National University of Singapore

# FT5005 Machine Learning for Finance

## Final Project Report I

March 2025

Group 6

| | |
|---|---|
| ALYA RIZKY PRASTITI | A0275710H |
| LIN XIAO | A0296234Y |
| WANG RUIJUE | A0211256R |
| XU RUI | A0296229R |
| ZOU ZEREN | A0297860M |

## Section I: Overview

This report focuses on forecasting revenue and EBITDA for real estate firms operating in the New York City market. To enhance our predictive capabilities, we integrate an additional dataset—the NYC Property Sales dataset—which provides detailed records of every property sold in NYC since 2003, updated monthly. This dataset includes key attributes such as sale price, property type, and location, offering a comprehensive view of market trends and valuation changes that are critical for assessing the performance of real estate companies.

Property sales data capture market sentiment, supply-demand imbalances, and asset value changes, serving as leading indicators for revenue and EBITDA. Academic research supports this; the *Journal of Real Estate Finance and Economics*[1] reported a 12% improvement in REIT earnings forecasts using granular property sales data over macroeconomic factors alone. Industry analyses by Empire State Realty Trust[2] similarly show residential price growth correlating with developer EBITDA margins, while commercial sales volume forecasts leasing revenues six months ahead. These findings align with broader financial economics literature emphasizing the predictive power of high-frequency, asset-specific data in reducing information asymmetry and enhancing forecasting precision.

### Feature Engineering Framework and the Evidence

Our feature engineering framework leverages publicly available data and industry-standard practices to enhance the predictive accuracy of our revenue and EBITDA forecasting models. The strategies we propose are grounded in evidence that is easily accessible online:

1. **Land and Improvement Value Separation**: Distinguishing between land and structural components of property sales enables a more accurate assessment of asset values. This approach aligns with methodologies that separate land and improvement values to enhance prediction accuracy.
2. **Location-Based(Geographical) Aggregation**: Aggregating sales data by borough and neighborhood facilitates the identification of location-specific trends and demand patterns, such as distance from Central Business District (CBD) or the influence of LRT, subways and other public transportations. Moreover, Haider and Miller (2000) investigated some cases in the United States and Australia and discovered that the house prices tend to drop sharply as the distance of a city's CBD's increases[3].
3. **Temporal Analysis**: Incorporating temporal features such as sale date and year built allows for the assessment of market cycles and the impact of property age on sales dynamics. Understanding these temporal factors aids in forecasting future

revenue streams. Ferlan, Bastic and Psunder (2017) revealed that the age of buildings is negatively correlated with the value of residential properties[4]. This indicates that as buildings get older, their price tend to drop.

4. **Property Characteristics Profiling**: Analyzing attributes like building class category, residential units, and commercial units provides insights into the types of properties driving sales and their alignment with market demand. As stated by the New York State Association for REALTORS (NYSAR), the properties are categorized as 2 bedrooms or fewer, 3 bedrooms, and 4 bedrooms or more[5].

In addition, the Annual Report on the New York State Market for residential real estate mentioned that the listing number of homes available for sale was lower by 2.6 percent comparing 2024 to the prior year. Moreover, home prices in 2024 were up with the overall median sales price increasing 7.8 percent compared to 2023.

## Section II: Existing Literature

### 2.1 Introduction

This section synthesizes theoretical and empirical literature to:

(1) Justify the selected features and features from engineering (X) based on their documented impact on target (Y).

(2) Identify methodologies from prior studies on similar prediction tasks.

Given the scarcity of studies directly addressing revenue and EBITDA prediction for real estate firms, where extant literature remains predominantly focused on property prices, our review systematically extended to cross-industry research examining financial performance prediction, including revenue, operating profit, net profit, ROA, and ROE.

The structural and functional linkages between financial metrics justify utilizing those literature for our revenue/EBITDA prediction. Revenue and EBITDA serve as foundational inputs for downstream profitability metrics in the income statement. Identical operational (e.g., asset turnover), financial (e.g., leverage), and macroeconomic drivers influencing later-stage variables inherently influence upstream dynamics through additive/multiplicative relationships. For instance, interest rate sensitivity affects both EBITDA (via financing costs) and ROE (via leverage amplification). Therefore, it is methodologically sound to utilize academic literature with analogous predictive objectives as reference

### 2.2 Theoretical Foundations: Feature-Target Relationships

Existing literature on corporate financial performance prediction predominantly employs four categories of explanatory variables: (1) fundamental accounting metrics from financial statements, (2) classic financial ratios, (3) macroeconomic indicators, and (4) property sales data.

