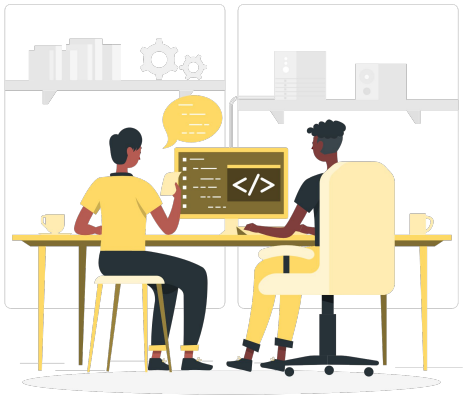
NYC Property Price Analysis & Prediction

by Godfried Junio Matahelemual

EXECUTIVE SUMMARY

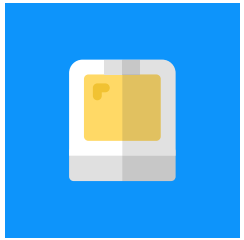
- The property sale prices shows an uptrend in general with a dip from the COVID-19 impact, and Manhattan remains the priciest market, although with fewer sales.
- The Random Forest Regressor (Default parameter) stands out as the most effective model for this particular prediction task, striking a balance between accuracy and the ability to generalize to new data.
- The physical size of the property (gross square feet) is the predominant factor influencing the model's predictions in the NYC real estate market.



Background

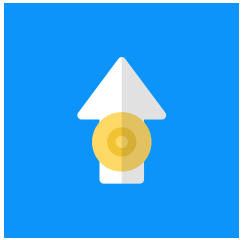
Early 20th Century	• Housing boom and immigration • Diversity of buildings
Late 21st Century	• Suburbanization • Economic recession
Beginning of the 21st Century	• Housing market paradigm shift • Affordable housing
Early 20th Century	• COVID-19

Problem Formulation



ML OBJECTIVE

Develop a predictive model to estimate property prices in NYC



ACTION

Use model predictions to guide clients in buying, selling, or investing in properties.

VALUES

Enhance the decision-making process for stakeholders through accurate predictions and strengthen trust in the real estate market with transparent and data-driven price estimations

Data

Category	Feature	Description
Property	borough	Borough name in NYC
	block	Block parcel number to identify location of buildings/property (for tax payment)
	lot	Lot parcel number to identify location of buildings/property (for tax payment)
	building_class_category	Category of building class
	gross_square_feet	Gross area of the property in square feet
	year_built	Year when the property is built
Infrastructure	subway_count_in_1km	Number of subway stops in 1 km radius where the property is located
	schools_count_in_1km	Number of school in 1 km radius where the property is located
	bus_count_in_1km	Number of bus_stops in 1 km radius where the property is located
	health_facil_count_in_1km	Number of health_facility in 1 km radiuse where the property is located
	parks_zones_count_in_1km	Number of parks_zones in 1 km radius where the property is located
	parks_properties_count_in_1km	Number of parks_properties in 1 km radius where the property is located

The dataset are collected from various sources. The complete metadata can be accessed here.

ML Development Process

01
Data Preparation
Combine, filter, cleanse, merge the data and buffer analysis, preparing it for further steps

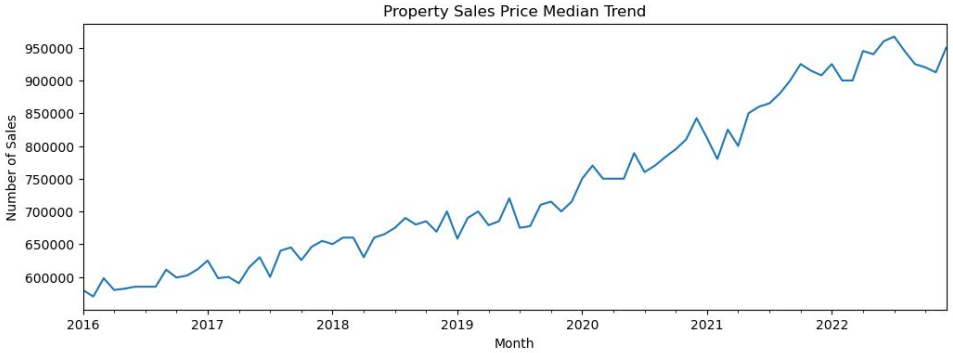
02
Data Inspection
Understand the data and determine data processing strategies for the data to be sufficient for modeling or further analysis

