Final Project : Initial Analysis Report

Group 9

Sharvil Kishor Wadekar
Mohit Ravindra Kamble
Soham Santosh Mane

College of Professional Studies – Northeastern University

ALY6015: Intermediate Analytics
Prof. Roy Wada

February 6, 2024

Introduction

This project explores Boston's housing dataset through comprehensive analysis, offering insights into diverse facets of the property market. Through detailed examination, including summary statistics, variable transformations, and analytical techniques such as correlation matrices, regression modeling, ANOVA, and chi-square tests, this study uncovers underlying patterns, trends, and relationships within the dataset. These findings provide valuable guidance for stakeholders navigating Boston's dynamic housing landscape.

Analysis

Summary Statistics

Analysis: The subset analysis of the Boston housing dataset reveals distinctive insights into specific segments of the property market. Focusing on properties located on Beacon Street, built in 1990, and categorized under land use code "CD," allows for a deeper understanding of the characteristics, values, and trends within these subsets. These targeted analyses provide valuable context for evaluating property dynamics, identifying potential market niches, and informing strategic decision-making in the real estate sector.

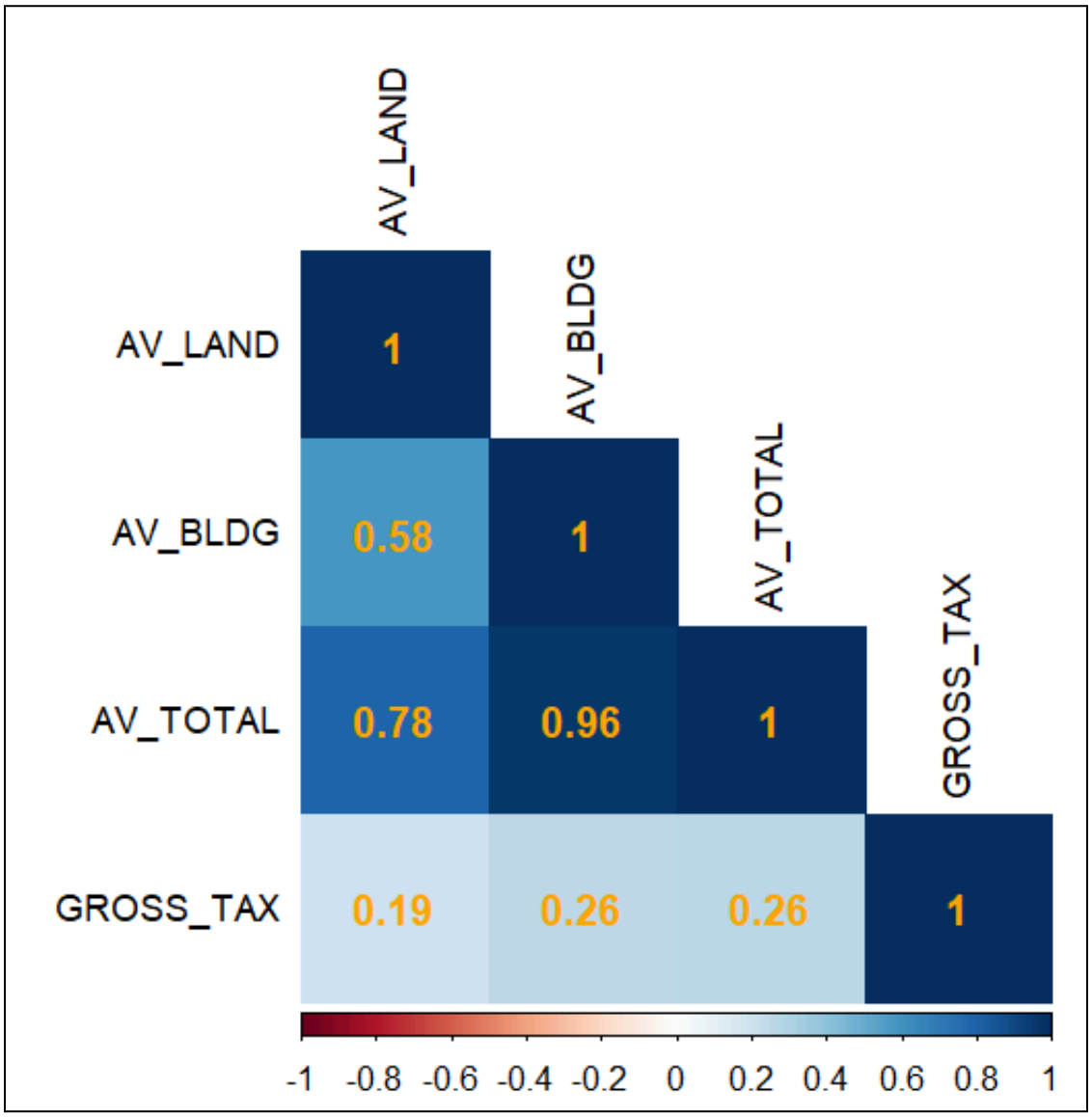
AV_LAND	AV_BLDG	AV_TOTAL
Min. : 0	Min. : 0	Min. : 0
1st Qu: 0	1st Qu: 402300	1st Qu: 416500
Median: 0	Median: 669200	Median: 682800
Mean: 3418769	Mean: 4478509	Mean: 7897278
3rd Qu: 0	3rd Qu: 1312400	3rd Qu: 1398550
Max : 253666900	Max: 418615600	Max: 498445500

