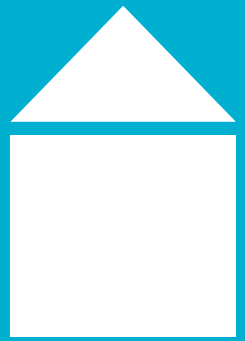
Zillow's Data Set (Group 1)

By: Roger, Naomi, Presley, Neftali, Michael, Luigi.

Abstract

Zillow is a website in which people can sell homes. But what makes a home sell faster? For example, are there certain characteristics in a description that make a home more favorable?

- Here, we examine the possible effect of several variables relating to Zillow postings on the time it takes to sell the home



Statement of the Problem

Our analysis focused on two questions:

1. **Spatial Analysis:** How do Zillow listings differ by space?
2. **Sentiment Analysis:** How do Zillow listings differ by their description?
3. **Prediction:** Can we reasonably guess how long a house will take to sell?

Dataset and Variables

Datasets: The Zillow Dataset

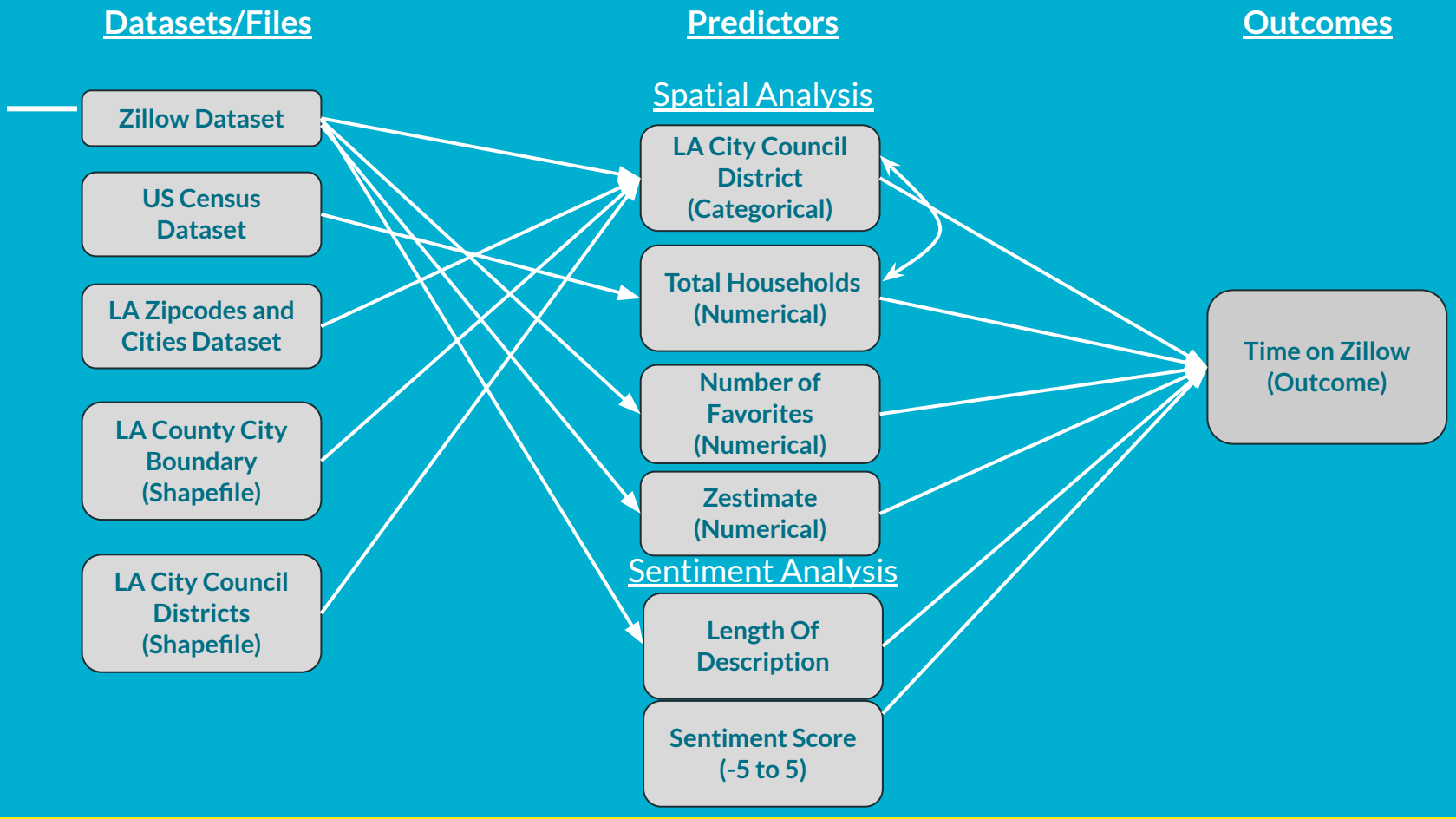
Overview

- **Before cleaning:** 34397 observations, 44 variables
- **After cleaning:** 32314 observations, 28 variables
  - Mutated variables, got rid of NAs
  - Initially was going to look at Foreclosed vs Non-foreclosed households but after cleaning there was only one foreclosed house left in the entire dataset!

Datasets: Other Datasets/Files Used

- Datasets
  - US Census Data (2018)
  - List of LA Zip Codes and Cities → Scraped from LA Almanac
- Shapefiles from LA Geohub (Open source platform by the City of LA)
  - LA County City Boundary Shapefile
  - LA City Council Districts Shapefile

