Introduction and data structure

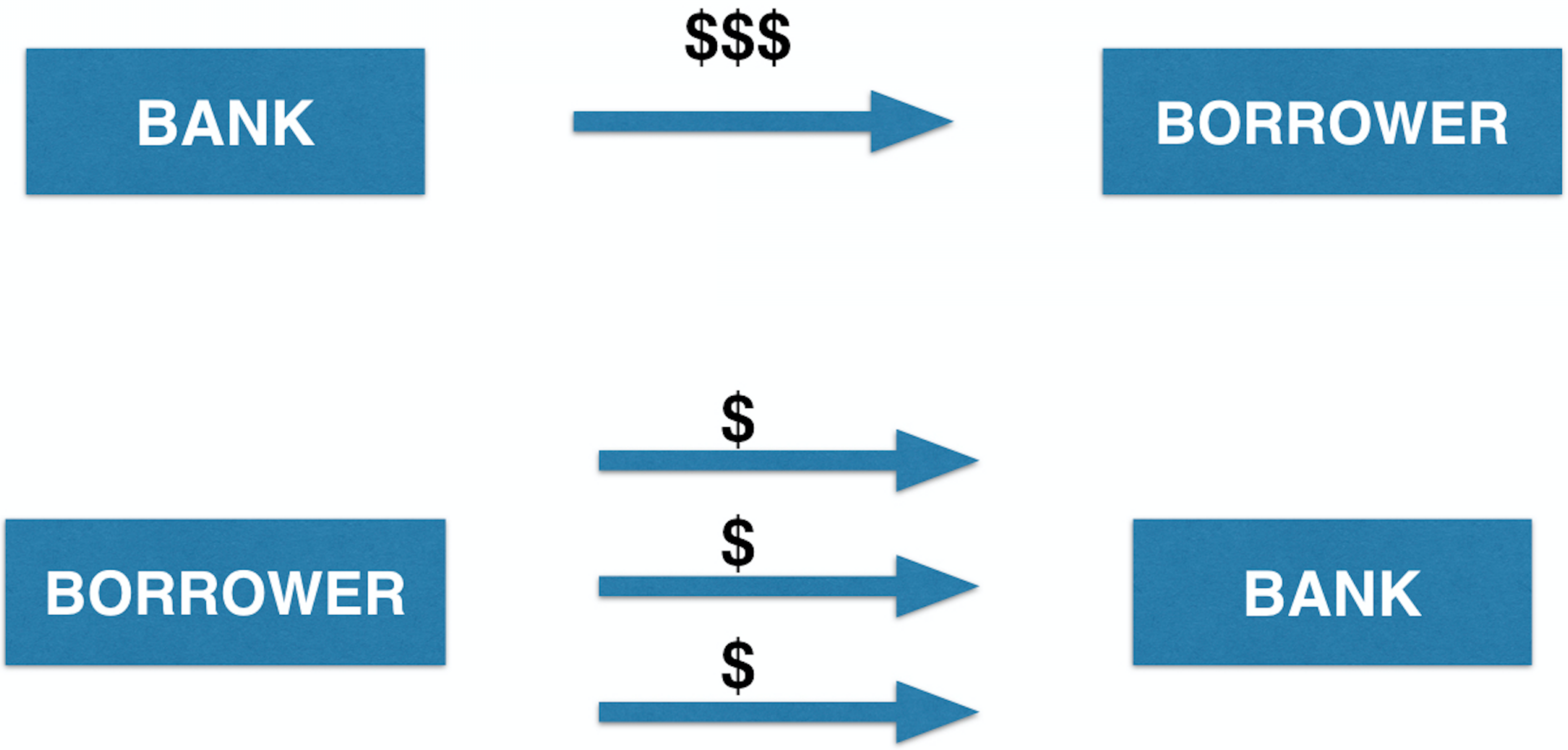
CREDIT RISK MODELING IN R

Lore Dirick

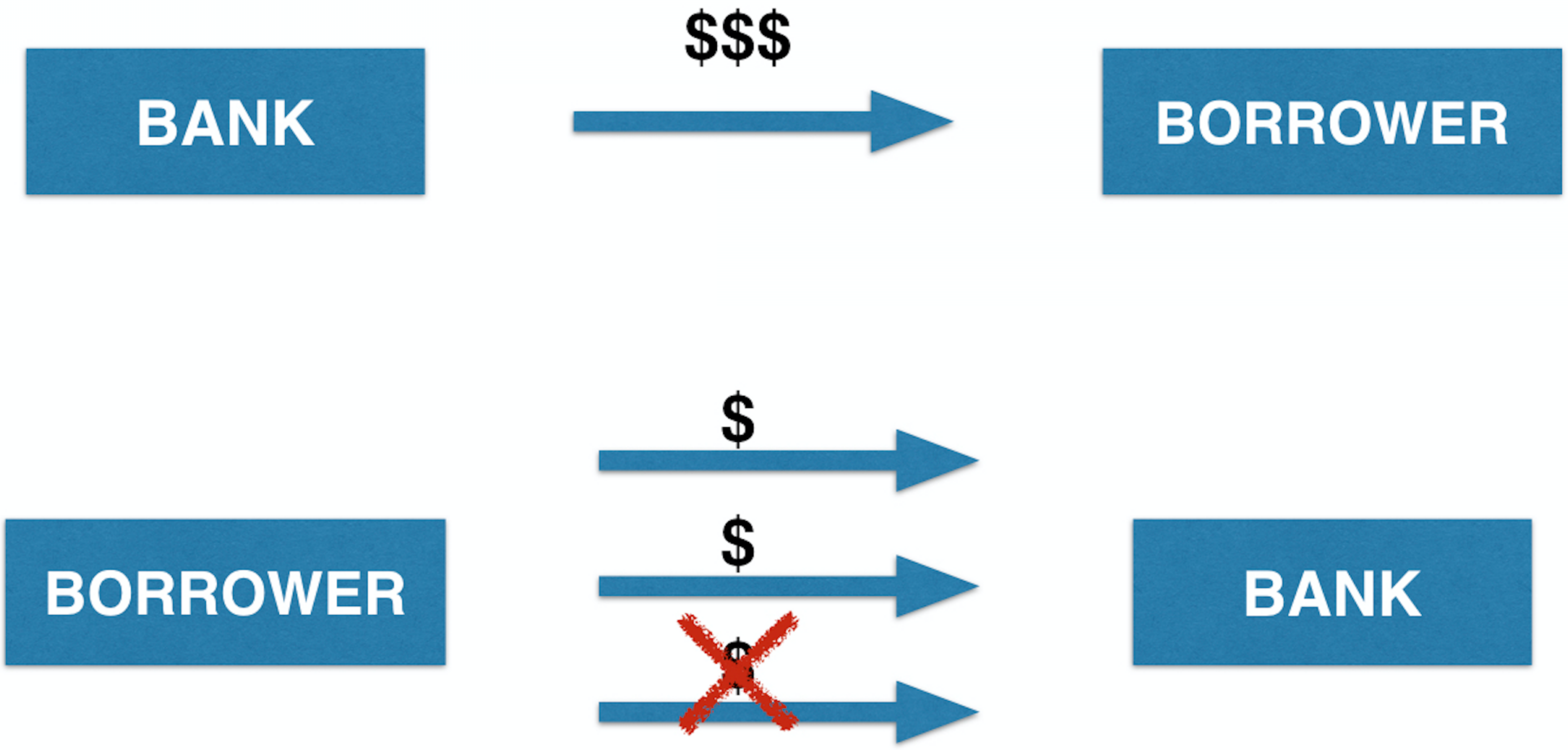
Manager of Data Science Curriculum at Flatiron School

What is loan default?

What is loan default?



What is loan default?



Components of expected loss (EL)

- Probability of default (`PD`)
- Exposure at default (`EAD`)
- Loss given default (`LGD`)

$$\mathrm{EL} = \mathrm{PD} \times \mathrm{EAD} \times \mathrm{LGD}$$

Components of expected loss (EL)

- **Probability of default (`PD`)**
- Exposure at default (`EAD`)
- Loss given default (`LGD`)

$$\mathrm{EL} = \mathrm{PD} \times \mathrm{EAD} \times \mathrm{LGD}$$

Information used by banks

- Application information:
  - Income
  - Marital status
  - ...
- Behavioral information
  - Current account balance
  - Payment arrears in account history
  - ...

```r
head(loan_data, 10)
```

```
   loan_status loan_amnt int_rate grade emp_length home_ownership annual_inc age
1            0      5000    10.65     B         10           RENT      24000  33
2            0      2400       NA     C         25           RENT      12252  31
3            0     10000    13.49     C         13           RENT      49200  24
4            0      5000       NA     A          3           RENT      36000  39
5            0      3000       NA     E          9           RENT      48000  24
6            0     12000    12.69     B         11            OWN      75000  28
7            1      9000    13.49     C          0           RENT      30000  22
8            0      3000     9.91     B          3           RENT      15000  22
9            1     10000    10.65     B          3           RENT     100000  28
10           0      1000    16.29     D          0           RENT      28000  22
```

```r
library(gmodels)
CrossTable(loan_data$home_ownership)
```

```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  29092


          |  MORTGAGE |     OTHER |       OWN |      RENT |
          |-----------|-----------|-----------|-----------|
          |     12002 |        97 |      2301 |     14692 |
          |     0.413 |     0.003 |     0.079 |     0.505 |
          |-----------|-----------|-----------|-----------|
```

```r
CrossTable(loan_data$home_ownership, loan_data$loan_status, prop.r = TRUE,
           prop.c = FALSE, prop.t = FALSE, prop.chisq = FALSE)
```

```
                         | loan_data$loan_status
loan_data$home_ownership |         0 |         1 | Row Total |
 ------------------------|-----------|-----------|-----------|
                MORTGAGE |     10821 |      1181 |     12002 |
                         |     0.902 |     0.098 |     0.413 |
 ------------------------|-----------|-----------|-----------|
                   OTHER |        80 |        17 |        97 |
                         |     0.825 |     0.175 |     0.003 |
 ------------------------|-----------|-----------|-----------|
                     OWN |      2049 |       252 |      2301 |
                         |     0.890 |     0.110 |     0.079 |
 ------------------------|-----------|-----------|-----------|
                    RENT |     12915 |      1777 |     14692 |
                         |     0.879 |     0.121 |     0.505 |
 ------------------------|-----------|-----------|-----------|
            Column Total |     25865 |      3227 |     29092 |
```

Let's practice!

