MONASH University

ETC5512: Wild Caught Data

Case Study: Mortgage Default

Lecturer: *Kate Saunders*

Department of Econometrics and Business Statistics

ETC5512.Clayton-x@monash.edu

Week 7

Have you ever gotten a loan from the bank?
Do banks offer the same interest rate?
Ever wondered why banks often do not interest rate match?
(think of it like price match!)

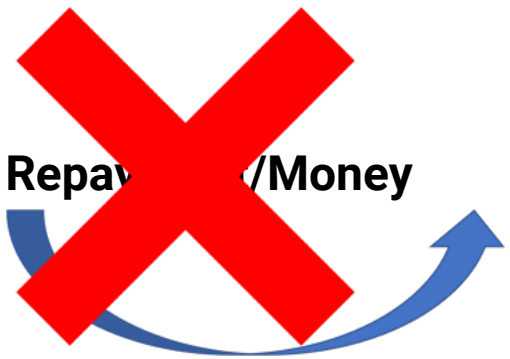
CREDIT RISK

Credit Risk

Lender

Mortgage

Repay /Money



Today you will:

- Look at a credit risk case study
- Use the Fannie Mae data set containing US loan data
- Learn some tricks for doing an analysis using big data

Coding Perspective:

- learn methods for wrangling a large data set
- learn about how to read / deal with zip files
- touch very briefly on functions
- learn about storage in R

Fannie Mae and Freddie Mac

Fannie Mae

- established in 1938
- Federal National Mortgage Association (U.S.)
- Fannie Mae is regulated by the Federal Housing Finance Agency (FHFA) and is a publicly traded company on the New Yock Stock Exchange (NYSE).
- initially it was a government sponsored enterprise in US and later switched to a private enterprise.
- Purpose: make loans and loan guarantees to low or middle income families.
- Purchases mortgages from lenders, which frees up the bank capital, allowing the bank to offer more loans and mortgages. Then later packages them into mortgage-backed securities, which are sold to investors on the secondary market.
- By purchasing mortgages from lenders and providing liquidity to the mortgage market, Fannie Mae helps to make homeownership more affordable and accessible for Americans.

Freddie Mac

- created in 1970
- Federal Home Mortgage Loan Corporation.
- same function as Fannie Mae.
- it is created to end Fannie Mae's monopoly on the secondary mortgage market.

Both are the two largest financial institutions in the world with the combined total mortgage assets of $1.4 trillion (as of 2020)

How they operate?

- Both of them own or guarantee just under half the total value of home loans in the U.S.
- They sell their mortgages as bonds and charge a fee.
- This bonds is called **Mortgage Backed Securities**.
  - an investment products that are created by pooling together a large number of individual mortgage loans into a single security.
  - the securities are then sold to investors on the secondary market.
  - the value of MBS depends on the underlying pool of mortgages

The more banks are able to sell mortgages to Fannie Mae and Freddie Mac, the more money banks can make!

What went wrong?

- Bank began making ***junk*** loans without checking the creditworthiness of the borrowers by simply selling it to the government sponsored enterprise to make more profits.
- Individuals can easily get loan regardless of they affordability.
- As the economy melt down, this caused the Government Sponsored Enterprise (GSE) to cover the difference for the investors and make a huge loss as more mortgage loan defaulted.
- Both GSE neared bankruptcy because of the sub-prime mortgage (sub-prime mortgage - the practice of lending money to people with low credibility at a high interest rate.)

Fannie Mae and Freddie Mac in COVID-19

- The federal government launched the Coronavirus Aid, Relief, and Economic Security (CARES) Act.
- This mortgage relief act offered protections for homeowners with mortgages backed by these GSE.
- This act expired on 31st July 2021. However, the borrowers under these GSE are eligible for 18 months of total forbearance as long as their plan is active by 28th February 2021.
- Due to this, you will observe "artificially" low default rate during the pandemic! The same happens in Australia!

Fannie Mae Data (Big Data!)

Fannie Mae Data Set

- Provides loan performance data on a portion of its single-family mortgage loans
- How can it helps? to gain insights into the drivers of mortgage default risk!
- The Single-Family Fixed Rate Mortgage (primary) dataset contains a subset of Fannie Mae's 30-year and less, fully amortizing, full documentation, single-family, conventional fixed-rate mortgages.
- Data available from 2000 onwards, mortgage loans originated prior to 1999 are excluded.
- Every quarter following the initial release, Fannie Mae updates acquisition and performance data as of the previous quarter.
- Fannie Mae releases updated information on or after the 20th of the month following the end of the quarter.
- Currently, they have about 100GB 😲 of historical data!

❓ What type of data is this?

Fannie Mae Single Family Loan Data

Fannie Mae Single-Family Loan Performance Data

Primary dataset: acquisition and performance data (~ 50 Gb)

Register a free account!

Download data for 2016 Q1 to 2023Q3 (zip files much smaller)

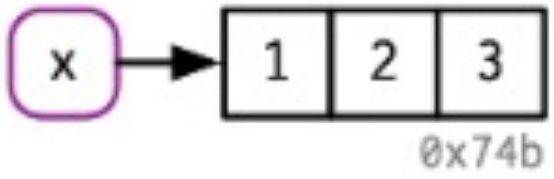
This data set is for mortgage loans, which is very similar to the bank's mortgage data. This helps you to understand how to estimate credit risk (at least as a start point 😁)

Digress

Names and Values in R

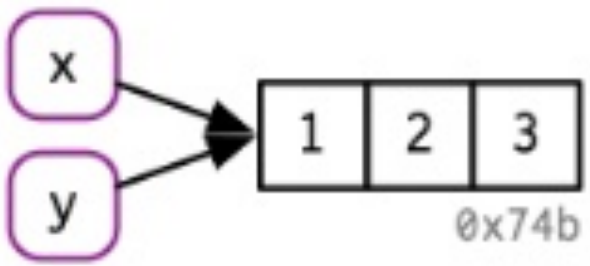
Names vs Values

```
x <- c(1, 2, 3)
```



