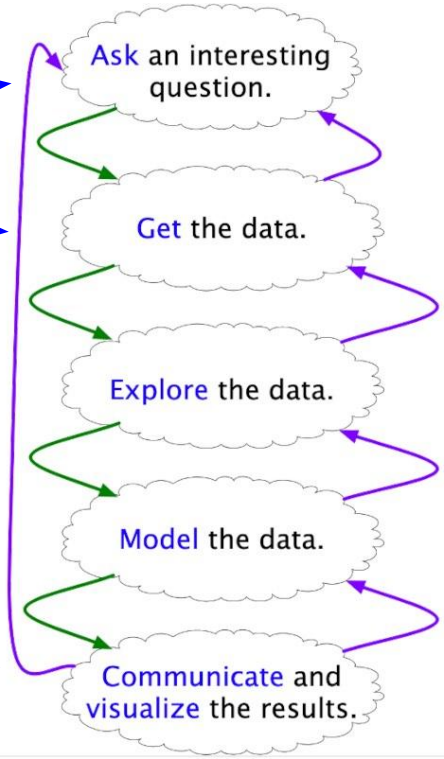
Housing Inventory Analysis

Tom Stuckey

Outline

- Problem Statement
- Data Gathering
- Data Exploration
- Data Modeling
- Summary Results

Guiding Analysis Framework

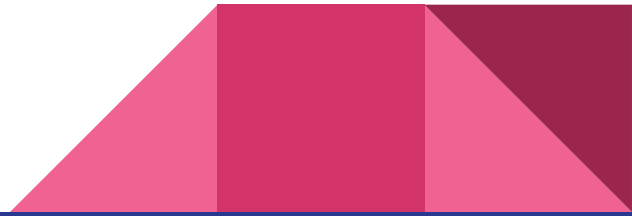


Data Science Process:
https://github.com/cs109/2015/blob/master/Lectures/01-Introduction.pdf

Problem Statement

Problem Statement

- Contemporarily, in Dec ‘21, housing inventory is a major topic of discussion. Prices are high, supplies (both completed housing and construction supplies) are short, and frustrations are plentiful (unless, perhaps, you are real-estate agent)
- Looking across recent history, can we explain housing inventory as a function of several other inputs?



Problem Statement

- Factors that influence housing inventory are numerous and include:
  - Mortgage rates
  - Employment rates
  - Employment trends (e.g. remote vs. in-person)
  - Planned housing construction
  - Federal Funds rates
  - Consumer debt/income ratios

