**Zillow Project Data Definitions and Analysis Notes**

*Data definitions*

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 all 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*

Data source: Zillow database exclusively

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

Summary of SQL query

* Column headings
  + propertylandusedesc,
  + bathroomcnt,
  + bedroomcnt,
  + calculatedfinishedsquarefeet,
  + taxvaluedollarcnt,
  + taxamount,
  + fips
* 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 in the variables of interested were filtered out

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 finishedsquarefeet50, which only had data for Ventura County, and the other two columns. Additionally, the data in finishedsquarefeet50 were often wildly 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 | Count of units |
| 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 | Number of units in query | County |
| 6037 | 12,388 | Los Angeles |
| 6059 | 2,780 | Orange |
| 6111 | 795 | Ventura |

Fips lookup:

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

https://www.latimes.com/opinion/livable-city/la-ol-sb50-single-family-20190424-story.html

Also need

axes are assessed at the county level, we would loke to know what states and counties these are located

Like to know the distribution of tax rates for each county

The data should have the tax amounts and tax value of the home

include in your report to us the distribution of tax rates for each county so that we can see how much they vary within the properties in the county and the rates the bulk of the properties sit around[[1]](#footnote-1)

For the first iteration of your model, use only square feet of the home, number of bedrooms, and number of bathrooms to estimate the properties assessed value, 'taxvaluedollarcnt'.

do some data validation or QA (quality assurance) to be sure the data you gather is what you think it is.

Your customer is the zillow data science team. state your goals as if you were delivering this to zillow.

1. This is separate from the model you will build, because if you use tax amount in your model, you would be using a future data point to predict a future data point, and that is cheating! In other words, for prediction purposes, we won't know tax amount until we know tax value. [↑](#footnote-ref-1)