CREDIT RISK MODELING IN R

Histograms and outliers

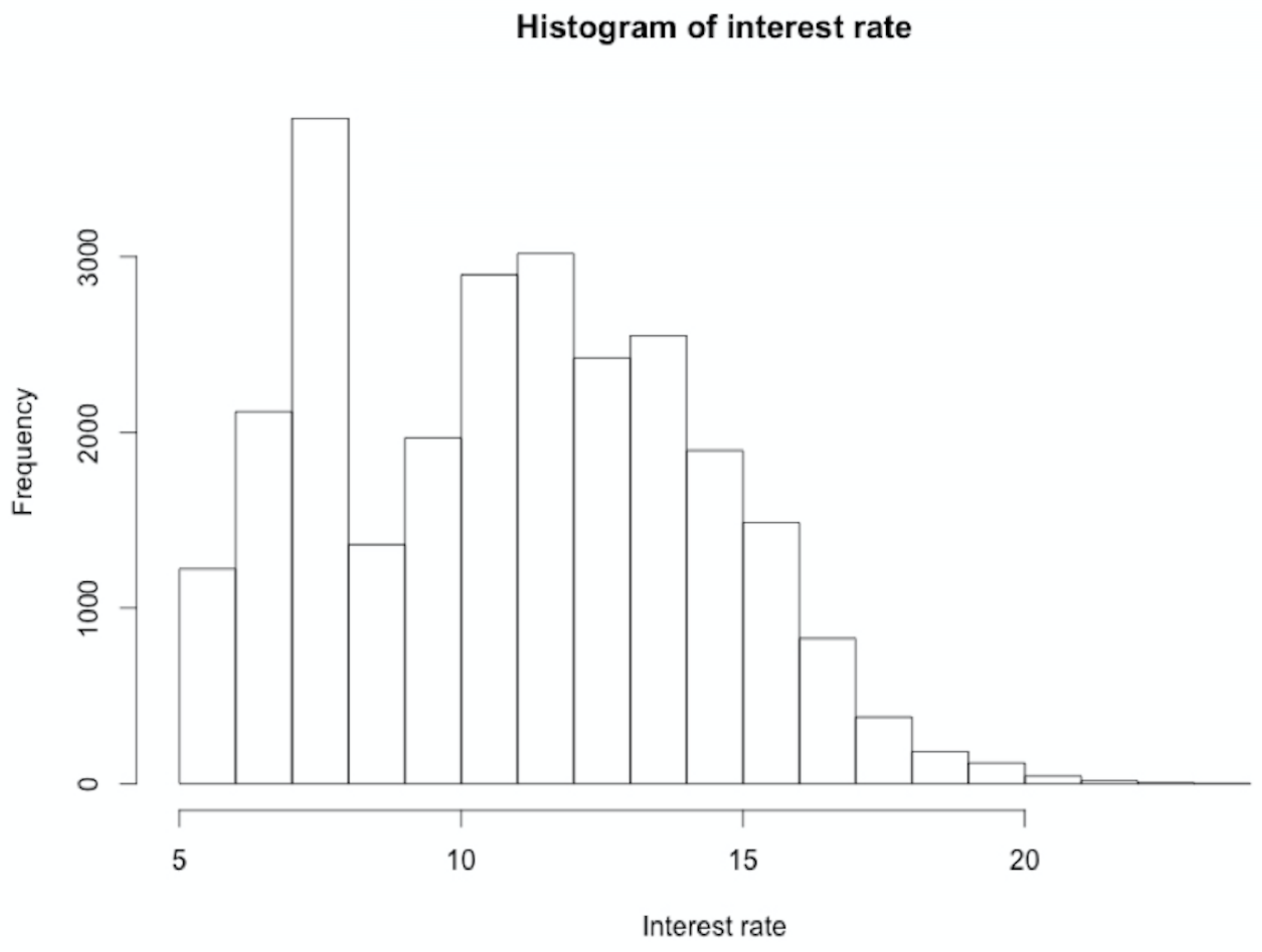
CREDIT RISK MODELING IN R

Lore Dirick

Manager of Data Science Curriculum at Flatiron School

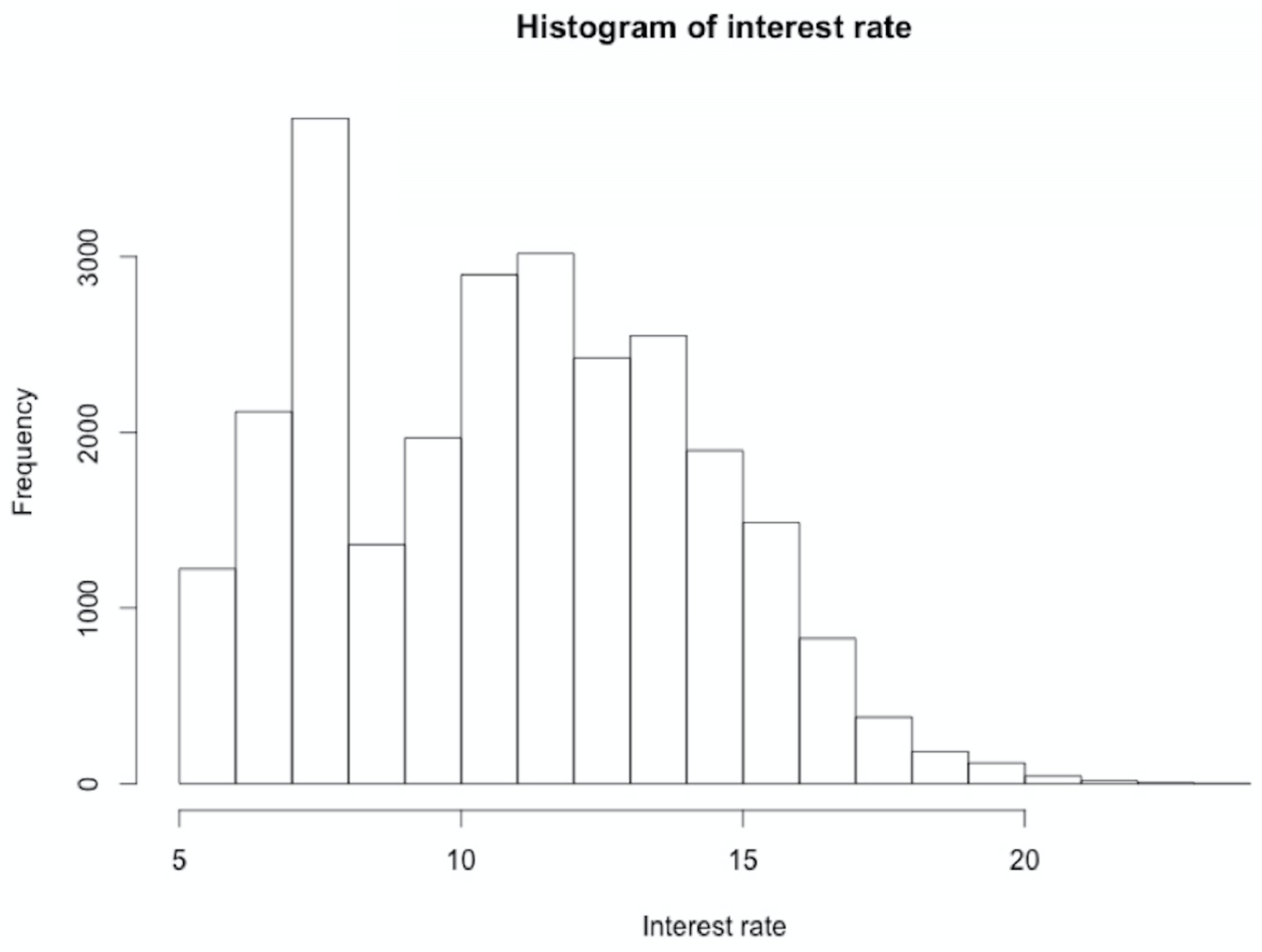
Using function hist()

```
hist(loan_data$int_rate)
```



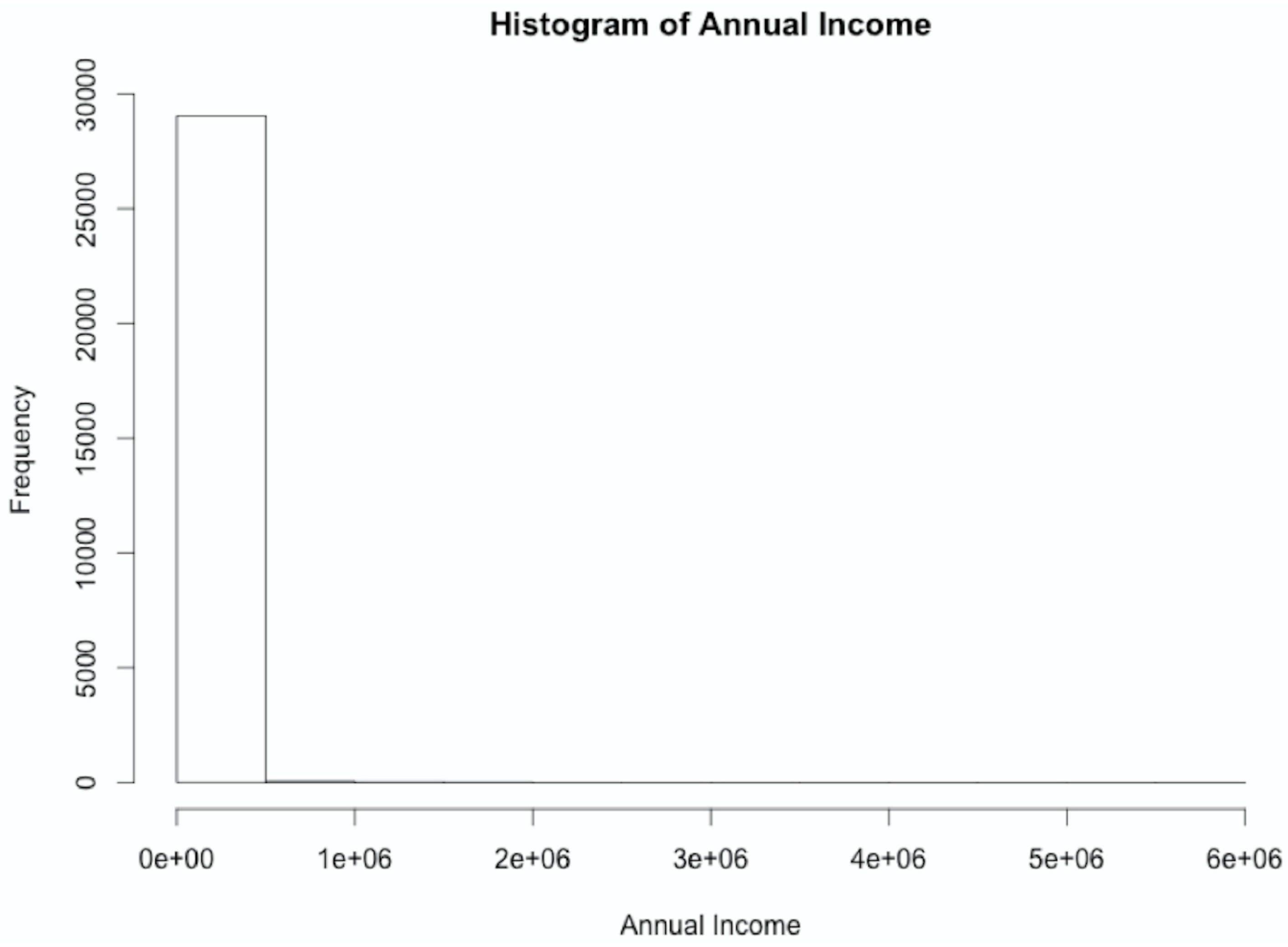
Using function hist()

```
hist(loan_data$int_rate, main = "Histogram of interest rate", xlab = "Interest rate")
```



