Rental Prediction Helper - Team 91

P. Chen (*pchen369*), C. Pomeroy (*cpomeroy6*), S. Arif (*sarif30*), N. A. Sadeeshkumar (*nsadeeshkumar3*)

I. Introduction

Our team is creating a new and improved renting helper. The tool will take a variety of features and make a prediction about fair rental prices for different types of units, estimated increase or decrease of rental prices per neighborhood and a summary about the area. There are a few applications similar to this already existing such as Zillow, Craigslist, and etc. We aim to improve and expand on these current applications. The application is designed to provide renters with information about different neighborhoods in a new city, including predicted rental prices and proximity to amenities, displayed on a map. This is different from rental listing sites like Zillow, as the focus is on helping users identify their ideal areas and neighborhoods based on their needs, which they can then use on rental listing sites to find the perfect rental property. We hope to serve people who are looking to rent without a lot of knowledge about the area.

Our project will introduce a one-stop-shop that shows fair price prediction and neighborhood information in an interactive dashboard. Prediction model performance will be evaluated using regression metrics based on a pre-prepared out-of-time testing dataset. The dashboard will only serve New York City neighborhoods in the beta version because of the high demand of rental properties and more available data. The model carries the risk of inaccurate price forecasts due to the volatility of the rental market and the inability to curate a diverse dataset. However, the payoff is significant, as the project will help individuals and families make informed rental decisions, save money, and find suitable rentals based on their geographic needs. We are developing using community edition tools and free tiers so we will not have any costs unless we scale the level of this project.

In the subsequent sections, we will delve into the methodologies employed for the three primary components of the rental helper. Following this, we will explore the project's innovative aspects, and discuss the experiments and measurements we will undertake to ensure the tool's quality. Finally, we will present the finding and conlusion of the project

II. Problem Definition

The majority of existing websites and tools presently offer only basic information about rental properties, such as listings, photographs, and asking prices. However, prospective renters require a more comprehensive understanding of various factors when choosing a place to rent. These factors include neighborhood characteristics, nearby amenities such as restaurants and stores, and public transportation access. Additionally, renters are keen to determine whether the asking price represents a good deal and anticipate potential market fluctuations if they decide to wait before committing to a rental agreement. Our tool addresses these concerns by providing a comprehensive, all-inclusive platform that consolidates and visually presents this vital information to renters in a user-friendly manner.

III. Literature Survey

Understanding rental prices in the UK

In this paper [1] McCord et al. look at the variation in rental prices in different areas of Belfast, Northern Ireland and identify sub-markets using several spatial modelling approaches including hedonic regression, SAR models, and geographically weighted regression. The results show that geographically weighted regression produced the most accurate predictions and that the market was heavily stratified into sub-markets. In our project, we want to focus on different geographic areas and could make use of the spatial approaches described in this paper, but we intend to extend this by creating an interactive map of the sub-markets.

Towards increasing residential market transparency

This article [2] uses hedonic regression with ordinary least squares, quantile regression, and geographically weighted regression to predict apartment prices in Poznań, Poland and display them on a map. The authors presented their findings on the map in many different ways and we can use these as potential inspiration for our project, as well as trying their regression method. The biggest limitation to this research was that the market in Poznań is quite small and some areas of the city have very little data. In our case, we intend to repeat this type of project in large markets in North America such as New York and Toronto, and we intend to create map visuals that are interactive.

Beyond Spatial Auto-Regressive Models

This study [3] used a CNN to model housing prices in several cities in the UK based on satellite imagery and compared it to the results from a traditional SAR model. We don't intend to use image data for our project, so the modelling methodology will not be helpful for us, however the study also focused on points of interest such as nearby amenities and found that combining either the features extracted from the CNN or the point of interest features with the property data produced similar results. For us, the important part of this study to focus on is the point of interest features they collected and we will extend this by creating an interactive interface.

Forcasting Spatially Correlated Targets

The paper [4] presents a methodology for forecasting the housing price and volume of multiple areas simultaneously. The model takes times series data of housing prices in multiple target areas and utilizes 3 stacked LSTM layers with multiple targets to predict prices of

all areas simultaneously. The spatial relationship is reflected implicitly using fully connected layers. The limitation of this method is that it can only make predictions for existing locations. However, we can use this method to bring information of existing locations up-to-date, then train new models to make predictions for new locations.

