Doing Customer Intelligence with R

By Jim Porzak
Director of Analytics
Loyalty Matrix, Inc.

October 19, 2004

Presented at the SDForum Business Intelligence SIG
Palo Alto, CA

Loyalty Matrix, Inc., 580 Market St., Suite 600, San Francisco, CA 94104 (415) 296-1141
www.LoyaltyMatrix.com R.LoyaltyMatrix.com

- Introduction to Customer Intelligence at Loyalty Matrix
- Introduction to R
- R for Exploratory Data Analysis (EDA)
- R for Statistics & Data Mining
- Summary and Q&A

What is Customer Intelligence (CI)?

- Twenty years ago a mentor in the Valley told me: “Jim, take care of your customers. They will take care of you.”
  - The fundamentals have not changed
  - But the data, tools and techniques have become *much* richer
- Definition: “Doing Customer Intelligence is analyzing customer behavioral and motivational data to build actionable business insights that can change marketing strategy and tactics to better align the organization with it’s customers needs and expectations.”
- Customer Intelligence is multifaceted
  - It is quantitative – insights must be based on facts
  - It is perceptive – insights are business interpretation of the numbers
  - It is actionable – insights must usable for the business
  - It is iterative – application generates more data & more insights

Loyalty Matrix, Inc.

- A privately held marketing services company based on technology
- A combination of seasoned marketers, techies and proprietary technology providing our clients with unparalleled customer intelligence solutions focused on:
  - Customer Acquisition
  - Customer Retention
  - Product Cross-sell and Up-sell
- Founded in 2001
- A San Francisco based firm with offices in Dallas, Chicago, and (soon) New York
- Key Goal: ***Transform Customer Data into Actionable Insights!***

Customer Intelligence (CI) in Action

- Case study: Apple .MAC Subscribers
  - Business Challenge: Maximize renewal rates
  - CI Insight: Early usage drives loyalty
  - Business Impact: Shift marketing effort to drive new subscriber usage (depth then breadth)

- Case study: St. Regis Hotel Guests
  - Business Challenge: Reduce high customer attrition
  - CI Insight: Decline in loyal guests masked by boom economy single visit guests.
  - Business Impact: Focus on regaining former loyal guests

ST. REGIS
Hotels & Resorts

The MatrixOptimizer®: Data → Insights → Action

The data the describes the relationship

Customer

- Transactional
- Web site usage data
- Survey data
- Research data
- 3rd party data

Company

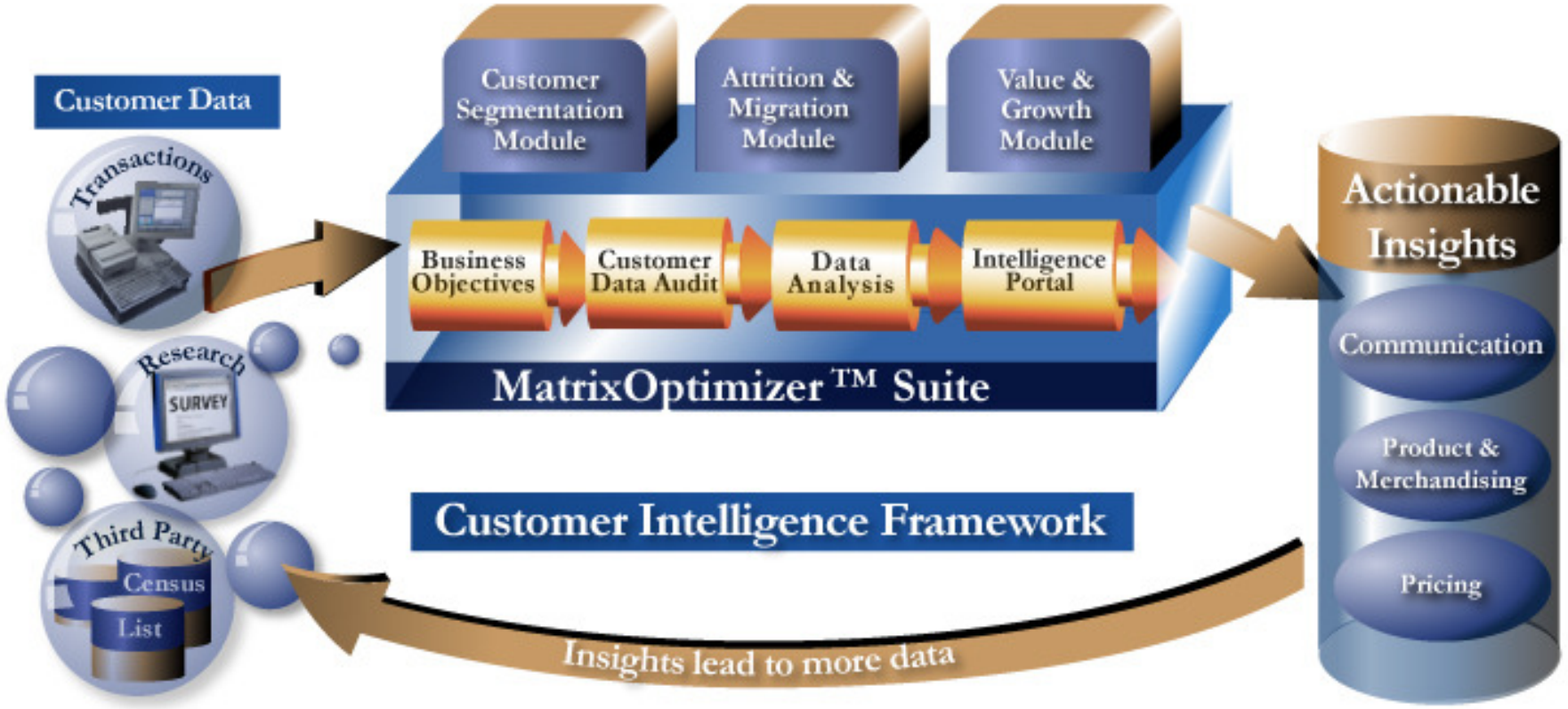
MatrixOptimizer®

The results

- Communication
- Pricing
- Product
- Distribution
- Data Strategies & Collection
- Loyalty Program Design & Evaluation

Loyalty Matrix

The CI Framework at Loyalty Matrix



Inside the CI solution, the MatrixOptimizer®

- Data Mart in Microsoft SQL 2000
  - Holds “ready to report data” in dimensional schema.
  - Major effort to load client data, quality checks, cleanse
  - Client data staged in SQL
- OLAP cubes built with Microsoft Analysis Services
  - Built off of Data Mart
  - Each cube focuses on single set of business problems
- OLAP presentation layer built with eBlocks
  - Web access to standard reports
  - Allows slice and dice
- Version 3.0 released October, 2004

How can R help build Customer Intelligence?

Challenges with Classical CI

- Aggregations limited to counts, sums & means.
- Nature of customer data
  - Events with related # and/or $ values
  - #'s, $'s & intervals all highly right skewed distributions
  - Data quality often suspect – especially in dimensions (factors)
- No rigorous tests
- No advanced methods

R to the Rescue!

- Visualization
  - Take first look at raw data
  - EDA on "clean" data
- Classical stats
  - Differences really significant?
  - Efficacy of marketing efforts?
- Prediction & Modeling
  - Classification
  - Meaningful customer attributes

- Introduction to Customer Intelligence at Loyalty Matrix
- Introduction to R
- R for Exploratory Data Analysis (EDA)
- R for Statistics & Data Mining
- Summary and Q&A

Evolution of R from S

- R is the free (GNU), open source, version of S
  - S developed by John Chambers and colleagues while at Bell Labs in 80's
  - For "data analysis and graphics"
  - Version 4 defined by the "Green Book" *Programming with Data*, 1998
  - S-Plus now owned and developed by Insightful Corp., Seattle, WA
- R was initially written by Robert Gentleman and Ross Ihaka
  - In early 1990's
  - Statistics Department of the University of Auckland
  - GNU GPL release in 1995
- Since 1997 a core group of 17 developers has had write access to the source
  - V1.0 released in February, 2000
  - New 0.1 level release ~ 6 months

Current state of R

- V2.0 Released October, 2004
- Windows, Mac OS, Linux & Unix ports
- Over 400 submitted packages from “abind” to “zoo”
- 12th newsletter (Volume 4/2) published September 2004
- The first useR! – R User Conference held in Vienna May 2004
- ~400 R-help newsgroup messages per week
- ~ Dozen texts specifically on R or with R examples and code
- R language generally accepted to be more powerful than S-Plus
- Some interesting GUI work in progress

R Resources

- R Homepage: http://www.r-project.org/
  - The official site of R
- R Foundation: http://www.r-project.org/foundation/
  - Central reference point for R development community
  - Holds copyright of R software and documentation
  - Support it!
- Local CRAN:
  - Mirror site
  - Current Binaries
  - Current Documentation
  - Link to related projects and sites
- R.LoyaltyMatrix.com blog

- Introduction to Customer Intelligence at Loyalty Matrix
- Introduction to R
- **R for Exploratory Data Analysis (EDA)**
- R for Statistics & Data Mining
- Summary and Q&A

Visualization is Key to EDA

Getting information from a table is like extracting sunlight from a cucumber.

– Farquhar & Farquhar, 1891.

If I can't picture it, I can't understand it.

– Albert Einstein

You can see a lot, just by looking.

– Yogi Berra

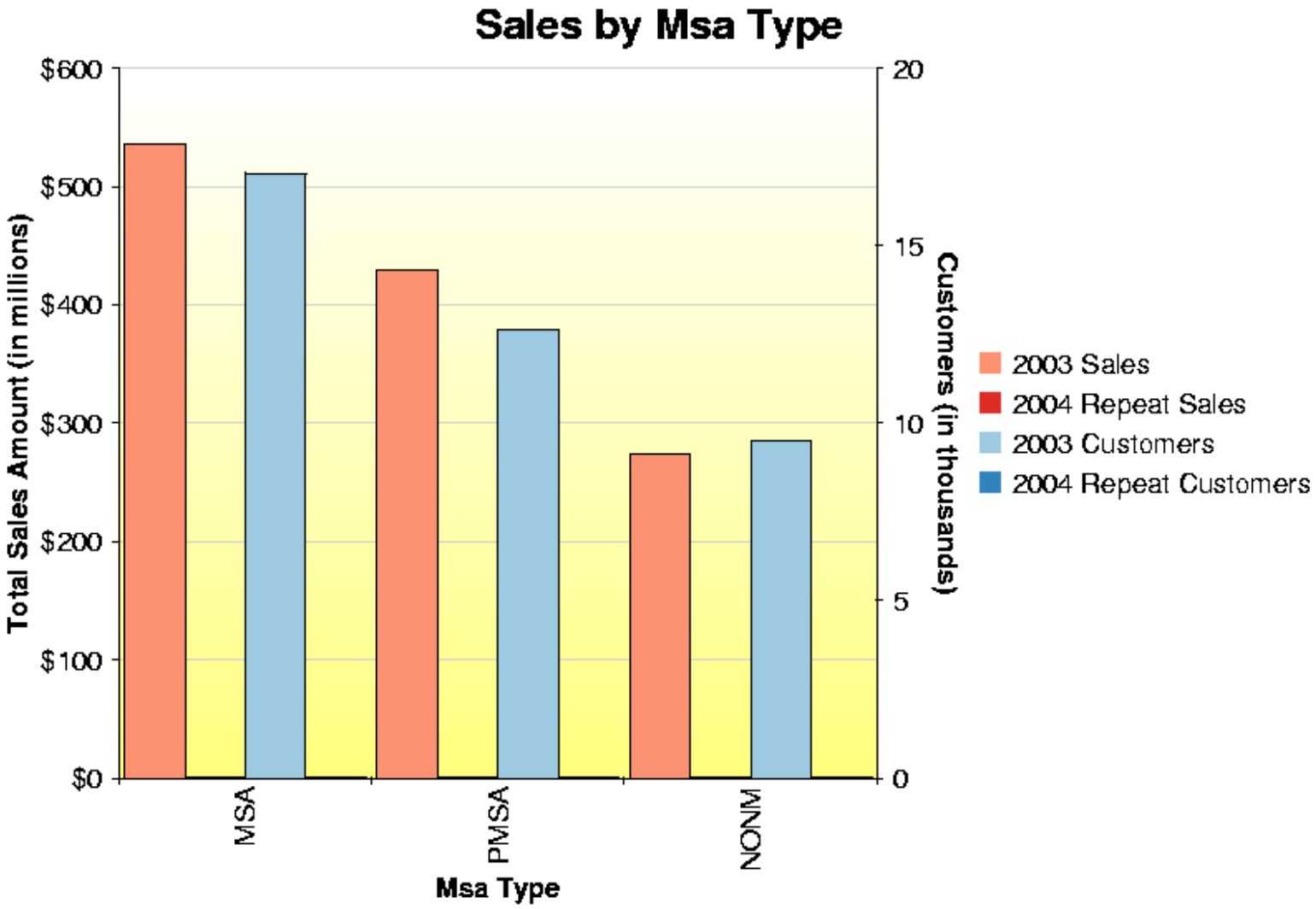
Thanks to Michael Friendly of VCD fame.

