Review and Preliminary Mortgage Analysis

SCALABLE DATA PROCESSING IN R

Michael Kane
Assistant Professor, Yale University

datacamp

Overview of the chapter

- Compare proportions of people receiving mortgages
- Missingness in the data
- Changes in
  - Mortgage demographic proportions over time
  - City vs rural mortgages
  - Proportion of people securing federally guaranteed loans

United States Census Bureau Race and Ethnic Proportions

Category	Percentge
American Indian or Alaska Native	0.9
Asian	4.8
Black or African American	12.6
Native Hawaiian or Other Pacific Islander	0.2
Two or more races (Not included)	2.9
Other race (Not included)	6.2
Hispanic or Latino ethnicity	16.3

Note: Hispanic or Latino is a designated ethnicity

Proportional Borrowing

We know that most mortgages went to people who identify as white.

> Is this group borrowing more proportionally?

Let's practice!

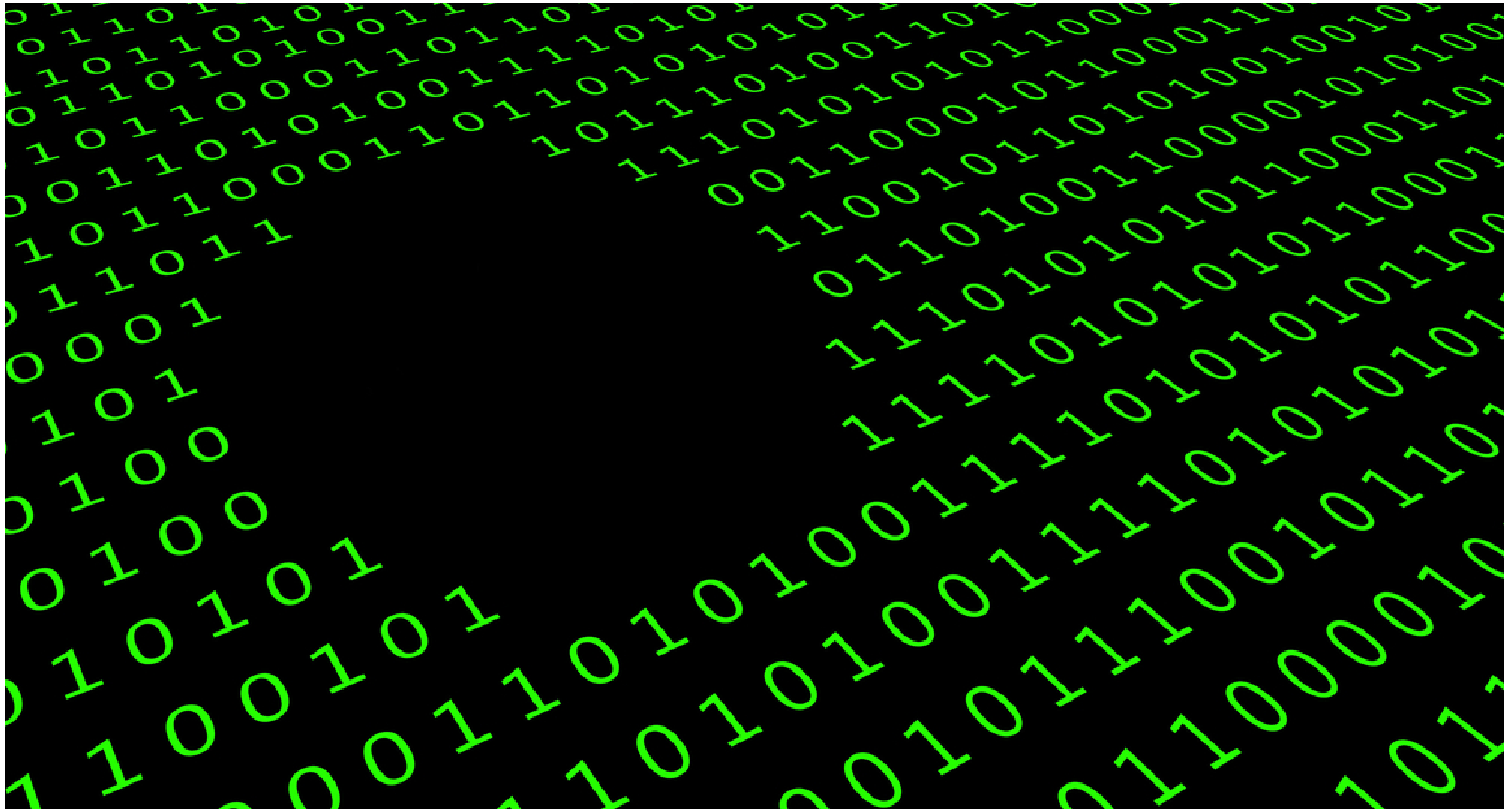
SCALABLE DATA PROCESSING IN R

Are the data missing at random?

