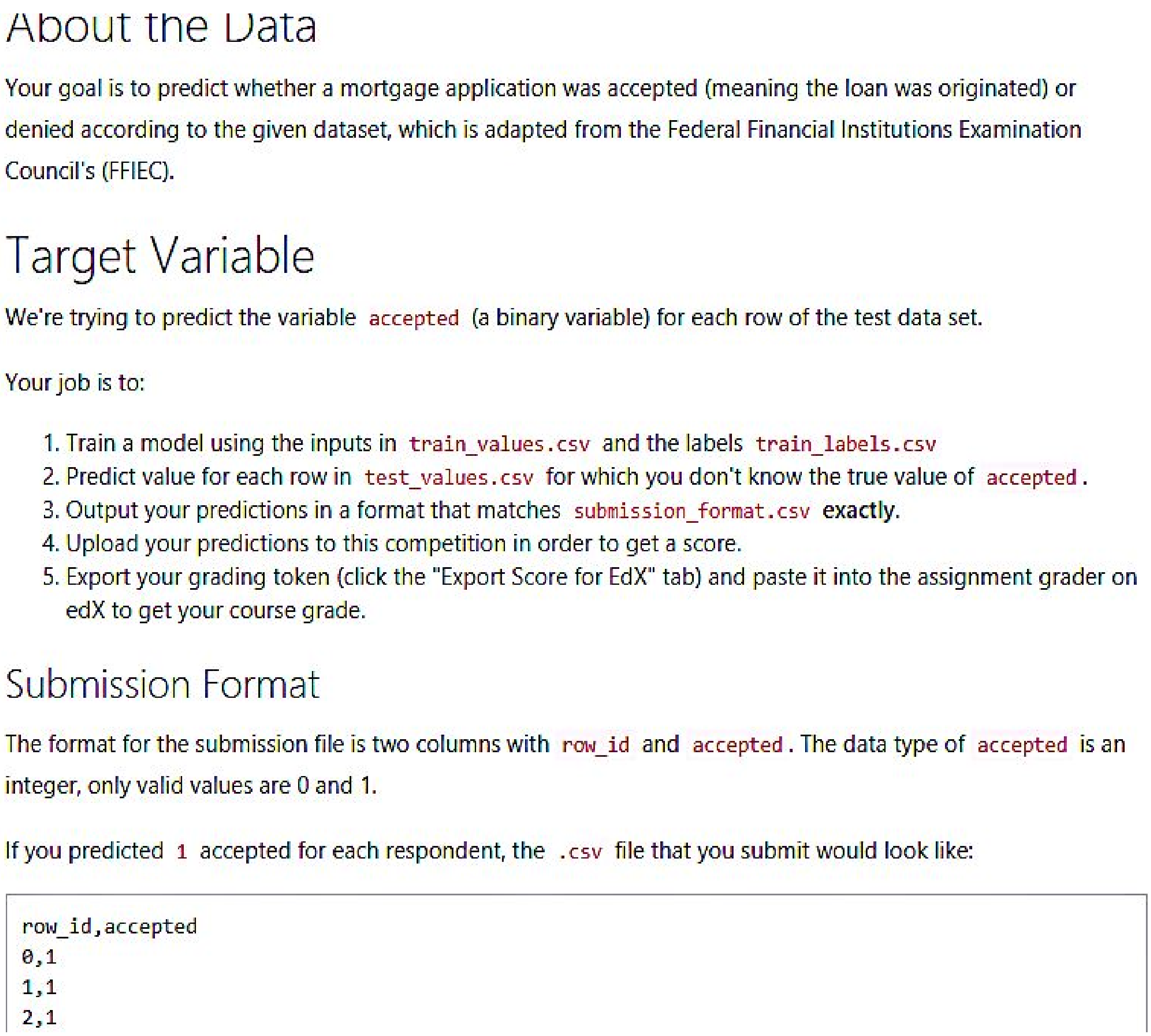


About the Data

Your goal is to predict whether a mortgage application was accepted (meaning the loan was originated) or denied according to the given dataset, which is adapted from the Federal Financial Institutions Examination Council's (FFIEC).

Target Variable

We're trying to predict the variable `accepted` (a binary variable) for each row of the test data set.

