Bethany Thompson

Thompson.bethany.01@gmail.com

Abstract

Quality of Zillow data and next steps for the preparation stage of the data science pipeline. Contains predictions\_2017 and properties\_2017 tables from the Zillow database.

Data quality report

Predicting Log Error Project

# About the Data

Data is acquired after connecting to the company SQL database and selecting the necessary data with a query. The query selects observations (properties) that are considered single units using the property land use ID and domain knowledge. Before any data prep, there are 61 columns and 70,364 rows. This report will detail the quality of each column and note initial observations, next steps, etc.

Necessary preparation for modeling may include removing outliers, scaling, creating dummy variables, or removing the feature. Any features with more than 75% of nulls will be dropped.

# Quality by Features

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column | Null Counts | Null % | Min/Max | Prep Needed |
| 'id' | 0 | 0 | 1 - 77613 |  |
| 'parcelid' | 0 | 0 |  |  |
| 'airconditioningtypeid' | 50,658 | 71.9 | 1 - 13 | drop (high nulls) |
| 'architecturalstyletypeid', | 70,213 | 99.8 | 2 - 21 | drop (high nulls) |
| 'basementsqft' | 70,320 | 99.9 | 63 – 2,443 | drop (high nulls) |
| 'bathroomcnt' | 0 | 0 | 0 - 20 | check outliers, scale |
| 'bedroomcnt' | 0 | 0 | 0 - 25 | check outliers, scale |
| 'buildingclasstypeid', | 70,364 | 100.0 |  | drop (high nulls) |
| 'buildingqualitytypeid' | 25,913 | 36.8 | 1 - 12 | scale |
| 'calculatedbathnbr' | 2,187 | 3.1 | 1 - 20 | scale |
| 'decktypeid', | 69,877 | 99.3 | 66 | drop (high nulls) |
| 'finishedfloor1squarefeet' | 64,695 | 91.9 | 47 – 15,998 | drop (high nulls) |
| 'calculatedfinishedsquarefeet', | 1,019 | 1.5 | 20 – 26,345 | check outliers, scale |
| 'finishedsquarefeet12' | 1,428 | 2.0 | 20 – 26,345 | drop (only need 1 sqft) |
| 'finishedsquarefeet13' | 70,154 | 99.7 | 224 – 2,400 | drop (high nulls) |
| 'finishedsquarefeet15', | 70,357 | 99.9 | 462 – 8,348 | drop (high nulls) |
| 'finishedsquarefeet50' | 64,695 | 91.9 | 47 – 15,998 | drop (high nulls) |
| 'finishedsquarefeet6' | 70,172 | 99.7 | 368 – 5,254 | drop (high nulls) |
| 'fips' | 0 | 0 | 6037, 6111 | create separate models |
| 'fireplacecnt', | 62,158 | 88.3 | 1 - 9 | drop (high nulls) |
| 'fullbathcnt' | 2,187 | 3.1 | 1 - 20 | scale |
| 'garagecarcnt' | 47,676 | 67.8 | 0 - 13 | determine if drop later |
| 'garagetotalsqft' | 47,676 | 67.8 | 0 - 3,774 | determine if drop later |
| 'hashottuborspa', | 69,108 | 98.2 | 1 | drop (high nulls) |
| 'heatingorsystemtypeid' | 25,015 | 35.6 | 1 - 24 | scale |
| 'latitude' | 0 | 0 | 33.3 – 34.8 | create clusters |
| **Column** | **Null Counts** | **Null %** | **Min/Max** | **Prep Needed** |
| 'longitude' | 0 | 0 | -119.4 – -117.6 | create clusters |
| 'lotsizesquarefeet', | 6,960 | 9.9 | 167 - 6,971,010 | check outliers, scale, create clusters |
| 'poolcnt' | 56,474 | 80.3 | 1 | drop (high nulls) |
| 'poolsizesum' | 69,588 | 98.9 | 28 - 2,176 | drop (high nulls) |
| 'pooltypeid10' | 69,941 | 99.4 | 1 | drop (high nulls) |
| 'pooltypeid2' | 69,531 | 98.8 | 1 | drop (high nulls) |
| 'pooltypeid7' | 57,323 | 81.5 | 1 | drop (high nulls) |
| 'propertycountylandusecode' | 0 | 0 |  | no description |
| 'propertylandusetypeid', | 0 | 0 | 260 - 266 | scale |
| 'propertyzoningdesc' | 25,378 | 36.1 |  | no description |
| 'rawcensustractandblock' | 0 | 0 | 60,371,011 - 61,110,091 | check outliers, scale, create clusters |
| 'regionidcity', | 1,402 | 2.0 | 3,491 - 396,556 | check outliers, scale, create clusters |
| 'regionidcounty' | 0 | 0 | 1,286 – 3,101 | check outliers, scale, create clusters |
| 'regionidneighborhood' | 43,255 | 61.4 | 6,952 - 764,167 | check outliers, scale, create clusters |
| 'regionidzip' | 259 | 0.4 | 95,982 - 399,675 | check outliers, scale, create clusters |
| 'roomcnt', | 0 | 0 | 0 - 86 | check outliers, scale |
| 'storytypeid' | 70,320 | 99.9 | 7 | drop (high nulls) |
| 'threequarterbathnbr' | 62,249 | 88.5 | 1 - 4 | drop (high nulls) |
| 'typeconstructiontypeid' | 70,198 | 99.8 | 4 - 6 | drop (high nulls) |
| 'unitcnt' | 25,652 | 36.5 | 1 - 13 | check outliers, scale |
| 'yardbuildingsqft17' | 68,064 | 96.7 | 11 - 4,839 | drop (high nulls) |
| 'yardbuildingsqft26' | 70,299 | 99.9 | 14 - 1,497 | drop (high nulls) |
| 'yearbuilt' | 1,008 | 1.4 | 1862 - 2016 yr | check outliers, scale |
| 'numberofstories' | 53,249 | 75.7 | 1 - 3 | drop (high nulls) |
| 'fireplaceflag' | 70,239 | 99.8 | 1 | drop (high nulls) |
| 'structuretaxvaluedollarcnt' | 1,148 | 1.6 | $5 - 66,404,932 | check outliers, scale, create clusters |
| 'taxvaluedollarcnt' | 921 | 1.3 | $9 - 149,139,154 | check outliers, scale, create clusters |
| 'assessmentyear' | 0 | 0 | 2014 - 2016 yr | create dummies |
| 'landtaxvaluedollarcnt', | 1,607 | 2.3 | $4 - 94,011,079 | check outliers, scale |
| 'taxamount' | 439 | 0.6 | $2.54 - 1,824,154.85 | check outliers, scale, create clusters |
| 'taxdelinquencyflag' | 69,043 | 98.1 |  | drop (high nulls) |
| 'taxdelinquencyyear', | 69,043 | 98.1 | 6 - 15 | drop (high nulls) |
| 'censustractandblock' | 1,590 | 2.3 |  | check outliers, scale |
| 'logerror' | 0 | 0 | -4.66 - 5.26 | scale, create clusters, target |
| 'transactiondate' | 0 | 0 |  |  |

# Conclusions

Many of the features have a large amount of null observations - up to 100%. In theory, some could seem useful to predict tax value, but for now we can not use them if the data quality is poor. I will be removing columns with a null percentage of 75% or higher. Leftover nulls will be dealt with later. I will either replace nulls with mean, median, etc. or remove the column/row.

After exploring with this data, I can further prep the features for clustering and modeling. Features may need outliers removed if any found and to be scaled. Additional features can be created, such as dummy variables or creating calculations based on other columns.

Overall, the data is sufficient enough to continue the project.