# Case Overview

**Problem Statement:** The project undertaken addresses the critical challenge of accurately assessing property values in Cook County, the second most populous county in the United States, encompassing over 130 municipalities. The problem lies in the inconsistency and complexity of property data provided by the Cook County Assessor’s Office (CCAO). To tackle this, the project involves processing and analyzing this intricate data through advanced machine learning techniques.

**Objective:** The primary objective as a data scientist is to develop a sophisticated model that enhances the precision of property valuations. This model aims to overcome historical valuation challenges and ensure that property tax assessments are fair, transparent, and reflective of real-world market values, thereby providing stakeholders in Cook County with reliable, data-driven insights for real estate decision-making.

# Methodology

**Data Preprocessing:**

The data preprocessing strategy for "predict\_property\_data.csv" and "historic\_property\_data.csv" encompassed a meticulous six-step procedure:

1. **Variable Selection**: Non-predictor variables, as indicated by the 'var\_is\_predictor' column in the code book, were excluded to focus on essential variables. Additionally, variables related to neighborhood characteristics, tax rates, and school district boundaries were omitted due to their interconnected nature.
2. **Missing Data Handling**: Variables with over 10% missing data were discarded to uphold data integrity.
3. **Uniqueness Filter**: Variables with less than 5 or more than 95% of rows containing a single value were removed, balancing the need for diversity and consistency in data.
4. **Imputation Strategy**: Missing values within similar location groups were imputed based on 'meta\_town\_code' - using median for continuous and mode for categorical variables.
5. **Winsorization**: To mitigate the impact of extreme outliers, values below the 5th percentile and above the 95th percentile were replaced with their respective percentile values.
6. **Variable Transformation**: Certain numeric variables were converted to factors as per guidelines in the code book.

**Variable Selection:** Initially, the project employed LASSO regression on historic property data to identify key factors influencing property sale prices. LASSO was chosen for its ability to handle complex datasets with many variables, providing a balance between model simplicity and predictive accuracy. It automatically highlights impactful factors while diminishing the influence of less significant ones. The model's interpretability and the use of cross-validation for parameter tuning were crucial in optimizing model performance.

**Modelling Method:** In this project, our journey began with the task of selecting the right variables to accurately assess property values in Cook County. We combined statistical techniques with real-world knowledge, ensuring our choices were not just data-driven but also meaningful in the context of the property market.

Initially, we started with LASSO regression and Random Forest algorithms. LASSO was great for pinpointing key factors due to its straightforward linear approach. However, we soon realized that the complexity of our data required a more nuanced method. The relationships between variables were intricate, not just linear, leading us to explore models that could delve deeper into these complexities.

This exploration led us to XGBoost, an advanced tool known for handling large datasets effectively and flexibly working with various data types. Its ability to process data in parallel, manage missing information, and provide clarity on which features really drive property values made it an ideal choice. When we compared XGBoost with the initial models, it stood out with its lower error rates and better explanation capabilities – essentially, it was more accurate and gave us clearer insights.

Throughout the modeling process, we focused on ensuring the model's reliability. We used cross-validation, a technique to test the model's predictions on new data, to prevent it from just memorizing the data it was trained on. This rigorous testing was vital for us to trust its predictions.

Finally, our approach was iterative. We continually refined the model, adapting it to new data and feedback, ensuring it stayed relevant and accurate in the ever-changing world of real estate. This adaptability was key in our efforts to provide the most accurate property value assessments for Cook County.

# Conclusion

The project's analytical efforts culminated in the precise estimation of property values across Cook County, leveraging the predictive strength of the XGBoost model. The model's accuracy was quantified, showing a Mean Squared Error (MSE) of 8,854,282,353.57697, while the R-squared values stood at 0.8710 for training and 0.7862 for the test set, indicating a strong model performance.

The distribution of the assessed property values, as predicted by the model, is summarized by key statistics, including minimum, maximum, mean, and quartile values, which are detailed below. This serves as a structured repository of the model's outputs, ready for review and application by the Cook County Assessor’s Office.

**Summary Statistics of Assessed Property Values:**

* Minimum Value: 8,162
* First Quartile: 160,385
* Median Value: 233,269
* Mean Value: 285,080
* Third Quartile: 351,336
* Maximum Value: 997,087

In conclusion, it is recommended that the CCAO consider the integration of the XGBoost model into their valuation processes due to its demonstrated accuracy and efficiency. The continual updating of data and regular model recalibration are advised to maintain the relevance and precision of property value assessments. This approach is expected to enhance the fairness and transparency of property tax assessments in Cook County.