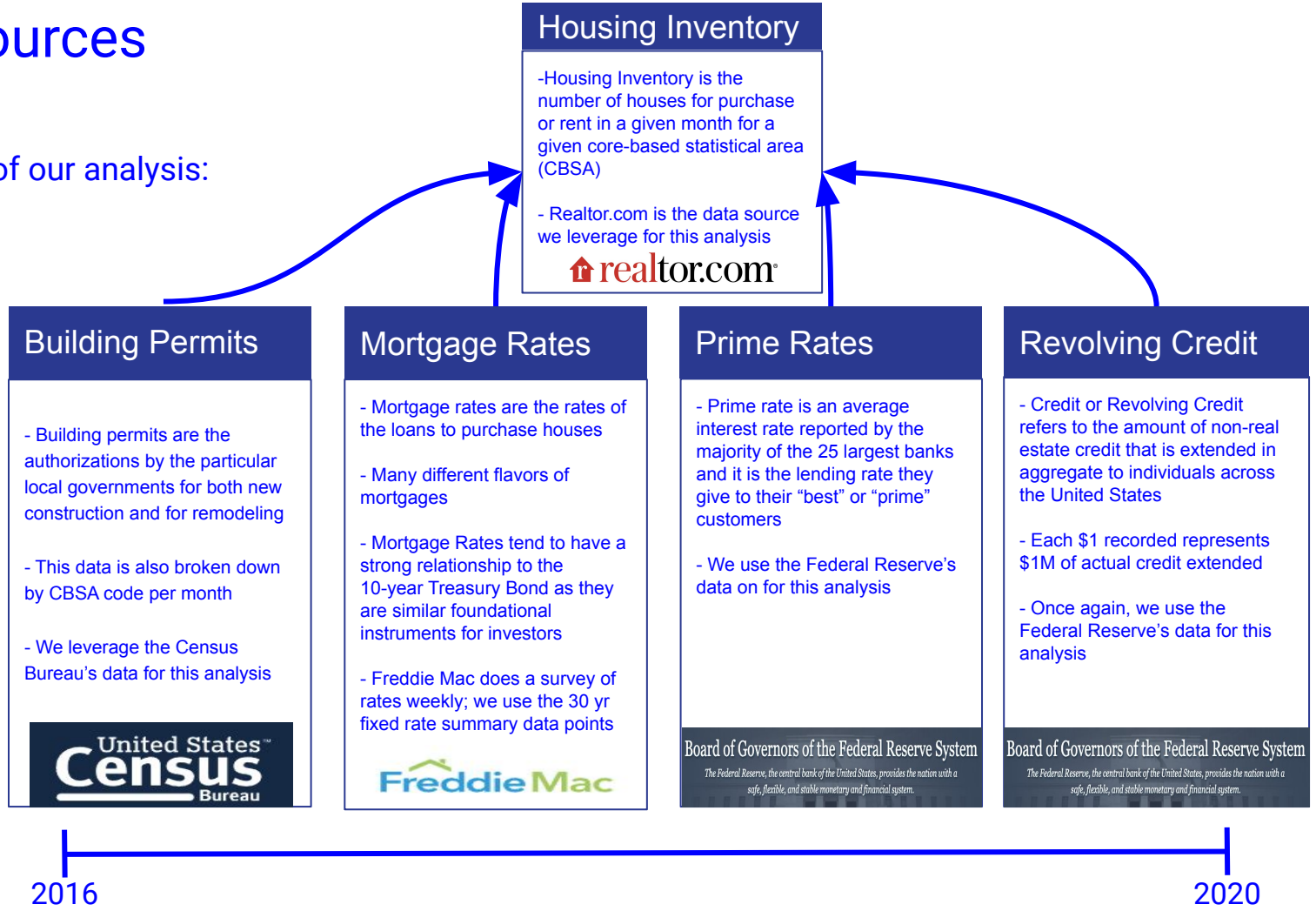
- What would we do if we had all the data?
  - Build a precise and accurate model, of course!
  - Make beaucoup

  - Develop more efficient governance
  - Predicting housing inventory could facilitate:
    - policy decisions
    - lending rates
    - employment offers
    - tax revenue forecast
    - general optimization across the whole housing supply chain

Data Gathering

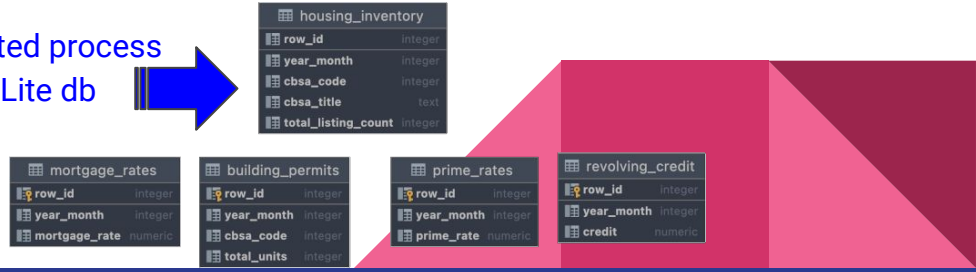
Data Sources

Scope of our analysis:



Extract, Transform, and Load

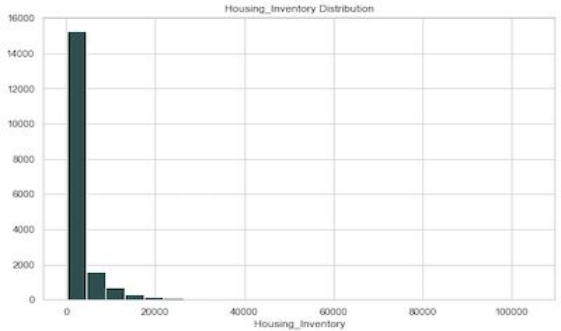
- Time challenges
  - Time is essentially a key value into into each observation, but we are focused on general explanation of housing inventory, not time series analysis
  - Time is not uniform across all the datasets; some are provided in weekly increments, others are in monthly or yearly increments
- Time Solutions
  - Bound everything to 2016-2020 (49 total months observed)
  - Convert all observations such that there is a composite key of *year+month*
  - Aggregate weekly observations into monthly observations (average values)
  - Disaggregate annual observations into monthly observations (divide values)
  - Drop the time key from the analysis (not doing time series)
- General challenges:
  - Five different classes of data sources, nine discrete data files, two types of files (txt and spreadsheets)
  - Not all data lines up
- General solutions:
  - Convert tabular text data -> Excel with semi-automated process
  - Standardize on Excel-base import routines into a SQLite db
  - Leverage runtime CTE with INNER JOINs
  - 18,000+ observations

