# FIN 550: Final Project

# EXECUTIVE SUMMARY

# Your Team Name (be creative): Team Omega \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Select whether this is an individual or group submission. No more than 3 members per group.

Tejaswi Pallapothu, Nilesh Choudhary, Subham Shaurav

# Case Overview

The case describes the significance of property taxes as a major revenue source for local governments in the United States, supporting essential services such as police, schools, and pensions. The focus is on the Cook County Assessor's Office (CCAO), responsible for assessing property values in the second most populous county in the U.S., including Chicago and over 130 municipalities.

The valuation task for CCAO is complex due to the diversity of properties across numerous municipalities. Before 2018, the valuation process was criticized for its lack of transparency and accuracy, leading to public and political pressures. The 2018 election resulted in the election of Fritz Kaegi, who pledged to reform CCAO and enhance transparency. To address challenges, the new Assessor invested in technology, increased data transparency, and expanded the data science team. The valuation pipeline shifted from outdated technology to a more capable implementation in R. CCAO emphasizes transparency by making modelling and data decisions publicly available.

The objective is to predict the value of the sale price for each Cook County home in “predict\_property\_data.csv” as closely as possible to its actual value.

# Methodology

**1. Data Exploration:**

Objective: Gain a comprehensive understanding of the datasets and variables.

Approach:

1. Load and explore "predict\_property\_data.csv" and "historic\_property\_data.csv." Referred to "codebook.csv" for variable definitions. Identified missing values (nulls) and non-numeric (categorical, logical, character) variables to convert to factors. Total #rows = 50000 Total #columns = 63, Average sale price = 298741.5834 USD, Max Sale Price = 9400000 USD, Min Sale Price = 100 USD, number of nulls are shown in R script.
2. From codebook.csv, use the “var\_is\_predictor” variable to filter out the columns which are set to FALSE. Removed 23 columns from both historic\_property\_data and predict\_property\_data.

**2. Handling Missing Values:**

Objective: Address missing data to ensure the robustness of the predictive models.

Approach:

1. For categorical/character columns that have a significant amount of nulls (>1000) which were all categorical, we replaced them with the string “unknown”.
2. We have directly removed rows for specific columns where the frequency of occurrence of null values is negligible.

**3. Feature Selection:**

Objective: Identify relevant predictors to enhance model interpretability and performance.

Approach:

1. Transformed all non-numerical variables into factors and subsequently expanded these factors into individual columns (one hot encoding).
2. Used Lasso regression to identify which variables have good predictive power.

**4. Model Building:**

Objective: Develop predictive models for property values based on historical data.

Approach:

1. Split the "historic\_property\_data.csv" into training and testing sets. Train: Test = 4:1.
2. Trained the data using Lasso Regression since the number of variables is high. Lot of these variables will not have enough predictive power and lasso automatically reduces coefficient for such variables.

**5. Model Evaluation:**

Objective: Evaluate model performance using Mean Squared Error (MSE).

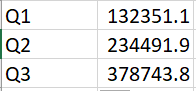
We trained the model on 80% of the historical data and then used the 20% of the data to find the MSE.

Lasso regression MSE: ~ $20B

# Conclusion

|  |  |  |
| --- | --- | --- |
| **Average Price (USD)** | **Max Price (USD)** | **Min Price**  **(USD)** |
| 308456.1 | 5192500 | -219157 |

Quartiles



Further study needs to be conducted to improve the accuracy of the model. A lot of predicted prices are negative which will need to handled separetely.