Five Repositories, One Dataset

USING EXPLORATORY DATA ANALYSIS TECHNIQUES TO TRACK PATTERNS OF USE

Mark Custer
Noah Huffman
Jennie Levine Knies
Kyle Rimkus
Sara Snyder

Outline of Today's talk

1. Introduction: Exploratory and preliminary nature of the study
2. Overview of website / EAD-portal metrics for three years
3. The path to an aggregate data set and difficulties
4. Collection-level metrics: one year, in depth
5. Visits from Mobile devices over the years
6. Wikipedia referrals over the years
7. Conclusion: Next Steps

1: Introduction

News from Google

Images and B-roll

Blog directory

Google+ directory

Twitter directory

Facebook directory

YouTube directory

Web Analytics Free of Charge, Courtesy of Google

Subscribe

Powerful web analytics service now available to all businesses

Mountain View, Calif. Nov. 14, 2005 – Today, Google Inc. (NASDAQ: GOOG) today announced that its hosted web analytics service, Google Analytics, is now free. Formerly known as Urchin from Google, Google Analytics helps businesses use performance data to improve their online marketing campaigns and websites. With Google Analytics, businesses can determine what keywords attract the most visitors, which email campaigns create more customers, and how to design web pages that hold people's attention.

"We want to give all online marketers and publishers access to powerful web analytics to help them better understand what their customers want. With this knowledge, businesses can create more accurate advertising and build better websites," said Paul Muret, Google engineering director, and one of the founders of Urchin. "By making this powerful service free, we aim to give all websites – large and small – the tools they need to better serve their customers, make more money, and improve the web experience for everyone."

Google Analytics can enhance every aspect of online marketing – from selecting and bidding on effective keywords, to determining the most relevant offers in email campaigns, to optimizing website design. By acting on this information, businesses of all sizes can attract more visitors, convert more prospects to customers, and improve the overall return on their marketing investment. Google Analytics is simple enough for businesses new to web analytics to get started quickly, and sophisticated enough for the most advanced online marketers.

In addition to being free, Google Analytics includes:

- Integration with Google AdWords, and works with any online ad network: Users of Google AdWords can access web analytics from a new interface within their AdWords account. Google Analytics automatically tags keyword destination URLs (which saves time and reduces the potential for errors), and imports cost data for ROI reports (for fast set-up and ease-of-use). Google Analytics is also able to track the results of any online marketing campaign, including banner ads, referral links, email newsletters, and organic and paid search.
- New reporting dashboards: Google Analytics executive summary reports for the three most common decision makers – executive, marketer, webmaster – will ease access to crucial information across departments.
- Global reach: Immediately available in 16 international languages: UK English, French, Italian, German, Spanish, Dutch, Japanese, Korean, Simplified Chinese, Traditional Chinese, Portuguese, Danish, Finnish, Norwegian, Swedish, and Russian.

Google Analytics runs on the same computing infrastructure that powers Google.com so it can support the traffic demands of any site, from those with a few visitors a week to hundreds of millions. Google Analytics is already used by many of the top properties on the web, including dozens of Fortune 500 companies. Businesses such as The Financial Times, National Semiconductor, Ritz Interactive,

Subject: Google analytics, and a small step toward open data (EAD Dredge)
Content-Type: multipart/alternative;

Hi all,

I have just recently created a Google Document with a single goal in mind: to share specific datasets more freely amongst archival practitioners and researchers.

At my institution, we have been using Google Analytics for our finding aid website since June 9, 2008. Because of that, we now have a dataset that's been consistently collected for over two years time, but it rarely gets analyzed (and the analysis that I am currently doing is only on a small subset of this data). Further, an advantage of using Google Analytics to collect this data (i.e., offloading the responsibility of collecting and storing that data, which of course has negative consequences, as well) is that it is stored in the "cloud" and can be made easily accessible to anyone else who has a Google account. Therefore, I would like to propose that:

1. If you are a Google Analytics Administrator who has been collecting data on your EAD pages, and

2. If you would be interested in sharing access to your Google Analytics EAD report, and receiving limited access to similar reports;

3. Or, if you are a researcher who would like access to such data;

4. Please send me an email off-list -- to [log in to unmask]<mailto:[log in to unmask]> -- with the name of your Google account username.

5. I will then give you access to this particular Google Spreadsheet, so that you can

6. Fill in your information, and also so that you can

7. Indicate if you would like to be added to other Google Analytics accounts (which would ultimately be left up to the discretion of that account administrator)

Right now, the spreadsheet is very basic, but it could be expanded or moved to a new format altogether if interest warrants. However, I have included one specific column, titled "Notes on data", which can be used to provide additional information in order to ensure that proper analyses could be done. For instance, the default variable that Google Analytics uses to indicate session timeouts is 1800 seconds (30 minutes), but some sites could have changed this default value.

Additionally, I will be happy to provide more information to anyone who is interested in participating (such as how to add new users to a GA report). Right now, I'm just curious if people will be interested and willing to share their web metrics data, and if others are interested in using that data for research. Thanks so much,

Mark Custer

Google Analytics

http://digital.lib.ecu.edu - http://digit...
digital.lib.ecu.edu/special/ead

Home | Standard Reporting | Custom Reporting

ACCOUNTS LIST

Find accounts & profiles

- abbott.lib.ecu.edu — 1 Properties
- CSI — 1 Properties
- digital.lib.ecu.edu — 10 Properties
- digital.lib.umd.edu/archivesum — 1 Properties
- halfalibrarian.com — 1 Properties
- library.duke.edu — 1 Properties
- Paraprofessional Conference website — 2 Properties
- richter_digital — 1 Properties
- web.lib.ecu.edu — 1 Properties
- www.aaa.si.edu — 1 Properties
- www.ecu.edu/cs-lib — 1 Properties

rt ▾ Add to Dashboard

viewed a total of 119,093 times

Pageviews: 119,093

Unique Pageviews: 87,084

Avg. Time on Page: 00:01:51

Bounce Rate: 76.33%

Conversions

Help

The Content Overview Report
Comparing Metrics
Using the Interactive Table

Create account Log in

Article Talk Read Edit View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Exploratory data analysis

From Wikipedia, the free encyclopedia

In statistics, **exploratory data analysis (EDA)** is an approach to analyzing data sets to summarize their main characteristics in easy-to-understand form, often with visual graphs, without using a statistical model or having formulated a hypothesis. Exploratory data analysis was promoted by John Tukey to encourage statisticians visually to examine their data sets, to formulate hypotheses that could be tested on new data-sets.

Tukey's championing of EDA encouraged the development of statistical computing packages, especially *S* at Bell Labs: The *S* programming language inspired the systems 'S'-PLUS and *R*. This family of statistical-computing environments featured vastly improved dynamic visualization capabilities, which allowed statisticians to identify outliers and patterns in data that merited further study.

Tukey's EDA was related to two other developments in statistical theory: Robust statistics and nonparametric statistics, both of which tried to reduce the sensitivity of statistical inferences to errors in formulating statistical models. Tukey promoted the use of five number summary of numerical data—the two extremes (maximum and minimum), the median, and the quartiles—because these median and quartiles, being functions of the empirical distribution are defined for all distributions, unlike the mean and standard deviation; moreover, the quartiles and median are more robust to skewed or heavy-tailed distributions than traditional summaries (the mean and standard deviation). The packages *S*, *S*-PLUS, and *R* included routines using resampling statistics, such as Quenouille and Tukey's jacknife and Efron's bootstrap, that were nonparametric and robust (for many problems).

