ALY6015 Intermediate Analytics

Final Project: Initial Analysis Report

Boston Property Assessment Analysis

Professor Wada Roy

Team Members: Sriram Kannepalli, Abhinav Girish, Ruilian Tao

Date: Dec 1, 2024

# Introduction

The goal of this project is to analyze property valuation data across multiple years to understand trends, relationships, and drivers of property values. Using datasets from 2018 and 2022, we aim to address the following key questions:

1. How do property values differ across years and ZIP codes?
2. What factors significantly influence property values?
3. Are there observable correlations between living area, building value, and total property value?

To answer these questions, the methods used include summary statistics, regression modeling, ANOVA, and correlation analysis.

# Data Analysis

## Data Description

The analysis is based on property valuation data from two years (2018 and 2022). Key variables include:

* TOTAL\_VALUE: Total property value
* LIVING\_AREA: Living area in square feet
* BLDG\_VALUE: Building value
* ZIPCODE: Geographic ZIP code of the property

## Analysis Techniques

1. **Descriptive Statistics and Subset Analysis:** To explore trends in the data, summary statistics were computed by year and ZIP code.
2. **Regression Analysis:** Linear and logistic regression models were used to identify predictors of total property value and categorize properties into high and low value groups.
3. **Correlation and ANOVA:** Correlation analysis was performed to understand the strength of relationships between variables. ANOVA was conducted to assess the impact of year and ZIP code on property values.

## Descriptive Statistics Results

**Overall Summary:**

| **Metric** | **Value** |
| --- | --- |
| Mean Total Value | $1,925,415 |
| Median Total Value | $507,400 |
| Standard Deviation (Total Value) | $24,109,989 |
| Mean Living Area (sq. ft.) | 3,805.58 |
| Standard Deviation (Living Area) | 31,772.26 |

**By Year:**

| **Year** | **Mean Total Value** | **Median Total Value** | **SD (Total Value)** | **Mean Living Area (sq. ft.)** | **SD (Living Area)** |
| --- | --- | --- | --- | --- | --- |
| 2018 | $2,446,157 | $451,000 | $31,616,847 | 3,561 | 29,389 |
| 2022 | $1,420,961 | $574,500 | $13,248,762 | 4,070 | 34,556 |

**By ZIP Code:**

| **ZIP Code** | **Mean Total Value** | **Median Total Value** | **SD (Total Value)** | **Mean Living Area (sq. ft.)** |
| --- | --- | --- | --- | --- |
| 2026 | $439,050 | $539,350 | $367,637 | 1,715 |
| 2110 | $5,848,896 | $936,000 | $39,559,016 | 12,968 |

## Regression Analysis

**Linear Regression:**

* Formula: TOTAL\_VALUE ~ LIVING\_AREA + BLDG\_VALUE + ZIPCODE
* Adjusted R-squared: 0.932
* Key Results:
  + Living Area: Coefficient = 56.02 (p < 0.001)
  + Building Value: Coefficient = 1.22 (p < 0.001)
  + ZIP Code: Not statistically significant (p = 0.229)

**Logistic Regression:**

* Formula: Value\_Category ~ LIVING\_AREA + BLDG\_VALUE + ZIPCODE
* AIC: 168,118
* Key Results:
  + Living Area: Negative association with high property value (p < 0.001)
  + Building Value: Strong negative association (p < 0.001)
  + ZIP Code: Weak positive association (p = 4.03e-07)

## ANOVA Results

**Total Property Value by Year and ZIP Code**

| **Source** | **Df** | **Sum Sq** | **Mean Sq** | **F value** | **p-value** |
| --- | --- | --- | --- | --- | --- |
| Year | 1 | 9.229e+16 | 9.229e+16 | 158.83 | <0.001 |
| ZIP Code | 1 | 3.591e+15 | 3.591e+15 | 6.18 | 0.013 |
| Residuals | 351253 | 2.041e+20 | 5.810e+14 |  |  |

Our ANOVA analysis highlights the statistical significance of differences in property values based on the year of assessment and ZIP code. Specifically, the results show that both the year (F = 158.83, p < 0.001) and ZIP code (F = 6.18, p = 0.013) significantly affect total property values. This suggests that temporal and geographic factors play a role in property valuation trends. The year-to-year differences reflect broader economic or market influences, while ZIP code-based disparities likely indicate localized variations such as neighborhood development, demand, or socio-economic conditions.

