Northeastern University

D'Amore-McKim School of Business

Project: Final Report

MISM6212

Data Mining and Machine Learning for Business

NYC Property Sales

Professor: Xiaoping Liu

Background Information of data:

The world of real estate is a dynamic landscape, where property values fluctuate, neighborhoods evolve, and transactions occur at a rapid pace. To gain a basic understanding of these market dynamics and uncover valuable insights, I chose a comprehensive real estate sales dataset for New York City. This dataset comprises a rich array of information, including property details, geographic identifiers, transaction specifics, and other key indicators. I hope I can draw insightful conclusions that can be helpful for real estate professionals and investors.

In this dataset, which I obtained from Kaggle (NYC Property Sales (kaggle.com)), columns such as BOROUGH, NEIGHBORHOOD, TAX CLASS, PROPERTY SIZE, SALE PRICE, and DATE, among others, serve in navigating the details of the real estate market. By examining the BOROUGH and NEIGHBORHOOD columns, we can assess the spatial distribution of real estate activity, identifying which areas are experiencing heightened demand and which may present untapped opportunities. The dataset of NYC Property Sales contains information about real estate transactions in New York City for the financial year 2016-2017. Government agencies, such as the NYC Department of Finance or similar organizations, typically collect and maintain this dataset. This dataset contains the location, address, type, sale price, and sale date of building units sold.

Through the SALE PRICE column and time-based analysis using SALE DATE, we can track how property prices have evolved over time and whether certain periods exhibit price surges or declines. The TAX Class columns allow us to understand the tax implications on sales, potentially revealing trends in property investments and ownership changes.

The analysis of this real estate sales dataset promises to provide a comprehensive view of the market and offers actionable insights that can guide investment decisions, policy formulation, and market strategies.

NYC Department of Finance: This is a government agency responsible for collecting property-related taxes and maintaining property records in New York City. They collect data on property sales as part of their tax assessment and collection process.

Real Estate Industry: The real estate industry in New York City is a critical part of the economy. It includes real estate developers, brokers, property management companies, and investors. Access to accurate and comprehensive property sales data is crucial for various stakeholders in this industry to make informed decisions.

Objectives:

The objective of the project is to present a Sale Price variation and time-based analysis using Sale Dates or years, and how property prices have evolved over time. Also, I could identify which types of properties are in high demand or have higher sales prices during the year by plotting a box plot of Borough versus sales price. Columns like RESIDENTIAL UNITS, COMMERCIAL UNITS, LAND SQUARE FEET, GROSS SQUARE FEET, and YEAR BUILT could offer insights into property characteristics.

Data Description:

Name of the Attribute	Data type	Descriptions of the Attribute
BOROUGH		The name of the borough in which the property is located. Manhattan (1), Bronx (2), Brooklyn (3), Queens (4), and Staten Island (5).
NEIGHBORHOOD	Object	Department of Finance assessors determine the neighborhood name while valuing properties

BUILDING CLASS CATEGORY	object	This is a field that we are including so that users of the Rolling Sales Files can easily identify similar properties by broad usage (e.g., One Family Homes) without looking up individual Building Classes.
TAX CLASS AT PRESENT	Object	Every property is assigned to one of four tax classes (Classes 1, 2, 3, and 4)
BLOCK	int64	sub-division of the borough on which real properties are located
LOT	int64	A Lot is a subdivision of a Tax Block and represents the property's unique location
EASE-MENT	Object	An easement is a right, such as a right of way, which allows an entity to make limited use of another’s real property
BUILDING CLASS AT PRESENT	Object	The first position of the Building Class is a letter that is used to describe a general class of properties that signifies office buildings. The second position, a number, adds more specific information about the

		property’s use or construction style.
ADDRESS	Object	The street address of the property as listed on the Sales File
APARTMENT NUMBER	Object	Apartment Number allocated
ZIP CODE	int64	The property’s postal code
RESIDENTIAL UNITS	int64	The number of residential units at the listed property
COMMERCIAL UNITS	int64	The number of commercial units at the listed property.
TOTAL UNITS	int64	The total number of units at the listed property.
LAND SQUARE FEET	Object	The land area of the property is listed in square feet.
GROSS SQUARE FEET	Object	The total area of all the floors of a building as measured from the exterior surfaces of the outside walls of the building,
YEAR BUILT	int64	Year the structure on the property was built.
TAX CLASS AT THE TIME OF SALE	int64	tax classes at the time of sales
BUILDING CLASS AT THE TIME OF SALE	Object	Building classes at the time of sales
SALE PRICE	Object	Price paid for the property.
SALE DATE	Object	Date the property sold.

Data Mining Process:

I have refined a dataset that originally contained 20 columns and nearly 84,000 rows. To prepare it for a time series model, I took several steps. First, I noticed that the sale prices and dates showed limited variation but indicated a gradual increase, particularly between 2016 and 2017. I addressed this by creating separate columns for the sale month and sale year.