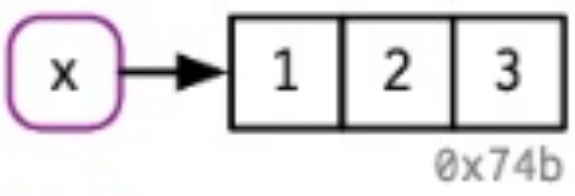

```
y <- x
```

--



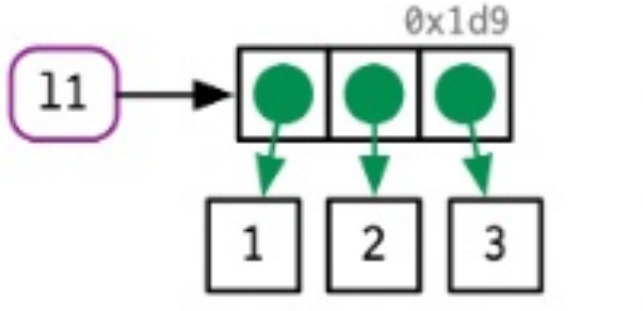
Copy on modify

```
x <- c(1, 2, 3)
y <- x

y[[3]] <- 4
x

## [1] 1 2 3

y

## [1] 1 2 4
```

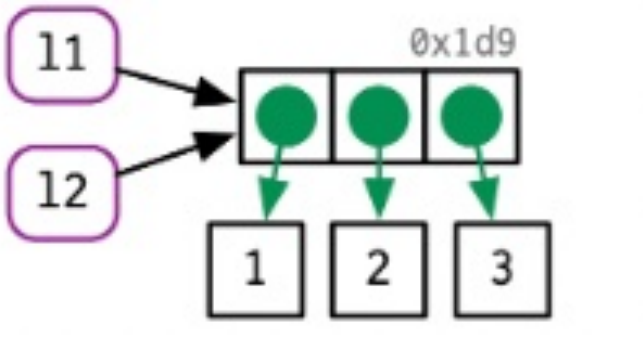


List

```
l1 <- list(1, 2, 3)
```

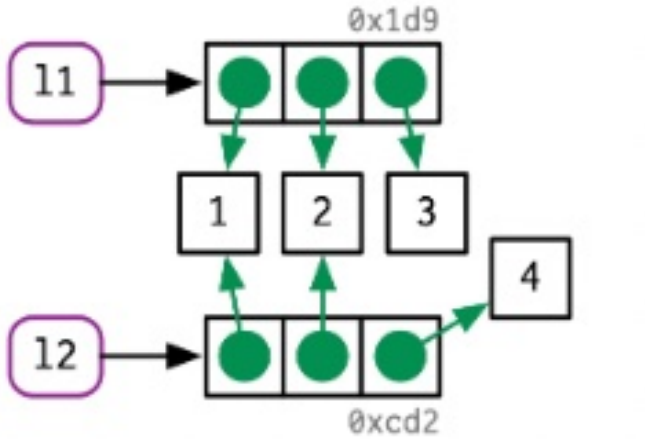



```
l2 <- l1
```



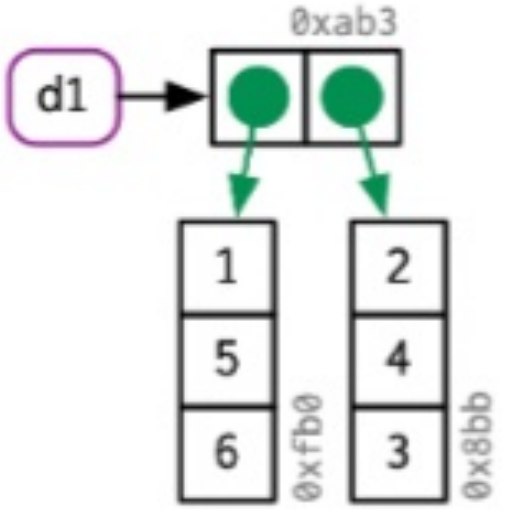
List (cont.)

```
l2[[3]] <- 4
```



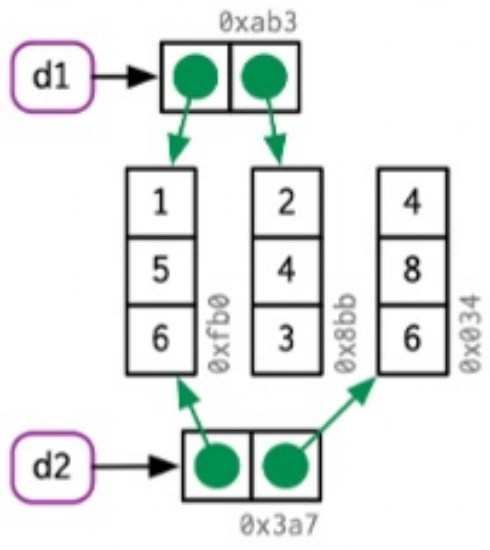
Data Frames

```
d1 <- data.frame(x = c(1, 5, 6), y = c(2, 4, 3))
```



