# 

# 

# 

# 

# The Neighborhoods of New York City

# 

# By

# Gavi Book, Joshua Trauring, Andy Katz, Noah Watman, and Joseph Greenberg

# Team Overview

This project was a collaborative effort, leveraging the strengths and interests of team members to create a well-rounded outcome. A core group managed hard coding and data to ensure a solid technical foundation. Others focused on reporting, making complex information accessible. Brainstorming and individual research refined ideas into actionable outcomes. In the final presentation, tasks were divided among general and detailed content, presentation style and design, and business considerations, ensuring a cohesive and effective delivery.

In the course of executing this project, our team adopted a flexible approach to task assignments and responsibilities. Rather than adhering to rigidly defined roles, we leveraged our collective strengths and prioritized the completion of tasks based on immediate project needs. As reflected in our [work break down spreadsheet](https://docs.google.com/spreadsheets/d/172HXvb5K58MryY7LeHVOtHL8vvZ9RIOWWbr0CH_Xpao/edit?usp=sharing), roles evolved organically over time, ensuring that each team member contributed effectively to various aspects of the project. This adaptive strategy allowed us to respond dynamically to emerging challenges and capitalize on individual strengths.

# Business Scenario

The hunt for the right New York City neighborhood often overwhelms young adults and families. The multitude of considerations, especially when responsible for more than oneself, creates a pressing need for a decisive solution. We want to introduce a tool dedicated to simplifying this daunting process. Our platform specializes in catering to the specific needs of young families, offering a personalized approach to identify the perfect NYC neighborhoods. This platform aims to empower young families in finding their ideal NYC neighborhood, alleviating stress and uncertainty from this crucial decision-making process.

Our project's success depends on the utilization of our datasets to create a list of neighborhoods ideally suited for young families. We measure our success by the utility of our neighborhood scoring system and its ability to provide accurate, actionable insights that empower families to make informed decisions about where to start their families. We aim to eventually implement a way for our users to provide insights into how useful our service was. Customer feedback can track the rate of successful relocations based on the model. With this new feedback data we will improve our model with comparison of predicted vs. actual neighborhood satisfaction over time.

# Data & Tools

## Explanation of Data

Our team embarked on a comprehensive data collection journey, harnessing a variety of sources to paint a full picture of the elements vital to young families contemplating a New York City residence. This rich tapestry of data forms the backbone of our analysis, ensuring that every recommendation we make is rooted in a deep, nuanced understanding of the urban familial landscape.

Our dataset covered various aspects for evaluating neighborhoods for young families. For education, we assessed school quality through English Language Arts (ELA) and math proficiency and the availability of after-school programs. Our community amenities research included arts, dining, health services, landmarks, retail, and more. Financially, we looked at the child care cost burden, median household incomes, median rents for different-sized apartments, and average commute times by ZIP code. Societally, we included noise complaints and major crime statistics per ZIP code. Overall, this dataset facilitated a detailed assessment of neighborhoods, addressing educational, financial, and societal considerations.

## Data Sources

Our comprehensive dataset draws from a diverse range of sources, each serving a distinct purpose crucial for young families considering a move to New York City. From NYC Open Data, we incorporate valuable insights, such as noise complaints recorded since 2016, geospatial data for census tracking in 2022 and 2010, and school quality reports for early childhood and elementary school years in 2020-2021 and Elementary/Middle, respectively. This information was collected by NYC to provide an open platform for non-urgent community concerns, track neighborhood demographics, and evaluate school quality through the Quality Review process.

Information gathered from NYC.gov includes town separation data from the 2010 Census, fair market rent values for 2023, and historical crime data from 2000 to 2022. The purpose of NYC collecting this data was to aid in town separation for census data, provide information to NYC residents about fair market rent values, and record crime statistics according to NY state penal law.

Additionally, Data.cccnewyork.org provided an in-depth analysis of the U.S. Census Bureau data, focusing on the affordability of child care costs across various age cohorts from birth to age 12. This was collected to track and report the increasing child care costs in NYC. Each dataset is a valuable piece of the puzzle, collectively contributing to a nuanced understanding of prospective neighborhoods for young families in the dynamic landscape of New York City.