Furthermore, I categorized the data into numeric and categorical variables to identify and manage empty values effectively. I replaced empty cells with NaN values and eliminated unnecessary columns, such as apartment numbers, which helped reduce the number of missing data points. I also removed duplicate entries to ensure a cleaner and more manageable dataset for subsequent algorithmic analysis.

Additionally, I filtered the data to retain only rows where the 'SALE PRICE' fell within the range of 10,000 to 10,000,000, and this filtered data was used to create a new data frame, saving the original dataset from modifications. To improve the interpretability of the data, I transformed numeric borough values into categorical names. These steps collectively prepared the dataset for more meaningful and effective data analysis. I also learned that this dataset has Sales price, square feet, and borough type as the most useful predictors.

Project Methods and Improvements:

Leveraging advanced machine learning on the NYC Property Sales dataset can provide valuable insights and improvements in various aspects of the real estate industry and government operations. These methods can lead to more informed decision-making, reduced risks, and increased efficiency in business operations.
Methods used are ***Clustering*** like ***k-means*** can group square feet used of properties with similar characteristics, assisting in market segmentation and targeted marketing strategies. Similarly, BOROUGH or NEIGHBORHOOD columns could extract insights into the geographic distribution of properties and sales prices. I specifically chose the clustering model because I wanted to identify outliers or anomalies in real estate data. These anomalies could indicate potential issues or opportunities, such as overpriced or undervalued properties. According to the

data, I was able to find the five different clusters as per the sales price versus gross square feet. I also found outlier points in the data points. It can be inferred from the plot that most of the gross square feet are located highest in the range of 6000 square feet to sale price with a range of $400000. We could analyze which boroughs or neighborhoods have the highest or lowest sales activity during the year by constructing the ***box plot*** variation.

The ***time series model*** could predict property price trends over time of the year, helping investors and developers make informed decisions about when to buy or sell properties. I attempted to enhance the model's ability to capture the temporal variation in the data by adjusting the dataset.

(K-NN) algorithm for regression is to predict real estate prices based on independent variables such as the year built, number of residential units, and number of commercial units. That also aimed to find the optimal number of neighbors (k) for the K-NN model by testing different values of K and selecting the one that yields the highest R-squared (R2) score. The R-squared score of the final model indicates how well the model fits the data and its ability to explain the variance in sale prices. A higher R-squared score suggests a better fit and predictive performance.

Managerial Implications and Business Improvement Aspects:

The various data mining techniques allow for more accurate customer segmentation and targeting. This reduces marketing spend while increasing the return on investment (ROI). Also, these can help real estate professionals identify emerging trends and opportunities in the market. This can guide investment decisions, such as when and where to buy or sell properties, potentially increasing sales revenue. While cost savings and increased sales are primary metrics for business improvement, other benefits include operational streamlining, enhanced customer experience, competitive advantage, and informed decision-making.

By clustering properties and understanding the characteristics of each cluster, you can create customer profiles. This helps in matching customers with properties that suit their preferences, improving customer satisfaction and potentially driving sales.

The performance of the K-Nearest Neighbors (K-NN) regression model on a real estate dataset was to find the best number of neighbors (k) for predicting property sale prices based on the independent variables. The result can be inferred that the optimal number of neighbors (k) for the K-NN regression model provided the best predictive performance for estimating property sale prices based on the given independent variables. K-NN also provides a structured approach to decision-making in real estate. It helps in property selection, pricing, and investment strategies by considering the surrounding properties.

Time series models can be used to predict future property price trends, which is valuable for both buyers and sellers in making investment decisions. Time series analysis can also reveal seasonal variations in real estate demand, helping real estate agents and investors plan their marketing and listing strategies. Time series data can provide insights into historical sales trends, allowing real estate professionals to adapt to changing market conditions.

Recommendations:

The recommendations can be for real estate agencies or organizations to apply K-NN to match buyers with properties that closely align with their preferences. providing personalized property recommendations and improving the overall customer experience would be beneficial, also using predictive models to assess the potential risk associated with real estate investments.

I found, as per the data, that the K-NN model performed well, although it could perform better. Similarly, for the time series model, I thought it was interesting to check how real estate demand evolved over time. In a business context, it is important to consider what the acceptable margin of error is for a company like a real estate agency. Making informed decisions about where and when to invest based on market forecasts and

using predictive models to conduct scenario analysis to assess the potential impact of different market conditions or economic factors on real estate investments can be the ultimate use of these algorithms.

In the same way, clustering can help real estate investors manage their portfolios by grouping properties with similar risk profiles or potential for growth. Clustering can guide investors and developers to identify emerging real estate markets with similar growth potential. Clustering can segment the real estate market into categories based on property types, neighborhood characteristics, and price ranges, enabling tailored marketing strategies.