Our Favorite Visualization Methods

- For first look and exploration
  - Frank Harrell's datadensity for quick look, quality checks, …
  - Scatter Plots for patterns, outliers, …
  - Box Plots for median, range, outliers
- To understand customer behavior
  - Interval Histograms for time between visit, purchase, …
  - Distance Histograms for customer to store travel
  - Geographical Maps for customer to store travel
- Exploration for correlations and associations
  - Scatterplot Matrices
  - Mosaic Plots for categorical associations

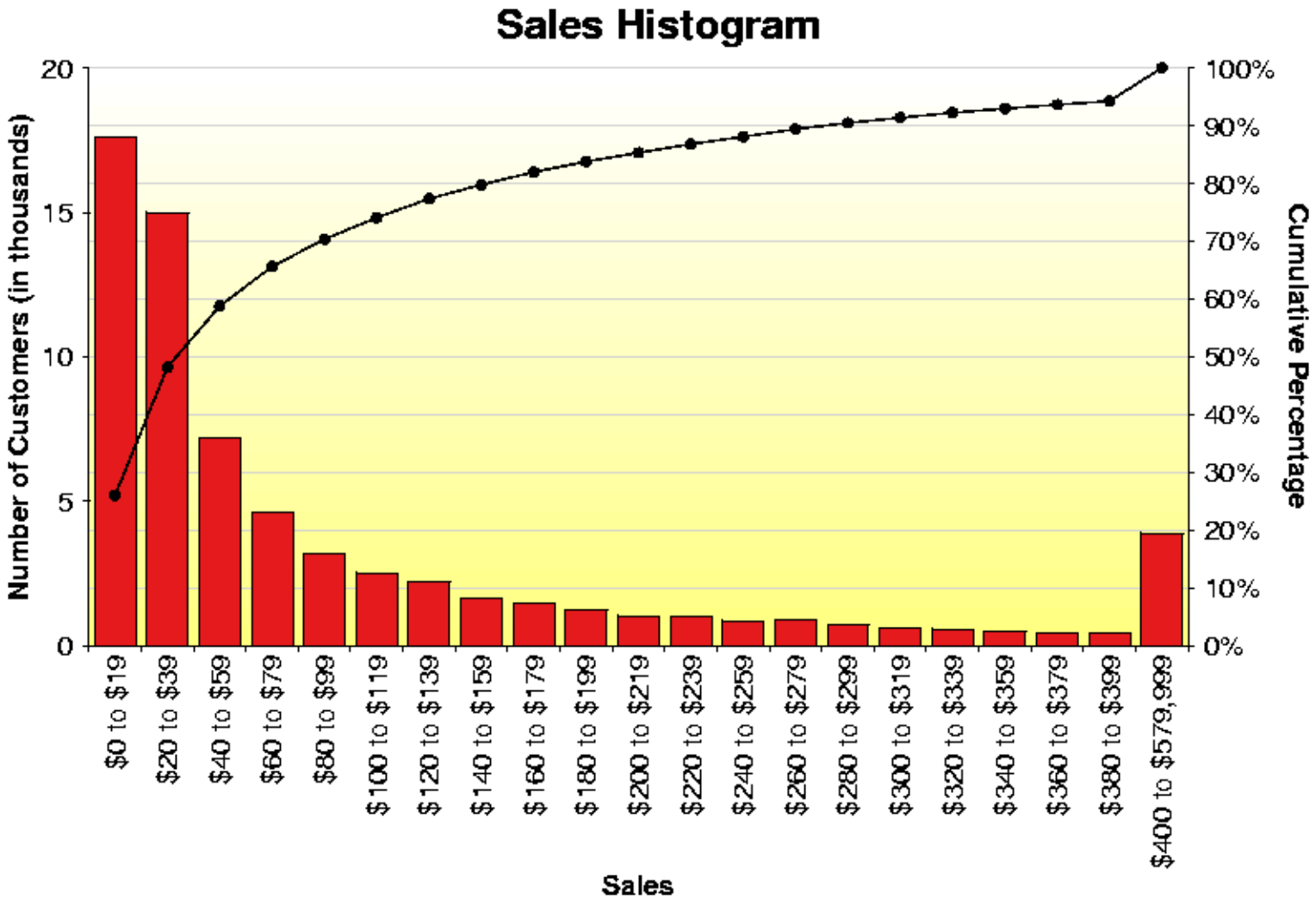
Basic Exploratory Data Analysis (EDA)

Everything We Know About the Fact Table – Example 1



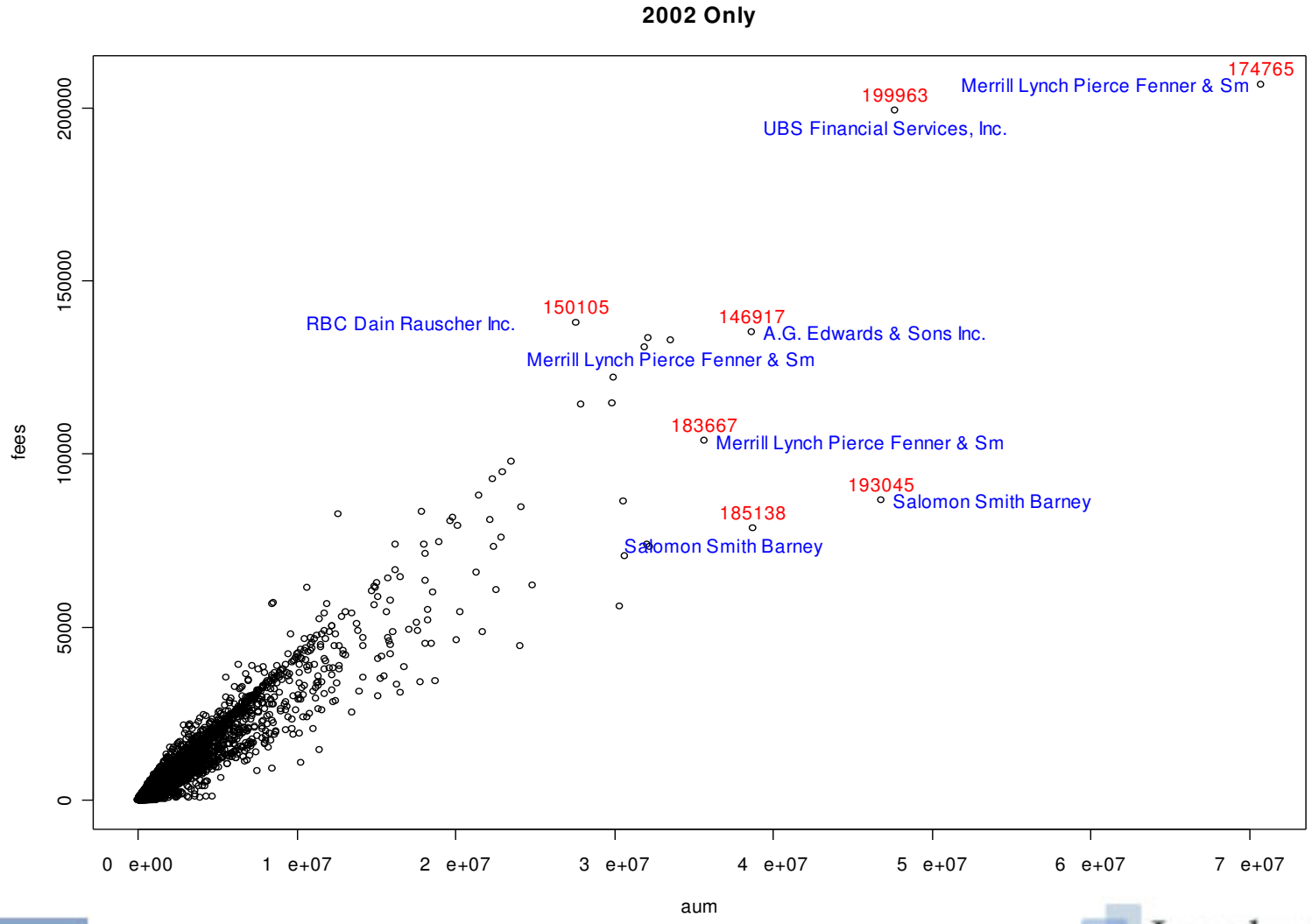
Loyalty Matrix

Everything We Know About the Fact Table – Example 2



Loyalty Matrix

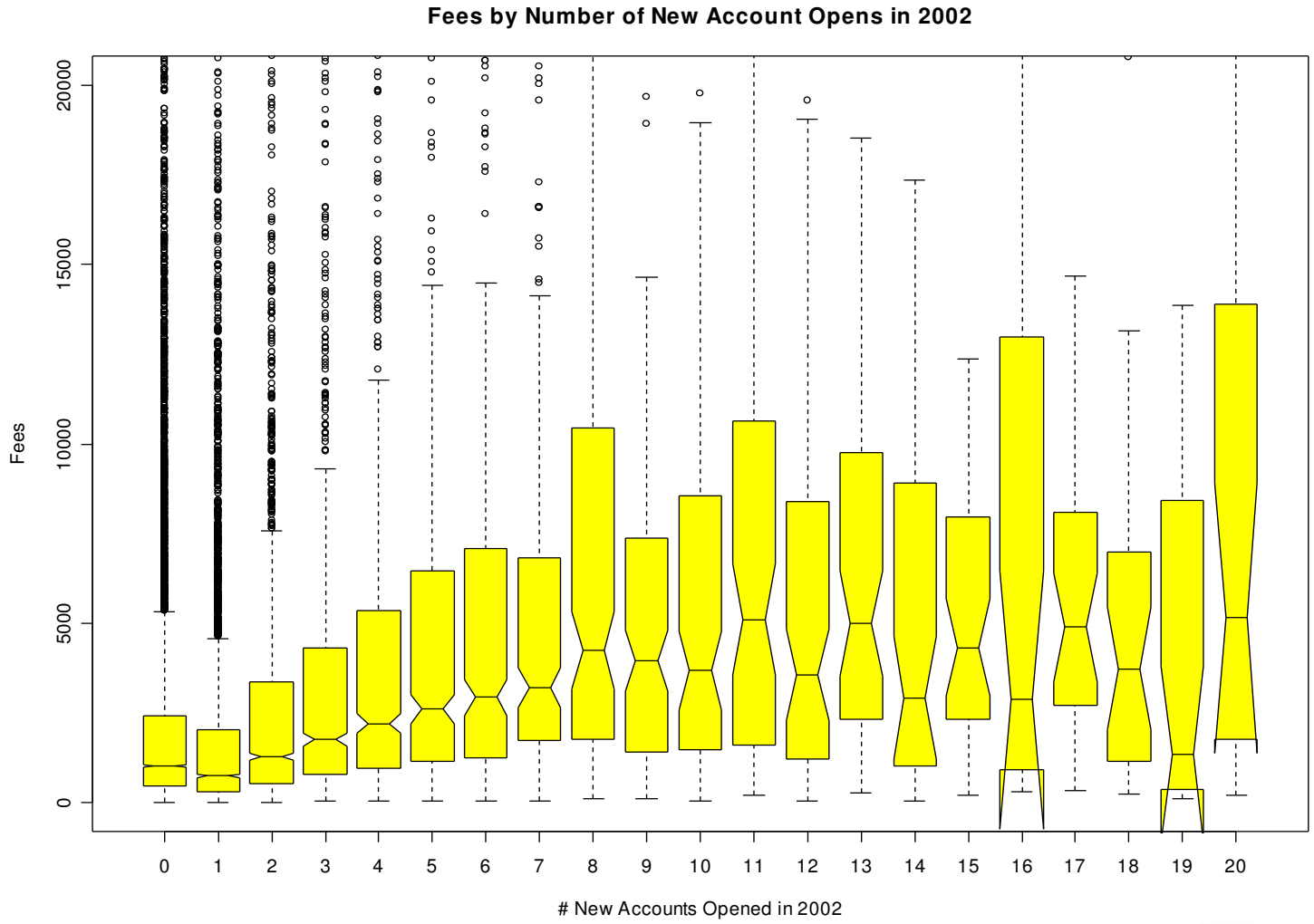
Scatter Plot Shows Correlation & Outlier Suspects.