Your job is to:

1. Train a model using the inputs in `train_values.csv` and the labels `train_labels.csv`
2. Predict value for each row in `test_values.csv` for which you don't know the true value of `accepted`.
3. Output your predictions in a format that matches `submission_format.csv` **exactly**.
4. Upload your predictions to this competition in order to get a score.
5. Export your grading token (click the "Export Score for EdX" tab) and paste it into the assignment grader on edX to get your course grade.

Submission Format

The format for the submission file is two columns with `row_id` and `accepted`. The data type of `accepted` is an integer, only valid values are 0 and 1.

If you predicted `1` accepted for each respondent, the `.csv` file that you submit would look like:

```
row_id,accepted
0,1
1,1
2,1
```

Submission Format

The format for the submission file is two columns with `row_id` and `accepted`. The data type of `accepted` is an integer, only valid values are 0 and 1.

If you predicted `1` accepted for each respondent, the `.csv` file that you submit would look like:

```
row_id,accepted
0,1
1,1
2,1
3,1
4,1
⋮
```

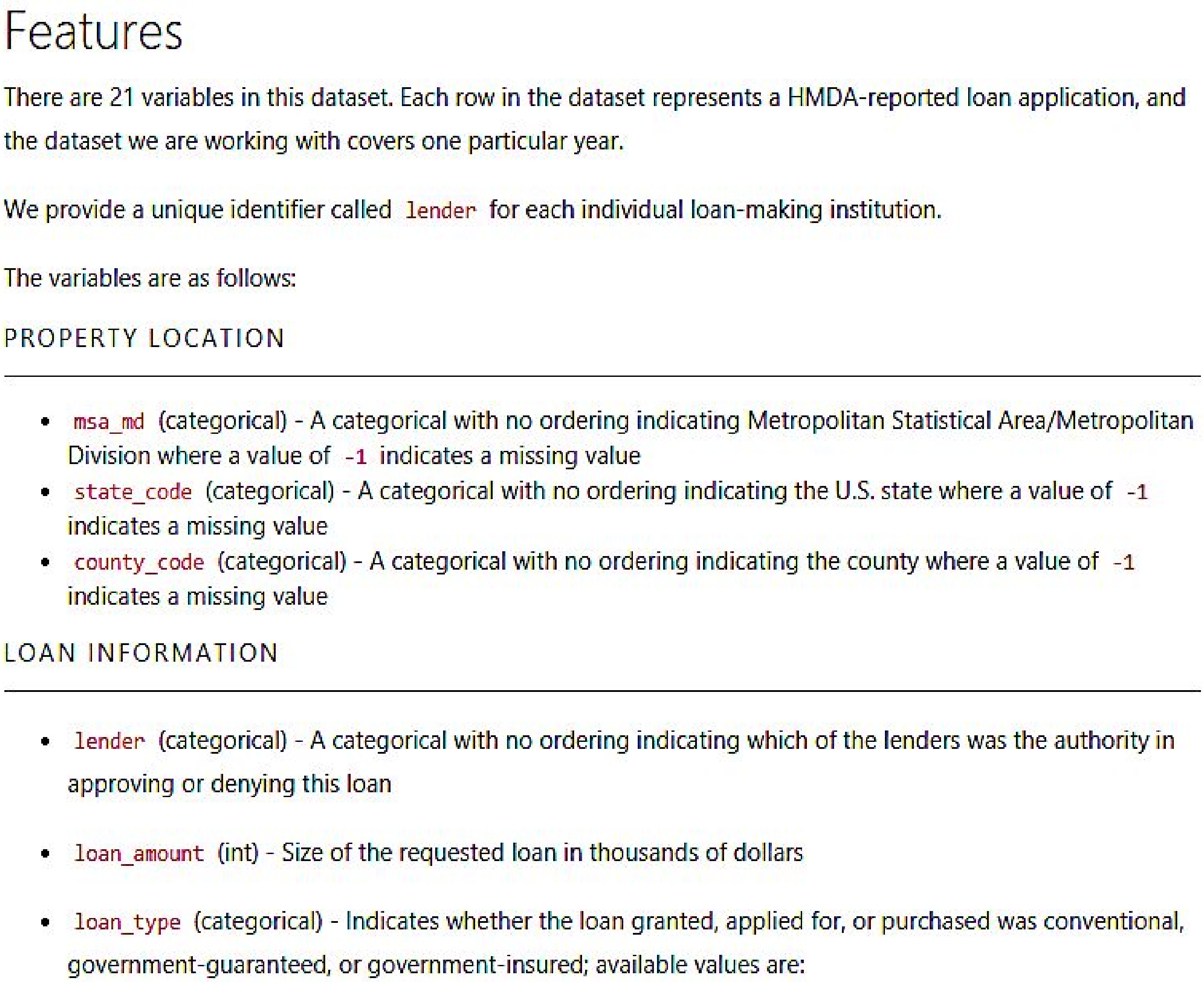
Performance Metric

We're predicting a binary variable, so this is a classification problem. To measure classification, we'll use a metric known as accuracy (also known as "classification rate"). Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. For binary classification, accuracy can be calculated in terms of positives and negatives as follows:

$$\text{accuracy} = (TP + TN)/(TP + TN + FP + FN)$$

where TP means True Positives, TN means True Negatives, FP means False Positives, and FN means False Negatives.

Features



Features

There are 21 variables in this dataset. Each row in the dataset represents a HMDA-reported loan application, and the dataset we are working with covers one particular year.

We provide a unique identifier called `lender` for each individual loan-making institution.

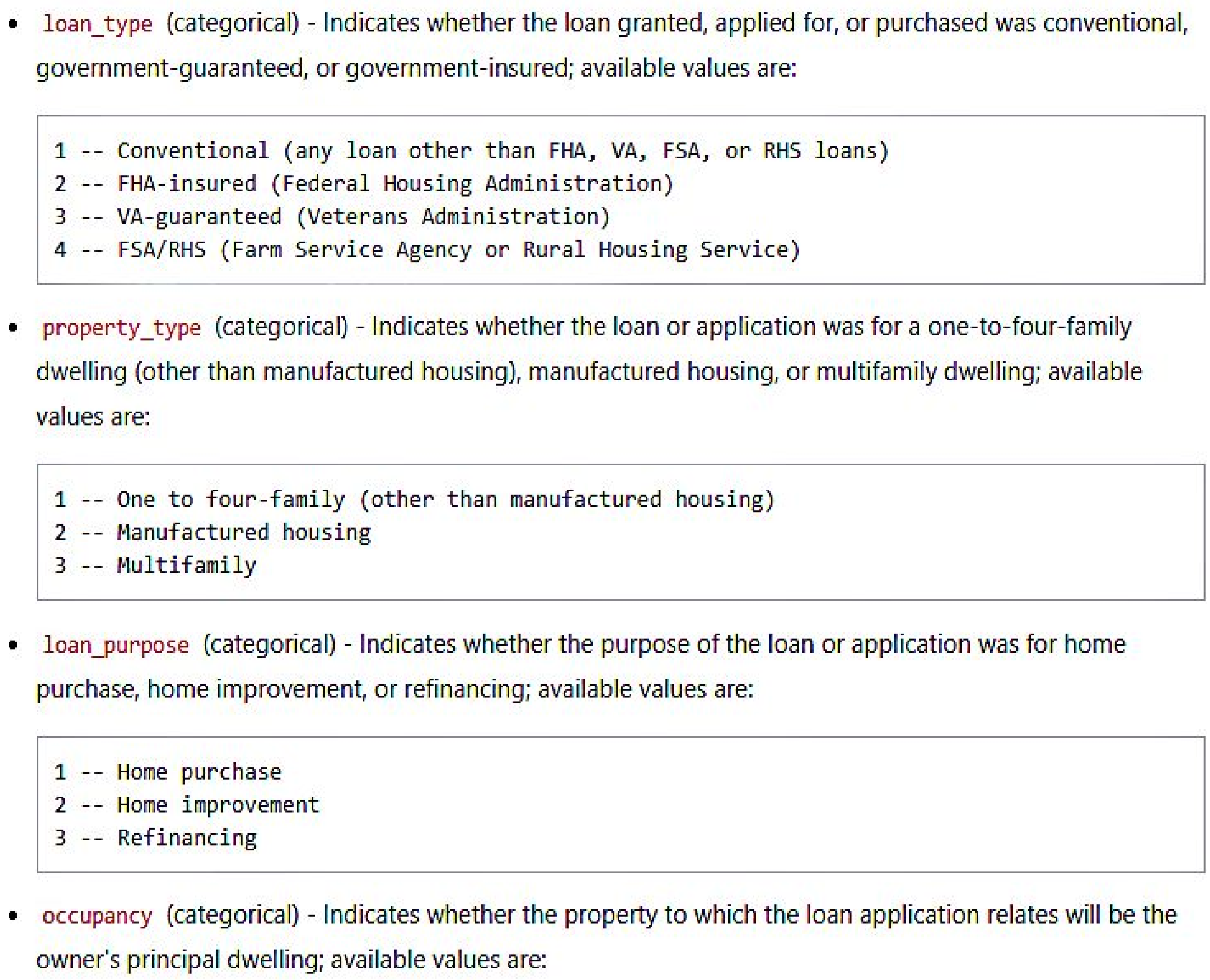
The variables are as follows:

PROPERTY LOCATION

- `msa_md` (categorical) - A categorical with no ordering indicating Metropolitan Statistical Area/Metropolitan Division where a value of `-1` indicates a missing value
- `state_code` (categorical) - A categorical with no ordering indicating the U.S. state where a value of `-1` indicates a missing value
- `county_code` (categorical) - A categorical with no ordering indicating the county where a value of `-1` indicates a missing value

LOAN INFORMATION

- `lender` (categorical) - A categorical with no ordering indicating which of the lenders was the authority in approving or denying this loan
- `loan_amount` (int) - Size of the requested loan in thousands of dollars
- `loan_type` (categorical) - Indicates whether the loan granted, applied for, or purchased was conventional, government-guaranteed, or government-insured; available values are:



- `loan_type` (categorical) - Indicates whether the loan granted, applied for, or purchased was conventional, government-guaranteed, or government-insured; available values are:

```
1 -- Conventional (any loan other than FHA, VA, FSA, or RHS loans)
2 -- FHA-insured (Federal Housing Administration)
3 -- VA-guaranteed (Veterans Administration)
4 -- FSA/RHS (Farm Service Agency or Rural Housing Service)
```

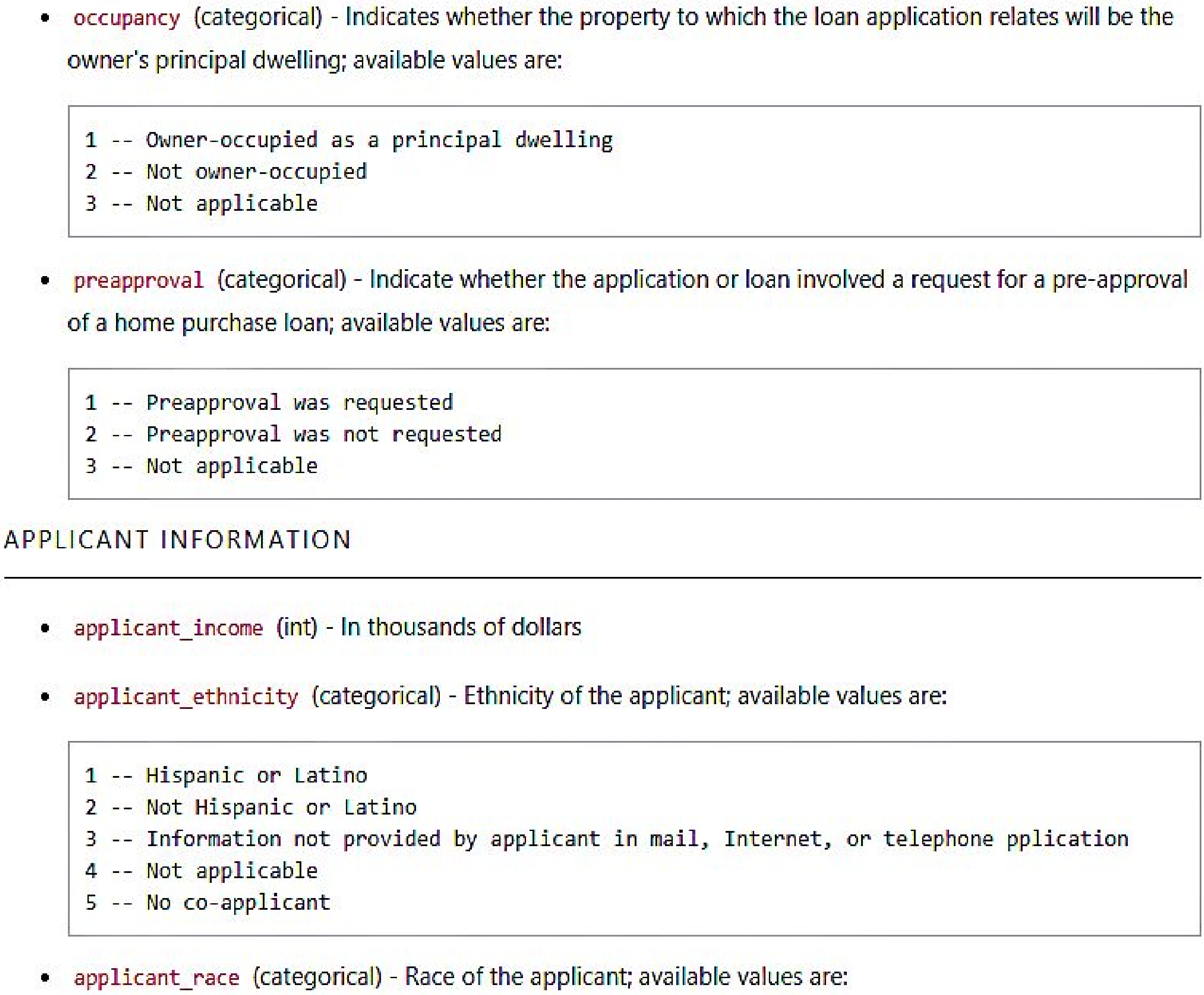
- `property_type` (categorical) - Indicates whether the loan or application was for a one-to-four-family dwelling (other than manufactured housing), manufactured housing, or multifamily dwelling; available values are:

```
1 -- One to four-family (other than manufactured housing)
2 -- Manufactured housing
3 -- Multifamily
```