We also utilized median household income data from the American Community Survey (ACS) by the United States Census Bureau, focusing on New York City's Public Use Microdata Areas (PUMAs) aligned with community districts. This data was key to evaluating the economic well-being of neighborhoods, a crucial aspect of assessing their family-friendliness. The latest ACS data was meticulously filtered and aligned with NYC's community districts, ensuring accuracy and relevance in our socio-economic analysis. We addressed any data inconsistencies through median imputation and integrated this dataset with other demographic and geographic information, using ZIP codes as a unifying factor. This comprehensive approach provided vital insights into each neighborhood's economic status, contributing significantly to our overall assessment of family-friendly neighborhoods in NYC.

Similarly, we leveraged average commute time data sourced from the American Community Survey (ACS) by the United States Census Bureau, specifically focusing on Public Use Microdata Areas (PUMAs) corresponding to NYC's community districts. Recognizing that commute time is a significant factor for family life quality, we utilized this data to provide insights into daily travel burdens. The data was averaged for each PUMA and aligned with corresponding ZIP codes to ensure a coherent geographic analysis. This approach allowed us to integrate commute times into our broader neighborhood assessment, providing a more nuanced understanding of convenience and accessibility in family-friendly neighborhood evaluations.

Our project included after-school program data from NYC's Department of Youth & Community Development, detailing programs for elementary and middle school students in various ZIP codes. This vital data covers program timing and type, focusing on education, sports, and cultural activities. Where data was absent, we imputed zeros, presuming no programs existed. The comprehensive dataset, now restricted on the NYC Open Data website, highlights the unique value of our analysis in offering insights into educational and developmental resources for families evaluating different neighborhoods.

For amenities, our project's dataset encompassed various venue categories across New York City ZIP codes. Using the Foursquare API, we gathered information on venue locations and types, including Arts and Entertainment, Dining and Drinking, and more. We aimed to analyze venue distribution across ZIP codes, offering insights into each area's accessibility and availability of services for business and community planning. Data for each ZIP code and category was collected via the Foursquare API, recorded as counts in a CSV file for analysis.

This data collection was crucial as a predictor variable for our final model. Each dataset contributed to our predictive capabilities, allowing our models to identify patterns and provide insights for young families looking for homes in New York City's dynamic environment.

## Data Cleaning, Missing Values, and Errors

There were many issues with the data. Some of the issues included:

1. The most important component for joining our various messy data sets from multiple sources was to join them on a primary key. Without merging the individual data, finding data driven insights would be impossible. Our dataset utilized ZIP codes as the primary key, but as some datasets lacked ZIP codes, we employed the Google Maps API to match longitude and latitude coordinates with the corresponding ZIP codes. We then continued to utilize geospatial data integration and established connections among diverse area codes including neighborhood tabulation areas (NTAs), community district tabulation areas (CDTAs), public use microdata areas (PUMAs), and census tracts.
2. Our dataset contained numerous N/As and empty cells. We needed to determine the significance of these problematic cells for our final analysis. In some instances we imputed these cells with a value of 0, for others we factored in median calculations.
3. Although not directly related to data quality, our individual datasets each consisted of diverse predictor variables from different datasets. These variables often existed in varying units, making it challenging to integrate them for overall scoring. To address this, we employed a normalization process. This method standardized the metrics across different units, allowing us to integrate into the analysis without affecting the results due to differing scales. This normalization approach enabled equal consideration of each metric, allowing us to make a comprehensive analysis of the compiled data.
4. Our data had many different column headers which forced us to standardize them within our dataset. Standardizing these headers was crucial to ensure uniformity across the dataset.
5. We encountered inconsistencies in data structures, ranging from strings to lists and booleans. We ensured and checked that each factor's data maintained the correct data structure to join with the other datasets.

## Tools

**RStudio**: RStudio was essential for two main tasks: cleaning and aggregating datasets and converting longitude and latitude data into ZIP codes with the ggmap package and Google API (revgeocode). Securing API access was initially challenging and costly, nearly exhausting our $50 budget from our professor. Consequently, we shifted to using unique longitudes and latitudes directly, reducing but still exceeding our budget. Integrating ggmap in RStudio was more efficient than using Tableau directly, as Tableau performed better with ZIP code-based data

**Python:** Using Python, we faced challenges due to its steep learning curve, particularly for team members less experienced in coding. We initially tried seaborn for data exploration and visualization but found its requirements for detailed coding and formatting too time-consuming and complex. As a result, we dropped seaborn from our toolkit.