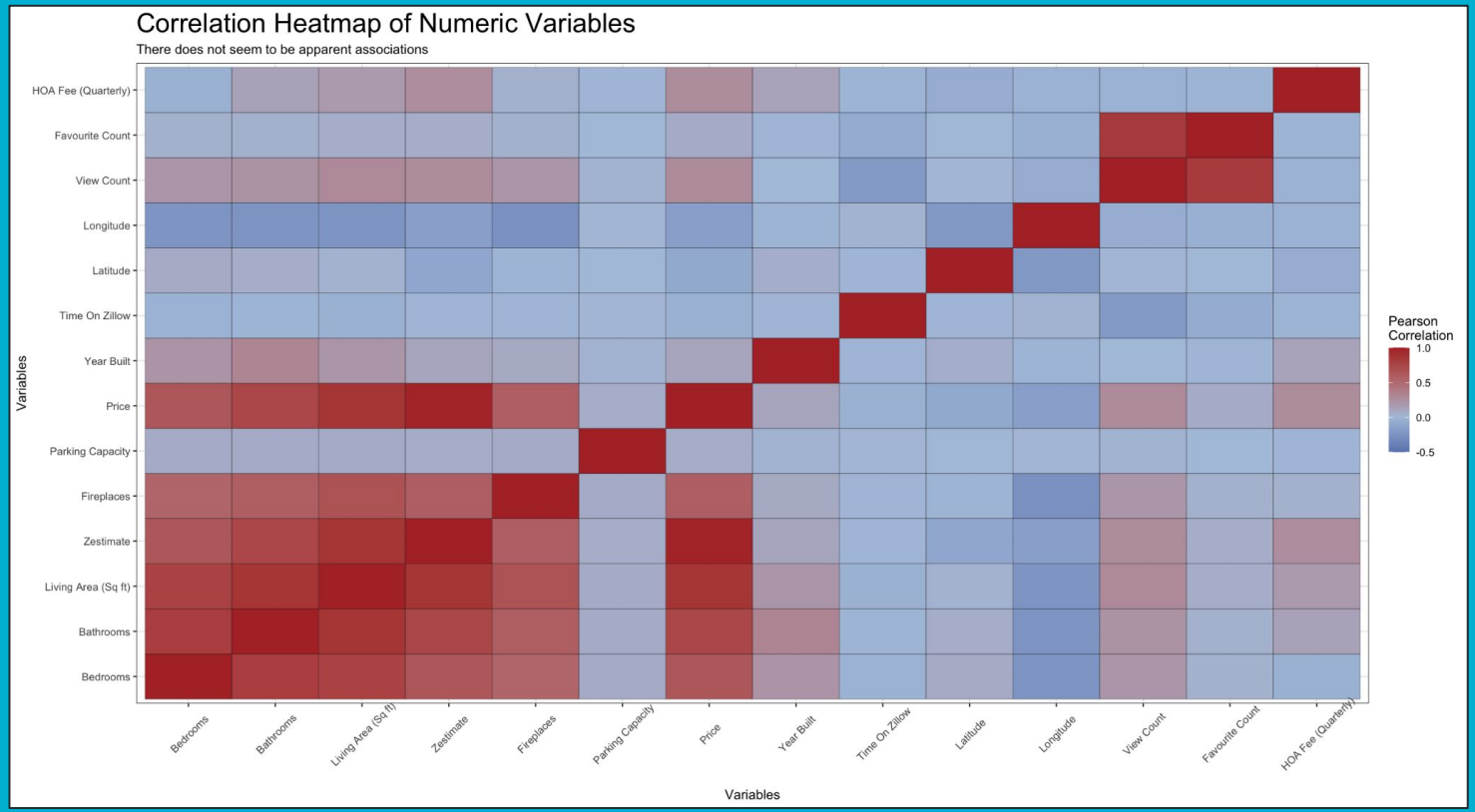
Schematic Diagram



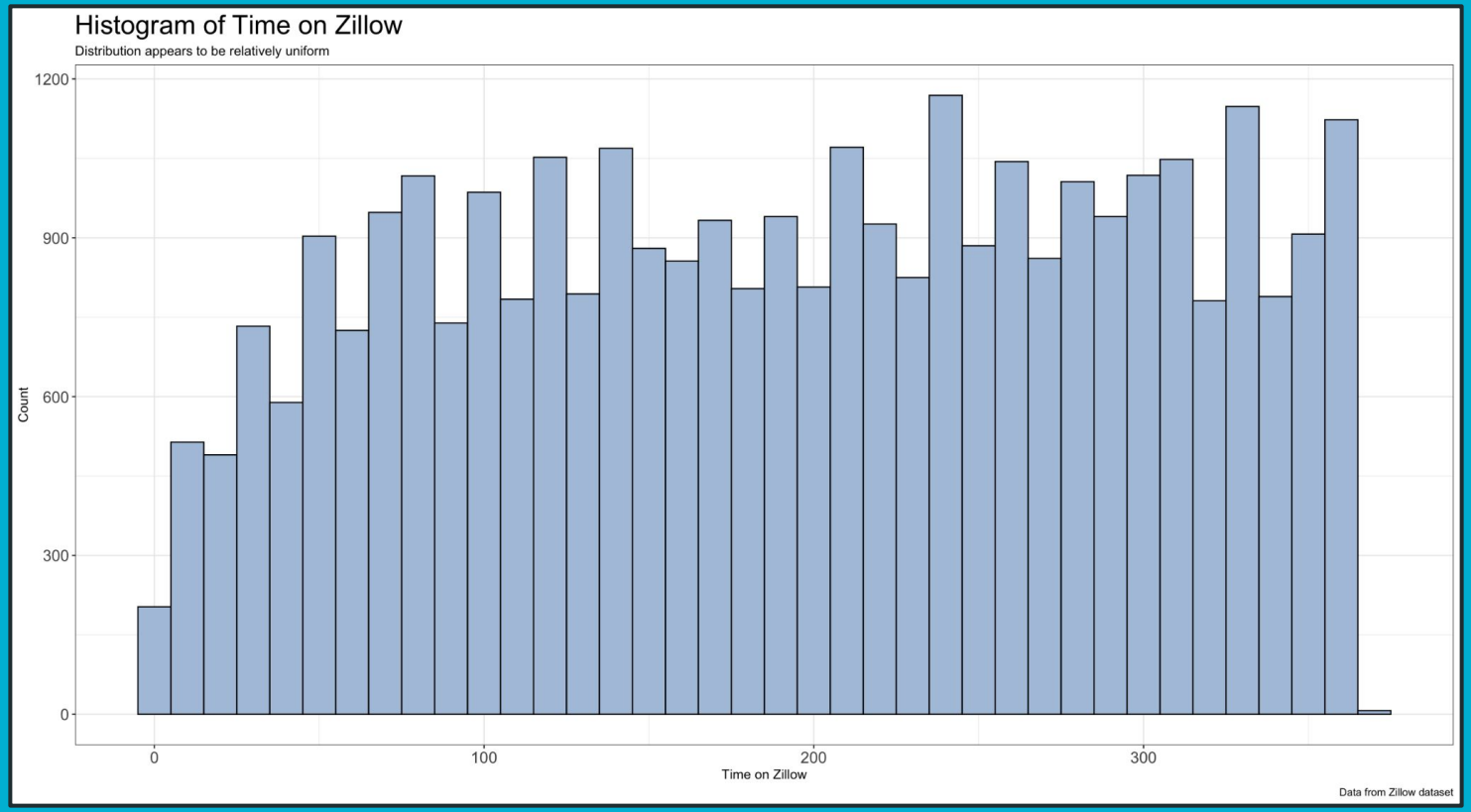
Exploratory Data Analysis

Spatial Analysis

Heatmap of Numeric Variables



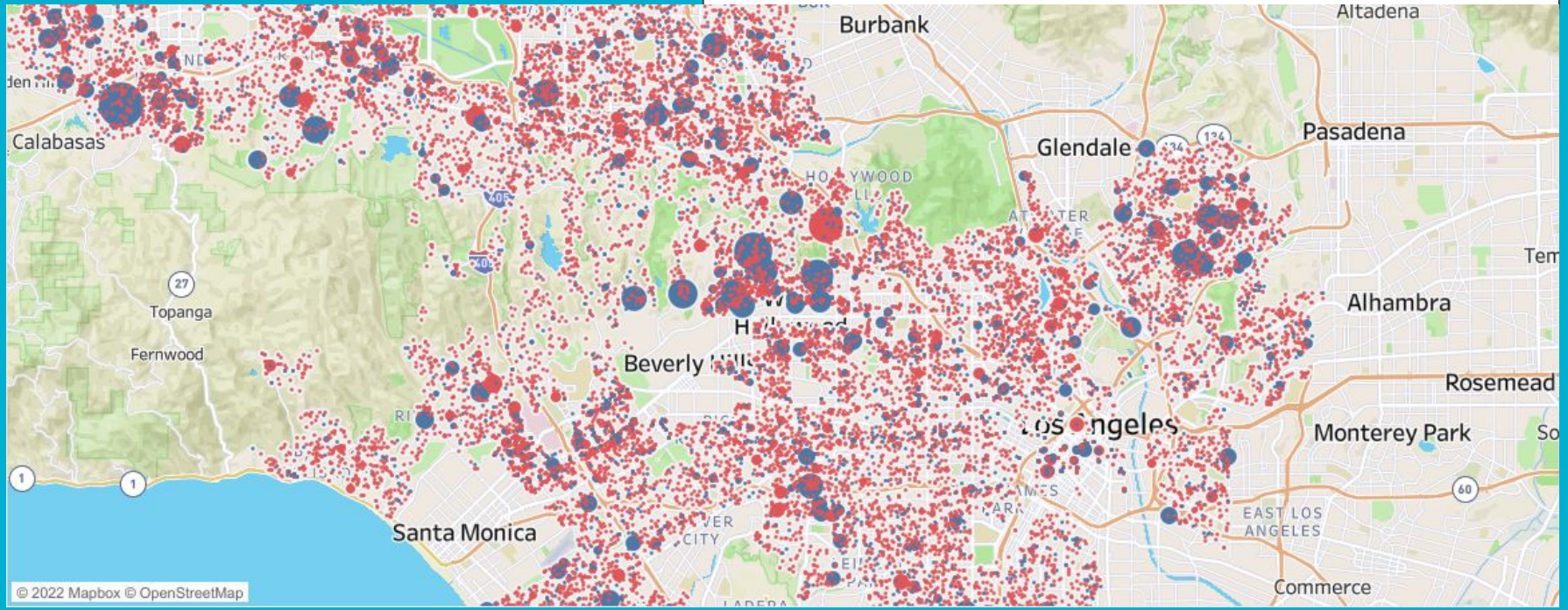