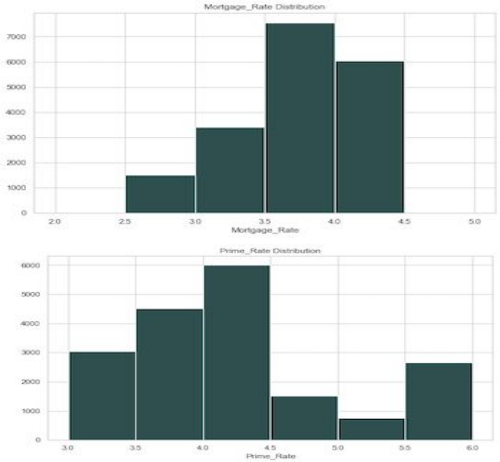


Data Exploration

Exploratory Data Analysis

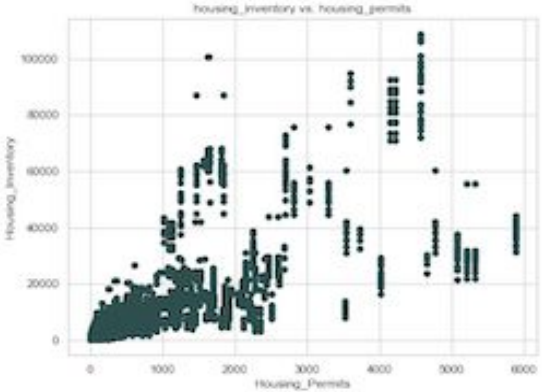
- Single Variable Analysis (key items)

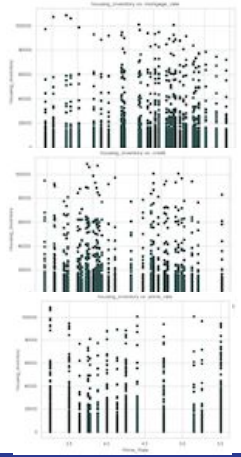




Total CBSA codes	935
Unique CBSA codes in housing inventory specific data	917
Unique CBSA codes in unified data for analysis	383

- Paired Variable Analysis Inventory v. * (key items)





** See notebooks on GitHub for in-depth analysis*

CBSA Title
new york-newark-jersey city, ny-nj-pa
chicago-naperville-elgin, il-in-wi
miami-fort lauderdale-west palm beach, fl

Data Modeling

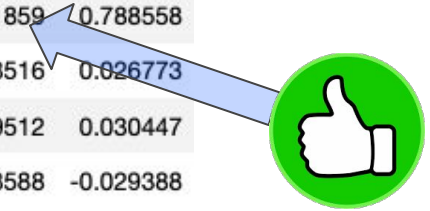
Data Modeling

- CLD (table)

Variable Name	Expected CLD sign to Housing Inventory	comment
housing permits	positive	More building permits mean more higher inventory
mortgage rate	positive	Higher mortgage rates should yield higher inventory
credit	negative	Higher credit showed a slight decrease in inventory
prime rate	positive	Slight increase in prime rate increased inventory
cbsa code	N/A	categorical variable for car origin

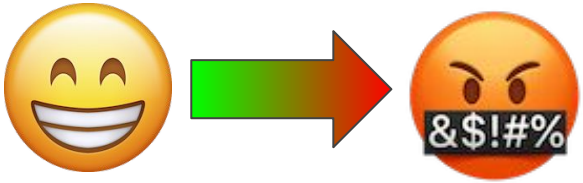
- Correlations

	feature	r	rho
0	housing_permits	0.771859	0.788558
1	mortgage_rate	0.013516	0.026773
2	prime_rate	0.009512	0.030447
3	credit	-0.003588	-0.029388