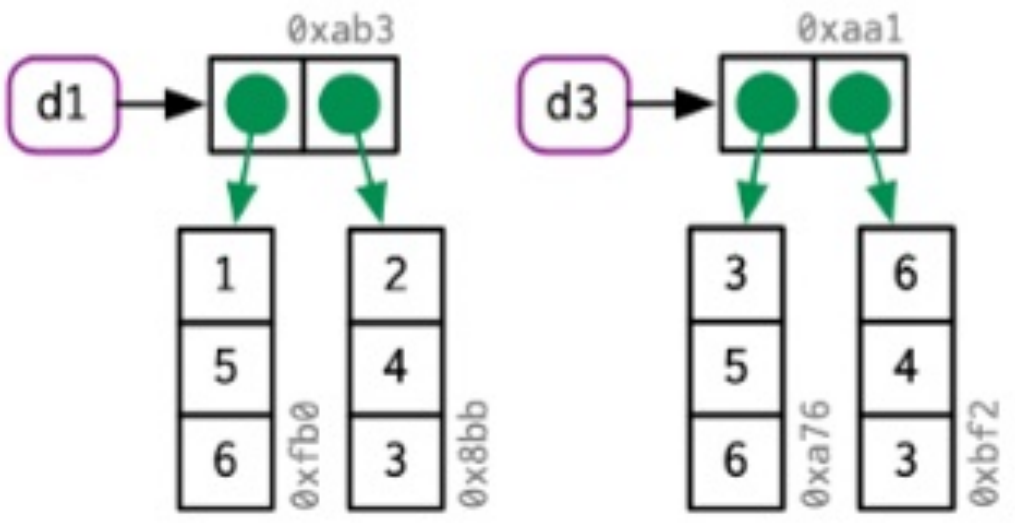
Data Frames (cont.)

```
d2 <- d1
d2[, 2] <- d2[, 2] * 2
```



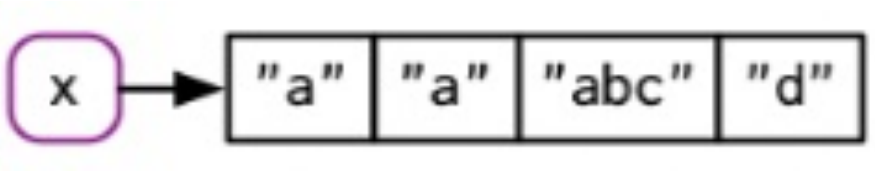
Data Frames (cont.)

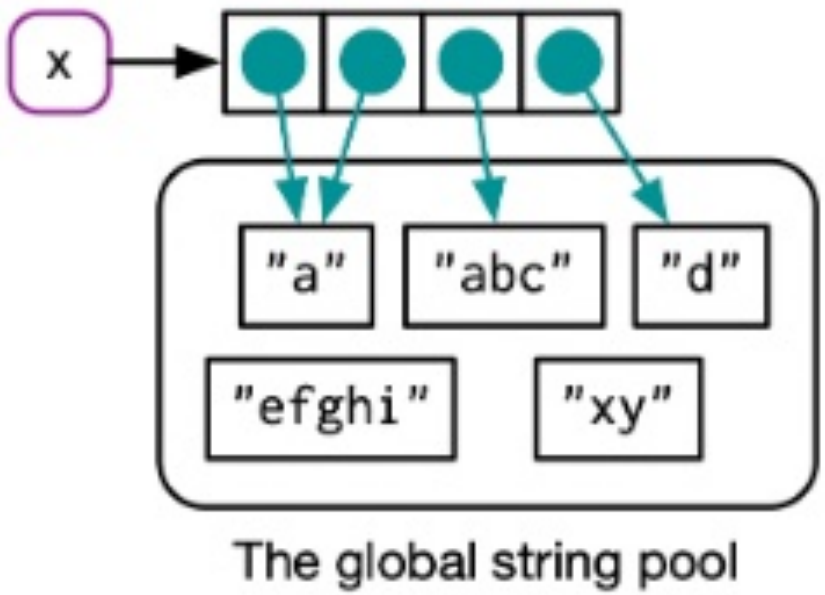
```
d3 <- d1
d3[1, ] <- d3[1, ] * 3
```



Character vectors

```
x <- c("a", "a", "abc", "d")
```





Some more tips

If you have big data it can be very important to think about how to fast your code runs. This requires understanding which parts take a *llllooooonnnnnnnnggggggggg* time.

A useful function for checking how big an object is in R (S3) is `object.size()`.

You can also time parts of your code to see which parts take time - here are 5 different examples

Now back to Fannie Mae

Getting Started

We will only use the single family fixed rate mortgage (primary) data set not the HARP!

1. Read the overview in the main page to get an idea about the data - Fannie Mae Single-Family Loan Performance Data
2. Understanding the variables. Choose the "Glossary and File Layout" file in the homepage (either csv or pdf, I prefer csv as it is easier to navigate, but the choice is yours!).
3. Read their FAQs.
4. Take a look at their data set using the sample provided in their website (sample File).

Fannie Mae Data Sample

POOL_ID	LOAN_ID	ACT_PERIOD	CHANNEL	SELLER	SERVICER	MASTER_SERVICER	ORIG_R
	100023020488	082009	R	Other	Other		5
	100023020488	092009	R	Other	Other		5
	100023020488	102009	R	Other	Other		5
	100023020488	112009	R	Other	Other		5
	100023020488	122009	R	Other	Other		5
	100023020488	012010	R	Other	Other		5
	100023020488	022010	R	Other	Other		5
	100023020488	032010	R	Other	Other		5
	100023020488	042010	R	Other	Other		5

Fannie Mae Data

Breakout session

- What's is stored in each zip file?
- What are the variable names and what do they mean?
- Is there anything in the FAQs I need to be aware of?

Fannie Mae Data

How to make the data at account level? One account one row!

Steps:

1. For each quarter make the file smaller by reducing the current transactional level data to single account data.
2. Read the raw data quarter by quarter.

R script that are used:

1. `00_read_data.R`.
2. `01_import_data.R`.
3. `02_combine.R`.

Cont.