Loyalty Matrix

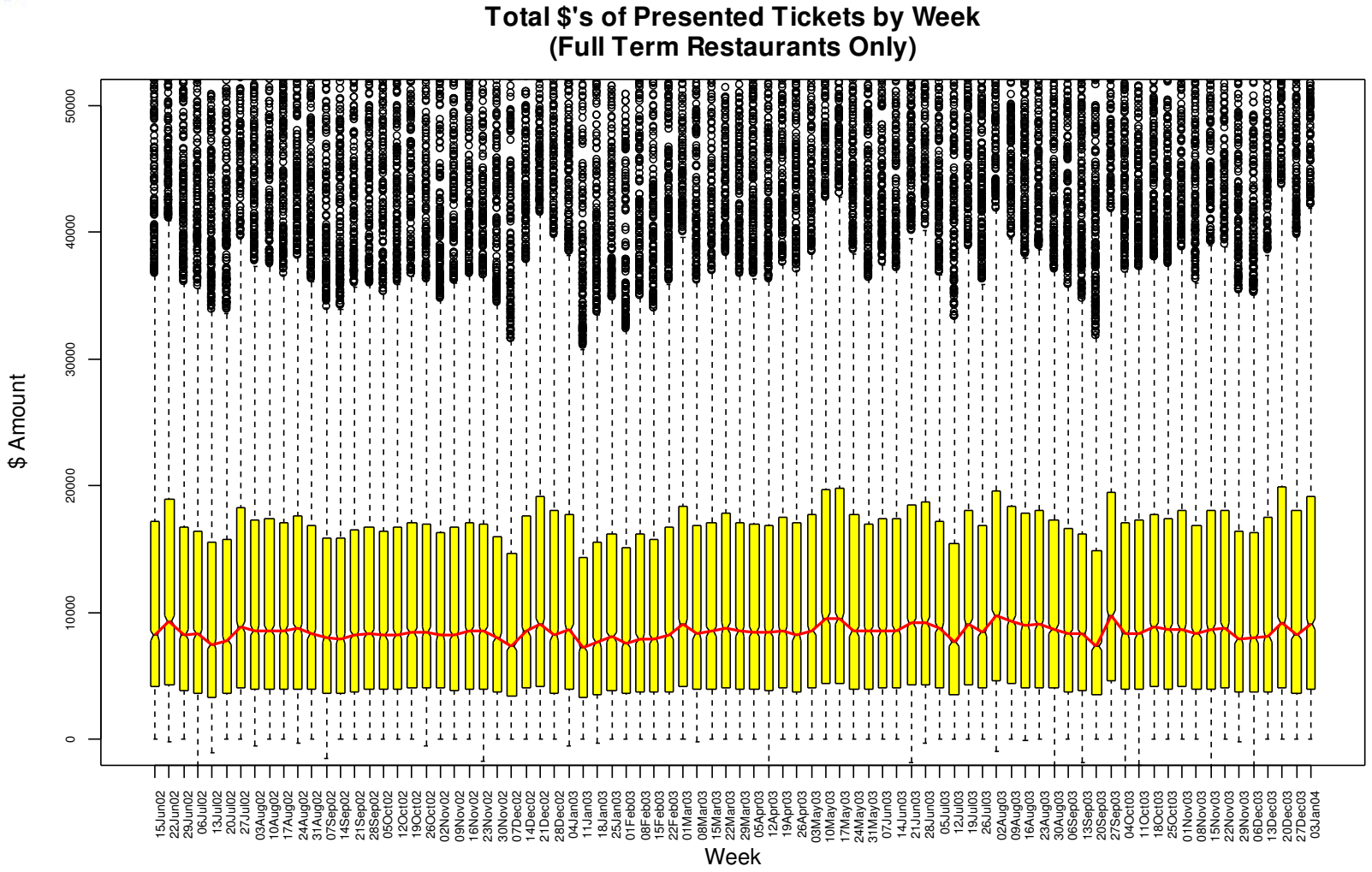
Box Plots for Skewed Data like $’s & #’s

Boxplots for Strongly Skewed $ Fees by #'s of Accounts



Loyalty Matrix

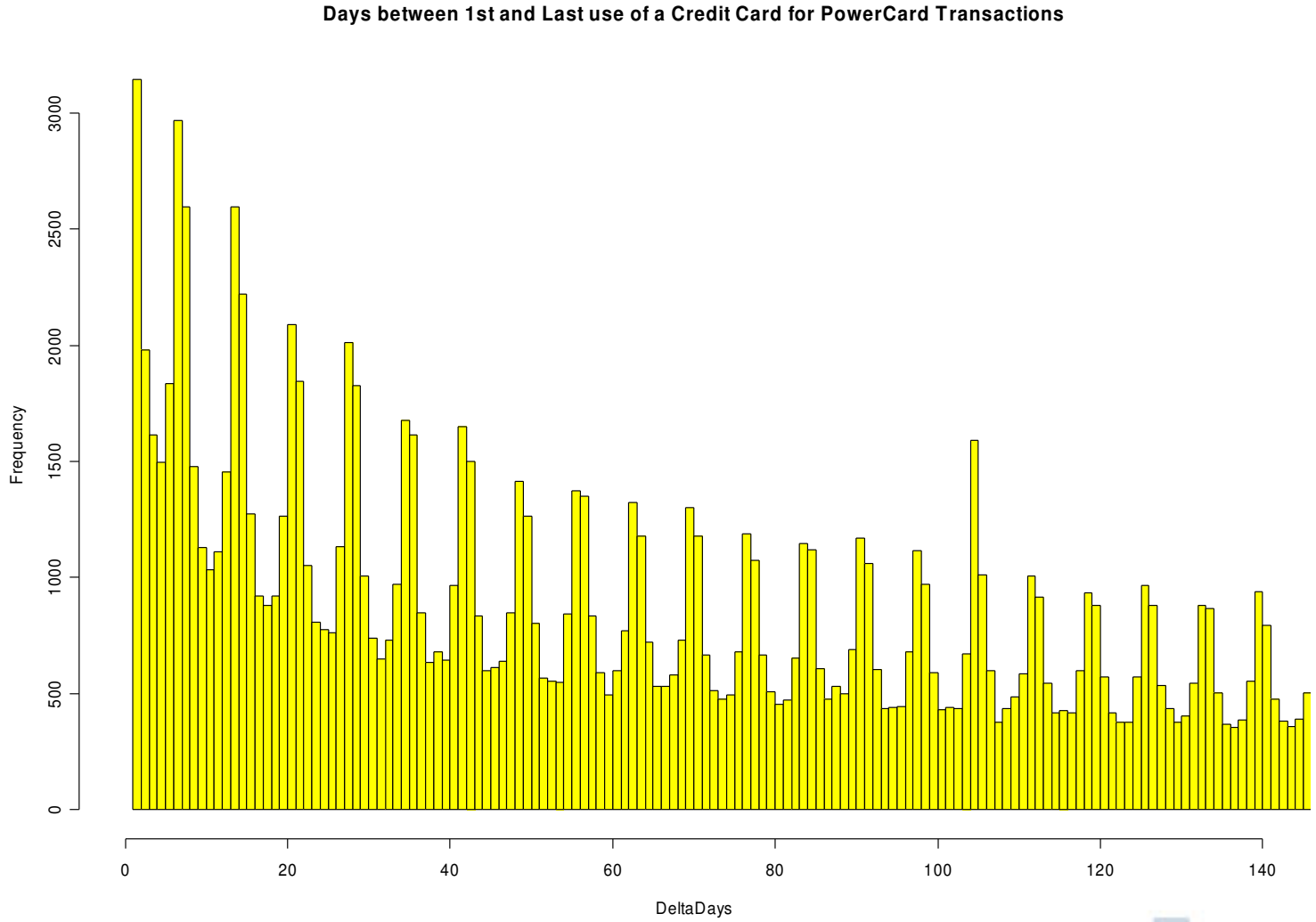
Boxplots for Restaurant Group $ Sales over Time



Loyalty Matrix

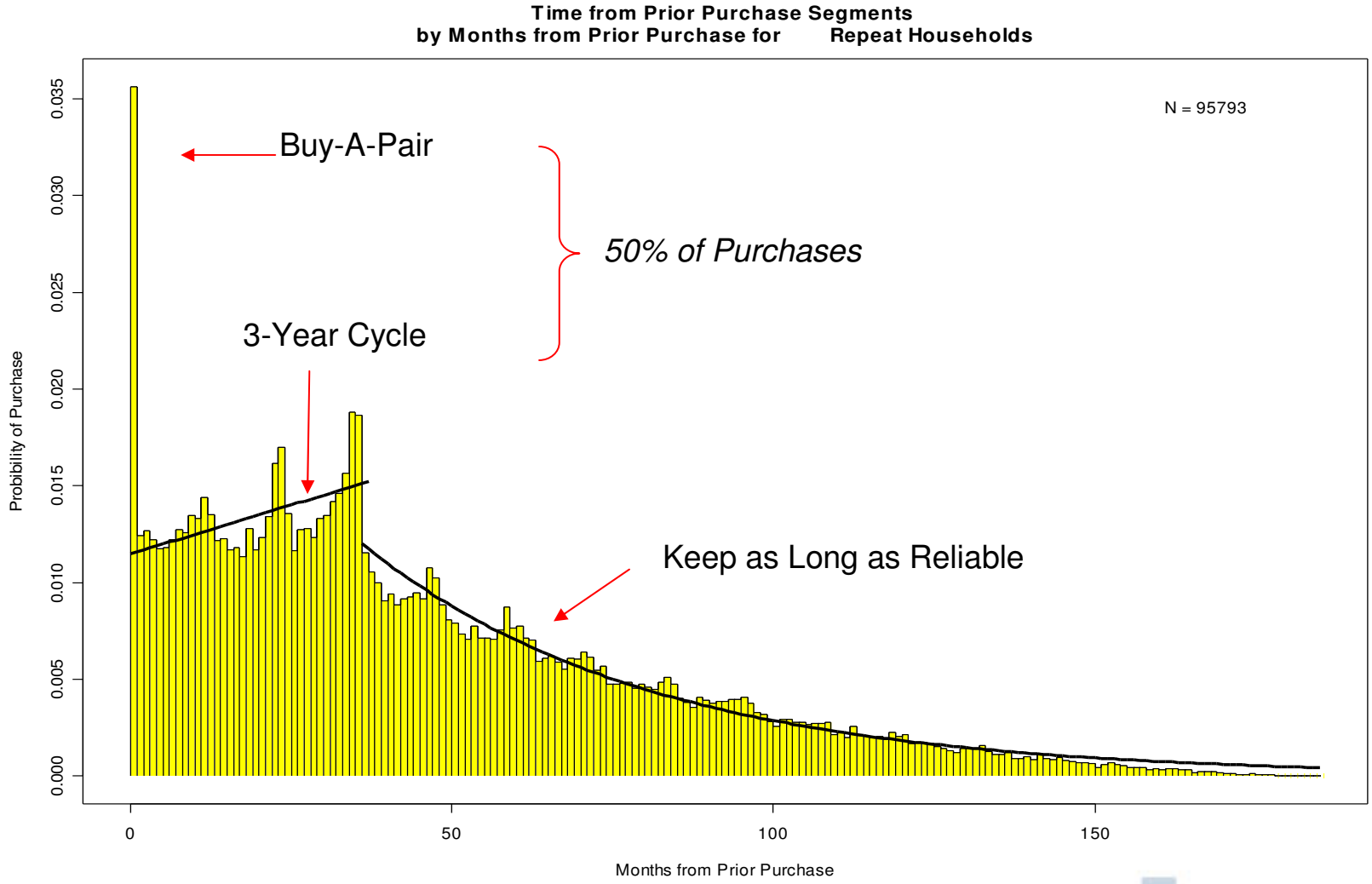
Customer Purchase Intervals

Visit Interval Histogram for Restaurant/Bar/Arcade



Loyalty Matrix

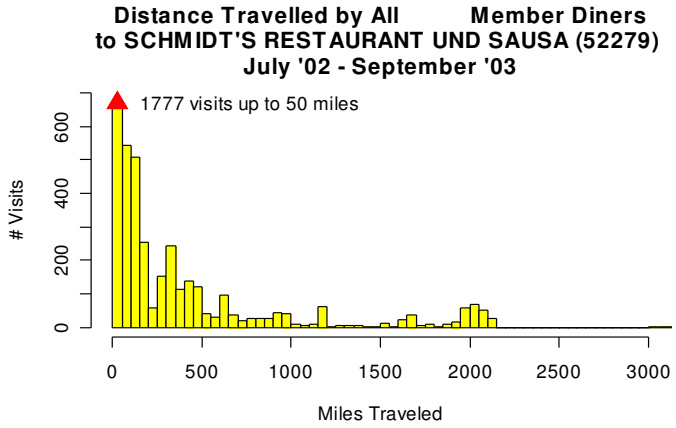
Purchase Interval for Vehicles Shows Key Customer Types

