CRAVENS - Part 2 Excercies

Conoly Cravens (mcc4443)

8/3/2021

# **Question 1** Visual Story Telling Part 1: *Green Buildings*

# **Question 2** Visual Story Telling Part 2: *Flights at ABIA*

# **Question 3** Portfolio Modeling

We assessed three different portfolios VaR:

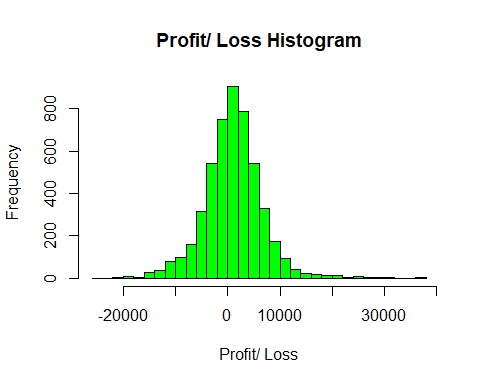
* Income Focused Portfolio
* A portfolio made up of *5 EFTs* with high dividend payout.
* Tech Portfolio
* A portfolio made up of *5 EFTs* focused on technology equities.
* As a reminder, the tech industry is very volatile.
* Balanced Portfolio
* A portfolio made up of *7 EFTs* focused on balanced and diversification.

## Income Focused Portfolio

The five EFTs:

* **SDIV** {Global Equities} - offer exposure to a basket of dividend-paying equities on a global scale
* **YLD** {High Yield Bonds} - provide “current income with diversified risk” by investing in companies with a “defensive quality bias.”
* **IJJ** {Mid Cap Value Equities} - exposure to mid-cap stocks that exhibit value characteristics and fine tune their domestic equity exposure
* **REZ** {Real Estate} – known for distributing 90% of their income to investors
* **SPYD** {Large Cap Blend Equities} – top 80 dividend-yielding companies in the S&P 500

A quick glimpse of the profit/loss histogram for the 5,000 bootstrap samples:

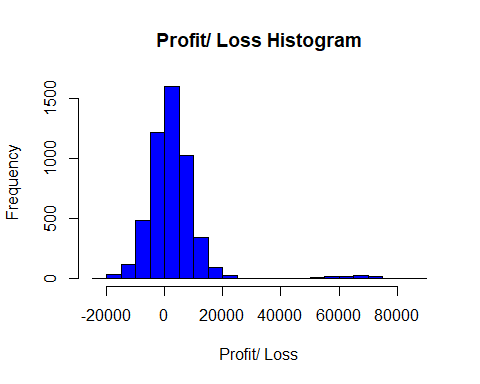
 Using the 5,000 bootstrap samples, we estimate the the 4-week value at risk for the *income driven portfolio* at the 5% level to:

## [1] "$8,156"

## Tech Portfolio

* **SPYG** {Large Gap Growth Equities} - over 300 holdings and exposure is tilted most heavily towards technology
* **QQQ** {Large Gap Growth Equities} - useful as part of a buy-and-hold approach for investors looking to maintain a tilt towards the potentially volatile tech sector
* **XLK** {Technology Equities} - it invests in companies from all across the technology sector
* **TDIV** {Technology Equities} - First Trust NASDAQ Technology Dividend Index Fund
* **FXL** {Technology Equities} - looking for a more qualitative approach to the tech sector

A quick glimpse of the profit/loss histogram for the 5,000 bootstrap samples {note a couple of huge wins since the tech industry is so volitle}:



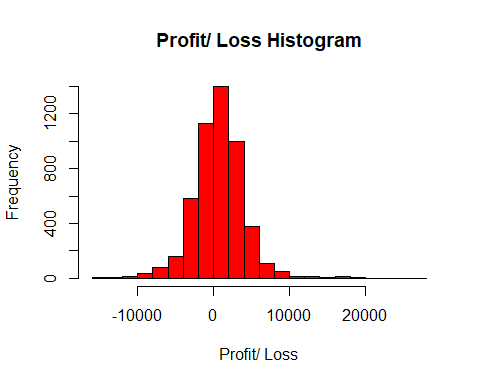
Using the 5,000 bootstrap samples, we estimate the the 4-week value at risk for the *tech portfolio* at the 5% level to:

## [1] "$8,205"

## Balanced Portfolio

* **RSP** {Large Cap Blend Equities} - considerably more balanced than other alternatives such as SPY, and a methodology that some investors believe will add value over the long haul
* **BSV** {Total Bond Markets} - great safe haven to park assets in volatile markets
* **RYT** {Technology Equities} - exposure that is considerably more balanced
* **SPEM** {Emerging Market Equities} - well-diversified option for long-term investors building a balanced portfolio
* **SCHF** {Foreign Large Cap Equities} - close to 1,000 individual holdings, this ETF brings immediate diversification
* **VXF** {All Cap Equities} - extremely diversified in small and mid caps
* **AOA** {Diversified Portfolio} - seeking an aggressive strategy that tilts towards equities and away from fixed income

A quick glimpse of the profit/loss histogram for the 5,000 bootstrap samples:



Using the 5,000 bootstrap samples, we estimate the the 4-week value at risk for the *Balanced portfolio* at the 5% level to:

## [1] "$4,451"

So in summary, we have three portfolios. One focused on dividend payout, one focused on tech (volatile industry), and another one with a well balanced portfolio. For each portfolio, we estimated the 4-week VaR at the 5% level:

## Portfolio VaR  
## 1 Income-Drive $8,156  
## 2 Tech $8,205  
## 3 Balanced $4,451

To no surprise, the balanced portfolio has the lowest amount of risk for a 4-week (20 trading day) period. Tech and income-driven turn out to have about the same amount of risk.

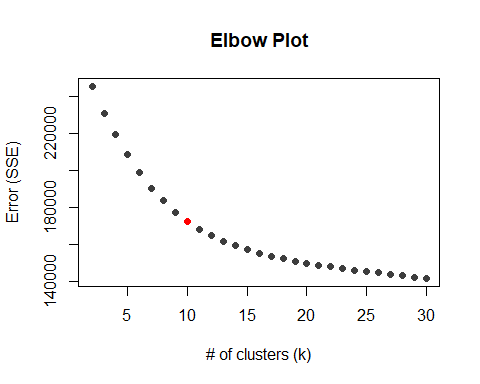
# **Question 4** Market Segmentation

FOR: NutrientH20 Executives OBJECTIVE: Identify market segments that appear in your social-media audience

## Methodology: Kmeans Clustering

After some data exploration and fitting a variety of models, we determined that K-means clustering was an effective (and efficient) way to determine market segments.

### Determining the Optimal Number of Clusters (k)



From the elbox plot, we determined that **10** clusters optimizes the bias-variance trade off in our error (slope from 10 -> 11 is less steap than the slope from 9 -> 10).

## Result

After dividing our followers into 10 segments, we get the following table of centers (conditional formatted to highlight the factors that each cluster scored well in).