**2.2.1 Fundamental Accounting Metrics**. Financial statements directly quantify a firm's operational efficiency and resource allocation through revenue recognition and expense matching, inherently reflecting key performance drivers. Although forecasting based solely on historical revenue or EBITDA represents a traditional time series method, it often overlooks critical external factors. Therefore, this basic approach serves as a baseline to evaluate more comprehensive multivariate models. Most studies incorporate multiple accounting measures, aligning with financial theory's emphasis on the interconnectedness of financial statements. Lee et al. (2017) used eleven financial indicators—including profit growth, operating expenses, shareholder’s equity, and total assets—to successfully predict revenue, operating profit, and net profit for 22 U.S. biopharmaceutical firms[6].

**2.2.2 Financial Ratios**. Financial theory suggests that ratios simplify complex financial statements into standardized metrics, providing advantages like improved readability, comparability across different-sized firms, stability, and early risk detection. Ratios effectively capture liquidity, profitability, leverage, and operational efficiency—critical factors influencing revenue generation and earnings capacity. Shen and Tzeng (2014) used 25 financial ratios, covering dimensions such as Capital Sufficiency, Asset Quality, Earnings, Liquidity, Interest Rate Sensitivity, and Growth, in a DRSA-based Neuro-Fuzzy model predicting ROA in Taiwanese banks[7]. Key predictive ratios included Capital Adequacy (Net Worth to Assets), Profitability (e.g., NIBT to Assets), Liquidity (Liquidity Ratio, Loan-to-Deposit Ratio), and Growth (Deposit, Investment, Guarantee Growth Rates). Thus, our feature engineering primarily involves computing these financial ratios as model inputs.

**2.2.3 Macroeconomic Indicators**. Macroeconomic variables significantly influence real estate companies' revenues through channels such as demand dynamics, financing conditions, and asset valuation effects. Factors like GDP growth, interest rates, inflation, unemployment, and monetary or fiscal policies shape market fundamentals and corporate earnings. Trang et al. (2022) found that money supply strongly impacts real estate profitability during crises like the COVID-19 pandemic, while GDP growth partially influences performance[8]. However, macroeconomic factors don't always enhance revenue predictions. Lam (2004) noted that incorporating macroeconomic variables into neural network models did not significantly improve forecasting accuracy[9]. Given these inconclusive findings and pandemic-related disruptions, we conservatively limit macroeconomic inputs in our models.

**2.2.4 Property sales data**. As noted in Section I, property sales constitute one of the primary revenue streams for real estate enterprises. Consequently, property sales data can serve as an indirect indicator of a real estate company's financial performance. And this relationship has been well-documented in the existing literature, thus warranting no further elaboration here.

The specific variables selected as model inputs for this project are enumerated in the Appendix.

### 2.3 Empirical Studies Results in Financial Performance Prediction

Table 1 summarizes key findings from two academic studies: *"A Contrastive Study of Machine Learning on Energy Firm Value Prediction"* by Zhang et al. (2020)[10] and *"Composite financial performance index prediction – A neural networks approach"* by Popa et al. (2021)[11].

Zhang et al. (2020) compared six machine learning methods—K-Nearest Neighbors (KNN), Decision Trees (DT), Support Vector Regression (SVR), Artificial Neural Networks (ANN), AdaBoost, and Random Forest (RF)—for predicting energy firm values, focusing on oil and power sectors. The best-performing method, performance metric (MAPE), and key predictive features based on empirical importance or theoretical relevance are highlighted.

Popa et al. (2021) proposed a neural network-based approach utilizing Principal Component Analysis (PCA) to create a composite financial performance index. Their study combined traditional accounting and value-added indicators for Romanian-listed companies (2011–2018), testing both feed-forward and recurrent neural networks to evaluate prediction accuracy based on historical financial data.

