Statistical Model to Predict Housing Prices in Astoria

Imran Sabur
BUA 633

I. Introduction

Astoria's proximity to Manhattan and its dense, walkable streets has made it an attractive neighborhood for many to reside in. This paper aims to develop a predictive model that helps determine the price of homes in Astoria. The analysis outlined in this paper will be beneficial not only to prospective home buyers, but real estate companies looking to invest in the area.

II. Previous Research

This is the first time this kind of research is being done.

III. Methodology

The research is cross-section. Data was collected from the NYC mayor's office. Among thousands of homes sold in NYC between 2021- 2022, roughly 350 homes were sold in Astoria. The analysis will employ two graphical techniques: histograms and scatterplots. The statistical analysis in this research will include descriptive statistics, correlation matrix and linear regression. The program used to compute statistical methods is R.

		+	+	+	+	-	-

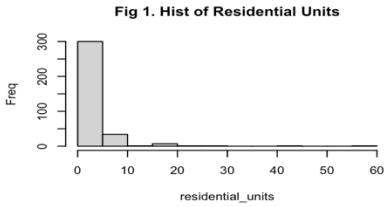
Eqn. 1 Price = f(ResidentialUnits,TotalUnits,LandSqFt,GrossSqFt,TaxClass,Age)

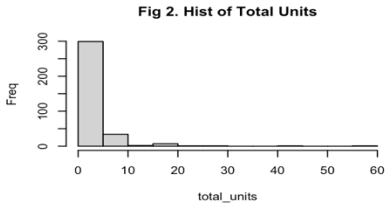
Eqn. 2 $Price = \alpha + \beta_{ru}*RU + \beta_{tu}*TU + \beta_{LSF}*LSF + \beta_{GSF}*GSF + \beta_{TC}*TC + \beta_{Age}*Age$

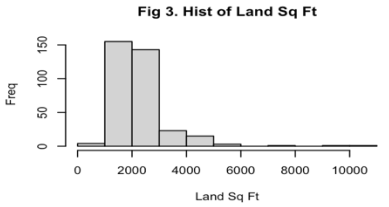
Eqn. 3 $Price = a + b_{ru}*RU + b_{tu}*TU + b_{lsf}*LSF + b_{gsf}*GSF + b_{tc}*TC + b_{age}*Age$

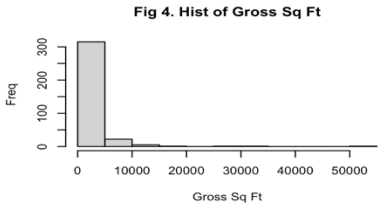
IV. Results

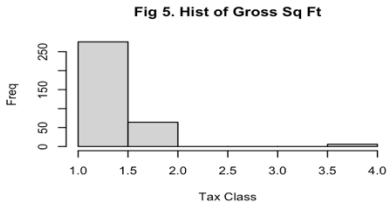
Figures 1-7 are histograms. Fig 1 is a histogram of residential units. Fig 2-7 are histograms of each independent variable. Figs 1-5, and 7 skew to the right. Fig. 6 seems symmetric, there is a slight left skew.