Predict Italian Real Estate Market Prices

In this paper [5], the authors discuss their use of three different machine learning algorithms - ElasticNet, Xgboost, and Artificial Neural Network - to predict Italian real estate prices. Their work demonstrates the potential of ML algorithms for predicting real estate market prices and points out the model limitations on predicting prices for the most expensive houses for all ML models they experimented. In our project, we should try data normalization or transformation to mitigate this limitation.

Housing Price Prediction Using ML Algorithms

The paper [6] discusses the development and testing of three bagging and two boosting machine learning algorithms for predicting housing prices using 28 features that fall into 6 categories. Their feature engineering and selection process provided considerable insights on data preprocessing for our project.

A new approach for predicting rents using ML

The study [7] uses over 4,000 rental listings, including variables such as location, size and amenities to train a model to predict rental prices. The results suggest that machine learning techniques can accurately predict rental prices in the Riyadh market with an average prediction error of around 7%. This can be useful for this project because we can use a similar approach using ARIMA, ANN and SVM models to generate predictions and apply it to other networks. It also mentions that unemployment rate, household size and interest rate will influence the prices in the area.

Data mining: practical machine learning tools

This chapter [8] focuses on regression, which is a type of machine learning algorithm that can be used to predict numerical values, such as real estate prices. The chapter covers a variety of regression techniques, including neural network regression. It covers a variety of steps that can be helpful for this project including data pre-processing, feature selection, and model evaluation. It is a bit general to data mining, but additional steps can be considered to apply to this project.

Combining structured and unstructured info sources

This article [9] discusses a study of user-contributed information on Zillow.com, a leading real-estate information service in the US. The study found that user contributed information improves the completeness and level of details of the information on the site, but the accuracy of user-contributed facts may not be high. The authors identified several weaknesses in the user-contributed information, including conceptual challenges, information integration failures and design deficiencies. It could be helpful to understand the shortcomings of Zillow and improve on that. It will be especially helpful to focus on building a user-friendly interface.

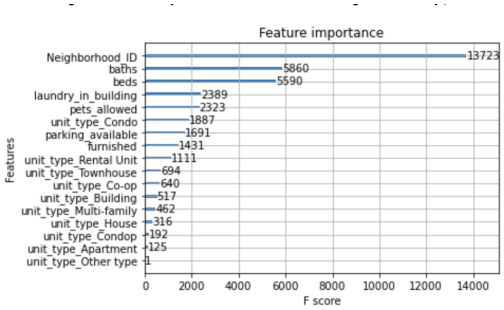

Figure 1: Feature Importance

Predicting the rental value of houses

This research aims to predict the rental value of houses in various countries. The authors compare the traditional Ordinary Least Squares (OLS) method for predicting rental prices against machine learning algorithms such as Ridge/LASSO regression and Random Forest. The study found that ML methods are able to account for spatial autocorrelations better and hence predict rental prices better than traditional OLS methods. For our project, we can build upon similar algorithms that are more robust and complex.

An extrapolative model of house price dynamics

In this paper, the authors conclude that many salient features of house prices such as excess volatility, momentum, and mean reversion can be explained by a model which incorporates naïve inference. Naïve inference occurs when sellers underestimate the momentum of the housing market and post prices too low during booms and too high during busts. The researchers outline an algorithm that uses the OLS forecast rule while considering naïve inference to predict prices. For our model, we can also use these principles in combination with more advanced ML algorithms to create an accurate and precise model.

Differences in housing price forecastability

In this research, housing prices across US for 1995-2006 are forecasted using an autoregressive benchmark model and models based on regional and national economic variables. The researchers discovered that models that take into account geographical economic variables can more accurately predict prices. However, states with high housing price growth, such as coastal states, are more difficult to forecast despite accounting for economic fundamentals for these states. These findings are important because for our model we can also take into account regional economic differences across states to create a more heterogeneous and unbiased model.

IV. Method

a. Visualization

To create the map interface we found a GeoJSON file that defines the neighborhood outlines in New York City[10] and used this with d3.js to create a map projection of the city with the neighborhood outlines as paths. Each neighborhood is coloured by it's predicted rental price from the regression model in a chloropleth map.

