      i) Lists all of the businesses found by the less successful MidFlorida branches and summarizes the businesses by more than one branch. Includes the business name, number of branches out of 15 that business is by as well as the percentage.
   e) Analyzed Results Manual Successful MidFlorida.xlsx
      i) Lists all of the businesses found by the successful MidFlorida branches and summarizes the businesses by more than one branch. Includes the business name, number of branches out of 15 that business is by as well as the percentage.
   f) Analyzed Results Manual Other Banks.xlsx
      i) Lists all of the businesses found by the other bank branches and summarizes the businesses by more than one branch. Includes the business name, number of branches out of 15 that business is by as well as the percentage.
3) Automated Folder
   a) G-Maps-Extractor-10….xlsx
      i) Top ten results of G Maps Extractor for Restaurants or Stores category listing the business info including name, address, categories, website, etc.
   b) Automated Scrape….xlsx
      i) Combines the results of all the G-Maps-Extractor-10 files for that part (successful, less successful, other banks).
   c) Analyzed Results….xlsx
      i) Lists all of the businesses found by that part and summarizes the businesses by more than one branch. Includes the business name, number of branches out of 15 that business is by as well as the percentage.
4) All Results Combined for Automated and Manual.xlsx
   a) Combines all of the “Analyzed Results” files with the method used and summarizes the businesses by more than one branch. Includes the business name, number of branches out of 15 that business is by as well as the percentage.
5) Summarized Results.xlsx
   a) Summary based on automated or manual method, the top three businesses that are by all banks or businesses that are by successful MidFlorida but not by less successful MidFlorida.

6) Selected Businesses Scraping Folder
    a) … Full Scrape.xlsx
        i) Exported data file from the WebScraper.io results for each webpage.
    b) … Reorganized.xlsx
        i) Results of using the Ablebits add-in in Excel to transform the data from the full scrape file for each webpage.
    c) … Locations Florida.xlsx
        i) The final file with all the locations in Florida and the business info for each of the businesses.
    d) … _Locations_Florida.xlsx
        i) All the locations in Florida and the business info with the latitude and longitude added for each of the businesses.
7) Outscraper 50 Results by Arcadia MIDFLORIDA.xlsx
    a) Shows the first 50 results of all categories by the Arcadia MIDFLORIDA using the Outscraper website.

I. Introduction

MIDFLORIDA Credit Union requested new techniques towards solving their problem of selecting a location for a new branch site. My part of this project used Web Scraping as a new tool in data collection for MIDFLORIDA. The purpose of my part is to use Web Scraping along with a manual approach to determine a convenient location for customers based on the businesses they frequent. Time equals money; therefore, a majority of people will choose to be a member of a bank that is in the most convenient location for them. A convenient location could be by their favorite restaurant (maybe to get cash for a tip), next to the pharmacy where they get their prescriptions, or next to the grocery store where they do all their shopping. Most people will not make a special trip just to go to the bank, so in order to make a branch successful it has to be by common businesses. The businesses that were selected based on the data collection were then added to the visualization map alongside the other sub team data including demographics, market saturation, and traffic analysis to help determine a new branch site location.

II. What is Web Scraping?

Web scraping is a technique used to collect content and data from websites using bots that are programmed to visit websites, grab the relevant pages, and extract useful information. Web scraping works by making an HTTP request to a server, extracting and parsing the website's code, and saving the relevant data locally. There are many different categories of Web scrapers, but for this project specifically I used scraping extensions. Scraping extensions allow the user to scrape web pages as they navigate in real-time and are the simplest web scraping tools. Typically, free and require little prior knowledge of web scraping.

III. Selecting Bank Locations

To begin this project, I first needed to decide on which banks I would consider as part of my dataset. To do this, I used the "Model_CSV" file where Madi added a column of one's and zero's indicating whether the MIDFLORIDA branch had an average contribution that was greater than or equal to one-hundred thousand. Due to the limit of time and resources I picked 15 successful MIDFLORIDA branches (greater than 100K) as well as 15 less successful MIDFLORIDA branches (less than 100K). I also randomly selected five Suncoast Credit Union locations, five Truist locations, and five Wells Fargo locations to give a total of 45 bank locations that I would be comparing. I did this by using Google Maps and saving the branches to a list to make sure I had a wide variety over the state of Florida as seen in figure 1. The list of branch names and addresses selected can be found in file 1.

