NTNU

Kunnskap for en bedre verden

Department of Computer Science

TDT4259 - Applied Data Science

Starbucks Coffee Shop Location Recommendations

Group: 67

Author:

Name	**Student ID**	**Email**
Amin Seffo	105222	ahmadase@stud.ntnu.no
Damian Roggensinger	105331	damianro@stud.ntnu.no
Dominik Nicolai Hahn	105546	domininh@stud.ntnu.no
Eng Chong Yock	105427	chongye@stud.ntnu.no
Kim Roger Gjertsås	512960	kimrgj@stud.ntnu.no
Philipp Peron	105892	philippp@stud.ntnu.no

November, 2023

Table of Contents

List of Figures

List of Tables

1 Introduction

This chapter sets the stage for the report, introducing the scenario, outlining the context, contributions, and the defined problem. We further introduce the team and their roles, concluding with a brief overview of the report's structure.

1.1 Context

Poor business location selection is a leading cause of business failures in America, underscoring the importance of meticulous site selection [1, 2]. It is especially the case for industries like traditional manufacturing firms, knowledge-based ones, and fast-growing companies as well as small retail shops such as coffee shops [1, 3]. For the mentioned industries, the criteria affecting a good site location vary based on their specific needs and target markets. Factors such as proximity to suppliers, accessibility for transportation, and local labor availability play a crucial role in the success of traditional manufacturing firms. Conversely, companies that rely on knowledge and experience, and are expanding rapidly, might give preference to areas with a dynamic pool of skilled individuals, proximity to innovation centers, and a business environment that fosters support. Meanwhile, small retail shops like coffee shops might focus on the overall ambiance of the area and nearby competitors [1, 3]. For them, it is also important to consider that factors like demographics and urban planning are dynamic and change over time [4].

A reason why the coffee shop business is particularly interesting is the existence of significant competition due to the high number of cafés. This makes choosing the right location especially important, as Ardhi [5] explained. The competition doesn't just come from independent local coffee shops but also from big franchise chains. Because of this reason, the location and the criteria that affect the decision are getting much attention in research [1, 4, 6].

For this group project, we assumed to work for Starbucks and were tasked with identifying new potential locations for a Starbucks branch in Philadelphia using Data Science. After gaining a first data understanding, this scenario was being chosen as Starbucks is a highly represented franchise in the Yelp dataset [7] as well as in the literature. Furthermore, we chose Philadelphia since the city represents a location for many Starbucks' coffee shops. This works main contributions are as following:

- We analyze different approaches mentioned in related literature, which provide criteria for the site location and use the most relevant criteria to identify optimal store locations.
- We present an interpretable data-driven approach to select new Starbucks coffee shop locations.
- We develop a pipeline for location analysis with visualization tools such as an interactive online map.
- We provide actionable recommendations for the site location for new Starbucks locations, which can also be transferred in a broader way to comparable businesses in the coffee industry.

1.2 Problem Definition

As mentioned in the introduction, the store location is one of the most critical factors for the success of a business in the coffee industry. A well-located business can attract a steady stream of customers, while poorly-located businesses may need help to stay profitable. As more consumers turn to online resources to decide which business they choose, the importance of digital platforms cannot be underestimated. A relevant platform, which many users trust for this case is Yelp [8]. Yelp is a popular review and recommendation platform that provides valuable information about businesses, stores, and services. It contains a wealth of data that can be useful for both consumers and the business owners themselves [9].

This project delves into a data-driven approach to address this need. Leveraging the datasets from Yelp [7] and other sources, this project aims to provide actionable recommendations for selecting optimal store locations, and with this increase the prospects of business success. As the criteria for an optimal store location vary for different industries, this report focuses on a single industry. The Yelp dataset contains many entries, especially in the food and drinks business category. Furthermore, the focus will be narrowed down to the coffee industry and within this industry, we will focus on Starbucks as a franchise. This is because Starbucks is a well-established company and the most represented franchise within the dataset. The Yelp dataset contains data on 11 metropolitan areas in North America. In this report, we only focus on one metropolitan area from which the findings are then derived. Philadelphia is chosen from the 11 areas in the dataset as the representative area because the density of Starbucks shops is the highest in this city.

Starbucks corporation is a 1971 founded multinational chain of coffeehouses with its headquarters in Washington [10]. By November 2022, the company's stores are located in 80 countries with an total amount of 35,711 stores [10]. As Starbucks faces strong competition in the coffee industry, the company has to evaluate different vital factors for business success to create a competitive advantage [11]. To investigate these factors, we have utilized data analytics along with machine learning and geographical insights, to offer a solution to the traditional challenge of selecting locations for business expansion or establishment from the view of Starbucks in the coffee industry.

1.3 Team Description

Name	**Experience**
Amin Seffo	Amin is an exchange student from TUM Germany studying a Master's degree in Robotics, Cognition, Intelligence. He also works as a Data Scientist at IBM.
Damian Roggensinger	Damian is an exchange student from Winterthur Switzerland studying a Bachelor's degree in electrical engineering. He works as a project manager in an engineering company.
Dominik Nicolai Hahn	Dominik is an exchange student from Aalen Germany, studying a Master's degree in business informatics with a focus on Data Science.
Eng Chong Yock	Eng is an exchange student from Nanyang Technological University Singapore studying a Master's degree in electrical engineering.
Kim Roger Gjertsås	Kim studies a Master's degree in informatics at NTNU, Trondheim.
Philipp Peron	Philipp is an exchange student from TUM studying a Master's degree in electrical engineering with a focus on AI and Machine Learning.

Table 1: Studies and experiences of the different group members.

1.4 Roles and Responsibilities

To prepare this report, several roles had to be occupied, and those changed throughout the project's progress. First, the dataset and the overall objective of the report had to be decided. This decision also involved us in conducting an initial data analysis to evaluate its suitability for the identified business objective. As this step was crucial for the understanding of the overall objective of the report as well as the whole project, we made sure everybody in the team was responsible for performing this step on their own and then the findings were discussed. After this, the team was divided into two sub-groups with different responsibilities:

- **The research group** was responsible to analyze related approaches and identify criteria, which are relevant for the achievement of the overall objective. Furthermore, the research group was responsible for outlining the Introduction and the Background section. This group consisted of Kim and Dominik.
- **The analysis group** was responsible to take the criteria identified by the research group and evaluated them based on the values of the datasets. Furthermore, the analysis group was also responsible for creating the store location recommendation algorithm. This group consisted of Amin, Damian, Eng, and Philipp.

Finally, after the two groups had performed their tasks in parallel, we reunited as a team and wrote the remaining sections of the report based on the data knowledge of the analysis group together with the business knowledge of the research group.

1.5 Outline

The report is structured as follows, in chapter 2, we outline the background including the objective definition and the used data strategy. Chapter 3 highlights the method and analysis used by describing the evaluated datasets and the related preprocessing steps, the recommendation engine, and the methods and tools used for its creation. Chapter 4 describes the evaluation and interpretation. It includes the analysis of the different criteria, which are evaluated as a base for the recommendation engine, and the evaluation of the engine itself. This is followed by an overview of an implementation plan and recommended actions for the approach in Chapter 5. This chapter also includes the limitations and possible future work, which emerged throughout the creation of this approach. Then in chapter 6, the possibilities for monitoring and maintenance are outlined. Those include various key performance indicators, a risk management and assessment plan, necessary human intervention steps, and lessons learned throughout the creation of the recommendation engine. Finally, in chapter 7, we conclude our report by summarizing the key findings derived from our analysis. These findings hold significant implications for retail strategies, offering valuable insights into effective location-based decision-making.

2 Background

This section addresses the objectives and purpose of our work. Furthermore, it offers an overview of the design strategy, along with justifications for the decisions taken and tools used for its execution.

2.1 Objective Definition

The success of a store, mainly Starbucks coffee shops in the Philadelphia area, hinges significantly on its location. In this report, we rely on the Yelp dataset as the cornerstone for evaluating various store location criteria. These criteria draw inspiration from industry best practices and offer a comprehensive understanding of prime locations for new store openings.

Our analysis addresses the following question: How can data be harnessed to provide recommendations for optimal Starbucks store locations in Philadelphia? Store location's critical role in restaurant success is underscored in this project, where digital platforms like Yelp serve as vital tools for informed decision-making. Leveraging data analytics and geographical insights, we present actionable recommendations that empower restaurateurs to identify locations with high potential for a profitable return on investment derived from our interpretable recommendation engine. These insights not only enhance accessibility but also help capture a larger market share, ultimately boosting overall profitability.

With this in mind, our project's objective is clear: **How can we leverage data to suggest optimal locations for new Starbucks stores in Philadelphia for optimizing business results?**

2.2 Store Location Criteria

In addressing this issue, it was necessary to identify various criteria for determining the optimal store location. Consequently, we analyzed existing research on store location recommendations, and the prevalent standards were extracted and incorporated into the proposed recommendation engine. An overview of the most common location factors and their occurrence in literature is shown in table 2.

	Foot Traffic			Demography			Competitors
Source	Park	University	Bus/Train	Young	Population	Income	Competitors
[1]			X	X	X	X	X
[12]	X	X	X		X		X
[13]		X					X
[14]		X		X		X	X
[15]	X	X	X				
[16]				X			
[17]						X	X
[18]					X	X	X
[19]			X		X	X	X
[20]			X		X		X
[21]					X	X	X
[22]					X		X

Table 2: Incorporated criteria for coffee shop locations mentioned in related works.