- `loan_purpose` (categorical) - Indicates whether the purpose of the loan or application was for home purchase, home improvement, or refinancing; available values are:

```
1 -- Home purchase
2 -- Home improvement
3 -- Refinancing
```

- `occupancy` (categorical) - Indicates whether the property to which the loan application relates will be the owner's principal dwelling; available values are:



- occupancy (categorical) - Indicates whether the property to which the loan application relates will be the owner's principal dwelling; available values are:

```
1 -- Owner-occupied as a principal dwelling
2 -- Not owner-occupied
3 -- Not applicable
```

- preapproval (categorical) - Indicate whether the application or loan involved a request for a pre-approval of a home purchase loan; available values are:

```
1 -- Preapproval was requested
2 -- Preapproval was not requested
3 -- Not applicable
```

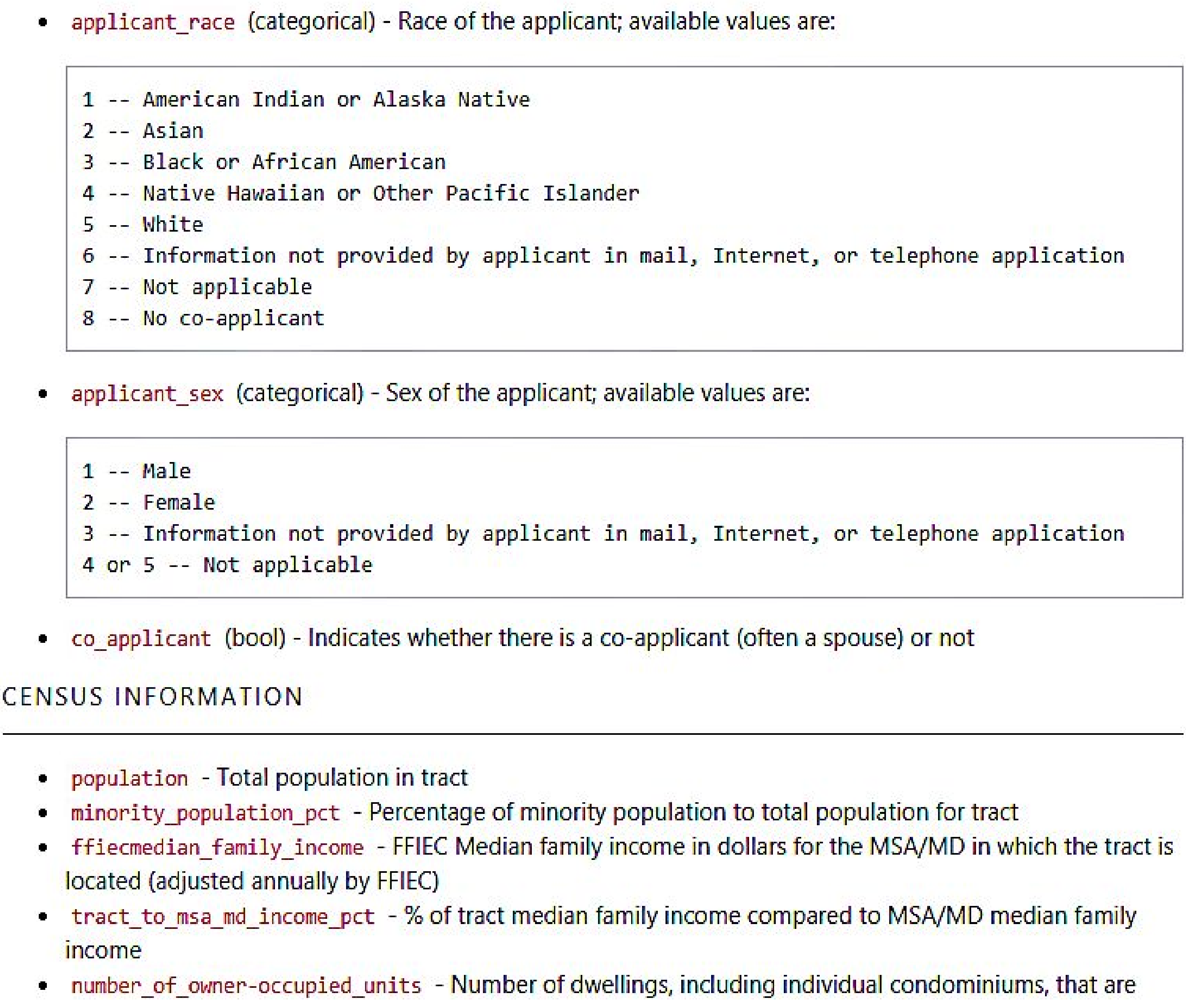
APPLICANT INFORMATION

- applicant_income (int) - In thousands of dollars

- applicant_ethnicity (categorical) - Ethnicity of the applicant; available values are:

```
1 -- Hispanic or Latino
2 -- Not Hispanic or Latino
3 -- Information not provided by applicant in mail, Internet, or telephone pplication
4 -- Not applicable
5 -- No co-applicant
```

- applicant_race (categorical) - Race of the applicant; available values are:



- applicant_race (categorical) - Race of the applicant; available values are:

```
1 -- American Indian or Alaska Native
2 -- Asian
3 -- Black or African American
4 -- Native Hawaiian or Other Pacific Islander
5 -- White
6 -- Information not provided by applicant in mail, Internet, or telephone application
7 -- Not applicable
8 -- No co-applicant
```