Exploratory data analysis, robust statistics, nonparametric statistics, and the development of statistical

Dimensions & Metrics Reference

This document lists and describes all the dimensions and metrics available through the Core Reporting API. Use this reference to:

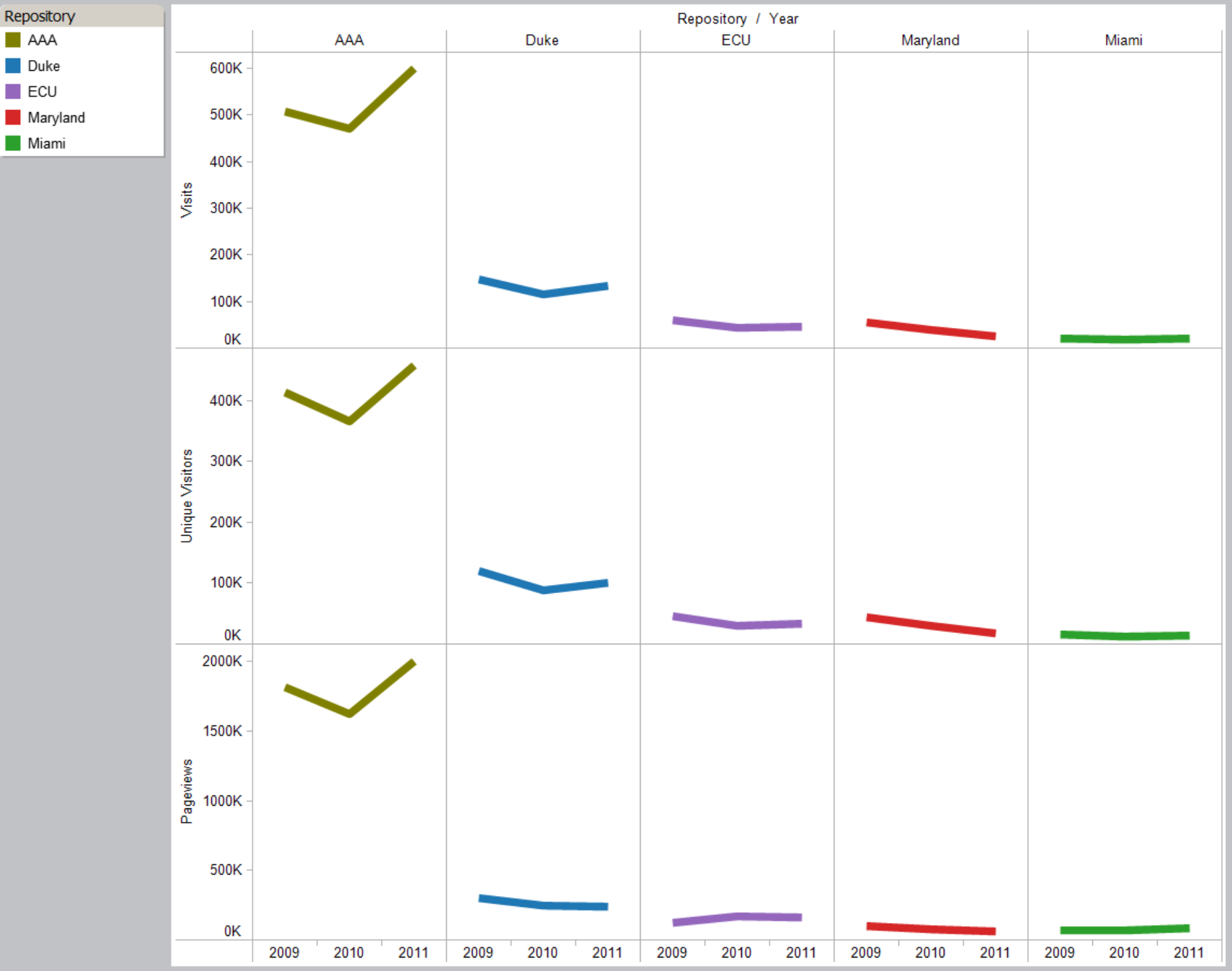
Explore all the dimensions and metrics – Click the plus box to see dimensions and metrics by feature. Search to quickly find the name you're looking for.

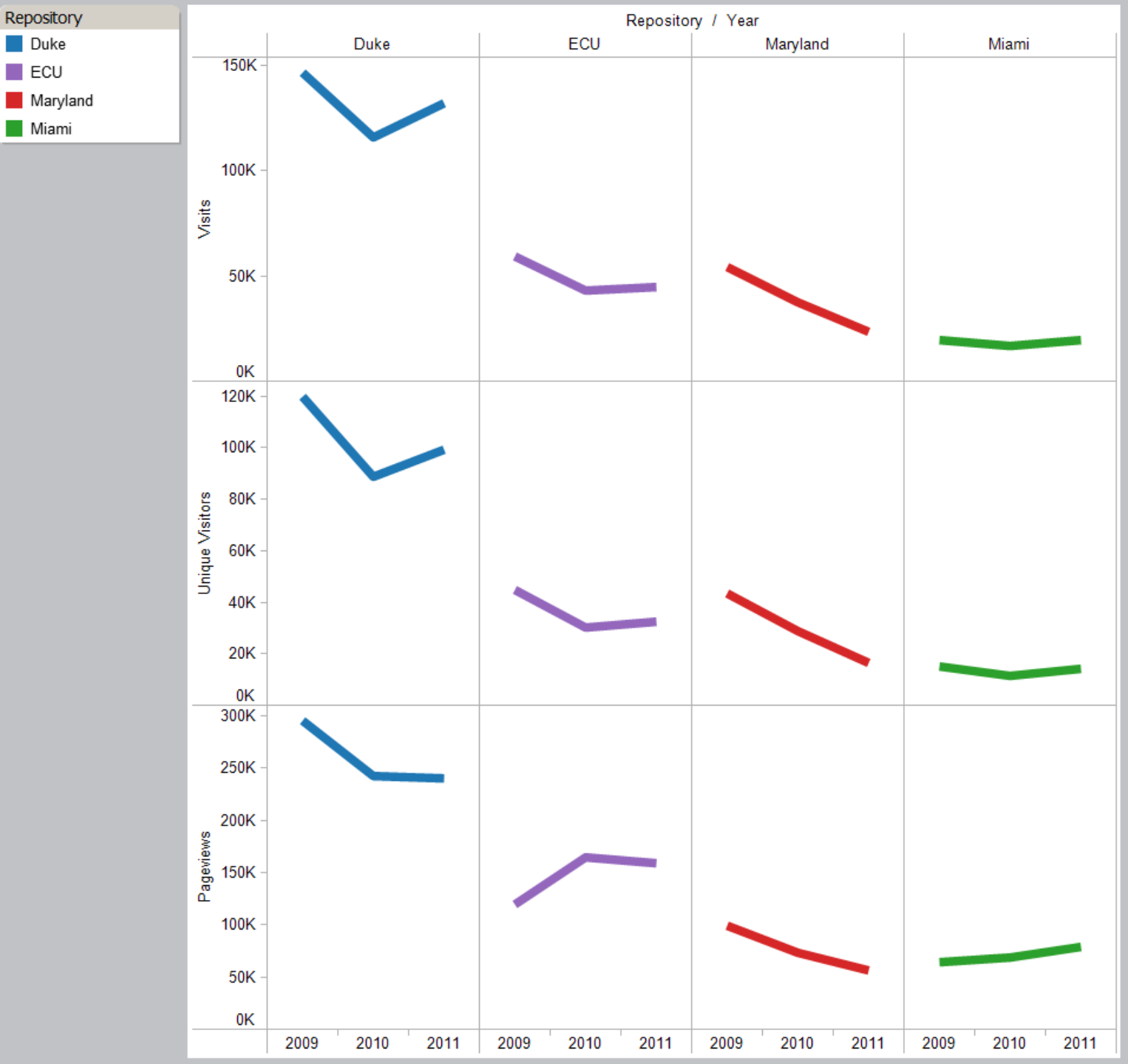
Valid Combinations – Not all dimensions and metrics can be queried together. Only certain dimensions and metrics can be used together to create valid combinations. In the table of contents, roll over each name and select it's checkbox to see all the other values that can be combined in the same query.

