Class 5: Predicting Response with RFM analysis

Professor Tae Jung Yoon (윤태중)
KAIST College of Business

Marketing Research & Analytics

Marketing Research & Analytics Course Structure

Getting Ready for Marketing Research and Analytics

- Marketing Research and Analytics Overview (Class 1)
- How to Tell Good From Bad Data Analytics (Class 2)
- Using Stata for Marketing Research and Analytics (Classes 2 & 3)
- Statistics Review (Class 4)

Understanding Customers and Markets

- Quantifying Customer Value (Class 1)
- Case Analysis: "Home Alarm, Inc.: Assessing Customer Lifetime Value" and Testing (Class 3)
- Measuring Customers' Willingness to Pay (Class 6)
- Valuation of Products: Conjoint Analysis (Classes 8 & 9)
- Market Segmentation: Cluster Analysis (Class 10)
- Survey, and Qualitative Research (Class 10)

Prospecting and Targeting the Right Customers

- Predicting Response with RFM analysis (Class 5)
- Case Analysis: "Tuango: RFM Analysis for Mobile App Push Messaging"; Lift and Gains (Class 6)
- Predicting Response with Logistic Regression (Class 7)
- Case Analysis: "BookBinders: Predicting Response with Logistic Regression" (Class 8)
- Predicting Response with Neural Networks (Class 9)
- Predicting Response with Decision Trees (Class 10)

Developing Customers

- Case Analysis: "Intuit: Quickbooks Upgrade" (Class 11)
- Next-Product-To-Buy Models: Learning From Purchases (Class 11)
- Recommendation Systems: Learning From Ratings (Class 12)

Retaining Customers

- Predicting Attrition (Class 12)

Selecting the Right Offers

- Design of Experiments / Multi Variable Testing (Class 13)
- Case Analysis: "Capital One: Information-Based Credit Card Design" (Class 14)

Limitations of Marketing Analytics

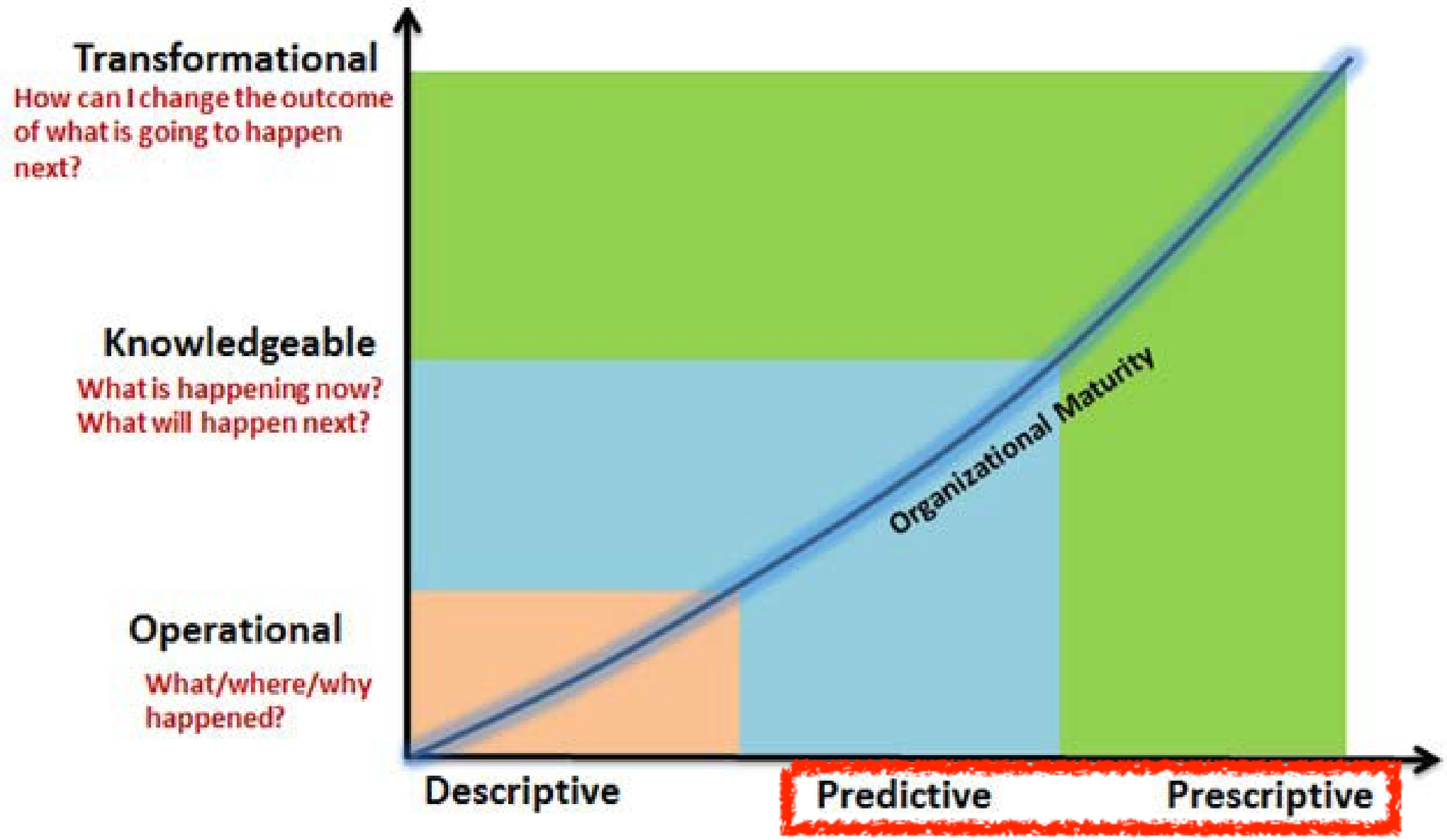
- When Marketing Analytics, CRM, and Databases Fail (Class 14)

Wrap-up

- Wrap-Up (Class 14)

Predictive and prescriptive analytics are at the top of many “analytics maturity curves”

EXAMPLE 1



Source: bdisys.com

“Length of stay” is an important hospital performance metric

LENGTH OF STAY (LOS)

- Number of days of inpatient care utilized by a patient

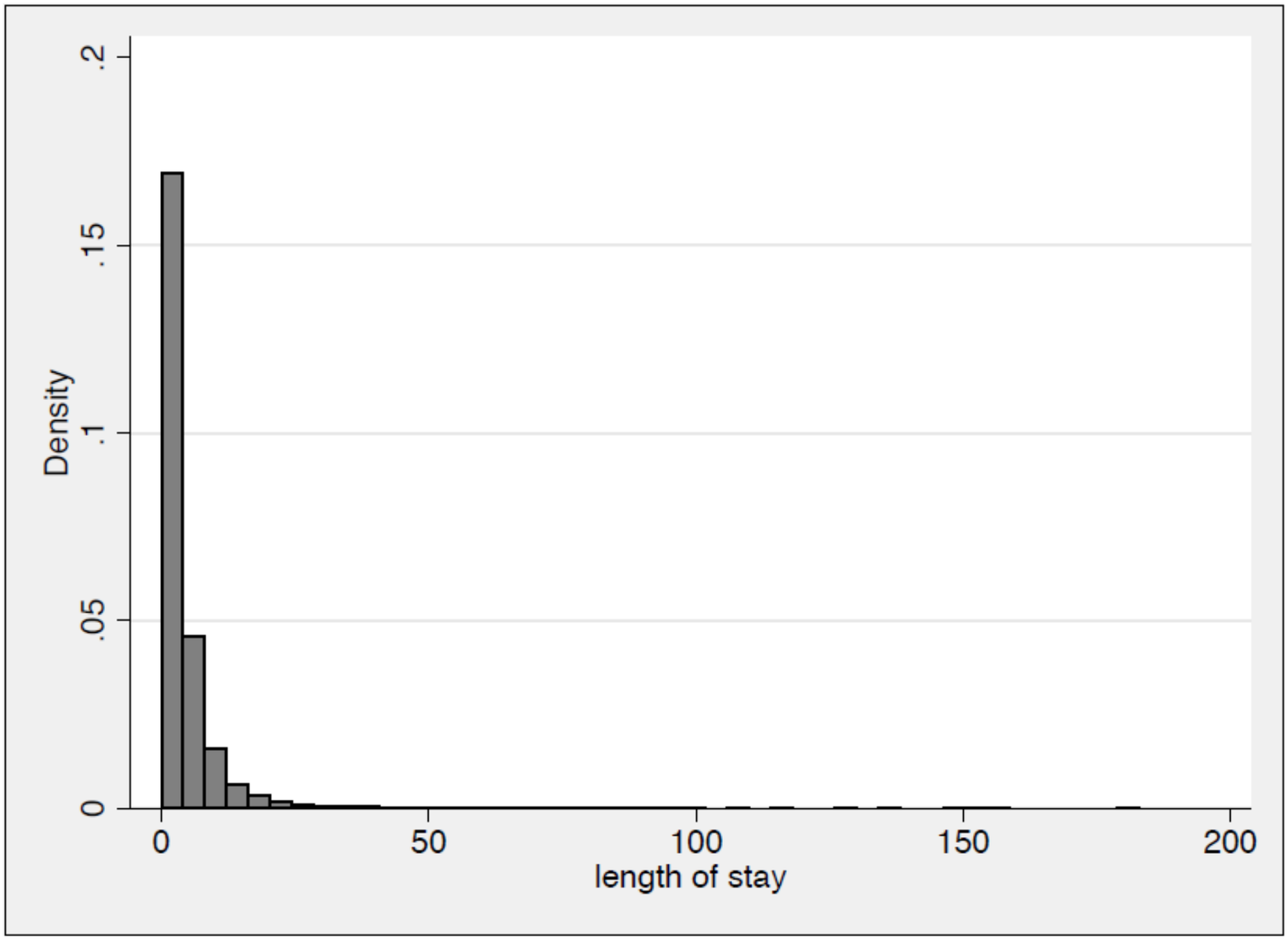
IMPORTANT WHY?

- Improving efficiency is key hospital concern
- Inpatient **efficiency** is largely a function of LOS
- Hospitals are typically paid a set fee for treating all patients in an “Diagnosis-Related Group,” **regardless of the actual cost** (e.g. LOS) for that case
- LOS analysis is crucial for **elective patient admission and assignment planning**

Source: “Measuring Hospitalist performance: Metrics, reports, and dashboards,” white paper, The Society of Hospital Medicine’s Benchmarks Committee

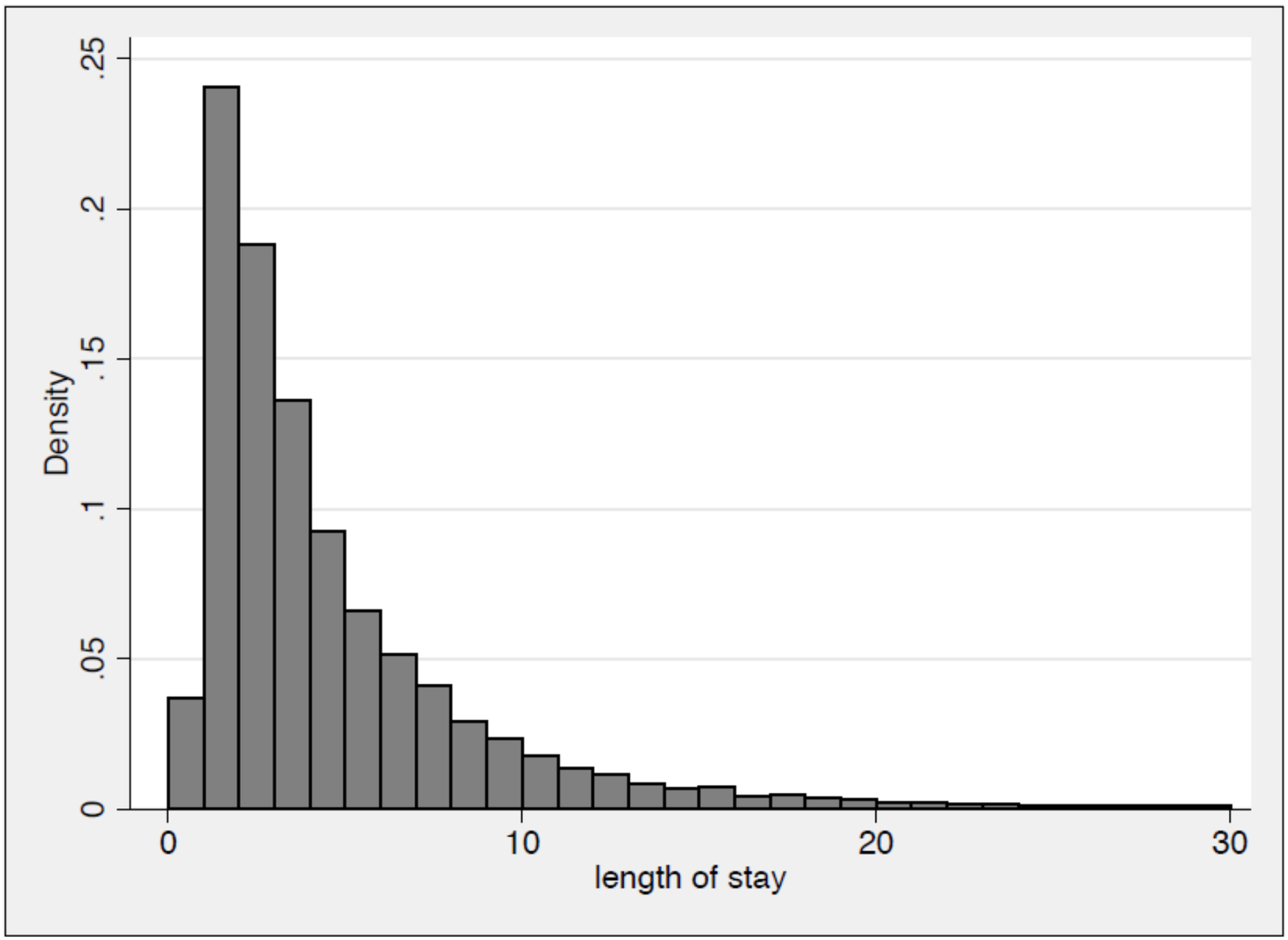
We use patient discharge data for San Diego county to understand length-of-stay