Python remained the cornerstone for data manipulation, analysis, and modeling, utilizing Pandas for data cleaning and preprocessing, NumPy for numerical computations, and scikit-learn for machine learning. Its wide-ranging libraries and community support greatly enhanced our efficiency, especially in handling missing data and merging datasets within Pandas.

**Tableau:** After selecting the optimal weighting method, we imported the resulting neighborhood scores into Tableau for dashboard creation. Tableau was also key in producing a user-friendly visualization map, utilizing heat coloring based on ZIP codes to provide an intuitive and visually engaging interface for our project.

**Microsoft Excel:** Excel played a crucial role in our data cleaning process, especially during the initial stages when comprehensively assessing the data. Using filters we efficiently organized and categorized diverse datasets, offering an insightful overview. Additionally, Excel served as a tool to streamline our dataset by removing redundant columns, allowing for the deletion of highly irrelevant data that didn't align with our model's objectives.

**Power BI:** After careful consideration, we opted against utilizing Power BI due to Tableau's superior integration capabilities for geospatial data. We found Tableau to offer a more seamless and effective integration, particularly when handling geospatial data, aligning better with our project requirements and objectives.

# Approach

## Ideation

Our approach evolved dynamically, beginning with a pivot from a real estate valuation model to a unique, family-centric neighborhood evaluation tool. Acknowledging the saturation of market estimators like Zillow's 'Zestimate,' we sought to carve a niche by focusing on the nuanced needs of young families in NYC. This shift was not merely a change in direction but a refinement towards a more impactful and differentiated service

We then went back to the drawing board and discussed an idea regarding Emergency Medical Services and allocation of resources. After that idea didn't pan out, we shifted to creating an accessible platform to assist young parents with financial management, particularly for baby-related expenses. Initially conceived as a dashboard, it evolved into an interactive app designed to offer real-time, actionable insights for smarter purchasing decisions. Unfortunately, the lack of pertinent datasets impeded the development of accurate models and predictions.

Nevertheless, after discussions with Professor Shin, we adapted our strategy to focus on specific items, like diapers, gathering real-time data through retail APIs for a browser extension that provides live price reports, akin to Honey. This aimed to help parents save on essentials. Despite our commitment, data quality issues persistently challenged the project's feasibility.

Continuing our dialogue with Professor Shin, we merged the initial real estate concept with aiding new parents, leading to a project focused on identifying ideal living locations tailored to young families' needs. This concept incorporated geolocation and real estate elements with various parental support aspects, adapting our initial ideas into a cohesive solution.

## Finalization of Our Idea

Our strategy focused on gathering data relevant to parks, schools, and crime rates to define family-friendly neighborhoods. We sourced varied datasets, standardizing formats through individual team efforts to clean missing or erroneous data. After refining the data, we integrated it by ZIP codes, often using Boolean expressions to sort observations. This resulted in interconnected datasets with ZIP codes as the primary key, ready for analysis.

After consolidating all our datasets into a unified table and resolving the N/As, we proceeded to implement our weighting methodologies. Our objective was to discern the most effective approach for establishing the optimal "family-friendly score". Once identified, this score would enable us to generate a comprehensive list of the most suitable ZIP codes in NYC for families looking to relocate.

Upon consolidating the initial three datasets, with our professors guidance, iit became evident that additional factors influence the choice of a neighborhood for relocation beyond the parks, crime, and schooling. To capture a comprehensive array of decision-influencing variables for young families, we sought more relevant datasets. This necessitated a similar process to that of the first three datasets, involving the sourcing and integration of additional datasets.

The outcome of this process yielded a more comprehensive, data-driven insight into the decision-making dynamics for young families when selecting a neighborhood. This better understanding reflected a more realistic and holistic perspective, encompassing a wider range of factors influencing their decision-making process.

## 

## Methodology

### Objective

The objective of our project was to identify family-friendly neighborhoods in New York City (NYC) using a comprehensive, data-driven approach. By analyzing various metrics critical to urban living and families, we aimed to provide actionable insights into which neighborhoods of NYC offer the best conditions for families.

### Data Collection

Our initial step in analyzing our original datasets was to pivot to more data collection, where we sourced datasets from official NYC Open Data repositories, APIs, and other credible sources; each offered comprehensive insights into the various factors affecting neighborhood desirability. These datasets included information on commute times, rent costs, academic performance, amenities, household income, noise complaints, and crime rates across different ZIP codes.