Outcome Variable: Time on Zillow



Mapping the Data

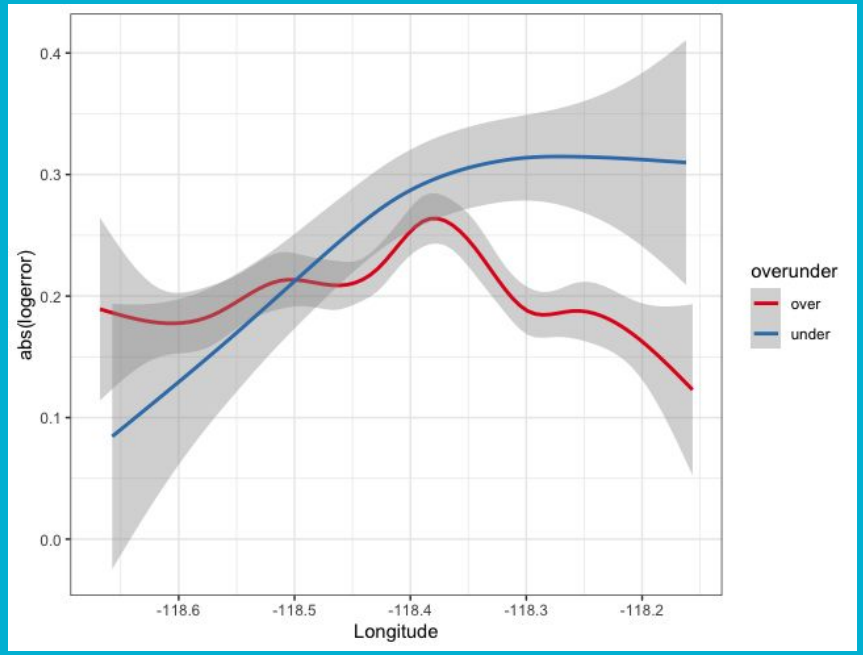
Speed
- <60 days to sell
- 60+ days to sell

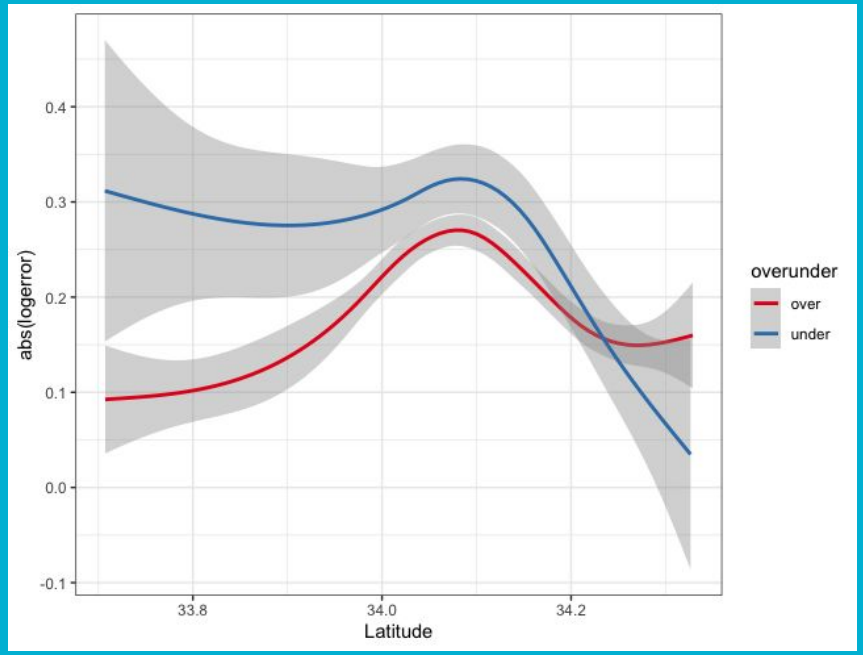
Favorites
- The bigger the point, the more favorites