The residuals from the model (Sum of Squares = 2.041e+20, Mean Square = 5.810e+14) suggest that while the independent variables capture some variability, a large proportion remains unexplained. This points to the potential impact of unaccounted factors, such as specific property characteristics, regional policies, or external economic factors. Despite these limitations, the ANOVA results provide robust evidence of significant group differences and underline the need for further granular analyses to pinpoint underlying drivers of these trends.

**Correlation Analysis**

| **Variable** | **Total Value** | **Living Area** | **Building Value** |
| --- | --- | --- | --- |
| Total Value | 1.000 | 0.445 | 0.963 |
| Living Area | 0.445 | 1.000 | 0.400 |
| Building Value | 0.963 | 0.400 | 1.000 |

The correlation analysis reveals critical relationships among key variables—Total Property Value, Living Area, and Building Value. Total Property Value demonstrates a very strong positive correlation with Building Value (r = 0.963), indicating that the building's valuation is a primary driver of the overall property value. This aligns with expectations, as building components often comprise the largest share of property assessments.

Conversely, the correlation between Total Property Value and Living Area is moderate (r = 0.445), suggesting that while larger living areas generally correspond to higher property values, the relationship is not as dominant as that of Building Value. The relatively weaker correlation between Living Area and Building Value (r = 0.400) further indicates that building quality or features might outweigh sheer size in influencing value. These findings emphasize the dominant role of Building Value in determining Total Property Value, while also highlighting that Living Area, though influential, may be moderated by other factors such as location or architectural quality.

## Visualizations

1. **Distribution of Total Property Values by Year:**

A graph of a distribution of property values

Description automatically generated

The image shows a scatter plot comparing the distribution of total property values across two years: 2018 and 2022.

Key Features

**Value Range**

* The y-axis represents total property values, ranging from 0 to approximately 2.0e+09 (2 billion)
* Most data points are concentrated in the lower range, below 5.0e+08 (500 million)

**Distribution Pattern**

* Both years show a similar vertical distribution pattern
* There are dense clusters of points near the bottom
* Several outliers appear at higher values, particularly around 1.0e+09 to 2.0e+09

**Temporal Comparison**

* The distribution patterns between 2018 and 2022 appear relatively consistent
* Both years exhibit a right-skewed distribution with numerous outliers above the main cluster
* The majority of properties in both years are valued in the lower range

**Outliers**

* Notable outliers exist in both years
* The highest values reach approximately 2.0e+09
* There's a visible gap between the main cluster of values and the outliers

This graph suggests that the property value distribution remained fairly stable between these two time periods, with most properties concentrated in lower values and a small number of high-value outliers.

1. **Living Area vs. Total Property Value:**

A graph with numbers and a diagram

Description automatically generated with medium confidence

This scatter plot visualizes the relationship between Living Area (in square feet) and Total Property Value, comparing data points from 2018 and 2022. Here are the key observations:

Distribution Pattern

The plot shows a generally positive correlation between living area and property value, with larger homes typically commanding higher values. The majority of data points cluster in the lower left portion of the graph, indicating that most properties have living areas under 2 million square feet and values under 500 million.

Time Comparison

* **2018 Data** is shown in coral/red points
* **2022 Data** is shown in turquoise/blue points

Value Range

The property values range from near zero to approximately 2 billion, with most properties concentrated below 500 million. There are several notable outliers with extremely high property values, particularly a few properties showing values around 2 billion.

Living Area Range

The living areas primarily range from 0 to about 6 million square feet, with the vast majority of properties having less than 2 million square feet of living space. There appears to be a sparse distribution of properties with extremely large living areas.

Outliers

Several notable outliers appear in both dimensions:

* A few properties with exceptionally high values (around 2 billion)
* Some properties with unusually large living areas (around 6 million square feet)
* These outliers deviate significantly from the main cluster of data points

The visualization effectively highlights the evolution of property values and sizes between these two time periods, though the relationship appears relatively consistent across both years.

1. **ANOVA Results (Year and ZIP Code):**

A graph of a property value

Description automatically generated

This scatter plot visualizes Total Property Values across different ZIP codes for two years: 2018 and 2022. The analysis is presented as an ANOVA (Analysis of Variance) comparison.