LOS for "Diseases & Disorders Of The Circulatory System"



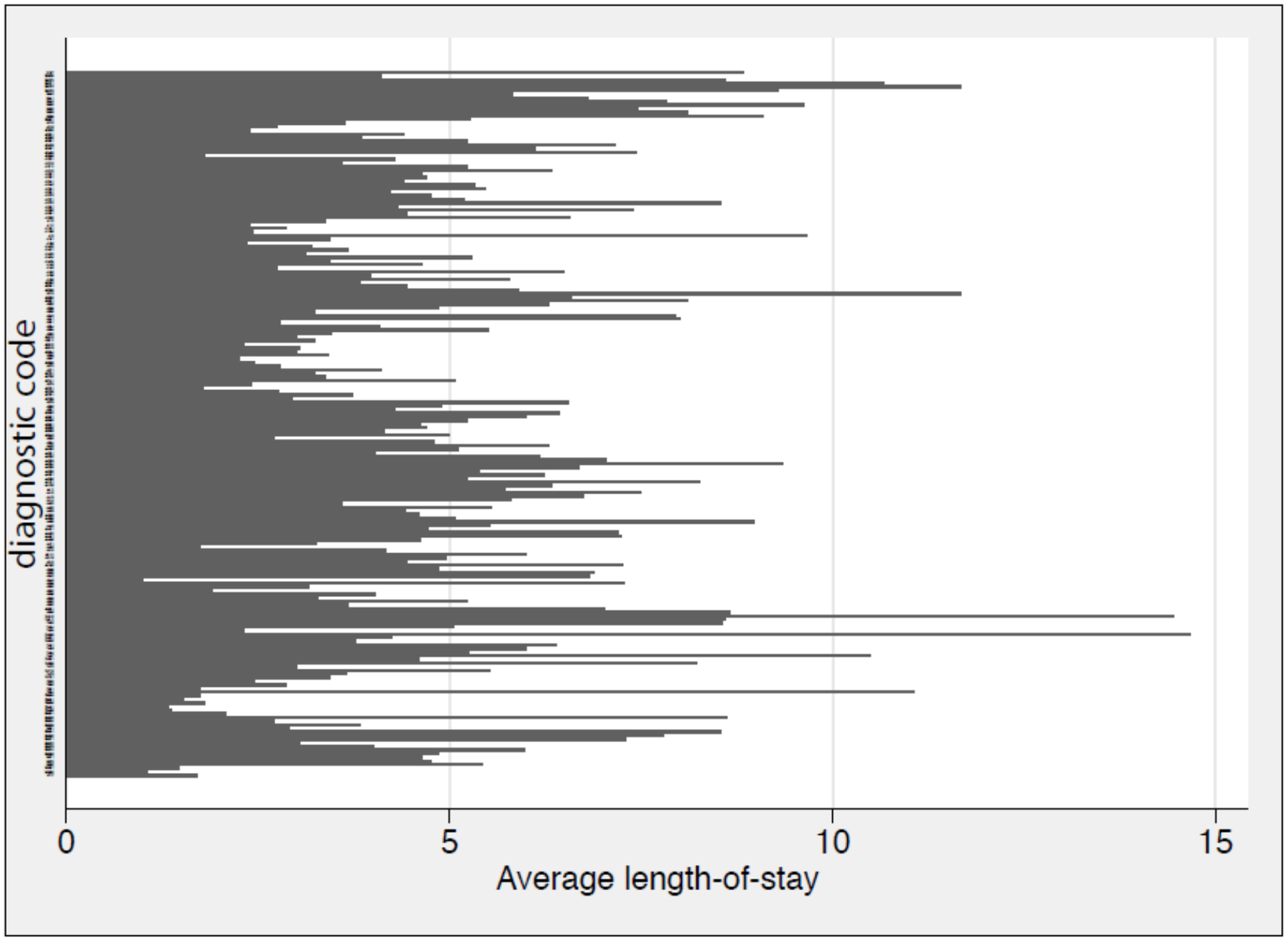
We use patient discharge data for San Diego county to understand length-of-stay

LOS≤30 for "Diseases & Disorders Of The Circulatory System"



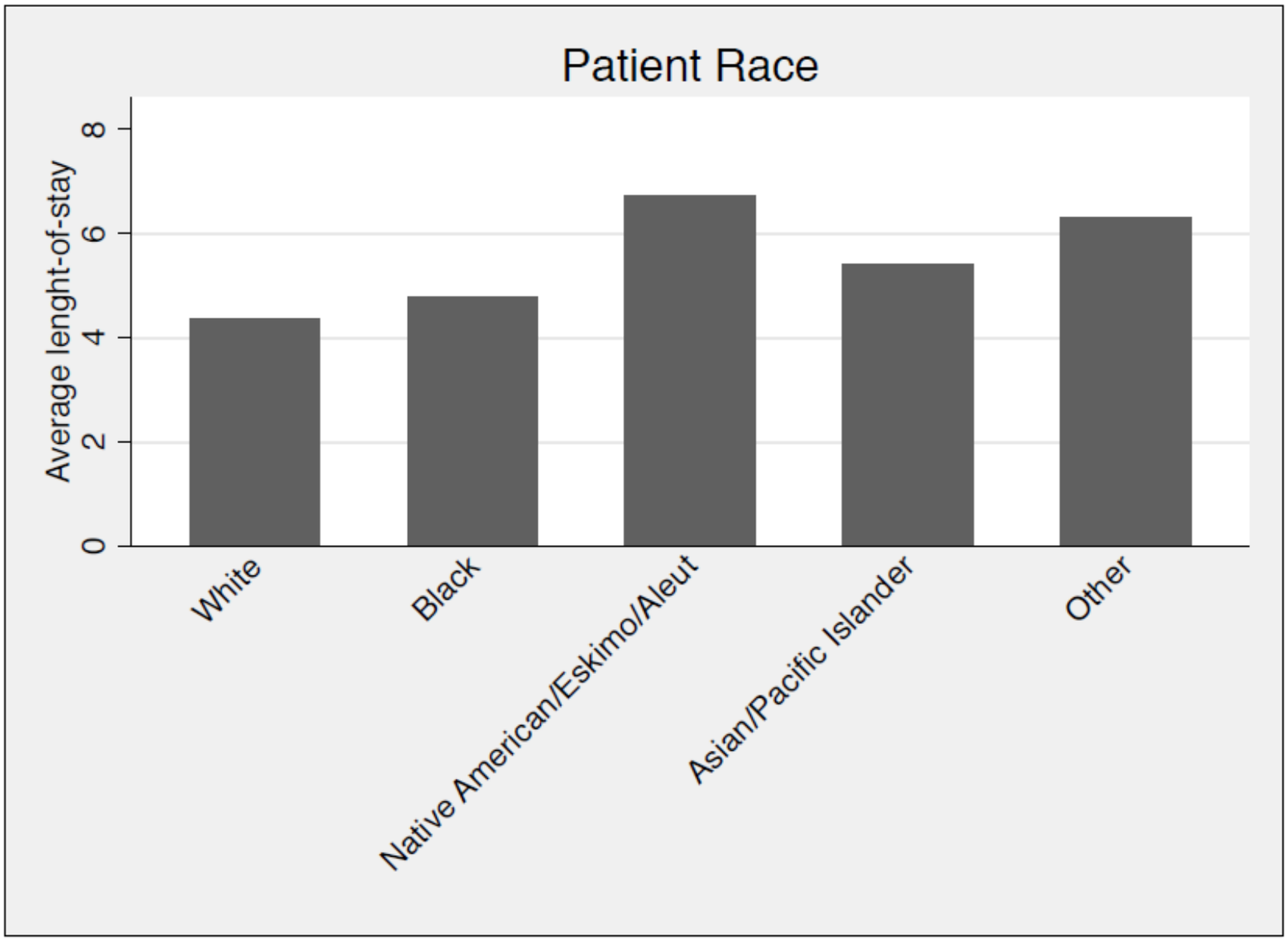
Length-of-stay varies strongly by diagnosis

LOS BY ICD-9-CM diagnostic code



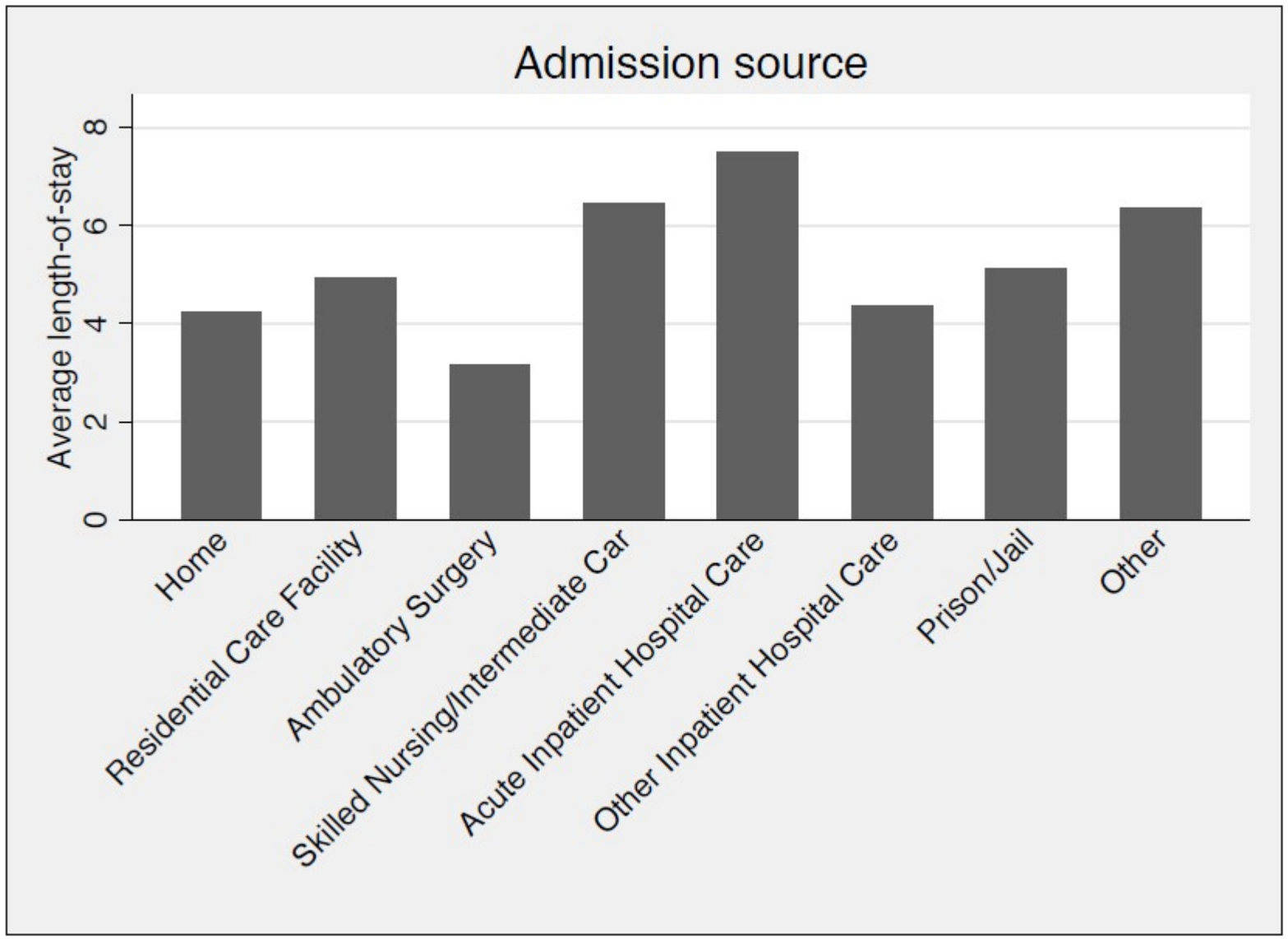
Length-of-stay varies by the race of the patient

LOS BY PATIENT RACE



Length-of-stay also varies by admission source

LOS BY PATIENT ADMISSION SOURCE



If you had only data on LOS by diagnosis, how would you form a prediction for LOS for a newly diagnosed patient?