Using function hist() on annual_inc

```r
hist(loan_data$annual_inc, xlab = "Annual Income", main = "Histogram of Annual Income")
```



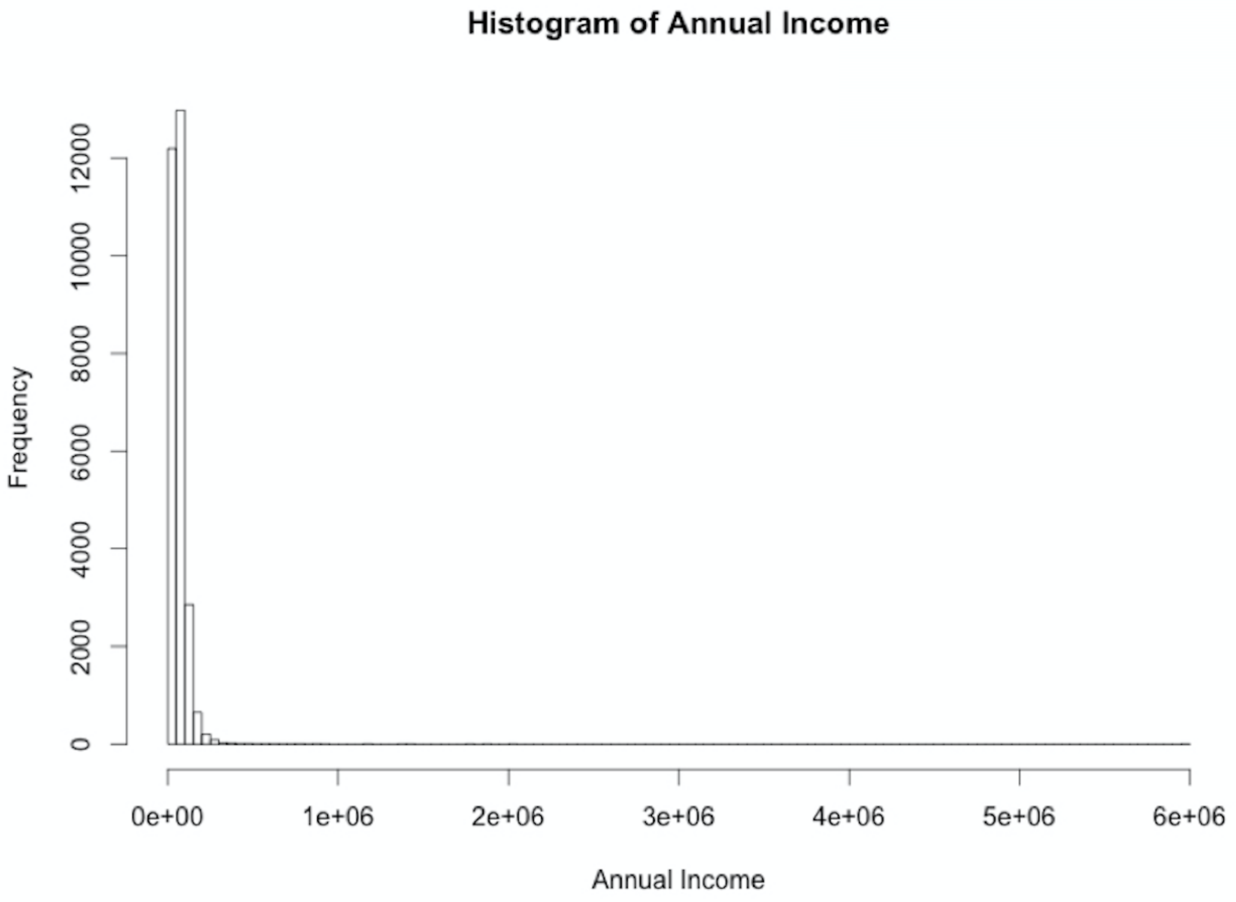
Using function hist() on annual_inc

```
hist_income <- hist(loan_data$annual_inc,
                    xlab = "Annual Income",
                    main = "Histogram of Annual Income")
hist_income$breaks
```

```
0  500000 1000000 1500000 2000000 2500000 3000000 3500000 4000000 4500000 ...
```

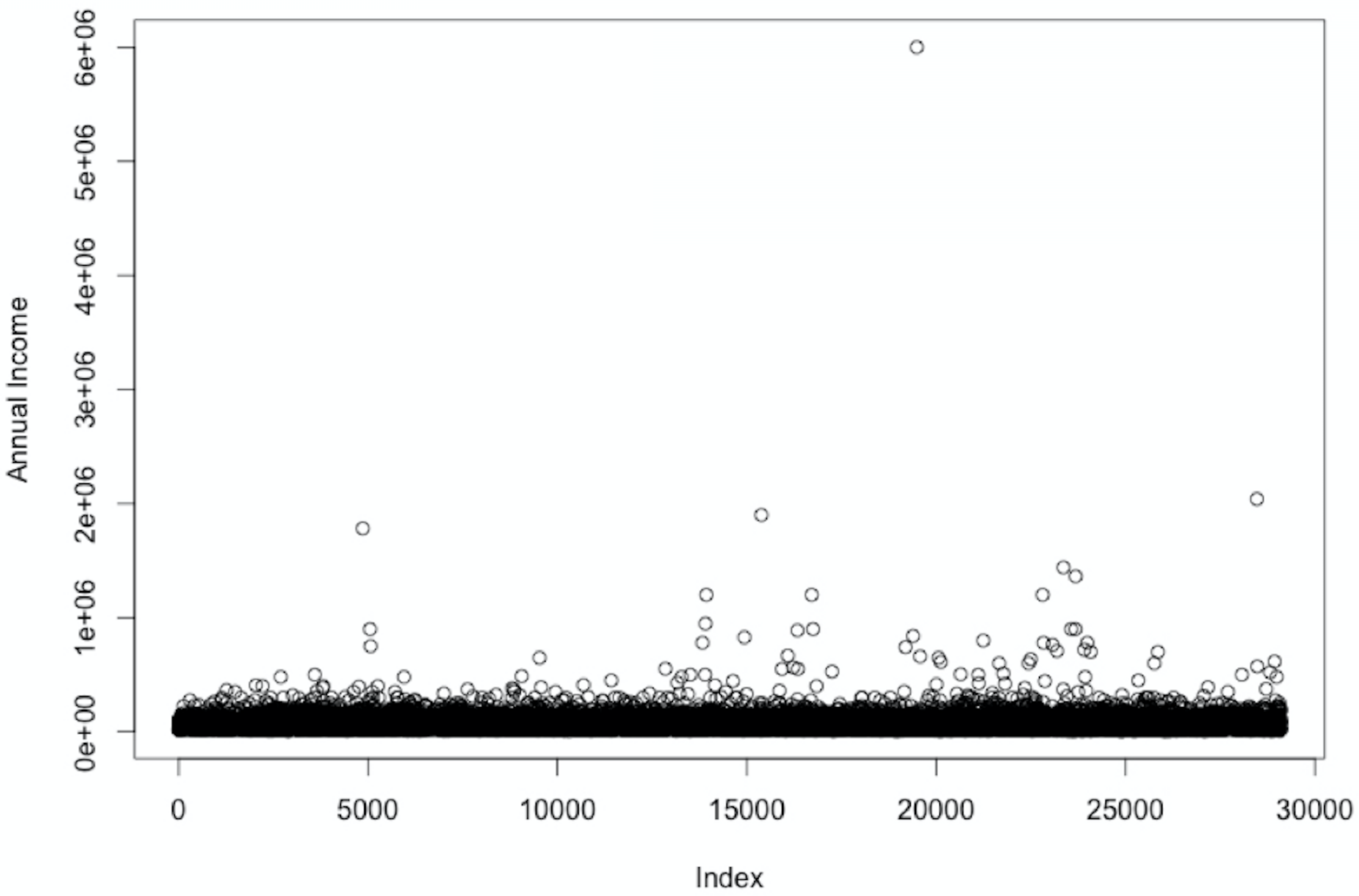
The breaks-argument

```
n_breaks <- sqrt(nrow(loan_data)) # n_breaks = 170.5638
hist_income_n <- hist(loan_data$annual_inc, breaks = n_breaks,
                      xlab = "Annual Income", main = "Histogram of Annual Income")
```