# Appendix

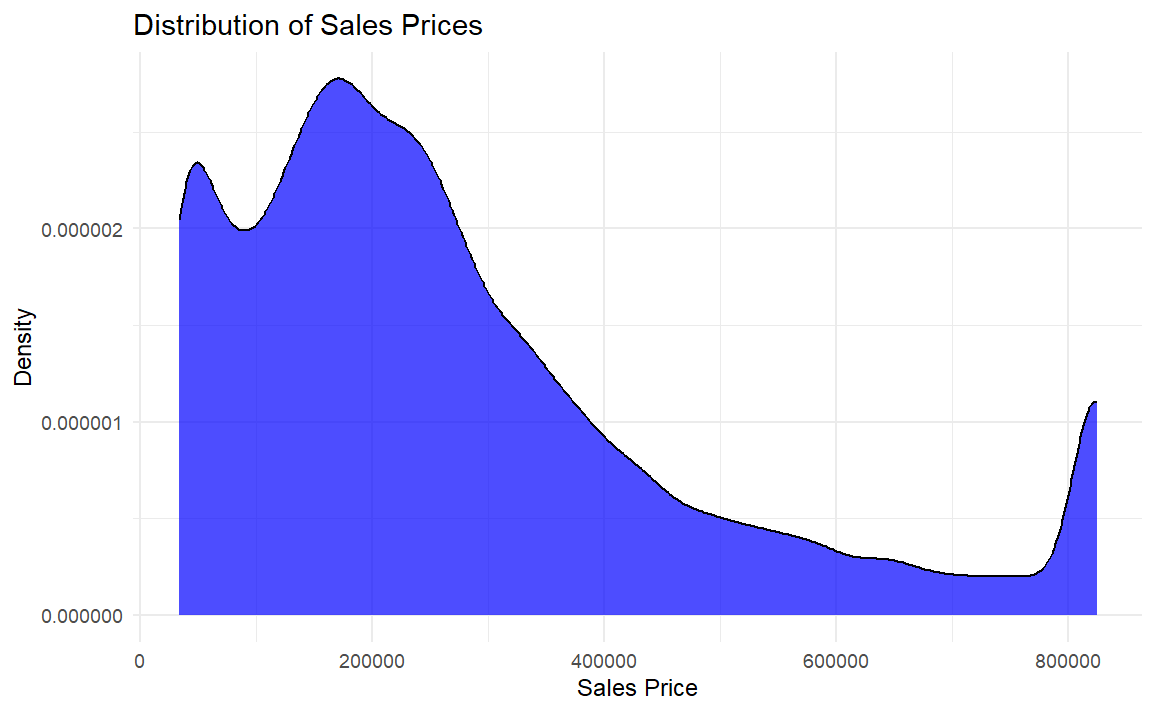
Include tables and plots here. Enumerate each table and plot, give each a descriptive title, and make sure elements are labeled clearly. Tables and plots should correspond to output from your code.

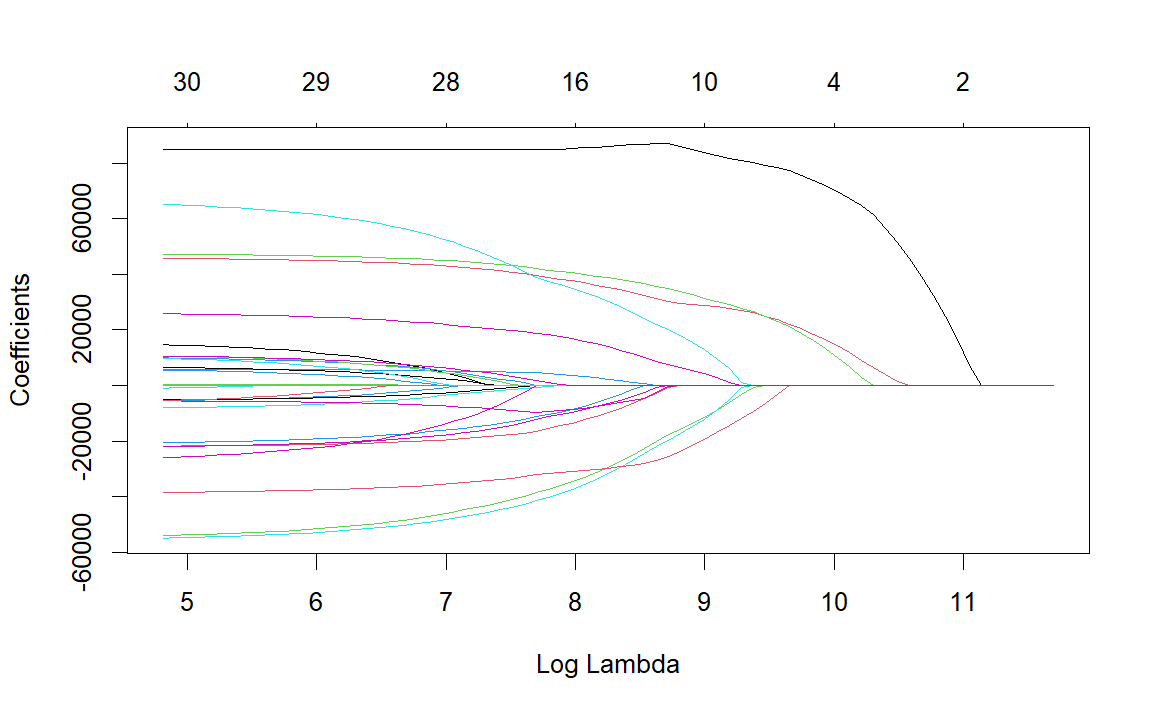
**Table 1: Summary of Model Performance Metrics**