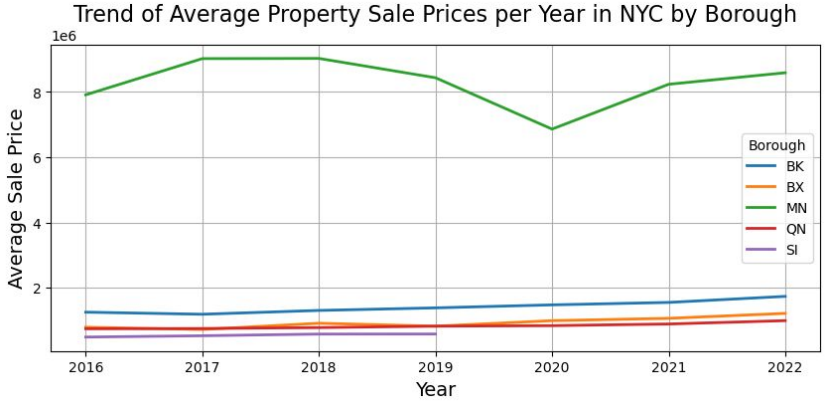
03
Exploratory Data Analysis
Dive deeper on the data and generate business insights

04
Linear Regression Analysis
Try the most simple model for prediction and find out what factors affecting the housing price analysis

05
ML Modeling
Try the most simple model for prediction and find out what factors affecting the housing price analysis

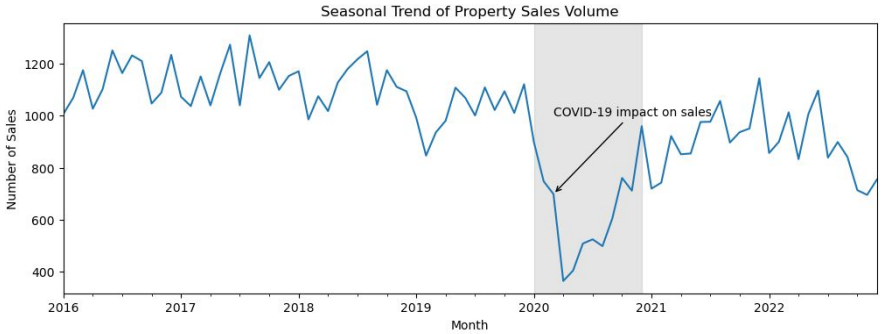
Key Findings: Sustained Growth in Property Values with Manhattan as the Market Lead

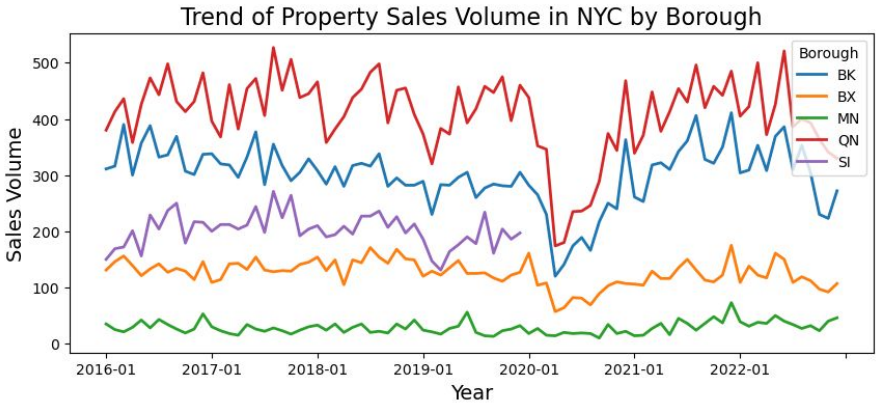




- Growth in Property Values: The median and average sales price of properties has seen a upward trend from 2016 through 2022, indicating a consistent increase in the value of the median segment of the market.
- Manhattan Leads Diverse Borough Pricing: Manhattan maintaining a considerably higher average sale price compared to other boroughs, reiterating its position as a premium real estate market. The other boroughs (BK, BX, QN, SI) show more stable and closely grouped price trends, indicating similar market behaviors and potential for those seeking more stable investment options.