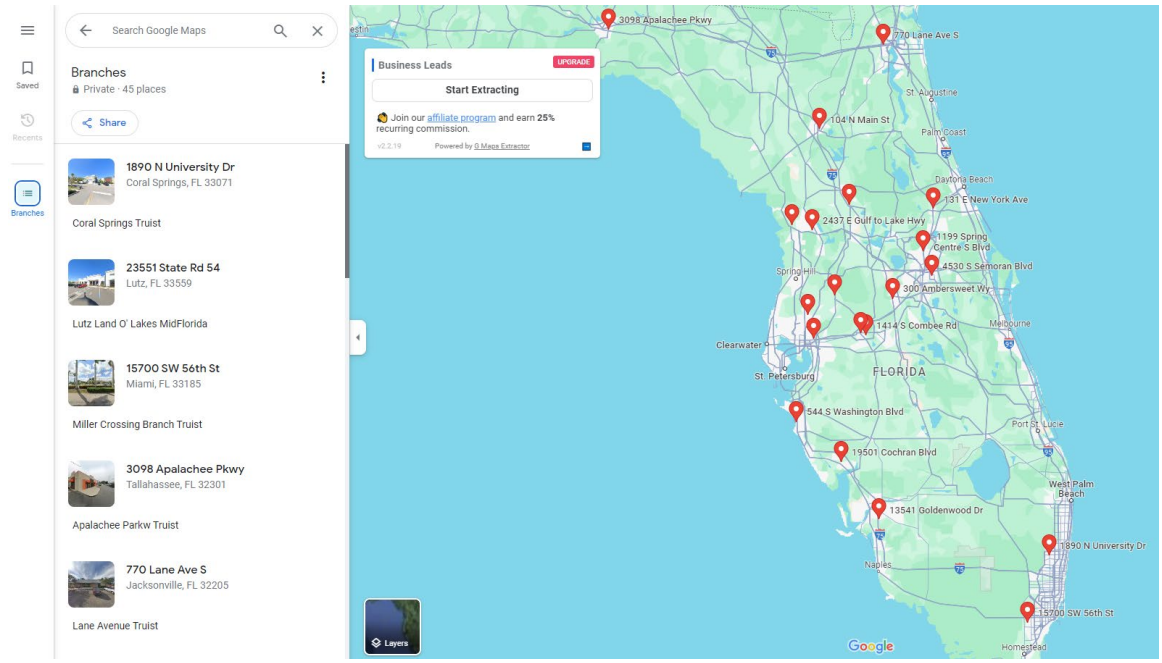

Figure 1: Adding all branches to Google Maps for visualization.

IV. Solution Paths

In the ideal scenario, there would be a website that exists where you could enter an address and a radius, and the website would output a list of all franchised businesses within that radius. Unfortunately, during my first semester of research I was unable to find such an application. Instead, I found many applications that work based off categories since Google Map uses categories to search for businesses. I decided to use one of these applications as my automated approach, and another website where you could select the radius but not the category as my manual approach. The reason I decided to use both an automated and manual approach, is because the manual approach is slower but more precise (since the radius is specified), while the automated approach is faster and gathers less data (limited to categories).

V. Manual Approach

The manual approach involved using a website called Google Map Developers using their "Draw a Circle Tool". The first step was entering the bank location address and then I decided to set the radius to two miles. The second step is clicking the "Zoom to Address" button along with the "New Circle" button. The demonstration of the first two steps can be seen in figure 2. As a note when using this website, in a rare few cases sometimes the center of the circle was not at the correct address, so make sure to zoom in all the way to verify when using this tool. The final step was zooming into the top of the circle and working my way left to right/top to bottom recording the franchises in an Excel document. I repeated these steps for all 45 bank locations. The data was broken up into three parts (15 locations each): less successful MIDFLORIDA, successful MIDFLORIDA, and other banks (Truist, Suncoast, Wells Fargo) and can be found in the Manual folder.