Step 1: Load the raw data and split it into acquisition table and performance table.

```
######################################################################
# Here the Loan Performance data is modified into a one-loan-per-row dataset including key analytic data fields.
# We encourage exploration of this code to understand how certain fields in the statistical summary are derived.
######################################################################

#----Setup----
library(data.table)
library(tidyverse)

load_lppub_file <- function(filename, col_names, col_classes){
  file <- unzip(paste0("raw data/", filename))
  df <- fread(file, sep = "|", col.names = col_names, colClasses = col_classes)
  file.remove(paste0(substr(filename, 1, 6), ".csv"))
  return(df)
}

#----Define Tables----

lppub_column_names <- c("POOL_ID", "LOAN_ID", "ACT_PERIOD", "CHANNEL", "SELLER", "SERVICER",
                        "MASTER_SERVICER", "ORIG_RATE", "CURR_RATE", "ORIG_UPB", "ISSUANCE_UPB",
                        "CURRENT_UPB", "ORIG_TERM", "ORIG_DATE", "FIRST_PAY", "LOAN_AGE",
                        "REM_MONTHS", "ADJ_REM_MONTHS", "MATR_DT", "OLTV", "OCLTV",
                        "NUM_BO", "DTI", "CSCORE_B", "CSCORE_C", "FIRST_FLAG", "PURPOSE",
                        "PROP", "NO_UNITS", "OCC_STAT", "STATE", "MSA", "ZIP", "MI_PCT",
                        "PRODUCT", "PPMT_FLG", "IO", "FIRST_PAY_IO", "MNTHS_TO_AMTZ_IO",
```

Cont.

Step 2: Read the data quarter by quarter

```
##################################################
# Obj: Read in Fannie Mae Data from large zip file
##################################################

#----Setup----

list.of.packages <- c("MASS", "data.table", "tidyverse", "here", "stringr",
                      "lubridate", "ggplot2", "usmap", "gganimate",
                      "glmnet")
new.packages <- list.of.packages[!(list.of.packages %in% installed.packages()[,"Package"])]
if(length(new.packages)) install.packages(new.packages)
invisible(lapply(list.of.packages, require, character.only = TRUE))

#----Importing data----

setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

files <- list.files("raw data")

for (file in files){
  # Set up file names
  fileYear <- stringr::str_sub(file, 1, 4)
  fileQtr <- stringr::str_sub(file, 5, 6)
  FileName <- file
  flush.console()
  print(FileName)

  # Read data one by one
  source('00_read_data.R')
```

Cont.

Step 3: Combine all quarters data into 1 file.

```
########################################
# Obj: Combine all stat.csv files into 1
#      for analysis purposes
########################################

#----Define column headers and classes----
stat_column_names <- c(
  "LOAN_ID", "ORIG_CHN", "SELLER", "loan_age", "orig_rt", "orig_amt",
  "orig_trm", "oltv", "ocltv", "num_bo", "dti",
  "CSCORE_B", "FTHB_FLG", "purpose", "PROP_TYP", "NUM_UNIT",
  "occ_stat", "state", "zip_3", "mi_pct", "CSCORE_C",
  "relo_flg", "MI_TYPE", "AQSN_DTE", "ORIG_DTE", "FRST_DTE",
  "LAST_RT", "LAST_UPB", "msa", "FCC_COST", "PP_COST",
  "AR_COST", "IE_COST", "TAX_COST", "NS_PROCS", "CE_PROCS",
  "RMW_PROCS", "O_PROCS", "repch_flag", "LAST_ACTIVITY_DATE",
  "LPI_DTE", "FCC_DTE", "DISP_DTE", "SERVICER", "F30_DTE",
  "F60_DTE", "F90_DTE", "F120_DTE", "F180_DTE", "FCE_DTE",
  "F180_UPB", "FCE_UPB", "F30_UPB", "F60_UPB", "F90_UPB",
  "MOD_FLAG", "FMOD_DTE", "FMOD_UPB", "MODIR_COST", "MODFB_COST",
  "MODFG_COST", "MODTRM_CHNG", "MODUPB_CHNG", "z_num_periods_120", "F120_UPB",
  "CSCORE_MN", "ORIG_VAL", "LAST_DTE", "LAST_STAT", "COMPLT_FLG",
  "INT_COST", "PFG_COST", "NET_LOSS", "NET_SEV", "MODTOT_COST"
)
```

Fannie Mae Data

Default (Response variable):

- defaulted is defined as if the borrower fails to pay back the money in 90 days.

Notes:

- 90+ days past due shows borrower distress -> serious delinquencies.
- However the definition of default can varies.
- According to the law, once the payment is past due 60 days -> defaulted account.