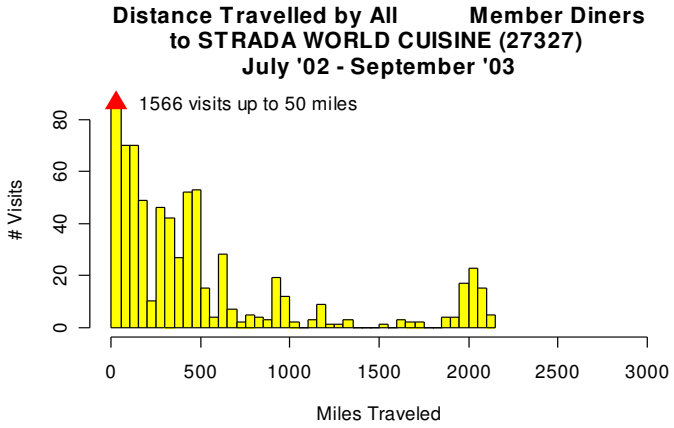


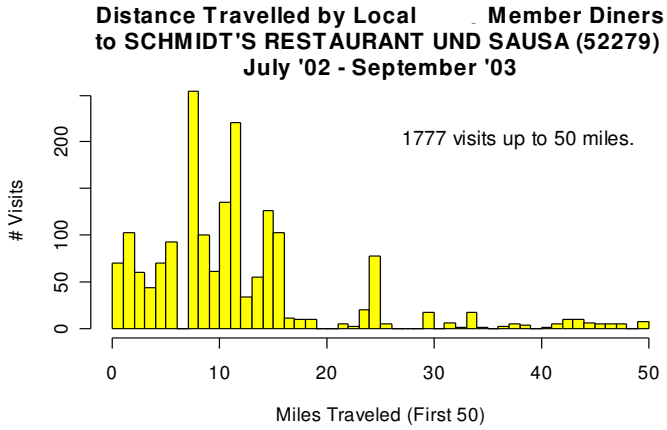
Loyalty Matrix

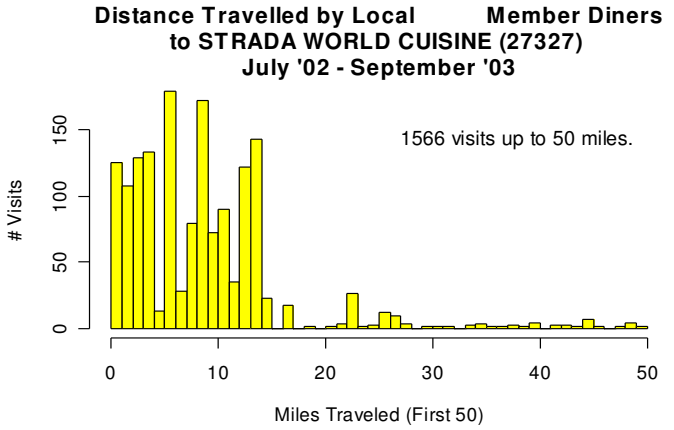
Distance Traveled and Geography

Distance Traveled by Customer to Restaurant





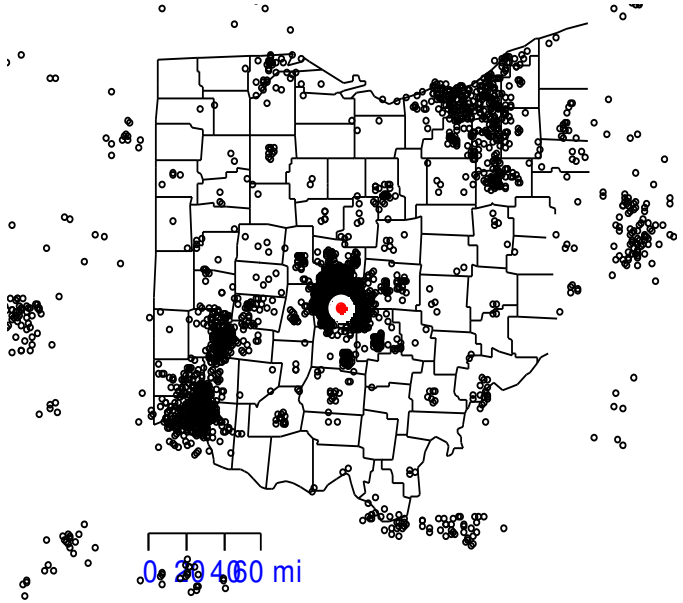




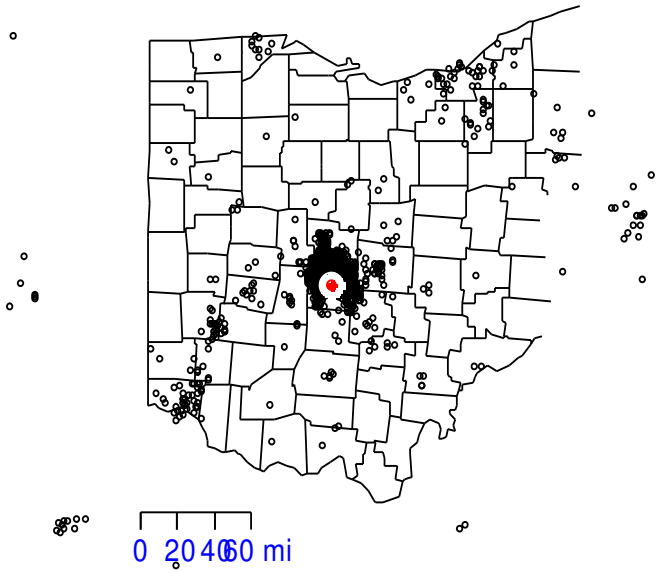
Loyalty Matrix

Map Home ZIP's for Customers Dining at Restaurants

SCHMIDT'S RESTAURANT UND SAUSA (52279)
Visits in Last 18 Months

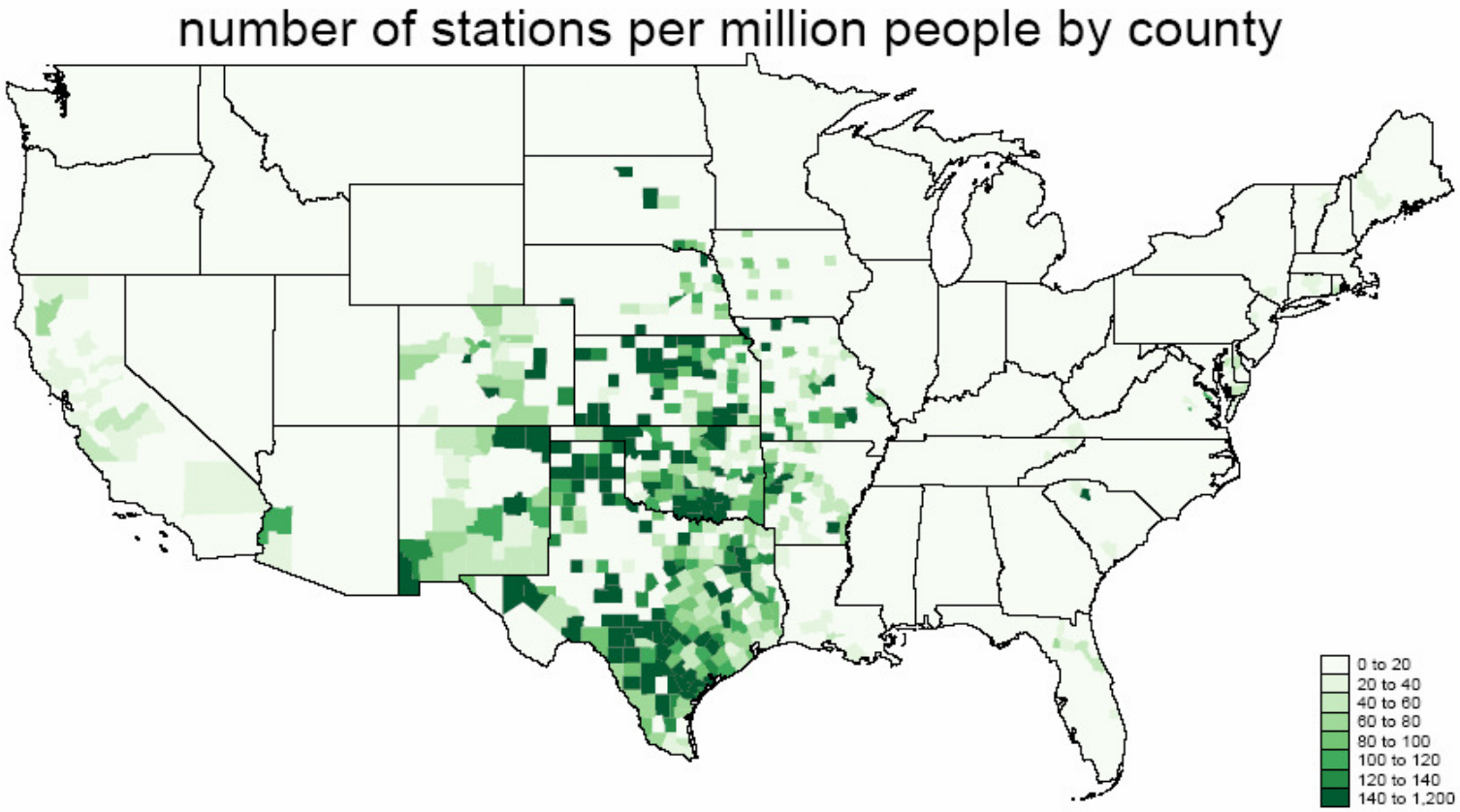


STRADA WORLD CUISINE (27327)
Visits in Last 18 Months



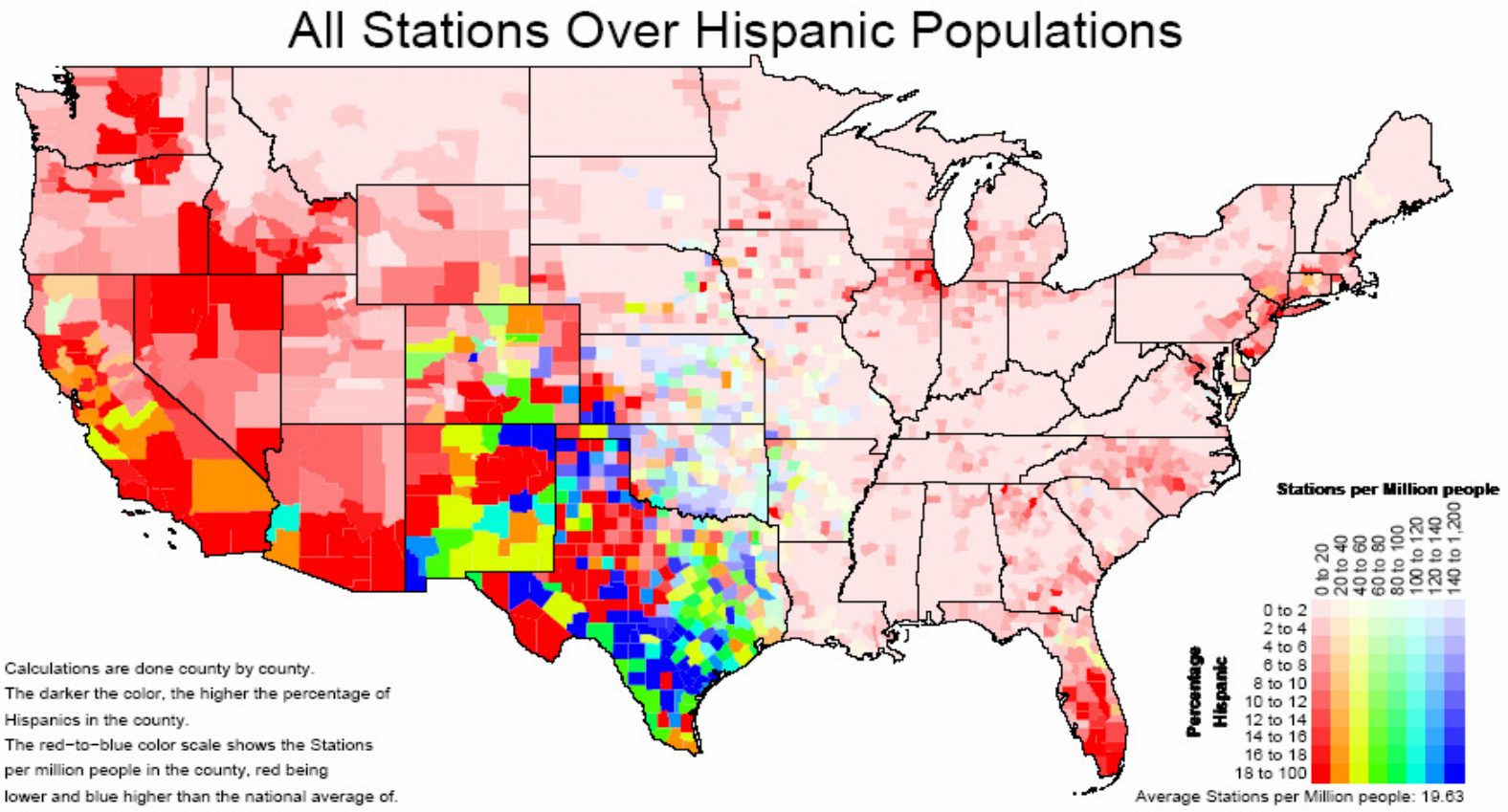
Loyalty Matrix

Geographic Density of Outlets Relative to Population



Loyalty Matrix

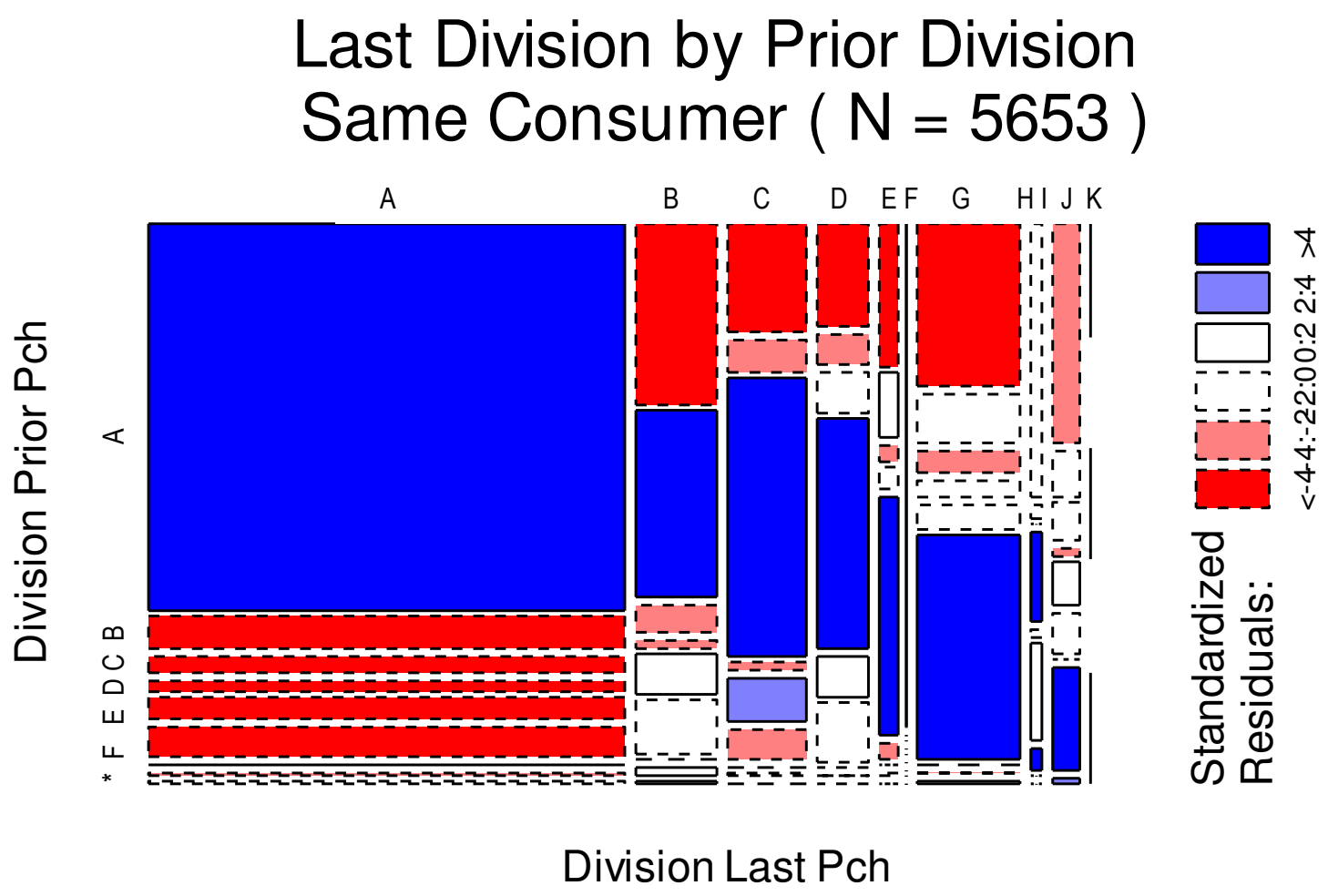
Geographic Density with Ethnic Targeted Coverage