### Data Preparation and Normalization

#### Integration and Cleaning:

In the data preparation and normalization phase, we collected various datasets that required extensive cleaning and standardization, particularly concerning standardizing ZIP code columns for consistent analysis across different datasets. The raw data varied in format, necessitating a thorough cleaning process. Our team standardized ZIP code formats to ensure consistency dropped duplicate entries based on ZIP codes to maintain data integrity, and filtered the data to only include operating facilities, which provided a more accurate representation of available amenities. In instances where multiple area codes corresponded to a single ZIP code, we aggregated the data carefully to ensure a fair representation of each neighborhood's characteristics. This included calculating average values where appropriate and considering the geographic size of the areas involved.

#### Averaging Underlying Data:

Many factors are based on averaged data to provide a stable and generalized view of the neighborhoods. The factors incorporated into our analysis, such as academic performance scores, rent costs, commute times, and crime rates, represent averaged data of underlying statistics specific to each factor. For example, the academic performance score is an average of ELA and Math proficiency scores from various schools within the ZIP code. In crime rates, there is an average of yearly crime data, providing a more stable and generalized view of the neighborhoods. With this generalized view of each data set, we ensure that our model's insights are based on consistent, representative figures rather than outliers or anomalies.

#### Imputation Strategies:

For missing data, we used two imputation strategies: median imputation for continuous variables and zero imputation for indicators where absence likely indicates a null value, such as amenities or after-school programs. Additionally, many of the features were averaged for reasons already mentioned.

#### Normalization:

The MinMaxScaler method was employed to normalize the features, ensuring all data sets were on a common scale – crucial for effective composite score comparison.

#### Geospatial Integration:

Due to the heterogeneity of geographic identifiers across datasets, we utilized geospatial joins and mapping to align various geographic markers like community districts, NTAs, and ZIP codes to a common ZIP code basis, ensuring that the analysis was consistent and robust. Utilizing geopandas, a Python library for geographical data manipulation, we read and wrote geospatial data. We applied spatial joins to connect disparate geographic identifiers with their corresponding ZIP codes based on their spatial relationship and ensured that our geospatial data was consistent and standardized to the same Coordinate Reference System (CRS) for accurate spatial joining.

#### Data Enrichment:

By enriching ZIP code data with additional geographic dimensions, we were able to incorporate a wider range of datasets into our analysis. This enriched data provided a more detailed and nuanced understanding of each neighborhood’s profile, allowing for a more comprehensive analysis of factors contributing to neighborhood desirability.

### Score Computation and Feature Engineering

Each neighborhood’s characteristics were categorized and quantified into individual scores across various dimensions crucial to family living, including academic performance, income levels, commute time to work centers, median rent cost for housing, noise complaints registered in the area, availability of amenities, number and quality of after-school programs, cost burden of living, and crime rates.

We engineered features to represent various aspects of urban living, including academic performance, income level, commute time, rent cost, noise complaints, amenities, after-school programs, cost burden, and crime rates. For metrics where a lower value is more desirable, such as noise complaints, rent costs, commute times, cost burden percentages, and crime rates, we inverted the scores. This inversion ensured that for all categories, a higher score consistently indicated a more favorable outcome for family-friendliness. For instance, an area with *fewer* noise complaints or *lower* crime rates would score higher after inversion, aligning with our goal of identifying tranquil, safe neighborhoods.

### Weighted Scoring

We recognized that defining a “family-friendly” neighborhood can vary significantly based on individual or policy priorities. To accommodate these diverse needs and provide a more nuanced understanding of neighborhood rankings, we implemented three different weighted scoring methods: prescribed, equal, and iterative weights. Each method offers a distinct approach to combining and prioritizing the various metrics we analyzed. We assigned specific weights to each category based on presumed importance to family living conditions in the Prescribed Weights Method. This method allows for a targeted approach where certain factors are emphasized according to common or established priorities. To ensure a balanced consideration of all factors, we also applied equal weights across all categories in the Equal Weights Method. This approach treats every dimension as equally important in contributing to the overall family-friendliness of a neighborhood, providing a uniform assessment strategy. Recognizing that different families may prioritize certain aspects of living over others, we implemented an Iterative Weights Method to explore a range of weight combinations. This method iteratively adjusts the weight of one category while proportionally distributing the remaining weight among other categories, allowing stakeholders to see how varying emphases on certain criteria can change the ranking of neighborhoods while providing a customizable and exploratory tool for understanding the impact of individual preferences on the overall assessment.