The derived criteria were the most common within the existing literature. Those criteria were then further clustered into three main groups: foot traffic analysis, demography analysis, and competitors. However, the literature also mentioned criteria, like car parking availability, which are not included in this report as they were infrequent in the literature or no datasets were publicly

available. For each criterion incorporated from the relevant approaches in this study, a dedicated dataset is integrated into the recommendation engine. The outline of those datasets can be found in section 3. The following aspects of the defined criteria increase the probability of a successful location according to the literature:

- **Foot Traffic**: A short distance to parks, universities, and bus stops / train stations is an indicator of increased foot traffic, which in turn is advantageous for companies like Starbucks.
- **Demography**: Starbucks' marketing is targeted at a younger population with a high socioeconomic status (measured by income). The more people living in an area, the better.
- **Competitors**: Larger distances to competitors are advantageous for the business as potential customers have fewer options to choose from.

2.3 Data Strategy

Developing a robust data strategy is essential for gaining valuable insights and determining the necessary variables to adjust in pursuit of our objectives. The credibility of the findings presented in this paper also hinges on the judicious selection of an appropriate data framework.

In our evaluation, we considered three prominent data strategies: the Business Analytics Methodology (BAM), the Cross-Industry Standard Process for Data Mining (CRISP-DM), and the Sample-Explore-Modify-Model-Assess (SEMMA). All three are widely recognized analytical approaches, and we carefully weighed their respective merits and drawbacks before making our choice.

Ultimately, we opted for CRISP-DM as the data strategy for this project. BAM was deemed inadequate in a technical sense, as it primarily addresses objectives at a strategic level [23]. On the other hand, SEMMA was viewed as excessively focused on technology, neglecting the strategic dimension of our objectives and concentrating solely on technical aspects such as data mining [24, 25].

While SEMMA excels in data sampling, CRISP-DM encompasses a more comprehensive understanding of strategy. It is noteworthy that CRISP-DM may demand regular updates and is often documentation-intensive. However, given the constraints of our project's timeline, we deemed these concerns manageable.

2.3.1 CRISP-DM

The CRISP-DM reference model provides an overview of the life cycle of projects in the context of data mining [26]. This model contains the different phases of a project, the respective tasks, as well as the created outputs, in this case, the entire life cycle of the data mining project is split into six phases [26]. It has to be noted that the sequence of the phases is not strict. The phases of the CRISP-DM, according to Wirth and Hipp [26], are the following and can be seen in figure 1:

- **Business Understanding**: The first phase centers on grasping the project's business objectives and requirements. Subsequently, the acquired knowledge is transformed into a data mining problem definition and an initial project plan crafted to attain these objectives.
- **Data Understanding**: The data understanding phase begins with an initial data-gathering step and progresses with endeavors to acquaint oneself with the data. These efforts involve identifying data quality issues, gaining preliminary insights from the data, and pinpointing intriguing subsets that can lead to the formulation of hypotheses about concealed information.
  There is a close link between Business Understanding and Data Understanding as the formulation of the problem in data mining projects, as well as the definition of the project plan, require at least a minimal understanding of the data.

- **Data Preparation**: The data preparation phase encompasses all tasks aimed at crafting the prepared dataset that will be used as input for modeling tools, starting from the initial raw data. The data preparation activities may be executed iteratively and without a specific sequence. They involve tasks such as selecting tables, records, and attributes, cleaning the data, creating new attributes, and transforming the data to make it compatible with modeling tools.
- **Modeling**: During this stage, various modeling techniques are chosen and applied while fine-tuning their parameters to achieve the best results. Typically, for a given data mining problem, there exist several techniques to choose from. It is important to note that certain techniques may require specific data formats.
- **Evaluation**: This phase follows the development of one or possibly multiple models that seem to exhibit high quality in terms of data analysis. However, before moving forward with the final deployment of the model, it is crucial to conduct a more comprehensive evaluation of the model and review the steps taken to build it, ensuring that it effectively aligns with the business objectives. A primary aim is to ascertain whether any critical business aspect may have been overlooked. By the conclusion of this stage, a determination regarding the utilization of the data mining results should be made.
- **Deployment**: Typically, crafting the model is not the project's result. Usually, the insights acquired must be structured and delivered in a manner that allows the customer to put them to practical use. The deployment phase can vary significantly based on the requirements, ranging from a straightforward report generation to the establishment of a recurring data mining procedure. In many instances, the user, rather than the data analyst, will be responsible for executing the deployment steps. Nevertheless, it is essential to have a clear understanding upfront of the actions required to effectively leverage the models that have been created.

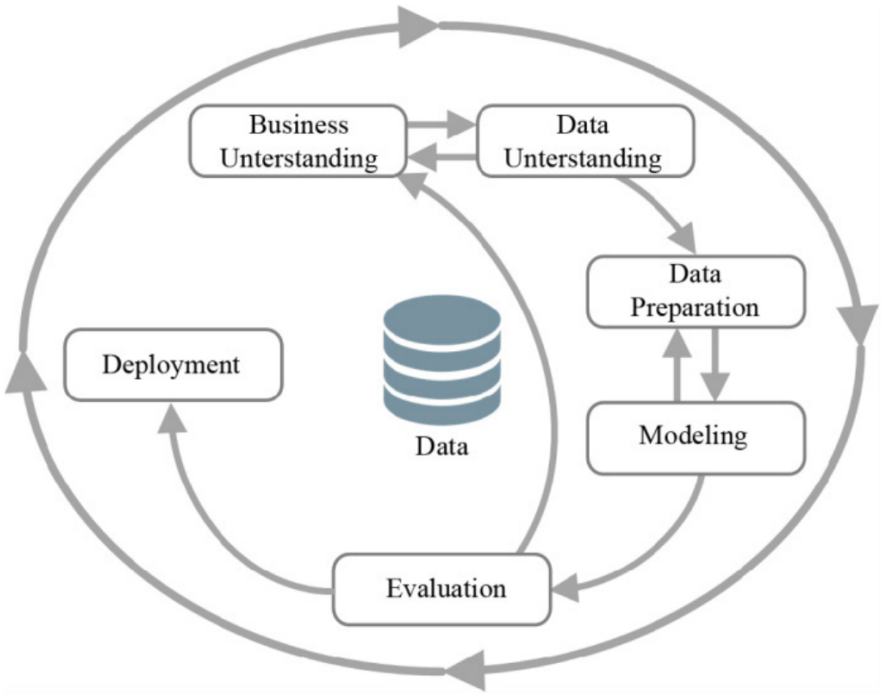


Figure 1: The Cross-Industry Standard Process for Data Mining [27].

The structure of this report is further based on the different phases of the CRISP-DM framework.

2.3.2 Incorporating Design Thinking Principles

In the initial stages, we utilized design thinking principles to gain a deeper understanding of problem-solving. Design thinking is a process of understanding the users, redefining problems, and creating innovative solutions in an iterative way to create prototypes or test scenarios [28]. This should lead to alternatives or strategies which deviate from existing solutions. For the incorporation, a 5-stage process is used, which includes the following steps: empathize, define, ideate, prototype, and test [28].
In this project, design thinking is crucial as it helps to understand the motives of customers to visit a coffee shop. As the goal of this report is to give store location recommendations, the behavior of the customer and the impact of it regarding the different identified criteria has to be understood. We employed design thinking alongside the CRISP-DM framework to enhance the ability to grasp individual preferences and viewpoints, which could be difficult to deduce from data patterns. Nevertheless, considering the constraints of the project, we found that engaging in discussions with commuters and stakeholders or gathering pertinent information to gain deeper insights into their decision-making processes was not a feasible option.
The design thinking process was the foundation for brainstorming activities throughout this project. The team held regular meetings to brainstorm the datasets and business objectives, with all members participating and deliberating for an hour. The discussion always concluded with a decisive agreement on a specific result.

2.4 Project Management Tools & Communication

For the creation of this report, we utilized various tools to increase the productivity and creativity of the individual team members. Additionally, various communication channels were used, and meetings were scheduled to improve feedback loops.

- **Mural**: Mural is a collaborative platform designed for teamwork with a focus on providing a secure and adaptable visual workspace. In the early stages of our project, we utilized Mural as a base for design thinking and brainstorming activities. We also used it to identify the dataset that we have chosen as the focal point for this report, as well as to identify various business goals derived from the dataset. Furthermore, we utilized Mural to delineate the report's scope, particularly in specifying what should be excluded from the content.
- **Trello**: Trello is an online task management service based on the Kanban method. This tool formed the foundation for implementing the agile project development approach, functioning as a task board. On this board, the advancement of different tasks and their corresponding responsibilities were systematically monitored.
- **Weekly Meetings**: Every Wednesday, we convened for an hour throughout the project's duration to review the progress of individual tasks, address any obstacles, and work towards their resolution virtually.
- **Hackathons**: These meetings occurred face-to-face and were scheduled as needed, typically lasting several hours to facilitate effective communication among group members and promote project advancement.
- **Webex**: This served as the communication platform for video-based online meetings, primarily utilized for weekly meetings.
- **WhatsApp**: WhatsApp served as the primary communication platform for prompt feedback opportunities and maintained a continuous flow of communication among group members.

3 Method and Analysis

This section provides an overview of the datasets employed in the analysis. We delve into the methods and tools utilized, along with a detailed exploration of data preprocessing and cleaning procedures. All the included steps can be found in our GitHub repository [29].

3.1 Description of Datasets

Yelp Dataset for Business Data
Based on the literature review, the proximity to competitors is one of the most important factors. In order to get data on businesses in Philadelphia we used the Yelp dataset [7] which was published by Yelp and received its latest update in 2022. It consists of five files for different categories, including for example businesses, reviews, and user data.
For this project, we only used the Yelp business data file. For every of the 150,346 included businesses it provides information like a unique business ID, the name of the business, its address, geographical coordinates (latitude and longitude), star rating, review count, and an indicator of whether the business is open. Additionally, there are columns for attributes, capturing details such as delivery options and outdoor seating, as well as categories that classify the business, such as restaurants, food, bubble tea, and coffee. We used all mentioned categories for our analysis except for the rating and the review count.
The Yelp business dataset demonstrates good reliability as businesses have a strong incentive to register, contributing to comprehensive data coverage. However, there is uncertainty regarding the up-to-dateness of the information. A notable strength is the absence of NaN values in the relevant columns, indicating completeness and enhancing the overall reliability of the dataset.