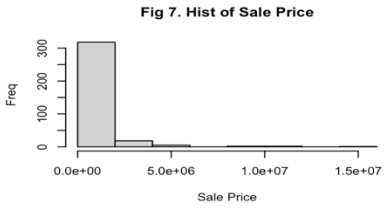




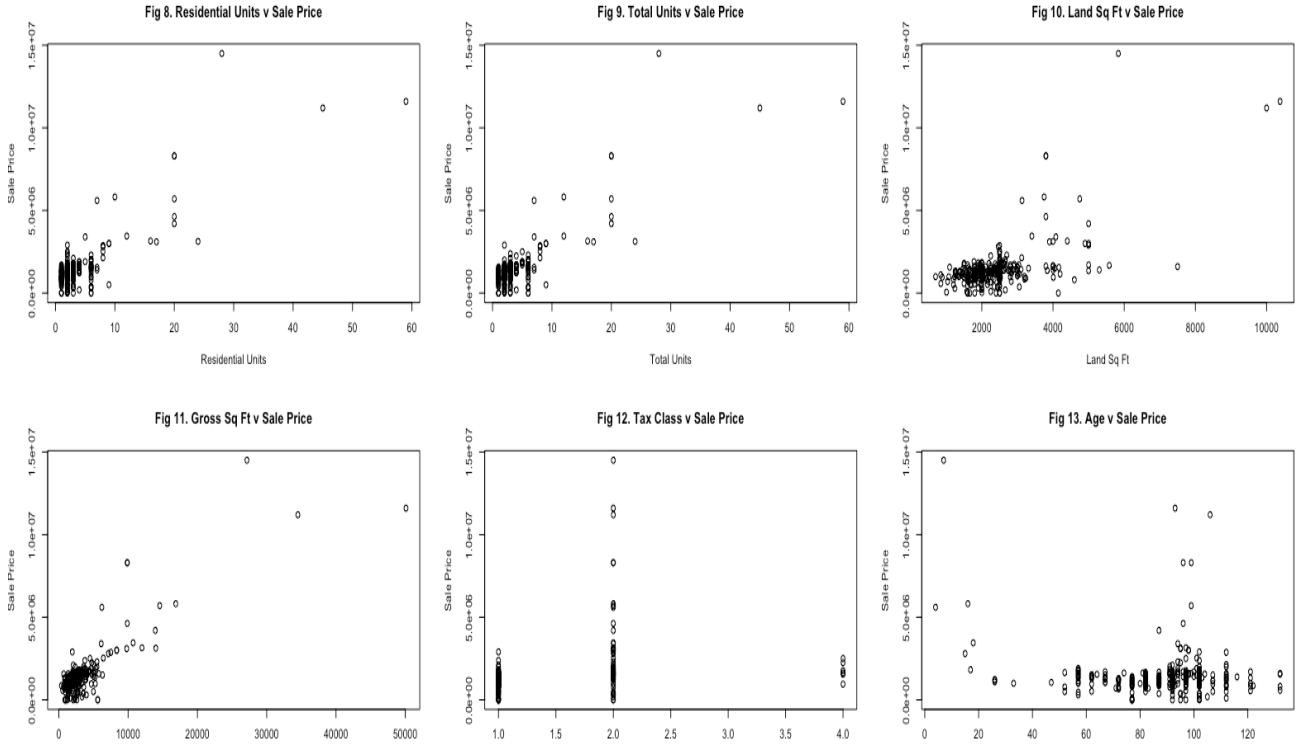








Figs 8-13 are scatterplots. The dependent variable or x-axis for all the graphs is the sale price. All figures demonstrate relationship between independent variables and sale price. Figs 8 – 10 Positive direction; weak. **Fig 11:** Strong positive. **Fig 12-13:** Weak Negative



Descriptive statistics are shown in table 1.

Table 1 **Descriptive Statistics**

	N	Mean	median	StdDev	Skewness	Kurtosis
residential_units	346	3	2	5	7	61
Total_units	346	3	2	5	7	60
Land_sq_ft	346	2384	2164	1036	3	23
Gross_sq_ft	346	2985	2102	3941	8	77
Tax_class	346	1	1	1	3	13
Age	346	88	92	19	-1	7
Sale_price	346	1437571	1250000	1370460	6	45

A correlation matrix is displayed in table 2.

Table 2 **Correlation Matrix**

	residential_units	total_units	land_sq_ft	gross_sq_ft	tax_class	age	sale_price
residential_units	1.00	1.00	0.68	0.96	0.38	-0.03	0.83
total_units	1.00	1.00	0.68	0.96	0.42	-0.02	0.84
land_sq_ft	0.68	0.68	1.00	0.71	0.23	0.07	0.64
gross_sq_ft	0.96	0.96	0.71	1.00	0.40	-0.09	0.85
tax_class	0.38	0.42	0.23	0.40	1.00	0.10	0.33
age	-0.03	-0.02	0.07	-0.09	0.10	1.00	-0.16
sale_price	0.83	0.84	0.64	0.85	0.33	-0.16	1.00

Multicollinearity seems to be present in the data. Specifically, between gross square feet and the total number of units, a correlation of 0.96 is observed. Overall, variables agree with the hypothesis. Multicollinearity is minimal.