We also created a tooltip for each neighborhood that shows its name and predicted rent upon mouseover, with a form on the left allowing users to customize their rent search based on important features such as the number of bedrooms and bathrooms. Any form selection triggers a real-time update of the map colors and tooltip.

We added on click behaviour to the neighborhoods so that when a neighborhood is clicked, it is highlighted and the neighborhood summary information comes up on the right side of the screen. This summary includes the number of stores, number of restaurants, average income, and crime rate, as well as the predicted rent based on the user's selection, which is accompanied by an arrow based on our ARIMA model that indicates whether the rent is expected to increase or decrease as of February 2023. To make it easier to compare neighborhoods, we added a colored box beside each of the statistics, which uses d3's scaleQuantile function to show how the selected neighborhood compares to other neighborhoods in the city.

Note that while some neighborhoods from the GeoJSON file lack rental data, such as JFK Airport, we have ensured that they are still visible on the map. These areas are colored in grey and show a predicted rent of "–".

b. Modeling

For the Renting Helper Tool, we have built a two-part model that predicts rental prices based on NYC neighborhoods in 2016, then uses a time series model to normalize them to today's date for predicting current rental prices based on neighborhoods. In addition, we have developed a time series model that forecasts rent for the next 6 months, indicating whether it will rise or fall. To build our models, we used necessary libraries such as pandas, NumPy, and scikit-learn. We loaded the dataset [11] from a CSV file into a pandas data frame and cleaned the data, removing missing values and unnecessary data. We then performed feature engineering and mapped the values in the dataset to the shape file mapping NYC to make predictions. During the exploratory data analysis, we found that the StreetEasy dataset had 52 features, including price based on 2016 data. We narrowed down the most correlated features to price and added a few more selections such as laundry, pets, parking, and furnished to make our Renting Helper Tool more useful.

For the price prediction model, we tested out various different models such as Random Forest Classifier, Linear Regression Model, Gradient Boost Regressor, XGBoost model and Hedonic Regression. After doing some hyper tuning, the XGBoost model provided the best performance, with an R2 score of 0.84.

PERFORMANCE METRICS		
RANDOM FOREST	Accuracy	0.21
LINEAR REGRESSION	MAE	1302
	RMSE	3149
	R2	0.4
GRADIENT BOOST	MAE	838
	RMSE	2123
	R2	0.72
XGBOOST	MAE	632
	RMSE	1725
	R2	0.84

Figure 2: Performance Metrics

If you take a look at figure 1, which shows the feature importance of the best XGBoost model, we can see similar results to our exploratory analysis efforts, indicating that neighborhoods, number of bedrooms, and number of bathrooms are the most important features for price prediction.

It is important to note that the linear regression model was not able to capture the complexity of the data, as shown by its high MAE and MSE and relatively low R2 score. The Gradient Boost model did improve the R2 score to 0.72, but the XGBoost model after hyper-tuning provided the best results. We will discuss more about the modeling experiments in the Experiments Section.

Regarding the time adjustment portion of the Renting Helper Tool, we used an ARIMA model with aggregated data from StreetEasy for each neighborhood going back 10 years from 2013 to February 2023. We used one neighborhood to best estimate the hyperparameters, where p=3, d=0, and q=1. The p parameter represents the number of autoregressive terms, the d parameter represents the number of times the time series needs to be differenced to make it stationary, and the q parameter represents the number of moving average terms. We calculated the inflation factor for each neighborhood by dividing the 2023 April projection value by the 2016 value and used it to multiply the price prediction to get the normalized time.

Finally, we created a projection of whether the rent will increase or decrease by averaging out the projection for the next year divided by the current date. We added an up arrow if the rent will increase and a downward arrow if the rent will decrease. We understand that this projection does not cover all scenarios since inflation is also determined by economic conditions, significant events, and other factors, but it serves as a baseline for this project due to limited data available.

c. Neighborhood Summary

Our dashboard displays neighborhood details, such as nearby restaurants or stores, income [12] and crime rates [13]. We used the Google Maps API to find restaurants

and stores in a given neighborhood. Specifically, we utilized the Google Maps Places library. Within the Places service we are using the **nearbySearch()** endpoint to search for places within a specified neighborhood. Using this endpoint, we created an API client that takes latitudinal and longitudinal points and a radius to define the area of interest in which to search for places. Additionally, the client takes a *searchtype* parameter, which we use to define what type of places we want the *nearbySearch* endpoint to return. For our project, we had two searchtypes, which are 'restaurant' and 'store'. We used our Maps API client to find restaurants and stores in each precinct of New York City until we had a list of thousands of restaurants and stores, which we then grouped by Neighborhood Tabulation Areas (NTA) to create an aggregated list that displays total number of restaurants and stores within a given neighborhood in NYC.