Datasets for demographics
According to the literature review conducted, three demographic criteria are crucial for finding a good location: age (a younger population is advantageous), population density (the denser, the better), and income (the more financial resources available, the better). We got this publicly available data from the U.S. Census Bureau. Specifically, three files were used:

1. **Census Tract Data for Pennsylvania**: The dataset [30] contains geometrical information for each census tract in the U.S. state of Pennsylvania. Census tracts are the smallest areal unit for which the U.S. Census Bureau has all the needed data available. Each census tract has its own geographical ID which allows assigning data from other datasets to a certain census tract.

2. **Population, Age and Sex Distribution for Philadelphia**: This dataset [31] contains general statistical information of each census tract within Philadelphia: Total number of population, age distribution, and sex distribution. By cross-referencing the total population with the geometrical data, the population density of each census tract can be calculated. Additionally, we isolated the percentage of the population between 10 and 34 years old, which is used as an indicator of how young the population in each census tract is.

3. **Household Income for Philadelphia**: In this dataset [32], information about household income for census tract is contained. Apart from detailed information for each type of household (how many people living in a house, rental or owned object, etc.), median household incomes are provided. We used only the median income for our analysis.

The data from the U.S. Census Bureau can be considered highly reliable, as it comes from an official source, and most demographic analyses in the USA are based on this data.

Datasets for foot traffic analysis
Based on the conducted literature review, the identification of an optimal location relies on three criteria inducing increased foot traffic: parks, public transport, and university locations. To access accurate and comprehensive data for these criteria, official datasets from the MyGeodata Cloud and OpenDataPhilly were integrated into the analysis.

The primary goal in using the following datasets for foot traffic analysis was to acquire the geographic locations of our intended amenities. These datasets are crucial for our analysis because these locations often serve as central hubs for community activities, gatherings, and commuting. Consequently, we expect higher human and foot traffic in these areas.

1. **Park Data:** The parks dataset [33], released in 2015, provides information on 167 parks like their names, geographical coordinates (latitude and longitude), and polygon boundaries.
2. **University Data:** The dataset on universities [34], released in 2017, provides information on 629 university locations. It includes columns such as the university names, geographical coordinates, and polygon boundaries.
3. **Bus Stops in Philadelphia Data:** The dataset [35], consisting of 23,178 bus stops in Philadelphia, provides diverse details for each location. Created on May 30, 2019, and updated multiple times a year, it includes geographic coordinates, road or path types, location names, bus service operators, shelter availability, associated routes, network or system information, etc. Some features had missing values (NaN) but we only used the geographic coordinates anyway which is why this did not pose a problem.
4. **Railway Stations in Philadelphia Data:** The dataset [36], updated on April 6, 2023, includes 156 railway stations and originates from the Southeastern Pennsylvania Transportation Authority. It includes railway station details, including their names and geographical coordinates.

Particularly for the park and university datasets, the utilization of polygon boundaries is more advantageous than a single coordinate because it enables a comprehensive representation of its spatial layout.

The datasets provided by OpenDataPhilly (offered by the City of Philadelphia) and MyGeodata Cloud (Geographic Information System (GIS) Software by GeoCzech, Inc.) were highly informative and proved to be valuable assets for our analysis. The reliability of these datasets further enhanced their significant contribution to our research.

3.2 Data Preprocessing

We opted to concentrate our analysis on the city of Philadelphia. Consequently, we needed to extract data specific to this chosen region from the broader dataset. We applied a Coordinate Reference System (CRS) to our geographical data to maintain uniformity across our spatial information. The dataset was narrowed down to Philadelphia, and standard preprocessing steps, including removing duplicates and handling anomalies, were applied across all datasets.

Yelp dataset specific preprocessing
We isolated coffee shops by filtering businesses with 'Coffee & Tea' in their categories, generating data tailored to this specific business category. Within this refined dataset, we further differentiated between Starbucks and non-Starbucks coffee shops, enhancing our ability to explore distinct patterns and characteristics within the coffee shop landscape. It is also crucial to emphasize that certain establishments classified as restaurants were included, as long as they offered coffee and tea.

Demographics datasets specific preprocessing
In addition to the already described reduction of the data to the city of Philadelphia, further considerations were necessary in dealing with demographic data:

First, not all census tracts contained data on population, age, and income. For example, a university campus might constitute its own census tract but has no permanent residents. Invalid values in calculations, such as population density, needed to be corrected, and missing data had to be marked to prevent it from skewing further analysis results.

Second, we found that some census tracts exhibited extreme outliers in variables such as income, leading to a non-optimal distribution as shown in figure 2.

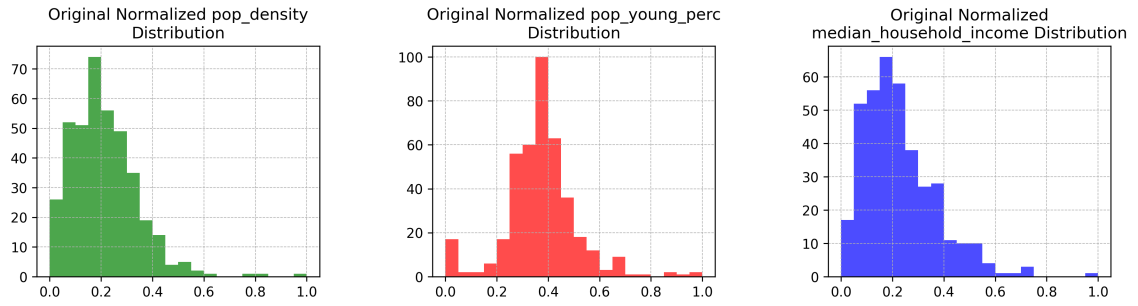


Figure 2: Distribution of demographic metrics with min-max-normalization.

Therefore, the data was normalized, and the lowest and highest percentiles were excluded to achieve a more uniform distribution as shown in figure 3. This allows better classification of the different geographical areas.

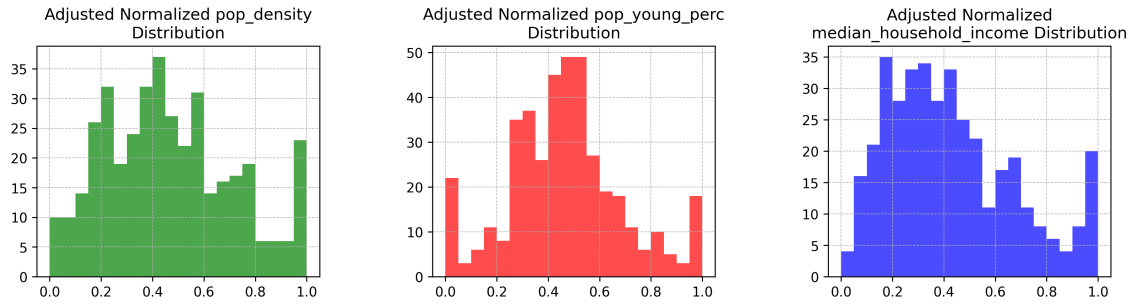


Figure 3: Distribution of demographic metrics with excluded lowest and highest percentiles.

Foot traffic analysis specific preprocessing
Notably, the data exhibited a relatively clean state, necessitating no further additional steps on top of standard data preprocessing steps.

3.3 Recommendation Engine

We designed the recommendation engine to assist in the identification of optimal locations for new Starbucks coffee shops. It evaluates potential locations in the Philadelphia area based on various criteria and scores them accordingly. The recommender takes coordinates (latitude and longitude) as input and outputs a score between 0.0 and 1.0 for each location, indicating how well it is suited as a new Starbucks location.
The recommendation pipeline, shown in figure 4, consists of three stages: Data Retrieval, Scoring, and Weighting which are explained in the following. We also discuss how the score can be evaluated and what role interpretability and customizability play.

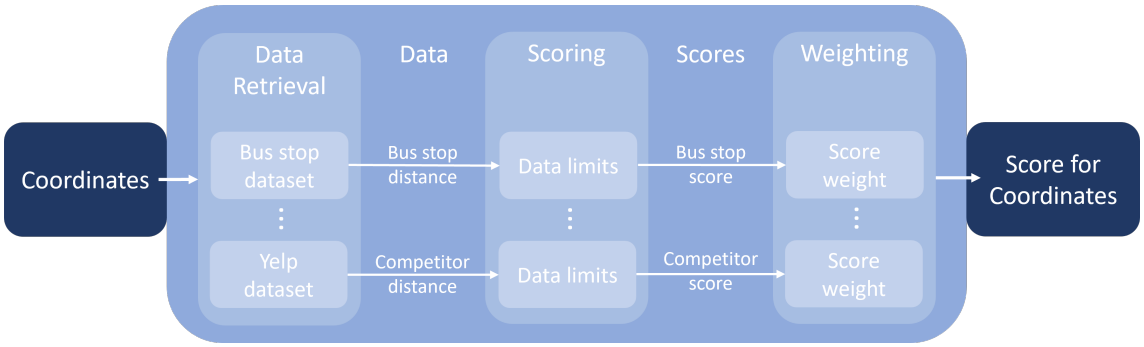


Figure 4: Recommendation engine pipeline from inputting coordinates to the scoring output, indicating the location's suitability for a coffee shop.

3.3.1 Step 1: Data Retrieval

The first step in the recommendation pipeline is retrieving the data for the given coordinates from the processed datasets. During this step, the recommender calculates nine different data values for the given location:

- **Competitors (4.1):** Distance to closest competing coffee shop
- **Demographics (4.2):** Population density, percentage of young people, and median household income
- **Geographical Characteristics (4.3):** Distance to a park, distance to a university, distance to a bus stop, distance to a train station

3.3.2 Step 2: Scoring

During the scoring step, the pipeline transforms the nine raw data values per location into nine scores between 0.0 and 1.0. A higher score is better and indicates favorable conditions for a coffee shop. There are two different mechanisms to get from the raw data values to the scores.

