Table of Contents

Basic Description1

Location2

Housing5

Agents\_Owner9

Selling12

Tax17

No Data18

Geographic Spread19

**Basic Description**

Number of observations: 9,717

Unique Identifier: “Parcel I.D”[[1]](#footnote-1)

|  |  |  |
| --- | --- | --- |
| **Year** | **Number of observations** | **ACS Estimate of Housing Units for PUMA (B25001)** |
| 2005 | 1463 |  |
| 2006 | 1358 |  |
| 2007 | 1103 |  |
| 2008 | 791 |  |
| 2009 | 792 | 61,463 |
| 2010 | 764 | 64,565 |
| 2011 | 760 | 65,727 |
| 2012 | 957 | 66,398 |
| 2013 | 1162 | 66,978 |
| 2014 | 1096 |  |
| 2015 YTD | 428 |  |

Duplications: 1 duplicated entry (Listing.ID 22010703 and 22010704)

* \* All variables the same besides Listing ID

**Location**

**\*\*\* Needs Cleaning**

**\*\*\* Cleaned**

“List.Number”

* The number used in the MLS database
* XXXXXXXX
* All Values are unique and there are no NAs

“Data\_Year”

* Original data set year (SDAL created to easily extract years of data)
* XXXX

“str\_number”

* The number of the street
* Numeric
* No missing values

“st\_direction”

* The direction of the street
* Validity: Should be N, W, S, E (or some consistent derivative). There are not. Current levels are:
  + "" "2832" "3600" "E" "E." "East" "J" "Lot 142" "n" "N" "N." "North" "s" "S" South" "w" "W" "W." "West" "WEST"

“Street.Name”

* Name of street of the property
* Character
* No missing values
* Some rows are combined variables with street suffix (e.g Winterberry Ct)

“Street.Suffix”

* Type of street the property is located on
* “Ct/Blvd/Dr/St/Pl/Cir/Ln/Trail/Way/Crescent/Point/Loop/Rd/Ave”
* Some missing values are because value is combine with “Street.Name”
* N/As are coded as “”
* 45% Missing

“Legal.Description”

* Description of the location of the property
* String
* 2 NA’S

“City”

* City name of where the property is located
* Factor
* “Lanexa, Newport News, Toano, Williamsburg”
* None missing

“State”

* State name of where the property is located
* Factor
* “VA”
* None Missing

“county”

* County name of where the property is located
* Factor
* 3 “James City” and the rest are N/A
* Some N/As coded as “”

“country”

* Country name of where the property is located
* Factor
* 9% “United States” and the rest are N/A
* Some N/As coded as “”

“Zip.Code”

* Zip code of the property
* Factor
* Some are the 5 digit code, others include the +4, and some include all with no -

“Area”

* Area the property is located on
* “JCC”
* No missing Values

“Subdivision”

* Name of the subdivision the property is located on
* Factor
* No missing values

“Neighborhood”

* Name of neighborhood of the property
* Character
* 81% are missing
* NA’s are coded as “”

“Zoning”

* The zoning of the property
* Factor
* In consistant coding
  + Over 286 levels used
  + Many are the same level just entered differently (e.g. 205 SF-Suburban, 502 single fam, 502 Single Fam, 502-SF Suburban, 502, Single Fam”

“Parcel.ID”

* Identification number for identification or record keeping of the property
* Factor
* 7% Missing/Uknowns and are coded different: “”, 0, 9999, 999999, 99999,
* Three TBD categories
* Not all are parcel ID’s (i.e.address)
* Not consistent
  + Some had “-“ in the parcel

“geo\_lat”

* Geographic latitude of the property
* Numeric
* 157 missing
* Multiplied by .000001

“geo\_lon”

* Geographic longitude of the property
* Numeric
* 157 missing: Same ones missing “geo\_lat”
* No obvious pattern to why these are missing
* Multiplied by .000001

“TractID”

* Census tract id that listing is withing
* Numeric
* 157 missing: Same ones missing “geo\_lat”

“Directions”

* Directions to the property
* String
* Some Missing values are coded n/a rather than N/A

“Relocation”

* Was the property relocated?
* “Yes/No”
* 7% Missing

**Housing**

**\*\*\* Needs Cleaning**

**\*\*\* Cleaned**

“List.Number”