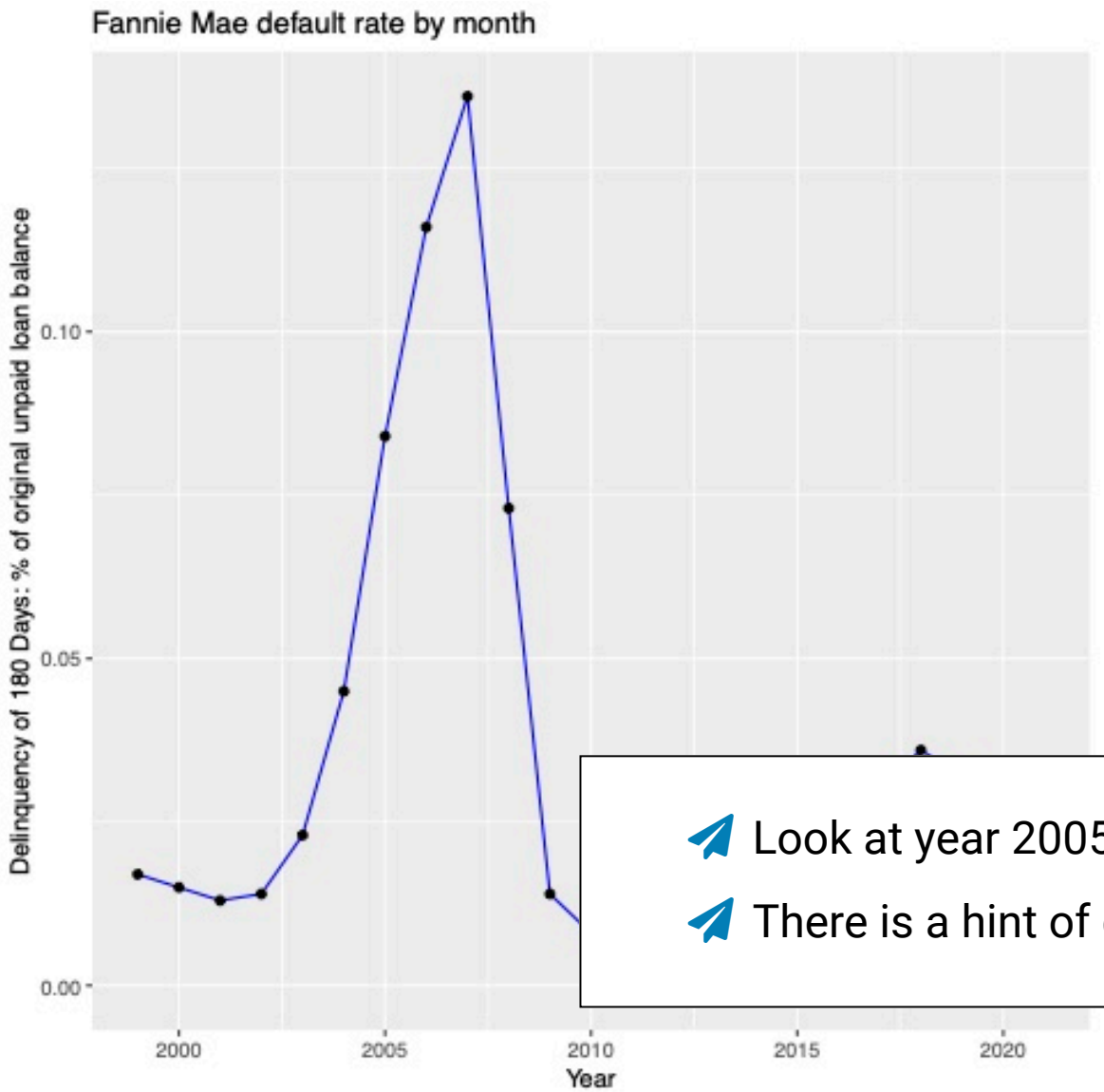
Borrower Characteristics: 👫

- FICO scores
- state
- age
- occupation

Loan/Property Information: 🏠

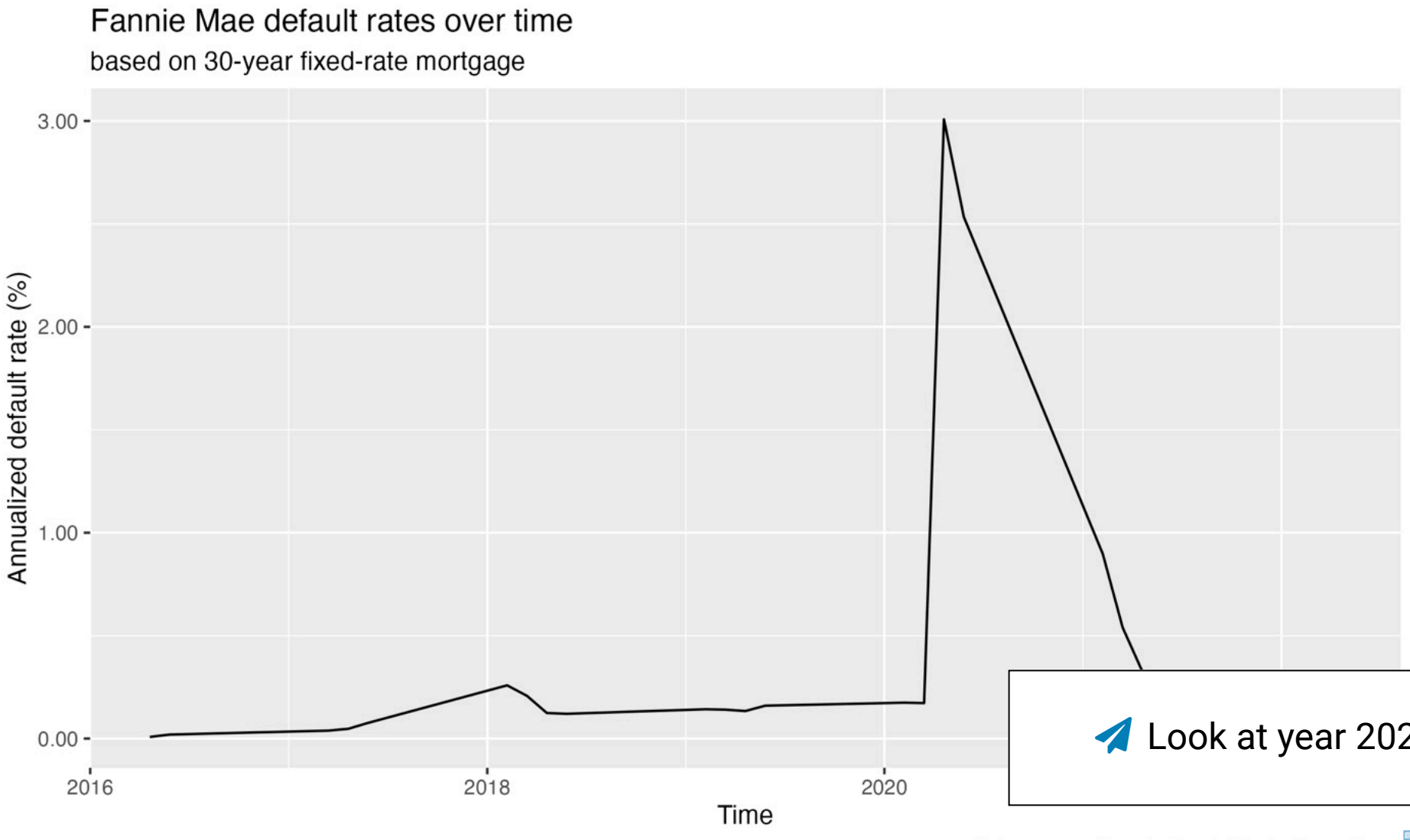
- occupancy status
- interest rate at origination
- term of loan
- loan to value ratio (LTV)
- debt to income ratio (DTI)
- mortgage insurance percentage (LMI)

What can you learn from the Fannie Mae data?