Additionally, we used the IRS SOI Tax Income and NYPD Crime data sets for each neighborhood. The crime data set has counts of major felony offenses for each precinct in New York, while the income data set has median incomes for each precinct. We used the pandas python library to join these data sets with the data we obtained using Google Maps API. The final table contains statistics for each neighborhood and was integrated with the d3.js library to create our interactive map dashboard.

d. Neighborhood Mapping

To integrate all the necessary data, we require multiple mapping table that connects the Area Name or IDs used in modelling and neighborhood summary to the one used in visualization. Our visualization incorporates the Neighborhood Tabulation Areas (NTA) for New York City (2020). Meanwhile, the neighborhood summary relies on both Police Precinct and Zip Code. The modelling data employs a more granular version of neighborhood names that reflect the names people use in their daily lives.

To link the Police Precinct or Zip Code to the NTA code, we utilized a GIS software called QGIS. We use this software to performs a spatial join of two sets of areas which requires the shape file of both sets. After loading the shape files into QGIS, we used the "Join attributes by location" tool to perform the spatial join. For this process, we chose "Intersects", "Within", and "Overlap" as geometric predicates and selected "Take attributes of the feature with largest overlap only (one-to-one)" as the "Join type."

However, the process of mapping the neighborhood name in the modeling data to the NTA is a bit more complex. First, we separated the NTA name by "-" since some NTAs contain multiple neighborhoods grouped together. Then, we used the character level n-gram Jaccard similarity with n=2 and similarity threshold=0.4 to compare the name in the modeling data to the separated NTA name. This helped us to link two-thirds of the neighborhoods in the training data. For the remaining neighborhoods, we manually mapped them using Google Maps.

V. Innovation

The project's first innovation lies in the comprehensive information provided through the interactive map interface, surpassing the limited information displayed on existing applications like Zillow and Craigslist. These current applications solely display the listing price and interior images of rental properties, leaving out essential details. Our tool addresses this gap by presenting additional information that renters need to make informed decisions. Firstly, it provides the predicted fair price given the requirements (e.g. number of bedrooms) of the unit, which is not currently available on existing platforms. Secondly, it provides neighborhood information such as average income and crime rate, which are factors renters typically consider before making a decision. Thirdly, it displays the proximity of various amenities, adding value to the rental property search. The ideal outcome would be to include the actual listing price on the map and provide a link to the listing to allow renters to view further details such as images of the interior of the property.

The project's second innovation is the combined approach we take to predict the current fair price of a rental property given the data constraints. Although it would be ideal to have recent data for New York City, we only have 2016-2017 data on rental prices based on property features, such as size, number of bedrooms, and number of bathrooms, etc. However, we do have aggregate time-series data on rental prices at the regional level for the past few years. To overcome this data limitation, we decided to use a regression model to predict the price initially and then use an ARIMA model to adjust the price to the current market value. This approach allows us to provide accurate rental price predictions despite the data constraints.

VI. Experiments

For the renting helper tool, we built a models one to predict renting prices based on neighborhood and other amenities. For the following experiments, we used Mean Absolute Error, Mean Squared Error and R2 Score to evaluate each model. Mean Absolute Error (MAE) is a measure of the average absolute difference between the actual rental prices and the predicted rental prices. Mean Squared Error (MSE) is another measure of the accuracy of the predictions, and it is calculated as the average of the squared differences between the actual and predicted values. R2 Score is a measure of the proportion of the variance in the target variable (rental prices) that is explained by the independent variables (features) in the model.