**Table 1.** Key Finding of Two Academic Papers

| **Aspects** | **Zhang et al. (2020) – "*A Contrastive Study of Machine Learning on Energy Firm Value Prediction"* paper** | **Popa et al. (2021) – Details from "*Composite financial performance index prediction – A neural networks approach"* paper** |
|---|---|---|
| **Research context** | Predicting energy firm value specifically in Oil and Power industries. | Predicting a composite financial performance index for Romanian listed companies. |
| **Methodology** | Compared multiple ML methods: ANN, RF, SVR, KNN, Decision | Developed and compared Feed-forward NN and RNN (LSTM) |

| | | |
|---|---|---|
| | Trees, and AdaBoost. | methods. Utilized Principal Common Analysis (PCA) to construct a composite index. |
| **Best method** | Artificial Neural Network (ANN) | Recurrent Neural Network (RNN), specifically LongShort-Term Memory (LSTM) Sequence-to-Sequence model. |
| **Performance metric and values** | - Oil: ANN MAPE* = 5.15%.<br>- Power: ANN MAPE = 8.64%. | RNN (LSTM) achieved MSE** (0.04 train, 0.09 test), MAE*** (0.15 train, 0.19 test). |
| **Important features** | Firm size, Asset turnover, EBIT, Net profit margin, ROA, ROE, Debt-to-asset ratio, Cash-to-debt, Firm growth, Acquisition type, Share % acquired. | Cash Value Added (CVA), Market Value Added (MVA), ROA, EPS, Economic Value Added (EVA), Solvency (SOL), Cash Flow Return on Investment (CFROI), ROE (identified through PCA). |

* Mean Absolute Percentage Error

** Mean Squared Error

*** Mean Absolute Error

Zhang et al. (2020) clearly demonstrated ANN's predictive effectiveness within homogeneous sectors, making it potentially suitable for real estate firms, which often share similar business models and financial indicators like asset turnover, debt levels, and profitability. However, the original model's specificity to energy firms necessitates adjustments to include real estate-specific variables such as occupancy rates, rental yields, and property valuations. Popa et al. (2021) illustrated the advantage of combining PCA and LSTM to capture historical trends and economic cycles, valuable for real estate forecasting. Still, PCA's interpretative complexity and limited sample scope require careful consideration of local market conditions and sector-specific metrics.

## Section III: Data Sources and Preprocessing

Our primary dataset is *Compustat Daily Updates - Fundamentals Quarterly.csv*, containing essential financial data. During preprocessing, we focus exclusively on real estate companies identified through Standard Industrial Classification (SIC) codes. Additionally, we restrict the dataset to records from 2010 onward, as earlier data may carry biases due to lasting impacts from the 2008 financial crisis. The training dataset covers the period from 2010 to 2020, while data from 2021–2024 is reserved for model testing.

For the target variable (Y), we focus on quarterly revenue and EBITDA. Since EBITDA is not directly available, we reconstruct it using the formula:

$$\text{EBITDA} = \text{Operating Income} + \text{Depreciation and Amortization}$$

We exclude interest expense due to its unavailability in our dataset.

Feature selection for model inputs (X) is based on domain knowledge, where we include fundamental accounting variables and construct financial ratios relevant to real estate prediction. Details on selected features are outlined in Table 2 in Appendix.

### Feature Construction Considerations

For certain features, we use a structured approach to improve consistency and interpretability. Instead of applying a simple moving average, we calculate average asset value by considering all previous quarters' total assets within the same fiscal year. This ensures that the NI_to_Asset feature accurately reflects a company's financial standing without being influenced by a rolling window that may span across different fiscal periods.

For the Relative Strength Index (RSI) calculation, we first determine quarterly gains and losses based on the closing price recorded on the last day of the quarter. We then apply the same fiscal-year-aligned averaging approach as used for total assets to compute average gains and losses. This methodology improves the consistency and comparability of financial indicators, making them more relevant for forecasting.

To prevent information leakage, we differentiate the feature pools for revenue and EBITDA predictions. Certain variables that directly influence one target variable are excluded when predicting the other. For example, operating income is used as a feature for revenue prediction but not for EBITDA prediction, as it is already embedded within the EBITDA calculation.