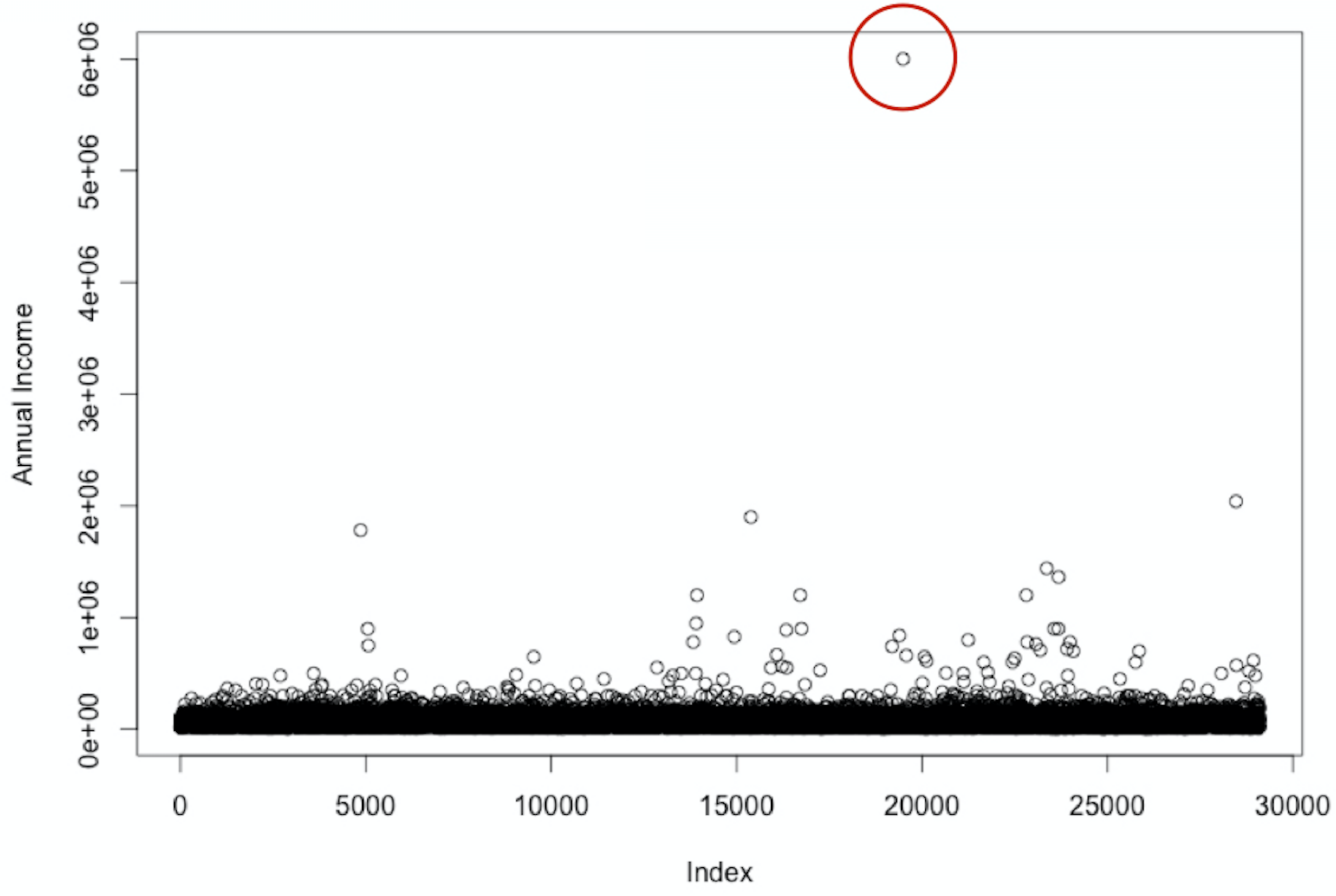
annual_inc

```
plot(loan_data$annual_inc, ylab = "Annual Income")
```



annual_inc

```
plot(loan_data$annual_inc, ylab = "Annual Income")
```



Outliers

- When is a value an outlier?
  - Expert judgment
  - Rule of thumb, e.g.,
    - Q1 - 1.5 * IQR
    - Q3 + 1.5 * IQR
  - Mostly: combination of both

Expert judgment

"Annual salaries > $3 million are outliers"

```
# Find outlier
index_outlier_expert <- which(loan_data$annual_inc > 3000000)

# Remove outlier from dataset
loan_data_expert <- loan_data[-index_outlier_expert, ]
```

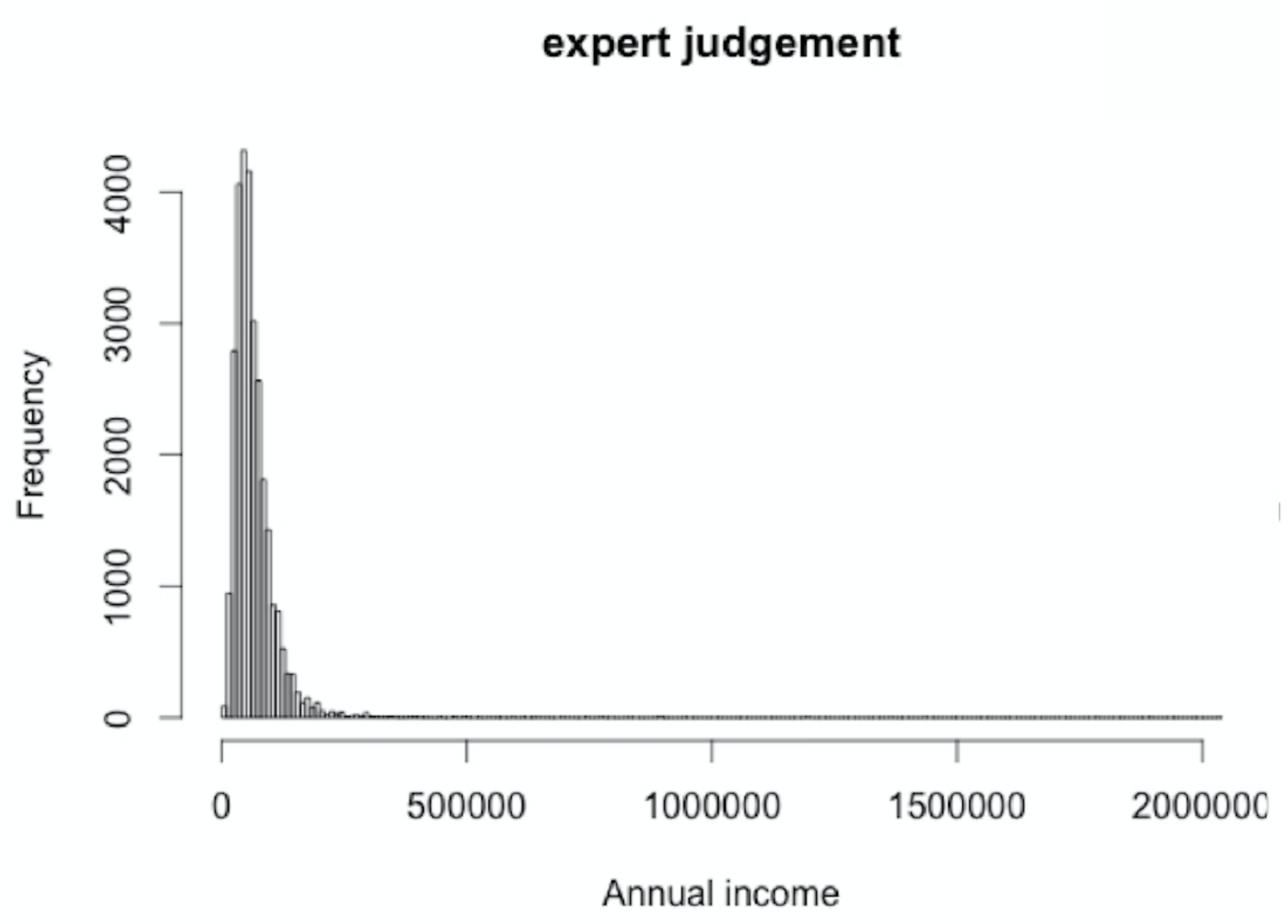
Rule of thumb

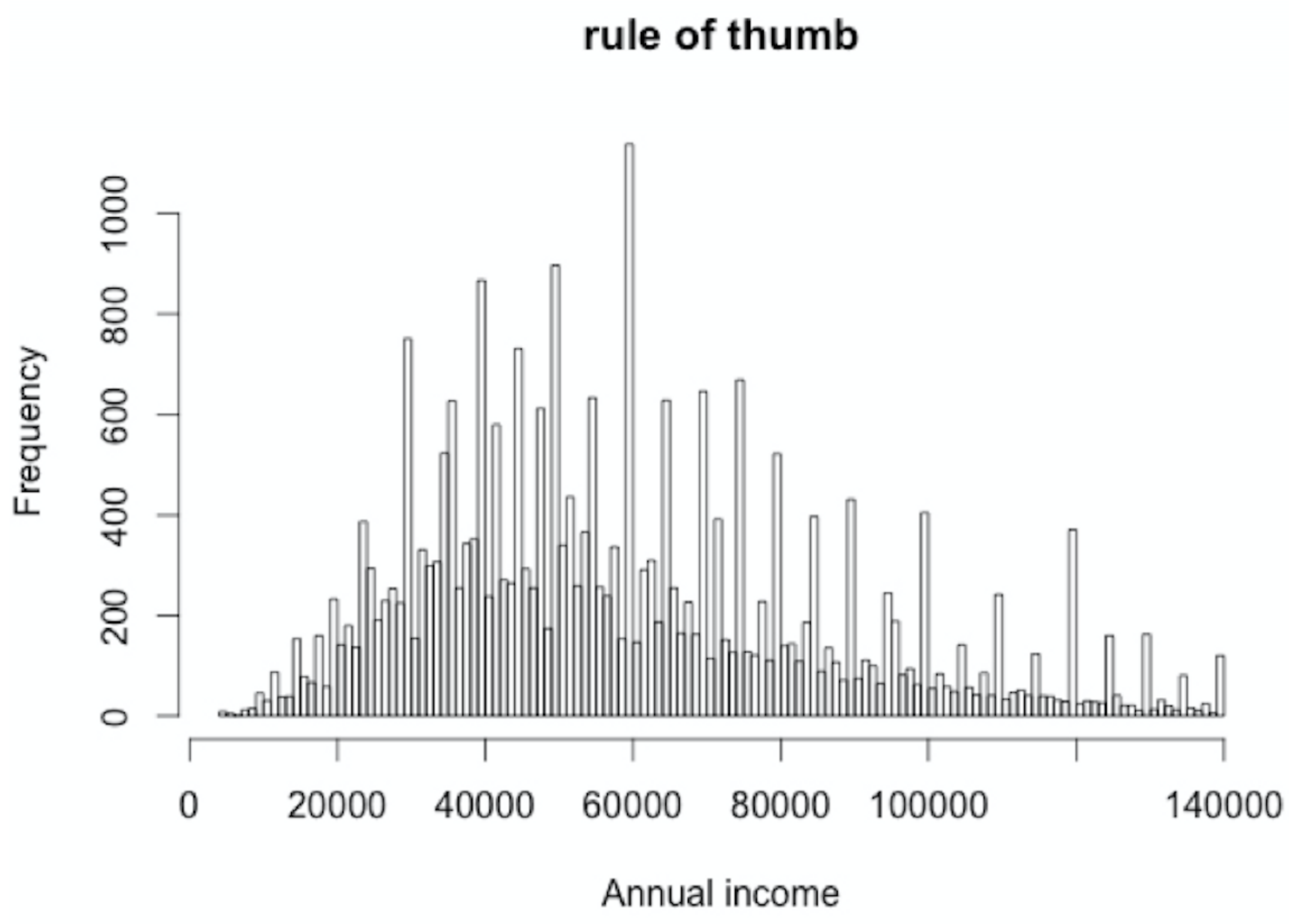
Outlier if bigger than Q3 + 1.5 * IQR

```
# Calculate Q3 + 1.5 * IQR
outlier_cutoff <- quantile(loan_data$annual_inc, 0.75) + 1.5 * IQR(loan_data$annual_inc)
# Identify outliers
index_outlier_ROT <- which(loan_data$annual_inc > outlier_cutoff)
# Remove outliers
loan_data_ROT <- loan_data[-index_outlier_ROT, ]
```

```r
hist(loan_data_expert$annual_inc,
     sqrt(nrow(loan_data_expert)),
     xlab = "Annual income")
```