For the distances to a park, a university, a bus stop, or a train station, we define values for the minimum score and the maximum score. The values are listed in table 3. For these categories, a lower value is better as foot traffic is on average higher, closer to parks and other points of interest. We chose the values based on logical thinking which is a limitation, see section 5.3. In the future, these values should be improved through further analysis of customer habits. Distances between the highest and lowest values given in the table are uniformly mapped between the score 0.0 and 1.0. Distances outside the range are mapped to the closest limit, i.e. the lowest or highest value. For example, a distance to a park of 400m would be mapped to 200m and receive a score of 0.0. Similarly, we calculated the score for the distance to the closest competitor but here a higher value is better.

For the the population density, the percentage of young people, and the median household income a higher value is preferable. These indicators already got normalized and scaled into the 0.0 to 1.0 score range during the data preprocessing, see section 3.2.

	Distance from location to a...			
	Park	University	Bus stop	Train station
Lowest data value (equiv. to score 1.0)	10m	10m	10m	10m
Highest data value (equiv. to score 0.0)	200m	300m	80m	400m

Table 3: Score conversion limits for distances of the different distance categories.

3.3.3 Step 3: Weighting

After the nine different data values for the potential location have been retrieved and transformed into the nine intermediate scores, the scores are weighted and combined into a single output score. To combine the scores the pipeline uses a weighted average calculation. The total weights for all intermediate sums up to 1. The final score is the sum of all the intermediate scores multiplied by their respective weight. The weights for the nine intermediate scores are shown in table 4. They are based on the number of occurrences of the respective criterion in the literature, mentioned in section 2.2, and personal thoughts. Like the min-max values used in the scoring step, determining these values to maximize recommendation accuracy requires detailed industry knowledge and customer studies. In future work, a machine learning algorithm could be used to adjust the weights to fit existing Starbucks locations and maximize revenue, see section 5.4.
The result of the weighting is the final recommendation score for the location.

	Distance from location to a...				
	Competitor	Park	University	Bus stop	Train station
Weight	0.2	0.05	0.1	0.15	0.1
Paper occurrence	8	2	4	5	5
	Percentage of young people	Median household income	Population density	**Sum**	
Weight	0.1	0.15	0.15	1.0	
Paper occurrence	3	5	6		

Table 4: Score weights for the different distance categories.

3.3.4 Interpretability and Customizability

Our recommendation engine does not use machine learning methods, like deep learning, and instead relies on traditional scoring and manual weighting of different criteria. This results in complete interpretability and customizability of the system. Both are crucial as they help build trust, ensure recommendations fit the business's needs, and allow the adaptation to changes. When people using the system understand why a suggestion is made, they're more likely to trust and follow it. The customizability lets us tweak the system to match our specific needs, making it more useful over time. Nevertheless, the customizability is also a limitation because detailed industry knowledge is necessary to optimize the scoring and weighting, see section 5.3.

3.4 Methods and Tools

3.4.1 Data Methods

The Geographic Information Systems (GIS) technology was utilized across the datasets to visually represent the distribution of various characteristics as heatmaps. This approach facilitated a spatial analysis of demographic data, enabling the visualization of data concentration and patterns across various regions of Philadelphia. It proves especially impactful in pinpointing geographic trends and

hotspots about population density, age, and income distributions. Moreover, it is instrumental in identifying hotspots related to foot traffic density and conducting competitor analysis.

- **Yelp dataset specific methods**
  We analyzed the geographic concentrations of competitors in Philadelphia by calculating the Euclidean distance between Starbucks and competitors in the Yelp dataset. This helped to identify the nearest direct competitors.
- **Demographics specific methods**
  To examine demographic characteristics in Philadelphia, we used descriptive statistics. This method allowed us to determine metrics like average population density and distribution across census tracts, providing an overview of the demographic landscape.
- **Foot traffic analysis specific methods**
  To understand foot traffic in Philadelphia, we employed Kernel Density Estimation (KDE). KDE helped evaluate metrics like the average density and spatial distribution of bus stops, schools, and parks. This enabled us to visualize the underlying distribution of our data in a continuous and visually informative manner.

3.4.2 Tools

We used many different tools for the analysis and the recommendation engine. Some noteworthy include the following:

- **Power BI**: We used Power BI for the initial exploration of the datasets. Power BI is very intuitive and allows for quick analysis without coding.
- **Kaggle**: Kaggle is a collaborative data science platform for Jupyter Notebooks and datasets that we used for data analysis and to write the recommendation engine. Kaggle provides more computing resources than Google Colab and also allows easier access to datasets.
- **Python**: To code the recommendation engine and perform more detailed data analysis we used Python. For Python, many data science libraries are available that allow for the easy handling of data and visualizing it.
- **Geoplot Python Library**: To create static plots of the geographical data we used the *Geoplot* library. It allows easy visualization of polygon, point, and density data. Geoplot has more fine-grain control over the plots than Power BI.
- **Folium Python Library**: For the interactive data visualization on a map we used the *Folium* library and hosted the resulting HTML file on GitHub. Folium is widely established and has numerous plugins to customize the maps. It also integrates well with the other Python libraries we used for the data analysis.

4 Evaluation and Interpretation

In this section we delve into the comprehensive evaluation of various factors influencing the selection of Starbucks locations in Philadelphia. We initiate by examining the distribution of coffee shops, delving into competitor-relations and demographic considerations, analyzing foot traffic, and culminate in an assessment of the recommendation engine's outcome.

4.1 Competitors

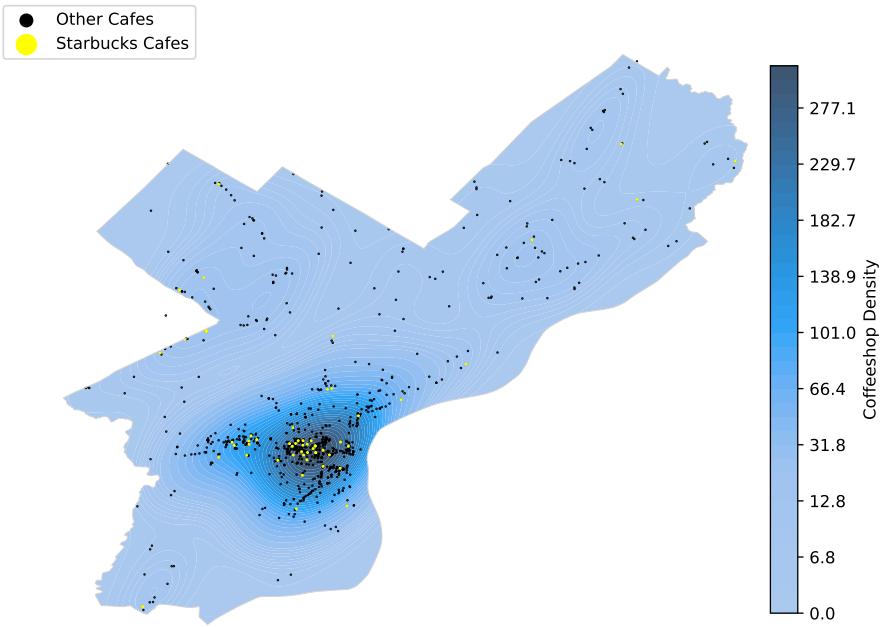


Figure 5: A coffee shop density plot with coffee shops as black points and Starbucks as yellow points. The plot is based on data from the Yelp dataset and shows how the coffee shops are distributed among the city of Philadelphia. Most coffee shops are located within one area.

Figure 5 shows the coffee shop distribution in Philadelphia. The plot highlights a significant clustering of coffee shops in the city centre of Philadelphia. This concentration suggests that the optimal location for a coffee shop may vary depending on specific business goals. For instance, Starbucks might strategically choose to establish a presence in an area without direct competitors to monopolize the market demand and cater to an under-served customer base. On the contrary, situating a Starbucks in an area with existing competitors could be advantageous, fostering healthy competition and attracting a demographic already inclined towards coffee consumption.

4.2 Demographics

When focusing solely on demographic characteristics, the goal is to identify regions in Philadelphia that exhibit high population density, a large proportion of young people (ages 10 to 34), and high socioeconomic potential (indicated by high income). Each of these three factors has been evaluated independently and can be visualized through a heatmap. However, it is important to note that while demographics are crucial, they are just one aspect of the overall assessment. Other evaluation criteria must also be considered to provide a well-rounded recommendation.

4.2.1 Population Density

Population density is an important factor for a successful location. In areas with high population density, stores and facilities tend to be visited more often than in areas with low population density. Figure 6 shows the population density within Philadelphia, the darker the blue the denser the population. There are several census tracts with increased population density noticeable, with a concentration visible in the lower right part of Philadelphia. Another indication of this is that the census tract areas there are smaller: The U.S. Census Bureau defines more areas in regions with high population density than in areas with low density.

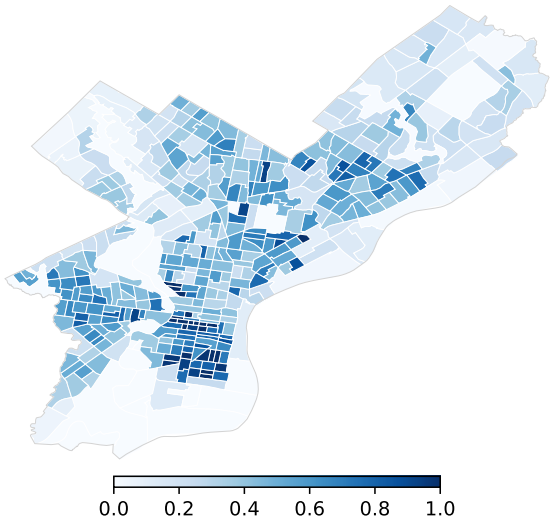


Figure 6: Normalized population density within each census tract in Philadelphia.