### Additional Data Sources

Beyond financial and accounting data, we incorporate information from *Compustat Daily Updates - Industry Specific Quarterly.csv*. The data collection methodology is implemented in the accompanying Jupyter notebook file for full transparency and reproducibility. Instead of using all available features, we selectively include variables directly related to real estate performance, such as room revenue, expenses, homebuilding, and construction activity. Some features in the dataset have ambiguous names, making their relevance unclear. To avoid introducing noise, irrelevant correlations, or overfitting, we discard such variables.

We explore text-based data sources, including Earnings Conference Call (ECC) transcripts and public news articles, which provide qualitative insights that enhance revenue and EBITDA prediction.

- ECC transcripts provide management insights into financial performance, growth strategies, and market conditions beyond numerical statements. Sentiment analysis evaluates executive confidence or uncertainty, while topic modeling identifies critical themes like cost-cutting or expansion plans. Incorporating linguistic features, tone analysis, and historical market reactions from ECC transcripts enhance forecasting accuracy.

- Public news articles capture real-time market sentiment, industry trends, and unforeseen events such as mergers, regulatory shifts, supply chain disruptions, and competitive pressures—all directly influencing revenue and EBITDA. Sentiment analysis assesses investor perceptions, while event-driven modeling quantifies financial impacts of key announcements. Integrating news data with financial metrics strengthens model robustness and responsiveness to market dynamics.

We also incorporated geographical data from the NYC Department of Finance, including sale price, borough, neighborhood, land and gross square footage, building class, and sale date, all essential for real estate forecasting. Recent trends indicate a notable decline in prices for single-unit, two-to-four-unit properties, and condominiums over the past five years, ending a decade-long appreciation trend. Despite this downturn, New York City's real estate market remains resilient, historically recovering and continuing to offer diverse opportunities across boroughs and property types.

### Data Merging and Integration

Companies in our dataset are uniquely identified by CIK, CUSIP, or Ticker codes. We merge all datasets using these identifiers along with the quarterly date rather than the exact release date, as financial statements are published on different dates across companies.

However, textual data integration poses challenges since companies in the ECC and news dataset are identified using a proprietary company_id without a clear reference to other identifiers. Due to this limitation, we currently use the ECC sentiment analysis as a market-wide positive/negative signal rather than linking it directly to specific companies.

For geographical data, while the dataset provides detailed location information, associating individual properties with real estate firms is complex. To maintain a macro-level perspective, we incorporate only the average home sale price per quarter rather than individual-level data. This ensures a high-level market trend representation without introducing excessive granularity.

After merging datasets, we thoroughly cleaned the data to ensure quality before modeling. Infinite values were replaced with NA to avoid computational issues. Missing data was addressed by forward-filling valid observations, dropping columns with over 50% missing values, and removing rows with more than 10 missing entries. Duplicate records were eliminated, reducing the revenue dataset from 2,688 rows and 29 columns to 2,593 rows and 20 columns, and streamlining the EBITDA dataset from 2,405 rows and 28 columns to 2,414 rows and 19 columns. Finally, all training features were normalized to ensure comparability and balanced model influence.

By applying these structured preprocessing steps, we ensure that our dataset is clean, relevant, and optimized for accurate and efficient predictive modeling. The summary statistics of processed x variables before scaling are tabulated in table 3 in the appendix.

## Conclusion

The correlation heatmap as shown in Figure 1 in the appendix reveals that most of the independent variables (X variables) in the dataset exhibit low correlation with each other, suggesting minimal multicollinearity and ensuring that the features provide unique information for predictive modeling. However, there are notable exceptions where strong correlations exist: operating income and net income (indicative of their direct financial relationship), equity and total assets (reflecting their interconnectedness in balance sheet metrics), and revenue growth versus net income growth (highlighting their shared dependence on business performance trends). These high correlations suggest that these variable pairs are closely related and may require careful consideration during model development to avoid redundancy or overemphasis. Overall, the dataset's feature set appears diverse and well-suited for robust predictive modeling, provided these highly correlated pairs are managed appropriately.

# Reference

[1] D. Geltner, A. Kumar, and A. M. Van De Minne, “Estimating Commercial Property Fundamentals from REIT data,” *J Real Estate Finan Econ*, Oct. 2023.