1

2

3

4

5

6

7

8

9

10

current\_events

0.10914581

0.05848016

-0.1977810

0.10629933

0.034542847

0.316328500

0.186429132

-0.077951523

0.33235797

-0.0254193582

travel

3.27190413

-0.18542261

-0.2263512

-0.10176818

0.048679256

0.238104631

-0.057760074

-0.031582744

-0.20185833

-0.1569172296

photo\_sharing

-0.11137698

-0.21574178

-0.4192754

-0.09574796

-0.174684076

-0.073458797

1.257802892

-0.009980486

1.10547322

-0.1171317611

tv\_film

-0.07173772

-0.00575064

-0.2118607

-0.09321536

-0.194493863

2.781838781

-0.135118842

0.107052269

-0.12401476

-0.1395029923

sports\_fandom

-0.20726375

0.67118752

-0.3212934

2.11490950

-0.155953351

-0.126612069

-0.206660261

-0.131180557

-0.20041004

-0.1935502056

politics

3.13063324

1.22750715

-0.3023948

-0.22390866

-0.190805114

-0.082009680

-0.122715847

-0.166187127

-0.15074810

-0.1920881061

food

0.16182288

-0.15675589

-0.3650572

1.87203059

-0.121226702

0.142870547

-0.190170650

-0.098710261

-0.30456109

0.4780753983

family

-0.08219719

0.22953691

-0.3113647

1.53239366

0.033554124

-0.133051231

0.042846335

0.218305145

-0.05049952

-0.1008332847

home\_and\_garden

0.04388347

0.15904958

-0.2048995

0.17431502

0.086104074

0.348989815

0.145956798

0.055075352

0.10846984

0.1535806743

music

-0.04190820

-0.08557909

-0.2207815

0.02334339

-0.015259800

0.971167965

0.537503715

-0.056491035

0.14655760

-0.0001695902

news

1.13789009

2.66037399

-0.3111708

-0.11073982

-0.131201280

0.014095015

-0.064270775

-0.190055004

-0.27347973

-0.0699468911

online\_gaming

-0.15553666

-0.12986855

-0.2349684

-0.07658452

-0.011631837

-0.174816112

-0.015688979

3.591328301

-0.16680430

-0.1178383548

shopping

-0.07427601

-0.17779910

-0.3907513

-0.02292988

-0.192007570

0.036162385

0.199371223

-0.117605377

1.38412244

-0.0636136271

health\_nutrition

-0.16121257

-0.24360412

-0.3078685

-0.13755045

-0.197874386

-0.153173220

-0.059152858

-0.178755187

-0.23420340

2.2416016645

college\_uni

-0.03834259

-0.19933946

-0.2547455

-0.13144109

-0.125144845

0.319614434

-0.007490816

3.327887289

-0.08702659

-0.2112635232

sports\_playing

0.04090674

-0.08758388

-0.2613880

0.11205967

-0.176202123

0.099697127

0.201218663

2.163068433

-0.03653960

-0.0246004068

cooking

-0.18827971

-0.24130535

-0.3234069

-0.09608011

-0.156156334

-0.127553306

2.835033912

-0.119923124

-0.22608448

0.4162971978

eco

0.17199992

-0.10401578

-0.2787385

0.16792980

0.190421266

0.089757822

0.020099723

-0.065956112

0.28043572

0.5747774678

computers

2.94554722

-0.18978301

-0.2652666

0.08301652

0.099437004

-0.166922641

0.080135353

-0.079289519

-0.03861193

-0.0885720956

business

0.54332109

-0.11860750

-0.2394405

0.11738548

-0.152352723

0.336647885

0.234304130

-0.096937484

0.34401672

0.0502804027

outdoors

-0.04537119

0.29559536

-0.3363297

-0.07183751

0.237642118

-0.098047424

0.022455618

-0.138310971

-0.25115567

1.7463144775

crafts

0.19280650

-0.16529131

-0.2948734

0.69778561

0.020629751

0.788822602

0.095682697

0.024182746

0.06888754

0.0873579483

automotive

-0.13344136