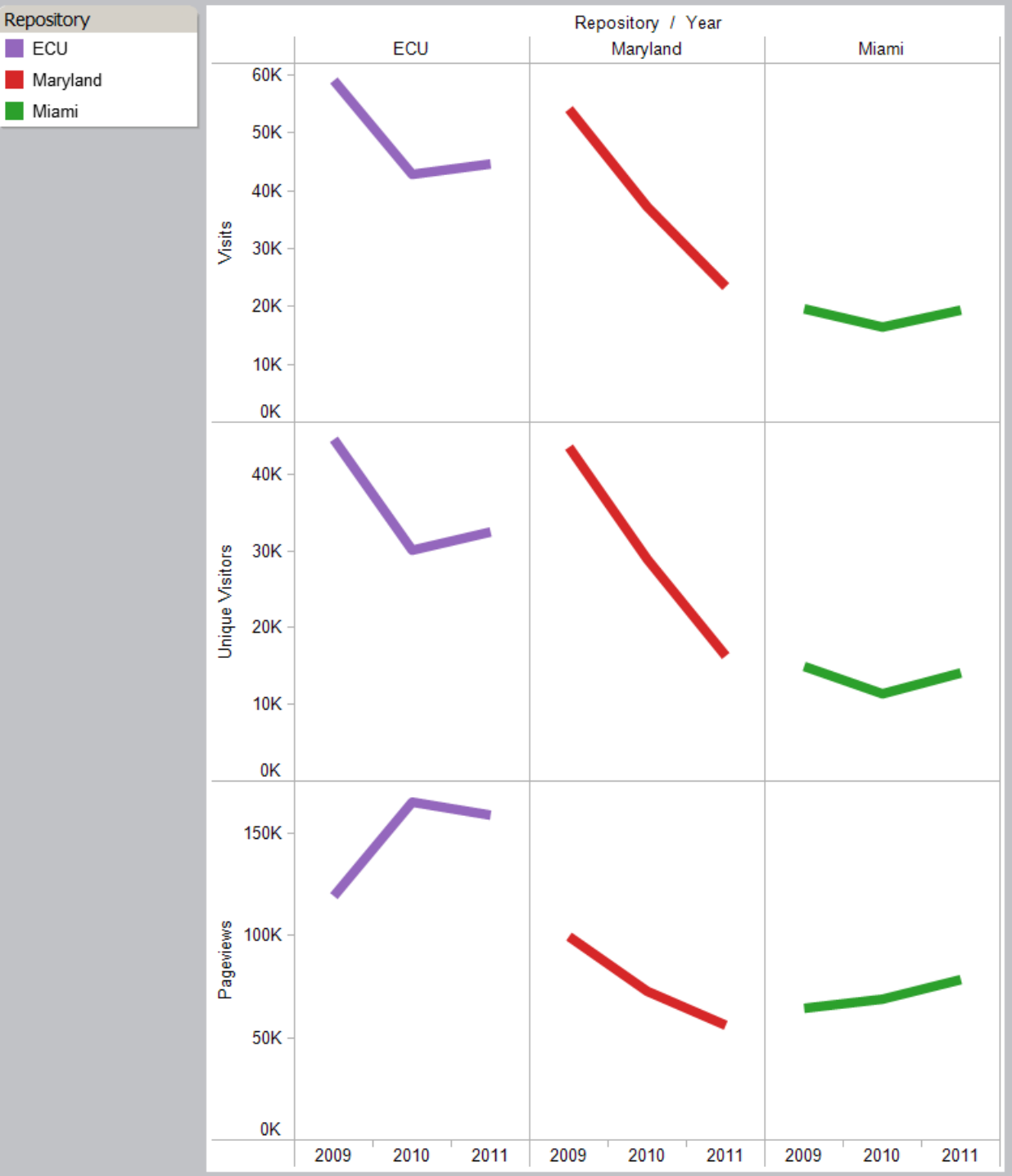
Expand All Search:

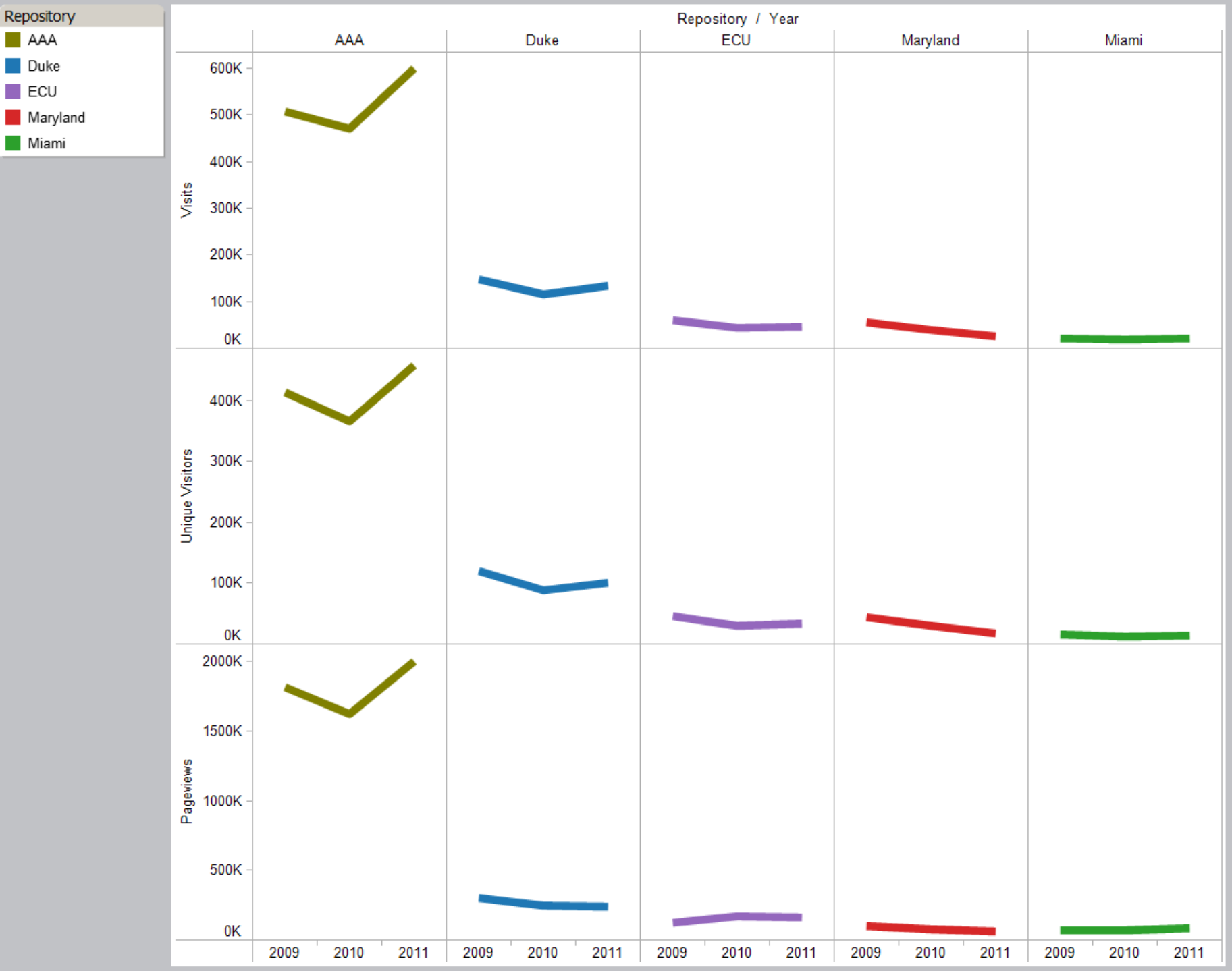
Dimensions **Metrics**

2: Website Metrics, 2009-2011


Repository
AAA
Duke
ECU
Maryland
Miami
Repository / Year
AAA
Duke
ECU
Maryland
Miami
Visits
600K
500K
400K
300K
200K
100K
0K
Unique Visitors
400K
300K
200K
100K
0K
Pageviews
2000K
1500K
1000K
500K
0K
2009
2010
2011


Repository
Duke
ECU
Maryland
Miami
Repository / Year
Duke
ECU
Maryland
Miami
Visits
150K
100K
50K
0K
Unique Visitors
120K
100K
80K
60K
40K
20K
0K
Pageviews
300K
250K
200K
150K
100K
50K
0K
2009
2010
2011


Repository
ECU
Maryland
Miami
Repository / Year
ECU
Maryland
Miami
Visits
60K
50K
40K
30K
20K
10K
0K
Unique Visitors
40K
30K
20K
10K
0K
Pageviews
150K
100K
50K
0K
2009
2010
2011


Repository
AAA
Duke
ECU
Maryland
Miami
Repository / Year
AAA
Duke
ECU
Maryland
Miami
Visits
600K
500K
400K
300K
200K
100K
0K
Unique Visitors
400K
300K
200K
100K
0K
Pageviews
2000K
1500K
1000K
500K
0K
2009
2010
2011

3: The Path and its Difficulties

Home
AAA Exports
Duke Exports
ECU Exports
Maryland Exports
Miami Exports
Sitemap

Edit sidebar

Home

In due time, I hope that this website will develop more structure (more pages, too), but for now, there's this:

Repository:	**Date of GA installation:**	**Date(s) of data gaps:**
Archives of American Art	2008-07-24	2011-04-16 Website down 2011-06-30 Firewall-related site crash No data for PDF versions of finding aids prior to 2011-05-16 (event tracking set up) Note: major web redesign launched 2011-01-14
Duke University	2009-04-28	2012-01-20: Filtering error in adding tracking for new finding aids interface to this profile resulted in four days where no data was collected.
East Carolina University	2008-06-09	"Events" weren't tracked correctly circa 2011-07
University of Maryland	2007-08-31	Anything? Everything looks normal
University of Miami	2009-05-27	There looks to be a gap circa 2011-05.

1st question: Sara has updated the information about AAA, so it looks like this question should be a lot easier to answer now.

Since it looks like all of the data sets should be pretty good for July 2009 - June 2010, how about we start out with that date range for each data set?

When looking at 'mobile' and other metrics, though, we should certainly combine more recent data, too.

2nd question: what data do we want to explore? I'm definitely going to look at UPVs and EPVHs (estimated page view hours per year), and both of those can be derived from GA's 'Top Content' report.

After talking with Sara and Noah, it sounds like we should also look at metrics pertaining to 'Referrals' and 'Mobile'. I also think that it might be interesting to look at 'Keywords,' since most web traffic is generated by search engines, but I doubt we could find out too much information by doing this.

In any event, what else about visitors, social, etc., should we export so that we can investigate the data across all five reports, even if in only a shallow sense?

Home
AAA Exports
Duke Exports
ECU Exports
Maryland Exports
Miami Exports
Sitemap

Edit sidebar

Duke Exports

Same deal as the Miami Exports page. I didn't filter out any of the URLs, so that will still need to be done later for the Top Content report (or, "Pages," as it's now referred to in the new GA interface).

However, Duke had over 80k organic keyword phrases during this annual period, and the GA interface would only allow me to export 50k at a time. So, I'd either need to figure out how to use the GA API, or go back later and export keywords 50,001 to the end of the list. Of course, I don't know if it's even worth looking at the keywords, but putting all 5 data sets together would make one heck of a tag cloud!

(edit: rather than mess with the API, which I couldn't completely figure out, I've exported the rest of Duke's keywords in the file that contains "part2". I'll combine those two files later on.)

+ Add file + Add link Move to Delete Subscribe to changes

Analytics_Duke_200907-201006_(ReferringSourcesReport).tsv Download	120k	v. 1	Apr 13, 2012 1:57 PM	Mark Custer
Analytics_Duke_200907-201006_(ReferringSource_Wikipedia).tsv Download	27k	v. 1	Apr 13, 2012 1:57 PM	Mark Custer
Analytics _Duke_ Organic Search Traffic 20090701-20100630.tsv Download	3146k	v. 1	Apr 13, 2012 1:57 PM	Mark Custer
Analytics _Duke_ Organic Search Traffic_part2 20090701-20100630.tsv Download	2329k	v. 1	Apr 18, 2012 12:56 PM	Mark Custer
Analytics _Duke_ Pages 20090701-20100630.tsv Download	3008k	v. 1	Apr 13, 2012 1:57 PM	Mark Custer

Comments

/collections/findingaids/downgall.htm%20and%20http:/www.aaa.si.edu/collectionsonline/downgall/overview.htm

/collections/oralhistories%20/tranSCRIPTs/levine02.htm

/search?q=cache:zqG_DxtU1AIJ:proust.library.miami.edu/findingaids/?p=collections/controlcard&id=480+orestes+miami&cd=13&hl=en&ct=clnk&gl=us

/translate_c?hl=ar&sl=en&u=http://proust.library.miami.edu/findingaids/%3Fp=collections/controlcard&id=247&prev=/search%3Fq=batista%2Bcollection&hl=ar&client=firefox-a&channel=s&rls=org.mozilla:ar:official&sa=N&rurl=translate.google.com.eg&usg=ALkJrhiuq78PNcimpnEph3V5gEnNNUZuNw

