**Zillow Project Data Definitions and Analysis Notes**

Sean Oslin

*Data definitions for ambiguous or unfamiliar variables*

Single unit property: A dwelling unit that is designed for occupancy by one household, located on a single parcel that does not contain any other dwelling unit (except an accessory dwelling unit, where permitted), and not attached to another dwelling unit on an abutting parcel. This classification includes individual manufactured housing units installed on a foundation system pursuant to Section 18551 of the California Health and Safety Code.

<http://www.qcode.us/codes/santamonica/view.php?topic=9-5-9_51-9_51_020>

This definition appears to be for the State of California, not just the city of Santa Monica.

Property types used for this analysis were selected based on standard real estate definitions for a single unit property.

Fips: The Federal Information Processing Standard Publication 6-4 (FIPS 6-4) was a five-digit [Federal Information Processing Standards](https://en.wikipedia.org/wiki/Federal_Information_Processing_Standards) code which uniquely identified [counties](https://en.wikipedia.org/wiki/County_(United_States)) and county equivalents in the [United States](https://en.wikipedia.org/wiki/United_States), certain U.S. possessions, and certain freely associated states.

*Data acquisition and preparation*

Data source: Zillow database, exclusively

Primary data wrangling used MySQL (file: zillow\_sql\_code.sql) that was then imported into Python

Summary of SQL query

* Column headings
  + propertylandusedesc as Property Description,
  + bathroomcnt as Number of Bathrooms,
  + bedroomcnt as Number of Bedrooms,
  + calculatedfinishedsquarefeet as Square Feet
  + taxvaluedollarcnt as Assessed Value
  + taxamount as Tax Amount
  + fips as County Number
* Data were filtered to the five Property Land Use Type Ids for single-unit residence
* Data were filtered to May and June of 2017
* All rows containing null values or 0s in the variables of interested were filtered out (120 rows)

An additional variable, County Tax Rate was created by dividing the Tax Amount by Assessed Value. This variable was created to facilitate a comparison of tax rates among the 3 counties.

*Data quality assurance*

The data set presented multiple square feet variables with no explanation of the differences among them. The columns with data were calculatedfinishedsquarefeet, finishedsquarefeet12, and finishedsquarefeet50.

No differences were found between calculatedfinishedsquarefeet and inishedsquarefeet12. Substantial differences were found between these first two columns and finishedsquarefeet50, which only had data for Ventura County. The data in finishedsquarefeet50 were often widely different from the other columns and, at times, not probable values for square feet (e.g. 200).

Because of the completeness of the data in the first two columns, I selected the first column, calculatedfinishedsquarefeet, for analysis.

The size of the final data sample was 15,963,

Split among the following unit types:

|  |  |  |
| --- | --- | --- |
| Property Land Use Type ID | Type of single unit property | Number of units in query |
| 261 | single family | 15956 |
| 262 | rural residence | 0 |
| 273 | bungalow | 0 |
| 275 | manufactured modular prefab | 7 |
| 279 | inferred single family | 0 |

And the following counties in California:

|  |  |  |
| --- | --- | --- |
| FIPS number | County | Number of units in query |
| 6037 | Los Angeles | 12,388 |
| 6059 | Orange | 2,780 |
| 6111 | Ventura | 795 |

Fips County lookup:

<https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_013697>

*Choice of data scaler*

Because each of the dependent variables were normally distributed, the standard scaler was used.