Brooklyn Housing Predictions

Prabhu Avula and Team

Introduction

What’s this About?

1

"Insert inspirational New York City quote because no one thinks of Brooklyn first; and because this quote slide was too pretty to delete from the theme, so we're keeping it!"

What's this About?

Introduction to Brooklyn's Housing Market:

- Notable for its dynamic nature and significance within the larger New York City real estate landscape.
- Subject to influences from demographic shifts, economic trends, and urban development projects.

Research Objective:

- Unravel the complex web of variables that have shaped housing prices in Brooklyn through rigorous data analysis and interpretation.

Goal

What's the objective?

2

What are we trying to do?

- Brooklyn's housing effect in the NYC metropolitan area.
- Enhance the ability to predict housing prices in brooklyn.
- Analyze variables impacts on market fluctuations
- Visualize these results using graphs, plots, etc.
- Acknowledge the potential for this goal to evolve in the future.

Methodology

What's the plan?

3

What's the Plan?

Objective:

- ❑ Apply linear regression to analyze the housing market from 2016-2020 (Pre-Covid).

Data Segmentation:

- ❑ Divide each year into quarters (Q1-Q4) for detailed analysis.
- ❑ Explore quarterly changes to understand socio-economic impacts.

Data Handling:

- ❑ Perform data split for training and validation.
- ❑ Allocate a major portion for linear regression model training.

The Plan - Part 2

Data Processing:

- ❑ Implement data cleaning techniques.
- ❑ Exclude unnecessary columns by verifying formats and removing anomalies.

Feature Engineering:

- ❑ Conduct feature engineering using linear regression model.
- ❑ Enhance the model's predictive capabilities for housing market analysis.

Visualization:

- ❑ Utilize graphs, boxplots, histograms, etc., for visual representation.
- ❑ Illustrate patterns and trends to enhance data interpretation.

TIMELINE - YEARS IN FOCUS

2016

First year of focus for this project. A lot of economic growth about to kickstart in the area.

2017

Continued economic growth, slow changes in the real estate sector, perhaps because the nets are terrible.

2018

Nets, still terrible. Continued growth. Fresh influx of companies, students and money.

2019-20

Coronavirus happened. Didn't need to add another point in the timeline for an extra year. At least the nets caught a break.

OUR PROCESS IS EASY, LITERALLY

Analyse & Process

Simple. We go through the data and get a good sense of it. Then, we change it to our requirements.

Train & Validate

Apply multiple models to see which one best aids our progress in this project. Then, select that model.

Visualize & Conclude

Visualize results using graphs, plots, etc just so it is easier for everybody to understand.

Data Analysis

4

Data Analysis

Dataset Preparation (2016-2020):

- ❑ Meticulous preprocessing of Brooklyn housing data.
- ❑ Removal of redundancies and alignment with project objectives.
- ❑ Enhanced interpretability through column name revisions.

Data Refinement:

- ❑ Elimination of null values and optimization of data types.
- ❑ Conversion of string values to numeric types.
- ❑ Removal of commas for numerical consistency.

Dataframes with observations and features

Data	
df1	25523 obs. of 21 variables
df2	24796 obs. of 21 variables
df3	23669 obs. of 21 variables
df4	23669 obs. of 21 variables
df5	21717 obs. of 21 variables
Functions	
read_return_df	function (path)

Data Analysis

Temporal Analysis:

- Restructuring dates using the date class.
- Quarter-wise segmentation for detailed trend exploration.

Consolidation and Outcome:

- Integration of data frames for a comprehensive dataset.
- Reveals insights into five-year housing trends in Brooklyn.

Data Processing

What do we need?