### Method Adjustments

Throughout the project, we adjusted our methods in response to data insights. For instance, the inversion of scores where lower values are desirable was refined after initial exploratory data analysis (EDA) suggested that simply inverting the scale did not accurately reflect desirability.

### Machine Learning Models and Evaluation

#### Model Selection and Optimization:

To classify neighborhoods into tiers of desirability, we initially experimented with various machine learning models, including Logistic Regression, Random Forest, KNN, SVM, and Decision Trees. However, after careful evaluation, we opted for Logistic Regression and Random Forest classifiers due to their exceptional performance in handling both linear and non-linear data, demonstrating robustness and effectiveness. Known for its efficiency and interpretability, Logistic Regression was optimized using GridSearchCV, which systematically worked through multiple combinations of parameter tunes, cross-validating as it went to determine which tune gives the best performance. This provided the best regularization strength ‘C’ as 10. This parameter helps prevent overfitting by penalizing larger coefficients, making the model more generalized. Random Forest, this ensemble method known for handling complex datasets with multiple features, was tuned to identify the optimal ‘max\_depth’ and ‘n\_estimators’. The depth of trees was left unbounded (None) and the number of trees was set to 100, allowing the model to learn as much as possible from the training data.

#### Model Evaluation Metrics:

We evaluated the models on precision, recall, F1-score, and accuracy. These metrics were chosen to provide a comprehensive understanding of model performance, considering aspects like the balance between precision and recall, the harmonic mean of these two metrics (F1-Score), and the overall accuracy rate. Precision indicates the proportion of positive identifications that were actually correct, implying a low false positive rate crucial for ensuring that only truly desirable neighborhoods are recommended. Recall indicates the proportion of actual positives that were correctly identified, critical for ensuring that most of the desirable neighborhoods are not missed out. F1-Score, the harmonic mean of precision and recall, provides a single score that balances both the concerns of precision and recall, particularly important when the class distribution is uneven. Overall, accuracy measures how often the model is correct, while it's not the sole focus due to the potential imbalance in class distribution.

### Model Outputs and Comparative Analysis

In the model outputs and comparative analysis, with the best parameter ‘C’: 10, the Logistic Regression model showed an excellent balance between precision and recall across different classes. Its overall accuracy was commendable, demonstrating its effectiveness in the prescribed method more prominently. The model’s interpretability is an added advantage, providing clear insights into feature importance and impact. Best parameters {‘C’: 10} showed an average precision of **0.86**, recall of **0.87**, and accuracy of **0.86** for the prescribed method. A confusion matrix was used to describe the performance of the classification model, showing the actual versus predicted values in a matrix format. The matrix for Logistic Regression looks like this: [[ 8 2 0] [ 1 12 2] [ 0 0 11]], where the diagonal elements (8, 12, 11) represent correct predictions while off-diagonal elements show incorrect predictions.

The Random Forest model, with its complexity and ability to model non-linear boundaries, also performed well, showing a slightly lower performance in terms of precision and recall but was particularly effective in capturing the complex patterns and relationships in the data, as evidenced by its confusion matrix and F1-scores. Best parameters {‘max\_depth’: None, ‘n\_estimators’: 100} with an average precision of **0.80**, recall of **0.81**, and accuracy of **0.78** for the prescribed method.

When comparing both models, Logistic Regression generally showed higher precision and recall, outperforming the Random Forest model. This indicates its suitability for this specific task where a balance between false positives and negatives is crucial. The models are evaluated across three different classes (0, 1, 2), which could represent different levels of neighborhood desirability based on the scoring criteria. The choice of Logistic Regression was further solidified by its performance stability across different weighting methods, particularly with the prescribed weights, where it consistently provided reliable classifications.

### Quantitative Summary and Insights

#### Top Neighborhoods and Accuracy:

The analysis identified specific ZIP codes as top candidates for family-friendly neighborhoods, with scores reflecting the composite measure of all factors considered. The range of model accuracy from **67%** to **86%** indicates a substantial ability to predict neighborhood tiers accurately. The prescribed method with its carefully considered weights consistently provided the most reliable and nuanced classification.