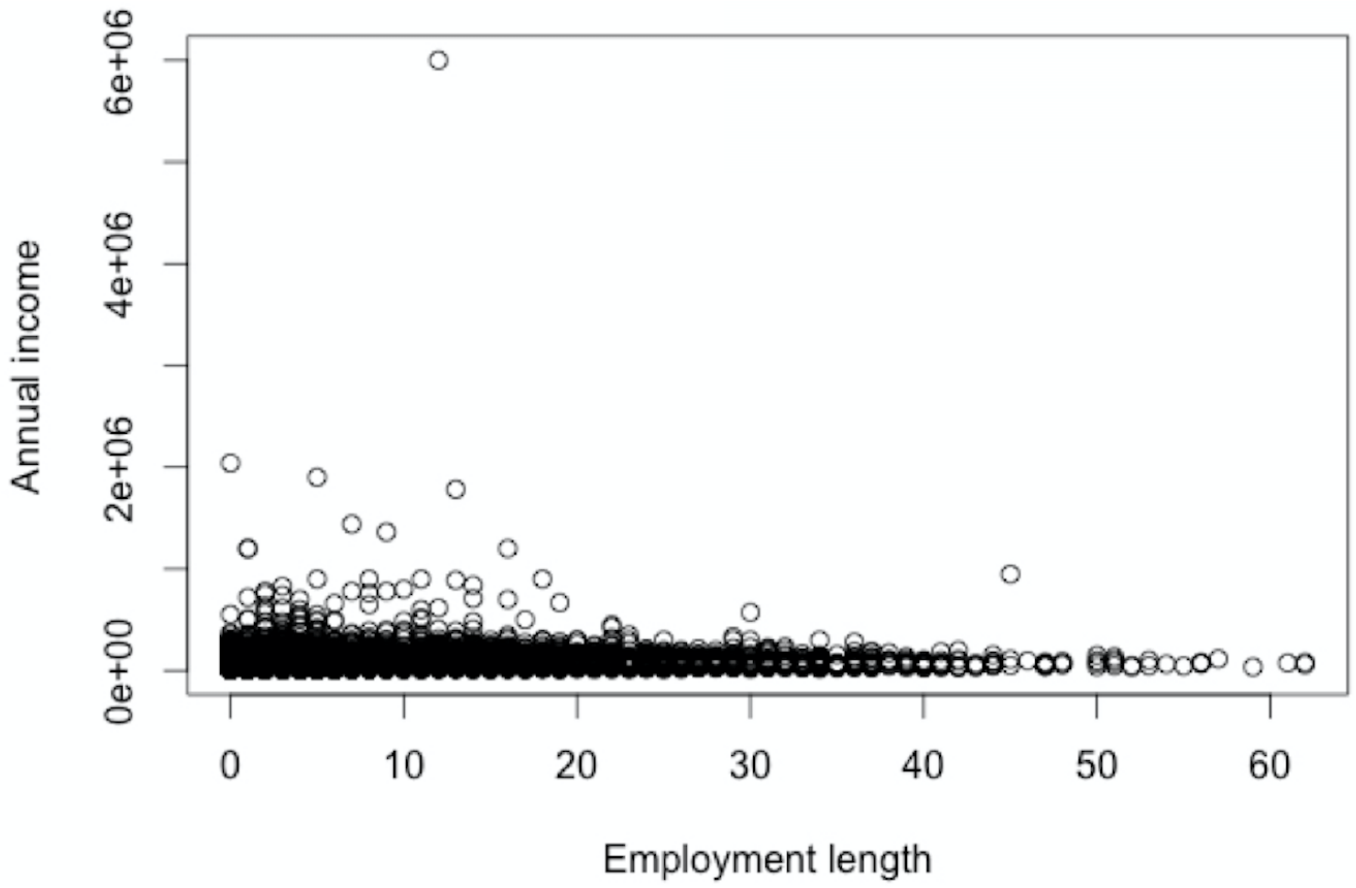
```r
hist(loan_data_ROT$annual_inc,
     sqrt(nrow(loan_data_ROT)),
     xlab = "Annual income")
```





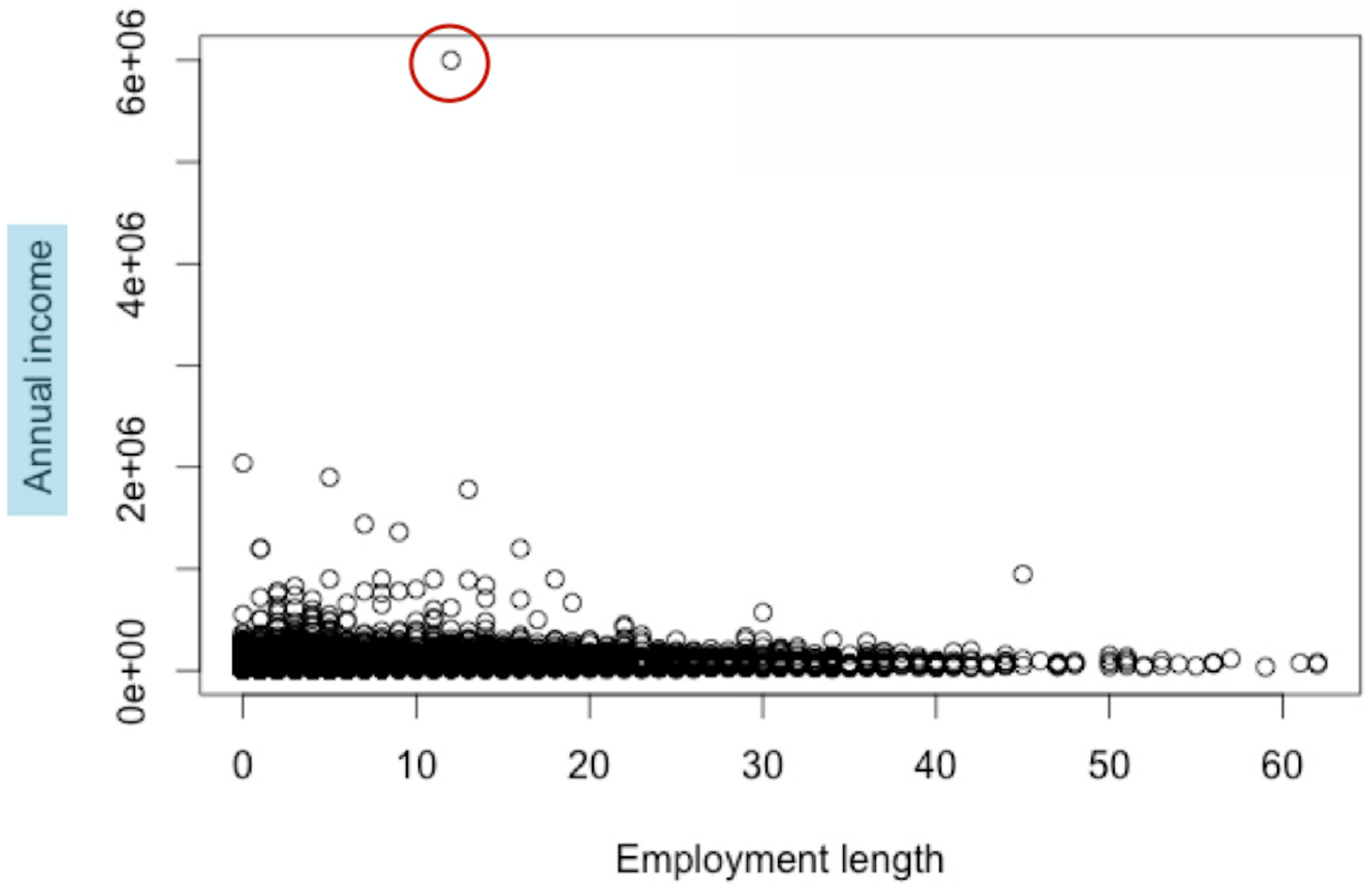
Bivariate plot

```
plot(loan_data$emp_length, loan_data$annual_inc,
     xlab= "Employment length", ylab= "Annual income")
```



Bivariate plot

```
plot(loan_data$emp_length, loan_data$annual_inc,
     xlab= "Employment length", ylab= "Annual income")
```



Let's practice!

CREDIT RISK MODELING IN R

Missing data and coarse classification

CREDIT RISK MODELING IN R

Lore Dirick
Manager of Data Science Curriculum at Flatiron School

Outlier deleted

```
loan_status  loan_amnt  int_rate  grade   emp_length  home_ownership   annual_inc    age
     0         5000      12.73      C         12         MORTGAGE        6000000     144
```

Missing inputs

	loan_status	loan_amnt	int_rate	grade	emp_length	home_ownership	annual_inc	age
...	...	...	...	...	...	...	...	...
125	0	6000	14.27	C	14	MORTGAGE	94800	23
126	1	2500	7.51	A	NA	OWN	12000	21
127	0	13500	9.91	B	2	MORTGAGE	36000	30
128	0	25000	12.42	B	2	RENT	225000	30
129	0	10000	NA	C	2	RENT	45900	65
130	0	2500	13.49	C	4	RENT	27200	26
...	...	...	...	...	...	...	...	...
2108	0	8000	7.90	A	8	RENT	64000	24
2109	0	12000	8.90	A	0	RENT	38400	26
2110	0	4000	NA	A	7	RENT	48000	30
2111	0	7000	9.91	B	20	MORTGAGE	130000	30
2112	0	7600	6.03	A	41	MORTGAGE	70920	28

Missing inputs

```
summary(loan_data$emp_length)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   2.000   4.000   6.145   8.000  62.000     809
```