4.2.2 Age Distribution

Another important criterion for Starbucks coffee shop locations is a low average age in the surrounding area. Starbucks primarily targets teenagers and young adults in its branding. Figure 7 shows the areas with the highest percentage of 10 to 34-year-olds in Philadelphia. Even though 10-year-olds would not be considered the primary target group of Starbucks, we included them to anticipate them growing up in the next few years. Again, advantageous areas are primarily found in the lower part of the city, but unlike population density, they are more located towards the bottom left and around the city center in tendency.

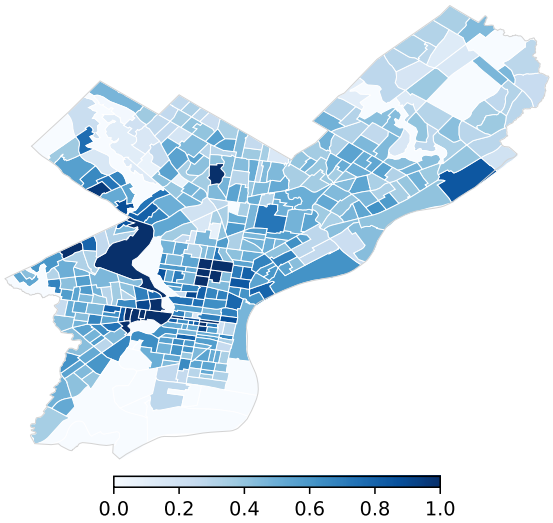


Figure 7: Normalized distribution of 10 to 34-year-olds within each census tract in Philadelphia.

4.2.3 Socioeconomic Status

The final demographic criterion considered was socioeconomic status: A higher median income in an area is advantageous for Starbucks, as more money can and tends to be spent there. The situation for Philadelphia is as depicted in figure 8: Areas with high median incomes are found mainly on the city edges. This includes the lower right edge of the city, which likely has some overlap with the other two criteria.

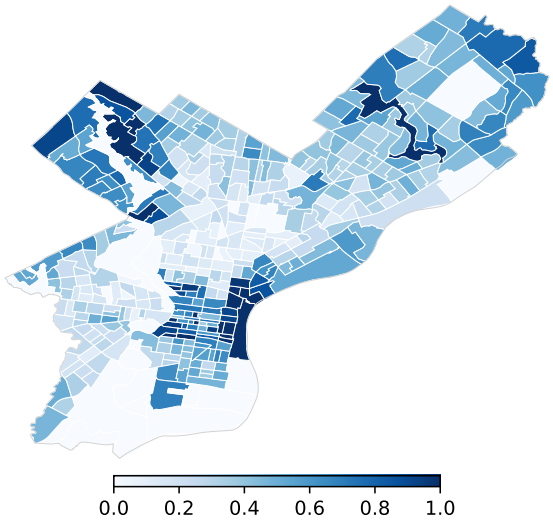


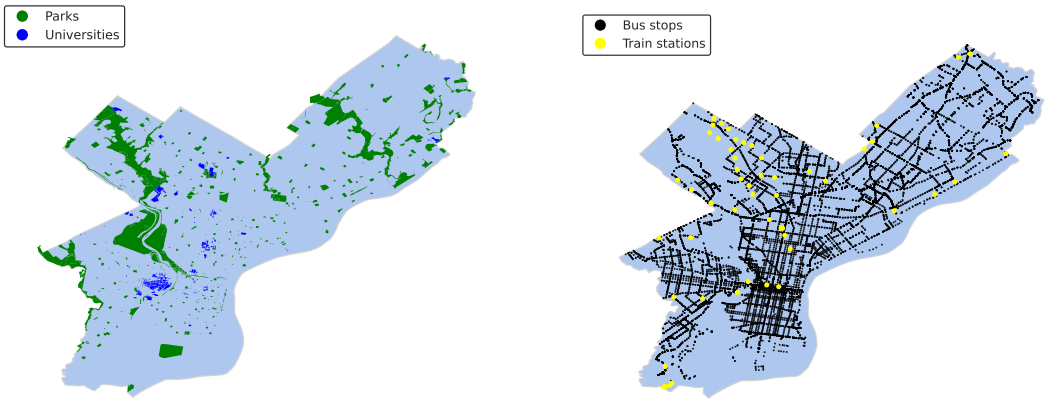
Figure 8: Normalized median household income for each census tract within Philadelphia.

4.2.4 Conclusion

Overall, a mixed picture emerges when evaluating the three demographic criteria in Philadelphia. Tendency-wise, the lower right part of the city seems to be an area that is demographically advantageous for Starbucks. However, two things must be noted:

1. The interpretation here is based on a visual evaluation. A final judgment must be made using all criteria (not just demographic).
2. The evaluation of the demographic criteria only considers the status quo and cannot provide information in this form about how the situation might develop in the future.

4.3 Foot Traffic Analysis



(a) Parks and universities in Philadelphia.

(b) Bus stops and train stations in Philadelphia.

Figure 9: Visualization of used datasets for the foot traffic analysis.

The proximity of nearby amenities can influence the success of a coffee shop, and similarly, the accessibility of bus stops and train stations plays a crucial role in determining the potential success of coffee establishments. Coffee shops near bus stops and train stations benefit from the increased foot traffic. Our analysis incorporated comprehensive datasets comprising 23,178 bus stops, 156 railway stations, 167 parks, and 490 university grounds.

Figure 9a shows scattered clusters of universities across Philadelphia, accompanied by the presence of a few big and any small parks. This spatial distribution suggests the potential for increased foot traffic near universities and parks, making these areas promising locations for businesses or initiatives that benefit from high pedestrian activity. Understanding the concentration of educational institutions and significant parks becomes crucial for strategic planning, as it offers valuable insights into areas where businesses may thrive due to heightened foot traffic generated by school-related activities and the recreational use of parks.

Figure 9b indicates that the southern region of Philadelphia is a good candidate for solid foot traffic, evident from the elevated concentrations of bus stops and railway stations in that vicinity. However, it remains crucial to emphasize that while foot traffic is a significant factor, it should not be the sole determinant for selecting an optimal store location. Previous considerations and discussed factors should be thoroughly considered to make a well-informed decision about the most suitable business location.

4.4 Recommendation Engine

4.4.1 Results and Interpretation

The recommendation engine outputs the weighted overall scores and the individual scores. We suggest two ways to use the recommendation engine:

1. Analyzing pre-selected locations, like available properties, by scoring them.
2. Find the best location within a defined area by analyzing a grid of equidistant points.

When we use the second method, for example for the city center of Philadelphia, we first define an area. For a center point like Philadelphia City Hall at (39.9523, -75.1635) and a radius of 1km, the recommendation engine finds the most promising locations for a new Starbucks coffee shop in the surrounding area. In this example, the recommendation engine evaluates a total of 6400 individual points in the 2km by 2km area when using a resolution of 25m.
By filtering for the point with the highest overall score we get the best location in the area, according to the analysis. In this case this is 1530 Locust Street at the coordinates (39.9485 -75.1677) with a score of 0.69 or 69%.

To further analyze the results for this location, the recommender provides two main visualization tools: a plot showing the contribution of the different categories to the score and an interactive map to understand the surrounding area of the suggested location.
The bar plot, as shown in figure 10, shows the overall and individual scores. It helps to understand the strengths and weaknesses of the suggested location. For this location for example the overall score is 69%. It is visible that the scores are high or perfect in the categories of distance to a university, distance to a bus stop and demographics. Weak categories with scores of lower than 0.5 are the distance to a competing coffee shop, a park, and a train station.

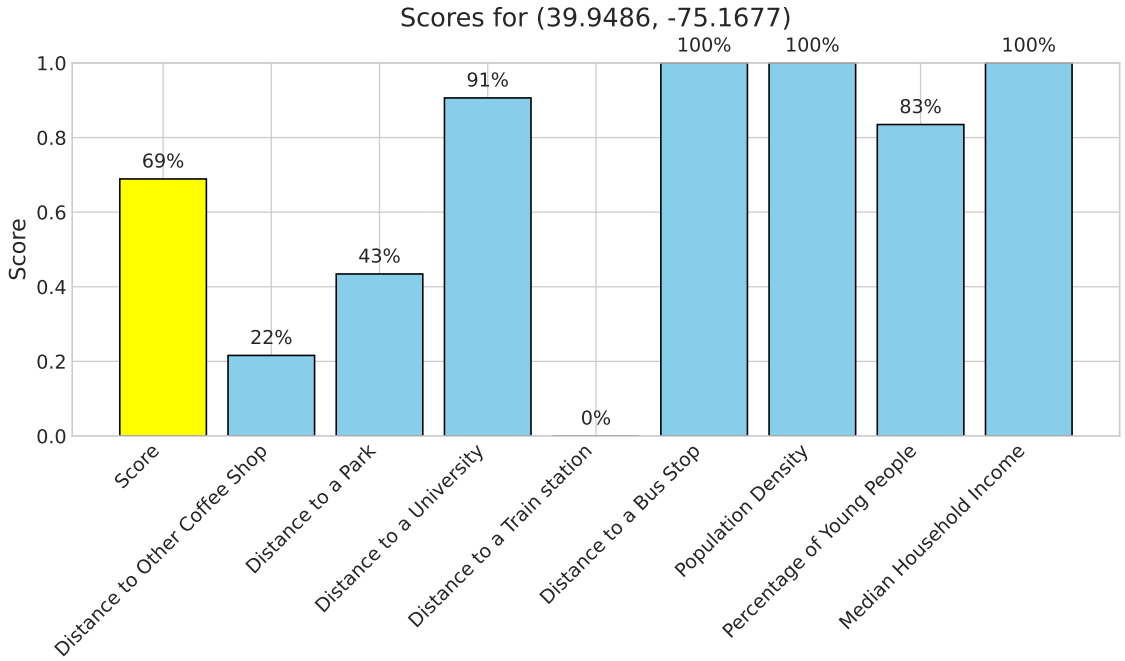
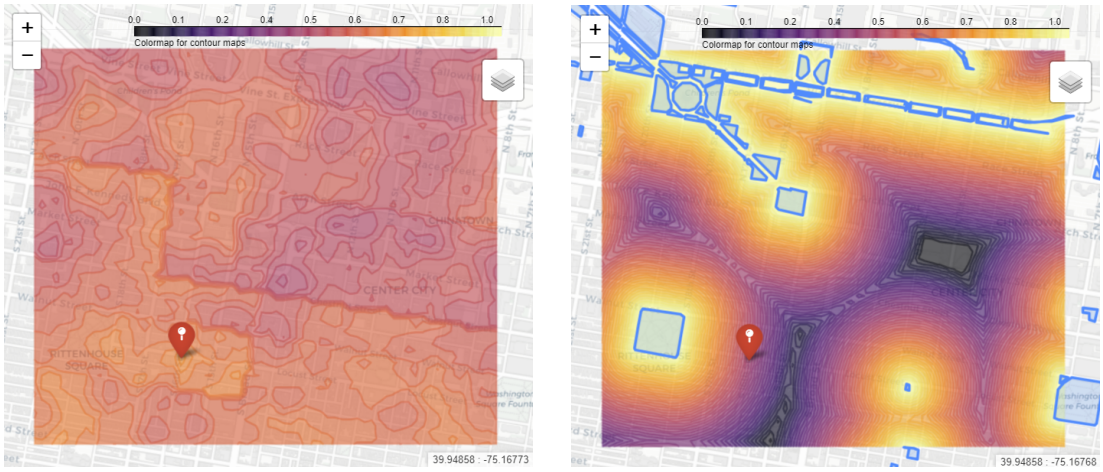