/search?q=cache:wkJ778Y-NEgJ:test.lib.umd.edu/archivesum/actions.DisplayEADDoc.do%3Fsource%3DMdU.ead.histms.0008.xml%26style%3Dead+historical+Davis+family+Texas&cd=6&hl=en&ct=clnk&gl=us

/digitalcollections/rbmscl/inv/results?q=testimonial+advertising&fq=duke.collection%3Ainv&start=0&rows=20&f=keyword&t=testimonial+advertising&btnG.x=0&btnG.y=0

/url_result?ctw_=sT,eCR-EJ,bT,hT,uaHR0cDovL3d3dy5saWIudW1kLmVkdS9hcmNoaXZlc3VtL2h0bWwvTWRVLmVhZC5saXRtcy4wMDA3Lmh0bWw=,qlang=ja|for=0|sp=-5|fs=100%|fb=0|fi=0|fc=FF0000|db=T|eid=CR-EJ,

/archivesum/actions.DisplayEADDoc.do?source=/MdU.ead.scpa.0078.test.xml&style=ead

downgall

levine02

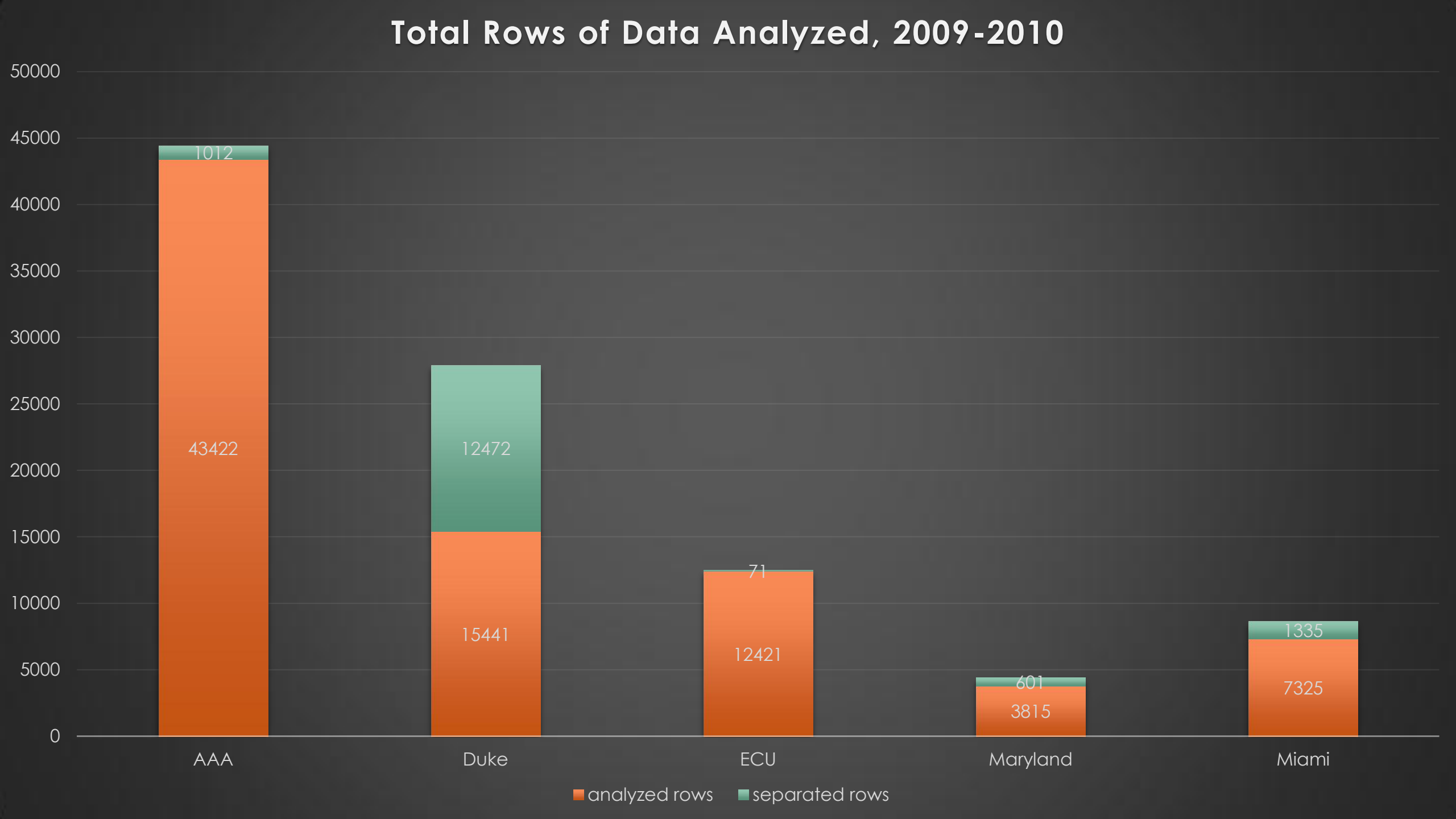
id=
480

id=247

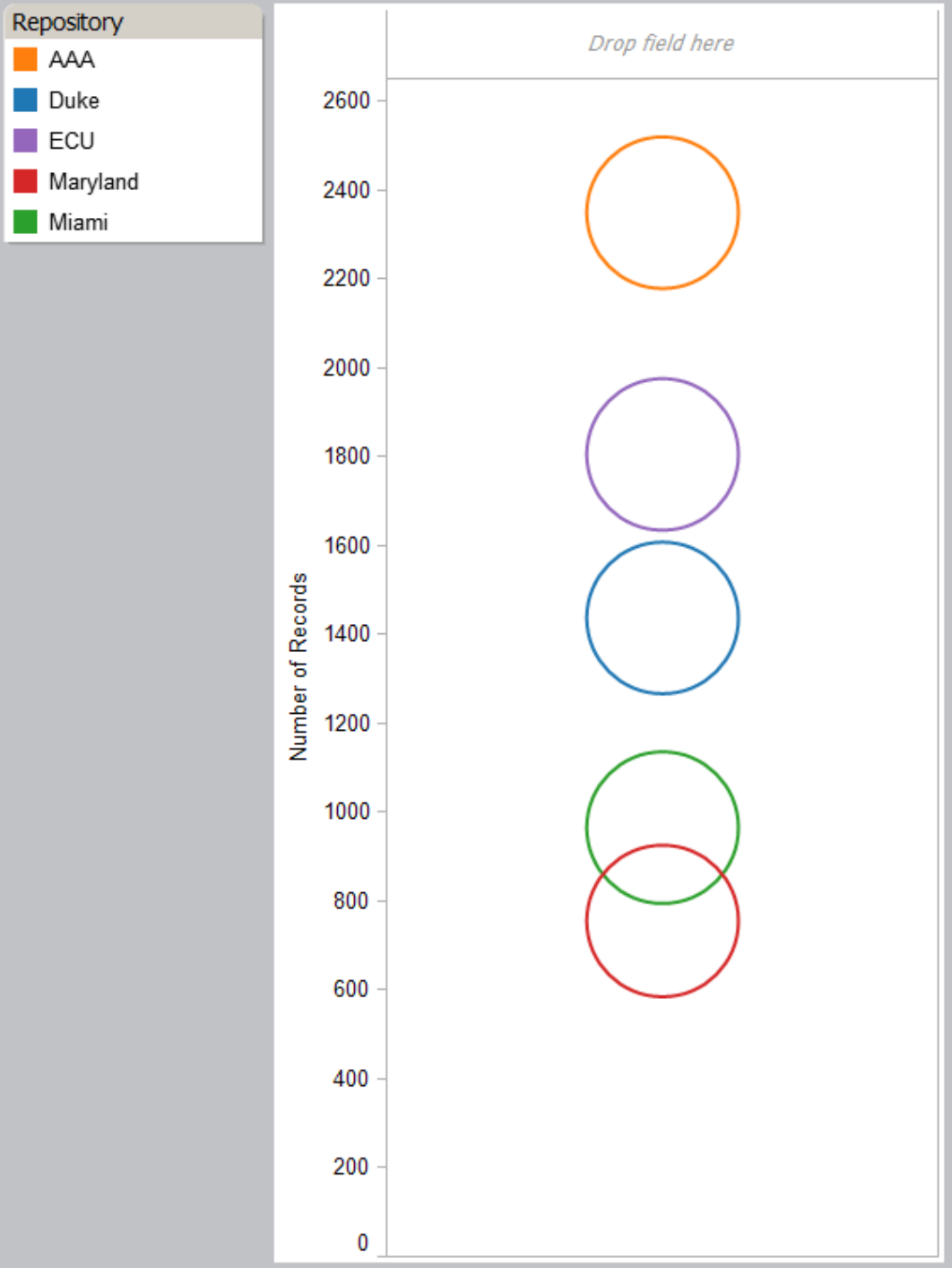
MdU.ead.histms.0008

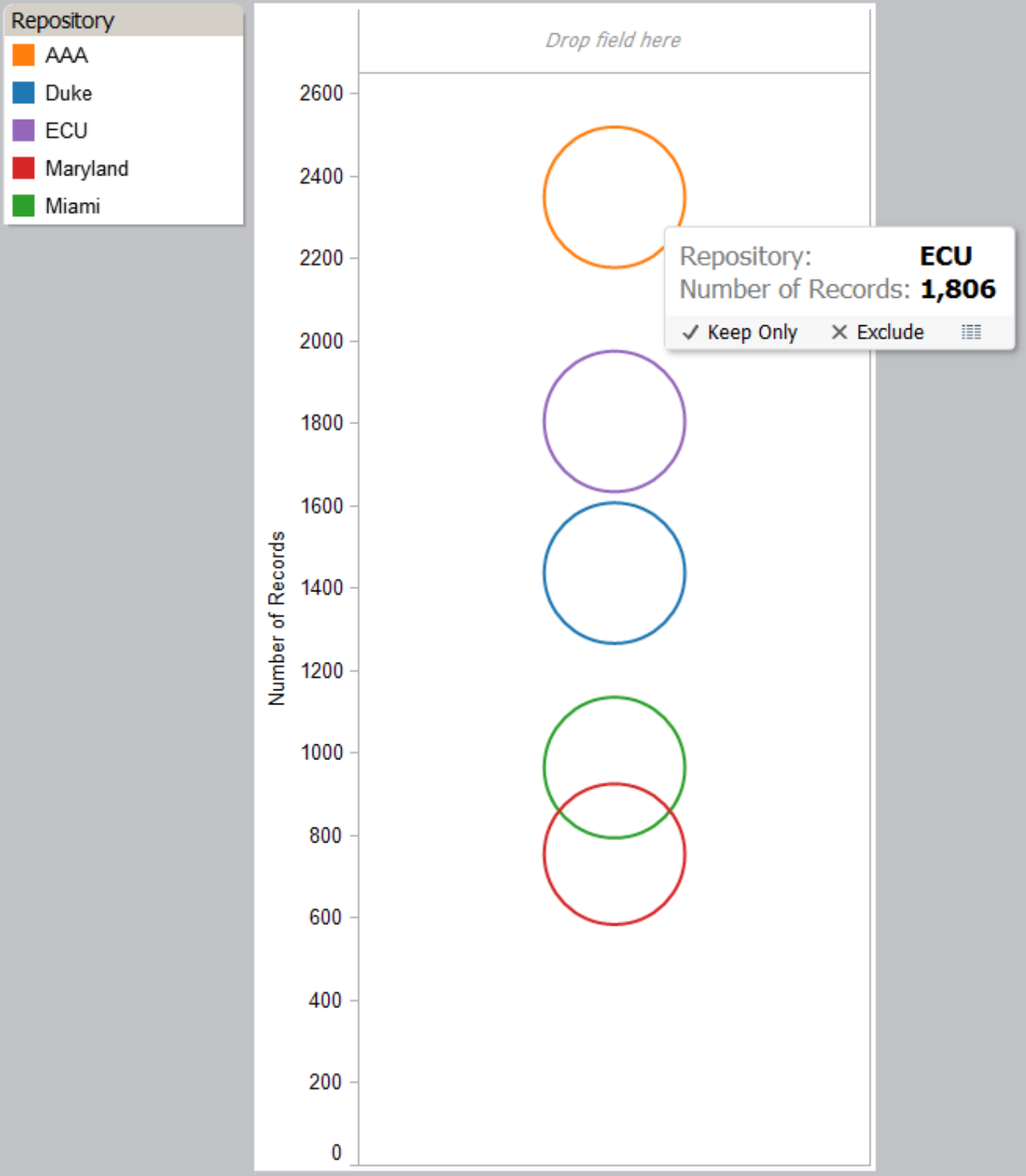
MdU.ead.scpa.0078.test

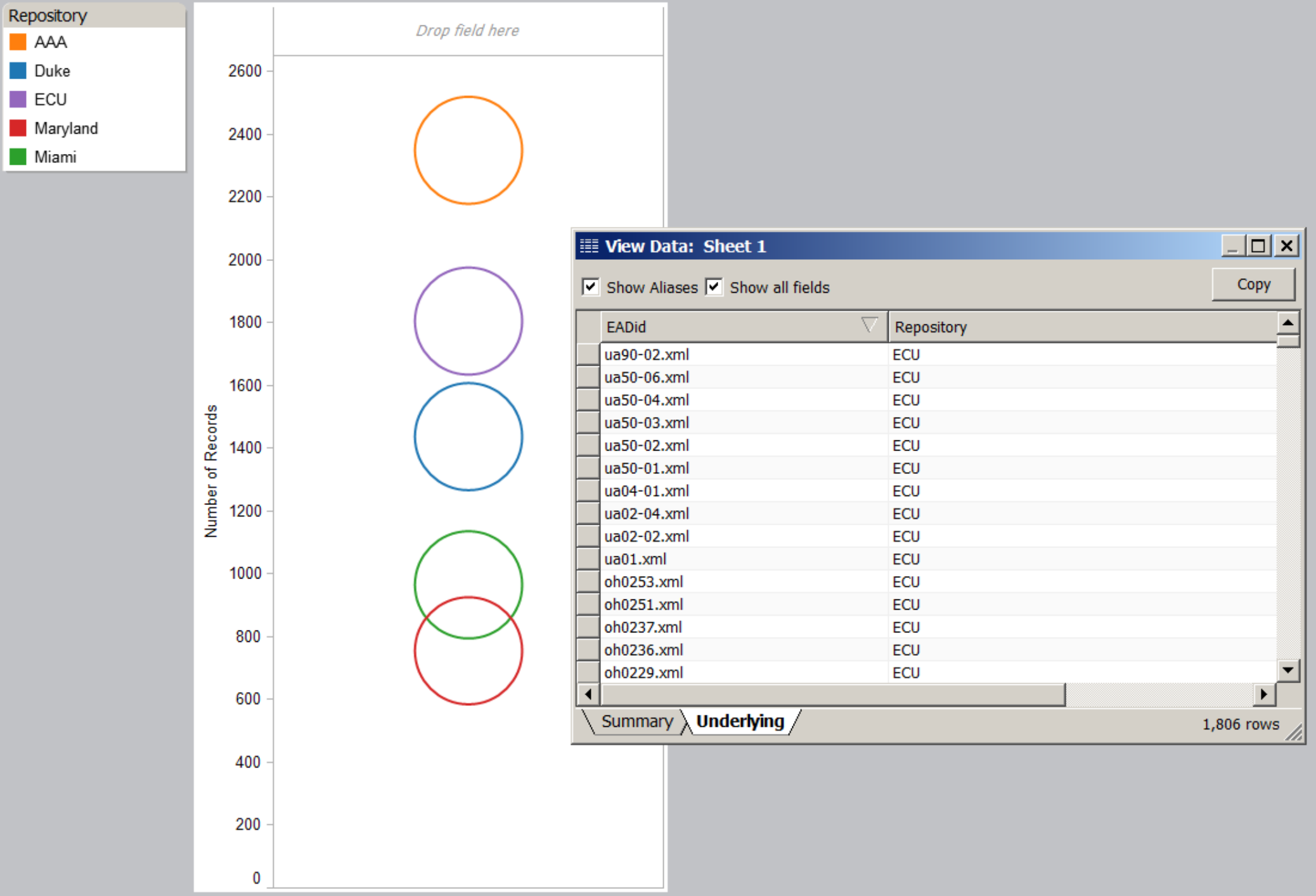
Total Rows of Data Analyzed, 2009-2010



4: Collection-level Data


Repository
AAA
Duke
ECU
Maryland
Miami
Drop field here
Number of Records
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0


Repository
AAA
Duke
ECU
Maryland
Miami
Drop field here
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
Number of Records
Repository: ECU
Number of Records: 1,806
Keep Only
Exclude