* The number used in the MLS database
* XXXXXXXX
* All Values are unique and there are no NAs

“Data\_Year”

* Original data set year (SDAL created to easily extract years of data)
* XXXX

“Assessed.Value”

* Estimated value of the property
* Numeric
* 60% Missing
* In valid entries: Some are listed at 0 or near 0

“Est\_Fin\_SqFt”

* Estimation of the final square footage
* Numeric
* 1 observation listed at 0 square feet

“Year.Built”

* Year the property was built
* XXXX (Needed to convert to factor
* Invalid entries (e.g. 0, 3, 19, 2942)

“ 2plus\_Bdrms\_1s\_Flr “

* Is there 2 bedrooms on first floor
* “Yes/No”
* Inconsistent missing value coding: “”, NA

“Realtor.com.Type”

* Used for Realtor.com to differentiate between Residential property types, i.e. Condo/Townhouse, Single Family, etc.
* “1, 2, 3”
  + Unknown what each stands for

“lot\_size”

* Size of the property’s lot
* String
* Missing are coded as “”
  + 77% are missing
* Inconsistent coding: some are lot size (100 x 100), others are descriptions of lots (e.g. corner lot”

“Total.Acres”

* Total amount of acres within the property
* Numeric (0 - 9999)
  + High value invalid (too large for area)
* 1 missing

“Condo.Level”

* The number of levels in the condo
* “1, 2, 3,4, Other”
* Inconsistent missing value: “”, N/A, NA

“Total.Rooms”

* Total amount of rooms in the unit (0 – 19)
* Numeric
* 7% Missing

“Total.Bedrooms”

* Total amount of bedrooms in the unit
* Numeric (0-9)
* None missing

“total\_bath”

* Total amount of bathrooms in the unit
* Numeric (0-10)
* None missing

“Baths\_Full”

* Total amount of full bathrooms in the unit
* Numeric (0-7)
* None missing

“Baths\_Half”

* Total amount of half bathrooms in the unit
* Numeric (0-7)
* None Missing

“Garage.Type”

* Type of garage of the property
* Factor
* Missing coded as “”
  + 1% Missing

“garage\_stall”

* Amount of garage stalls on the property
* “0, 1, 2, 3, 3+”
* Missing coded as “”
  + 16% Missing

“Water.Frontage”

* Is there water frontage on the property?
* “Yes/No”
* Missing coded as “”
  + 4% Missing

“IDX”

* Is there an IDX
* “Yes/No”
* None missing

“Mster.Bdrm.1st.Floor”

* Is the master bedroom on the first floor?
* “Yes/No”
* None missing

“SqFt.Source”

* Source of the square footage information
* “Owner, Public Record, Appraiser”
* None missing

“Number\_Stories”

* The number of stories in the unit
* “1, 1.5, 2, 2.5, 3, 4+”
* None missing

“Number\_Fireplaces”

* The number of fireplace in the unit
* “1, 2, 3, 4+”
* None missing

“Golf.Frontage”

* Is there a golf frontage on the property
* “Yes/No”
* Invalid value of “Select one”
* Inconsistent missing code: “”, NA
  + 46% Missing

“Type”

* Does the house share a common party wall or is it isolated?
* “Mobile Home, Single Family Detach/Single Family Attach, Other”
* 1 missing
  + coded as “ ”

“Ownership”

* The type of ownership
* “Fee simple, Condominium, Coop”
* None missing

“Est.Unfinished.Sq.Ft”

* Estimated amount of square feet that will is unfinished
* Numeric (0 - 4972)
* 87% missing

“Garage.Bays”

* Number of garages for the unit
* “0, 1, 2, 2.5, 3, 4, 5”
* 1% Missing

“Lot.Description”

* Description of property’s lot
* String
* Similar entries to “lot\_size”

“Rooms”

* The amount of each type of room in the property
* String
* Varied length and order
* Each section separated by “|”

“Features”

* The characteristics of the property
* String
* Varied length and order
* Each section separated by “|”

**Agents\_Owner**

**\*\*\* Needs Cleaning**

**\*\*\* Cleaned**

“List.Number”

* The number used in the MLS database
* XXXXXXXX
* All Values are unique and there are no NAs