LOS BY DIAGNOSIS

ICD-9-CM diagnostic code	LOS	Frequency
428.0 chf nos	4.9	5,496
414.01 crnry athrscl natve vssl	3.7	4,700
786.59 chest pain nec	1.8	2,833
410.71 subendo infarct, initial	5.2	2,320
427.31 atrial fibrillation	3.2	1,309
780.2 syncope and collapse	2.5	1,184
786.50 chest pain nos	1.8	1,162
410.41 ami inferior wall, init	4.4	622
427.89 cardiac dysrhythmias nec	2.9	555
427.81 sinoatrial node dysfunct	3.7	533
453.8 venous thrombosis nec#	4.9	514
996.62 react-oth vasc dev/graft	7.8	446
402.91 hyp ht dis nos w ht fail	5.2	440
410.11 ami anterior wall, init	5.2	417
424.1 aortic valve disorder	8.0	358
…	…	…

How can LOS predictions help?

USING LOS ESTIMATES

Descriptive Analytics:

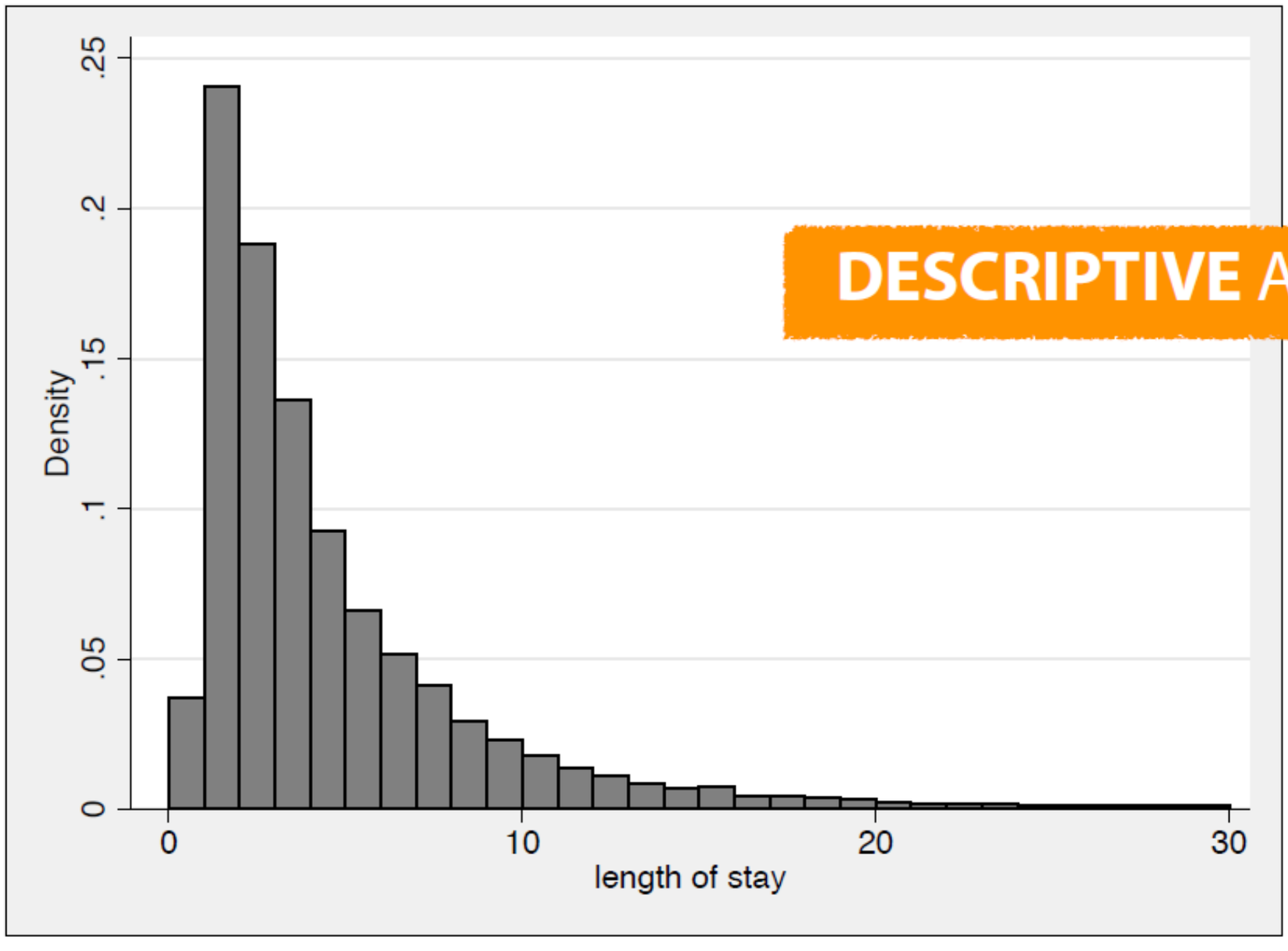
Descriptive statistics that **summarize data**,

often over time and /or by groups, geography, etc.

Examples: Averages, histograms, counts, sums, percentages, min max and simple transformations of the data.

We use patient discharge data for San Diego county to understand length-of-stay

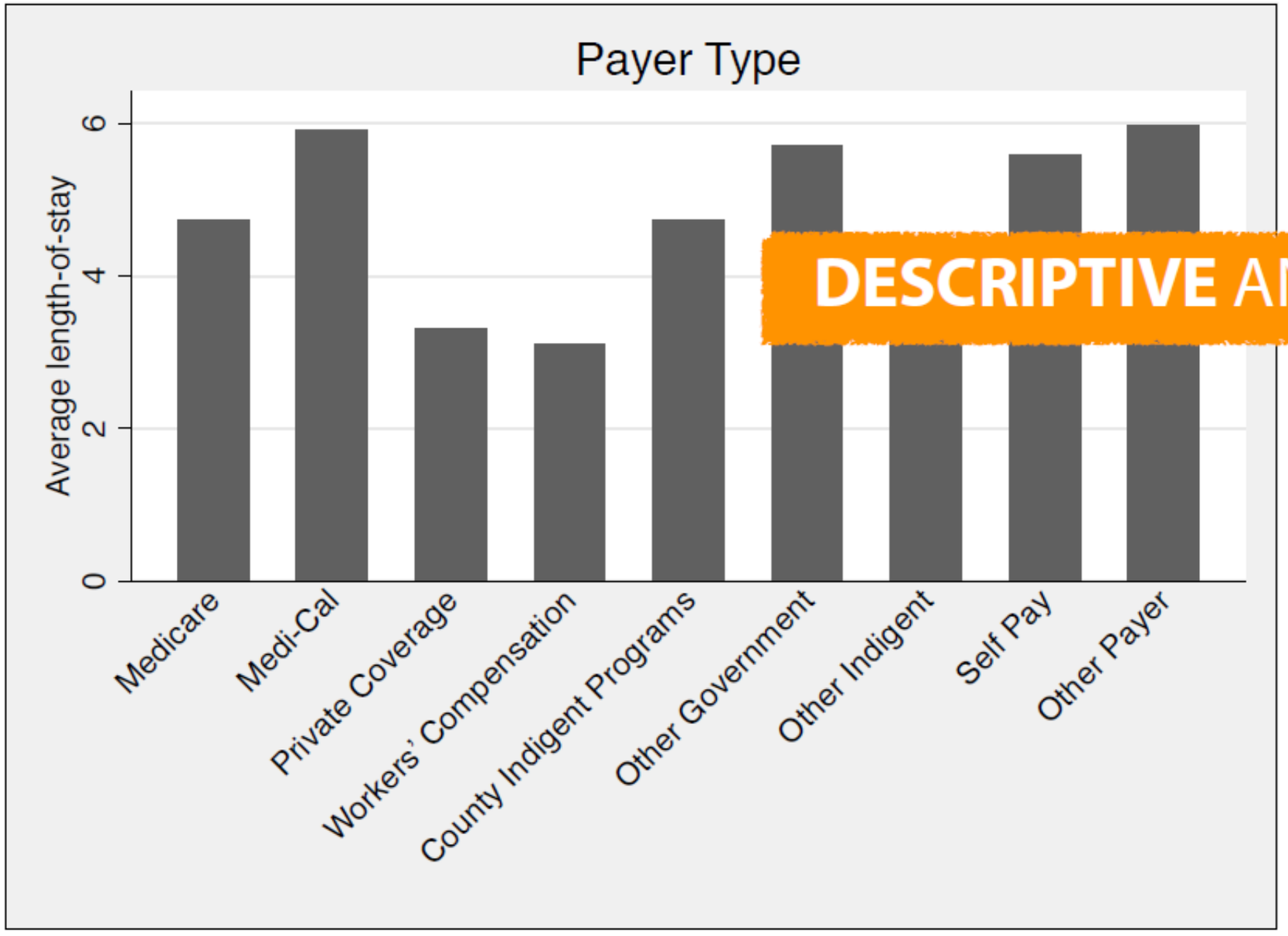
LOS≤30 for "Diseases & Disorders Of The Circulatory System"



DESCRIPTIVE ANALYTICS

Length-of-stay varies by type of payer

LOS BY PAYER TYPE



DESCRIPTIVE ANALYTICS

Predictive Analytics:

Using data **that you have** to predict data that **you don't have**, using **statistical** or **machine learning** approaches.

Examples: Sales probabilities, future inventory levels, customer churn rates in the future, part failures in the future, ...

If you had only data on LOS by diagnosis, how would you form a prediction for LOS for a newly diagnosed patient?