Creating/ Transforming new variables

Analysis: The Boston housing dataset was augmented with two new ratio variables, *"landtotal_ratio" and "bldgtotal_ratio,"* representing the proportion of assessed land value and building value, respectively, to the total assessed property value. This addition offers insights into the relative contributions of land and building values to overall property assessments. Understanding these ratios aids in assessing property market dynamics and devising effective valuation strategies for real estate investments and developments.

Analytical Methods

Correlation Matrix:



Analysis:

The analysis uncovers substantial positive correlations between AV_TOTAL and AV_BLDG (0.96), and a moderate correlation with AV_LAND (0.78), underscoring the importance of building and land values in determining property worth. Weak positive correlations between GROSS_TAX and AV_BLDG (0.26) and AV_TOTAL (0.26) indicate their impact on tax rates. Interestingly, AV_LAND exhibits a weaker correlation (0.19) with GROSS_TAX, suggesting a less direct influence.

Regression Table:

```
Multiple Linear Regression Results
================================================
==
                          Dependent variable:

------------------------------
                                 AV_TOTAL
------------------------------------------------
--
LAND_SF                           5.567***
                                  (0.034)

GROSS_AREA                       406.775***
                                  (2.401)

Constant                        423,343.100***
                                 (68,374.130)

------------------------------------------------
--
Observations                       168,494
R2                                  0.265
Adjusted R2                         0.265
Residual Std. Error  27,715,243.000 (df =
168491)
F Statistic         30,318.840*** (df = 2;
168491)
================================================
==
Note:                    *p<0.1; **p<0.05;
***p<0.01
```

Analysis:

The multiple linear regression model predicts AV_TOTAL based on LAND_SF and GROSS_AREA. Both predictors have significant effects on AV_TOTAL ($p < 0.01$). LAND_SF coefficient (5.567) indicates that, on average, a one-unit increase in land square footage is associated with a $5.567 increase in property value. GROSS_AREA coefficient (406.775) suggests a larger impact on property value. The model's R-squared value is 0.265.

ANOVA:

```
                Df                  Sum Sq            Mean Sq F
value              Pr(>F)
LU              16   7536996705730611200 471062294108163200
482.3 <0.0000000000000002 ***
Residuals   174651 170569685378288418816    976631598893155
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1
```

Analysis:
ANOVA results reveal a significant difference in mean AV_TOTAL across different levels of the categorical variable 'LU' ($p < 0.0000000000000002$). This suggests that the land use category significantly influences property values. The F-statistic of 482.3 indicates substantial variation in AV_TOTAL between land use categories, emphasizing the importance of considering land use when assessing property values.