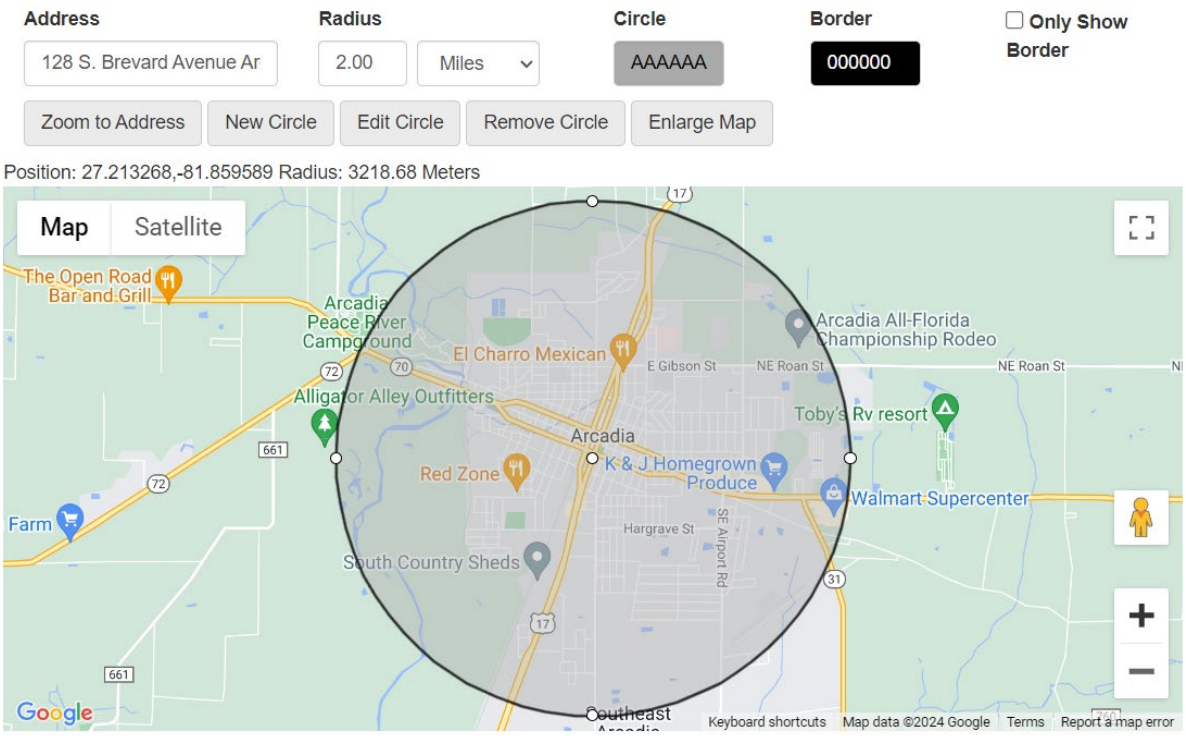

Figure 2: Demonstrating the manual method using Google Map Developers.

VI. Automated Approach

The automated approach involved using a Web scraping extension called G Maps Extractor and adding it to my Chrome browser. How this extension works is when you search a category in Google Maps the extension is able to extract the first ten search results to an Excel file. Unfortunately, the free version only allows you to scrape 1,000 results for free each month. Therefore, I decided to search two categories of "Stores" and "Restaurants" for all 45 bank locations leading me to scrape 900 results to leave room for error. Once I downloaded the extension, the first step was opening Google Maps and typing in an address of a branch. The second step was clicking the "Nearby" button on Google Maps to try and get closer businesses. The third step was entering the category "Restaurants" and clicking the "Start Extracting" button on the extension. The demonstration of the third step can be seen in figure 3. The fourth step was clicking the "Export Detailed List" button and then the "Reset" button to reset the extension. The final step was entering the category "Stores" and clicking the "Start Extracting", "Export Detailed List", and "Reset" buttons. I then repeated these steps for all 45 bank locations. The data was broken up into three parts (15 locations each): less successful MIDFLORIDA, successful MIDFLORIDA, and other banks (Truist, Suncoast, Wells Fargo) and can be found in the Automated folder.