LOS BY DIAGNOSIS

ICD-9-CM diagnostic code	LOS	Frequency
428.0 chf nos	4.9	5,496
414.01 crnry athrscl natve vssl	3.7	4,700
786.59 chest pain nec	1.8	2,833
410.71 subendo infarct, initial		
427.31 atrial fibrillation		
780.2 syncope and collapse	2.5	1,184
786.50 chest pain nos	1.8	1,162
410.41 ami inferior wall, init	4.4	622
427.89 cardiac dysrhythmias nec	2.9	555
427.81 sinoatrial node dysfunct	3.7	533
453.8 venous thrombosis nec#	4.9	514
996.62 react-oth vasc dev/graft	7.8	446
402.91 hyp ht dis nos w ht fail	5.2	440
410.11 ami anterior wall, init	5.2	417
424.1 aortic valve disorder	8.0	358
…	…	…

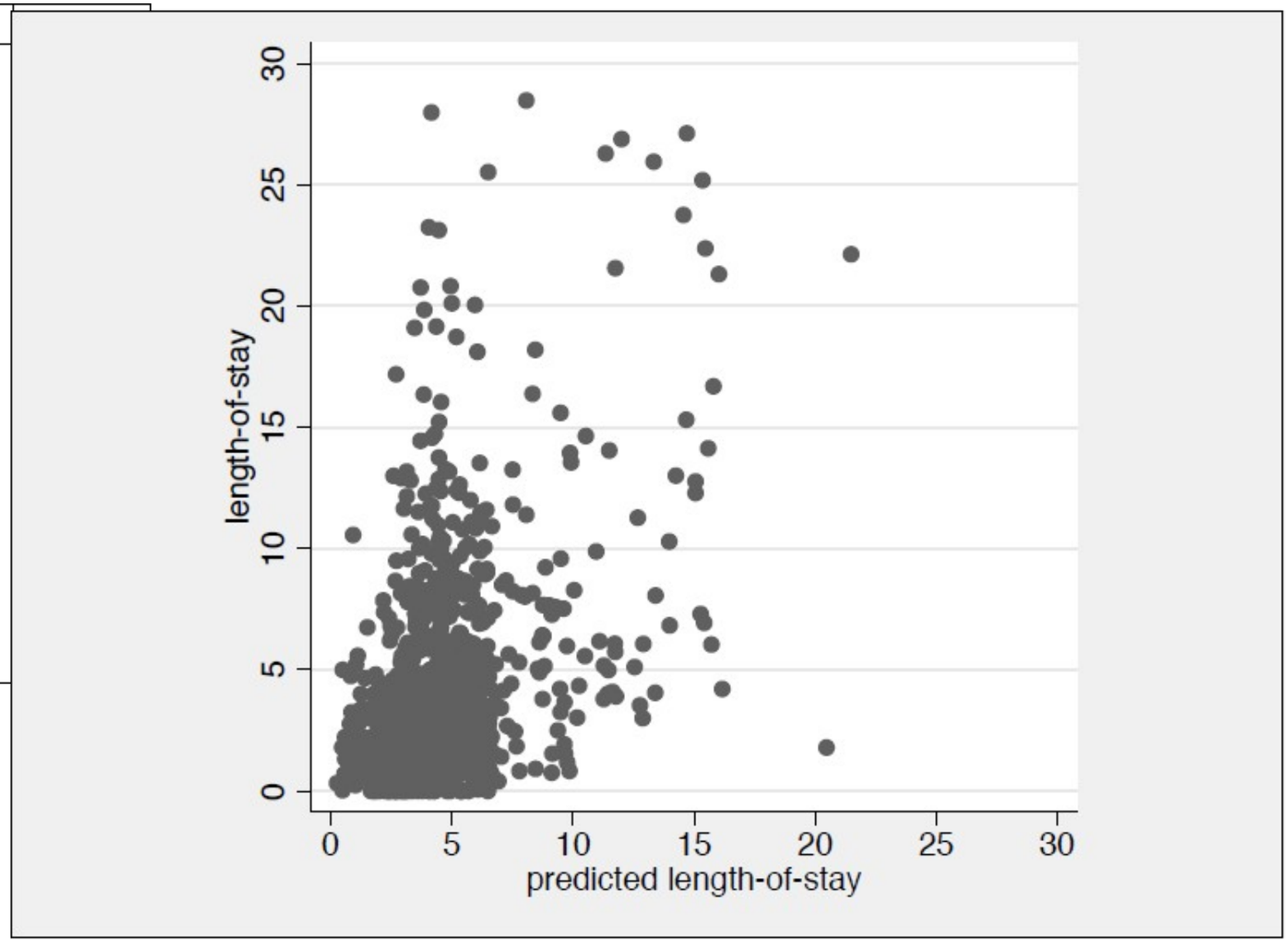
PREDICTIVE ANALYTICS

Predictions using "*average LOS for same diagnosis*" don't do very well

VARIABLE

ICD-9-CM diagnostic code	LOS
428.0 chf nos	4.9
414.01 crnry athrscl natve vssl	3.7
786.59 chest pain nec	1.8
410.71 subendo infarct, initial	5.2
427.31 atrial fibrillation	3.2
780.2 syncope and collapse	2.5
786.50 chest pain nos	1.8
410.41 ami inferior wall, init	4.4
427.89 cardiac dysrhythmias nec	2.9
427.81 sinoatrial node dysfunct	3.7
453.8 venous thrombosis nec#	4.9
996.62 react-oth vasc dev/graft	7.8
402.91 hyp ht dis nos w ht fail	5.2
410.11 ami anterior wall, init	5.2
424.1 aortic valve disorder	8.0
…	…

PREDICTIVE PERFORMANCE FOR TEST PATIENTS

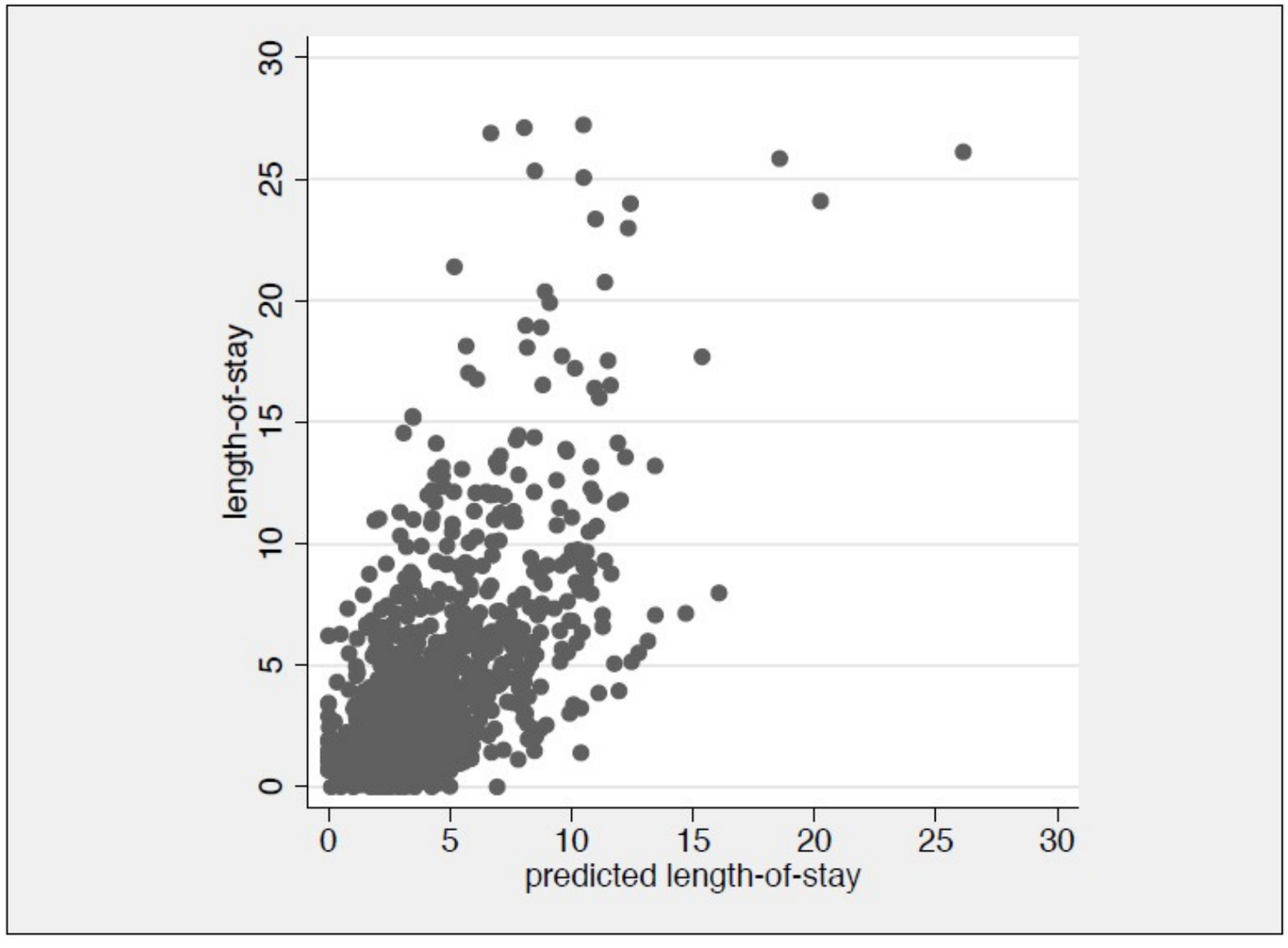


Using *linear regression model* with more variables

VARIABLES

- Diagnosis
- Procedure
- Race
- Age
- Sex
- Admission quarter
- Admission type
- Admission source
- Payer category
- DNR status

PREDICTIVE PERFORMANCE FOR TEST PATIENTS

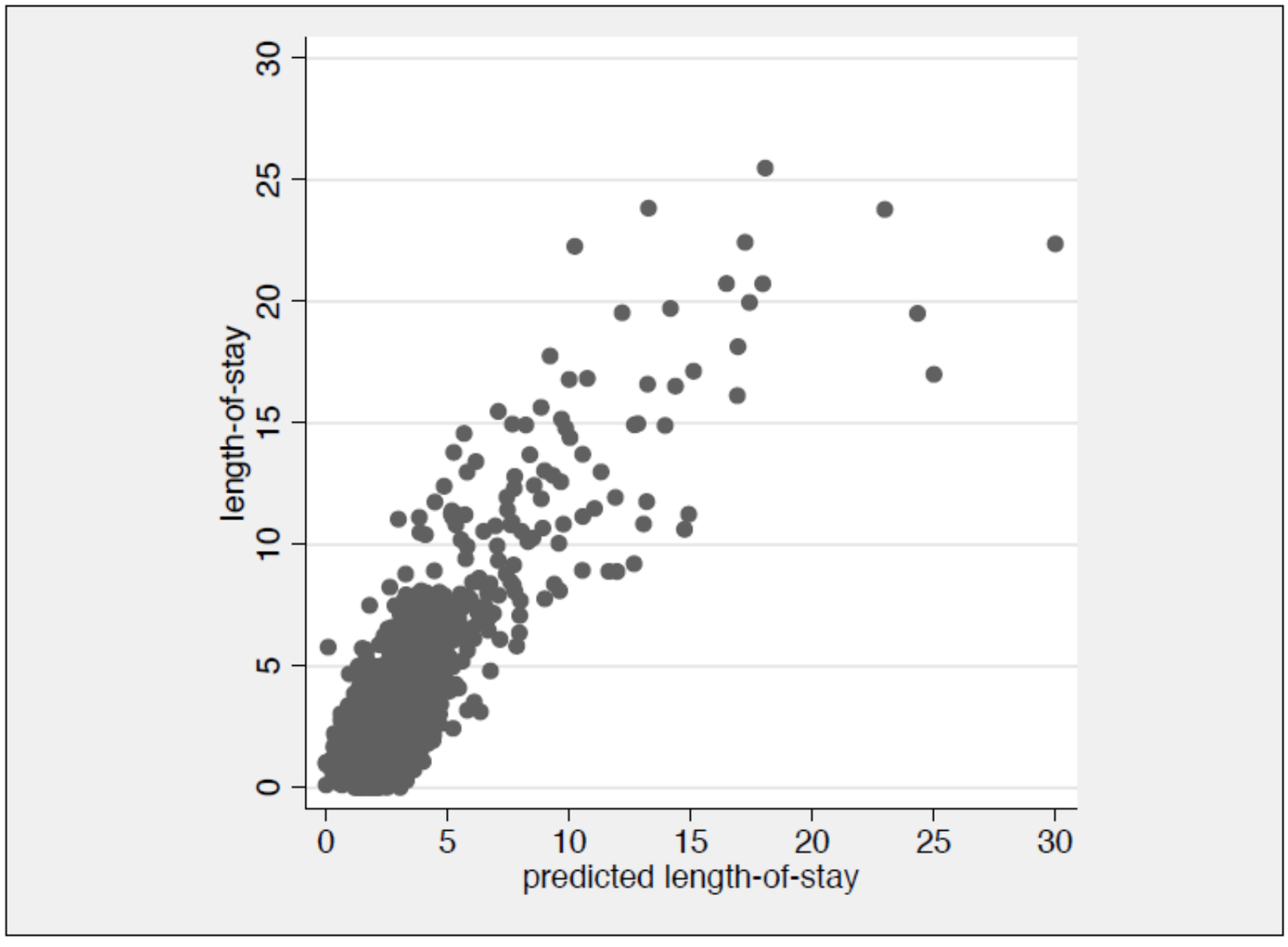


Hazard model (survival analysis) with more variables do much better

VARIABLES

- Diagnosis
- Procedure
- Race
- Age
- Sex
- Admission quarter
- Admission type
- Admission source
- Payer category
- DNR status

PREDICTIVE PERFORMANCE FOR TEST PATIENTS



Prescriptive Analytics:

Using predictive analytics to **take actions** based on the predicted outcomes

Examples: Maintenance schedules, sales calls, proactive churn interventions, revenue management, asset allocations ...

