Data Dictionary

Explanation of Important Variables included in Toy Dataset

Let's take a look at the summary of our toy data

```
#>   sequence_num     respondent_id_prop ImportParcelID     PropertyLandUseStndCode
#>  Min.   :1000060   Min.   :1.00e+10   Min.   :8.90e+06   Length:33411
#>  1st Qu.:1138342   1st Qu.:1.00e+10   1st Qu.:1.12e+07   Class :character
#>  Median :1259346   Median :2.00e+10   Median :1.20e+07   Mode  :character
#>  Mean   :1405396   Mean   :4.34e+10   Mean   :1.26e+07
#>  3rd Qu.:1574343   3rd Qu.:9.00e+10   3rd Qu.:1.29e+07
#>  Max.   :3141165   Max.   :9.00e+10   Max.   :1.70e+08
#>
#>  SalesPriceAmount    loan_amount    RecordingDate        InitialInterestRate
#>  Min.   :3.64e+03   Min.   :    1   Min.   :2005-01-03   Min.   : 0.000
#>  1st Qu.:2.50e+05   1st Qu.:  211   1st Qu.:2007-08-15   1st Qu.: 0.000
#>  Median :3.90e+05   Median :  326   Median :2010-07-15   Median : 0.000
#>  Mean   :5.21e+05   Mean   :  406   Mean   :2010-02-14   Mean   : 0.978
#>  3rd Qu.:6.06e+05   3rd Qu.:  485   3rd Qu.:2012-05-16   3rd Qu.: 0.000
#>  Max.   :1.54e+08   Max.   :78929   Max.   :2016-12-30   Max.   :11.600
#>                                                          NA's   :1
#>     TransId           YearBuilt    BuildingAreaSqFt PropertyCity       TotalBedrooms
#>  Min.   :6.96e+05   Min.   :1824   Min.   :     1   Length:33411       Min.   : 0.00
#>  1st Qu.:6.61e+06   1st Qu.:1949   1st Qu.:  1138   Class :character   1st Qu.: 2.00
#>  Median :7.73e+06   Median :1965   Median :  1524   Mode  :character   Median : 3.00
#>  Mean   :2.32e+07   Mean   :1967   Mean   :  1891                      Mean   : 3.22
#>  3rd Qu.:4.84e+07   3rd Qu.:1990   3rd Qu.:  2122                      3rd Qu.: 4.00
#>  Max.   :4.11e+08   Max.   :2019   Max.   :346537                      Max.   :72.00
#>                     NA's   :422    NA's   :783                         NA's   :65
#>  TotalCalculatedBathCount applicant_race_1    income          applicant_ethnicity
#>  Min.   : 0.00            Min.   :1.0      Length:33411       Min.   :1.0
#>  1st Qu.: 2.00            1st Qu.:5.0      Class :character   1st Qu.:2.0
#>  Median : 2.00            Median :5.0      Mode  :character   Median :2.0
#>  Mean   : 2.37            Mean   :4.5                         Mean   :1.9
#>  3rd Qu.: 3.00            3rd Qu.:5.0                         3rd Qu.:2.0
#>  Max.   :49.00            Max.   :7.0                         Max.   :4.0
#>  NA's   :140
#>  census_tract       rate_spread        applicant_sex  WFPC_risk2018   WFPC_risk2012
#>  Length:33411       Length:33411       Min.   :1.00   Min.   :    0   Min.   :    0
#>  Class :character   Class :character   1st Qu.:1.00   1st Qu.:    0   1st Qu.:    0
#>  Mode  :character   Mode  :character   Median :1.00   Median :    0   Median :    0
#>                                        Mean   :1.42   Mean   :   94   Mean   :  273
#>                                        3rd Qu.:2.00   3rd Qu.:    0   3rd Qu.:   20
#>                                        Max.   :4.00   Max.   :23972   Max.   :50063
#>                                                       NA's   :21
#>    wuiflag90       wuiflag00       wuiflag10       elevation
#>  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   : -16.9
```

The variables `sequence_num`, `respondent_id`, `TransID` and `ImportParcelID` are just what is used to identify unique loans/loaners/transactions/properties respectively.

Many of the variables included are self-explanatory, but a few are less so. As a note **zillow derived data** are in dollars, whereas **hmda derived data** are in thousands of dollars.

Main point of above - `SalesPriceAmount` is from ztrax, and is sale amount in dollars, while `loan_amount` and `income` are in thousands of dollars. The above issue should otherwise not be an issue.

A few variables are sort of inscrutable, let's go through these in a list.

`PropertyLandUseStndCode` - this is a ztrax variable that indicates land use. The data has been pre-cleaned to the following types of data - here is the mapping.

- 'RR101', **SFR**
- 'RR999', **Inferred SFR**
- 'RR102', **Rural Residence** (includes farm/productive land?)
- 'RR104', **Townhouse**
- 'RR105', **Cluster Home**
- 'RR106', **Condominium**
- 'RR107', **Cooperative**
- 'RR108', **Row House**
- 'RR109', **Planned Unit Development**
- 'RR113', **Bungalow**
- 'RR116', **Patio Home**
- 'RR119', **Garden Home**
- 'RR120' **Landominium**

`applicant_ethnicity` and `applicant_race` are categorical variables that code an applicant's race or ethnicity. These come with a list of code-mappings.

Ethnicity

- 1 - Hispanic or Latino
- 2 - Not Hispanic or Latino
- 3 - Information not Provided
- 4 - Not applicable
- 5 - No co-applicant (not relevant for us)

Race

- 1 - American Indian or Alaskan Native
- 2 - Asian
- 3 - Black or African American
- 4 - Native Hawaiian or other Pacific Islander
- 5 - White
- 6 - Information not Provided
- 7 - Not applicable
- 8 - No co-applicant (not relevant)

Hedonic Variables

I have also included a set of potentially useful hedonic/property characteristic variables. One of note is `elevation` which is in meters, and is the elevation of the point location of the property.

Square footage, bedrooms, year of home construction and bathrooms are also provided.

Location can be found through the `census_tracts` variable.

Fire Risk Variables

There are a few different measures of fire risk available for analysis, depending on preference.

- **WFPC** (either 2012 or 2018) - this is the wildfire potential variable. Most observations will have a value of 0 - but there are some significantly larger levels. From experience, running regressions with logged values of this work better, but I'm not sure if that applies for a sorting analysis.
- **wuiflag** (available for 1990, 2000, and 2010) - this is a binary variable for if the home's point location is in the wildland-urban-interface, as decided in 1990, 2000 or 2010 respectively.

Let me know if you need any more information