SCALABLE DATA PROCESSING IN R

Michael Kane
Assistant Professor, Yale University

datacamp



Types of Missing Data

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random (MNAR)

MCAR

Missing Completely at Random

- There is no way to predict which values are missing
- Can drop missing data

MAR

Missing at Random

- Missingness is dependent on variables in the data set
- Use multiple imputation to predict what missing values could be

MNAR

Missing Not at Random

- Not MCAR or MAR
- Deterministic relationship between variables

Dealing with missing data in this course

- Full treatment of missingness is beyond the scope of this course
- We will check to see if it's plausible data are MCAR and drop missing values

A Quick Check for MAR

- Recode a column with one if the data is missing and zero otherwise
- Regress other variables onto it using a logistic regression
- Significant p-value indicates MAR
- Repeat for other columns with missingness
- Some p-values can be significant by chance, so adjust your cutoff for significance based on the number of regressions

MAR Quick Check Example

```
# Our dependent variable
is_missing <- rbinom(1000, 1, 0.5)
# Our independent variables
data_matrix <- matrix(rnorm(1000*10), nrow = 1000,
                      ncol = 10)

# A vector of p-values we'll fill in
p_vals <- rep(NA, ncol(data_matrix))
```

MAR Quick Check Example

```
# Perform logistic regression
for (j in 1:ncol(data_matrix)) {
 s <- summary(glm(is_missing ~ data_matrix[, j]),
              family = binomial)
              p_vals[j] <- s$coefficients[2, 4]
 }
# Show the p-values
p_vals
```

```
0.5930082 0.7822695 0.7560343 0.3689330 0.8757048
0.8812320 0.8281008 0.4888898 0.4781299 0.5655739
```

Let's practice!

SCALABLE DATA PROCESSING IN R

Analyzing the Housing Data

SCALABLE DATA PROCESSING IN R

Simon Urbanek

Member of R-Core, Lead Inventive Scientist, AT&T Labs Research

So far ..

- Compare different demographic groups in data
- Quick check to see if data are missing at random

Adjusted Counts and Proportional Change by Year

- Adjusting group size lets you compare different groups as if they were the same size
- Proportional change shows growth (or decline) of a group

Let's practice!

SCALABLE DATA PROCESSING IN R

Other Lending Trends

SCALABLE DATA PROCESSING IN R

Simon Urbanek

Member of R-Core, Lead Inventive Scientist, AT&T Labs Research

In this lesson ...

- City vs rural
- Federally guaranteed loans vs. income

City vs. Rural

- City means a home is in a metropolitan area, otherwise rural
- In the mortgage data set, city has `msa` value of 1, 0 otherwise
- For a more precise definition see **FHFA website**

Federally Guaranteed Loans and Borrower Income

- Federally guaranteed loans protect the company issuing a loan
- If a lender can issue a federally guaranteed loan, then the lender is less worried about the loan defaulting as the government will buy the loan
- We'll use Borrower Income Ratio: borrower income divided by median income of people in the area

Let's practice!

SCALABLE DATA PROCESSING IN R

Congratulations!

SCALABLE DATA PROCESSING IN R

Michael J. Kane and Simon Urba...