Data Modeling

- Null Model
  - Null model is our basis to compare models; key metrics:
    - Mean: 3634.16 houses
    - Error: 8178.64 houses
  - Null model with 95% error bounds:
    - Theoretical: -12395.54 <= μ <= 10757.11
    - Actual: 0 <= μ <= 10757.11
- Linear Model 1(All-in Model)
  - housing_inventory ~ housing_permits + mortgage_rate + revolving_credit + prime_rate + (383 - 1) one-hot encoded cbsa_codes
  - Mean metrics:
    - Error: 1483.35 houses
    - R^2: 0.97

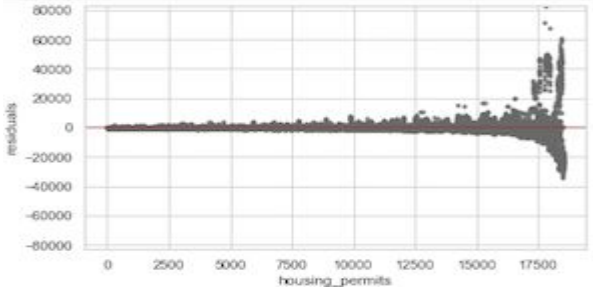


Great Model….but….it's too precise for general use!

Data Modeling

- Linear Model 2 (dropped cbsa codes)
  - housing_inventory ~ housing_permits + mortgage_rate + revolving_credit + prime_rate
  - Mean Metrics:
    - Error: 5194.67 houses
    - R^2: 0.60

Residual Analysis of of housing_permits indicates heteroscedasticity



Null Model:
Mean: 3634 houses
Error: 8179 houses

- Linear Model 3
  - housing_inventory ~ housing_permits + mortgage rate + revolving credit + prime rate + mortgage rate : prime rate
  - Mean Metrics:
    - Error: 5190.89 houses
    - R^2: 0.60

** See notebooks on GitHub for in-depth analysis*

Data Modeling

- Linear Model 4
  - housing_inventory ~ housing permits + mortgage_rate + revolving_credit + prime_rate + mortgage_rate:prime+rate + mortgage_rate:revolving credit + prime_rate:revolving_credit + mortgage_rate:revolving_credit
  - Mean Metrics:
    - Error: 5179.79 houses
    - R^2: 0.60

Null Model:
Mean: 3634 houses
Error: 8179 houses

- Linear Model 5
  - housing_inventory ~ lg(housing permits) + mortgage_rate + revolving_credit + prime_rate + mortgage_rate:prime+rate + mortgage_rate:revolving credit + prime_rate:revolving_credit + mortgage_rate:revolving_credit
  - Mean Metrics:
    - Error: 6770.86 houses
    - R^2: 0.31

** See notebooks on GitHub for in-depth analysis*

Final Model is Linear Model 4

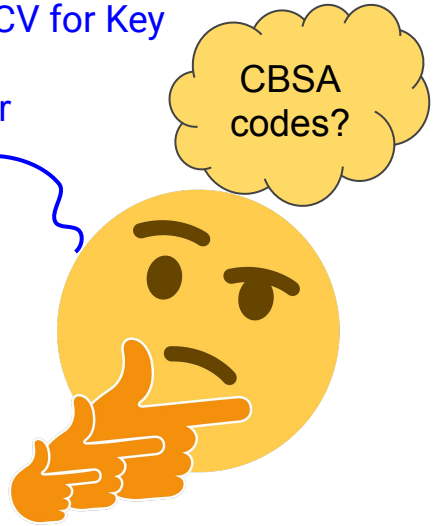
$$\hat{y} = 79342.25 + 10.37\beta_1 - 29330.08\beta_2 - 0.11\beta_3 + 17737.31\beta_4 - 2339.83\beta_5 + 0.04\beta_6 - 0.01\beta_7$$

		95% BCI	
Coefficients	**Mean**	**Lo**	**Hi**
	β_0	79342.25	-232330.92	270834.68
housing_permits	β_1	10.37	9.91	10.91
mortgage_rate	β_2	-29330.08	-79011.65	52656.40
credit	β_3	-0.11	-0.31	0.20
prime_rate	β_4	17737.31	-29320.82	95247.43
mortgage_rate:prime_rate	β_5	-2339.83	-22713.09	9766.44
mortgage_rate:credit	β_6	0.04	-0.04	0.09
prime_rate:credit	β_7	-0.01	-0.09	0.04
mortgage_rate:prime_rate:credit	β_8	0.00	-0.01	0.02