Chi-square Test:

```
> summary(mod3)
          Length Class  Mode
statistic   1    -none- numeric
parameter   1    -none- numeric
p.value     1    -none- numeric
method      1    -none- character
data.name   1    -none- character
observed  422    table  numeric
expected  422    -none- numeric
residuals 422    table  numeric
stdres    422    table  numeric
```

Analysis:
The chi-square test revealed no significant association ($p = 0$) between property type (PTYPE) and owner occupancy (OWN_OCC). This implies that property type does not significantly influence owner occupancy status. The lack of significance suggests no evidence to reject the null hypothesis, indicating independence between property type and owner occupancy. Therefore, property type alone may not be a determining factor in predicting owner occupancy status.

Conclusion

In conclusion, the analysis of the Boston housing dataset has provided valuable insights into various aspects of the property market. Through subset analysis, variable transformations, and analytical techniques, we gained a deeper understanding of property characteristics, market dynamics, and influencing factors. The correlations, regression results, ANOVA, and chi-square tests shed light on relationships between variables and highlighted significant factors impacting property values and ownership patterns. This comprehensive analysis equips stakeholders with actionable insights for informed decision-making in the real estate sector.

References

1. Anova : source
2. Chi-square: source
3. Regression: source
4. Correlation: source
5. Dataset: source

Appendix

```
# Final Project: Initial Analysis Report

library(dplyr)
library(tidyverse)
boston <- read.csv("D:/NEU STUDY/2nd Quarter/Intermediate Analytics (ALY 6015)/Group
Project/fy19fullpropassess.csv")

# Data Overview
str(boston)
summary(boston)

# Subset of street name BEACON
subset_beacon <- subset(boston, ST_NAME == "BEACON")
str(subset_beacon)
summary(subset_beacon)

# Subset of year 1990
subset_1990 <- subset(boston, YR_BUILT == 1990)
str(subset_1990)
summary(subset_1990)

# Subset of land used cd
subset_cd <- subset(boston, LU == "CD")
str(subset_cd)
summary(subset_cd)

# New Ratio variable of AV_LAND/AV_TOTAL
boston$landtotal_ratio <- round((boston$AV_LAND / boston$AV_TOTAL), 2)

# New Ratio variable of AV_BLDG/AV_TOTAL
boston$bldgtotal_ratio <- round((boston$AV_BLDG / boston$AV_TOTAL), 2)
names(boston)
View(boston)

# Correlation Matrix
library(corrplot)

# Replace missing values with the calculated mean
mean_yr_remod <- mean(boston$YR_REMOD, na.rm = TRUE)
boston$YR_REMOD[is.na(boston$YR_REMOD)] <- mean_yr_remod

selected_columns <- boston[c("AV_LAND", "AV_BLDG",   "AV_TOTAL", "GROSS_TAX")]
```

```
# Correlation matrix
cor_matrix <- cor(selected_columns)

# Correlation plot
corrplot(cor_matrix, method = "color", type = "lower", tl.cex = 0.9, tl.col = "black", addCoef.col
= "orange")

# Regression Table
# Predicting AV_TOTAL based on LAND_SF and GROSS_AREA
mod1 <- lm(AV_TOTAL ~ LAND_SF + GROSS_AREA, data = boston)
summary(mod1)

install.packages("stargazer")
library(stargazer)
stargazer(mod1, title = " Multiple Linear Regression Results", type = "text")
mod1$p.value


# ANOVA test
# Comparing the mean AV_TOTAL across different levels of a categorical variable 'LU'
anova_model <- aov(AV_TOTAL ~ LU, data = boston)
mod2 <- aov(AV_TOTAL ~ LU, data = boston)
summary(mod2)
mod2$p.value

# Chi-square Test
contgncy_table <- table(boston$PTYPE, boston$OWN_OCC)
mod3 <- chisq.test(contgncy_table)
summary(mod3)
mod3$p.value
```