LOS predictions determine elective patient admissions

USING LOS ESTIMATES

Predictive models come in non-statistical and statistical varieties

TYPES OF PREDICTIVE MODELS

- Heuristics (rule of thumb)
  - Recency, frequency, Monetary (RFM) analysis
- Analyst-driven models (Statistical)
  - Regression models
  - Discrete choice models
- Data-driven models (Machine Learning)
  - Neural Networks
  - Decision Trees (Boosted and Random Forests)
  - Recommender systems

RFM is easy to use and requires no statistical knowledge

RFM FACTS

- We know normally two things about customers
  - Who they are (demographic data)
  - What they do (behavioral data)

- RFM is purely based on behavior

- Applies only to existing customers, not to prospects

- Widely used, works well

- Easy to use (no analytics team required)

- Works for B2C and B2B

- RFM is about response rate, not profitability

The premise of RFM is that past behavior predicts future behavior

ELEMENTS OF AN RFM ANALYSIS

- **Recency**
  - How long ago did the customer make the purchase?
- **Frequency**
  - How many purchases has the customer made (in given time period)
- **Monetary**
  - How much has the consumers spent in total (in given time period)

We begin by coding RFM into N-tiles (quantiles)

START OF RFM ANALYSIS

- Pick variable of interest (recency, frequency, monetary)
- Sort database from best to worst on variable
- Decide into how many groups to classify consumers
  - Normally pick 5 groups --> split consumers into "quintiles"
  - If pick 10 groups --> "deciles"
- Assign consumers to groups
  - Top group is quintile 1, second is quintile 2, etc.
  - Make sure most desirable group for each variable is in the first group

Next we combine the N-tiles into an “RFM Index”

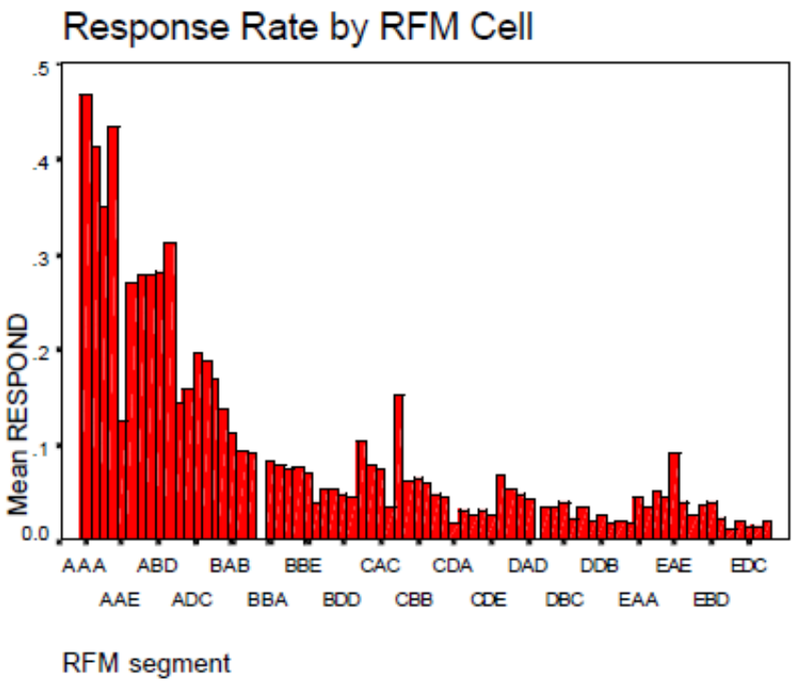
NEXT STEPS IN RFM ANALYSIS

- Assign every customer a 3-digit code, e.g. 125, 555, etc.
  (quintile for R, quintile for F, quintile for M)
  -> will have 125 cells for quintiles

- Take a random sample of customers

- Approach them with an offer

- Calculate response rate per RFM cell



Finally, we select profitable cells and extend the offer to them

FINAL STEPS IN RFM ANALYSIS: CALCULATE BREAK-EVEN RESPONSE RATE

- An offer is profitable if

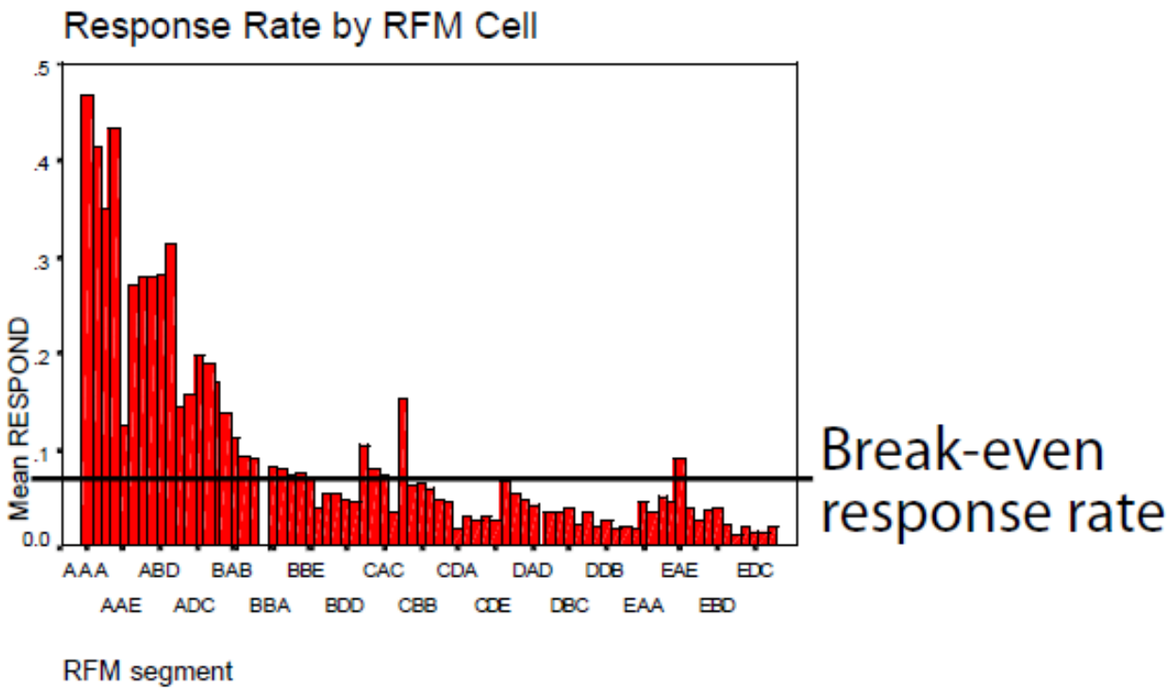
Expected Profits

Response Rate in each cell x Profit on Sale - Cost of Offer ≥ 0

<=>

Response Rate in each cell ≥ Cost of Offer/Profit on Sale

- The lowest probability of response for which the offer is still profitable is called the **"Break-Even Response Rate" = Cost of Offer/Profit on Sale**



- Select cells with above break even response rate from test

- Approach all consumers in the cells with an offer

We use the Bookbinders Book Club as an example of how to target an offer with RFM analysis

RFM TEST AT BOOKBINDERS

- Stan Lawton (marketing director) pulls a random sample of 50,000 customers from the Bookbinders database

- Stan mails "The Art History of Florence" to the entire sample

- 4522 customers buy the book

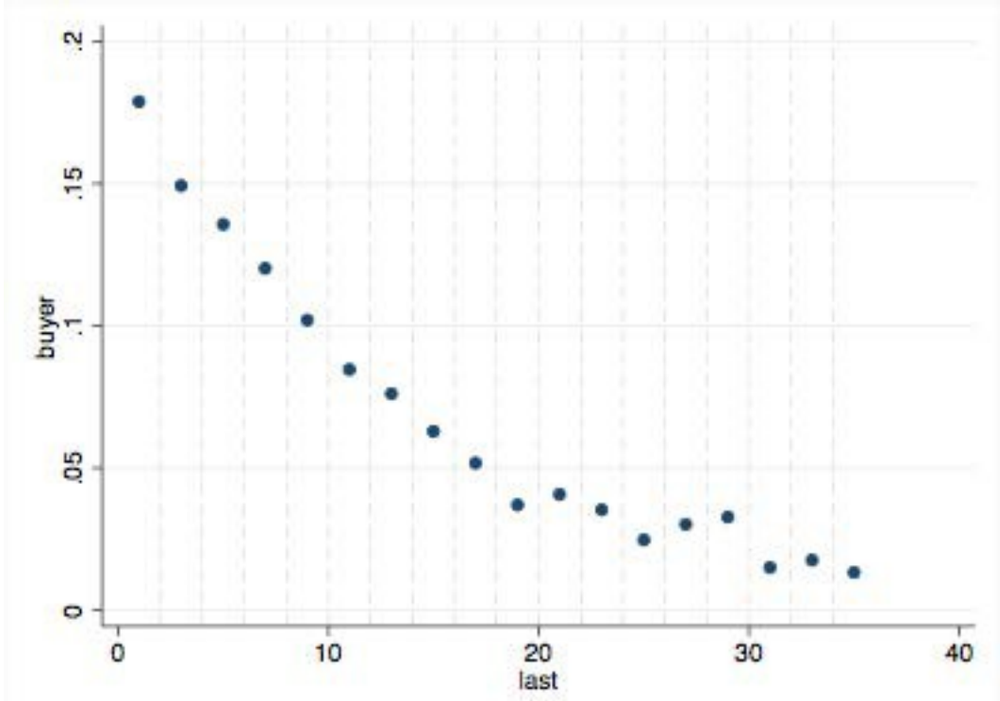
- Stan has information on the
  - recency of the last purchase,
  - the purchase frequency, and
  - total expenditure of each customers

- Plans to use test to determine which customers to target from the entire database (500,000 remaining customers, excluding test group)

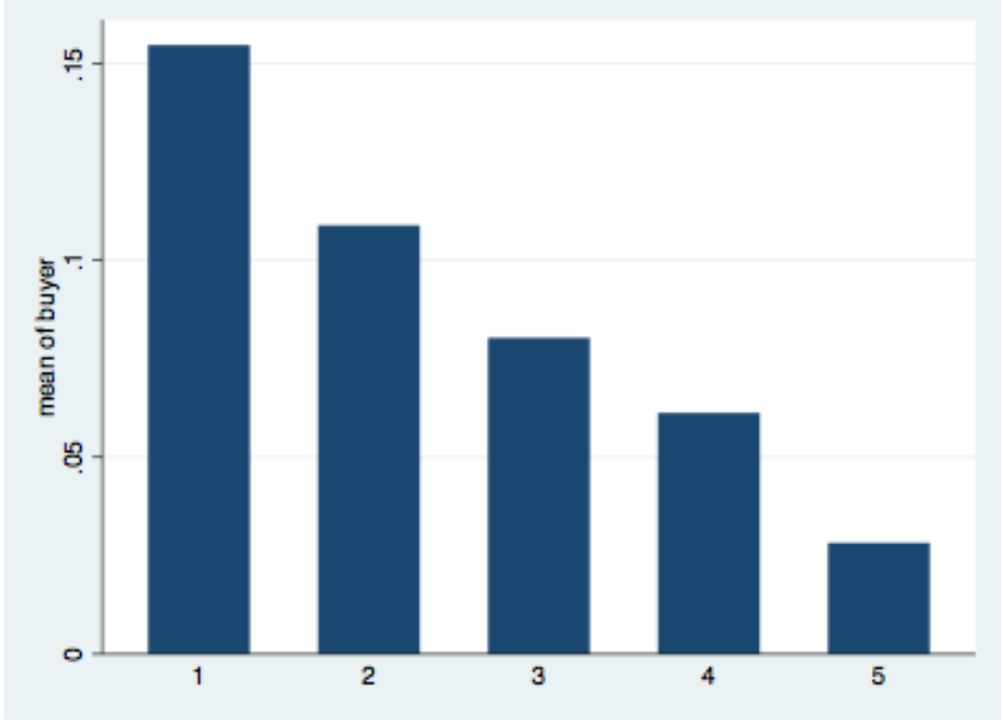
We begin by assessing the recency variable and forming quintiles

% BUYERS BY RECENCY

```
. qscatter buyer last
```




```
. xtile rec_quin=last, nquantile(5)
. graph bar (mean) buyer, over(rec_quin)
```



Does recency predict purchase prob?