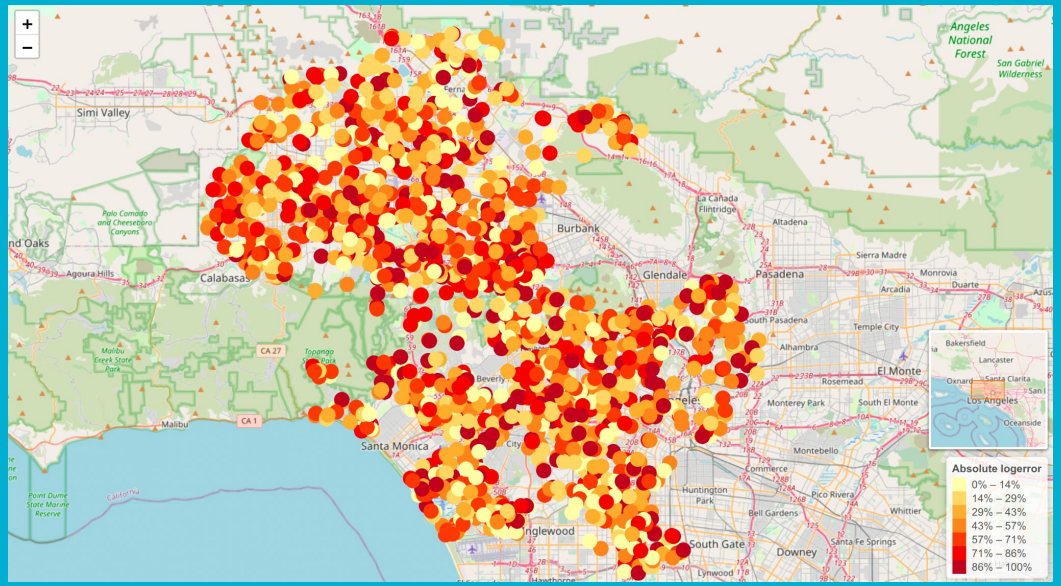


Link

Zestimate Accuracy by Coordinates





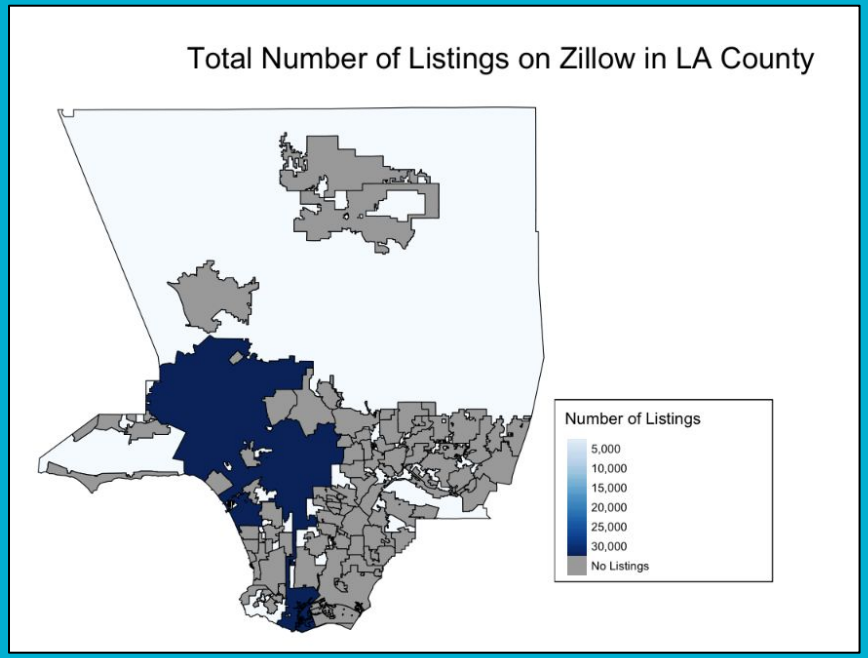


Zipcode	Abs. Log Error
90069	0.54997082
90038	0.54060919
90210	0.50070345
90021	0.08883130
90717	0.08782186
90732	0.08190552

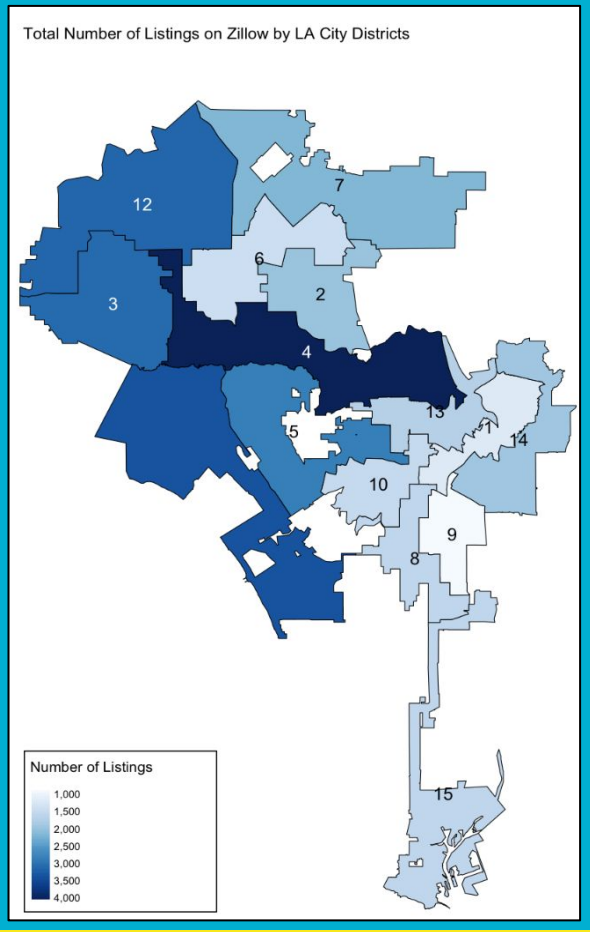
Narrowing Down the Cities

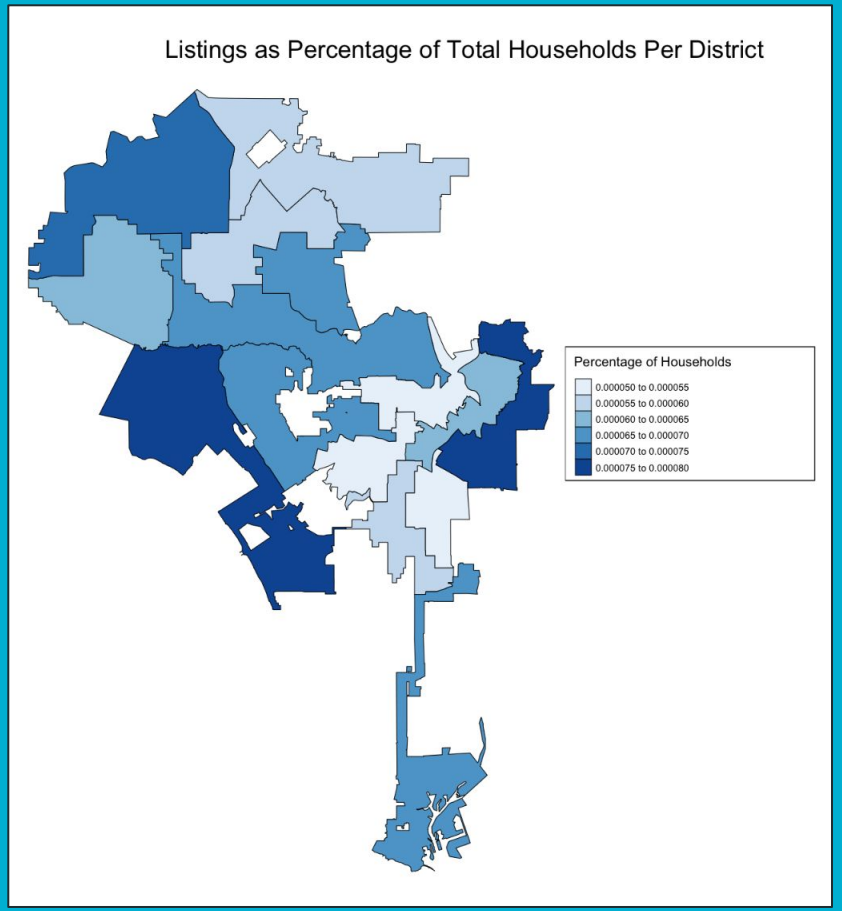
The “City” column in the Zillow dataset was misleading because the majority were neighborhoods within the city of LA itself

City	Number of Listings
Los Angeles	16040
Van Nuys	1204
Woodland Hills	1181
North Hollywood	1169
Sherman Oaks	888