Figure 10: Bar plot showing the overall score in yellow and the individual scores in blue for the example given in section 4.4.1.

The interactive map helps to get a better understanding of the surroundings. By default, the map shows a contour map of the score over the chosen area, as shown in figure 11a and also available

online. This helps to get an understanding of how the score behaves over the area. By turning on different layers in the user interface of the map it can also visualize the individual scores that influence the overall score. Figure 11b shows an example where the parks are displayed with the contour map of the park score.



(a) Overall score contour map with marked top location.

(b) Distance to park contour map with marked top location.

Figure 11: Screenshots of the interactive indicator and score visualization map from the recommender.

4.4.2 Suitability of Results

The results of the recommender, as scores for the different locations, are valuable for Starbucks' business goals, offering a data-driven approach to location selection. These results are derived from the most common store location criteria mentioned in the existing literature. With the exploitation of those criteria, the recommender provides the best store location for a specific area which should directly lead to an improvement of the business objectives related to the store location.
The interpretability and customizability of the system empower stakeholders to trust and understand the recommendations, fostering better decision-making. The visualization tools facilitate a clear understanding of the factors influencing each location's score, contributing to a strategic and informed expansion strategy for Starbucks in the Philadelphia area. The recommender aims to assist in identifying optimal locations by considering various factors such as closeness to competitors as well as demographic and geographical characteristics.

4.4.3 Shortcomings

While the recommendation engine offers valuable insights, it has several shortcomings:

- The distance ranges used for scoring are set manually. This requires detailed knowledge about the industry, the customer habits and also depends on the specific surroundings. This is a big problem for small independent coffee shops. Large chains like Starbucks can afford to invest in further studies to improve and essentially eliminate this limitation.
- Similar to the distance ranges the weighting of the different categories is also selected manually and comes with similar problems.
- The recommender is at best only as good as the quality and amount of data available. If information on important location factors like parking availability is not available, the recommender can not consider these.

Further limitations, which also focus on future steps of the project, are listed in section 5.3.

4.4.4 Reliability

To assess the reliability of the recommendation engine, we utilized several measures to validate its performance. These measures include scoring existing Starbucks locations, employing multiple criteria, and establishing a plan for continuous improvement.

Scoring Existing Starbucks Locations:
To gauge the reliability of the recommendation engine, we calculated scores for existing Starbucks locations in the Philadelphia area. The results demonstrated a consistent performance, with a minimum score of 0.31, an average score of 0.52, and a maximum score of 0.65. These scores indicate a generally reliable system, as the recommendations align with the presence of successful Starbucks locations. For further analysis, these recommended scores should be correlated with the business success of the individual locations. Because the business data is not available this is unfortunately not possible for us.

Utilization of Multiple Criteria
The reliability of the recommendation engine is further ensured by incorporating a diverse set of criteria (competitors, demographics, and foot traffic). By considering various factors, the system becomes more robust and adaptable. In cases where certain criteria may not be ideal for a specific location, alternative criteria can compensate for any shortcomings, enhancing the overall reliability of the recommendations.

Continuous Improvement and Adaptation:
To maintain reliability in the long term, the recommendation engine is designed for continuous improvement. Regular updates to the datasets, as well as adjustments to scoring weights based on changing trends or business strategies, contribute to the system's ability to adapt to evolving conditions. This ensures that the recommendations stay relevant and effective over time.

5 Deployment and Recommendations

This section details the steps and considerations necessary for implementing the recommendation engine's findings into Starbucks' decision-making process. Firstly, we outline the implementation plan's key stages and timeframes leading up to selecting a new location. Following this plan are recommended actions for various stakeholders within the company based on the engine's results. Moreover, we explore the limitations of our approach, highlighting its constraints and suggesting areas for future work and improvement.

5.1 Implementation Plan

With the help of the recommendation engine, the search for a suitable location can commence. The following implementation plan, also shown in figure 12, explains the necessary steps up to the decision for a new location, including approximate time frames:

1. **Discussion of the results with the responsible people (2 weeks)**: The results of the recommendation engine should be discussed with the management or project leader. At this point, additional criteria should be defined that need to be considered in the selection (criteria not covered by the analysis).
2. **Selection of several areas with high potential (8 weeks)**: In this step, the most promising areas should be selected. Based on the recommendation engine, various experts should contribute to the discussion: marketing, people with knowledge of the local conditions and history, urban planning, lawyers, supply chain managers, etc.
3. **Inspection of the selected areas (4 weeks)**: Once an agreement has been reached on some areas, specialists must assess the situation on-site. It should be determined whether there are locations available for purchase or rent, and other factors that influence the decision should be recorded.
4. **Evaluation of the inspected areas (4 weeks)**: In further discussion, the results from the last step should be incorporated into the previous considerations. The goal is a list of specific potential locations.
5. **Decision making for a new location (1 week)**: In the final step, the decision for a new location can be made based on all known facts and previous discussions.

Figure 12: Implementation plan of five phases to decide for a new Starbucks location.

5.2 Recommended Actions

The implementation and use of the recommendation engine to determine a new location requires discussion with a variety of stakeholders. The goal is to evaluate the recommendations made by the recommendation engine, taking into account further criteria that cannot yet or cannot at all, be assessed with data. We recommend the following actions for the various stakeholders to bring the most valuable contribution to the discussion:

- **Management / Project Leader**: Should review the results of the study and examine them in the context of Starbucks' overall business strategy in Philadelphia.

- **Marketing / Branding**: Should take note of the results of the study and possibly contribute further important factors for the subsequent decision-making process.
- **Legal Department**: Should be aware of the study results and point out known legal restrictions or concerns as early as possible. Arrangements with residents, etc., may also need to be made.
- **Construction Planning**: Should assess the situation in potential areas, considering criteria such as availability of locations, general urban development, suitability for setting up a Starbucks branch, etc., in collaboration with local experts.
- **Supply Chain Management**: The implications on the supply chain for different possible locations should be clarified, since an unobstructed supply chain is crucial for a new locations success.
- **Information Technology Department**: Incorporate the recommendation engine into existing processes and try to align current research investigations with the new technology. Additionally, think about adjacencies for further investigation.

5.3 Limitations

Our analysis, while being comprehensive, is subject to several identified limitations. These limitations compass various aspects covering our overall approach, ranging from constraints within the datasets used, to broader methodological or contextual constraints. This section outlines these limitations which are crucial for contextualizing the findings and acknowledging the boundaries of our analysis in guiding strategic decisions for Starbucks' location selection.

- **Challenges in Recommendation Weights Conversion**: Converting raw data into scores to determine location suitability requires a deep understanding of the industry and customers. Further research on the importance of location factors and insider knowledge is necessary for refining the recommendation weights and values.
- **Uncovering Non-linear Relations**: Identifying non-linear relationships that influence business location suitability without leveraging machine learning techniques was proven challenging. With limited data on Starbucks locations and the absence of revenue information, our exploration of intricate patterns affecting location suitability was hindered.
- **Unaccounted Influential Factors**: Several influential factors, such as parking availability, accessibility, availability of locations and cultural nuances remain unexplored within our analysis.
- **Dataset Timeliness**: Although the datasets reflect the current state of the city, they are not entirely up-to-date. Changes in areas like new parks, new commuter usage items, income levels or district popularity over time, impacting the accuracy of long-term predictions, are not accounted for.
- **Predicting Future Development**: The datasets used cannot predict future city developments. Areas classified as high or low income might undergo changes, and districts may transform within a few years, making long-term predictions challenging.
- **Assumptions in Foot Traffic Analysis**: In our foot traffic analysis, assumptions were made regarding an even distribution of users in parks, a consistent university population, and comparable commuter usage at bus stops. These assumptions might oversimplify actual user behavior patterns, potentially impacting the accuracy of our assessments.
- **Unknown Starbucks Plans**: Lack of business information regarding Starbucks' future plans for Philadelphia limits the direct applicability of our findings to their strategies in the city. However, by not using area-specific strategies, the methodology employed could serve as a valuable approach for other cities or expansion plans.