Will the best customers be in quintile 1?

Next, we form frequency quintiles

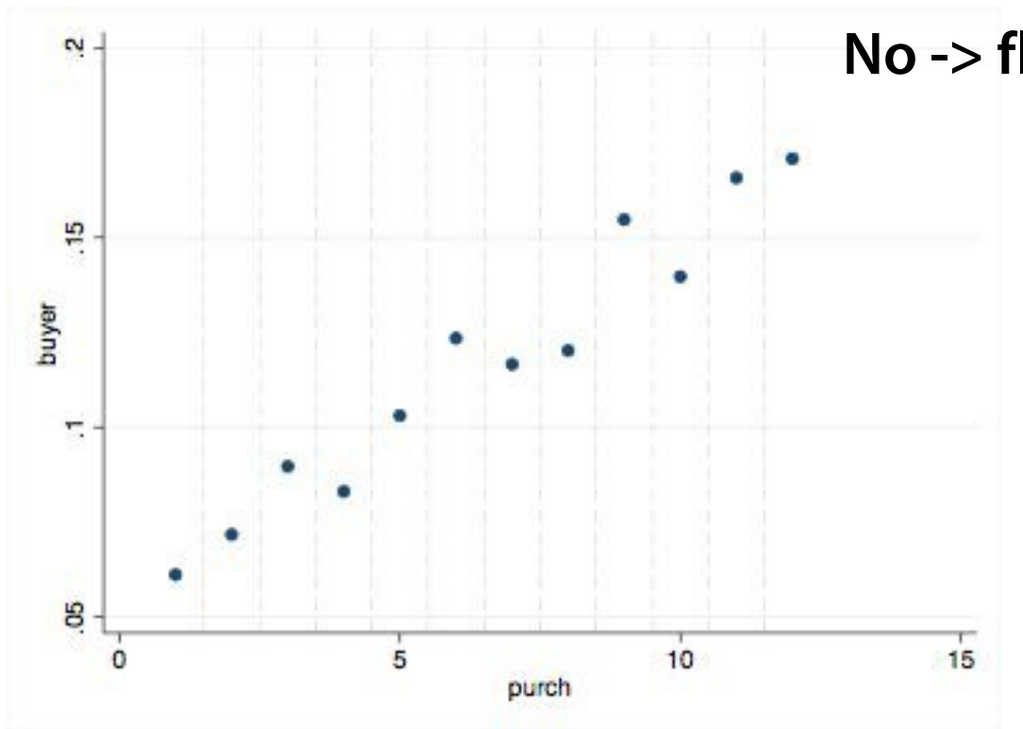
% BUYERS BY FREQUENCY

```
. qscatter buyer purch
```

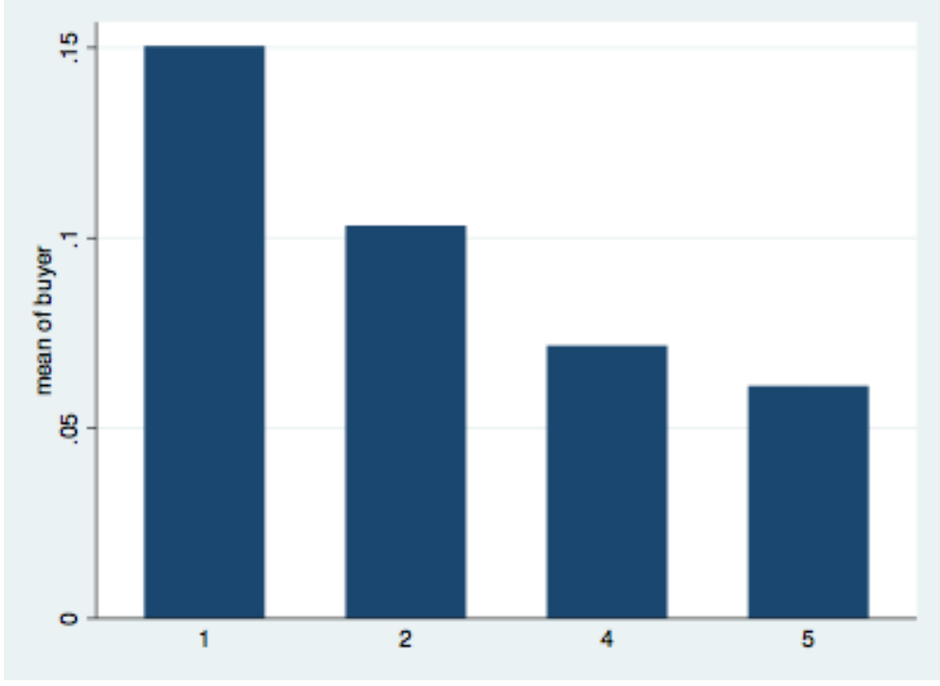
Does frequency predict purchase prob?

Will the best customers be in quintile 1?

No -> flip scale!




```
. xtile freq_quin=purch, nquantile(5)
. replace freq_quin=6-freq_quin
  graph bar (mean) buyer, over(freq_quin)
```



TABULATION OF NUMBER OF PURCHASES

```
. tabulate purch

  total # |
purchases |      Freq.     Percent        Cum.
----------+-----------------------------------
        1 |     15,120       30.24       30.24
        2 |     14,935       29.87       60.11
        3 |      2,019        4.04       64.15
        4 |      1,963        3.93       68.07
        5 |      2,018        4.04       72.11
        6 |      1,984        3.97       76.08
        7 |      2,058        4.12       80.19
        8 |      1,955        3.91       84.10
        9 |      1,945        3.89       87.99
       10 |      1,968        3.94       91.93
       11 |      2,033        4.07       96.00
       12 |      2,002        4.00      100.00
----------+-----------------------------------
    Total |     50,000      100.00
```

Finally, we form a monetary quintile

% BUYERS BY MONETARY QUINTILE

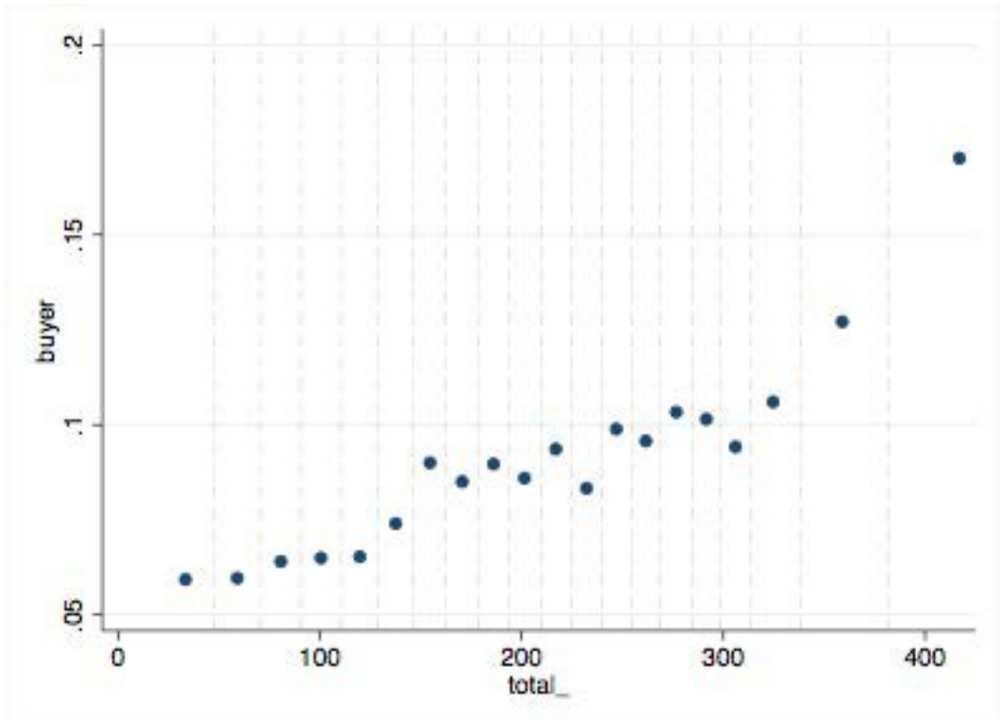
```
. qscatter buyer total_
```

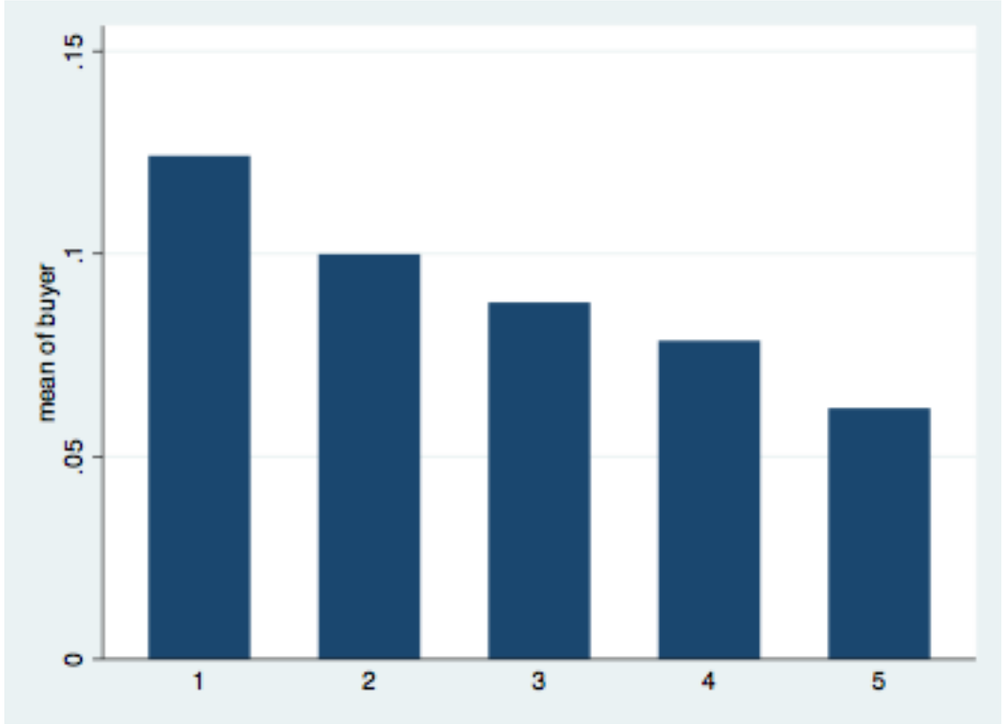
```
. xtile mon_quin=total_, nquantile(5)
. replace mon_quin=6-mon_quin
. graph bar (mean) buyer, over(mon_quin)
```

Does monetary predict purchase prob?

Are the best customers in quintile 1?

No -> flip scale!





Do RFM all capture the same underlying behavioral characteristic?