#### Insights from Weighting Methods:

The prescribed weighting method emerged as the most effective primarily due to its tailored approach to reflect the complex reality of urban living. The method considers empirical evidence, assigning weights to different factors according to their importance in determining a neighborhood's family-friendliness. This method's superiority was evident in the more accurate and consistent results it produced compared to the equal and iterative methods.

#### Successes and Failures:

In evaluating different weighting methods for scoring neighborhoods, the project revealed that while the Prescribed Weight Method delivered a top neighborhood score of 0.683976 with an 86% accuracy in logistic regression, the Equal and Iterative Methods yielded top scores of 0.658596, showing consistency in the top-ranked neighborhoods across methods. The top 10 neighborhoods largely overlapped across all weighting schemes, indicating a substantial agreement in what constitutes the top-tier neighborhoods. Despite the variations in weighting, the slight shifts in neighborhood scores did not lead to fundamentally different rankings, suggesting the robustness and dominance of the selected features in determining neighborhood quality. This consistency across different methods underscores the importance of a well-considered approach to assigning weights, relying on domain knowledge and the specific analysis objectives, as minor adjustments in weights did not significantly alter the outcome or provide new insights. The results highlight the strength of the underlying data and model, affirming the relevance of the chosen features and the stability of the prescribed method as a reliable baseline for neighborhood evaluation.

### Overall Implications

The quantified results and insights from this project provide a solid foundation for families looking for a neighborhood. The use of data-driven methodologies, rigorous evaluation, and careful consideration of various urban living factors ensure that the recommendations are both reliable and relevant. The project demonstrates the significant potential of applying advanced analytics and machine learning techniques to urban planning and community development, setting a precedent for future initiatives in this domain.

### Ranking Top Family-Friendly Neighborhoods

Based on the optimal weighting method, a comprehensive list highlighting the most family-friendly neighborhoods was generated. This ranking, established through our prescribed and most consistent weighing scale, offers a glimpse into the top 10 communities based on their “family-friendliness score” (see Appendix for an example of the Top 10 Neighborhoods list).

### Visualization: Tableau

In our pursuit to exceed the basic project requirements, our team, under the mentorship of our professor and leveraging her extensive experience in data analysis, ambitiously expanded the project’s scope. We integrated our findings into a Tableau dashboard, significantly enhancing the presentation and accessibility of our data-driven insights. This initiative was driven by our commitment to providing a practical, user-friendly tool for young families navigating the complexities of finding a suitable home in New York City. The dashboard, featuring an intuitive interface and interactive elements, allows users to explore the top neighborhoods based on various family-friendly criteria. This enhancement not only enriched our analytical skills but also significantly increased the utility and impact of our project, bridging the gap between complex data analysis and real-world applications. The dashboard allows users to customize their preferred communities within a specified range, providing a visual representation atop a map of NYC.

Tableau was used to create clear, concise bar charts displaying the top 10 family-friendly neighborhoods in NYC, ranked by an overall score considering our factors. Each bar was labeled with the neighborhood’s ZIP code and its respective overall neighborhood score. In addition, a map highlighted the top neighborhoods, marked with symbols that varied in color to represent each area's ranking and score. This visual tool allowed users to understand the geographic distribution of family-friendly areas within the city. (Please refer to the Appendix for a comprehensive explanation of the Tableau integration.)

### Individual Member Results/Analysis

Each team member contributed to different parts of the project, including data cleaning, feature engineering, model training, and evaluation. The combined effort resulted in a robust analysis and a set of models that could predict neighborhood tiers based on the defined criteria.

### Conclusion

The detailed methodologies, performance metrics, and comprehensive analysis presented in this section underscore the project’s success in utilizing data-driven approaches to identify family-friendly neighborhoods. The prescribed weighting method, combined with the Logistic Regression model, proved to be the most effective in providing accurate and reliable predictions. Our findings offer valuable insights and tools for families and policymakers alike, contributing to informed decisions that enhance urban living and community development. The project’s commitment to quantification, precision, and practical relevance sets a high standard for future works in urban data analysis and planning.