**Key Features**

**Distribution Pattern**

* The y-axis shows Total Property Value ranging from 0 to 2.0e+09 (2 billion)
* Property values are scattered vertically, with most values concentrated in the lower range
* Several high-value outliers appear at the top of both years

**Temporal Comparison**

* Data points are plotted for two specific years: 2018 and 2022
* The overall distribution pattern appears similar between both years
* Some properties show notably high values above 1.5 billion in both periods

**ZIP Code Classification**

* The legend shows ZIP codes from 2026 to 2467
* ZIP codes are color-coded for easy identification
* Includes an "NA" category for unclassified data points

**Visual Characteristics**

The plot uses a consistent point size for all data markers, making it easy to identify individual properties. The grid lines help in reading approximate values, while the color-coding system helps distinguish between different ZIP code areas. The distribution suggests a right-skewed pattern, with most properties having lower values and fewer extremely high-value properties.

# Interpretation & Conclusions

* **Summary:** The analysis revealed significant differences in property values by year, with 2018 showing higher average values. Building value is the strongest predictor of total property value, as shown by both regression and correlation analyses.
* **Limitations:** The dataset contains missing values, particularly in ZIP code analysis, which may impact conclusions.
* **Future Work:** To improve model performance, advanced methods such as stepwise regression and LASSO could be employed. Further external datasets (e.g., CPI or population) could be integrated for richer analysis.

# Appendices

**R Code**

|  |
| --- |
| # Load necessary libraries  library(dplyr)  library(ggplot2)  library(tidyr)  library(car)  # Load datasets  data\_2018 <- read.csv("ast2018full.csv", stringsAsFactors = FALSE)  data\_2022 <- read.csv("fy2022pa-4.csv", stringsAsFactors = FALSE)  # Data Cleaning and Standardization  data\_2018 <- data\_2018 %>%  select(PID, ZIPCODE, AV\_TOTAL, AV\_BLDG, LIVING\_AREA) %>%  rename(  TOTAL\_VALUE = AV\_TOTAL,  BLDG\_VALUE = AV\_BLDG  ) %>%  mutate(  TOTAL\_VALUE = as.numeric(gsub("[^0-9.-]", "", TOTAL\_VALUE)),  BLDG\_VALUE = as.numeric(gsub("[^0-9.-]", "", BLDG\_VALUE)),  LIVING\_AREA = as.numeric(gsub("[^0-9.-]", "", LIVING\_AREA)),  Year = 2018  )  data\_2022 <- data\_2022 %>%  select(PID, ZIPCODE, TOTAL\_VALUE, BLDG\_VALUE, LIVING\_AREA) %>%  mutate(  TOTAL\_VALUE = as.numeric(gsub("[^0-9.-]", "", TOTAL\_VALUE)),  BLDG\_VALUE = as.numeric(gsub("[^0-9.-]", "", BLDG\_VALUE)),  LIVING\_AREA = as.numeric(gsub("[^0-9.-]", "", LIVING\_AREA)),  Year = 2022  )  # Combine datasets  data\_combined <- rbind(data\_2018, data\_2022)  # Create additional variables  median\_value <- median(data\_combined$TOTAL\_VALUE, na.rm = TRUE)  data\_combined <- data\_combined %>%  mutate(Value\_Category = factor(ifelse(TOTAL\_VALUE > median\_value, "High", "Low")))  # Subset Analysis  # Overall Summary  overall\_stats <- data\_combined %>%  summarize(  Mean\_Total\_Value = mean(TOTAL\_VALUE, na.rm = TRUE),  Median\_Total\_Value = median(TOTAL\_VALUE, na.rm = TRUE),  SD\_Total\_Value = sd(TOTAL\_VALUE, na.rm = TRUE),  Mean\_Living\_Area = mean(LIVING\_AREA, na.rm = TRUE),  SD\_Living\_Area = sd(LIVING\_AREA, na.rm = TRUE),  Mean\_Building\_Value = mean(BLDG\_VALUE, na.rm = TRUE),  SD\_Building\_Value = sd(BLDG\_VALUE, na.rm = TRUE)  )  # By Year  year\_based\_stats <- data\_combined %>%  group\_by(Year) %>%  summarize(  Mean\_Total\_Value = mean(TOTAL\_VALUE, na.rm = TRUE),  Median\_Total\_Value = median(TOTAL\_VALUE, na.rm = TRUE),  SD\_Total\_Value = sd(TOTAL\_VALUE, na.rm = TRUE),  Mean\_Living\_Area = mean(LIVING\_AREA, na.rm = TRUE),  SD\_Living\_Area = sd(LIVING\_AREA, na.rm = TRUE),  Mean\_Building\_Value = mean(BLDG\_VALUE, na.rm = TRUE),  SD\_Building\_Value = sd(BLDG\_VALUE, na.rm = TRUE)  )  # By ZIP Code  zipcode\_stats <- data\_combined %>%  group\_by(ZIPCODE) %>%  summarize(  Mean\_Total\_Value = mean(TOTAL\_VALUE, na.rm = TRUE),  Median\_Total\_Value = median(TOTAL\_VALUE, na.rm = TRUE),  SD\_Total\_Value = sd(TOTAL\_VALUE, na.rm = TRUE),  Mean\_Living\_Area = mean(LIVING\_AREA, na.rm = TRUE),  SD\_Living\_Area = sd(LIVING\_AREA, na.rm = TRUE),  Mean\_Building\_Value = mean(BLDG\_VALUE, na.rm = TRUE),  SD\_Building\_Value = sd(BLDG\_VALUE, na.rm = TRUE)  )  # Regression Models  linear\_model <- lm(TOTAL\_VALUE ~ LIVING\_AREA + BLDG\_VALUE + ZIPCODE, data = data\_combined)  linear\_summary <- summary(linear\_model)  logistic\_model <- glm(Value\_Category ~ LIVING\_AREA + BLDG\_VALUE + ZIPCODE, data = data\_combined, family = binomial)  logistic\_summary <- summary(logistic\_model)  # ANOVA Analysis  anova\_model <- aov(TOTAL\_VALUE ~ Year + ZIPCODE, data = data\_combined)  anova\_summary <- summary(anova\_model)  # Correlation  correlation\_results <- data\_combined %>%  select(TOTAL\_VALUE, LIVING\_AREA, BLDG\_VALUE) %>%  drop\_na() %>%  cor()  # Visualizations  # 1. Boxplot of Total Property Value by Year  ggplot(data\_combined, aes(x = factor(Year), y = TOTAL\_VALUE)) +  geom\_boxplot() +  labs(title = "Distribution of Total Property Values by Year", x = "Year", y = "Total Property Value") +  theme\_minimal()  # 2. Scatter Plot of Living Area vs. Total Property Value  ggplot(data\_combined, aes(x = LIVING\_AREA, y = TOTAL\_VALUE, color = factor(Year))) +  geom\_point(alpha = 0.5) +  labs(title = "Living Area vs. Total Property Value", x = "Living Area (sq ft)", y = "Total Property Value") +  theme\_minimal()  # 3. ANOVA Plot (Mean Total Value by Year and ZIP Code)  ggplot(data\_combined, aes(x = factor(Year), y = TOTAL\_VALUE, fill = factor(ZIPCODE))) +  geom\_boxplot() +  labs(title = "ANOVA: Total Property Value by Year and ZIP Code", x = "Year", y = "Total Property Value") +  theme\_minimal()  # Outputs  print("Overall Summary Statistics:")  print(overall\_stats)  print("Year-Based Summary Statistics:")  print(year\_based\_stats)  print("ZIP Code-Based Summary Statistics:")  print(zipcode\_stats)  print("Linear Regression Summary:")  print(linear\_summary)  print("Logistic Regression Summary:")  print(logistic\_summary)  print("ANOVA Summary:")  print(anova\_summary)  print("Correlation Results:")  print(correlation\_results) |