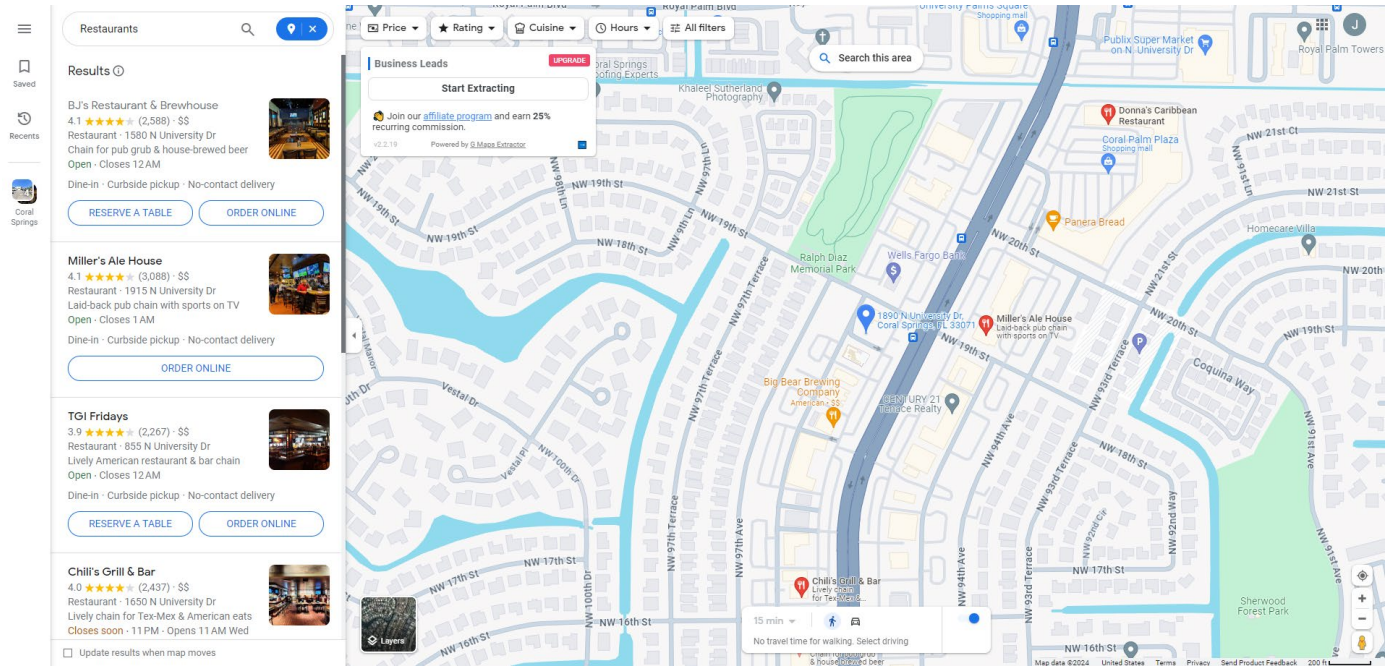

Figure 3: Demonstrating the automated method using G Maps Extractor browser extension.

VII. Analyzing Results

Once I had all my results analyzed individually for the automated and manual methods it was time to pull it all together. I first created the "All Results Combined" file (file 4) that essentially combined all of the "Analyzed Results…" files from both methods. File 4 consists of the automated and manual results for the successful and less successful MIDFLORIDA branches, as well as the other banks. These results include all the businesses that were by more than one branch and includes the business name, category, and number of branches that business is by out of 15, as well as the percentage. Once I created this file I found it difficult to decide on what businesses to select to add to the final visualization map, which is why I created "Summarized Results" file (file 5). After discussing with Dr. Dewey, I determined that the most important businesses were the businesses that were by all banks (successful MIDFLORIDA, Wells Fargo, Suncoast, Truist). In addition, the businesses that were found to be by successful MIDFLORIDA branches, but not by less successful MIDFLORIDA branches. Using file 4 I was able to create file 5 summarizing based on the automated or manual methods and listing the top three results. By looking at businesses that are by all banks, it can help to narrow down the census tracts we are looking at, by not focusing on census tracts that do not have these staple businesses. By looking at businesses that are by successful MIDFLORIDA but not by less successful MIDFLORIDA, we can help narrow down a specific area/intersection that contains the businesses that could make MIDFLORIDA more likely to be successful.