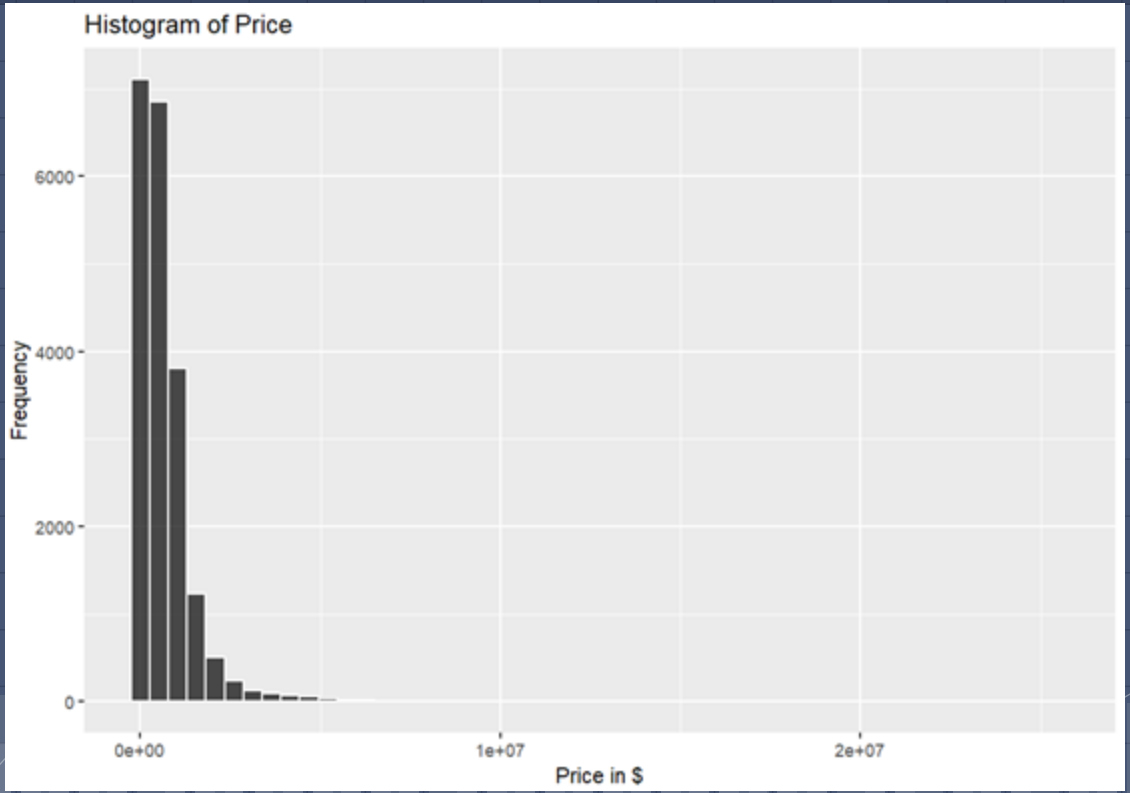
5

Data Processing:

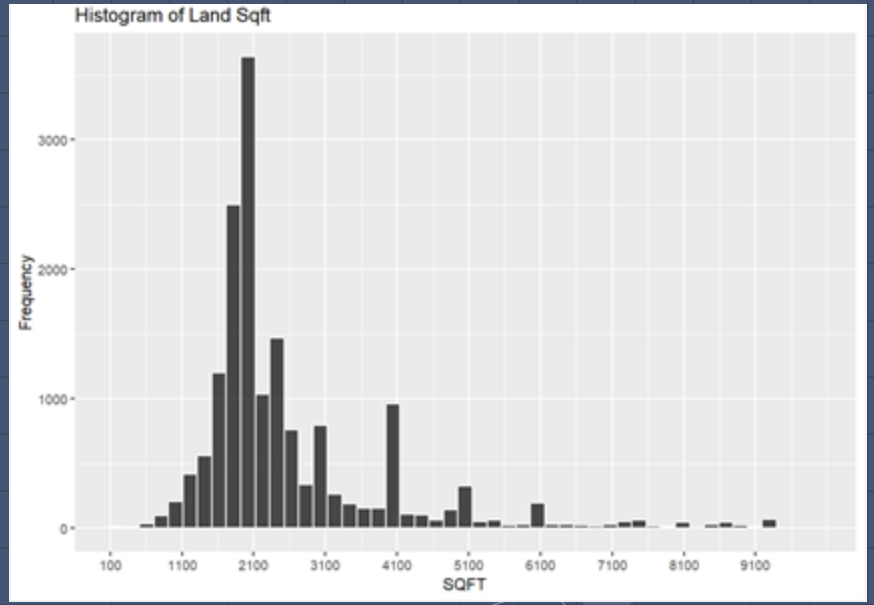
Steps involved in the data processing:

- ❑ Data Collection
- ❑ Data Cleaning
- ❑ Feature engineering
- ❑ Data Transformation
- ❑ Model Training [Subsequent Section]
- ❑ Model Evaluation [Subsequent Section]

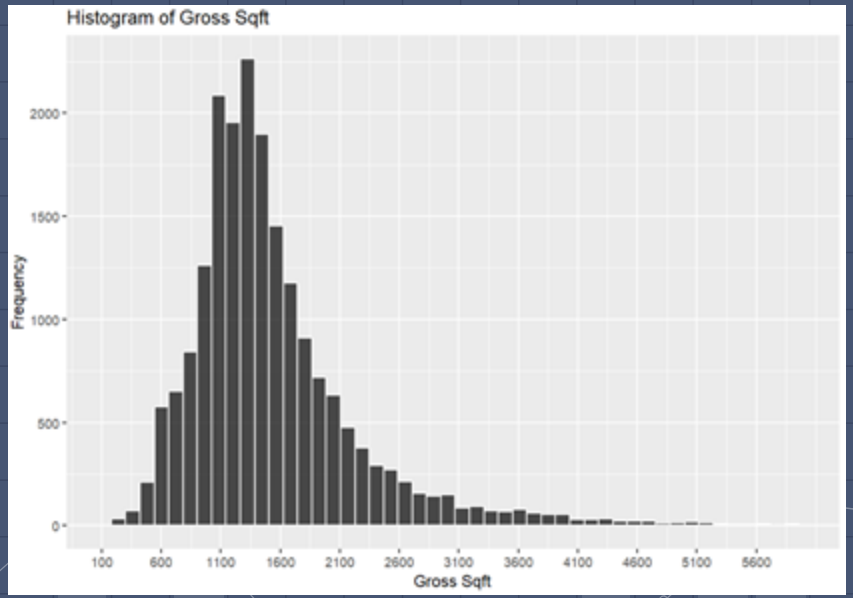
Histogram on Price vs Frequency



Histogram on Sqft vs Frequency



Histogram on Gross Sqft vs Frequency



119,374

Number of observations in the resultant data frame that we did not test on because that was before final processing. Although, big numbers are cool to look at.

119,374

Observations before final processing.

20,185

Observations after final processing based on constraints.

100%

Tested on those 20 grand!

Model Training

How do we get what we want?

6

Model Training

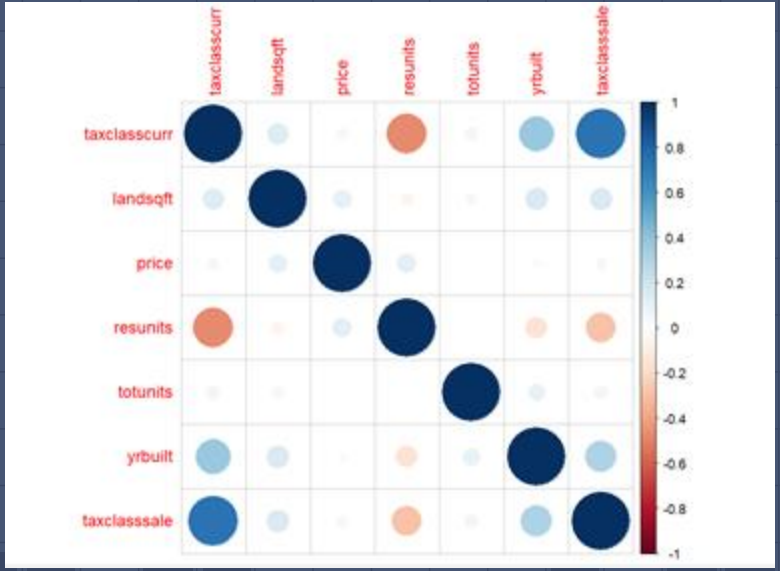
Data Analysis and Preprocessing:

- ❑ Completed data analysis and preprocessing phase.
- ❑ Linear regression model is utilized.