“Data\_Year”

* Original data set year (SDAL created to easily extract years of data)
* XXXX

“Agency.Name”

* The name of the agency (ID number)
* String
* None missing

“Agency.Phone”

* Phone number of the agency
* (XXX) XX-XXXX
* None missing

“Agency.Email”

* Email of the agency
* String
* None missing

“Listing.Agent”

* Name of the listing agent
* String
* Invalid entries: -2, 1084 1084
* None missing

“Listing.Agent.Phone”

* Phone number of the listing agent
* (XXX) XX-XXXX
* Missing Values coded as “”
  + 36% missing

“Listing.Agent.Email”

* Email of the listing agent
* String
* Missing Values coded as “”
  + 24% missing

“Co.Listing.Agent”

* Name of the co-listing agent
* String
* Missing Values coded as “”
* Invalid enteries: “-2, 665 665”

“Selling.Agency”

* Name of the selling agency (ID)
* String
* Missing Values coded as “”
* Invalid entries: “23, (23), 90, (90)
  + 20 missing

“Selling.Agency.Phone”

* Phone number of the selling agency
* (XXX) XX-XXXX
* Missing Values coded as “”
  + 28 misssing

“Selling.Agency.Email”

* Email of the selling agency
* String
* Missing Values coded as “”
  + 42 missing

“Selling.Agent”

* Name of the selling agent
* String
* Missing Values coded as “”
* Invalid entries: “-2, 1006 1006, 1240 1240, 382 382”

“Selling.Agent.Phone”

* Phone number of the selling agent
* (XXX) XX-XXXX
* Missing Values coded as “”
  + 25% missing

“Selling.Agent.Email”

* Email of the selling agenct
* String
* Missing Values coded as “”
  + 36% missing

“Co.Selling.Agent”

* Name of the co-selling agent
* String
* Missing Values coded as “”
  + 96% missing

Ltd.Service.Agent

* Limited Service Agent involved?
* “Yes/No”
* Missing Values coded as “”
  + 6 missing

“Owner.Phone”

* Phone number of the owner
* (XXX) XX-XXXX
* Missing Values coded as “”
  + 90% missing

“Owner.Name”

* Name of the owner
* String
* Inconsistent missing value codes: “”, NA
* Inconsistent name withheld code: WITHHELD, name withheld
* Inconsistent strings for same owner (e.g. Freddie Mac/Homesteps, Homesteps, Freddie Mac)

**Selling**

**\*\*\* Needs Cleaning**

**\*\*\* Cleaned**

“List.Number”

* The number used in the MLS database
* XXXXXXXX
* All Values are unique and there are no NAs

“Data\_Year”

* Original data set year (SDAL created to easily extract years of data)
* XXXX

“Property.Type

* Type of property/building
* “Residential”

“Card.Format”

* The card format
* “Residential”

“End.Date”

* The deadline date for the sale of the property
* m/d/yy
* None missing

“Listing.Date”

* Date when property was listed
* m/d/yy
* None missing

“Sold.Date”

* Date when property was sold
* m/d/yy
* None missing

“Status”

* Status of the property
* “C”
* None missing

“Status.Change”

* Date of status change of the property
* m/d/yy
* None missing

“Contingent”

* Offer made on a property; seller can accept or decline
* Factor
* Inconsistent factor possibly (n, N)
* Missing values are coded as “”
  + 57% missing

“Cont..Remarks”

* Comments made by the contingent
* Factor
* Missing are coded as “”
  + 38% missing

“Orig..List.Price”

* Original price of property
* Numeric (0 – 500000)
  + 2 are coded as 0
* 1 missing

“Price”

* Actual price of property
* Numeric (10000 – 350000)
* 1 missing

“Sold.Price”

* Price property sold at
* Numeric (24500 – 300000)
* None missing

“Low.Price”

* Lowest price offered
* Numeric (135000 – 849900)
* 99% Missing

“sell\_broker\_comm”

* The compensation offered to the selling broker/agent
* “Yes/No”
* Inconsistent missing code: “”, NA
  + 70% missing

“Variable.Commission

* Was there a variable commission?
* “Yes/No”
* Missing value coded as “”
  + 75 missing