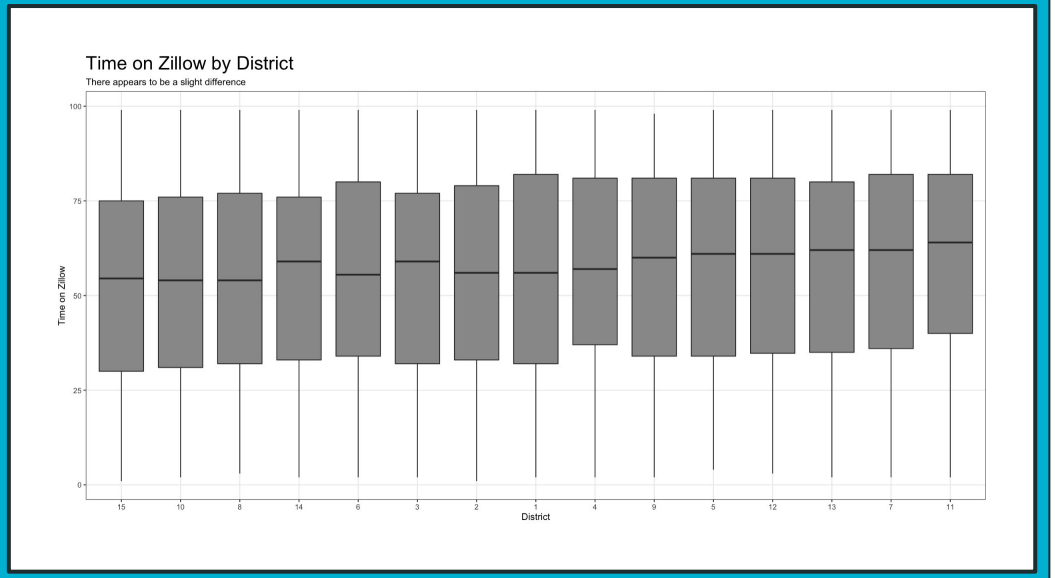


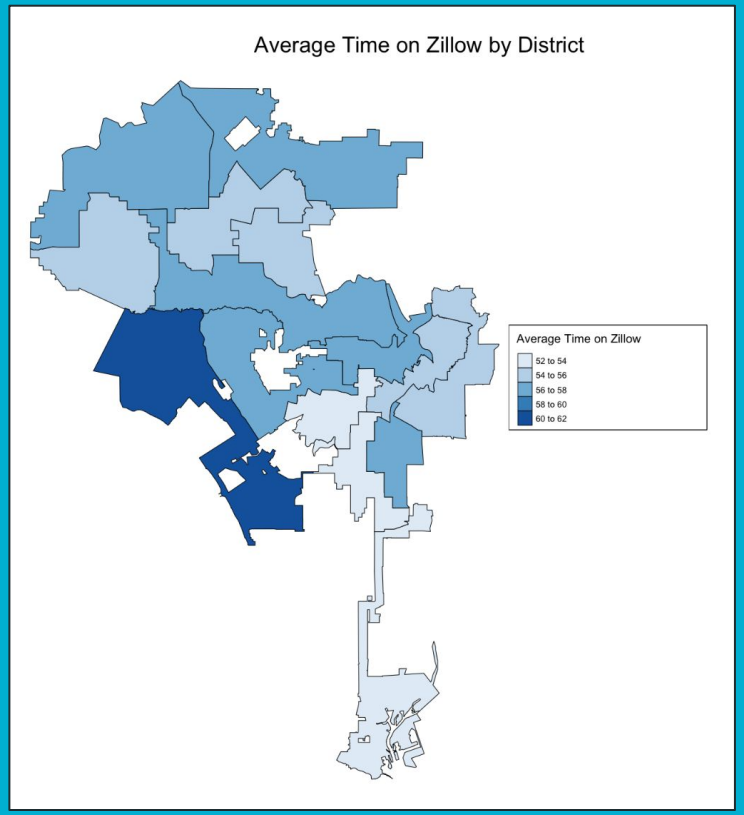
Listings by LA City Council Districts





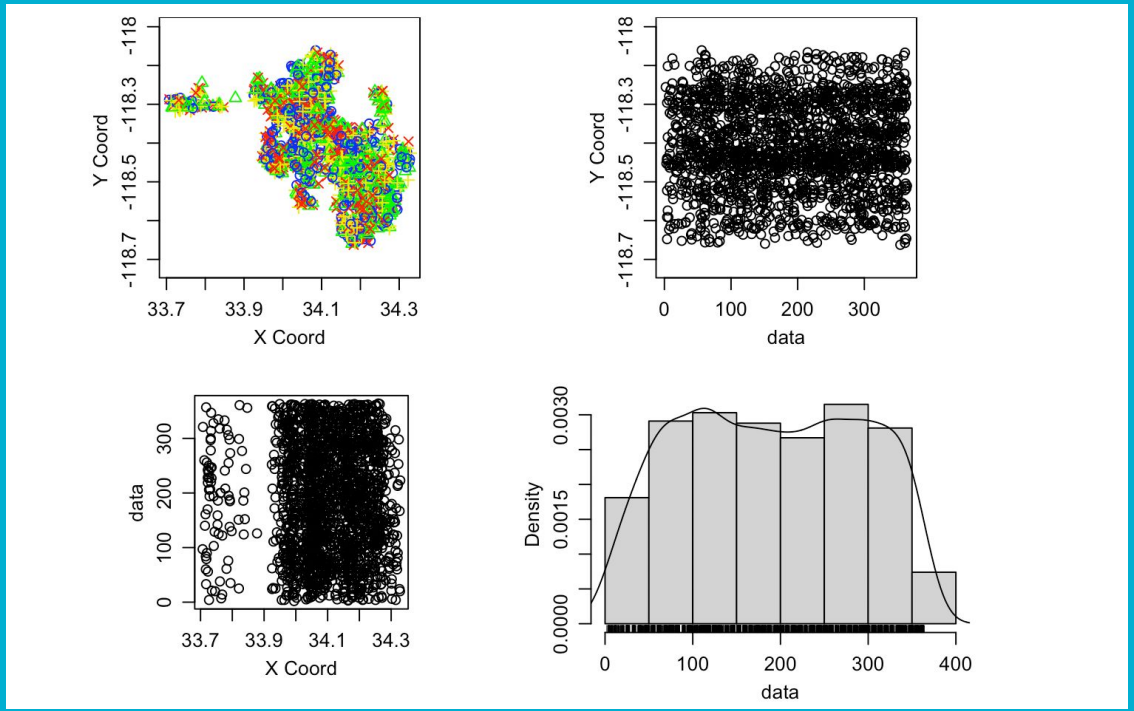
Time on Zillow by District





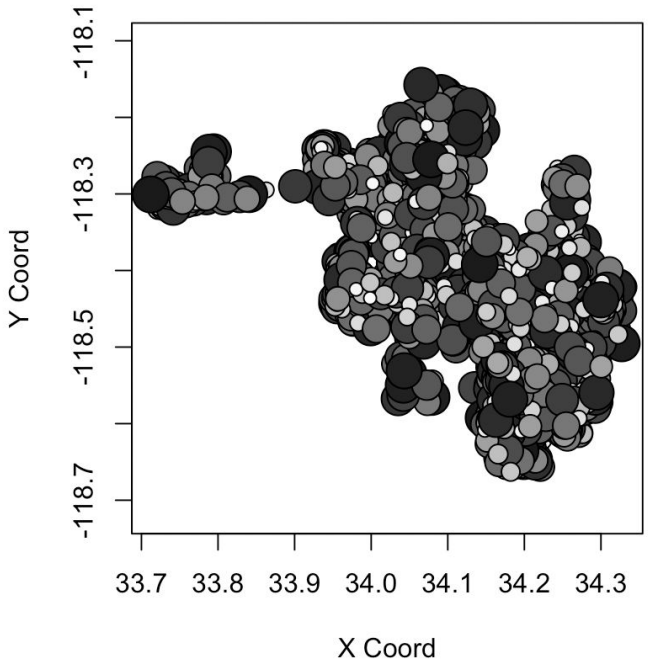
Spatial Visualization

Measuring ‘Days on Zillow’ variable against differences in latitude and longitude.



