Miriam Zuk and Sam Maurer, Jan 2015

Objective: we need to filter the Dataquick transaction records so that they include only true market-rate sales, with valid price-per-square-foot and without duplication

| **Issue** | **Possible Criteria** | **Analysis** | **Final criteria (SQL syntax)** |
| --- | --- | --- | --- |
| ***PART 1 – GENERAL FILTERS*** | | | |
| Remove transactions from outside the 9-county San Francisco Bay Area |  |  | 1. mm\_fips\_muni\_code IN (1,13,41,55,75,81,85,95,97) |
| Remove transactions from prior to 1988 |  |  | 1. (s.sr\_date\_transfer/10000) >= 1988 |
| Remove non-residential transactions |  | * These represent less than 10% of state-wide transactions provided by Dataquick, and only 2.2% after applying the other data filters | 1. SUBSTRING(a.use\_code\_std FROM 1 FOR 1) = 'R' |

|  |  |  |  |
| --- | --- | --- | --- |
| ***PART 2 – LINKING TRANSACTIONS TO ASSESSOR DATA*** | | | |
| Basic identifiers have to be present in order for us to link transactions to census tracts | 1. sr\_property\_id nonempty, > 0 2. sa\_census\_tract nonempty | * Census tracts are listed in the assessor table but not in the transactions table, so we match transactions to assessor records using the property id * 8% of transactions have a missing or 0 property id, and 0.3% of current assessor records are missing a census tract * These transactions will disappear automatically from the final statistics, but it’s probably best to explicitly remove them so they don’t affect how we’re judging the other data cleaning filters * There don’t seem to be any zero values for the census tract | 1. sr\_property\_id IS NOT null 2. sr\_property\_id > 0 3. sa\_census\_tract != ‘’   (varying syntax due to integer vs. character data fields) |
| Historical assessor data is sporadically incomplete, so it’s probably best to pull square footage and use codes from the current assessor table, even though they could have changed or the property may no longer exist | 1. sales.sr\_property\_id = assessor.sa\_property\_id  for matching the square footage and use codes | * Historical assessor data is missing for several entire counties in 2011 and 2012 * In general, the historical tables are also less complete than the current assessor table * When we match transactions to the next-year assessor tables, 1%–10% are missing, but when we match them to the current table, only < 1% are missing | 1. sales.sr\_property\_id = assessor.sa\_property\_id  for matching the square footage and use codes |
| Square footage and use codes have to be present in order to calculate final statistics | 1. sa\_sqft nonempty  and > 0 2. use\_code\_std nonempty | * After implementing the primary filters (arms-length, positive transfer value, property match in the assessor table), 3.5% of the Bay Area transactions have missing or zero square footage and < 1% are missing a use code * We’ll proactively remove these from the “clean” data tables | 1. sa\_sqft IS NOT null 2. sa\_sqft > 0 3. use\_code\_std != ‘’ |

|  |  |  |  |
| --- | --- | --- | --- |
| ***PART 3 – PROPERTIES OF INDIVIDUAL TRANSACTIONS*** | | | |
| Dataquick’s arms-length flag may not be accurate, because it includes transactions with a transfer value of 0 and excludes some with a transfer value > 0 | 1. sr\_arms\_length\_flag = 1 and/or 2. sr\_val\_transfer nonempty and > 0 | * Cross-tabulation of transfer value and arms-length flag:  **(A)** 38% - value > 0 and arms-length  **(B)** 48% - value = 0 and non-arms-length  **(C)** 12% - value > 0 and non-arms-length  **(D)** 2% - value = 0 and arms-length * Group D in particular calls Dataquick’s methodology into question, but examples from Group C look ok (sales to trusts and other things we should be filtering out) * All in all, it seems best to remove transactions Dataquick classifies as non-arms-length rather than trying to catch all of them using other filters * We have to remove transactions with missing or 0 transfer values anyway, in order to calculate meaningful price statistics | 1. sr\_arms\_length\_flag = ‘1’ 2. sr\_val\_transfer IS NOT null 3. sr\_val\_transfer > 0 |
| Only include resale and subdivision transaction types | 1. sr\_tran\_type = ‘R’   OR sr\_tran\_type = ‘S’ | * For transactions with value > 0:  89% = R (resale) 10% = S (subdivision) 0.5% = C (construction) 0.5% = T (timeshare) none refinance, none missing | 1. sr\_tran\_type = ‘R’ OR sr\_tran\_type = ‘S’ |
| Possibly should filter by transaction document type | 1. sr\_doc\_type = ? | * For transactions with value > 0:  46% = G (grant deed) 6% = U (trustees deed) 1% = Q (quitclaim) negligible H, W, T 47% missing * Too many missing values to use this field | 1. NONE |
| Only include transactions representing full sale amount | 1. sr\_full\_part\_code = ‘F’ OR sr\_full\_part\_code empty | * For transactions with value > 0: 79% = F (full) 3% = P (partial, excluding liens etc.) 4% other (C, U) 14% missing (data dictionary indicates missing = assumed full) * Overall, the data in this field doesn’t seem reliable enough to use | 1. NONE |
| Remove trust transactions that Dataquick misclassified as arms-length | 1. sr\_buyer without “trust” 2. sr\_seller without “trust” | * Pulled a sample of matching records and the filter works as expected | 1. sr\_buyer NOT ILIKE ‘% trust%’ 2. sr\_seller NOT ILIKE ‘% trust%’  (case-insensitive pattern matching where % matches any string of zero or more characters) |
| Remove public agency transactions because they’re often not at market prices | 1. Eliminate transactions that include “county,” “city,” “agency,” “redevelopment,” etc. | * Filter works as expected, with > 90% of the matches being public agencies * The false positives are entities with names like “First National Bank Daly City,” but there doesn’t seem be to any easy way to improve the pattern matching | 1. As above, with “county,” “city,” “agency,” “redevelopment” |