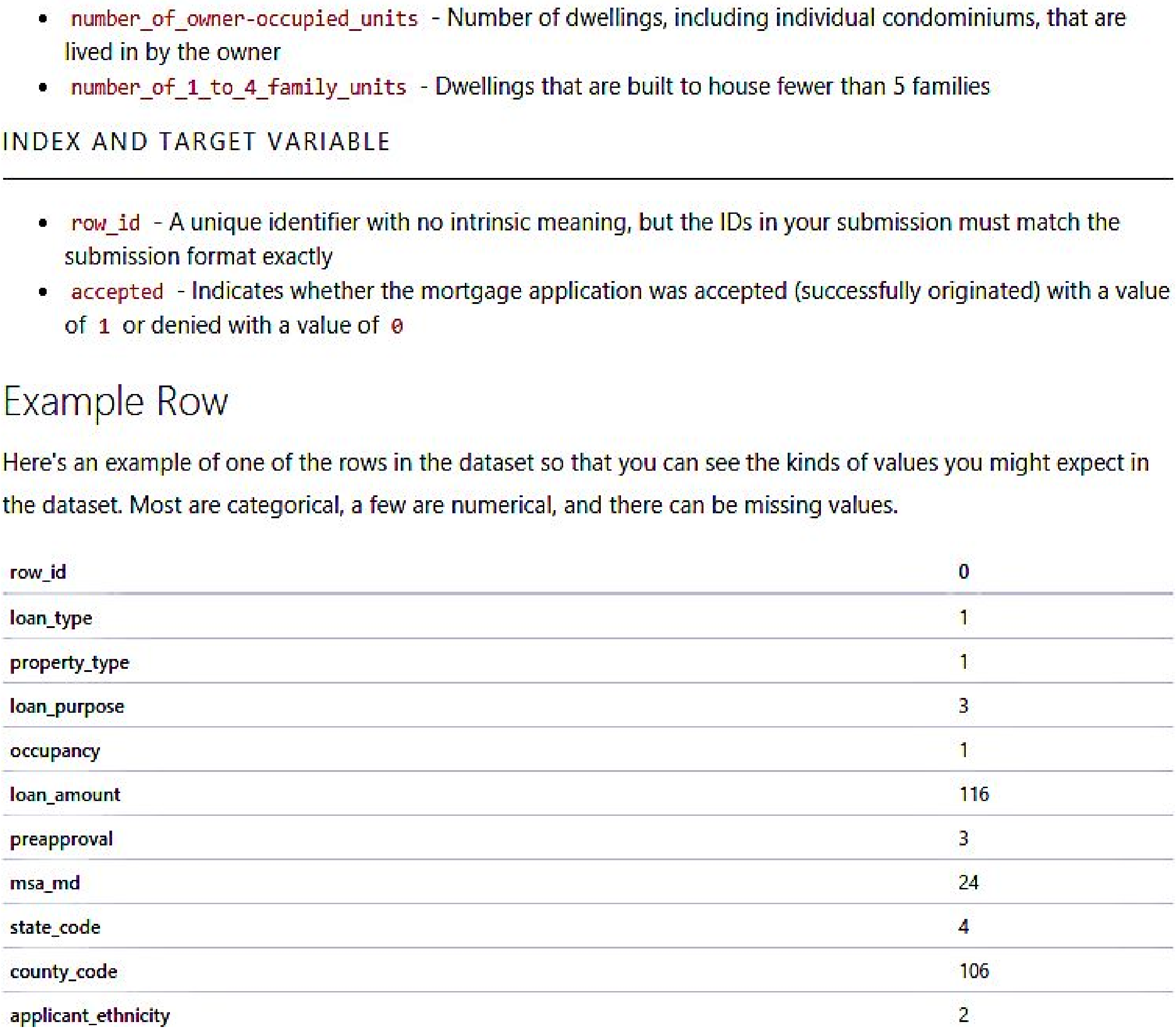
- applicant_sex (categorical) - Sex of the applicant; available values are:

```
1 -- Male
2 -- Female
3 -- Information not provided by applicant in mail, Internet, or telephone application
4 or 5 -- Not applicable
```

- co_applicant (bool) - Indicates whether there is a co-applicant (often a spouse) or not

CENSUS INFORMATION

- population - Total population in tract
- minority_population_pct - Percentage of minority population to total population for tract
- ffiecmedian_family_income - FFIEC Median family income in dollars for the MSA/MD in which the tract is located (adjusted annually by FFIEC)
- tract_to_msa_md_income_pct - % of tract median family income compared to MSA/MD median family income
- number_of_owner-occupied_units - Number of dwellings, including individual condominiums, that are



- number_of_owner-occupied_units - Number of dwellings, including individual condominiums, that are lived in by the owner
- number_of_1_to_4_family_units - Dwellings that are built to house fewer than 5 families