Regression results are displayed in table 3.

Table 3 **Regression Results**

Eqn. 4: P = 1242782.7 + -365162.97*RU + 501064.24*TU + 155.87*LSF + 101.01*GSF - 158290.94*TC -8868.88*Age

```
Call:
lm(formula = sale_price ~ residential_units + total_units + land_sq_ft +
    gross_sq_ft + tax_class + age, data = ast_df, na.action = na.omit)

Residuals:
     Min       1Q   Median       3Q      Max
-3197936  -178713    27520   237744  6179232

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1242782.72  213143.93   5.831 1.29e-08 ***
residential_units -365162.97  125126.56  -2.918 0.003754 **
total_units        501064.24  137508.14   3.644 0.000311 ***
land_sq_ft            155.87      51.84   3.007 0.002839 **
gross_sq_ft           101.01      39.68   2.545 0.011358 *
tax_class         -158290.94   92803.39  -1.706 0.088988 .
age                 -8868.88    2074.25  -4.276 2.48e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 684900 on 339 degrees of freedom
Multiple R-squared:  0.7546,    Adjusted R-squared:  0.7502
F-statistic: 173.7 on 6 and 339 DF,  p-value: < 2.2e-16
```

For every unit increase in residential units, sale price decreases by $365,166. For every unit increase in the total units, sale price increases by $501,064. For every foot increase in the land square feet, the sale price increases by $150. For every foot increase in the gross square feet, the sale price increases by $101. For one unit increase in the tax class, the sale price decreases by $158,290. As the age of the home increases yearly, the sale price decreases by $8868.

The F-stat of the whole equation is 173.7, rendering the equation significant. The coefficient of determination is 0.75. Therefore, 75% of the variation in home prices in Astoria can be attributed to the number of residential units, the number of total units, the land square feet, the gross square feet, the tax class, and the home's age. Each of the regression coefficients were statistically significant.

V. Conclusion

The conducted research was quite successful. There is statistical evidence to support that significant percentage of the variability in the prices of homes in Astoria can be explained by these 6 variables. Potential home buyers in Astoria can use this model to estimate the price of their home depending on these features. Real estate companies can use this model to aid in their asset purchases. The research can be further improved with the addition of more observations and or inclusion of other features such as the number of bathrooms, bedrooms etc.

VI. Appendix I

Residential units	total_units	l_sq_ft	gross_sq_ft	tax_class	age	sale_price
1	1	3,197	1,212	1	102	999,000
1	1	1,602	1,224	1	77	0
1	1	1,566	1,224	1	77	1,238,000
1	1	1,674	1,224	1	77	843,000
1	1	1,692	1,224	1	77	874,888
1	1	2,300	1,850	1	77	0
1	2	5,000	4,400	1	67	1,700,000
1	1	3,007	1,120	1	121	0
1	2	2,425	2,560	1	92	0
1	2	1,600	1,760	1	92	0
1	1	2,500	1,700	1	82	999,000
1	1	2,000	1,296	1	102	0
1	1	2,500	1,035	1	121	995,000
1	1	1,815	1,362	1	77	0
1	1	2,500	1,358	1	112	860,000
1	1	2,500	1,200	1	112	0
1	1	1,996	1,395	1	82	1,250,000
1	1	1,211	1,375	1	82	0
1	1	1,996	1,395	1	82	1,050,000
1	1	2,000	1,585	1	85	0
1	1	1,800	1,937	1	82	950,000
1	1	1,600	480	1	112	0
1	1	2,500	920	1	102	925,000
1	1	2,500	956	1	112	0
1	1	2,000	1,560	1	92	1,055,000
1	1	2,000	2,290	1	82	1,100,000
1	1	2,000	1,320	1	92	0
1	1	2,000	1,320	1	92	0
1	1	1,900	802	1	71	710,000
1	1	2,525	1,200	1	107	950,000
1	1	2,500	1,408	1	102	815,000
1	1	2,100	1,389	1	82	900,000
1	1	3,100	612	1	121	0
1	1	2,260	2,095	1	77	1,180,000
1	1	1,045	1,320	1	67	0
1	1	2,000	1,564	1	82	875,000
1	1	1,620	1,566	1	83	700,000
1	1	2,000	1,962	1	77	1,100,000