The first model we tried was a Random Forest Classifier. The ensemble was over fitting to the data. Even with hyper tuning, we achieved an accuracy of only 20/

The next model we tried was a Linear Regression Model. In this case, the MAE of 1301.78 means that, on average, the predicted rental prices differ from the actual rental prices by about 1301.78. In this case, the MSE of 9913106.42 means that the average squared difference between the actual and predicted rental prices is about

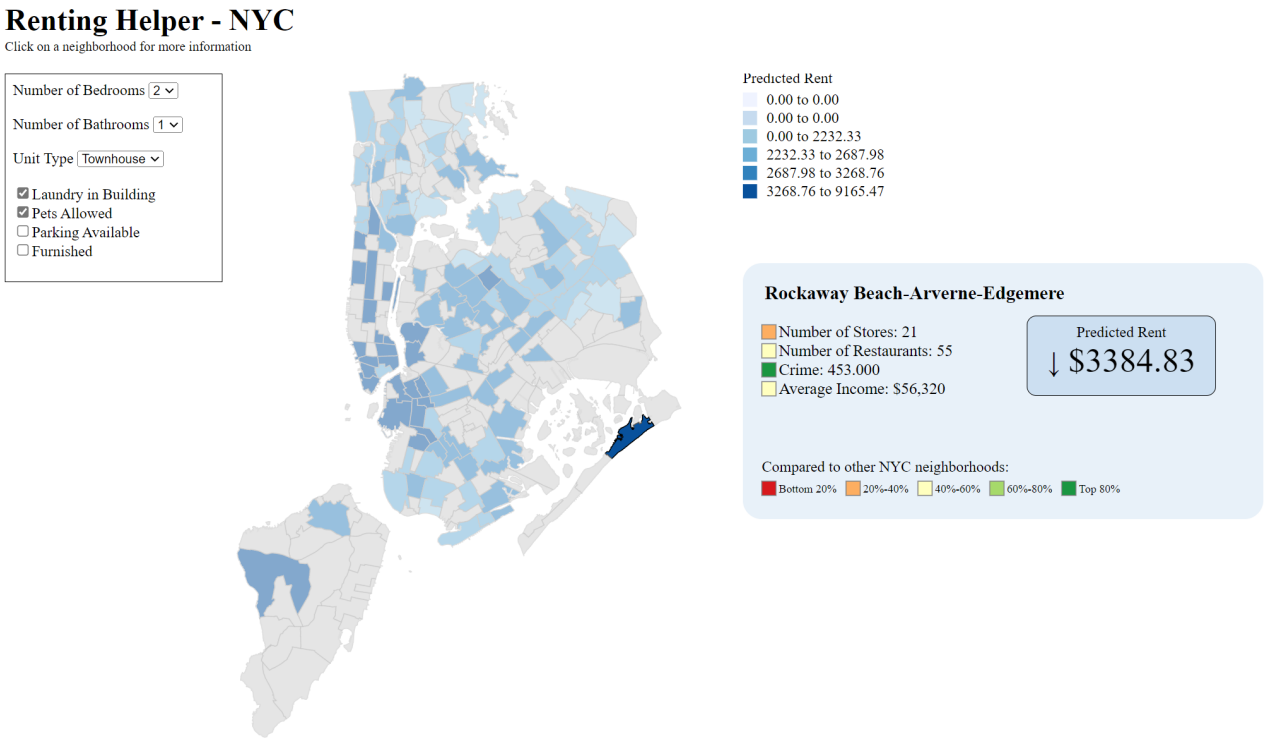

Figure 3: Rental Helper UI and Example Case

9,913,106.42.In this case, the R2 score of 0.4008 means that about 40 percent of the variation in the rental prices can be explained by the features used in the model. Overall, these performance metrics indicate that the linear regression model may not be the best fit for this particular problem, as the MAE and MSE are relatively high and the R2 score is relatively low.

The Gradient Boost model significantly increased the R2 score to .72 but the XGBoost Model after hyper tuning brought the R2 score all the way up to .84. As this was the best score, shown in Figure 2, we chose to implement the price prediction model as of 2016 with the XGBoost model.

While testing out each model, we utilized cross-validation using k-fold cross validation. This allowed us to break up our data to test how our data is performing and to ensure it is not overfitting or underfitting.

For the renting helper tool, we also used multiple time series ARIMA models to correct the predictions to 2023. We used one neighborhood to perform hyper parameter tuning. This p parameter represents the number of autoregressive terms and refers to the number of lagged observations included in the model. A high value of p means that the model is using a lot of past values to make predictions. The d parameter represents the number of times the time series needs to be differenced to make it stationary. Stationarity means that the mean and variance of the time series remain constant over time. The q parameter represents the number of moving average terms and refers to the number of lagged forecast errors included in the model. A high value of q means that the model is using a lot of past errors to make predictions. Ultimately, the best estimated parameters were: p as 3, d as 0 , and q as 1.