# Reference

Analytics Vidhya. (2023). What is Analysis of Variance (ANOVA)?. Retrieved from <https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/>

City of Boston. (n.d.). Property assessment data. Retrieved November 5, 2024, from <https://data.boston.gov/dataset/property-assessment>

DATAtab. (n.d.). ANOVA (Analysis of Variance) • Simply Explained. Retrieved from <https://datatab.net/tutorial/anova>

Pennsylvania State University. (n.d.). Lesson 10: Introduction to ANOVA | STAT 500. Retrieved from <https://online.stat.psu.edu/stat500/lesson/10>

Real Statistics Using Excel. (2013). ANOVA using Regression. Retrieved from <https://real-statistics.com/multiple-regression/anova-using-regression/>

Research Method. (2023). ANOVA (Analysis of Variance) - Formulas, Types, and Examples. Retrieved from <https://researchmethod.net/anova/>

Statology. (2021). ANOVA vs. Regression: What’s the Difference?. Retrieved from <https://www.statology.org/anova-vs-regression/>

Scribbr. (2020). Descriptive Statistics | Definitions, Types, Examples. Retrieved from <https://www.scribbr.com/statistics/descriptive-statistics/>

University of Connecticut. (n.d.). ANOVA, Regression, and Chi-Square. Retrieved from <https://researchbasics.education.uconn.edu/anova_regression_and_chi-square/>