Loyalty Matrix

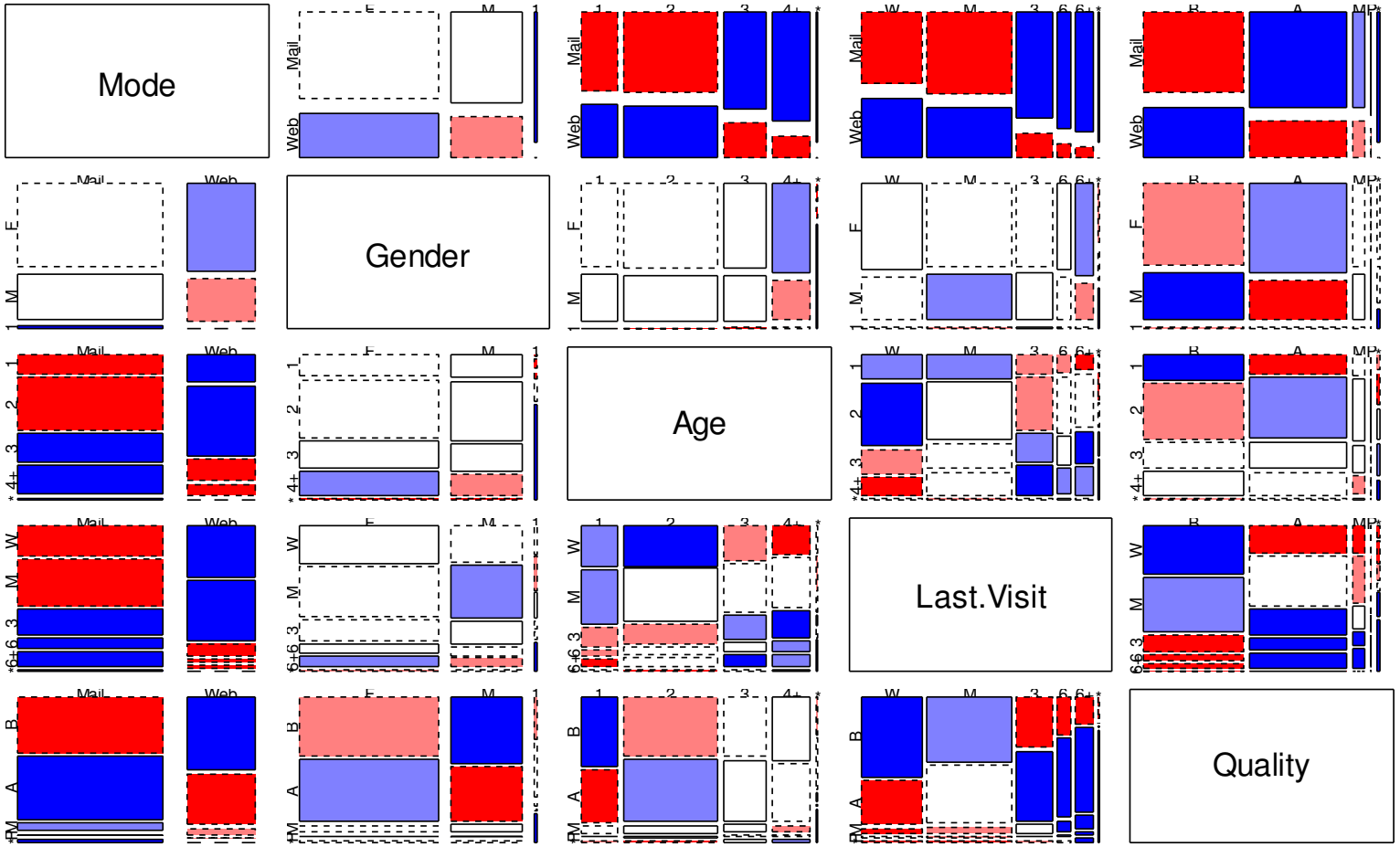
Understanding Categorical Variables

Mosaic Plots Show Automotive Brand Loyalty



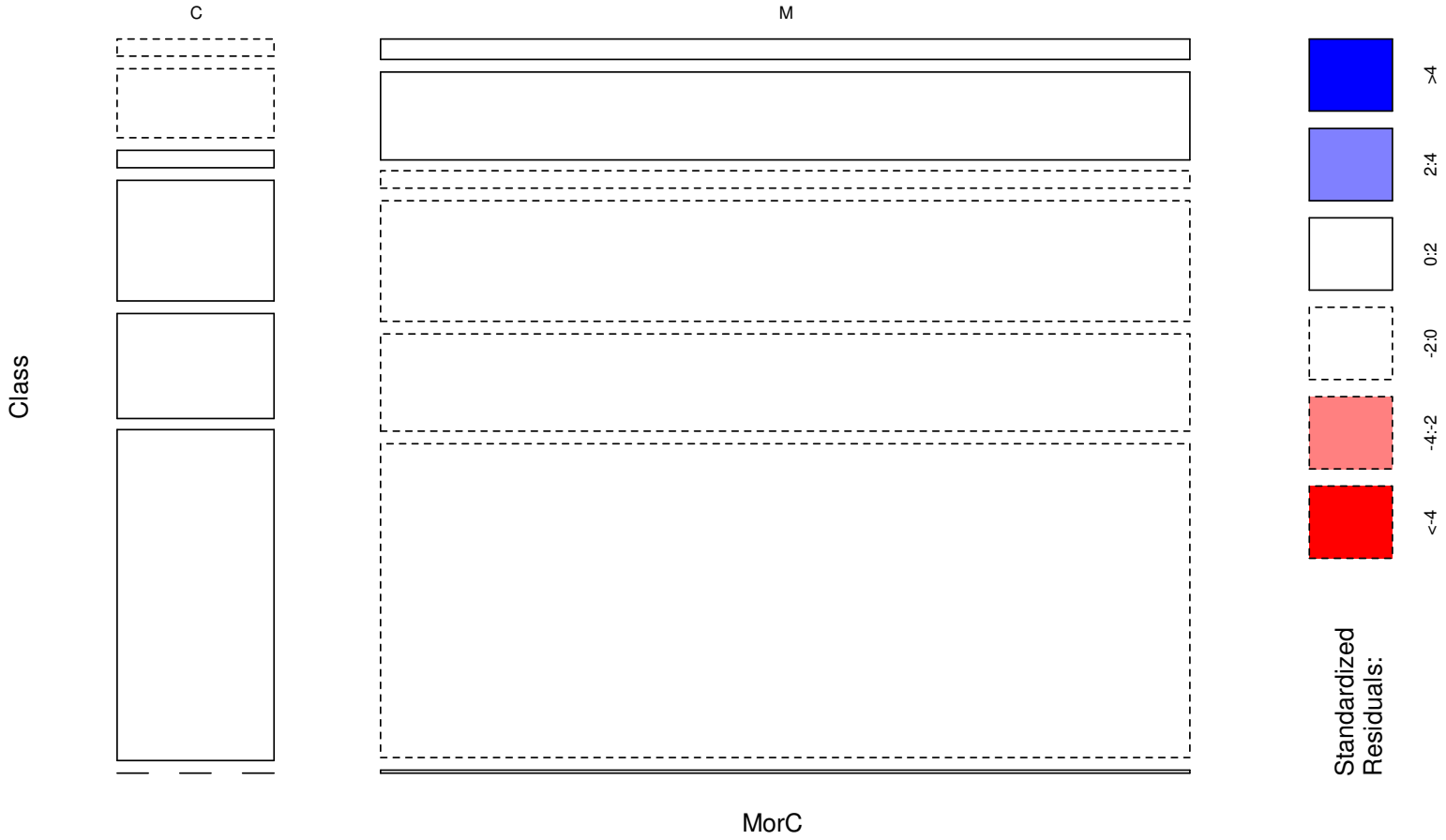
Loyalty Matrix

Mosaic Pairs Shows Categorical Responses in Survey



Loyalty Matrix

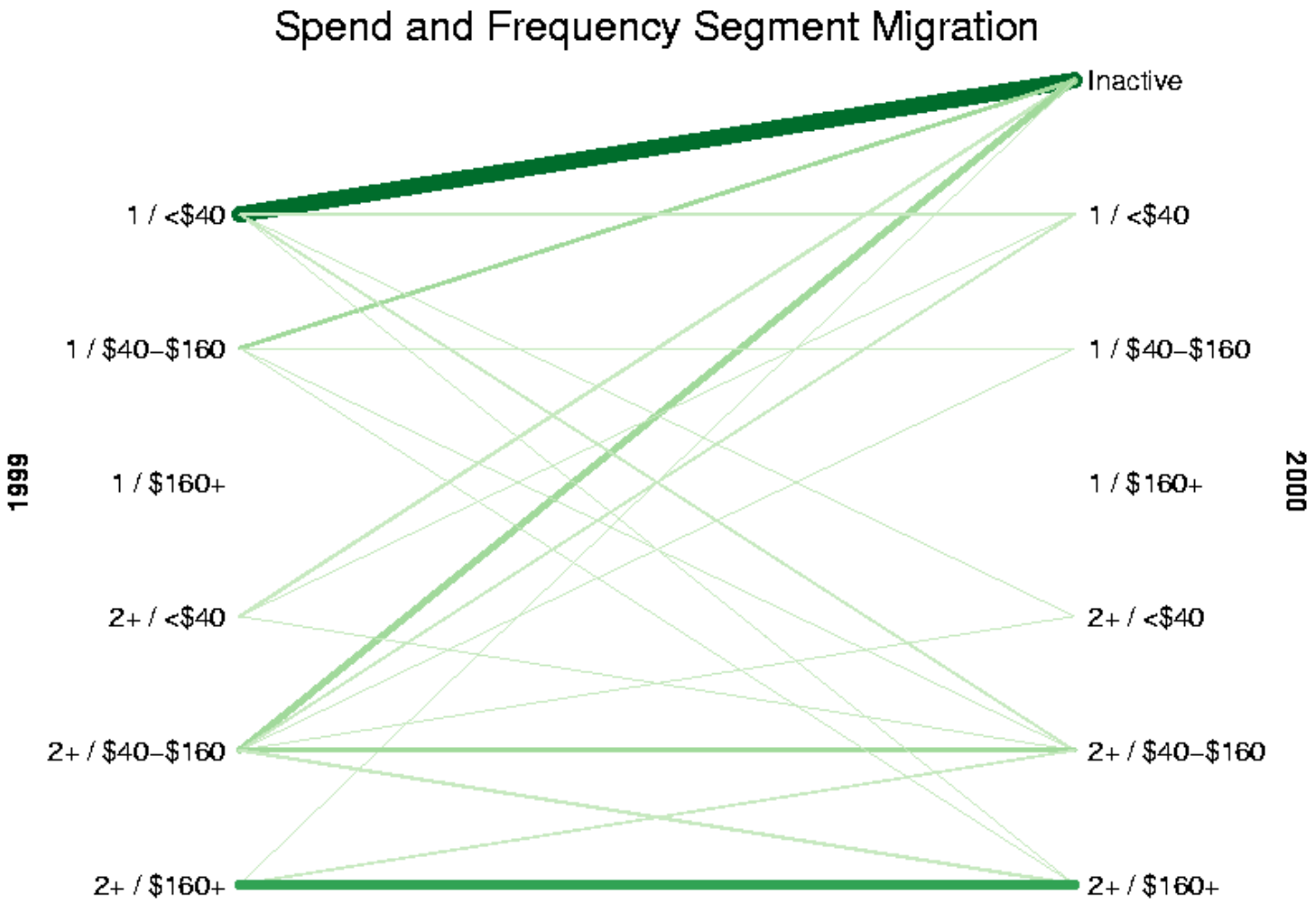
Mosaic Shows Mailed vs. Control Purchases are Alike



Loyalty Matrix

Customer Segment Migration

Customer Segment Migration



Loyalty Matrix

- Introduction to Customer Intelligence at Loyalty Matrix
- Introduction to R
- R for Exploratory Data Analysis (EDA)
- **R for Statistics & Data Mining**
- Summary and Q&A

Classical Statistics add Rigor

- Background
  - Business analysts mostly use Excel
  - Great for exploration and building presentations
  - Tests of significance at best an after-thought
- Example: Test Direct Mail Campaign Effectiveness
  - Offer: Test drive “RV” at your local dealer. Get a goodie.
    - ~$100,000 is not an impulse purchase!
    - Low conversion rates (proportions)
  - Control group is “hold out” from each list.
  - Are there incremental sales? How many? $ value?

Fisher.test Measures if Mail Campaign was Effective

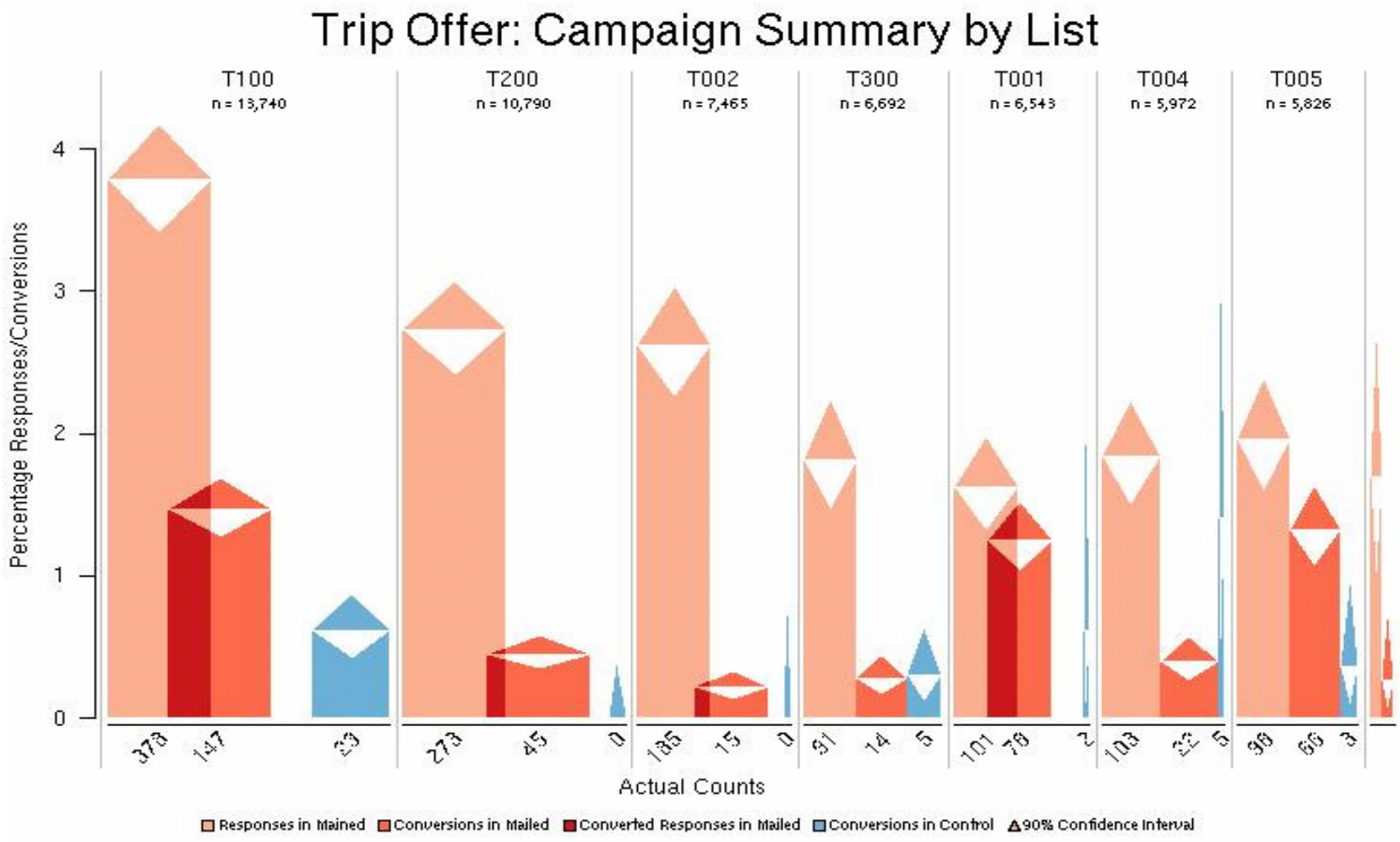
```
>PA.Group <- c(64443, 12060) # Size of Mailed, Control groups
>PA.Sales <- c(139, 15)      # Sales to Mailed, Control groups
>fisher.test(matrix(c(PA.Sales, PA.Group-PA.Sales), 2),
  alternative="greater")

        Fisher's Exact Test for Count Data

data:  matrix(c(PA.Sales, PA.Group - PA.Sales), 2)
p-value = 0.02122
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 1.095079      Inf
sample estimates:
odds ratio
  1.735754
```