|  |  |  |  |
| --- | --- | --- | --- |
| ***PART 4 – SETS OF RELATED TRANSACTIONS*** | | | |
| Sets of transactions involving the **same property id** on the same day often represent different parts of a single sale (refinance, multiple loans, trust transactions, one to many owners or vice versa, etc.) | 1. After applying all the prior filters, group remaining transactions by sr\_property\_id and sr\_date\_transfer 2. If the dollar amounts match, keep one transaction 3. If the dollar amounts differ, remove all the transactions | * After applying all the prior filters, these duplicates represent about 1.0% of the remaining transactions (0.6% same price, 0.4% differing prices) * The same-price duplicates are *mostly* transactions involving intermediaries, and the differing-price duplicates are *mostly* multi-part transactions, but the patterns aren’t consistent enough for us to get reliable prices from these records | 1. After applying all the prior filters, group remaining transactions by sr\_property\_id and sr\_date\_transfer 2. Remove all these transactions |
| Sets of residential transactions on a single day with the **same document number but differing property id’s** represent subdivision or condo building sales, which often have incorrect price or square footage data | 1. After applying all the prior filters, group remaining residential transactions by mm\_fips\_muni\_code, sr\_doc\_nbr\_fmt, and sr\_date\_transfer 2. If the dollar amounts match, calculate price per square foot as transaction price / total square footage 3. If the dollar amounts differ, calculate the price per square footage normally | * After applying all the prior filters, these duplicates represent about 1.2% of the remaining transactions * (We have to group transactions by county here because document numbers can repeat across jurisdictions) * Dataquick reps informed us that for residential condo and subdivision transactions involving multiple property id’s, they record the total transaction price separately for each unit * This looks correct based on the data, but it’s hard to be certain | 1. After applying all the prior filters, group remaining residential transactions by mm\_fips\_muni\_code, sr\_doc\_nbr\_fmt, and sr\_date\_transfer 2. If the dollar amounts match, only keep one of the transactions, and calculate price per square foot as transaction price / total square footage 3. If the dollar amounts differ, calculate the price per square foot normally |

|  |  |  |  |
| --- | --- | --- | --- |
| ***PART 5 – PRICE OUTLIERS*** | | | |
| Identify and filter out significant outliers in price per square foot, because these are likely to be errors that would bias aggregate calculations | 1. After applying all prior filters adjust the remaining prices for CPI inflation 2. Remove the top 0.1% of transactions by price per square foot, separately for each county | * We adjust prices to 2010 dollars using national headline CPI for the calendar year of the transaction[[1]](#footnote-1) * The residential price cutoffs work out to $1054 for Alameda, $794 for Contra Costa, $1788 for Marin, $1577 for Napa, $2014 for San Francisco, $1773 for San Mateo, $1354 for Santa Clara, $729 for Solano, and $1260 for Sonoma, in 2010 dollars | * After applying all prior filters, adjust the remaining prices for CPI inflation * Remove the top 0.1% of transactions by price per square foot, separately for each county |

1. http://www.bls.gov/data/inflation\_calculator.htm [↑](#footnote-ref-1)