Model Training and Comparison:

- ❑ Employment of four different models for training.
- ❑ A thorough comparison of models.

Correlation of features:



	taxclasscurr	landsqft	price	resunits	totunits	yrbuilt	taxclasssale
taxclasscurr	1.00000000	0.13593808	0.0452871801	-0.479745999	0.0566112622	0.37686996	0.73691132
landsqft	0.13593808	1.00000000	0.1067544908	-0.050763909	0.0436298232	0.15552551	0.15849301
price	0.04528718	0.10675449	1.0000000000	0.110065485	-0.0005511958	0.02380221	0.04234858
resunits	-0.47974600	-0.05076391	0.1100654850	1.000000000	0.0066524887	-0.14601830	-0.28157367
totunits	0.05661126	0.04362982	-0.0005511958	0.006652489	1.0000000000	0.09256420	0.05778891
yrbuilt	0.37686996	0.15552551	0.0238022128	-0.146018300	0.0925641964	1.00000000	0.30233786
taxclasssale	0.73691132	0.15849301	0.0423485765	-0.281573673	0.0577889064	0.30233786	1.00000000

Model Training

Evaluation Metrics:

- Evaluation of models using R-Squared, RMSE values,and degrees of freedom.
- Considering multiple metrics for comprehensive assessment.

Model Selection and Decision:

- The model that best suited project objectives is chosen among 4 models.
- Decision based on performance in terms of R-Squared, RMSE, and model complexity.

Model Validation

Which one’s the best?

7

Model Validation

- Model validation is a crucial step in assessing the performance and generalization capability of a predictive model, such as one used for Brooklyn housing pricing prediction.
- By rigorously validating the model, you ensure that it provides reliable predictions on new, unseen data, enhancing its utility in making accurate housing price predictions in Brooklyn.

This is how the Models were tested by Linear Regression:

```
#Starting the models
model1 <- lm(price ~ bldclasscat + bldclasssale + grosssqft + yrbuilt + quarter + zip_rk, data =
fullDf)
summary(model1)
############### r^2 = 0.5518 , Degrees of Freedom = 43 #####################

#RMSE
sqrt(mean(model1$residuals^2))
######### RMSE = 497361.6 ##################

#model2
model2 <- lm(price ~ bldclasscat + log(landsqft) + sqrt(grosssqft )*zip_rk + yrbuilt + quarter , data =
fullDf)
summary(model2)
##############  r^2 = 0.6146 , Degrees of Freedom = 39 ###################

#RMSE model2
sqrt(mean(model2$residuals^2))
################ RMSE = 460563.2 ############

#model3 square root of price
model3 <- lm(sqrt(price) ~ bldclasscat + yrbuilt + landsqft + grosssqft + quarter + zip_rk, data =
fullDf)
summary(model3)
############## r^2 = 0.5774 , Degrees of Freedom = 35 #########

#RMSE model 3
sqrt(mean((fullDf$price - model3$fitted.values^2)^2))
################ RMSE = 480768.5 ############

#final model, add interaction between zip_rk and grosssqft
final_model <- lm(sqrt(price) ~ bldclasscat + yrbuilt + quarter + sqrt(grosssqft)*(zip_rk)
+log(landsqft), data = fullDf)
summary(final_model)
############## r^2 = 0.6232 , Degrees of Freedom = 39 #############

#RMSE final model
sqrt(mean((fullDf$price - final_model$fitted.values^2)^2))
############ RMSE = 460493 ###############
```

Model Selection and Validation

We chose to go with the final model we trained because of:

- Subtle differences in R-Squared, RMSE values and Degrees of Freedom.
- Felt this model was best suited for our progress.