Metrics	**Mean**	**Lo**	**Hi**
σ	5179.79	4863.31	5487.70
R^2	0.60	0.58	0.62

95% Credible Interval from CV for Key Metrics:

- (4349 - 6031) for error
- (0.55 - 0.65) for R^2



Summary Results

Predictions

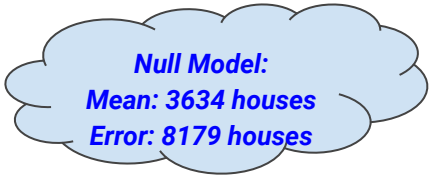
- Prediction 1: something reasonable with 5,000 building permits, a mortgage rate of 2.0%, revolving credit of 750,000 millions, and a prime rate of 4.0%
  - Linear model predicts 45,835 houses
  - With a 95% error bounds of 29,805 houses - 61,865 houses
  - Intuitively reasonable

Null Model:
Mean: 3634 houses
Error: 8179 houses

- Prediction 2: well beyond the far range of our data with 15,000 building permits, a mortgage rate of 10.0%, 1,000,000,000 millions in revolving credit, and a prime-rate of 8.0%
  - Linear model predicts -879,355 houses
  - With a 95% error bounds of -895,385 houses - -863,326 houses
  - Strong NO on available housing inventory in this scenario

Predictions

- Prediction 3: minimum prediction with 1 building permit, a mortgage rate of 0.5%, 1 million in revolving credit, and a prime rate of 0% (banks are just literally giving money away)
    - Linear model predicts 77,886 houses
    - With a 95% error bounds of 61,856 houses - 93,916 houses
    - Interesting theoretical scenario with only a single building permit



- Prediction 4: seek to maximize inventory by bumping up the housing permits a bit by taking the situation in scenario 3 and making the housing permits 5,000 for the month (we would expect this to be pretty close our maximum value observed in the dataset)
    - Linear model predicts 129,735 houses
    - With a 95% error bounds of 113,706 houses - 145,765 houses
    - Increasing in the building permits in the favorable scenario jumped housing inventory 66%

Closing Thoughts

- Precise Model was too focused
- General Model was “ok” … at best
- What’s a better compromise for future analysis? Probably some abstraction of CBSAs into larger regions

Fork Us on GitHub!: https://github.com/tstuckey/housing-inventory

References

- Ayers, C. (2021, August 13). How A Low Housing Inventory Impacts The Real Estate Market. Rocket Mortgage. Retrieved October 15, 2021, from https://www.rocketmortgage.com/learn/low-housing-inventory
- Mcleod, S. (2020, December 29). Maslow's Hierarchy of Needs. Simply Psychology. Retrieved November 24, 2021, from https://www.simplypsychology.org/maslow.html
- Pfister, H. P., Blitzstein, J. B., & Kaynig, V. K. (2015, December 5). CS109 Data Science. Https://Github.Com/Cs109/. Retrieved November 24, 2021, from https://github.com/cs109/2015/blob/master/Lectures/01-Introduction.pdf
- https://www.realtor.com/research/data/
- https://www.census.gov/construction/bps/msaannual.html
- http://www.freddiemac.com/pmms/
- Investopedia. (2020, June 30). Prime Rate Definition. Retrieved November 25, 2021, from https://www.investopedia.com/terms/p/primerate.asp
- https://www.federalreserve.gov/datadownload/Download.aspx?rel=H15&series=6fa2b8138e0eafe0ad6cde24ba2307f5&from=01/01/2004&to=12/31/2020&lastObs=&filetype=spreadsheetml&label=include&layout=seriescolumn
- https://www.federalreserve.gov/datadownload/Download.aspx?rel=g19&series=3f01d9dcee01c5a0459b2ed2450bd7de&filetype=csv&label=include&layout=seriescolumn&from=01/01/2004&to=12/31/2020
- Balaban, J. (2018, August 27). When and How to use Weighted Least Squares (WLS) Models. Medium. Retrieved December 4, 2021, from https://towardsdatascience.com/when-and-how-to-use-weighted-least-squares-wls-models-a68808b1a89d
- Federal Bank St. Louis. (2021, December 1). FRED Economic Data. FRED Economic Data. Retrieved December 6, 2021, from https://fred.stlouisfed.org/series/FEDFUNDS