Instructors, DataCamp

Split-Apply-Combine

- Break the data into parts
- Compute on the parts
- Combine the results

Split-Apply-Combine: Advantages

- Manageable parts don't overwhelm your computer
- Approach is easy to parallelize
- Process sequentially
- Process on serveral machines in a cluster

Split-Apply-Combine: R

- `split()` partitions set of row numbers or `data.frame`
- `Map()` computes on parts
- `Reduce()` combines results

bigmemory

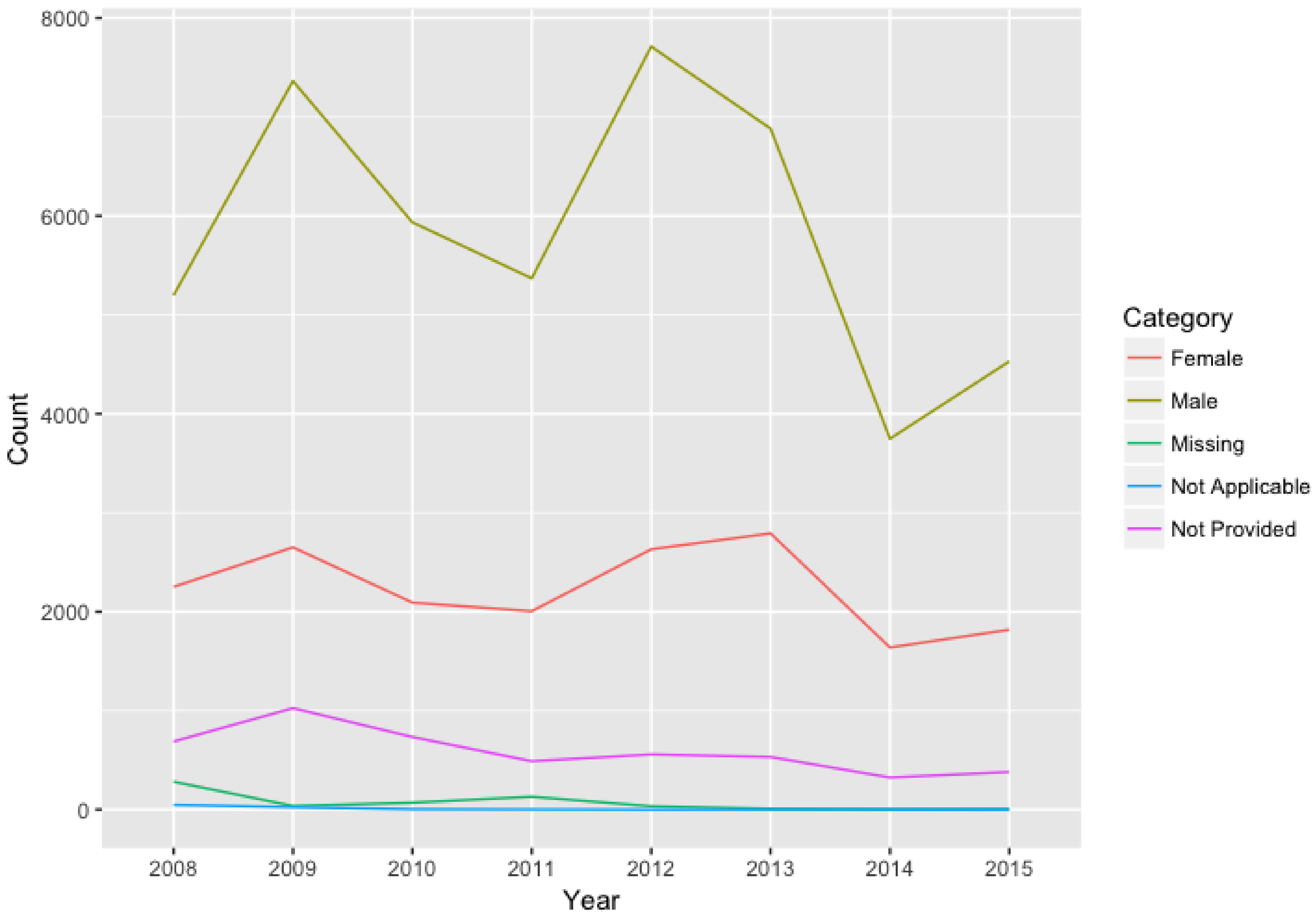
bigmemory

- Good for larger data sets that can be represented as dense matrices and might be too big for RAM
- Looks like a regular R matrix

iotools

iotools

- Good for much larger data that can be processed in sequential chunks
- Supports `data.frame` and `matrix`


8000
6000
4000
2000
0
Count
2008
2009
2010
2011
2012
2013
2014
2015
Year
Category
Female
Male
Missing
Not Applicable
Not Provided

Good luck!

SCALABLE DATA PROCESSING IN R