Missing inputs: strategies

- Delete row/column
- Replace
- Keep

Delete rows

```
index_NA <- which(is.na(loan_data$emp_length)
loan_data_no_NA <- loan_data[-c(index_NA), ]
```

```
loan_status   loan_amnt   int_rate grade emp_length  home_ownership  annual_inc  age
...     ...         ...        ...   ...        ...             ...         ...  ...
125       0        6000      14.27     C         14        MORTGAGE       94800   23
126       1        2500       7.51     A         NA             OWN       12000   21
127       0       13500       9.91     B          2        MORTGAGE       36000   30
128       0       25000      12.42     B          2            RENT      225000   30
129       0       10000         NA     C          2            RENT       45900   65
130       0        2500      13.49     C          4            RENT       27200   26
...     ...         ...        ...   ...        ...             ...         ...  ...
2112      0        7600       6.03     A         41        MORTGAGE       70920   28
2113      0       10000      11.71     B          5            RENT       48132   22
2114      0        8000       6.62     A         17             OWN       42000   24
2115      0        4475         NA     B         NA             OWN       15000   23
2116      0        5750       8.90     A          3            RENT       17000   21
...     ...         ...        ...   ...        ...             ...         ...  ...
```

Delete column

```
loan_data_delete_employ <- loan_data
loan_data_delete_employ$emp_length <- NULL
```

```
loan_status   loan_amnt   int_rate grade   home_ownership  annual_inc  age
...     ...         ...        ...   ...              ...         ...  ...
125       0        6000      14.27     C         MORTGAGE       94800   23
126       1        2500       7.51     A              OWN       12000   21
127       0       13500       9.91     B         MORTGAGE       36000   30
128       0       25000      12.42     B             RENT       225000  30
129       0       10000         NA     C             RENT       45900   65
130       0        2500      13.49     C             RENT       27200   26
...     ...         ...        ...   ...              ...         ...  ...
2112      0        7600       6.03     A         MORTGAGE       70920   28
2113      0       10000      11.71     B             RENT       48132   22
2114      0        8000       6.62     A              OWN       42000   24
2115      0        4475         NA     B              OWN       15000   23
2116      0        5750       8.90     A             RENT       17000   21
...     ...         ...        ...   ...              ...         ...  ...
```

Replace: median imputation

```
index_NA <- which(is.na(loan_data$emp_length))
loan_data_replace <- loan_data
loan_data_replace$emp_length[index_NA] <- median(loan_data$emp_length, na.rm = TRUE)
```

```
loan_status loan_amnt  int_rate  grade  emp_length home_ownership annual_inc age
...     ...         ...        ...    ...        ...            ...          ...  ...
125       0        6000      14.27      C         14       MORTGAGE        94800   23
126       1        2500       7.51      A         NA            OWN        12000   21
127       0       13500       9.91      B          2       MORTGAGE        36000   30
128       0       25000      12.42      B          2           RENT       225000   30
129       0       10000         NA      C          2           RENT        45900   65
130       0        2500      13.49      C          4           RENT        27200   26
...     ...         ...        ...    ...        ...            ...          ...  ...
2112      0        7600       6.03      A         41       MORTGAGE        70920   28
2113      0       10000      11.71      B          5           RENT        48132   22
2114      0        8000       6.62      A         17            OWN        42000   24
2115      0        4475         NA      B         NA            OWN        15000   23
2116      0        5750       8.90      A          3           RENT        17000   21
...     ...         ...        ...    ...        ...            ...          ...  ...
```

Replace: median imputation

```
index_NA <- which(is.na(loan_data$emp_length))
loan_data_replace <- loan_data
loan_data_replace$emp_length[index_NA] <- median(loan_data$emp_length, na.rm = TRUE)
```

```
loan_status loan_amnt  int_rate  grade  emp_length home_ownership annual_inc age
...     ...         ...        ...   ...        ...            ...          ...  ...
125       0        6000      14.27     C         14       MORTGAGE        94800   23
126       1        2500       7.51     A          4            OWN        12000   21
127       0       13500       9.91     B          2       MORTGAGE        36000   30
128       0       25000      12.42     B          2           RENT       225000   30
129       0       10000         NA     C          2           RENT        45900   65
130       0        2500      13.49     C          4           RENT        27200   26
...     ...         ...        ...   ...        ...            ...          ...  ...
2112      0        7600       6.03     A         41       MORTGAGE        70920   28
2113      0       10000      11.71     B          5           RENT        48132   22
2114      0        8000       6.62     A         17            OWN        42000   24
2115      0        4475         NA     B          4            OWN        15000   23
2116      0        5750       8.90     A          3           RENT        17000   21
...     ...         ...        ...   ...        ...            ...          ...  ...
```

Keep

- Keep `NA`
- *Problem*: will cause row deletions for many models
- *Solution*: coarse classification, put variable in "bins"
  - New variable `emp_cat`
  - Range: 0-62 years → make bins of +/- 15 years
  - Categories: "0-15", "15-30", "30-45", "45+", "missing"

Keep: coarse classification

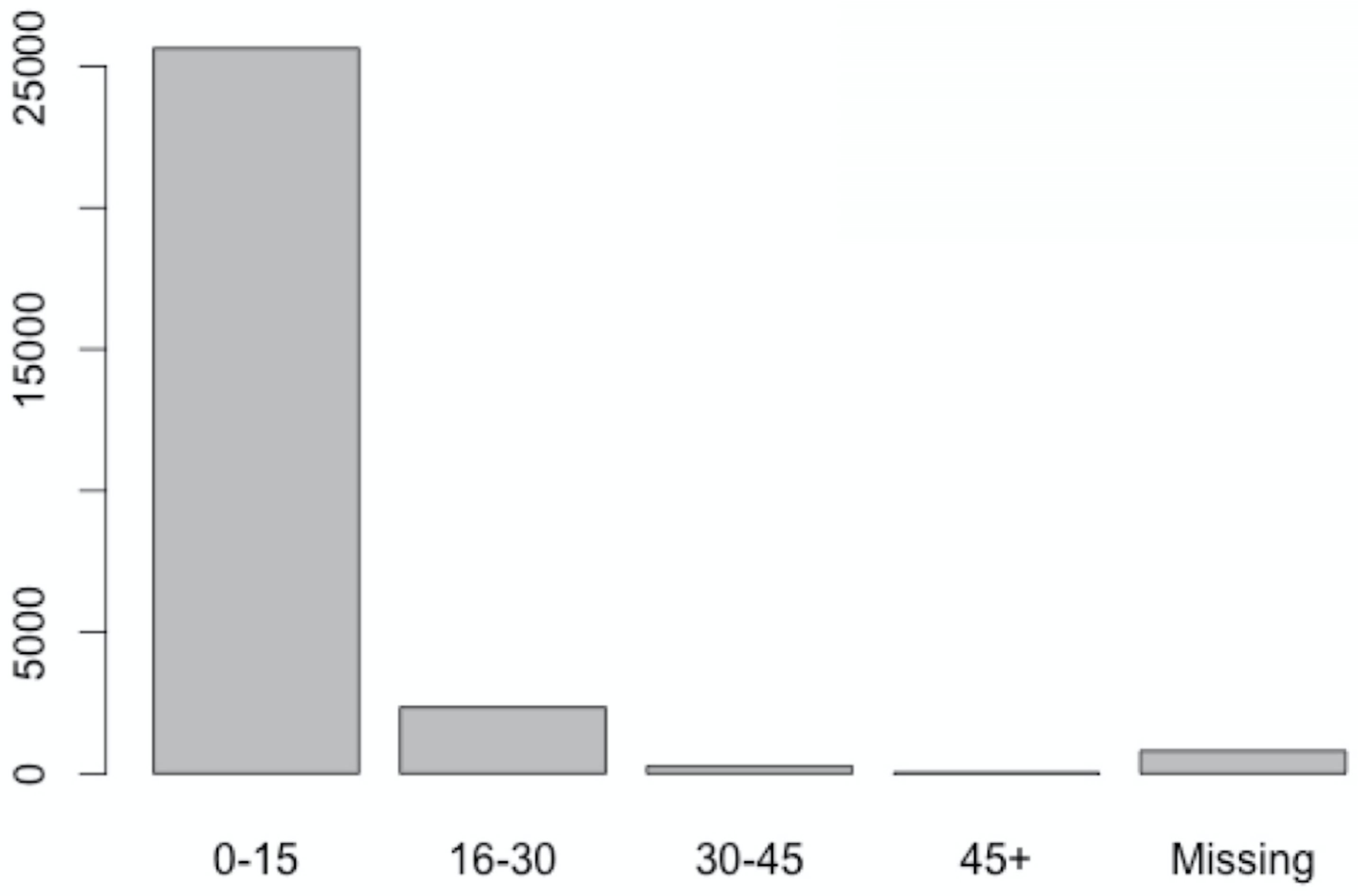
```
loan_status  loan_amnt  int_rate  grade  emp_length  home_ownership annual_inc age
...     ...         ...        ...   ...         ...           ...          ...  ...
125       0        6000      14.27     C         14        MORTGAGE        94800   23
126       1        2500       7.51     A         NA             OWN        12000   21
127       0       13500       9.91     B          2        MORTGAGE        36000   30
128       0       25000      12.42     B          2            RENT        225000  30
129       0       10000         NA     C          2            RENT        45900   65
130       0        2500      13.49     C          4            RENT        27200   26
...     ...         ...        ...   ...         ...           ...          ...  ...
2112      0        7600       6.03     A         41        MORTGAGE        70920   28
2113      0       10000      11.71     B          5            RENT        48132   22
2114      0        8000       6.62     A         17             OWN        42000   24
2115      0        4475         NA     B         NA             OWN        15000   23
2116      0        5750       8.90     A          3            RENT        17000   21
...     ...         ...        ...   ...         ...           ...          ...  ...
```

Keep: coarse classification

	loan_status	loan_amnt	int_rate	grade	emp_cat	home_ownership	annual_inc	age
...	...	...	...	...	...	...	...	...
125	0	6000	14.27	C	0-15	MORTGAGE	94800	23
126	1	2500	7.51	A	Missing	OWN	12000	21
127	0	13500	9.91	B	0-15	MORTGAGE	36000	30
128	0	25000	12.42	B	0-15	RENT	225000	30
129	0	10000	NA	C	0-15	RENT	45900	65
130	0	2500	13.49	C	0-15	RENT	27200	26
...	...	...	...	...	...	...	...	...
2112	0	7600	6.03	A	30-45	MORTGAGE	70920	28
2113	0	10000	11.71	B	0-15	RENT	48132	22
2114	0	8000	6.62	A	15-30	OWN	42000	24
2115	0	4475	NA	B	Missing	OWN	15000	23
2116	0	5750	8.90	A	0-15	RENT	17000	21
...	...	...	...	...	...	...	...	...