We will be wrapping the Fisher test with an easy-to-use interface and result visualization as a Campaign Effectiveness module.

All Mail Lists for a Campaign

	Mailed						Control					P-value	Incremental		
Name	sent	resp	conv	rate	lower	upper	held	conv	rate	lower	upper	p.value	conv	lower	upper
T001:List Source D	6,217	101	78	1.25	1.03	1.51	326	2	0.61	0.11	1.92	0.231	n/s	n/s	n/s
T002:List Source C	7,050	185	15	0.21	0.13	0.33	415	0	0.00	0.00	0.72	n/c	n/c	n/c	n/c
T003:List Source E	1,118	19	3	0.27	0.07	0.69	59	0	0.00	0.00	4.95	n/c	n/c	n/c	n/c
T004:List Source B	5,615	103	22	0.39	0.27	0.56	357	5	1.40	0.55	2.92	0.996	n/s	n/s	n/s
T005:List Source A	5,000	98	66	1.32	1.07	1.62	826	3	0.36	0.10	0.94	0.008	48	30	71
T100:List Source H	10,000	378	147	1.47	1.28	1.68	3,740	23	0.61	0.42	0.87	0.000	86	25	108
T200:List Source G	10,000	273	45	0.45	0.35	0.58	790	0	0.00	0.00	0.38	n/c	n/c	n/c	n/c
T300:List Source F	5,001	91	14	0.28	0.17	0.44	1,691	5	0.30	0.12	0.62	0.658	n/s	n/s	n/s

Marketing Campaign Visualization

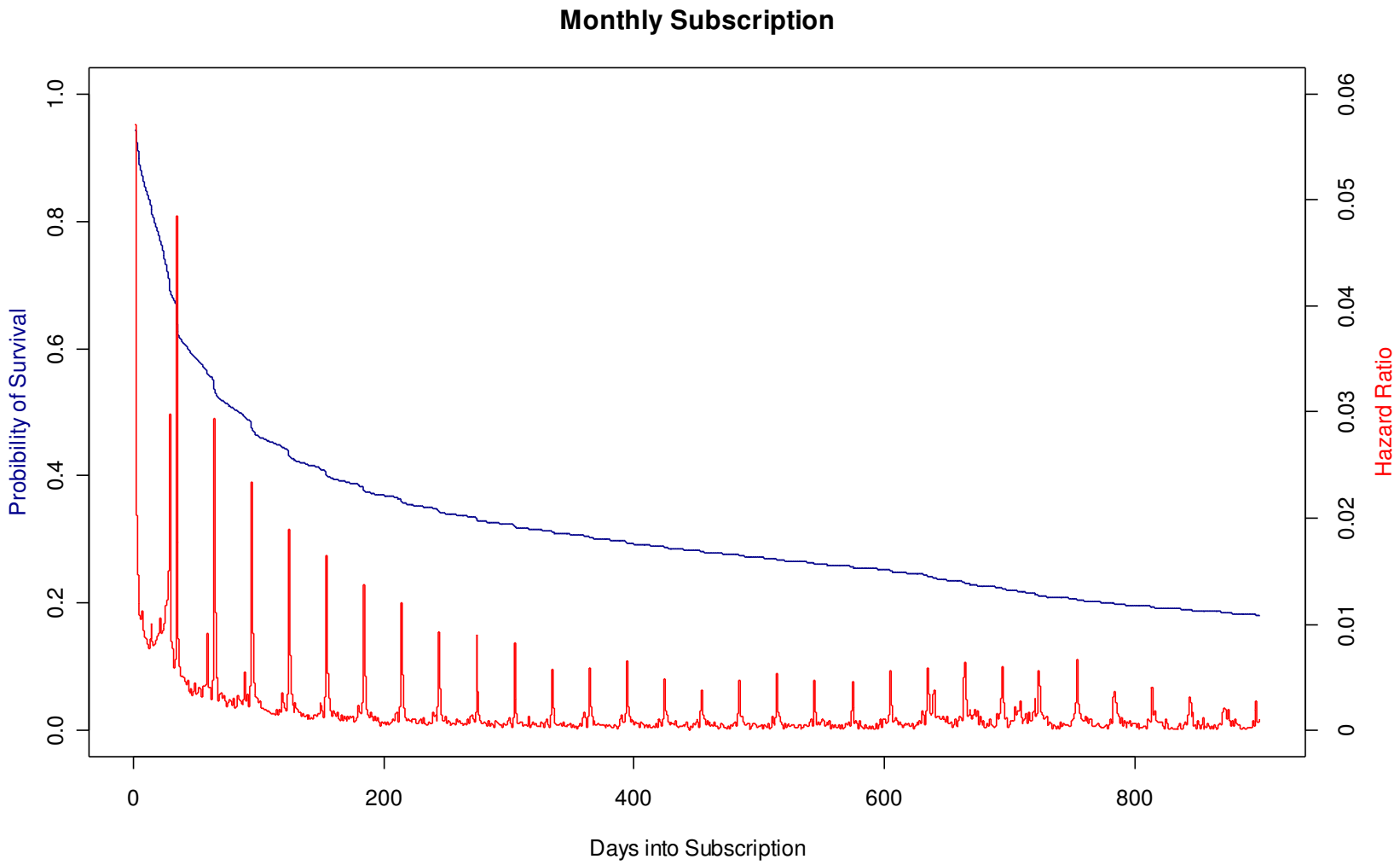


Loyalty Matrix

Subscription Survival Analysis

- Probability of surviving a certain number of days?
- What is half live of typical subscriber?
- Hazard Ratio = (# Expired) / (# at risk) at each time interval
  - H(t) = Ne(t) / Nr(t)
- Probability of Survival = product (1 – H(t)) up to time of interest
  - $Ps(t) = \Pi\ (1 - H(i)\ ;\ i = 1, t$
- Need to know
  - Number of days in subscription
  - Expired or censored?
- See
  - Gordon Linoff's article in Intelligent Enterprise, August 2004
  - Tableman & Kim, *Survival Analysis Using S, Chapman & Hall, 2004*
  - R library: survival and more…