5.4 Future Work

The current iteration of our recommendation engine presents a promising approach for assisting Starbucks identifying optimal locations for new coffee shops. However, to further elevate the efficacy of this system, we have identified several courses for future exploration and enhancements, which are presented below.

- **Conducting Customer Studies**: By engaging directly with Starbucks' customers through surveys and interviews we could gain a more comprehensive insight into customer preferences according to various location factors. This could further improve our knowledge of these location factors and cause more informed adjustment of weights assigned to location parameters, and ultimately improve our recommendation model.
- **Performance Evaluation of the Recommendation Engine using Financial Data from Existing Starbucks Coffee Shops**: An essential step in improving our recommendation engine involves its real-world performance outcomes. By utilizing financial data from existing Starbucks coffee shops, we could conduct more comprehensive analysis by comparing any recommended location against financial status of Starbucks shops.
- **Improve Existing Business Factors**: Improving the consideration of competitive dynamics within our model is critical for a more holistic recommendation system. This involves incorporating multiple competitors simultaneously and leveraging their popularity metrics derived from reviews or similar sources. Weighting these factors effectively within our recommendation engine will enable a deeper understanding of the competitive landscape, contributing to more informed location recommendations.
- **Additional Data and Location Factors**: Introducing supplementary data points and location factors is potentially key to refining the recommendation engine's performance. Factors such as available parking spots, accessibility, distance to major highways or other nearby amenities will be integrated. Analyzing the impact of these additional variables on store success will enhance the model's predictive capabilities and provide more comprehensive location recommendations.
- **Scaling Geographical Recommendations**: Expanding the scope of our recommendation system beyond the current study area is crucial for its applicability in diverse regions. Scaling the model to encompass other cities, metropolitan regions, and suburbs necessitates adaptation to varying demographics, market dynamics, and regional preferences. This expansion will facilitate strategic decision-making for Starbucks' store placement in different geographical landscapes.
- **Utilizing User Reviews**: An essential aspect of our future work involves extracting detailed insights from customer reviews. By systematically analyzing sentiments, preferences, and specific factors influencing customer satisfaction, we aim to gain valuable information about customer experiences at various locations. This in-depth analysis will focus on understanding the nuances of customer feedback. The goal is to extract actionable insights from reviews to enhance our recommendation engine's understanding of customer preferences and satisfaction levels. Integrating these insights into our model will contribute to more informed decision-making regarding new Starbucks store locations

6 Monitoring and Maintenance

In this section, we outline the different key performance indicators for the validation of the recommendation and the possible risks of the deployment as well as the need for human intervention and the lessons learned.

6.1 Key Performance Indicators

There are various key performance indicators, which can be used to validate whether the recommendation of a location for a coffee shop is good. The highlighted indicators can only be accessed from coffee shops after they are established based on the recommended location. So they can then implicitly be used to improve the selection of new shop locations by adjusting the recommendation engine and its weights. Those key performance indicators include:

- **Sales and Revenues**: The sales and revenues of the shop can be monitored based on a specific time horizon. Those values can then be compared to good performing coffee shops of the franchise itself or to anticipated values in the industry.
- **Customer Count**: The amount of customers can be counted and then be compared to good performing stores in comparable locations based on the defined criteria. The customer count in combination with the sales can then also lead to identifying the average transaction value.
- **Customer Retention Rate**: This rate shows the percentage of customers who return for repeating visits. This is a good indicator for a good store location as it shows the satisfaction of the customers in a broader sense.
- **Competitor Analysis**: The performance of nearby coffee shops of competitors should be always monitored. Therefore, their customer traffic as well as their revenues should be analyzed to identify the relative position of the coffee shop.
- **Customer Reviews and Feedback**: The recommendation engine has the potential for improvement in subsequent stages by integrating customer reviews. These reviews can be systematically mined and subjected to sentiment analysis to determine customer satisfaction with the coffee shop's location.
- **Market Penetration**: The market share of the coffee shop in the area can be analyzed and from this the relative performance to coffee shops nearby can be derived.
- **Return on Investment**: The return on investment should be always monitored as it shows the relation of the revenues in comparison to the investments into the coffee shop. This is an important measure as it can be a threshold set by the company upfront for the accepted performance of a shop.
- **Customer Churn Rate**: The rate at which customers stop visiting the shop should also be monitored. This can be done with member cards and the identification of visits of various customers. This is an important measure as it shows which customers are not satisfied with the coffee shop and it can be further used to identify if this correlates with location factors.
- **Supply Chain Efficiency**: It should be monitored whether the supply chain can satisfy the local demand of the different products. Therefore, it should be monitored how often various products are out of stock and if those bottlenecks are caused by positional effects.
- **Seasonal Variability**: It should be analyzed how seasonality effects the revenues of a new coffee shop, especially in comparison to other Starbucks branches in Philadelphia. This analysis should also include local events and holidays. This is necessary, as it shows the variability of revenues for the shop location.
- **Operational Costs**: Those costs include all costs which are associated with the operation of the coffee shop. Those should be monitored to show possible changes over time and the overall profitability of the coffee shop.

6.2 Risk Management

Every project includes risks and to address them early and implement mitigation strategies is crucial to ensure project success. Therefore, the things which can go wrong during and after deployment are outlined and possibilities on how they can be handled are shown in table 5.

Risk	Definition	Mitigation Strategy
Data Bias	If the training data used to build the recommendation engine is biased, the system may make recommendations that favor certain locations or demographics, potentially leading to unequal distribution of resources or missed opportunities in other areas.	Incorporate domain experts into the assessment of recommendations. This should be carried out iteratively post-deployment of the engine to safeguard against data bias by continuously updating the data source.
Privacy Concerns	The inclusion of customer reviews in the analysis by the recommendation engine or metrics may give rise to privacy concerns. If customers perceive that their personal information is not sufficiently protected, it has the potential to result in backlash and damage to the reputation.	Personal information utilized for the model will undergo anonymization. Moreover, the collection process will guarantee the acquisition of only the essential data required for recommendations. Additionally, transparency regarding data usage will be provided to customers, and adherence to pertinent data regulations will be maintained.
Security Risks	If the recommendation engine is vulnerable to cyber threats, it could be exploited, leading to unauthorized access, data breaches, or manipulation of the recommendation algorithm.	To address this issue, authorization concepts will be integrated, and users of the system will undergo trainings to guarantee the security of the engine.
Dependency on Historical Data	The recommendation engine relies on historical data to make predictions. If this data becomes outdated or no longer reflects current customer preferences, the system may provide less relevant recommendations.	The data used for store location recommendations will undergo iterative updates, ensuring the incorporation of changes for various criteria.
Legal and Regulatory Compliance	Failure to comply with changes in data protection regulations or other relevant laws may result in legal consequences for the engine. Therefore, it is crucial to ensure that the recommendation engine adheres to privacy and compliance standards.	Regularly conduction of a proactive assessments to stay abreast of evolving regulations. Implementation of a dynamic compliance framework that enables swift adjustments to ensure the recommendation engine aligns with the latest legal requirements.
Changing Business Goals	Changes in the business goals may impact the desired functionality of the recommendation engine. In this case, the weights or the criteria, influencing the final decision have to be adjusted.	Close alignment of the project team to the management team to identify changes to overall strategies and business goals. Furthermore, the modularity of the engine has to be ensured to faster adopt to changes.

Table 5: Risks which can occur during and after deployment and the appropriate mitigation strategies to show how those risks could be handled.

6.3 Human Intervention

The proposed recommendation engine is mainly a decision support system and human intervention is recommended at various points:

- **Decision Support**: The recommendation engine can propose the best location according to predefined criteria. However, for a conclusive decision, it is essential to involve human intervention, leveraging the expertise of domain experts or locals to enhance the location selection process.
- **Exceptions**: Human intervention may be necessary when there are unexpected events or situations that the recommendation engine may not have been adapted to. This could include sudden changes in local demographics, economic shifts, or unforeseen events that impact foot traffic and customer behavior.
- **Concept Drift**: As the recommendation engine relies on datasets that are continuously retrieved and updated with new entries, the attributes of the variables may undergo changes over time. Consequently, it becomes necessary to periodically assess whether the criteria used by the recommendation engine remain appropriate and if any adjustments to their weighting are required.
- **Regulatory Changes**: Changes in regulations or laws may impact the viability of a location. Human experts can assess the legal landscape and guide decisions based on these changes.
- **Changing Business Goals**: If the business objectives or strategies of Starbucks undergo a change, human experts can adjust the recommendation engine to align with the updated goals. This ensures that the system remains in sync with the evolving vision and continues to provide support accordingly.
- **Location Exclusion List**: A list of locations with restrictions not deducible from the data will be excluded from recommendations. In such cases, the option for a manual review by a domain expert will be accessible.
- **Infrastructure Changes**: Changes in local infrastructure, such as the opening of a new transportation hub or the closure of a major road, can affect foot traffic and may require human assessment for optimal decision-making.

6.4 Lessons Learned

Throughout the execution of the project and the writing of the report, we learned several lessons that can be applied to future projects or when implementing further functionalities within this project.

- **Business Context**: One of the valuable lessons we learned from the execution of the project is the necessity of embedding the project in a valuable business objective. In this project it was particularly helpful to define a specific scenario to allow coming up with business objectives.
- **Project Management**: We improved our project management expertise, with a notable impact on task distribution and time estimation skills. Furthermore, we gained valuable insights into implementing an effective data strategy for project execution this was done by incorporating Mural for the design thinking activity. These enhancements reflect a broader understanding of project dynamics. Especially, out capabilities in the task definition and time scheduling advanced significantly by incorporating Trello into the execution of our project.
- **Clear Problem Definition**: The lesson underscores the critical importance of clearly articulating the problem statement before embarking on the analysis. Without a precisely defined problem, the process becomes arduous, hindering the development of meaningful insights or solutions. A well-crafted problem statement serves as the foundation for effective analysis and problem-solving.