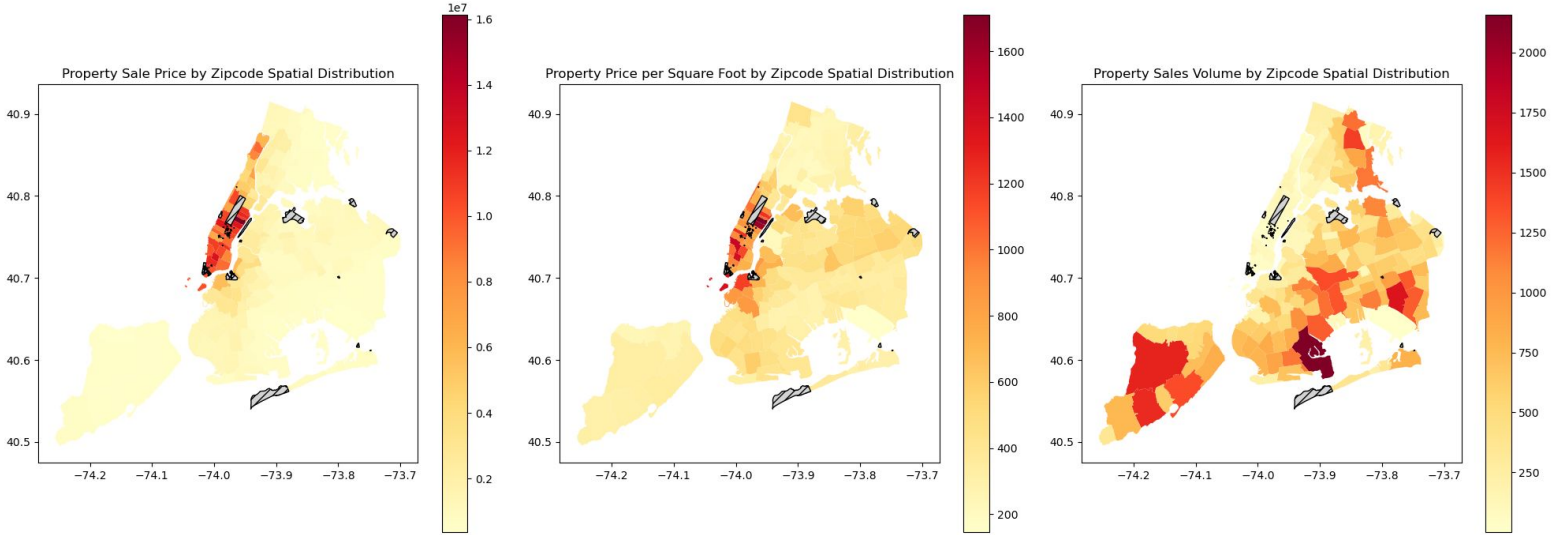
Key Findings: Pandemic Effects on Sales, Seasonal Patterns, and Manhattan's High-Price, Low-Turnover Dynamic





- Pandemic Disrupts Sales Volume Trend: The property sales volume were steady before they fell sharply in 2020 because of COVID-19. Post-pandemic recovery is observable, yet the market has not returned to pre-pandemic levels, indicating a lasting effect on the sales volume.
- Seasonality is exist: The sales volume highlights a declining long-term trend (due to COVID-19 with slow recovery). Seasonality is evident in sales, displaying regular patterns of peaks and troughs each year.
- High Prices, Low Sales: Manhattan Trails in Turnover Despite Leading in Cost, as Brooklyn and Queens Dominate Market Activity in sales over time.

Key Findings: Manhattan apartments come at a high cost but with fewer sales transactions



- Market Concentration: The dense clustering of high-value sales in Manhattan indicates a strong market concentration.
- Small Spaces, Big Value: Even smaller apartments in Manhattan are very pricey. This suggests that people might be willing to pay a lot for a prestigious address.
- Manhattan's Low Turnover Despite High Prices: Manhattan, even though it's the most expensive, has the fewest sales, which may mean there aren't many houses being bought or sold there.

Modeling: the Predictors

Category	Feature	Description
Property	building_class_category	Category of building class
	gross_square_feet	Gross area of the property in square feet
	year_built	Year when the property is built
Infrastructure	subway_count_in_1km	Number of subway stops in 1 km radius where the property is located
	schools_count_in_1km	Number of school in 1 km radius where the property is located
	bus_count_in_1km	Number of bus_stops in 1 km radius where the property is located
	health_facil_count_in_1km	Number of health_facility in 1 km radius where the property is located
	parks_zones_count_in_1km	Number of parks_zones in 1 km radius where the property is located
	parks_properties_count_in_1km	Number of parks_properties in 1 km radius where the property is located

- 9 features are used as predictors. In general there are 3 property related features and 6 infrastructure related features.

Modeling: Parametric Modeling

	Method	MAE	MAPE (%)	MSE	RMSE
0	Benchmark	657249.728474	136.700724	4.985636e+12	2.232854e+06
1	Standard Scaled Features	657249.728474	136.700724	4.985636e+12	2.232854e+06
2	Log Scaled Features	776457.095851	160.456488	5.552706e+12	2.356418e+06
3	Selected Features	650473.406486	134.842520	5.046779e+12	2.246504e+06
4	Ridge	665485.459255	138.147600	5.023297e+12	2.241271e+06
5	Lasso	644715.076848	130.750003	5.577787e+12	2.361734e+06