2.58899421

-0.3158703

0.11571656

0.127956124

-0.232944689

0.016614991

0.059887522

0.06778716

-0.1951588978

art

-0.15993898

-0.16520164

-0.2397956

-0.01966460

-0.037595455

2.801091104

0.002341985

0.289501860

-0.20751566

-0.0640105596

religion

0.11777009

-0.17916520

-0.2986564

2.31519685

-0.181605572

0.004152806

-0.124590545

-0.182245214

-0.25021381

-0.1548297146

beauty

-0.18362257

-0.18616351

-0.2748303

0.32751589

-0.077074345

0.004099199

2.680773249

-0.224220907

-0.18146690

-0.2081397276

parenting

0.02354578

0.03494837

-0.3290974

2.18071420

0.061159738

-0.183408021

-0.069356818

-0.122939241

-0.19871460

-0.0869272261

dating

0.30690965

-0.03440827

-0.1800022

0.02787447

0.007539990

-0.087134906

0.015965495

-0.008810570

0.36154019

0.1779129089

school

-0.10351228

0.01656593

-0.3279260

1.70372772

0.002801984

-0.045882464

0.166725756

-0.214497007

0.11456148

-0.1843845945

personal\_fitness

-0.14326485

-0.23582040

-0.3303525

-0.09380323

-0.079425378

-0.158803880

-0.035753996

-0.181447042

-0.18802745

2.1806483710

fashion

-0.17521031

-0.21791781

-0.2918598

0.01999443

-0.074797890

-0.010845632

2.730643123

-0.069802964

-0.05662950

-0.1139196947

small\_business

0.41077931

-0.15856783

-0.2211531

0.08691185

0.431075230

0.781020747

0.167977868

0.105757151

0.18459313

-0.1499010468

bot

-0.14955067

-0.17957302

-0.1570436

-0.06966897

4.956453864

-0.107578038

-0.105444145

-0.109702927

-0.15340703

-0.1185267170

unknown

-0.10029258

-0.08160946

-0.3933973

-0.15160878

0.083966218

0.048079368

0.058670606

-0.085460393

1.46129945

-0.0979588792

## Cluster Analysis

### Cluster 1: Working Professionals

Cluster one had high scores in *travel*, *politics*, and *computers*, and medium scores in business, small business, and news. Based on these characteristics, we classify these are working professionals.

### Cluster 2: Middle Age Men

Cluster two had high scores in *politics*, *news*, and *automative*, and medium scores in sports fandom and outdoors. Based on these characteristics, we classify these are middle age men.

### Cluster 3: Uncharacterized

Cluster three had all negative scores, meaning no category stood out. This is the cluster of people we could not classify.

### Cluster 4: Parents

Cluster four had high scores in *family*, *religion*, *parenting*, *school*, *sports fandom*, and *food*. Based on these characteristics, we classify this segment as parents.

### Cluster 5: BOTs

Cluster five are the bots that were not initially filtered out.

### Cluster 6: Artists

Cluster six had high scores in *tv film* and *art*, and medium scores in music, craft, and small businesses. Based on these characteristics, we classify these as artists.

### Cluster 7: Young Women

Cluster seven had high scores in *cooking*, *beauty*, and *fashion*, and medium scores in photo sharing and music. Based on these characteristics, we classify these as young women

### Cluster 8: College Students

Cluster eight had high scores in *online gaming*, *college university*, and *sports playing*. Based on these characteristics, we classify these as college students.