Spatial Visualization

Lighter, smaller circles indicate less time on Zillow while larger and darker circles indicate more time.

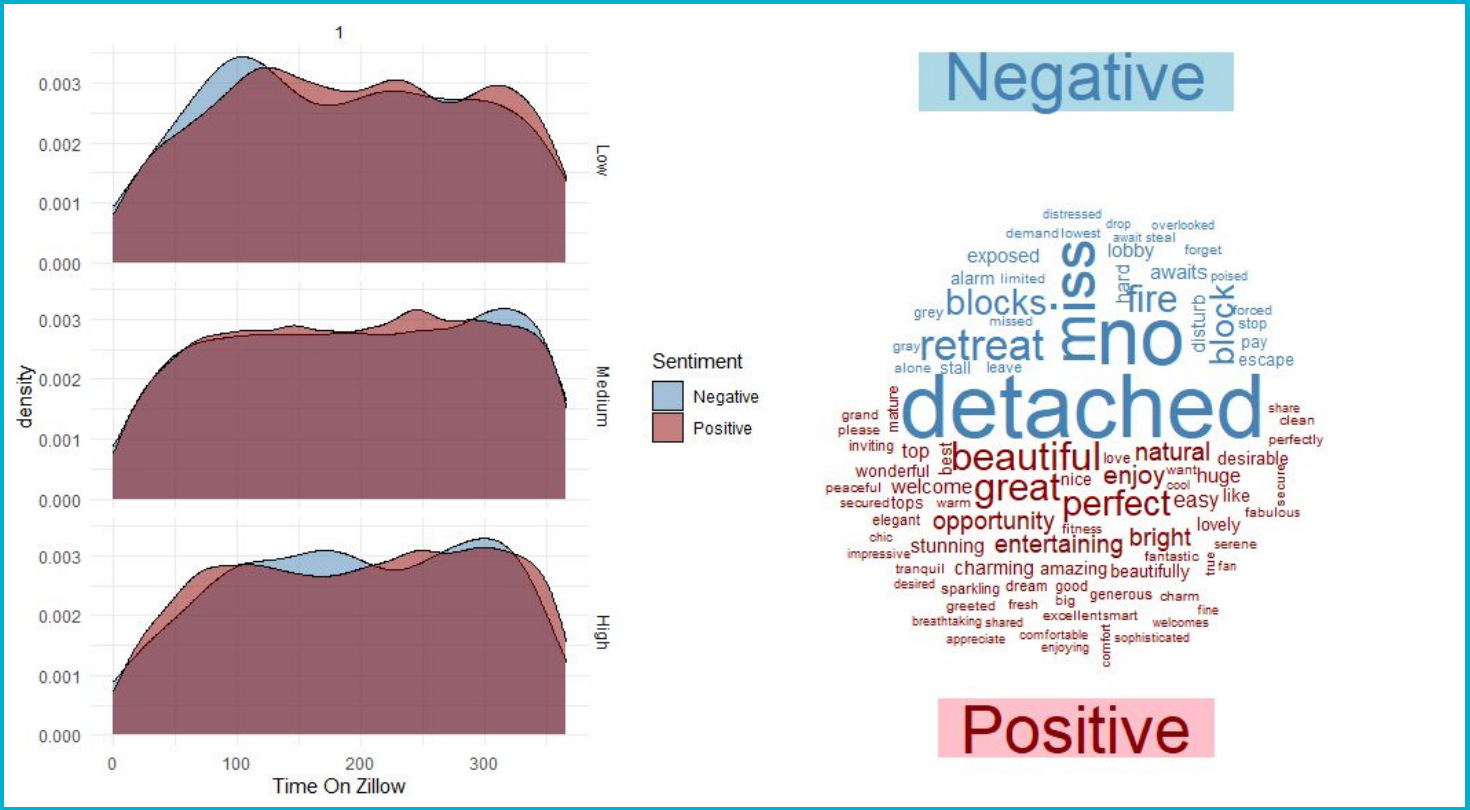


Sentiment Analysis

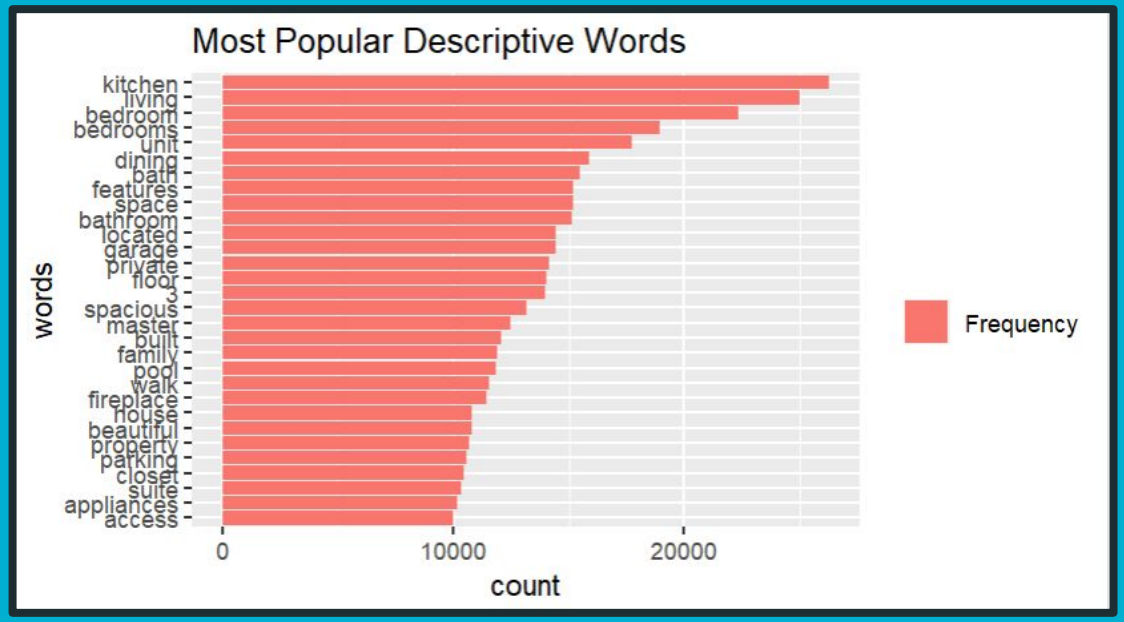
Sentiment Analysis

Left: Differences between Time on Zillow by Length of Description and Sentiment (AFINN Lexicon: Scores -5 to +5)

Right: Word Cloud of most common valued words in house descriptions



Sentiment Analysis (Property Features)