# Summary of Results

In our comprehensive study to identify family-friendly neighborhoods in New York City, we employed robust data analysis and machine learning techniques to quantify and rank neighborhood desirability based on factors important to family life. Our key findings are as follows:

We identified top neighborhoods for families using a prescribed weighting method, highlighting areas such as ZIP codes 10309, 11004, 10019, 10013, and 10021. These neighborhoods stood out for their blend of safety, educational quality, amenities, and affordability. The implementation of different weighting methods — prescribed, equal, and iterative — allowed us to consider various priority settings, with the prescribed method offering particularly balanced and insightful results.

One surprising insight from our study was the significant variation in family-friendliness scores across different neighborhoods, revealing the diverse range of living experiences within the city. Some neighborhoods traditionally considered highly desirable did not score as high when analyzed through the lens of family-centric metrics, emphasizing the nuanced nature of urban living and the importance of a data-driven approach to neighborhood evaluation.

In terms of machine learning efficacy, our models demonstrated strong predictive capabilities, with Logistic Regression, at a regularization strength of 'C': 10, providing high accuracy and interpretability. These models are not only useful for current analyses but also scalable for broader applications, such as adapting the approach for different cities or specific demographic needs.

This project highlights the value of employing data-driven methodologies in urban planning and personal decision-making. With refined models and expanded datasets, our approach can continue to offer nuanced insights and practical recommendations for families seeking the ideal neighborhood.

# Recommendations and Next Steps

In the culmination of our project, we recommend highlighting the best neighborhoods for families based on our comprehensive analysis. This recommendation will serve as a guide for families considering relocation and looking for communities that align with their needs. Moreover, our findings can significantly aid policymakers and urban planners in understanding neighborhood dynamics and making informed decisions geared towards family welfare and urban development.

To expand the project's impact, we propose extending the research scope to include a broader geographic range, encompassing not just additional cities but perhaps different urban and rural settings. This expansion aims to create a universal tool that families anywhere can rely on and provide valuable insights to policymakers across various regions.

Moreover, recognizing the diversity in family needs and urban dwellers' preferences, adapting our methodology to cater to different demographic requirements presents a valuable opportunity. Whether it's adjusting the model to identify elder-friendly areas, regions best suited for young professionals, or the most accessible neighborhoods, the tool's versatility can significantly increase its utility.

A critical next step involves enhancing the user experience by allowing for input and customization. By developing an interface where users can specify their preferences—such as proximity to schools, green spaces, or specific amenities—the tool can provide tailored recommendations, making it a personalized decision-making assistant.

Lastly, transforming our insights into a user-friendly web application will make the data accessible and interactive to a broader audience. This application would not only provide information but also engage users in an interactive experience, allowing them to visualize data, modify search parameters, and understand how different neighborhoods meet their unique criteria.

Implementing these recommendations and next steps will transform our project from a static analysis into a dynamic tool that continues to evolve and adapt to user needs and preferences. It will pave the way for informed decisions, better quality of life, and more cohesive communities, aligned with the diverse needs of urban families today and in the future.

# 

# Appendix

***In the interest of transparency and collaboration, all referenced code, datasets, and visualizations related to this project are conveniently accessible in our*** [***shared Google Drive folder***](https://drive.google.com/drive/folders/14Igs2jI-0MaUFQ8_6hD5RIjd0TkUX7Ux?usp=share_link)***.***

## Bibliography

**NYCOpenData**

* Noise complaints 2016<https://www.kaggle.com/datasets/somesnm/partynyc>
* Geospatial Data. 2022<https://data.cityofnewyork.us/widgets/9nt8-h7nd?mobile_redirect=true>
* Geospatial Data. 2010<https://data.cityofnewyork.us/City-Government/2010-Census-Tracts/fxpq-c8ku>
* School quality for early childhood. 2020-2021<https://data.cityofnewyork.us/Education/2020-2021-School-Quality-Reports-Early-Childhood-S/ks4m-pn8b>
* Police Precinct data. 2013.<https://data.cityofnewyork.us/Public-Safety/Police-Precincts/78dh-3ptz>
* Felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) 2006.<https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i>
* School quality for elementary school. 2020-2021.<https://data.cityofnewyork.us/Education/2020-2021-School-Quality-Reports-Elementary-Middle/4dxw-kr73>
* After school programs <https://data.cityofnewyork.us/login>