Bin frequencies

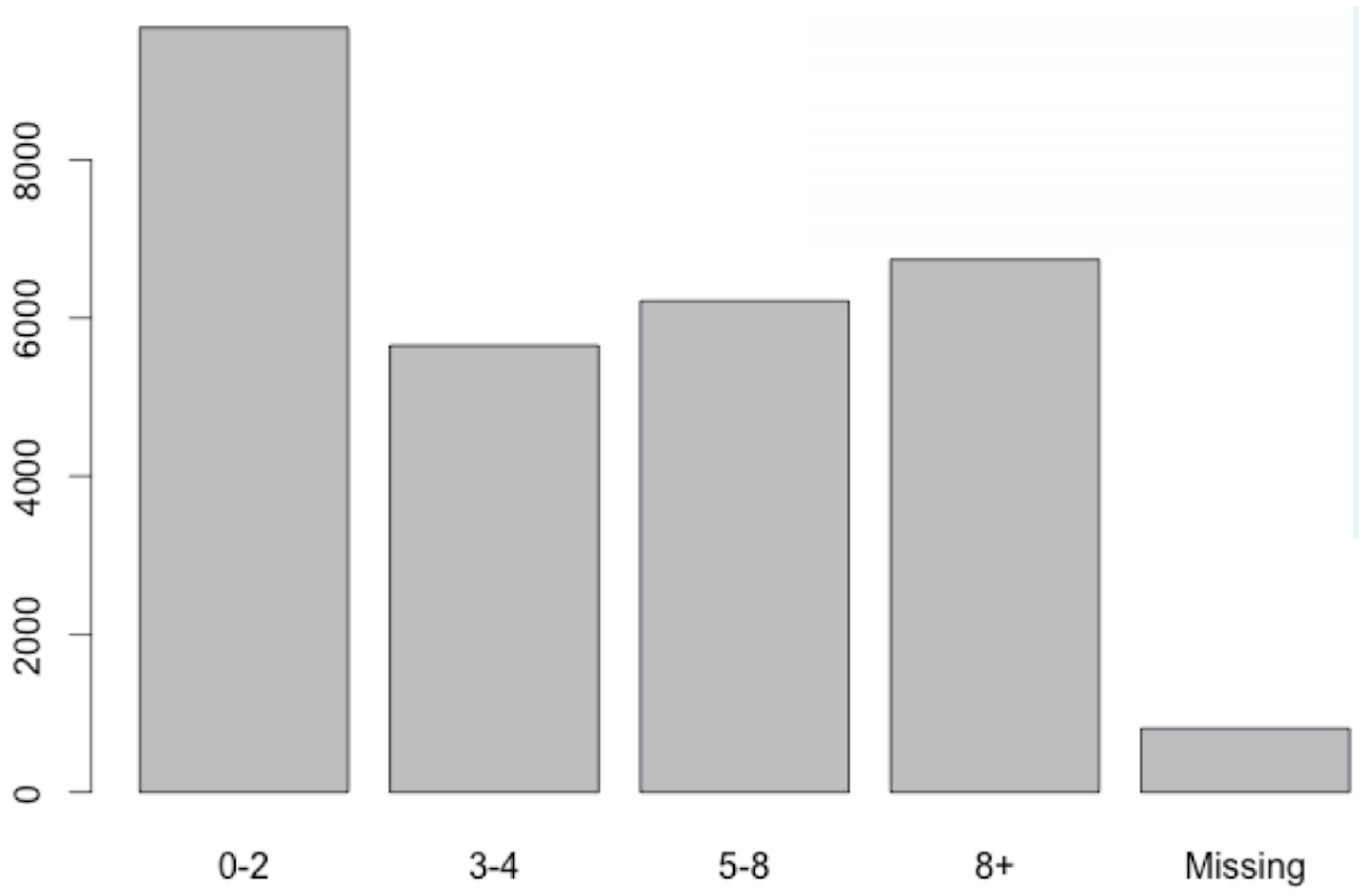
```
plot(loan_data$emp_cat)
```




```
emp_cat
  ...
  0-15
  Missing
  0-15
  0-15
  0-15
  0-15
  ...
  30-45
  0-15
  15-30
  Missing
  0-15
  ...
```

Bin frequencies

```
plot(loan_data$emp_cat)
```




```
emp_cat
  ...
  8+
  Missing
  0-2
  0-2
  0-2
  3-4
  ...
  8+
  5-8
  8+
  Missing
  3-4
  ...
```

Final remarks

- Treat outliers as `NA` s

Final remarks

- Treat outliers as `NA` s

	CONTINUOUS	CATEGORICAL
DELETE	Delete rows (observations with `NA` s) Delete column (entire variable)	Delete rows (observations with `NA` s) Delete column (entire variable)
REPLACE	Replace using median	Replace using most frequent category
KEEP	Keep as `NA` (not always possible) Keep using coarse classification	`NA` category

Let's practice!