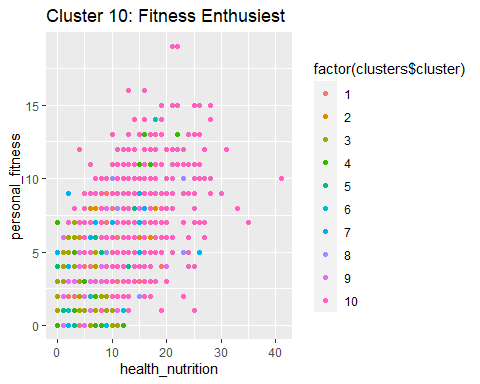
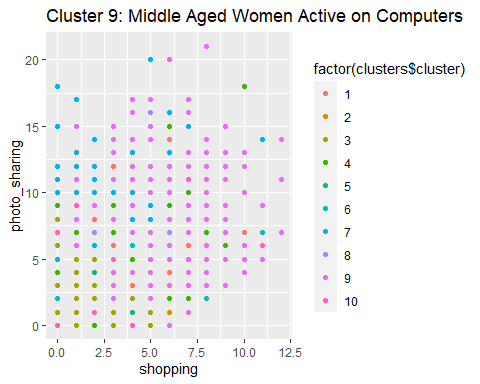
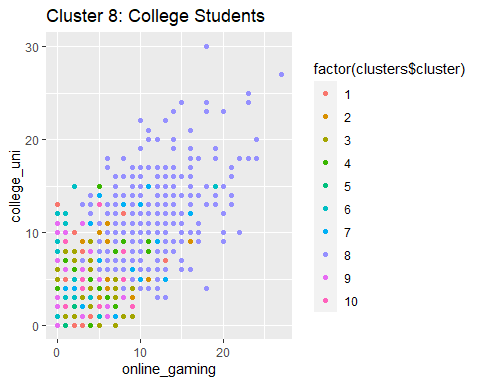
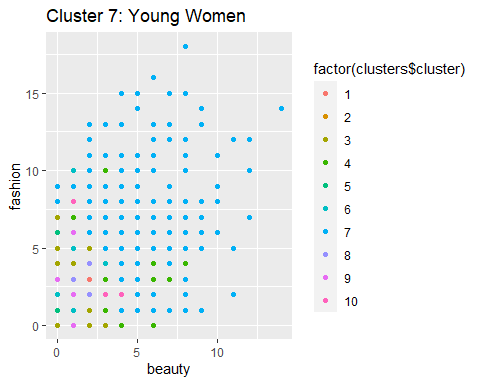
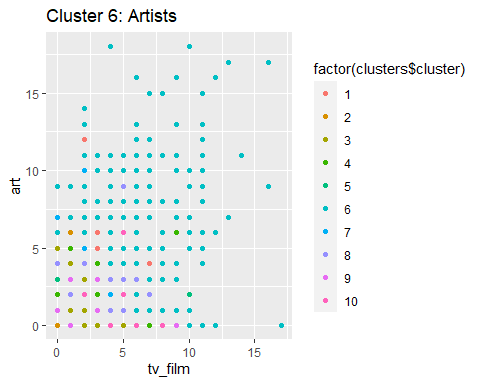
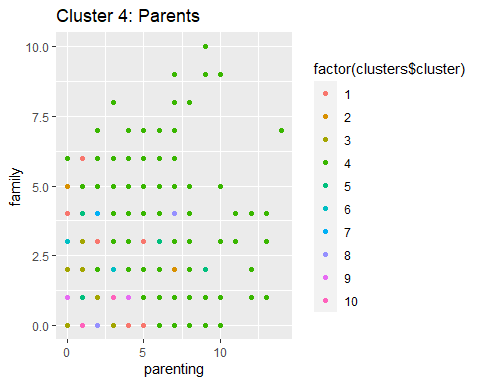
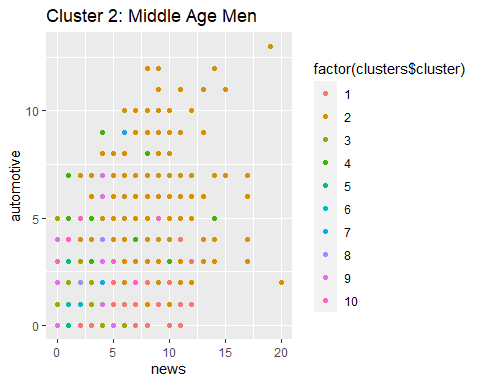
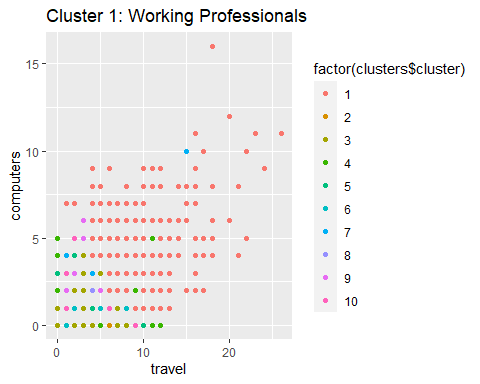
### Cluster 9: Middle Aged Women Active on Computers

Cluster nine had high scores in *photo sharing*, *shopping*, and *unknown (combination of chatter and uncategorized*. Based on these characteristics, we classify these as middle age women active on their computers.

### Cluster 10: Fitness Enthusiest

Cluster ten had high scores in *health nutrition*, *outdoors*, and *personal fitness*, and medium scores in food, cooking, and eco. Based on these characteristics, we classify this cluster as fitness enthusiast.

To better picture these clusters, I put together a plot for each cluster (excluding Bots & Uncharacterized) showing dots for different users along the scale of two primary factors. Each dot is colored in with the cluster. You will see for each graph, the cluster being highlighted dominates the graph!



## Summary

We saw a total of 10 different clusters, all with different sizes:

## Cluster Characteristic Cluster.Size Percent.of.Audience  
## 1 1 Working Professionals 349 4  
## 2 2 Middle Age Men 428 5  
## 3 3 Uncharacterized 3251 41  
## 4 4 Parents 664 8  
## 5 5 Bots 214 3  
## 6 6 Artists 387 5  
## 7 7 Young Women 468 6  
## 8 8 College Students 351 4  
## 9 9 Middle age women on computers 1019 13  
## 10 10 Fitness Enthusiest 751 10

While we would ideally trim down the uncharacterized users to a unique group, increasing the number of clusters makes other categories too small (less than 2%) and therefore take away from the overall effectivness of market segmentation for business implenetation.

# **Question 5** Author Attribution

# **Question 6** Association Rule Mining