“Foreclosure”

* Is it a foreclosure or not?
* “Yes/No”
* Missing value coded as “”
  + 21% missing

“mod\_timestamp”

* The time a listing was modified
* XX:XX.X
* None missing
* Invalid time: (e.g. 59:51.1)

“Occupied.By”

* The availability of the property
* “Owner, Tenant, Vacant”
* Missing value coded as “”
  + 13% missing

“Owner.Agent”

* Is there an owner/agent?
* “Y/N”
* Missing value coded as “”
  + 4 missing

“Listing.Type”

* Type of listing of the property
* “EA, ERS, OTH, RP, RP-ERS, VRC”
* None missing

“Seller.Concessions”

* Did the seller provide concessions
* “Yes, No, S”
* Missing value coded as “”
  + 45% missing

“Special.Assmnts”

* Are there any special assessments?
* Factor
* Inconsistent coding ($value, value, annually, club…)
* Inconsistent missing value coding: “”, -, --, ---, NA

“SellingBroker.Incent”

* Used if any incentives are offered to the selling broker/agent
* “Yes/No”
* Invalid code: Select Onr
* Inconsistent missing value code: “”, NA
  + 47% missing

“Describe.Concession”

* Used if the seller has allowed disclosure of any concessions given at closing ($5,000 carpet allowance for example)
* String
* Inconsistent missing value coding: “”, NA
* Inconsistent 0 value coding: 0, none

“How.Sold”

* Method of payment used to buy the property
* Levels: "Assumption" ,"Cash","Contract For Deed", "Conventional","Farmers Home","FHA","Other","Owner Held","VA"
* No Missing Values

“Selling.Broker.Comp”

* The compensation offered to the selling broker/agent
* Numeric (0-18300)
* 1 missing

“Est..Closing.Date”

* Estimated date of closing
* m/d/yy
* Inconsistent missing code: “”, NA
  + 42% missing

“Short.CompromiseSale”

* Was this a short or comprise sale
* “Yes/No”
* Inconsistent missing code: “”, NA
  + 81% missing

“Days.on.Market”

* The amount of days the property was on the market
* Numeric (-289 – 2192)
* No missing values
* Invalid entries (e.g -298)
  + 4 below 0

“Photo.URL”

* Link to photos of the property
* String
* Missing values coded as “”

“Public.Remarks”

* Comments made from the public
* String
* Missing values coded as “”

“Agent.Remarks”

* Comments made from the public
* String
* Missing values coded as “”

**Tax**

“List.Number”

* The number used in the MLS database
* XXXXXXXX
* All Values are unique and there are no NAs

“Data\_Year”

* Original data set year (SDAL created to easily extract years of data)
* XXXX

"Partial.Tax.Assmnt"

* Incomplete tax amount due to any of the following reasons: New or proposed construction, Incomplete construction, or Additional structures being built
* “Y/N”
* 3% Missing

"Tax.Year"

* Year taxes were paid
* Factor (Had to convert to a factor from a numeric)
* Invalid and inconsistent coding
  + 58 levels ranging from 0 to 8481

"Rollback.Taxes"

* The sum of the deferred tax plus simple interest
* Factor/Numeric
* Missing are coded as: “”, -, na, N/K, N/A, no, No, NO, none, None…”
  + 95% are coded as “”
* Invalid entries (e.g. by seller, current”

"Tax.Rate"

* The real estate tax rate per $100 of real estate assessed value
* Numeric
* 37% missing
* Invalid entries (e.g. rate in % or value of tax)

"Taxes"

* Taxes paid
* Numeric (0 – 563,000)
* None missing
* Unsure if 0 = missing or 0 paid
  + 7% are listed at 0

**No Data**

“box\_nbr”

“Address.2”

“Streetdirsuffix”

“carrier\_route”

“geo\_county”

“assessed\_val”

“stories”

“Book.Section”

“book\_sec”

“Withdraw.Date”

“Cancel.Date”

“high\_price”

“financing”

“Supplement.Attached”

"taxes"

“sqft1”

“sqft2”

“sqft3”

“sqft4”

“baths\_3\_4”

**Geographic Spread**

****

1. Will need to be cleaned (see below in Location Section) [↑](#footnote-ref-1)