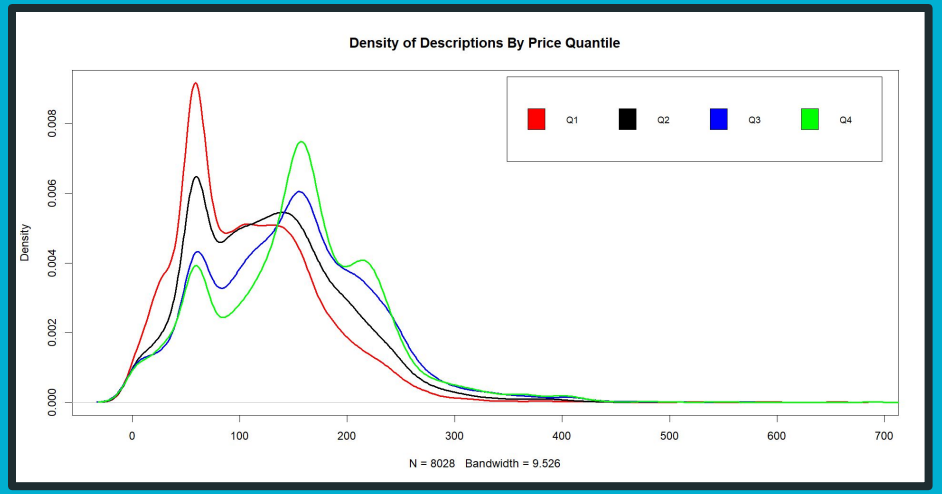
Frequencies over 10,000

Top Features

Sentiment Scores and Description Length by Quantiles

- Quantiles contain about 8,000 observations each
- Sentiment scores calculate the ratio between good and bad expressions.
- As property value increases, description length and positive sentiments increase.

Quantile	Price	Description Length (Means)	Sentiment Scores (Rounded)
1st	630,000	107	7
2nd	853,000	127	10
3rd	1350,000	147	11
4th	228,648,800	152	13



Sentiment Comparison Between Quantiles

POSITIVE

Q1	Q2	Q3	Q4
spacious	spacious	master	master
beautiful	beautiful	spacious	spacious
master	master	beautiful	perfect
stainless	perfect	perfect	beautiful
perfect	stainless	stainless	entertaining
bright	enjoy	enjoy	modern
easy	bright	entertaining	enjoy
enjoy	entertaining	modern	stunning
top	ready	bright	stainless
ready	upgraded	gorgeous	top

- How are these sentiment scores reflected in the positive terminology?

  - At first instance, an increase in property space is reflected (Spacious → Master)
  - Q1 compared to other groups has simpler terms.
  - The higher quantiles reflect a change in vocabulary as they more use 'sophisticated' descriptions (Modern, Stunning, Gorgeous)

Low to mid priced descriptions focus on home essentials while the higher categories focus on the 'quality' features of a property .

Word Clouds: Time on Zillow

Sold in less than 60 days

Sold in more than 60 days

Word Clouds: Time on Zillow

The words that were different among the word clouds:

Sold in less than 60 days

- PRIVATE
- SUITE
- HEART

Sold in more than 60 days

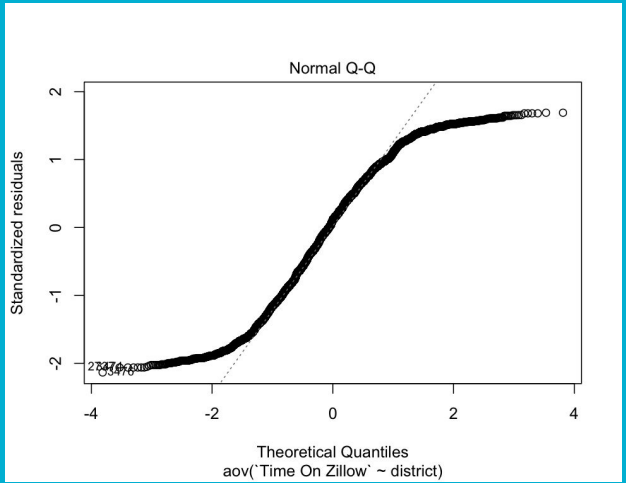
- CENTRAL
- INTERIOR
- STORAGE

Statistical Methods/ Summary of Results

ANOVA: Time on Zillow vs City of LA District

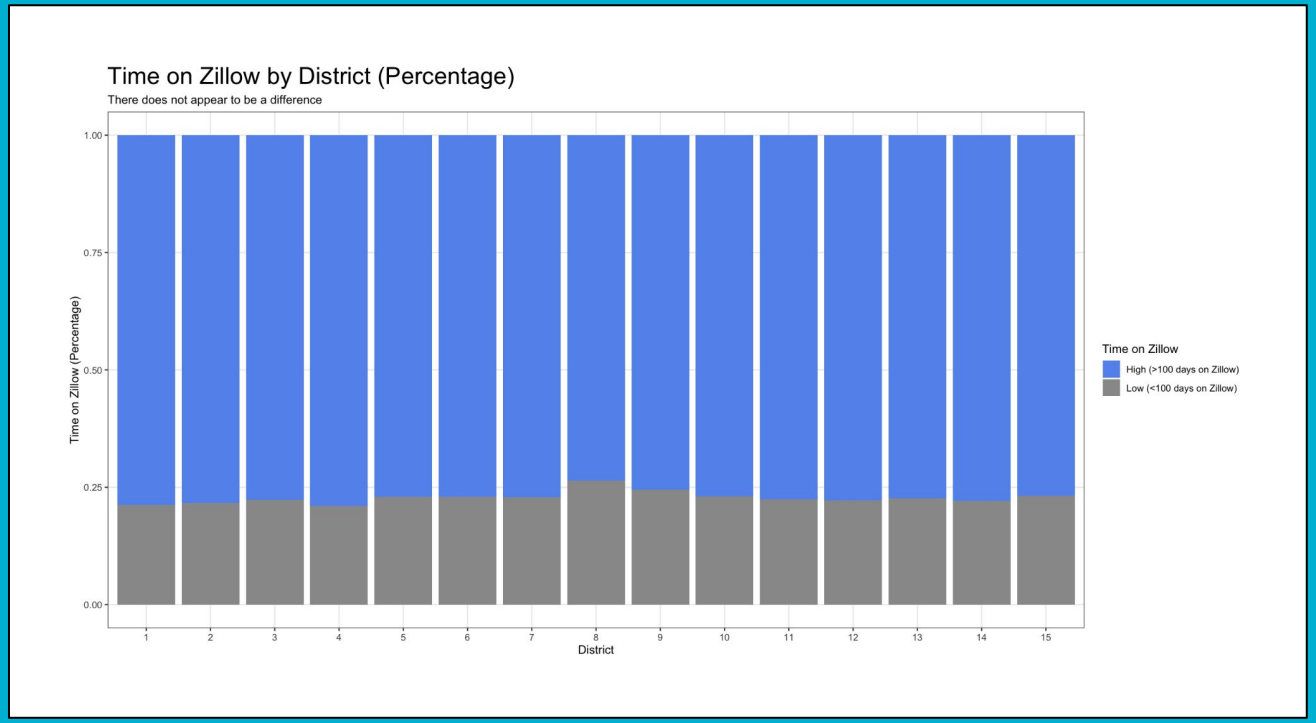
Important: Normality assumption is broken

	Df	Sum of Squares	Mean Square	F value	P-value
District	14	25647	1831.9	2.462	0.0018 **
Residuals	7257	5400566	744.2		