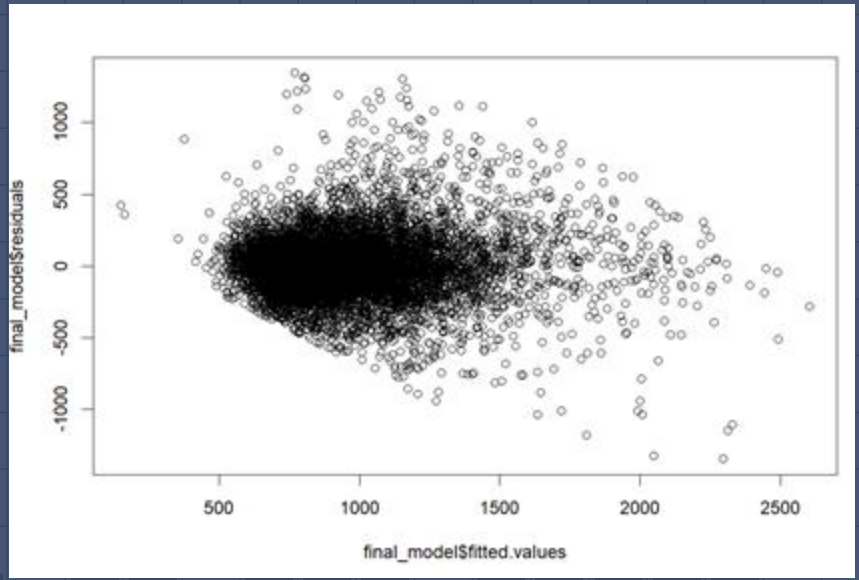
Conclusion

What does it conclude?

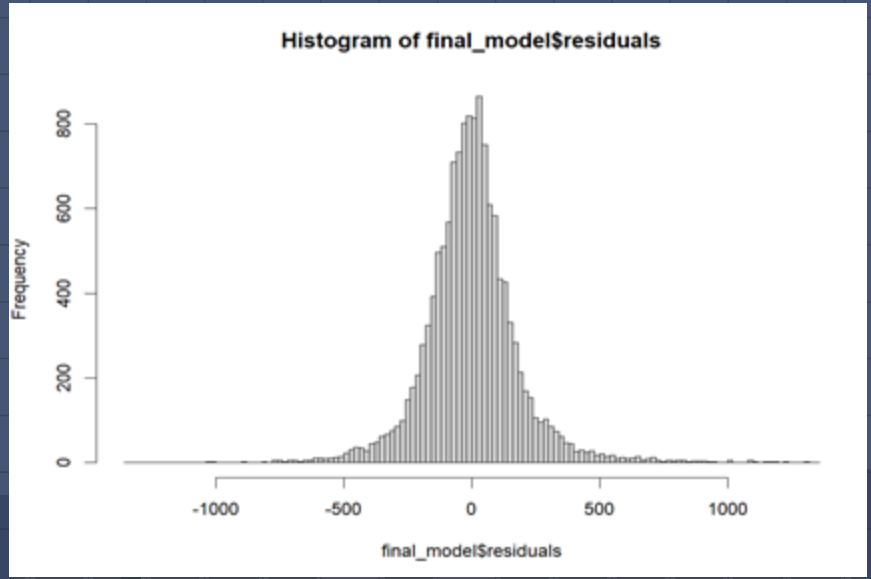
8

CONCLUSION

- **Objective:** Analyzing Brooklyn housing prices from 2016 to 2020 using linear regression.
- **Impact :** Informing legislators, investors, and urban planners for sustainable development.
- **Contribution:** Enriching urban economics discussions with insights into future price trends.
- **Prediction:** Applying a robust model for interpreting factors influencing market movements.
- **Narrative :** Dissecting Brooklyn's history to offer a dynamic perspective for decision-makers in the evolving real estate landscape.



Final Model Residuals vs Fitted Values



Histogram of Final Model Residuals

THANKS!