- **Data Quality**: The lesson highlights the crucial role of high-quality, relevant data in ensuring accurate and reliable results. Inaccuracies or gaps in the data, as we had in the demographics data, can compromise the integrity of analyses and undermine the reliability of model predictions. Utilizing precise and comprehensive data sets is essential for obtaining trustworthy insights and making informed decisions.
- **Effective Communication**: This lesson underscores the essential need for transparent communication between us and the supervisors (in this case the teaching assistant and the professor). Clear articulation of technical findings to the supervisors is imperative to ensure the success of the project. Additionally, a fast communication channel is important for the overall project's success. This could be seen through the rapid project progress after obstacles could be removed within the weekly meetings and the hackathons.
- **Project Documentation**: It was surprising to see how much work went into the project documentation compared to the work going into the actual analysis itself. In the end, working on the project documentation helped us to present our findings in a way understandable for everyone not familiar with our project.

7 Conclusion

In summary, the primary aim of this project is to leverage data to suggest optimal locations for new Starbucks stores in Philadelphia. This objective is achieved through the implementation of an interpretable, deterministic recommendation engine, which utilizes specific location data to propose suitable store locations. The exemplary result, based on key criteria derived from relevant literature, indicated that the recommendation engine effectively addresses the analyzed variables, ultimately proposing an optimal store location in alignment with the project's goal. However, it's essential to note that the final evaluation of the objective hinges on the actual establishment of the store at the proposed location, as many key performance indicators rely on the relative performance of existing coffee shops in that vicinity.
This is further supported by the contributions of this work. Through which, an examination of various approaches outlined in the relevant literature, three key criteria, foot traffic, demography, and competitors, were identified. These criteria form the basis for recommending a suitable location for the store. Then a recommendation engine pipeline has been developed using these criteria, generating scores that indicate the suitability of a given location for a coffee shop based on input coordinates. The output from this recommendation engine is subsequently utilized for an interactive online map and serves as the foundation for actionable recommendations guiding the site selection process for new Starbucks coffee shops.

Furthermore, this project has provided us with invaluable lessons. As the report pointed out, it is critically important to integrate projects into precise business objectives and to define problems clearly in order to drive meaningful analysis. In addition, it highlighted the importance of high-quality data, effective stakeholder communication, and the integration of domain expertise with data science initiatives.

Ultimately, we believe this project underscores the potential of data-driven approaches in enhancing strategic decision-making processes. The developed recommendation engine not only addresses the needs of Starbucks but also lays a solid foundation for further improvements and applications in the field of location-based business decisions.

Bibliography

[1] Vadym Sokol and Kristijan Jordanov. *Site selection for small retail stores using sustainable and location-driven indicators: Case study: Starbucks coffee shops in Los Angeles*. 2020.

[2] Jeremy YL Yap, Chiung Ching Ho and Choo-Yee Ting. 'Analytic Hierarchy Process (AHP) for business site selection'. In: *AIP Conference Proceedings*. Vol. 2016. 1. AIP Publishing. 2018.

[3] Kayla Gordon. 'Business site selection, location analysis, and gis'. In: (2017).

[4] Omar Ibrahim Aboulola. 'GIS spatial analysis: A new approach to site selection and decision making for small retail facilities'. PhD thesis. The Claremont Graduate University, 2018.

[5] M Khalifatul Ardhi, Jangkung Handoyo Mulyo et al. 'How does entrepreneurial orientation affect the business performance of coffee shop MSMEs in Indonesia?' In: *E3S Web of Conferences*. Vol. 306. EDP Sciences. 2021, p. 03011.

[6] Liyan Xu et al. 'A Two-Layer Location Choice Model Reveals What's New in the "New Retail"'. In: *Annals of the American Association of Geographers* 113.3 (2023), pp. 635–657.

[7] Inc. Yelp. *Yelp Dataset — kaggle.com*. https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset. 2022. (Visited on 29th Oct. 2023).

[8] Nabiha Asghar. 'Yelp dataset challenge: Review rating prediction'. In: *arXiv preprint arXiv:1605.05362* (2016).

[9] Michael Luca. 'Reviews, reputation, and revenue: The case of Yelp. com'. In: *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper* 12-016 (2016).

[10] Wikipedia. *Starbucks - Wikipedia — en.wikipedia.org*. https://en.wikipedia.org/wiki/Starbucks. 2023. (Visited on 29th Oct. 2023).

[11] Laura J Blake. 'Starbucks Enters India: The Indomitable competitor or underdog? Karen A. Berger, Pace University'. In: *Journal of Case Studies November* 34.2 (2016), pp. 75–91.

[12] Hui-Ping Ho, Ching-Ter Chang and Cheng-Yuan Ku. 'On the location selection problem using analytic hierarchy process and multi-choice goal programming'. In: *International Journal of Systems Science* 44.1 (2013), pp. 94–108.

[13] Wen-Hwa Ko and Chihwei P Chiu. 'A new coffee shop location planning for customer satisfaction in Taiwan'. In: *International Journal of the Information Systems for Logistics and Management* 2.1 (2006), pp. 55–62.

[14] Wen-Ko Liang and Rou-An Wu. 'Analysis of Coffee shop market'. In: (2012).

[15] Xiangyi Lin and Yuanyuan Zu. *Multi-criteria GIS-based procedure for coffee shop location decision*. 2013.

[16] Lisa Waxman. 'The coffee shop: Social and physical factors influencing place attachment'. In: *Journal of Interior Design* 31.3 (2006), pp. 35–53.

[17] Linder G Ringo. 'Utilizing GIS-Based Site Selection Analysis for Potential Customer Segmentation and Location Suitability Modeling to Determine a Suitable Location to Establish a Dunn Bros Coffee Franchise in the Twin Cities Metro, Minnesota'. In: *Papers in Resource Analysis* (2009), pp. 5–10.

[18] Viola Alfheny and Ali Ibrahim. 'Implementation of the simple multi-attribute rating technique (SMART) method for support selection of coffee shop business location'. In: *Jurnal Teknik Informatika (Jutif)* 3.4 (2022), pp. 963–968.

[19] Reflan Revife Purba et al. 'Decision Support System in the Best Selection Coffee Shop with TOPSIS Method'. In: *The IJICS (International Journal of Informatics and Computer Science)* 7.1 (2023), pp. 28–34.

[20] Sylvia Sylvia. 'Decision Making Method for KT Kopi Café Location Selection Using Analytical Hierarchy Process Method'. In: *Journal of Industrial Engineering & Management Research* 3.5 (2022), pp. 163–171.

[21] YY Wibisono and S Marella. 'A decision making model for selection of café location: an ANP approach'. In: *Journal of Physics: Conference Series*. Vol. 1477. 5. IOP Publishing. 2020, p. 052030.

[22] Tri Hendra Widadi and Dina Dellyana. 'Proposed business strategy for coffee shop based on customer preferences'. In: *Fair Value: Jurnal Ilmiah Akuntansi dan Keuangan* 5.8 (2023), pp. 3291–3305.

[23] Giles A Hindle and Richard Vidgen. 'Developing a business analytics methodology: A case study in the foodbank sector'. In: *European Journal of Operational Research* 268.3 (2018), pp. 836–851.

[24] Siti Aishah Mohd Selamat et al. 'Big data analytics—A review of data-mining models for small and medium enterprises in the transportation sector'. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.3 (2018), e1238.

[25] Umair Shafique and Haseeb Qaiser. 'A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)'. In: *International Journal of Innovation and Scientific Research* 12.1 (2014), pp. 217–222.

[26] Rüdiger Wirth and Jochen Hipp. 'CRISP-DM: Towards a standard process model for data mining'. In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Vol. 1. Manchester. 2000, pp. 29–39.

[27] Steffen Huber et al. 'DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model'. In: *Procedia Cirp* 79 (2019), pp. 403–408.

[28] Teo Yu Siang. *What is Design Thinking and Why Is It So Popular? — interaction-design.org*. https://www.interaction-design.org/literature/article/what-is-design-thinking-and-why-is-it-so-popular. 2022. (Visited on 4th Nov. 2023).

[29] Philipp Peron. *Project Github Repository*. https://github.com/PhilippPeron/applied-data-science. 2023. (Visited on 20th Nov. 2023).

[30] United States Census Bureau. TL_2020_42. https://www2.census.gov/geo/tiger/TIGER2020PL/STATE/42_PENNSYLVANIA/42/. 2020. (Visited on 15th Nov. 2023).

[31] United States Census Bureau. *American Community Survey — S0101 — Age and Sex*. https://data.census.gov/table/ACSST1Y2022.S0101?q=Age+and+Sex. 2020. (Visited on 15th Nov. 2023).

[32] United States Census Bureau. *American Community Survey — S2503 — Financial Characteristics*. https://data.census.gov/table/ACSST5Y2020.S2503?q=2020+Median+Household+Income. 2020. (Visited on 15th Nov. 2023).

[33] Philadelphia Parks and Recreation (PPR). *Parks Recreation Districts*. https://opendataphilly.org/datasets/parks-recreation-districts/. 2015. (Visited on 15th Nov. 2023).

[34] Philadelphia Universities and Colleges. *Philadelphia Universities and Colleges*. https://opendataphilly.org/datasets/philadelphia-universities-and-colleges/. 2017. (Visited on 15th Nov. 2023).

[35] MyGeoData Cloud. *Bus Stops And Stations in Philadelphia*. https://mygeodata.cloud/data/download/osm/bus-stops-and-stations/united-states-of-america--pennsylvania/philadelphia-county/philadelphia. (Visited on 15th Nov. 2023).

[36] Philadelphia Regional Rail Stations. *Philadelphia Regional Rail Stations*. https://gis-septa.hub.arcgis.com/datasets/SEPTA::regional-rail-stations/about. 2023. (Visited on 15th Nov. 2023).