[2] “Empire State Realty Trust Earnings | Q4 2024 Results & Analysis | ESRT Financial News,” Panabee. Accessed: Apr. 04, 2025. [Online].

[3] M. Haider and E. J. Miller, “Effects of Transportation Infrastructure and Location on Residential Real Estate Values: Application of Spatial Autoregressive Techniques,” *Transportation Research Record*, vol. 1722, no. 1, pp. 1–8, Jan. 2000.

[4] N. Ferlan, M. Bastič, and I. Pšunder, “Influential Factors on the Market Value of Residential Properties,” *EE*, vol. 28, no. 2, pp. 135–144, Apr. 2017.

[5] “NYSAR_ANN_2024.pdf.” Accessed: Apr. 04, 2025. [Online].

[6] J. Lee, D. Jang, and S. Park, “Deep Learning-Based Corporate Performance Prediction Model Considering Technical Capability,” *Sustainability*, vol. 9, no. 6, Art. no. 6, Jun. 2017.

[7] K.-Y. Shen and G.-H. Tzeng, “DRSA-Based Neuro-Fuzzy Inference Systems for the Financial Performance Prediction of Commercial Banks,” *International Journal of Fuzzy Systems*, vol. 16, no. 2, 2014.

[8] L. N. T. Trang, D. T. T. Nhan, D. N. T. Phuong, and W.-K. Wong, “The Effects of Selected Financial Ratios on Profitability: an Empirical Analysis of Real Estate Firms in Vietnam.,” *Annals of Financial Economics*, vol. 17, no. 1, pp. 1–29, Mar. 2022.

[9] M. Lam, “Neural network techniques for financial performance prediction: integrating fundamental and technical analysis,” *Decision Support Systems*, vol. 37, no. 4, pp. 567–581, Sep. 2004.

[10] C. Zhang, H. Zhang, and D. Liu, “A Contrastive Study of Machine Learning on Energy Firm Value Prediction,” vol. 8, 2020.

[11] D. C. S. Popa, D. N. Popa, V. Bogdan, and R. Simut, “Composite financial performance index prediction – a neural networks approach,” *Journal of Business Economics and Management*, vol. 22, no. 2, pp. 277–296, Feb. 2021.

# Appendix

**Table 2.** X Variables Construction

| **Target variables** | **Column Name/ Calculation** | |
|---|---|---|
| Revenue | revtq (quarterly) | Our target1 |
| EBITDA | Operating Income + Depreciation and Amortization | Our target2 |
| **Basic X variables** | **Column Name in dataset (Compstate)** | **Definition** |
| Number of common shares outstanding | cshoq | |
| Common Shares Traded | cshtrq | |
| Capital Expenditure | capxy | Funds used by a company to acquire, upgrade, and maintain physical assets |
| Operating Income (After Depreciation) | oiadpq | Total earnings from its core business functions (Not use it in EBITDA prediction due to target leakage) |
| Net Income | niq | |
| Operating Expense | xoprq | |
| Shareholder's Equity | teqq | |
| Total Assets | atq | |
| EPS | epsfiq | How much profit each outstanding share of common stock has earned |
| **Features Engineering** | **Calculation** | **Definition** |
| Debt to Total Asset | Total Debt / Total Assets<br>= 1 - Shareholder's Equity / Total assets | Grasping financial health and stability (Solvency) |
| NI to Asset | NI/Average asset | |
| ROA | net income / total assets | How profitable a company is relative to its total assets |
| ROE | net income / shareholder's equity | How profitable a company is relative to its net assets |
| Revenue Growth | (current period revenue – previous period revenue) / previous period revenue | |
| Operating Income Growth | (current period Operating Profit – previous period Operating Profit) / previous period Operating Profit | |
| Net Income Growth | (current period Net Income – previous period Net Income) / previous period Net Income | |
| Quick Ratio | current assets - inventories / current liabilities | Indicate liquidity |
| RSI (Relative Strength Index) | 100 – 100 / (1 + RS)<br>RS = Average gain/Average loss | A momentum oscillator that is widely used in technical analysis of stocks and commodities to identify changes in momentum and price direction |
| MVA (Market value-added) | Market Value of Shares – Book Value of Shareholders' Equity | The amount of wealth that a company is able to create for its stakeholders since its foundation |