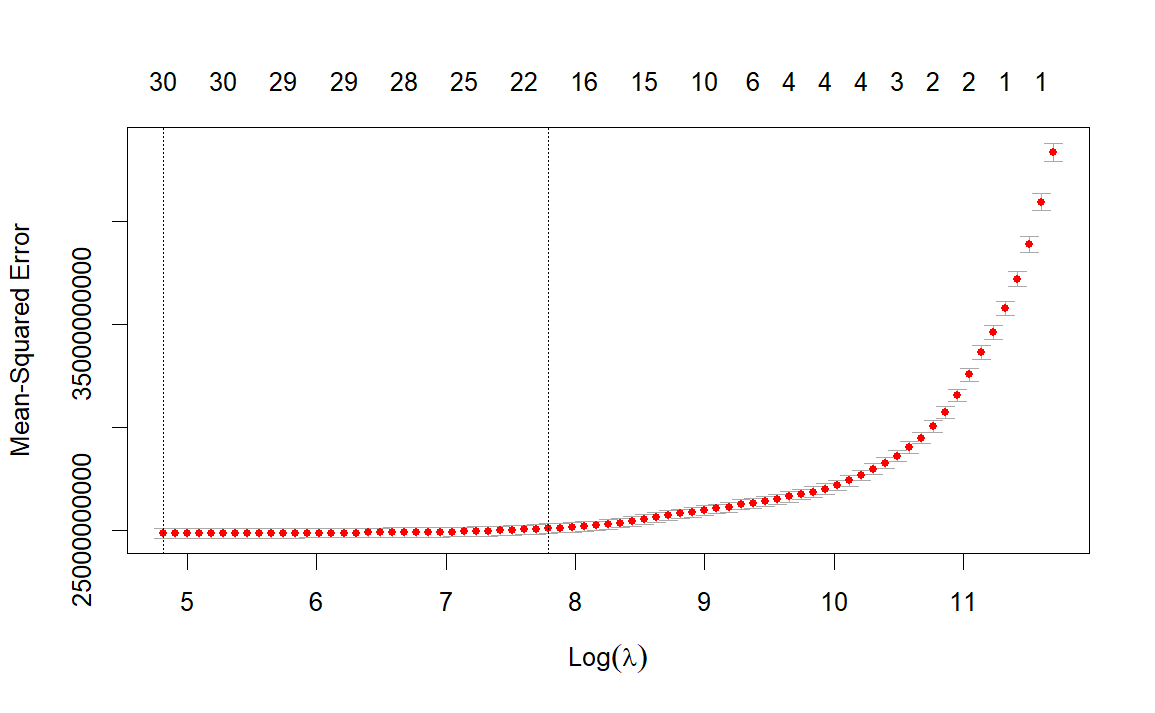
| **Model** | **MSE** | **R-Squared (Training)** | **R (Training)** | **R-Squared (Test)** | **R (Test)** |
| --- | --- | --- | --- | --- | --- |
| LASSO Regression | 24,416,814,042 | - | - | - | - |
| Random Forest | 14,146,125,947.81 | - | - | - | - |
| XGBoost | 8,854,282,353.57 | 0.8710 | 0.9333 | 0.7862 | 0.8867 |