Subscription Survival and Hazard

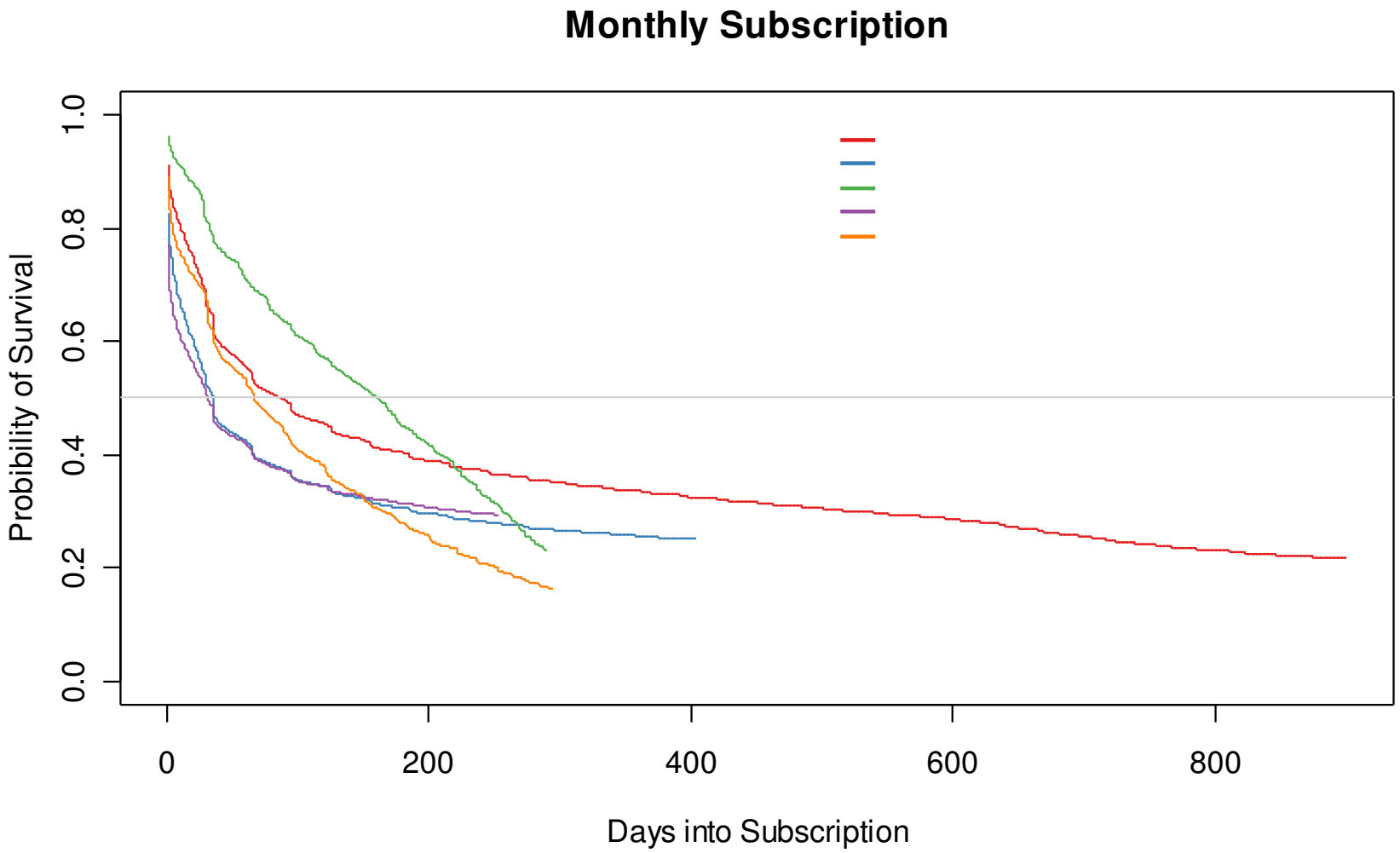


Loyalty Matrix

Some R Code – Read data set & plot survival curve

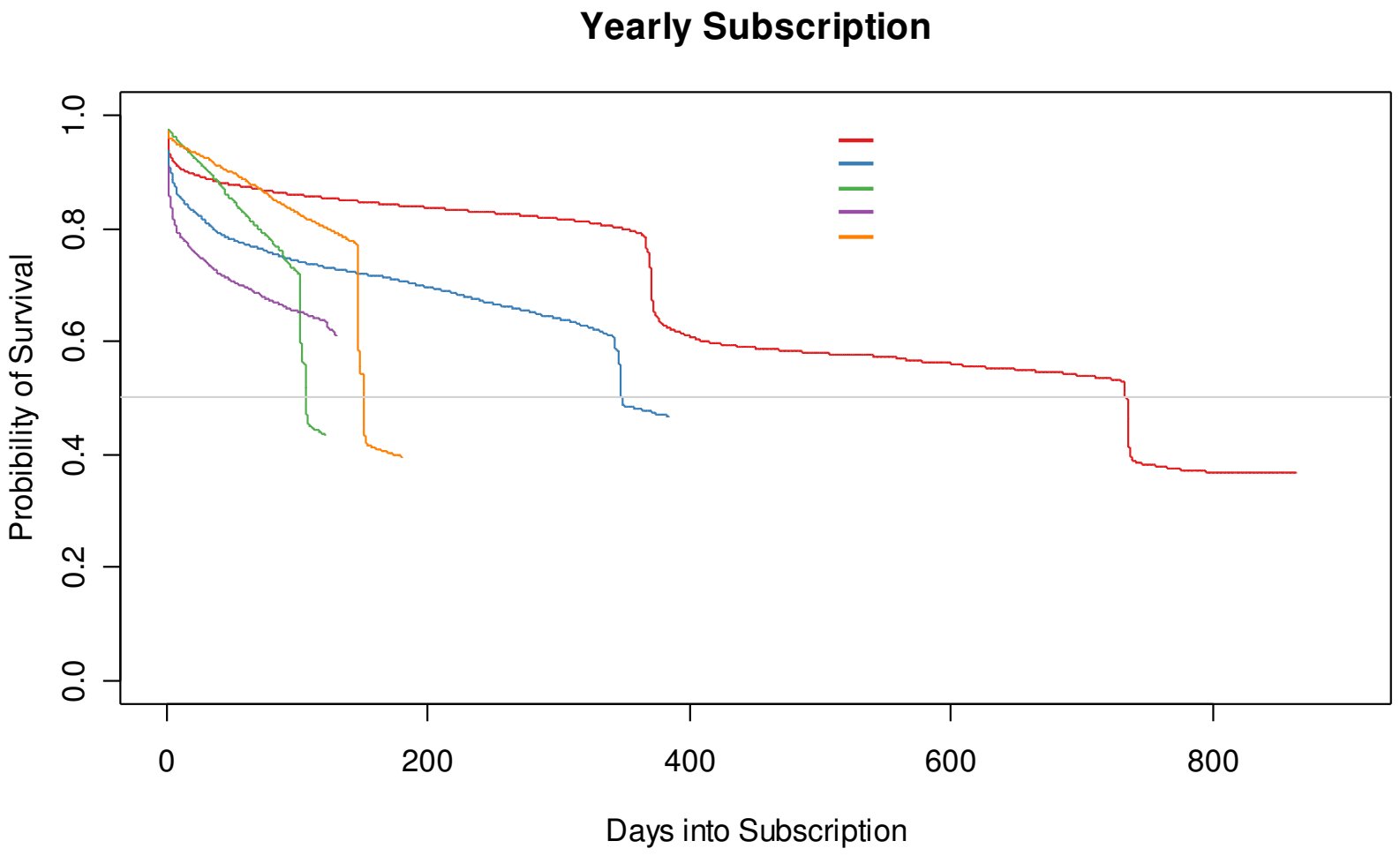
```
library(RODBC)
channel <- odbcConnect(“MyDataSet", uid = “me", pwd = “xx")
sSQL <- readLines("qCustomerSubscriptions4R.sql", n = -1)
PBSub <- sqlQuery(channel, sSQL, na.strings = c("", " "))

nDays <- 900
NumSubs <- length(SubStat)
NumSuccumbed <- table(NumDaysInSub[SubStat=="Expire"])
NumAtRisk <- NumSubs - cumsum(table(NumDaysInSub))
                     + table(NumDaysInSub)[1]
HazardProb <- NumSuccumbed[1:nDays]/NumAtRisk[1:nDays]
SurvivalProb <- cumprod(1-HazardProb)

plot(SurvivalProb, type = "s", ylim = c(0, 1),
                   xlab = "Days into Subscription",
                   ylab = "", main = “Subscription )
```

Survival by Product Type for Monthly Subscriptions



Loyalty Matrix

Survival by Product Type for Yearly Subscriptions



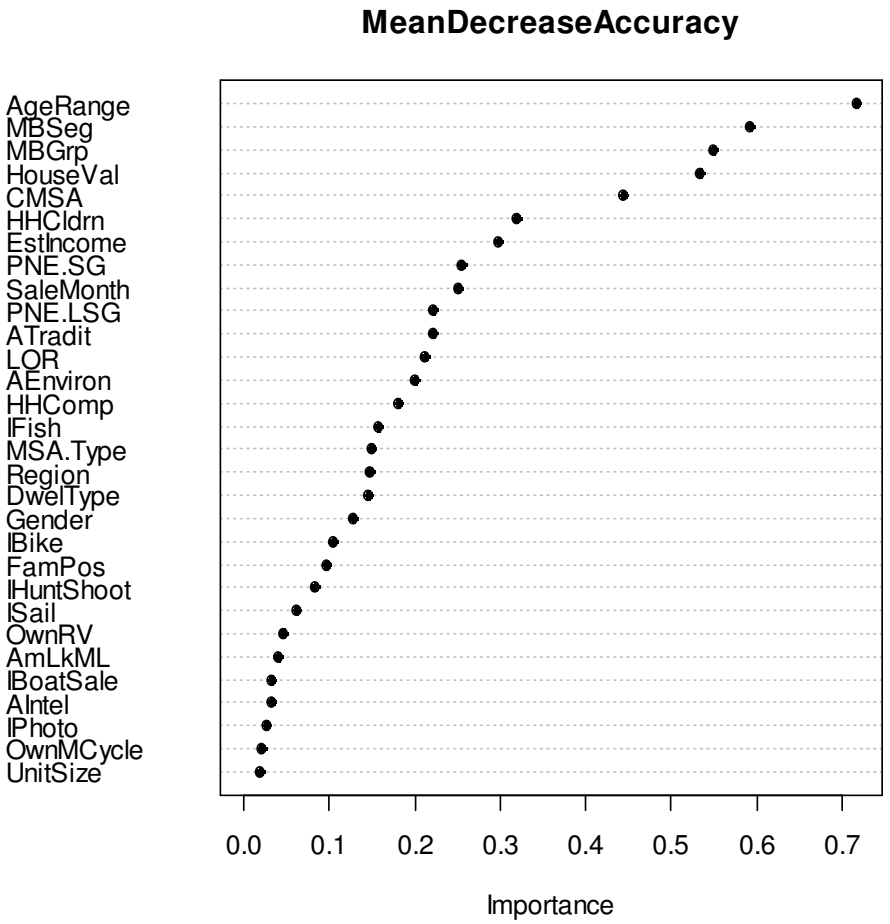
Loyalty Matrix

Advanced Methods Add CI Insight

- Background
  - High-end “data mining” usually done with SAS, SPSS, …
  - Not generally available to front line business analysts
- Example: Customers purchasing different types of vehicles
  - Two distinct product groups
    - High end RV’s
    - Entry level towables
  - How do customer profiles differ?
    - Continuous & categorical data
    - Sales data + 3rd party demographic matching
  - What can we learn for product design? Marketing?