VIII. Selected Businesses

I used the file I created "Summarized Results" (file 5) to decide which businesses to add to the visualization map. The first business I did was Publix since Publix had been brought up many

times by our sponsor as a business they consider when looking at a location. Based on the data I collected, I then had the evidence to support this as Publix was by 87% of all banks. Luckily, our sponsor was able to get all the locations of Publix in the US from a friend, so I just had to create a file with only the Florida locations. The next business I tried to do was McDonald's since it was by 100% of all banks but unfortunately I was not able to find all the locations in Florida available online. The next business I did was Dollar General using their website and another Web scraping browser extension Webscraper.io. Webscraper.io is by far the simplest and easiest Web scraper browser extension I have found. The Chrome plugin is completely free and there are many YouTube videos along with documentation on how to use the extension. The next business I did was Cheddar's Scratch Kitchen since it was by 53% of successful MIDFLORIDA branches and by no less successful. The next business I tried to do was CVS which I was able to find the website that listed all the locations, but unfortunately I could not get the Web scraper browser extension to work on this site due to the backend of how the website was made. The next business I did was Advance Auto Parts since it was by 80% of all banks. The final business I did was Beef 'O' Brady's since it was by 47% of successful MIDFLORIDA branches and by no less successful. To use Webscraper.io on Google Chrome select the top right corner drop down and select More tools > Developer tools and there will be a tab for Web Scraper. For the Dollar General, Cheddar's, Advance Auto, and Beef 'O' Brady's websites, in order to get the address data that I wanted, first I had to click the link to that specific location's website as it wasn't listed on the main page. Luckily, Webscraper.io has a function for this and I was able to find a YouTube video to demonstrate this process of using the Pagination Selector. While as I stated with CVS this extension is not able to work for every website, I highly recommend this free tool. For further explanation along with step-by-step tutorials on how to get started, visit their website for documentation. Once I exported the data from all these sites, I needed to get the addresses converted to latitude/longitude in order for Anthony to add the businesses to his visualization map. Thankfully, Jeffrey was able to do this as I was unable to find a free way to do the conversion myself. Because of the way some of the websites were set up, the extraction of data did not come out into the specific rows and columns that I wanted. To fix this I used an Excel add-in called Ablebits where I specifically used the "Merge Duplicates" and "Split Names" features.

IX. Further Recommendations

As mentioned in the Solution Paths section of this report, the ideal application for this problem does not exist yet. One further recommendation is to find a way to create this application or to continue to look into other Web scraping software as they advance. One website I did find promising was Outscraper which allows you to enter a address and search all possible categories and extract all the results. However, this website does not allow you to specify a radius and of course costs money. I have included a file which shows a test run I did of this website as proof of concept where I only extracted the first 50 results but searched all categories (file 7). Another recommendation is to compare the businesses around the rest of the successful and less successful MIDFLORIDA locations along with more locations for other banks. My final recommendation is to use a tool like G Maps Extractor or Outscraper where once a potential

location is determined, you can further look into the businesses around that location using these tools.

X. Conclusion

Overall, I believe Web scraping is a powerful tool that can be used in many different ways and can be used for many applications. I also think comparing the businesses around potential sites is a major component in the potential profitability of that site. Possibly learning more about MIDFLORIDA customers and when they tend to stop by the bank could also be useful in solving this problem. I appreciate this opportunity of being able to learn something completely new and to work with MIDFLORIDA.

XI. References

Chouinard, Jean-Christophe. "Web Scraping Tutorial (with Examples)." *JC Chouinard*, 7 May 2023, www.jcchouinard.com/web-scraping/.

Hillier, Will. "What Is Web Scraping? A Complete Beginner's Guide." *CareerFoundry*, 13 Aug. 2021, https://careerfoundry.com/en/blog/data-analytics/web-scraping-guide/.