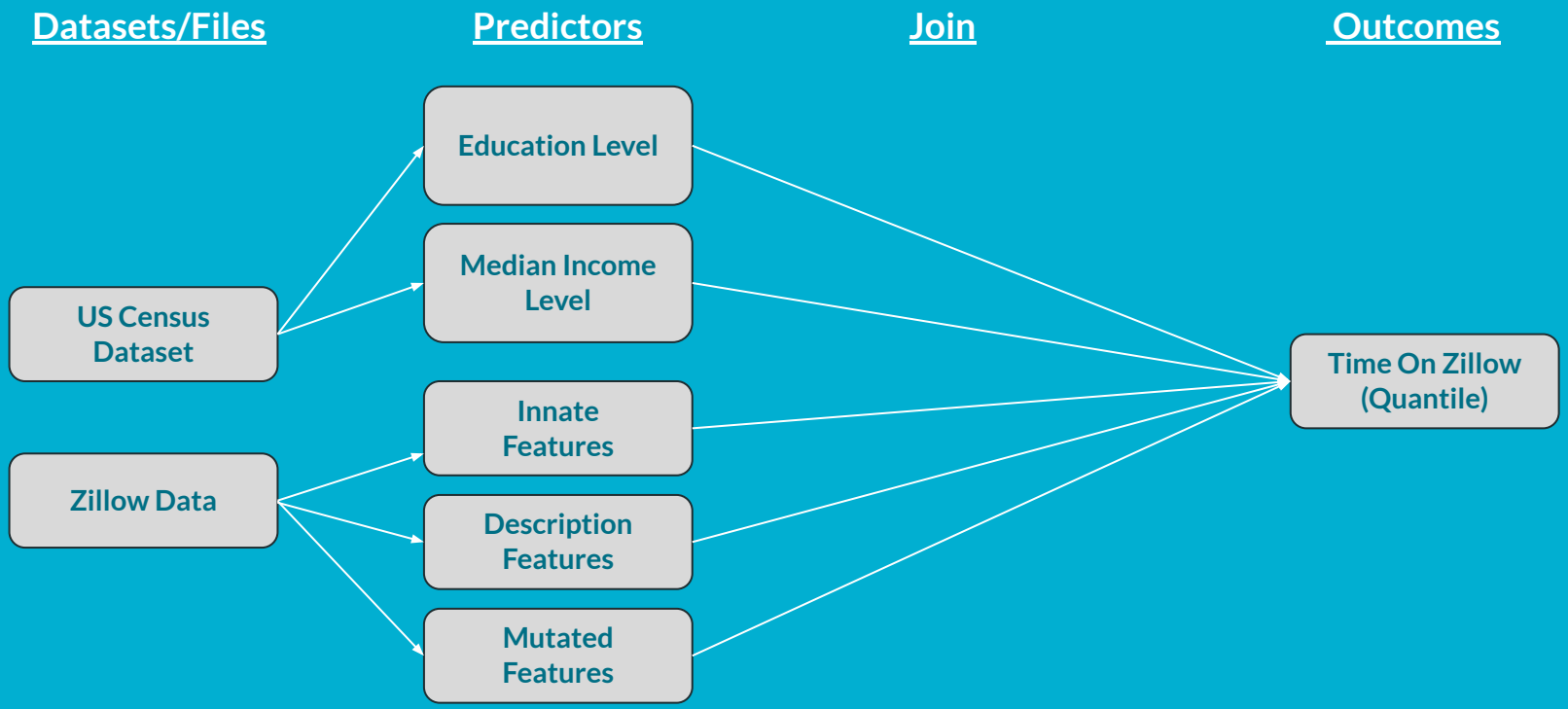


Pearson's Chi-squared Test: Time on Zillow as a Categorical Variable

- Time on Zillow categorized as below and above 100 days
- P-value = 0.04259 → Slightly under 0.05
- But plot suggests it's not that big of a difference



Random Forest - Schematic



Random Forest - Importance

Most Important Variables

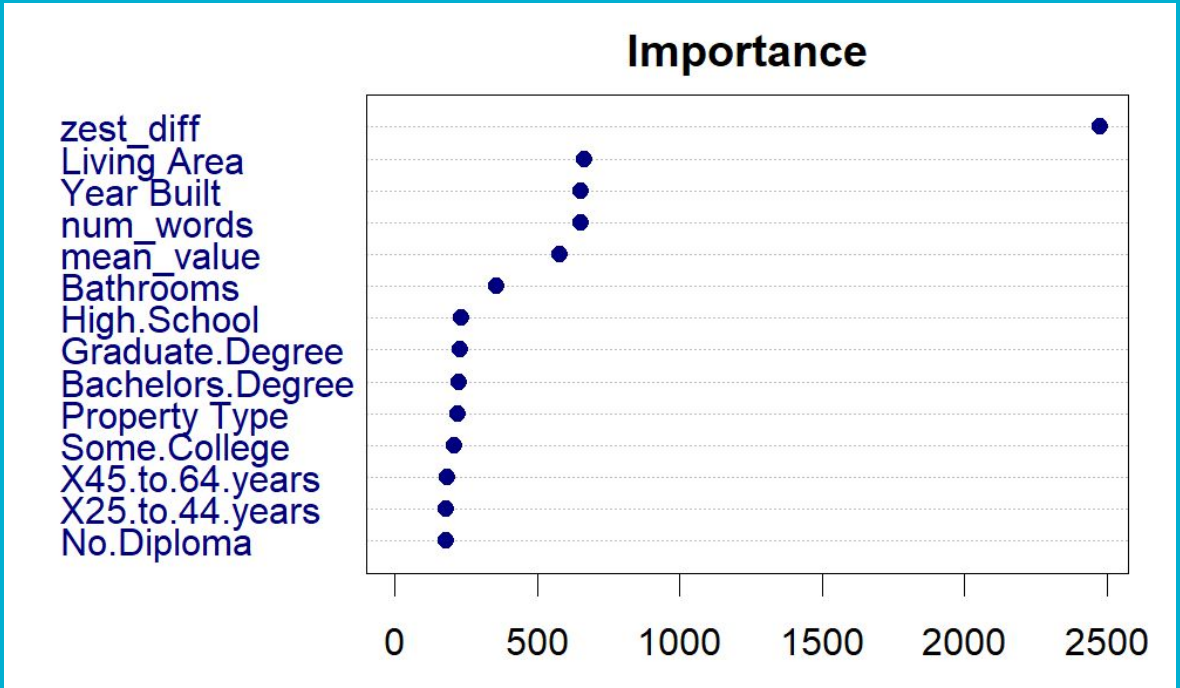
Difference Between Price and Zestimate

Living Area

Year Built

Length of Description

Sentiment Score of Desc.



Random Forest - Confusion Table

	Low	Medium	High	Very High
Low	999	180	58	120
Medium	164	743	234	156
High	55	211	598	404
Very High	51	61	281	862

Test set error rate: 38.15%
Groups are split by equally spaced quantiles

Recommendations and Shortcomings

- Recommendations:
  - Location, location, location
  - Biases in the description – Negatives are not really discussed
  - Look into relationships between “important variables” and time on Zillow
    - Difference between Zestimate and Price; can slightly lowball prices have a large influence
    - Length, Sentiment
- Shortcomings:
  - No “revolutionary” results
  - Brief implementation of the census data due to time.