CORRELATION BETWEEN RFM VARIABLES

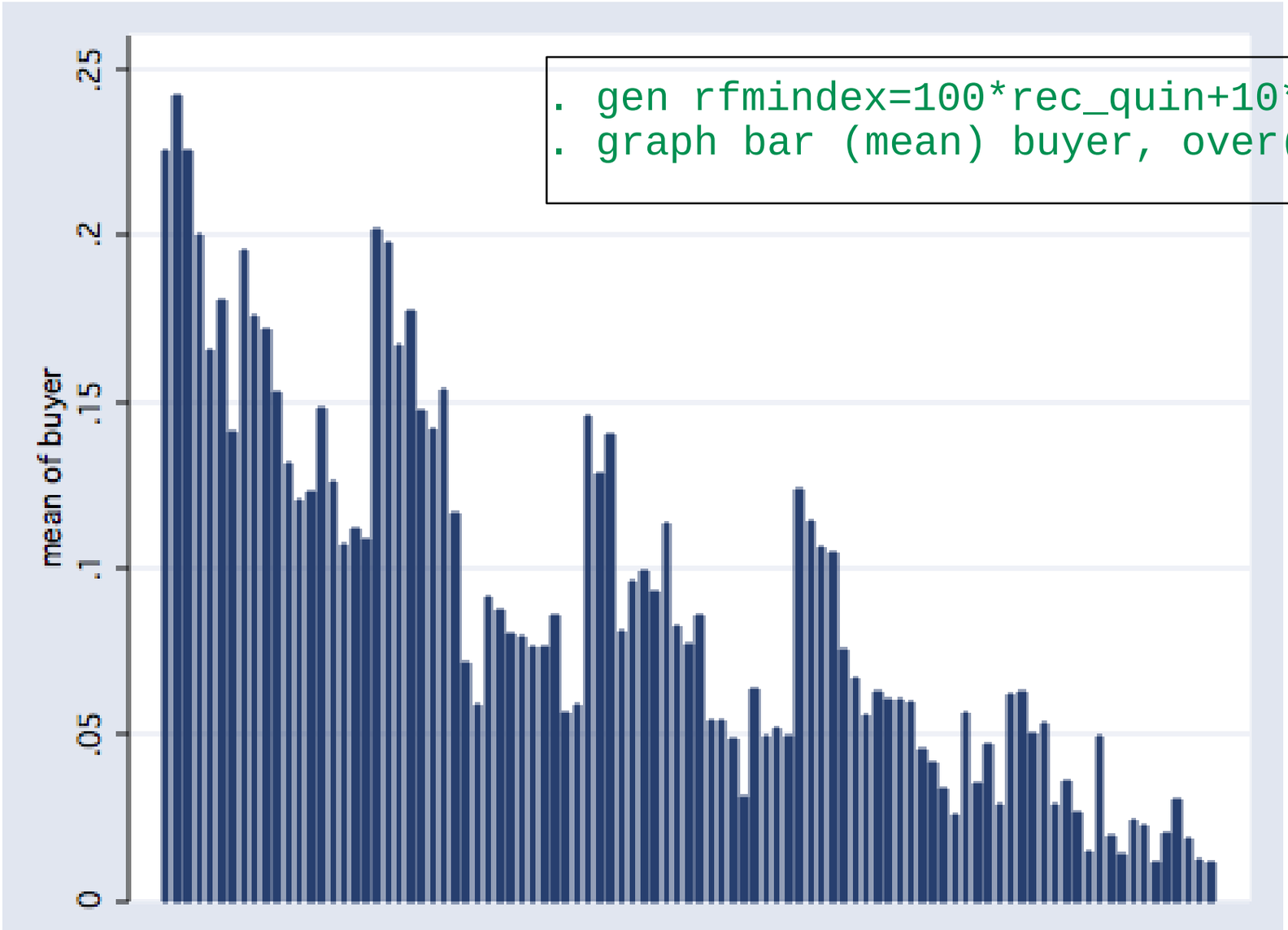
```
. pwcorr last purch total_, sig

             |     last    purch   total_
-------------+---------------------------
        last |   1.0000
             |
       purch |   0.0060   1.0000
             |   0.1791
             |
      total_ |  -0.0019   0.5153   1.0000
             |   0.6685   0.0000
```

- What does this say and mean?

The RFM-index is easy to calculate

RESPONSE RATE BY INDEPENDENT N-TILE RFM INDEX




```
. gen rfmindex=100*rec_quin+10*freq_quin+mon_quin
. graph bar (mean) buyer, over(rfmindex)
```

There are several ways of constructing an RFM index

TYPES OF RFM INDICES

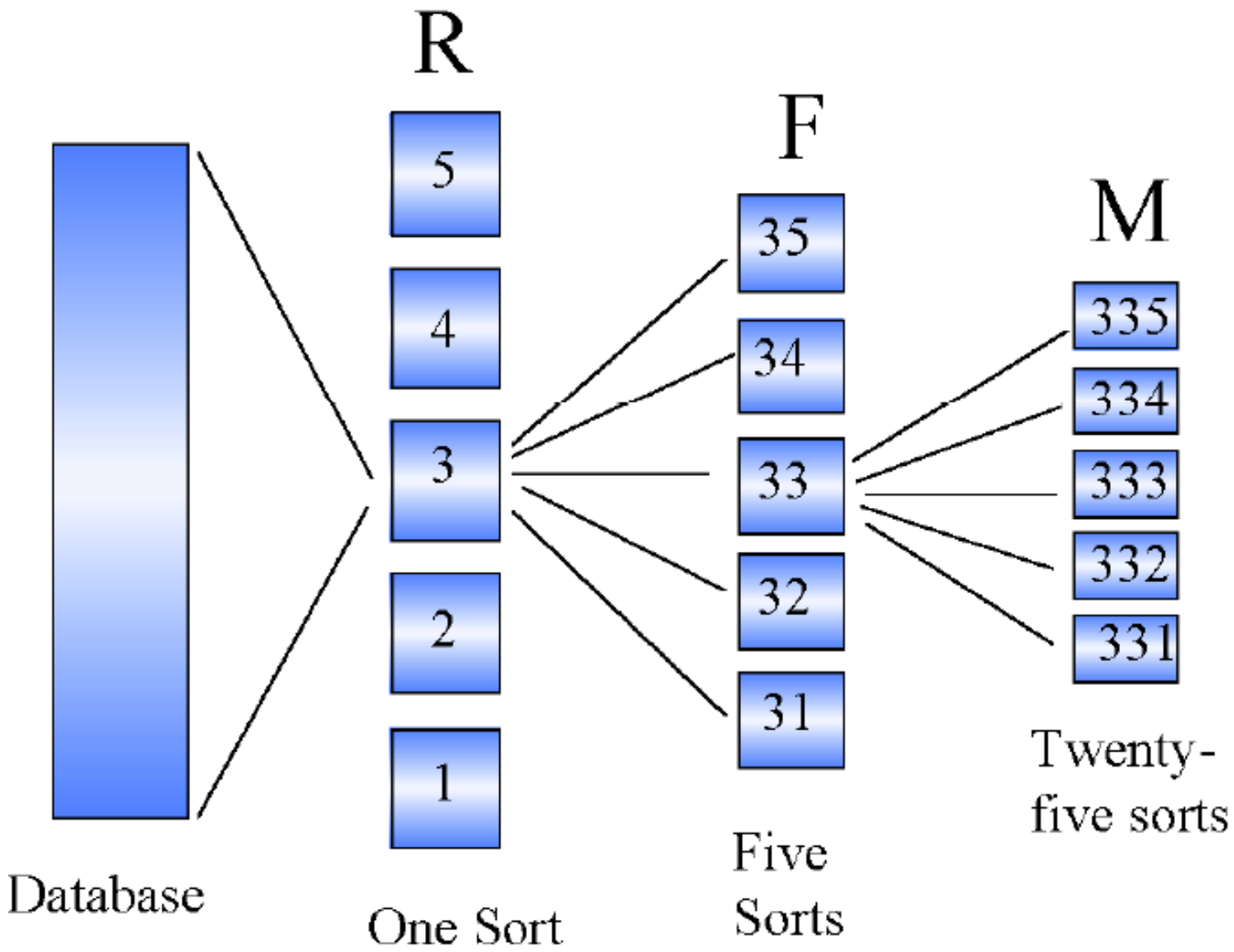
- **Independent N-tile approach (what we have done so far...)**
  - Create quintile for recency
  - Independently create quintile for frequency
  - Independently create quintile for monetary

- **Sequential N-tile approach**
  - Create quintile for recency
  - Within each recency quintile, create quintiles for frequency
  - Within each of 25 recency-frequency groups, create quintiles for monetary

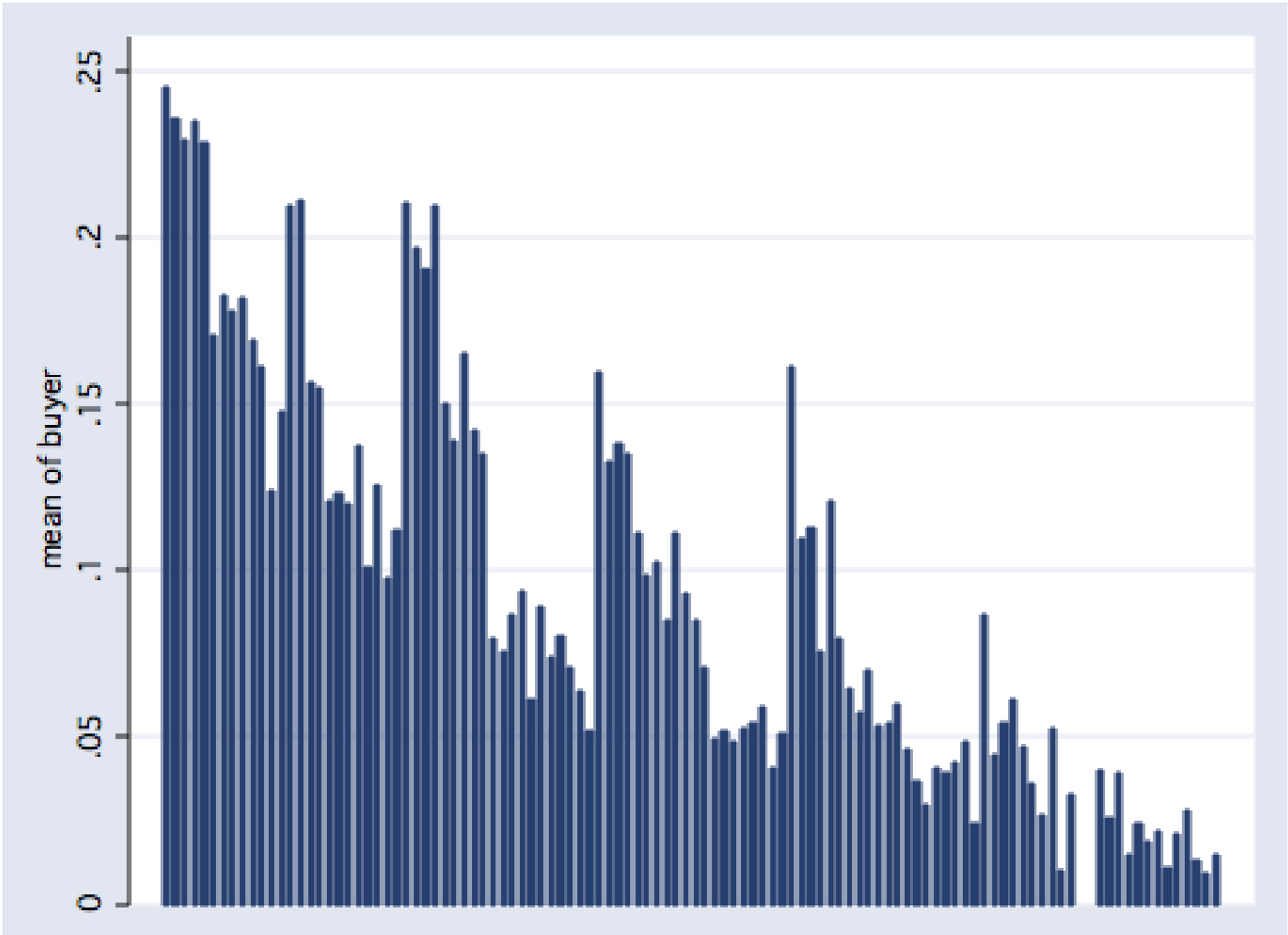
- **Intuitive groupings approach**
  - Pick intuitive cutoff points for R, F, and M
  - e.g. one-time buyers vs. repeat-buyers, less than 6 months, 6mo-1year, more than 1 year

The RFM index based on sequential N-tiles is substantially harder to calculate but is considered a better approach

CONSTRUCTION OF SEQUENTIAL N-TILES



RESPONSE RATE BY SEQUENTIAL N-TILE RFM INDEX



There are several ways of constructing an RFM index

TYPES OF RFM INDICES

- **Independent N-tile approach**
  - Create quintile for recency
  - Independently create quintile for frequency
  - Independently create quintile for monetary