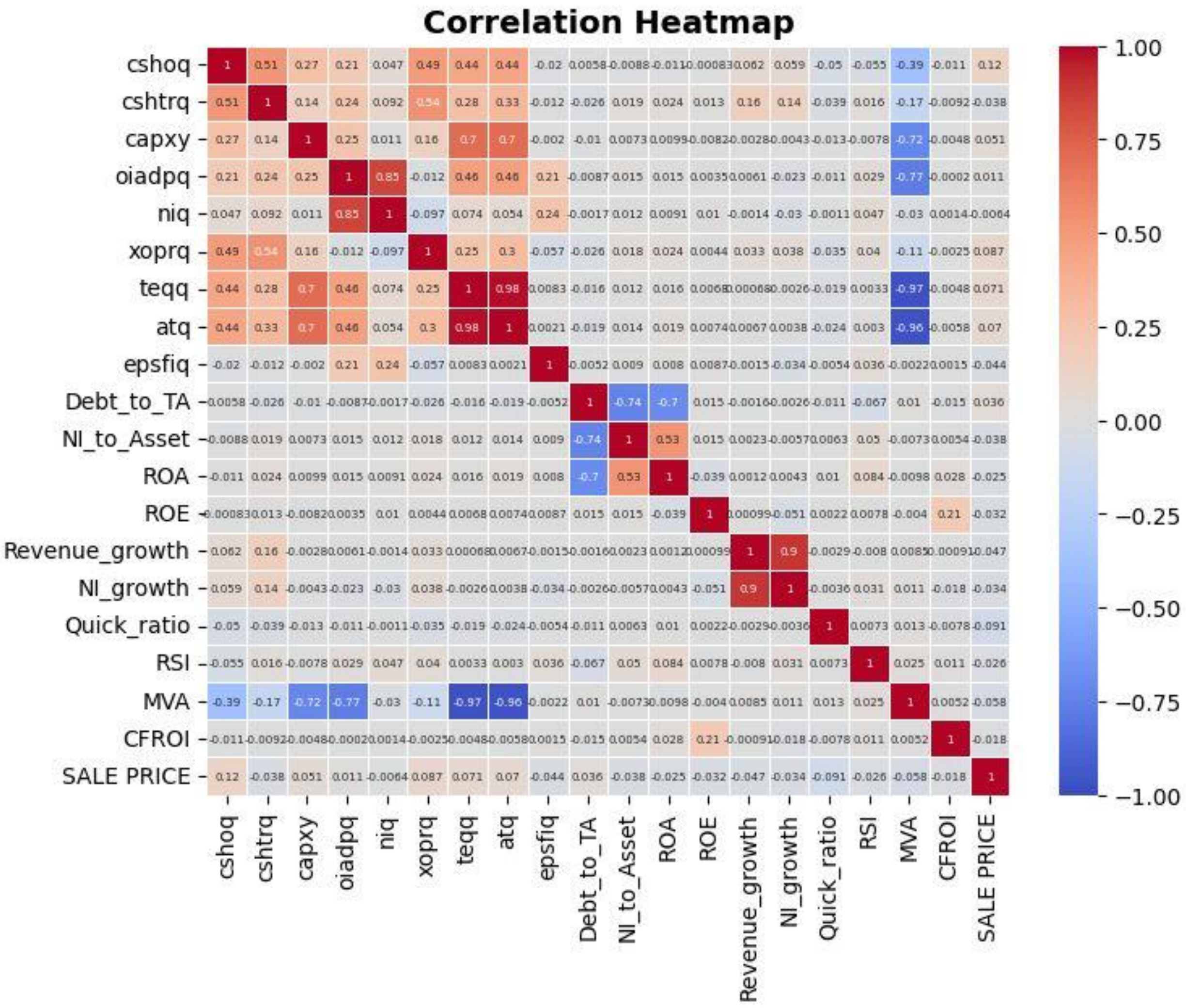
| | | |
|---|---|---|
| CFROI (Cash flow return on investment) | Operating Cash Flow (OCF) / Capital Employed<br>CE = Total Assets – Current Liabilities | A proxy for a company's economic return |
| **Additional X variables** | **Column Name/ Calculation (Compstate Specific)** | **Definition** |
| Rooms Expenses Quarterly | gmrmexpq | |
| Room Revenue Quarterly | gmrmrevq | |
| Finished Homes/Constr in Progress | hbinvfhq | |
| Land Under Development | hbinvludq | |
| Total Homebuilding Inventories | hbinvtq | |
| Homesites/Lots Owned | hbloq | |
| Sale Price | Mean sale price per quarter =sum(Sale Price)/\|Q\| | Quarterly aggregated sale price in ‘nyc-property-sales.csv’ |