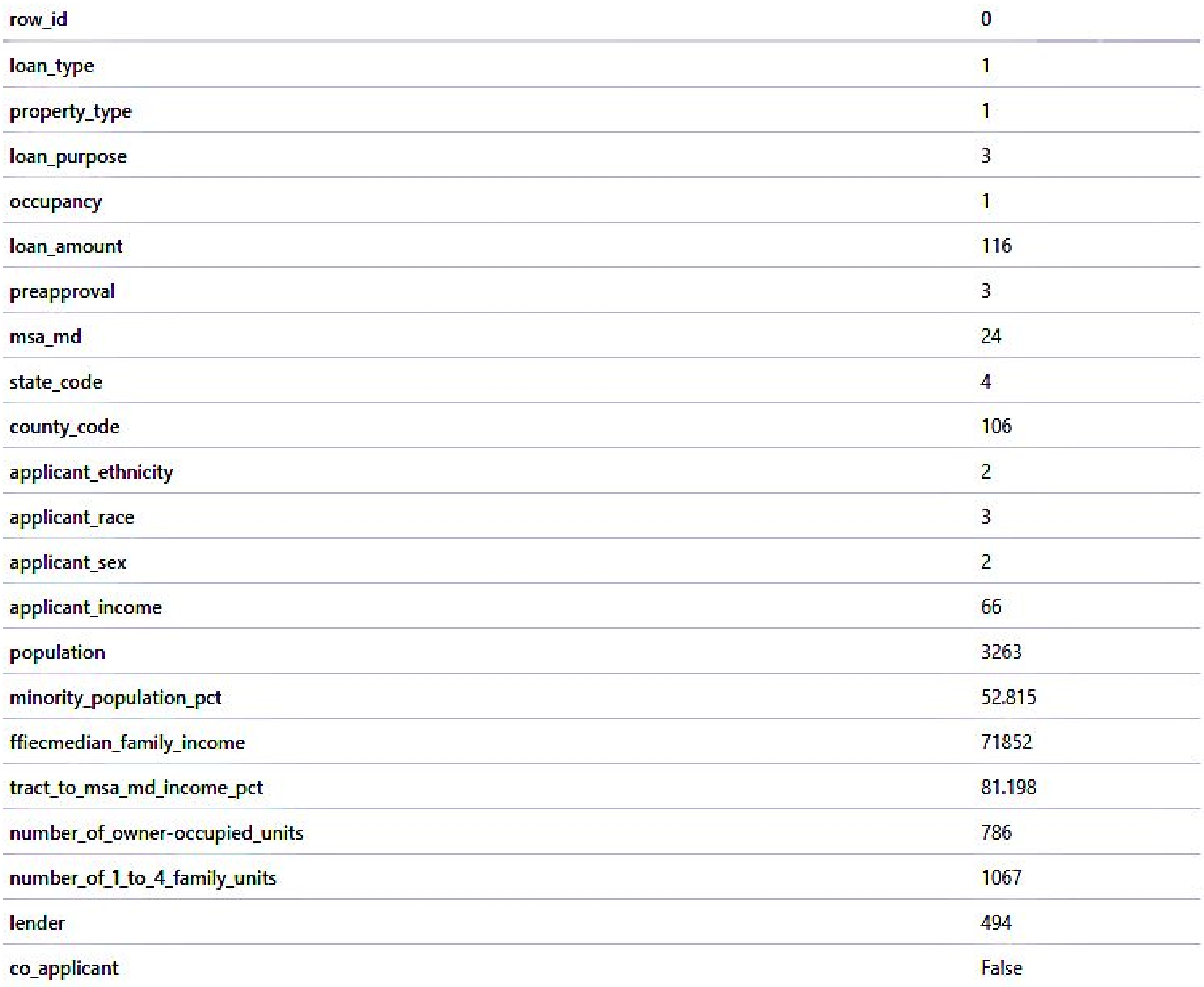
INDEX AND TARGET VARIABLE

- row_id - A unique identifier with no intrinsic meaning, but the IDs in your submission must match the submission format exactly
- accepted - Indicates whether the mortgage application was accepted (successfully originated) with a value of 1 or denied with a value of 0

Example Row

Here's an example of one of the rows in the dataset so that you can see the kinds of values you might expect in the dataset. Most are categorical, a few are numerical, and there can be missing values.

row_id	0
loan_type	1
property_type	1
loan_purpose	3
occupancy	1
loan_amount	116
preapproval	3
msa_md	24
state_code	4
county_code	106
applicant_ethnicity	2



row_id	0
loan_type	1
property_type	1
loan_purpose	3
occupancy	1
loan_amount	116
preapproval	3
msa_md	24
state_code	4
county_code	106
applicant_ethnicity	2
applicant_race	3
applicant_sex	2
applicant_income	66
population	3263
minority_population_pct	52.815
ffiecmedian_family_income	71852
tract_to_msa_md_income_pct	81.198
number_of_owner-occupied_units	786
number_of_1_to_4_family_units	1067
lender	494
co_applicant	False