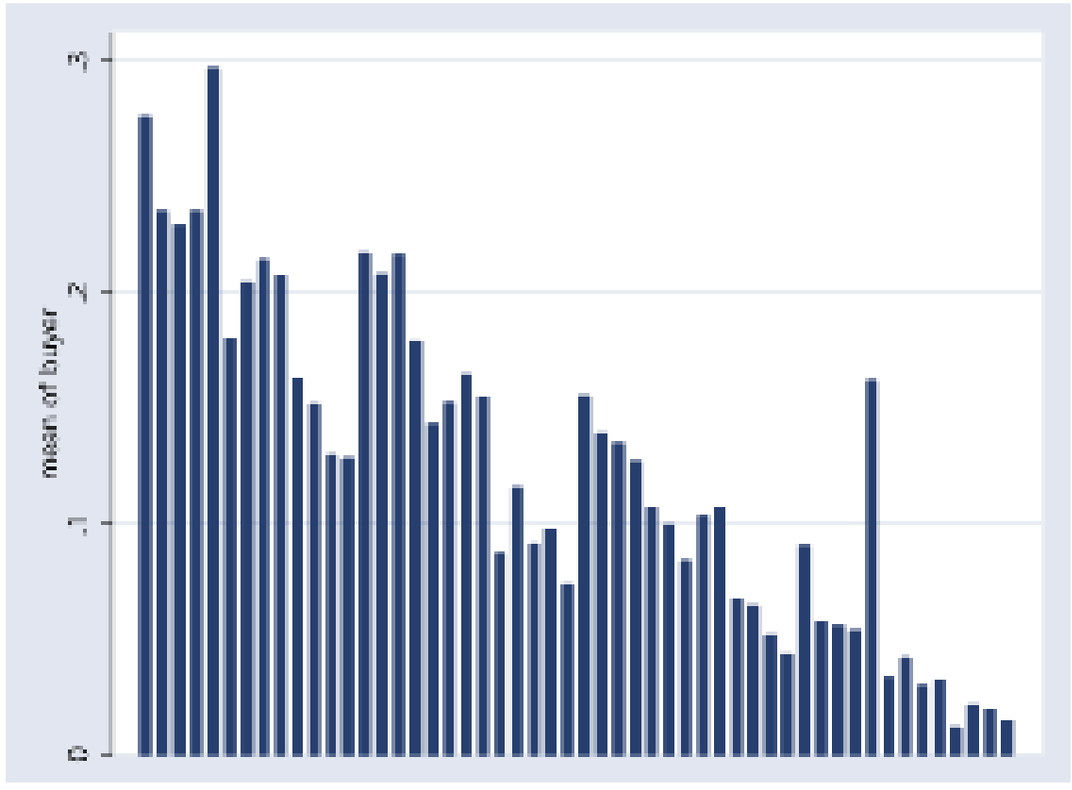
- **Sequential N-tile approach**
  - Create quintile for recency
  - Within each recency quintile, create quintiles for frequency
  - Within each of 25 recency-frequency groups, create quintiles for monetary

- **Intuitive groupings approach**
  - Pick intuitive cutoff points for R, F, and M
  - e.g. one-time buyers vs. repeat-buyers, less than 6 months, 6mo-1year, more than 1 year

The break-even response rate tells us to which cells to extend the offer

BREAK EVEN RESPONSE RATE

- Cost of mailing an offer = $0.50
- Selling price (includes shipping) = $18
- Wholesale price paid by Bookbinders = $9
- Shipping costs = $3
- Break-even =
  Cost to mail / net revenue per sale =
  .5 / (18-9-3) = **8.3%**



As a benchmark we calculate the profitability of mailing the full 500,000 consumers

PROFITABILITY (FULL SAMPLE)

- Mail to full sample: **500,000**
- Average response rate

```
sum buyer

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       buyer |     50000       .09044    .286814         0          1
```

Aver. response rate: 9.04%

Expected number of buyers:
9.04% * 500,000 = 45,200

- Profit = ($18 - 9 - 3)* **45,200** - 0.5***500,000** = **$21,200**
- Return on marketing expenditure = $21,200/$250,000 = **8.5%**

Using the RFM index we target fewer customers but with a higher response rate

PROFITABILITY (RFM-INDEX BASED ON INDEPENDENT N-TILE)

```
. rename rfmindex rfmindex_iq
. rfmpredict rfm_response_iq=buyer, indexvar(rfmindex_iq)

. generate mailto_iq=0
. replace  mailto_iq=1 if rfm_response_iq>0.083

. tabulate mailto_iq

  mailto_iq |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |     26,732       53.46       53.46
          1 |     23,268       46.54      100.00
------------+-----------------------------------
      Total |     50,000      100.00

. sum buyer if mailto_iq==1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       buyer |     23268    .1405364    .3475501         0          1
```

Mail to 46.54% of sample:
500,000 * 46.54%=232,700

Expected response rate: 14.05%
Expected number of buyers:
14.05% * 232,700 = 32,694

- Profit = ($18-9-3)* **32,694** - 0.5***232,700**
  = **$79,814**
- Return on marketing expenditure
  = $79,814/$116,350 = **68.6%**

“EGEN” EXAMPLE

```
. rfmpredict rfm_response_iq=buyer, indexvar(rfmindex_iq)

. generate mailto_iq=0
. replace  mailto_iq=1 if rfm_response_iq>0.083
```

acctnum	rfmindex_iq	buyer	rfm_response_iq	mailto_iq
10009	111	1	0.333	1
10034	111	0	0.333	1
10092	111	1	0.333	1
10185	111	0	0.333	1
10228	111	0	0.333	1
10236	111	0	0.333	1
10043	112	0	0	0
10293	112	0	0	0
10550	112	0	0	0
10618	112	0	0	0
10823	112	0	0	0
10875	112	0	0	0
10008	113	0	0.333	1
10055	113	1	0.333	1
10136	113	1	0.333	1
10152	113	0	0.333	1
10203	113	0	0.333	1
10269	113	0	0.333	1
10229	123	0	0.167	1
10381	123	1	0.167	1
10411	123	0	0.167	1
10452	123	0	0.167	1
10538	123	0	0.167	1
10598	123	0	0.167	1
10017	224	0	0	0
10066	224	0	0	0
10150	224	0	0	0
10453	224	0	0	0
10461	224	0	0	0
10502	224	0	0	0

PROFITABILITY (RFM-INDEX BASED ON SEQUENTIAL N-TILE)

```
. rfmpredict rfm_response_sq=buyer, indexvar(rfmindex_sq)

. generate mailto_sq=0
. replace mailto_sq=1 if rfm_response_sq>0.083

. tabulate mailto_sq

    mailto_sq |      Freq.     Percent        Cum.
--------------+-----------------------------------
            0 |     26,300       52.60       52.60
            1 |     23,700       47.40      100.00
--------------+-----------------------------------
        Total |     50,000      100.00

. sum buyer if mailto_sq==1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       buyer |     23700    .1396624    .3466438         0          1
```

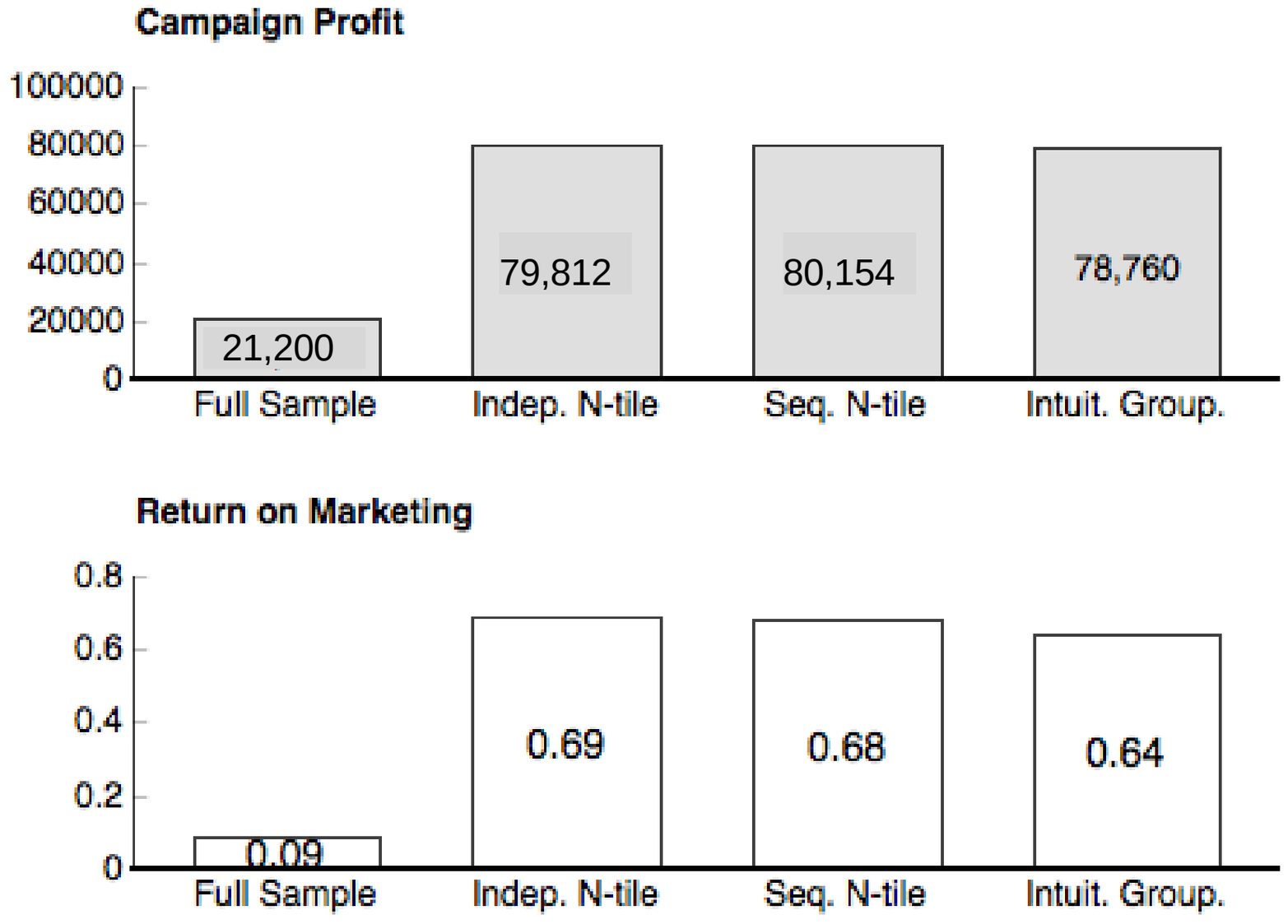
Mail to 47.4% of sample:
500,000 * 47.4%=237,000

Expected response rate: 13.97%
Expected number of buyers:
13.97% * 237,000 = **33,109**

- Profit = ($18 -9 -3)* **33,109** – 0.5***237,000**
  = **$ 80,154**
- Return on marketing expenditure
  = $80,154 / $118,500= **67.6%**

Using behavioral information with RFM analysis can dramatically improve the return on a marketing campaign

RFM IMPROVEMENTS



Marketing Research & Analytics Course Structure

Getting Ready for Marketing Research and Analytics

- Marketing Research and Analytics Overview (Class 1)
- How to Tell Good From Bad Data Analytics (Class 2)
- Using Stata for Marketing Research and Analytics (Classes 2 & 3)
- Statistics Review (Class 4)

Understanding Customers and Markets

- Quantifying Customer Value (Class 1)
- Case Analysis: "Home Alarm, Inc.: Assessing Customer Lifetime Value" and Testing (Class 3)
- Measuring Customers' Willingness to Pay (Class 6)
- Valuation of Products: Conjoint Analysis (Classes 8 & 9)
- Market Segmentation: Cluster Analysis (Class 10)
- Survey, and Qualitative Research (Class 10)

Prospecting and Targeting the Right Customers

- Predicting Response with RFM analysis (Class 5)
- Case Analysis: "Tuango: RFM Analysis for Mobile App Push Messaging"; Lift and Gains (Class 6)
- Predicting Response with Logistic Regression (Class 7)
- Case Analysis: "BookBinders: Predicting Response with Logistic Regression" (Class 8)
- Predicting Response with Neural Networks (Class 9)
- Predicting Response with Decision Trees (Class 10)

Developing Customers

- Case Analysis: "Intuit: Quickbooks Upgrade" (Class 11)
- Next-Product-To-Buy Models: Learning From Purchases (Class 11)
- Recommendation Systems: Learning From Ratings (Class 12)

Retaining Customers

- Predicting Attrition (Class 12)

Selecting the Right Offers

- Design of Experiments / Multi Variable Testing (Class 13)
- Case Analysis: "Capital One: Information-Based Credit Card Design" (Class 14)

Limitations of Marketing Analytics

- When Marketing Analytics, CRM, and Databases Fail (Class 14)

Wrap-up

- Wrap-Up (Class 14)

Next Time: Tuango RFM Case

- Use Bookbinders RFM "do-file" "RFM_BBB_stata.do" as the basis for this assignment
- It will not run as is -- you need to modify the analysis to fit the Tuango case