- Look at year 2005 onwards! 👀
- There is a hint of crisis before year 2008!

What can you learn from the Fannie Mae data (from our data set)?

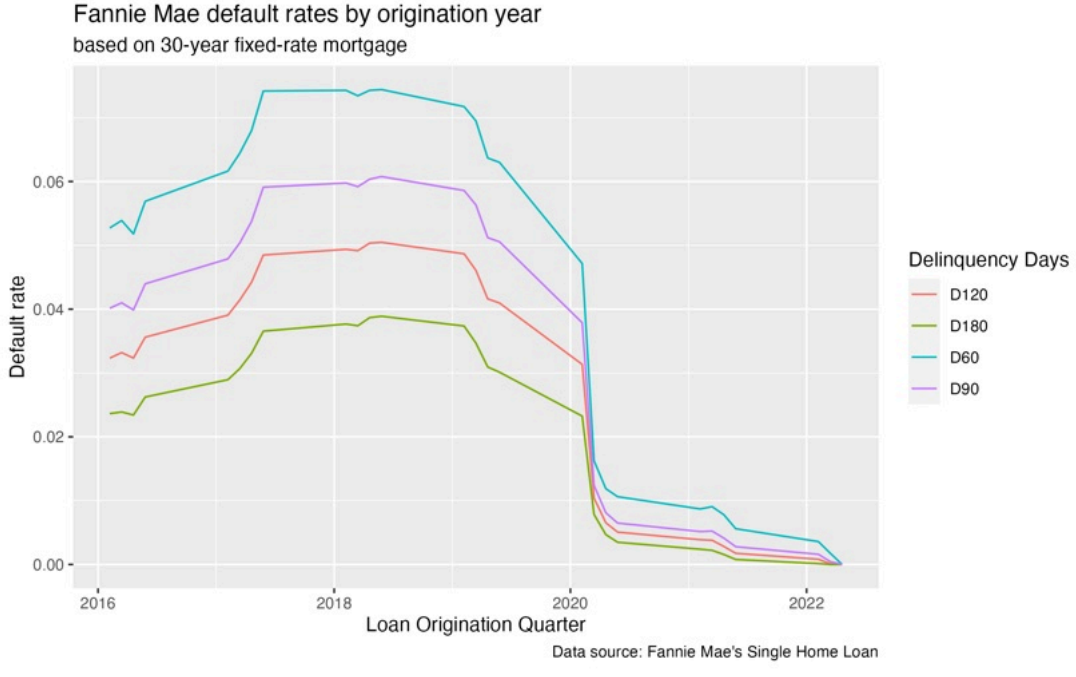


Data source: Fannie Mae's Single Home Loan

Delinquency (from 2016 onwards)

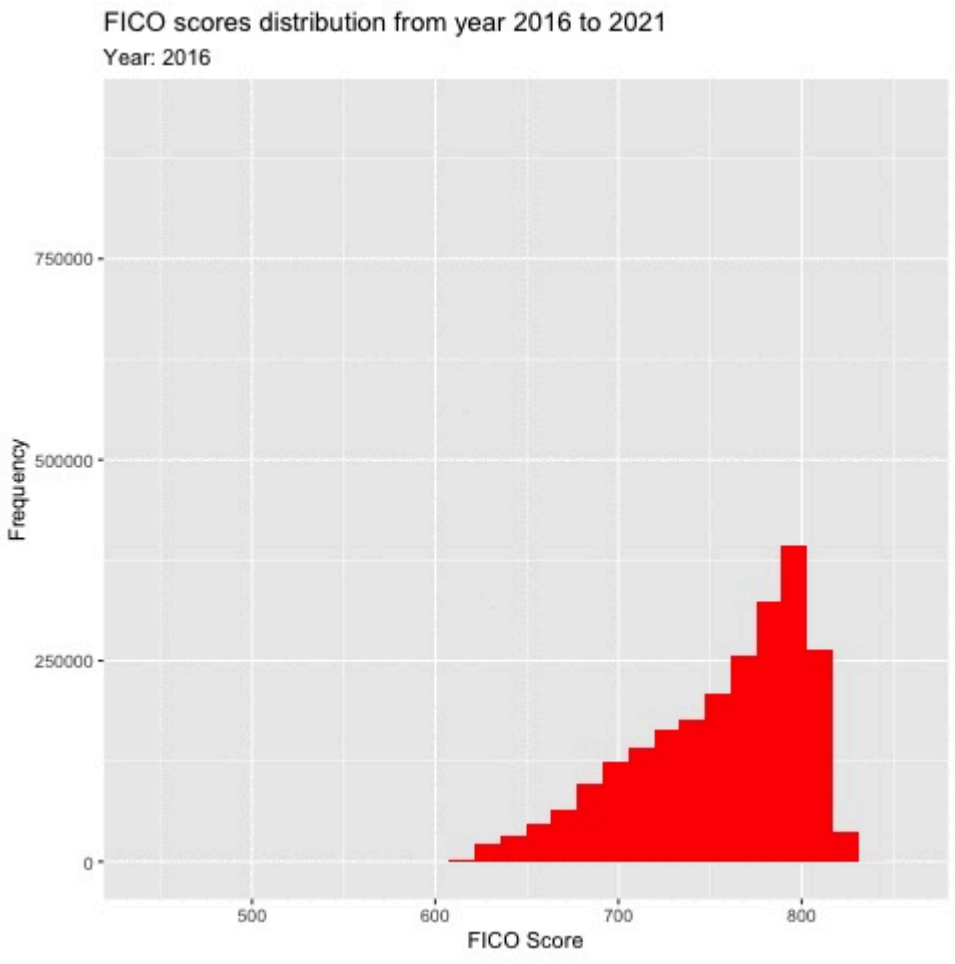
Field Position	Field short name	Field Name	Description	Date Bound Notes	Respective Disclosure Notes
40	DLQ_STATUS	Current Loan Delinquency Status	The number of months the obligor is delinquent as determined by the governing mortgage documents.	SF Loan Performance: Enhanced format with the October 2020 Release	For mortgage loans removed from the reference pool or historical data set, this field will be blank, subsequent to the month of removal. If the delinquency is unknown, the value 'XX' will display. In the event the loan is greater than or equal to 99 months delinquent, the field will report a '99'.

```
f60_table <- performancefile %>%
  filter(dlq_status >= 2 & dlq_status < 999, z_zb_code == '') %>
  group_by(LOAN_ID) %>%
  summarize(F60_DTE = min(period)) %>%
  left_join(performancefile,  by = c("LOAN_ID" = "LOAN_ID", "F60
  select(LOAN_ID, F60_DTE, act_upb) %>%
  rename(F60_UPB = act_upb)
```