CREDIT RISK MODELING IN R

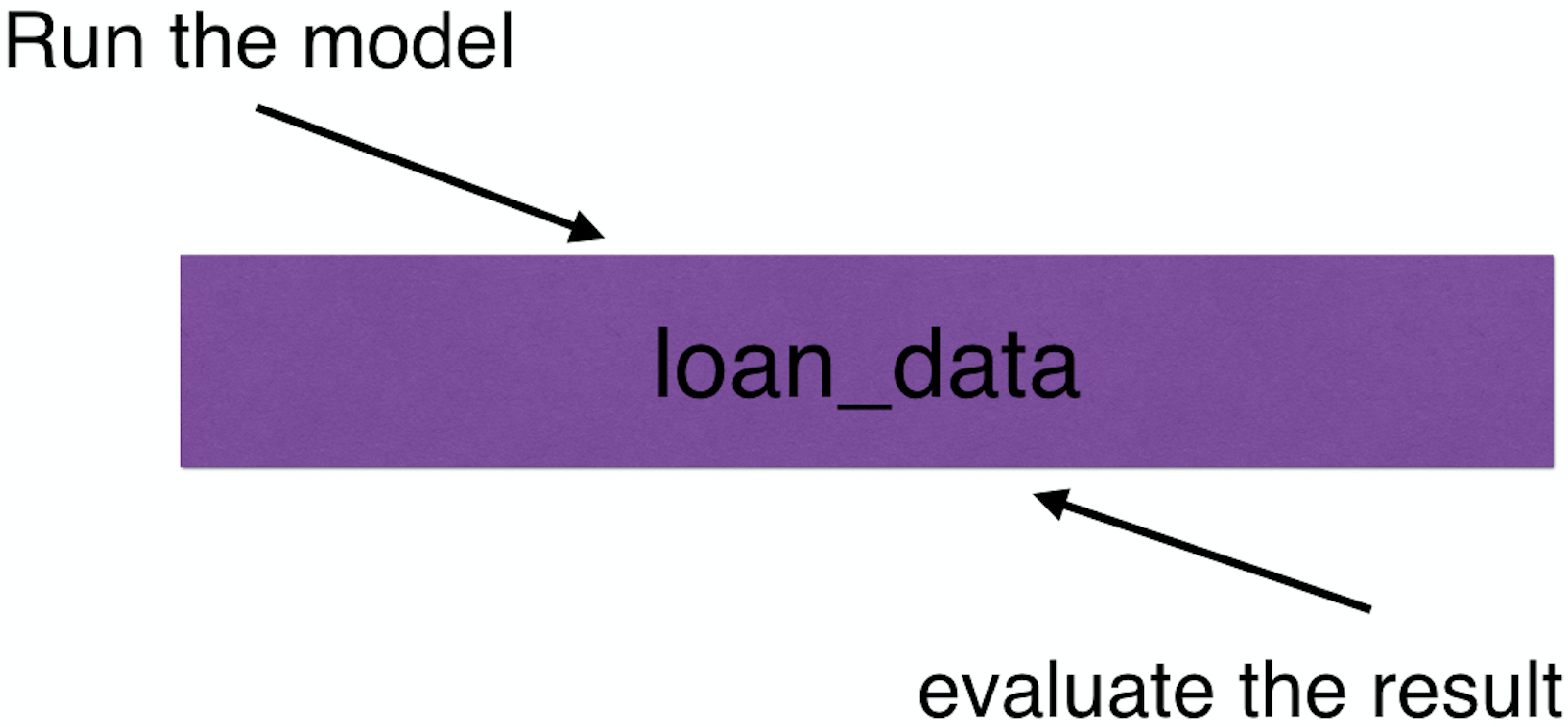
Data splitting and confusion matrices

CREDIT RISK MODELING IN R

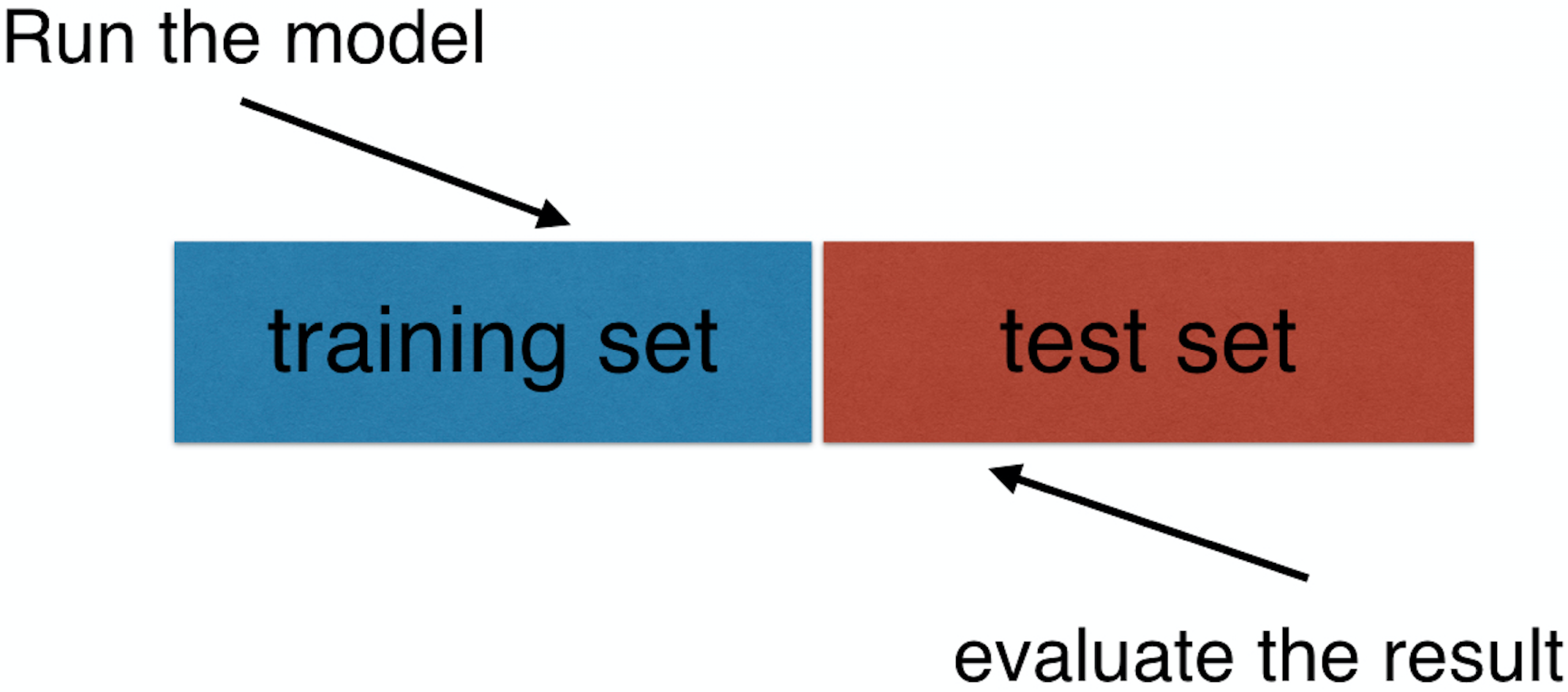
Lore Dirick

Manager of Data Science Curriculum at Flatiron School

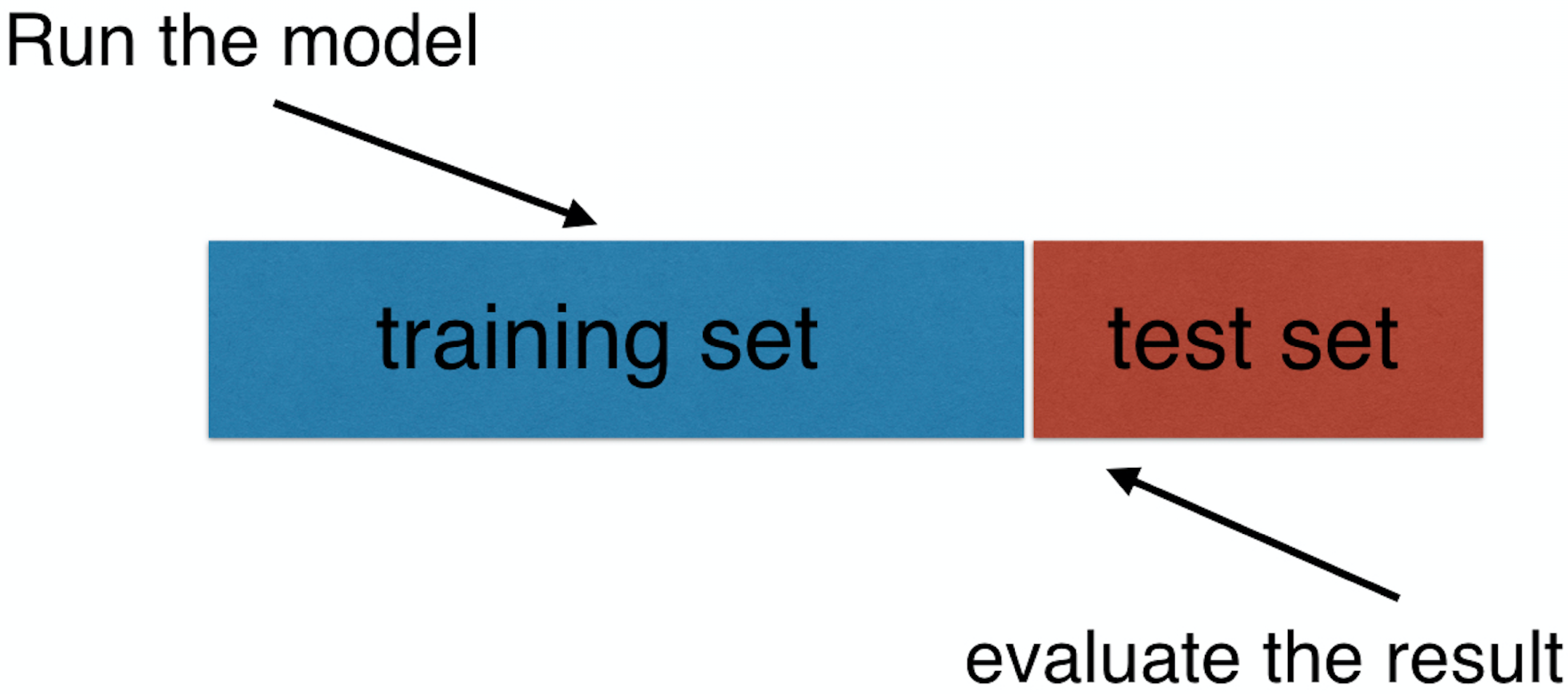
Start analysis



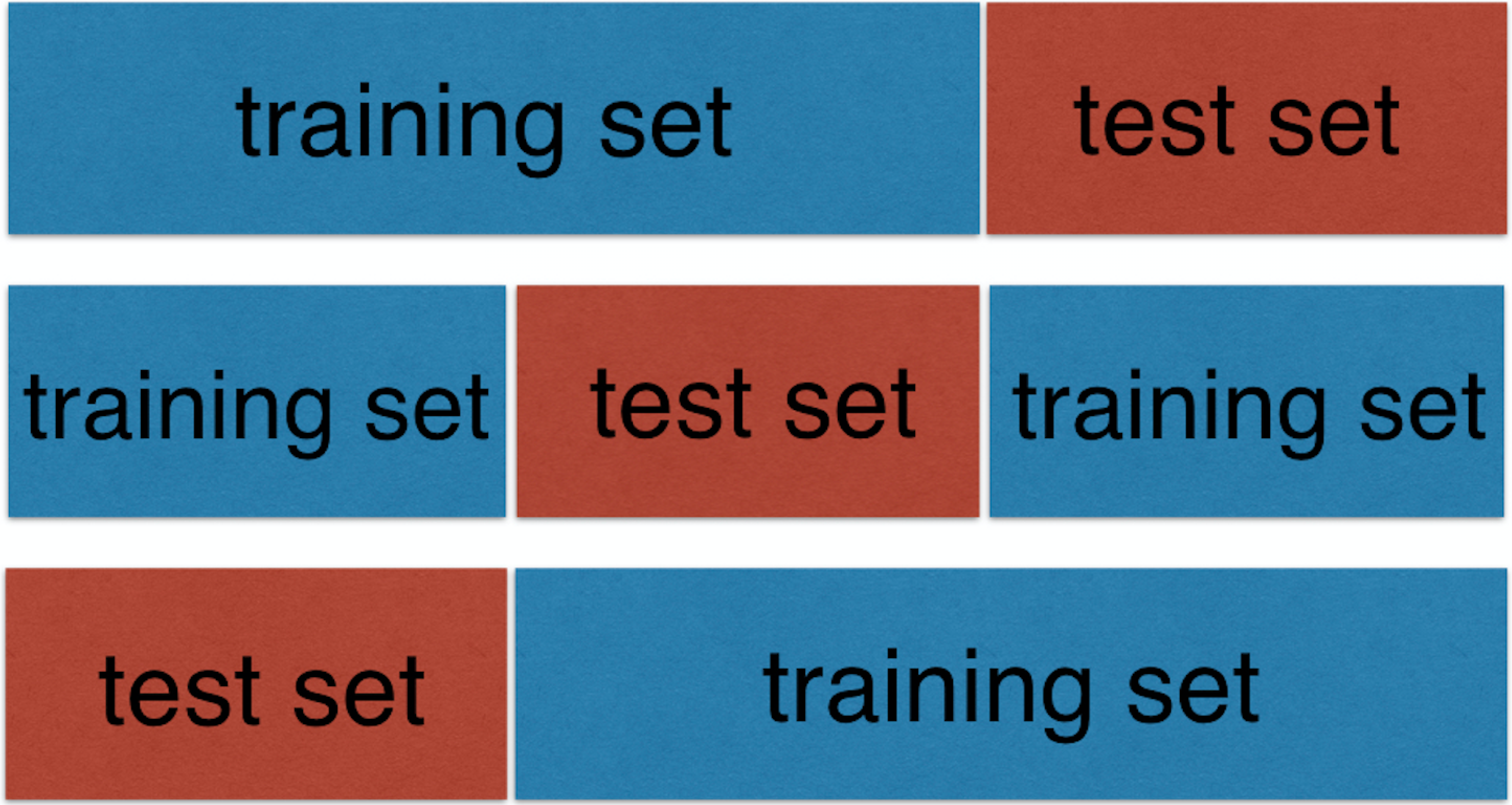
Training and test set



Training and test set



Cross-validation



Evaluate a model

```
      test_set$loan_status    model_prediction
            ...                     ...
 [8066,]      1                       1
 [8067,]      0                       0
 [8068,]      0                       0
 [8069,]      0                       0
 [8070,]      0                       0
 [8071,]      0                       1
 [8072,]      1                       0
 [8073,]      1                       1
 [8074,]      0                       0
 [8075,]      0                       0
 [8076,]      0                       0
 [8077,]      1                       1
 [8078,]      0                       0
        ...                        ...
```

Evaluate a model

```
        test_set$loan_status    model_prediction
               ...                    ...
[8066,]         1                      1
[8067,]         0                      0
 [8068,]        0                      0
 [8069,]        0                      0
 [8070,]        0                      0
 [8071,]        0                      1
 [8072,]        1                      0
 [8073,]        1                      1
 [8074,]        0                      0
 [8075,]        0                      0
 [8076,]        0                      0
 [8077,]        1                      1
 [8078,]        0                      0
 [8079,]        0                      1
```

Actual loan status v. Model prediction

	No default (0)	Default (1)
No default (0)	8	2
Default (1)	1	3

Evaluate a model

```
      test_set$loan_status    model_prediction
                ...                  ...
[8066,]           1                    1
[8067,]           0                    0
 [8068,]          0                    0
 [8069,]          0                    0
 [8070,]          0                    0
 [8071,]          0                    1
 [8072,]          1                    0
 [8073,]          1                    1
 [8074,]          0                    0
 [8075,]          0                    0
 [8076,]          0                    0
 [8077,]          1                    1
 [8078,]          0                    0
 [8079,]          0                    1
```

Actual loan status v. Model prediction

	No default (0)	Default (1)
No default (0)	TN	FP
Default (1)	FN	TP

Some measures...

- *Accuracy*

$$\frac{(8+3)}{14} = 78.57\%$$

- *Sensitivity*

$$\frac{3}{(1+3)} = 75\%$$

- *Specificity*

$$\frac{8}{(8+2)} = 80\%$$

Actual loan status v. Model prediction

	No default (0)	**Default (1)**
No default (0)	8	2
Default (1)	1	3

Let's practice!

CREDIT RISK MODELING IN R