*Note: The MSE values are indicative of the average squared difference between the estimated and actual values. The R-Squared and R values represent the proportion of variance explained by the model and the correlation coefficient, respectively, showcasing the model’s accuracy and predictive power.*

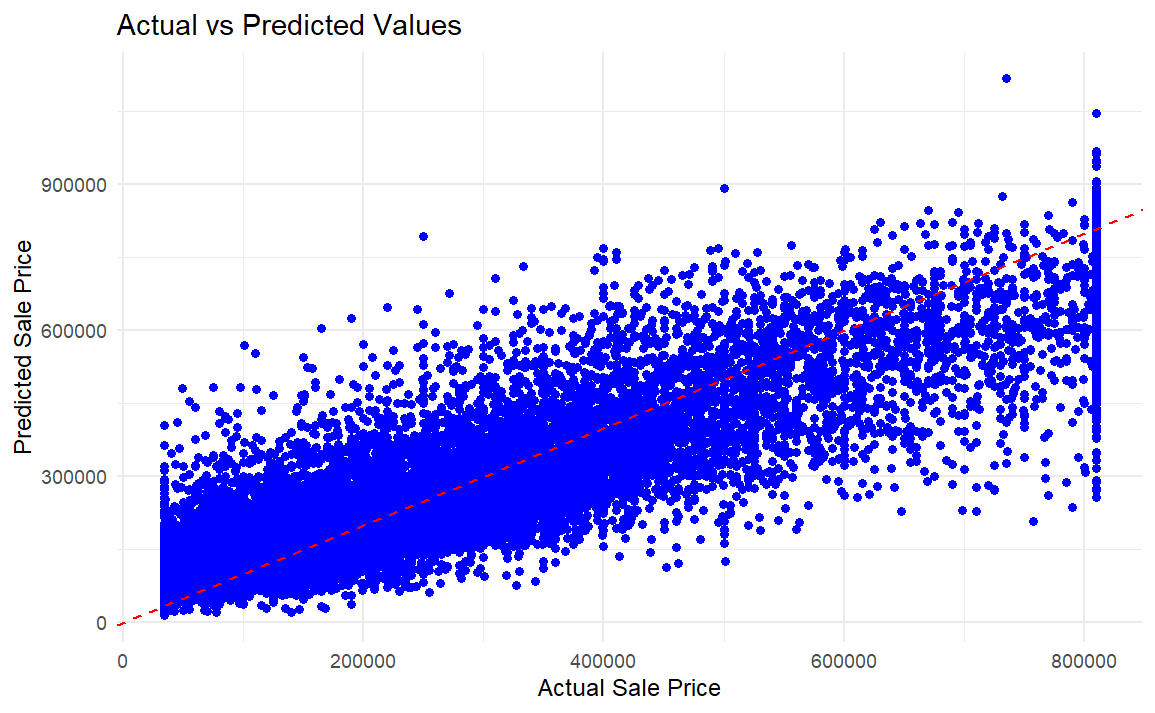
**Plot 1: Distribution of Sales Prices**