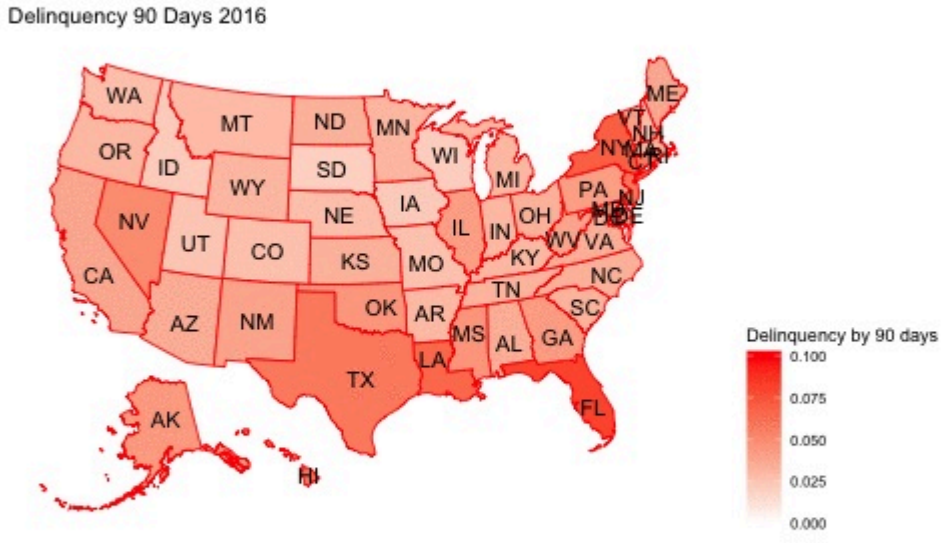
Delinquency by Fico Scores

```
ggplot(alldata %>% filter(orig_yr >= 2016),
       aes(x = CSCORE_B,
           group = orig_yr,
           )) +
  geom_histogram(fill = "red") +
  ggtitle("FICO scores distribution from year 2016 to 2021") +
  labs(subtitle = ("Year: {closest_state}"),
       ylab = "Frequency",
       xlab = "FICO Score") +
  transition_states(orig_yr,
                    transition_length = 6,
                    state_length = 1) -> plot1
anim_save("Ficohistogram.gif", plot1)
```