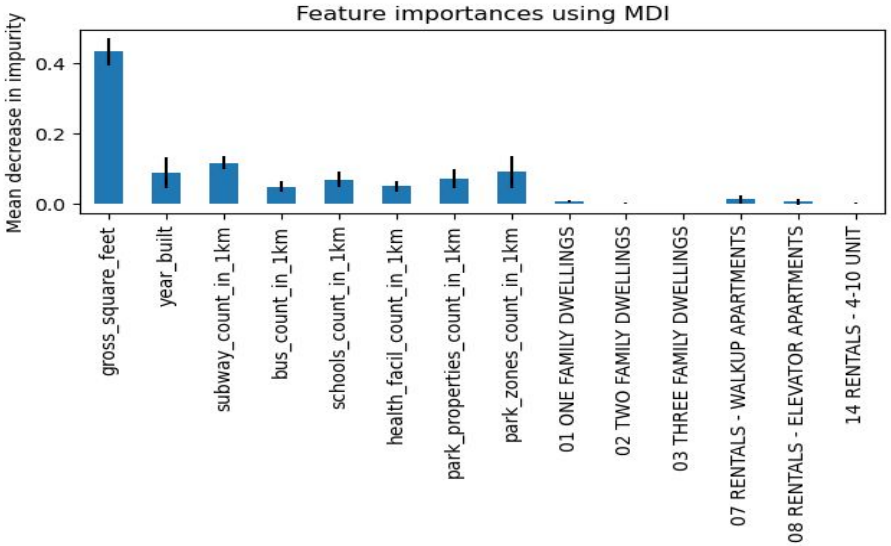
- Initially, the most simple algorithm are used for the modeling. But the result shows that the linear model appears to underperform. The high MAPE (greater than 100%) is often a sign that the model may not be suitable for the dataset or that the data requires transformation or cleaning before modeling. Therefore, machine learning models will be explored in the subsequent notebook.

Modeling: Machine Learning Modeling

	Train R2 Score	Valid R2 Score	Train RMSE	Valid RMSE	Train MAPE	Valid MAPE	Train MAE	Valid MAE
LinearRegression	0.407670	0.427538	2.232854e+06	2.016524e+06	1.367007	1.240148	657249.728474	646509.816885
DecisionTreeRegressor	0.983526	0.328640	3.723696e+05	2.183776e+06	0.105563	0.829772	40712.454319	562571.020324
RandomForestRegressor	0.926049	0.603168	7.889520e+05	1.678934e+06	0.354704	0.777577	182756.647582	435410.132840
XGBRegressor	0.904018	0.555254	8.988233e+05	1.777404e+06	0.884742	0.912936	364246.405283	474335.293125

- RandomForestRegressor offers better generalization than the Decision Tree model, with a training R2 score of around 0.96 and a validation R2 score of about 0.60. The RMSE, MAPE, and MAE values for the validation set are lower compared to the Decision Tree, suggesting improved prediction accuracy. However, the difference between training and validation scores still indicates a tendency towards overfitting.
- After the feature selection and hyperparameter tuning, the default Random Forest model still appears to be the best model.

Modeling: Feature Importance



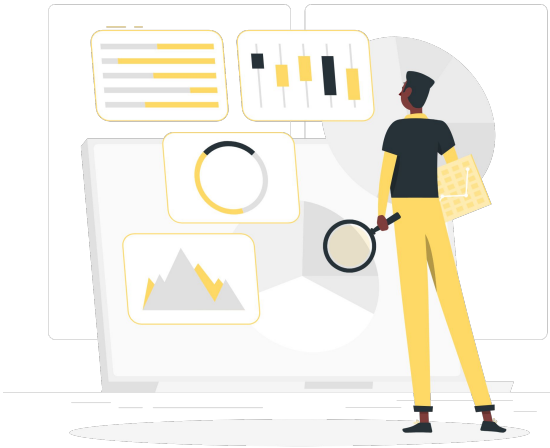
- The most significant predictor in the model. Is 'gross_square_feet' followed by 'subway_count_in_1km', 'park_zones_count_in_1km', 'year_built', 'schools_count_in_1km',
- Different types of dwellings like one, two, and three-family homes, as well as rental apartments with and without elevators, are represented with varying degrees of importance. While these factors are considered by the model, their impact is considerably less than that of gross square feet.

FUTURE IMPROVEMENT

- Remodeling with new strategies to overcome the overfit
- Model log using MLFLow
- Deploy an interface using streamlit

CONCLUSIONS

- The property sale prices shows an uptrend in general with a dip from the COVID-19 impact, and Manhattan remains the priciest market, although with fewer sales.
- The Random Forest Regressor (Default parameter) stands out as the most effective model for this particular prediction task, striking a balance between accuracy and the ability to generalize to new data.
- The physical size of the property (gross square feet) is the predominant factor influencing the model's predictions in the NYC real estate market.



“ There is no greater risk than being blind to the unknown ”

Thanks!