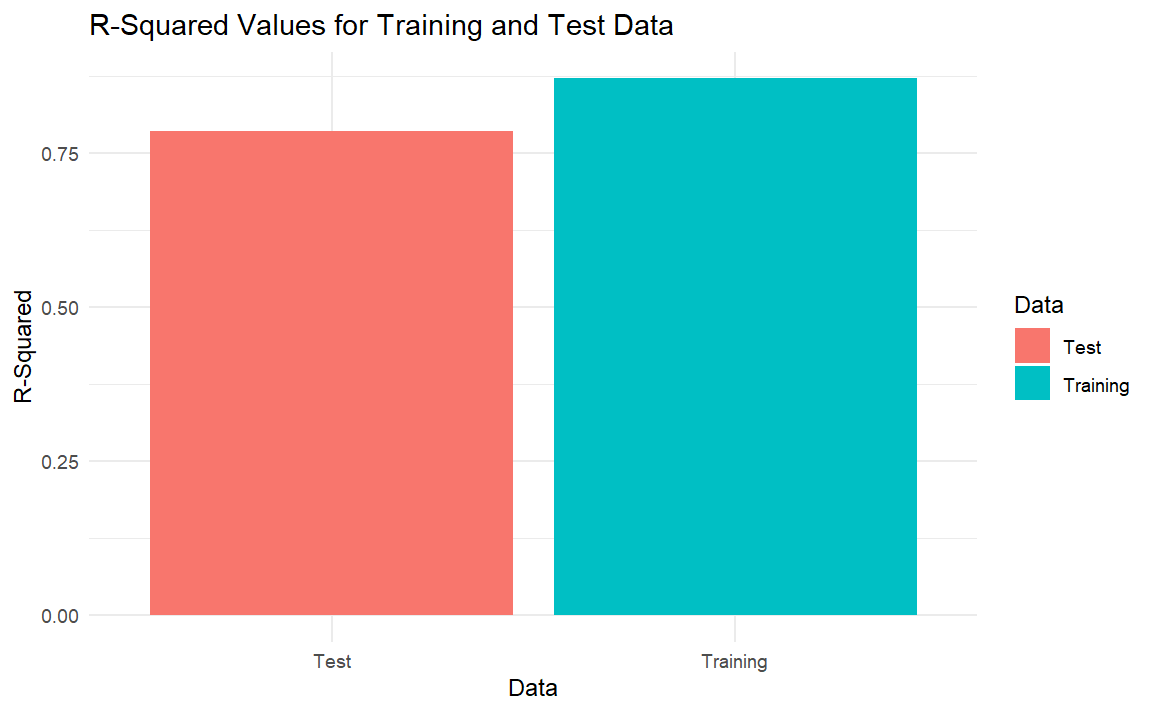


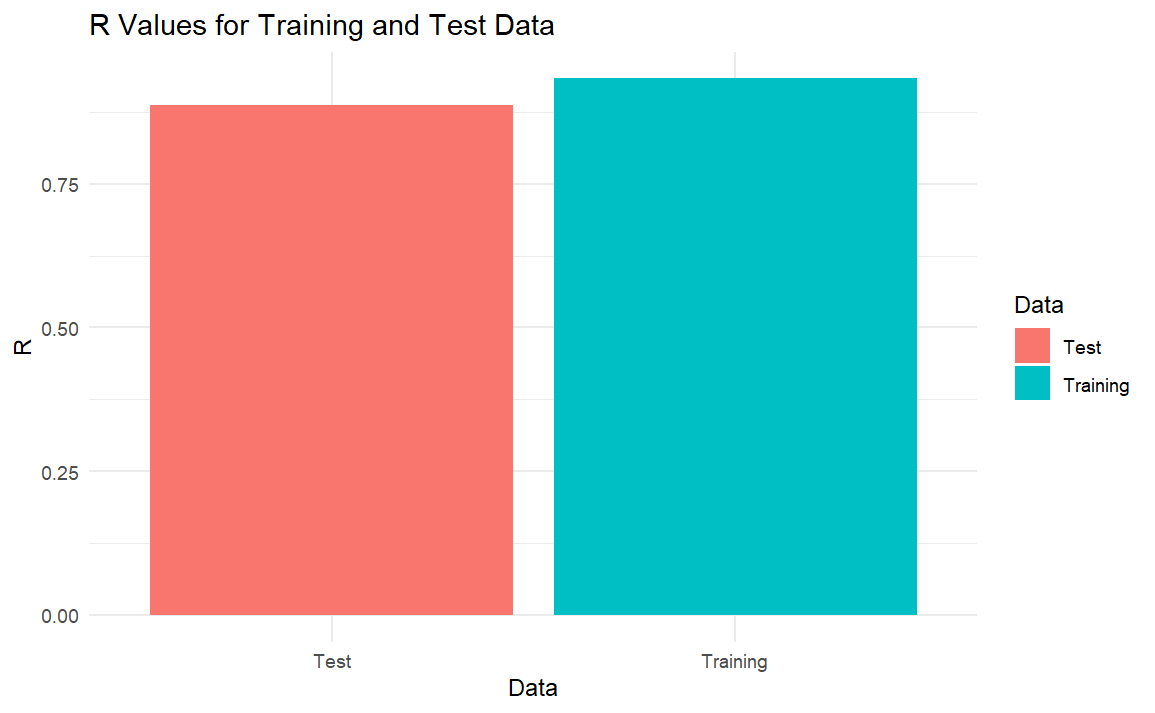
**Plot 2: LASSO Coefficient Paths**

**Plot 3: LASSO Cross-Validated MSE**

**Plot 4: Random Forest Variable Importance**A screenshot of a computer screen

Description automatically generated**Plot 5: XGBoost Actual vs Predicted Values**

**Plot 6: XGBoost R-Squared Values for Training and Test Data**

**Plot 7: XGBoost R Values for Training and Test Data**

**Plot 8: Distribution of Assessed Property Values**