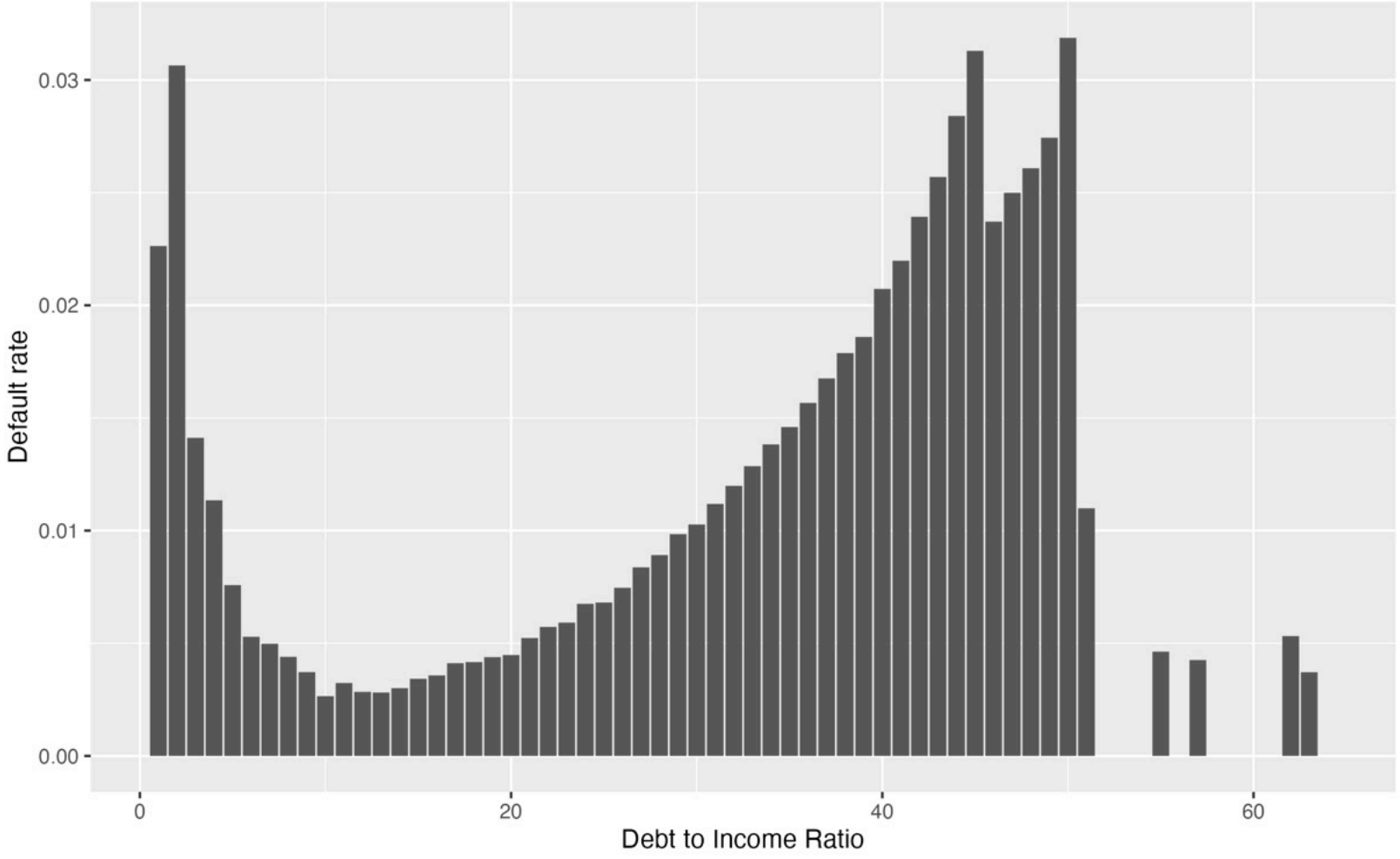


Default rate by States

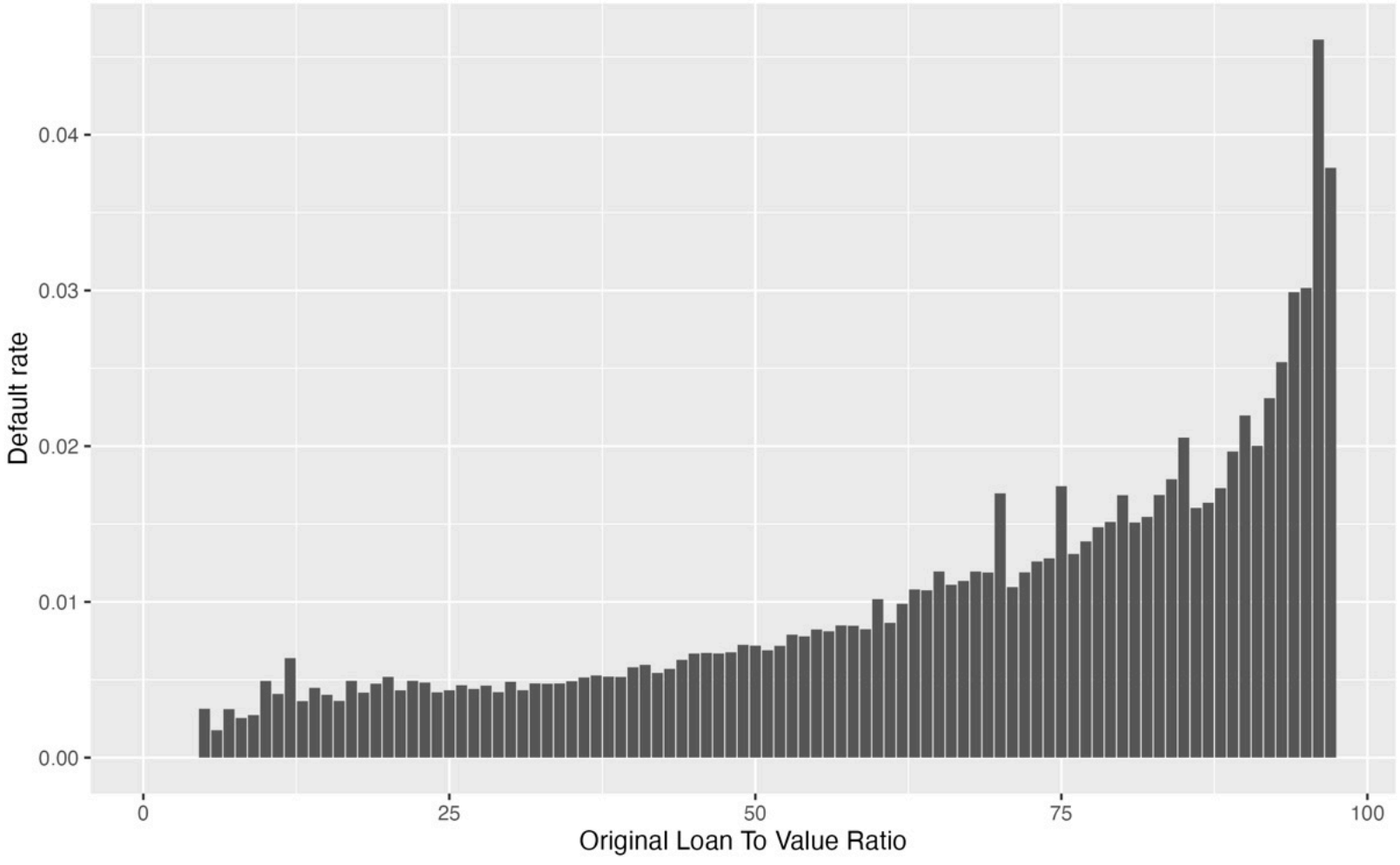
```
#library(gganimate)
datamap <- alldata %>%
  filter(orig_yq >= 2016.1) %>%
  group_by(state, orig_yr) %>%
  summarize(D90 = sum(!is.na(F90_DTE))/length(F90_DTE))

plot_usmap(
  data = datamap,
  labels = TRUE,
  values = "D90",
  color = "red"
) +
  scale_fill_continuous(name = "Delinquency by 90 days", low = "
  theme(legend.position = "right") -> mapPlot

transitionMap <- mapPlot +
  labs(title = "Delinquency 90 Days {as.integer(frame_time)}") +
  transition_time(orig_yr)

anim <- animate(transitionMap, fps = 10)
anim_save("map.gif", anim)
```



Are borrowers with higher debt-to-income ratios more risky?



Are those who borrow more as a proportion of their house values more risky?



Summary

What we learnt today:

- Introduced to credit risk data
- This was a big data case study
- Saw an example of how to wrangle different .zip files in R
- Learnt important lessons about memory constraints

Slides updated and maintained by Dr. Kate Saunders. Previous maintainer was Dr. Joan Tan

Lecturer: *Kate Saunders*

Department of Econometrics and Business Statistics

ETC5512.Clayton-x@monash.edu

Week 7