**Table 3.** Summary Statistics of the Processed Variables Before Scaling

| Metric | Count | Mean | Std Dev | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| **cshoq** | 1871 | 56.91 | 98.56 | 0.00 | 10.09 | 26.29 | 55.81 | 1184.09 |
| **cshtrq** | 1792 | 1.2e+07 | 3.4e+07 | 0.00 | 6.11e+04 | 5.6e+05 | 5.7e+06 | 4.0e+08 |
| **capxy** | 1920 | 60.87 | 523.16 | -0.45 | 0.00 | 0.72 | 10.79 | 15885.0 |
| **oiadpq** | 1922 | 14.47 | 150.86 | -3673.05 | -0.29 | 0.27 | 8.84 | 941.0 |
| **niq** | 1922 | 1.56 | 113.38 | -2828.92 | -0.88 | -0.00 | 4.50 | 727.66 |
| **xoprq** | 1922 | 163.84 | 542.53 | -0.62 | 0.65 | 5.80 | 66.46 | 6481.37 |
| **teqq** | 1926 | 662.55 | 3731.33 | -1774.91 | 2.99 | 46.00 | 263.96 | 46740.0 |
| **atq** | 1926 | 2042.76 | 9316.12 | 0.00 | 23.79 | 150.83 | 1116.79 | 122520 |
| **epsfiq** | 1904 | 0.41 | 7.08 | -56.45 | -0.06 | 0.00 | 0.17 | 245.10 |
| **Debt/TA** | 1902 | 7.51 | 91.65 | -0.99 | -0.64 | -0.37 | -0.09 | 1853.00 |
| **NI/Asset** | 1873 | -0.22 | 3.43 | -113.60 | -0.02 | 0.00 | 0.01 | 8.11 |
| **ROA** | 1898 | -0.69 | 8.06 | -142.00 | -0.02 | 0.00 | 0.01 | 3.72 |
| **ROE** | 1922 | -0.07 | 1.81 | -62.49 | -0.03 | 0.00 | 0.04 | 16.17 |
| **Revenue** | 1834 | 146.25 | 5990.29 | -2.00 | -0.16 | 0.00 | 0.16 | 256469.7 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **growth** | | | | | | | | 5 |
| **NI growth** | 1926 | 2.34 | 84.96 | -696.18 | -0.96 | -0.21 | 0.49 | 3312.50 |
| **Quick ratio** | 1504 | 24.69 | 227.83 | 0.00 | 0.41 | 1.12 | 2.54 | 4759.40 |
| **RSI** | 1701 | 49.26 | 39.73 | 0.00 | 0.75 | 50.00 | 94.55 | 100.0 |
| **MVA** | 1643 | -815.78 | 7681.59 | -<br>99548.29 | -59.70 | -0.11 | 51.13 | 11358.51 |
| **CFROI** | 1508 | 0.05 | 1.42 | -11.67 | -0.05 | 0.00 | 0.05 | 28.31 |
| **Sale Price** | 1926 | 1.1e+06 | 2.4e+05 | 5.9e+05 | 8.5e+05 | 1.1e+06 | 1.2e+06 | 1.6e+06 |

**Figure 1**. Correlation Heatmap of the datasets


Correlation Heatmap
cshoq
cshtrq
capxy
oiadpq
niq
xoprq
teqq
atq
epsfiq
Debt_to_TA
NI_to_Asset
ROA
ROE
Revenue_growth
NI_growth
Quick_ratio
RSI
MVA
CFROI
SALE PRICE