Overall, by conducting a variety of experiments and validations, we can ensure that our renting helper tool and models are accurate, effective, and user-friendly. This will provide valuable insights and assistance to renters and real estate professionals in making informed decisions about renting and leasing properties.

VII. Conclusion

Figure 3 illustrates the end product of Renting Helper tool. User inputs are located in the top left corner, where various rental property features can be selected. As users adjust their preferences, the colors on the map change to represent the predicted rent for different neighborhoods based on the chosen features. The color scale is displayed in the top right corner for easy reference.

By clicking on a specific neighborhood, users can access detailed information about the area, such as the number of stores and restaurants, crime rates, and average income. This data is presented in a blue window situated in the bottom right corner of the UI. Additionally, an arrow predicts whether the rent is expected to rise or fall in the upcoming month.

Our rental price prediction model has achieved an R-squared value of 0.84 and a mean absolute error (MAE) of 632, indicating that the predictions are relatively re-

liable. In future developments, we plan to also incorporate actual listing information into the tool and expand its coverage to include other major cities in North America.

VIII. Team Statement

All team members have put in a similar amount of effort.

References

[1] McCord, M., Davis, P.T., Haran, M., McIlhatton, D. and McCord, J. (2014), "Understanding rental prices in the UK: a comparative application of spatial modelling approaches", International Journal of Housing Markets and Analysis, Vol. 7 No. 1, pp. 98-128. https://doi.org/10.1108/IJHMA-09-2012-0043

[2] Cellmer, R., Trojanek, R. (2019). Towards Increasing Residential Market Transparency: Mapping Local Housing Prices and Dynamics. ISPRS International Journal of Geo-Information, 9(1), 2. https://doi.org/10.3390/ijgi9010002

[3] Bency, Archith and Rallapalli, Swati and Ganti, Raghu and Srivatsa, Mudhakar and Manjunath, B.. (2017). Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery. 320-329. 10.1109/WACV.2017.42.

[4] C. LEE, "Forcasting Spatially Correlated Targets: Simultaneous Prediction of Housing Market Activity Across Multiple Areas", International Journal of Strategic Property Management, Gangwon-do, Korea, Nov. 2021.

[5] L. Rampini and F. R. Cecconi. "Artificial Intelligence Algorithms to Predict Italian Real Estate Market Prices", Journal of Property Investment Finance. Milan, Italy, Aug. 2021.

[6] R.-T. Mora-Garcia , M.-F. Cespedes-Lopez and V. R. Perez-Sanchez. "Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times", MDPI. San Vicente del Raspeig, Spain, Nov. 2022.

[7] Alharthi, M., Aldakhil, A.M., Alghamdi, R., Al-Baiz, H. (2021). A new approach for predicting housing rents using machine learning: Evidence from the Riyadh market. International Journal of Housing Markets and Analysis, 14(1), 114-136.

[8] Witten, I. H., Frank, E., Hall, M. A. (2016). Data Mining: Practical Machine Learning Tools and Techniques (4th ed.). Morgan Kaufmann.

[9] Gelman, I.A., Wu, N. (2011). Combining Structured and Unstructured Information Sources for a Study of Data Quality: A Case Study of Zillow.Com. In Proceedings of the 44th Hawaii International Conference on System Sciences (HICSS).

[10] City of New York. (2020). 2020 Neighborhood Tabulation Areas (NTAs) Tabular Data. Retrieved April 9, 2023, from https://data.cityofnewyork.us/City-Government/2020-Neighborhood-Tabulation-Areas-NTAs-Tabular/9nt8-h7nd/data

[11] Purcell, Brendan. (2019). StreetEasy rental and sales data for New York City [streeteasy_db]. GitHub.

[12] SOI Tax Stats - Individual Income Tax Statistics - 2020 ZIP Code Data. (2020). https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2020-zip-code-data-soi

[13] New York City Police Department. (2022). Historical New York City Crime Data. https://www.nyc.gov/site/nypd/stats/crime-statistics/historical.page