VII. Appendix II

```
options(max.print=1000000)

data.class(astoriaprices)
astoria_df<-as.data.frame(astoriaprices)
data.class(astoria_df)
dim(astoria_df)

head(astoria_df)


head(astoria_df)
astoria_df[astoria_df==0] <- NA

head(astoria_df)
dim(astoria_df)
ast_df <- na.omit(astoria_df)

#na.omit(ast_df)

head(ast_df)

dim(ast_df)
dim(ast_df)

#Histogram
par(mfrow=c(3,3))
hist(ast_df$residential_units, main = "Fig 1. Hist of Residential Units", xlab = "residential_units",
ylab = "Freq") # display histograms
hist(ast_df$total_units, main = "Fig 2. Hist of Total Units", xlab = "total_units", ylab = "Freq")
hist(ast_df$land_sq_ft, main = "Fig 3. Hist of Land Sq Ft", xlab = "Land Sq Ft", ylab = "Freq")
hist(ast_df$gross_sq_ft, main = "Fig 4. Hist of Gross Sq Ft", xlab = "Gross Sq Ft", ylab = "Freq")
hist(ast_df$tax_class, main = "Fig 5. Hist of Gross Sq Ft", xlab = "Tax Class", ylab = "Freq")
hist(ast_df$age, main = "Fig 6. Hist of Gross Sq Ft", xlab = "Age", ylab = "Freq")
hist(ast_df$sale_price, main = "Fig 7. Hist of Sale Price", xlab = "Sale Price", ylab = "Freq")

#Scatterplot
par(mfrow=c(3,3))
plot(ast_df$residential_units,ast_df$sale_price, main = "Fig 8. Residential Units v Sale
Price",xlab = "Residential Units", ylab = "Sale Price")
plot(ast_df$total_units,ast_df$sale_price, main = "Fig 9. Total Units v Sale Price",xlab = "Total
Units", ylab = "Sale Price")
```

```
plot(ast_df$land_sq_ft,ast_df$sale_price, main = "Fig 10. Land Sq Ft v Sale Price",xlab = "Land Sq Ft", ylab = "Sale Price")
plot(ast_df$gross_sq_ft,ast_df$sale_price, main = "Fig 11. Gross Sq Ft v Sale Price",xlab = "Gross Sq Ft", ylab = "Sale Price")
plot(ast_df$tax_class,ast_df$sale_price, main = "Fig 12. Tax Class v Sale Price",xlab = "Tax Class", ylab = "Sale Price")
plot(ast_df$age,ast_df$sale_price, main = "Fig 13. Age v Sale Price",xlab = "Age", ylab = "Sale Price")

#Descriptive Statistics
summary(ast_df)
install.packages("YRmisc")
library("YRmisc")
ds.summ(ast_df)

#Correlation Matrix
round(cor(ast_df),2)

#Regression
fit <-
lm(sale_price~residential_units+total_units+land_sq_ft+gross_sq_ft+tax_class+age,data=ast_df,na.action=na.omit)
fit$coefficients
fit$coefficients[1]

summary(fit)
summary(fit)$r.squared

fit$fitted.values
fit$residuals

#Residuals
v<-data.frame(ast_df,p=fit$fitted.values,r=fit$residuals)
par(mfrow=c(1,1))
hist(v$r, main = "Hist of Residuals")

plot(v$p,v$sale_price)
```