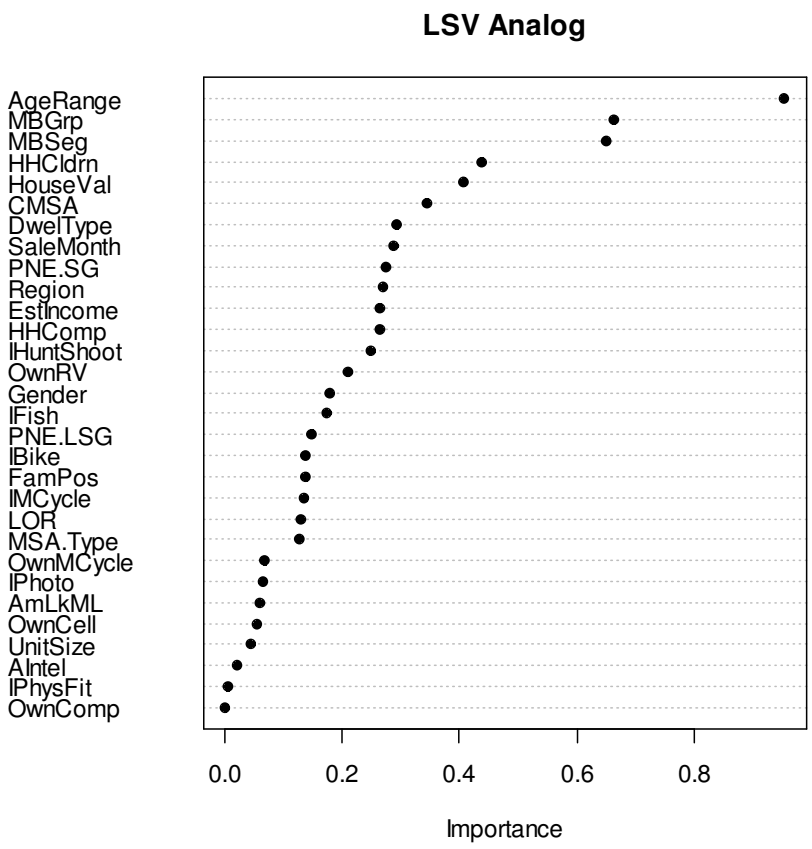
randomForest Differentiates Between Customer Groups

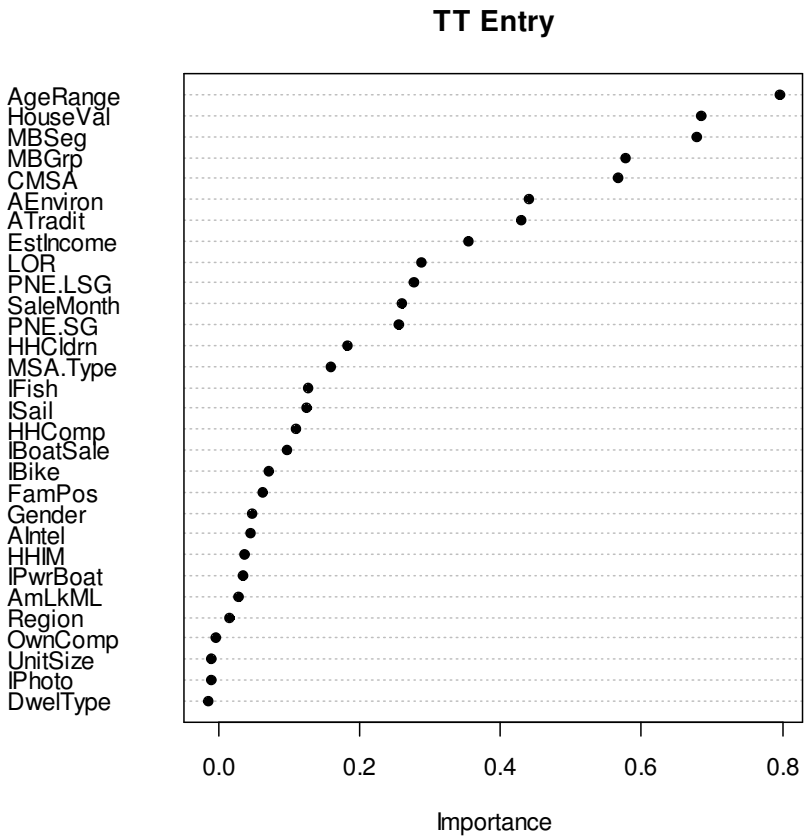
```
> FWE.rf

Call:
 randomForest.default(x = FWEi,
 y = Type, mtry = 6,
 importance = TRUE,
 proximity = TRUE,
 outscale = TRUE)

Type of random forest: classification

Number of trees: 500
No. of variables tried at each split: 6

OOB estimate of  error rate: 28.95%

Confusion matrix:

           LSV Analog  TT Entry  class.err
LSV Analog        622       222    0.26303
TT Entry          295       647    0.31316
```



randomForest package by Andy Liaw and Matthew Wiener based on original Fortran code by Leo Breiman and Adele Cutler

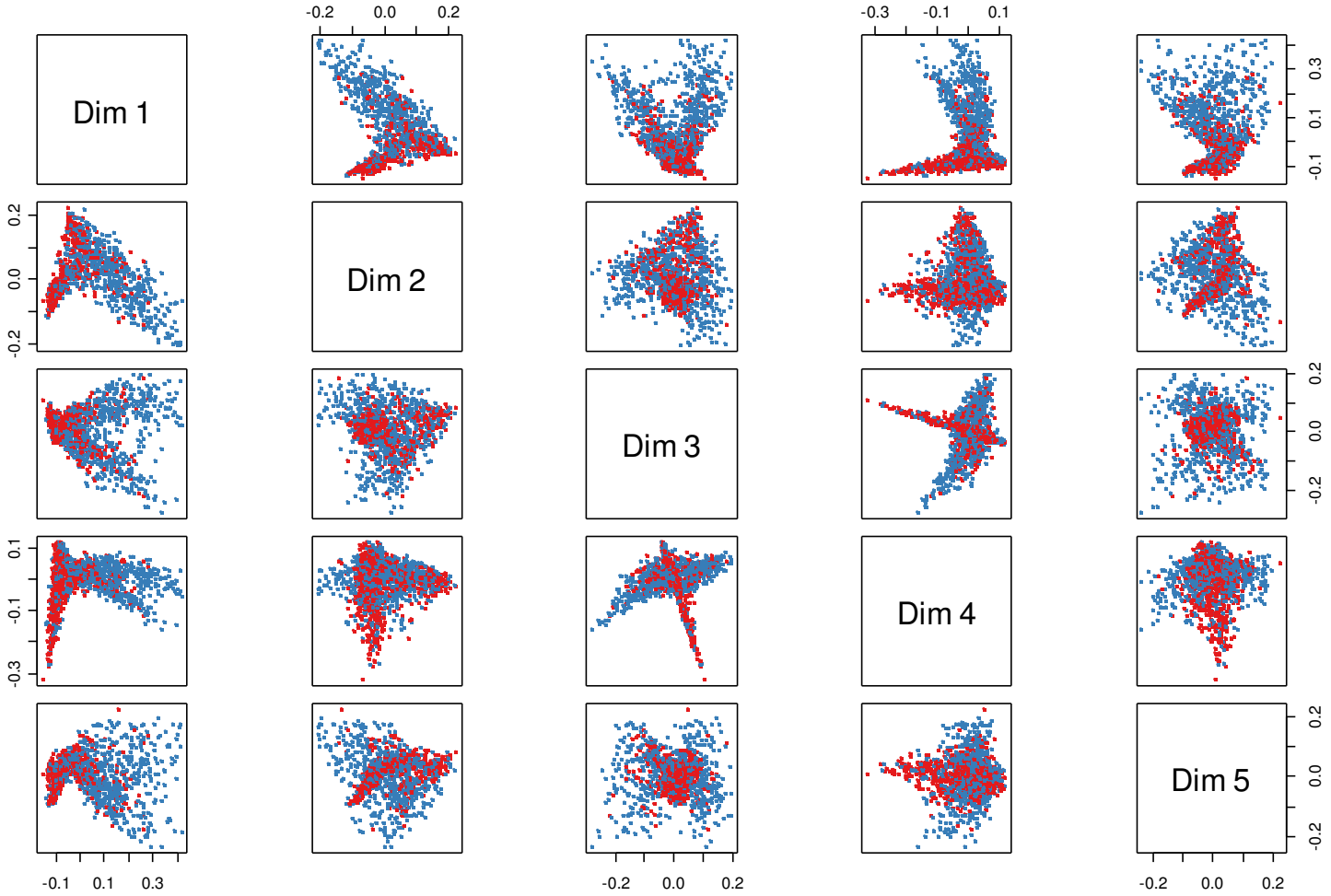
randomForest Shows Variable Importance by Group





Loyalty Matrix

Multi-dimensional Scaling Plot of rF Proximities

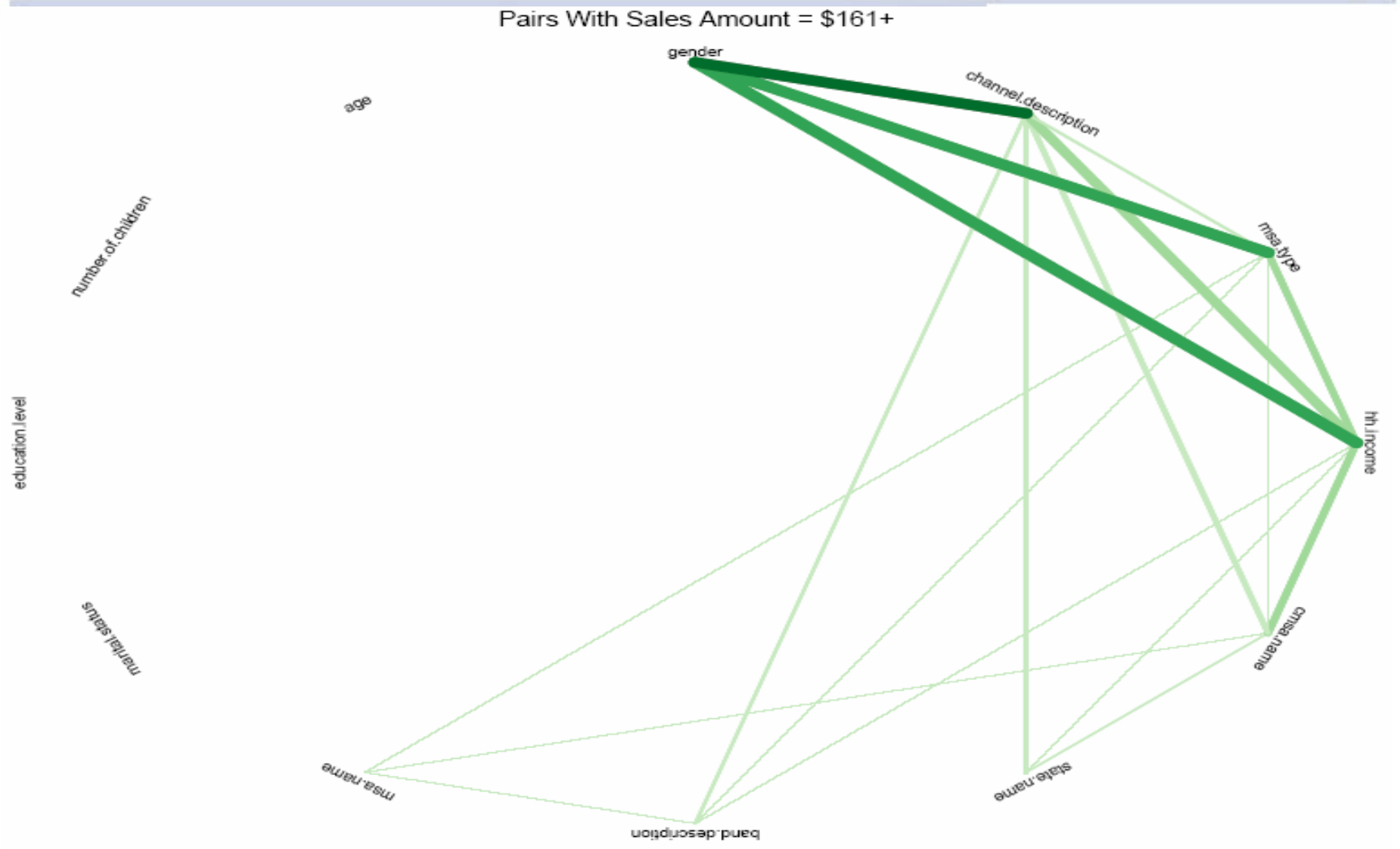


Loyalty Matrix

Association Analysis

- AKA “Shopping Basket” analysis
- Apriori algorithm by Christian Borgelt

Demographic Associations for High Spenders



Loyalty Matrix

R Lessons Learned at Loyalty Matrix

- R a flexible addition to “classical CI” methods
- Yes, there is a learning curve
- Very responsive R user community and support
- No problem with client acceptance

Questions, Comments?

- Now would be the time
- My email is JPorzak@LoyaltyMatrix.com
- Our R Blog is R.LoyaltyMatrix.com