**NYC.gov**

* Census Bureau Data 2010<https://www.nyc.gov/assets/planning/download/pdf/data-maps/nyc-population/census2010/puma_cd_map.pdf>
* Rent values 2023<https://www.nyc.gov/assets/hpd/downloads/pdfs/services/2023-fmr-and-safmr.pdf>
* Crime and offense data 2000-2022<https://www.nyc.gov/site/nypd/stats/crime-statistics/historical.page>

**Data.cccnewyork.org**

* Analysis of U.S. Census Bureau (Child burden), 2014-2021.<https://data.cccnewyork.org/data/bar/1450/cost-burden-for-infanttoddler-child-care#1450/454,455,456/1/1751/127>

**Data.Cencus.gov**

* Census data with income & commute time. 2022 [https://data.census.gov/table/ACSST1Y…](https://data.census.gov/table/ACSST1Y2022.S1903?q=new%20york%20city%20income&g=795XX00US3604103,3604104,3604107,3604108,3604109,3604110,3604111,3604112,3604121,3604165,3604204,3604205,3604207,3604208,3604209,3604210,3604211,3604212,3604221,3604263,3604301,3604302,3604303,3604304,3604305,3604306,3604307,3604308,3604309,3604310,3604311,3604312,3604313,3604314,3604315,3604316,3604317,3604318,3604401,3604402,3604403,3604404,3604405,3604406,3604407,3604408,3604409,3604410,3604411,3604412,3604413,3604414,3604501,3604502,3604503)

**Foursquare**

* Amenities <https://location.foursquare.com/developer/reference/place-search>

## Appendix: Top 10 Neighborhoods Example

| **ZIP\_Code** | **Borough** | **Neighborhood** | **Overall\_Score\_Prescribed** |
| --- | --- | --- | --- |
| 10309 | Staten Island | South Shore | 0.683976 |
| 11004 | Queens | Southeast Queens | 0.683115 |
| 10019 | Manhattan | Chelsea and Clinton | 0.678393 |
| 10013 | Manhattan | Greenwich Village and Soho | 0.672629 |
| 10021 | Manhattan | Upper East Side | 0.664826 |
| 11361 | Queens | Northeast Queens | 0.658896 |
| 10312 | Staten Island | South Shore | 0.658625 |
| 11365 | Queens | Central Queens | 0.658207 |
| 10306 | Staten Island | South Shore | 0.653907 |
| 10075 | Manhattan | Upper East Side | 0.643581 |

## Appendix: Tableau Workbook Walkthrough

### Top Neighborhoods Dashboard

This dashboard presents a visual representation of the top neighborhoods for families, combining a map and bar graph based on our prescribed weight scores for each ZIP code. The map utilizes a blue color gradient to denote the score intensity, with darker shades indicating higher scores. Selecting a ZIP code in either visualization highlights it in both, providing a cohesive view of its score and geographical position. Clicking on a ZIP code reveals essential details such as neighborhood name, borough, tabulation area, community district, precinct number, and PUMA code. Additionally, there's a hyperlink leading to a detailed breakdown of category-specific scores for the neighborhood.

### Adjustable Weights Dashboard

This interactive dashboard allows users to personalize the weight of each factor according to their preferences. It encompasses all ZIP codes, denoted by a color gradient from dark red to dark green, indicating the lowest to the highest scores. Users can manipulate the weights via sliders for each factor, with the visualizations updating in real-time. Selecting a ZIP code provides the same detailed information as the first dashboard and links to further category breakdowns. However, the unrestricted nature of weighting could lead to illogical total scores exceeding 100%, a limitation noted due to the absence of total weight constraints in Tableau.

### Dynamic Dashboard

Responding to the limitations of the Adjustable Weights Dashboard, this dashboard incorporates user interactivity with predefined scenarios. Users select their priority scenario from options like Family Friendliness Focused, Academic Focused, and more. Each scenario alters the weight distribution across different factors. Like the previous dashboards, this also features a map and bar graph indicating top neighborhoods based on the selected scenario, with a similar interactive and detailed data presentation. This approach ensures a dynamic yet controlled user experience, avoiding the unrestricted weighting issue.

Each dashboard is crafted to provide a different level of interaction and information depth, from direct neighborhood ranking to customizable weight distribution, offering various insights and user experiences. The third dashboard particularly addresses the limitations observed in the second, providing a more refined and realistic user interaction without compromising the integrity of the neighborhood scores. These visualizations serve as a powerful tool for families, policymakers, or researchers interested in the multifaceted aspects of neighborhood livability.