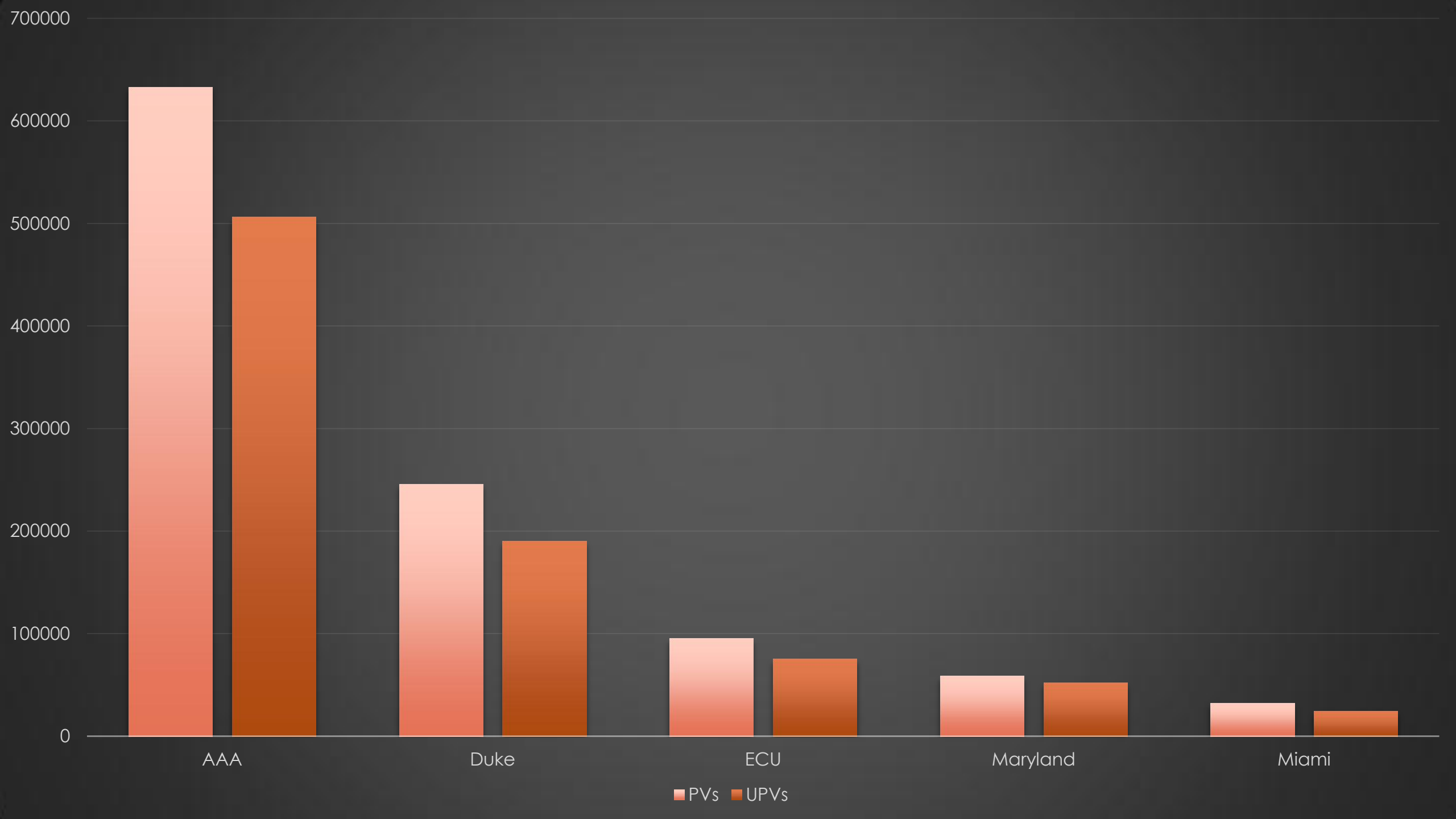


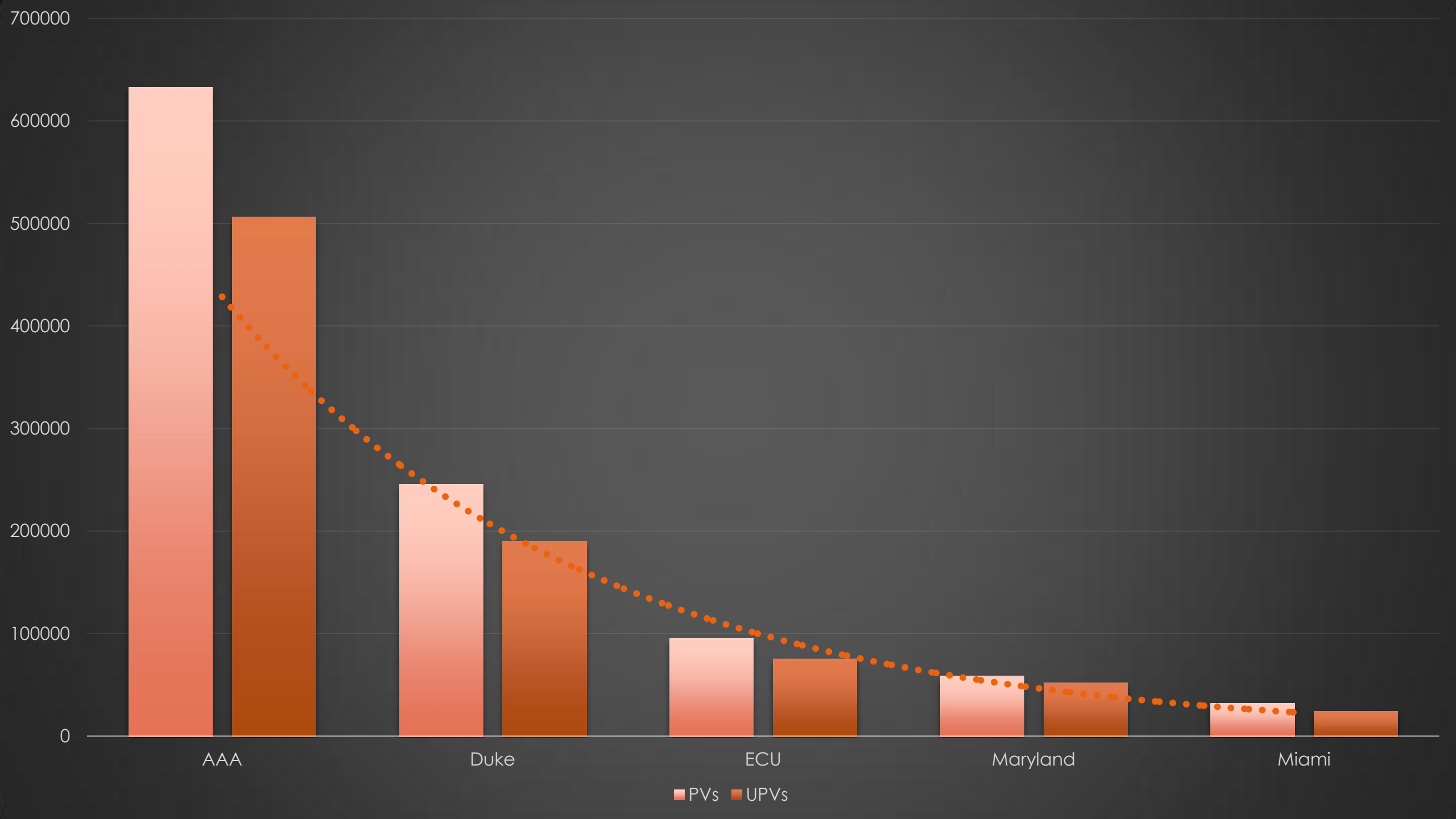
View Data: Sheet 1

Show Aliases Show all fields Copy

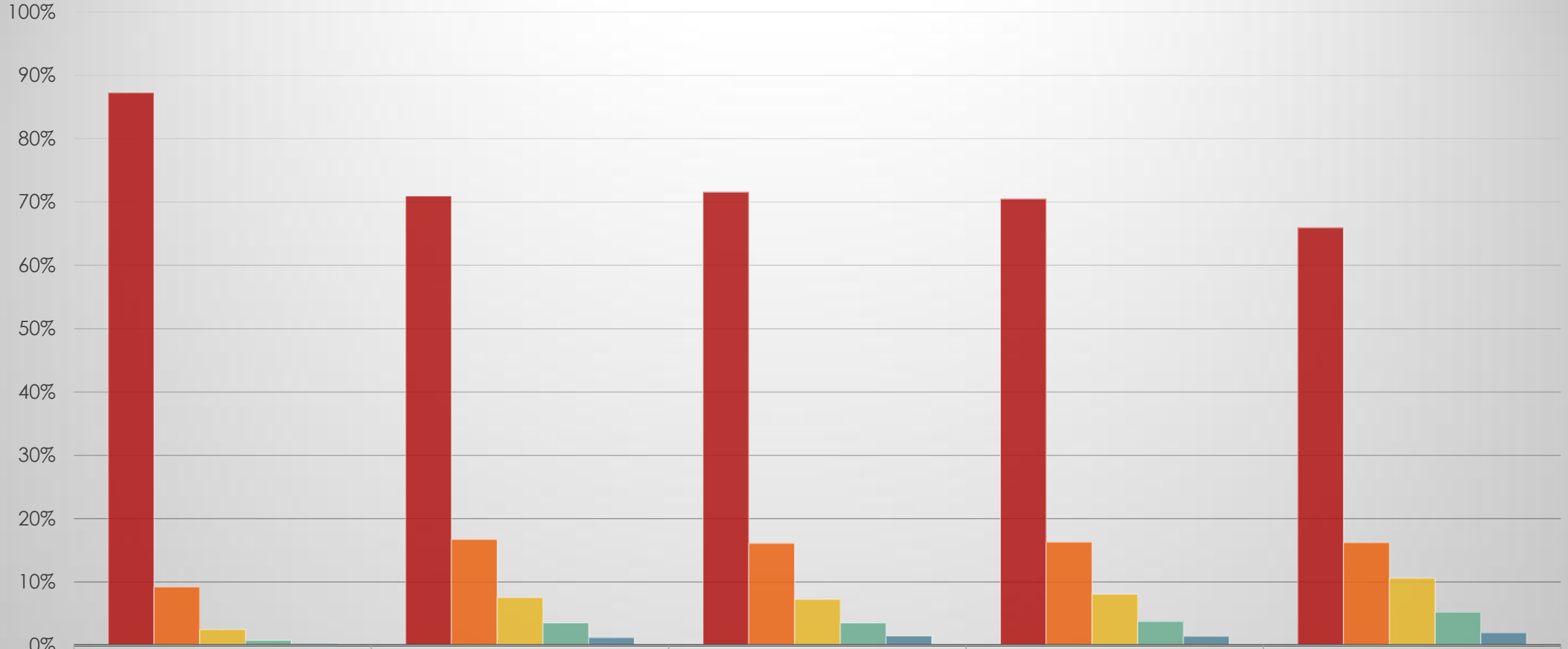
EADid	Repository
ua90-02.xml	ECU
ua50-06.xml	ECU
ua50-04.xml	ECU
ua50-03.xml	ECU
ua50-02.xml	ECU
ua50-01.xml	ECU
ua04-01.xml	ECU
ua02-04.xml	ECU
ua02-02.xml	ECU
ua01.xml	ECU
oh0253.xml	ECU
oh0251.xml	ECU
oh0237.xml	ECU
oh0236.xml	ECU
oh0229.xml	ECU

Summary Underlying 1,806 rows


700000
600000
500000
400000
300000
200000
100000
0
AAA
Duke
ECU
Maryland
Miami
PVs
UPVs

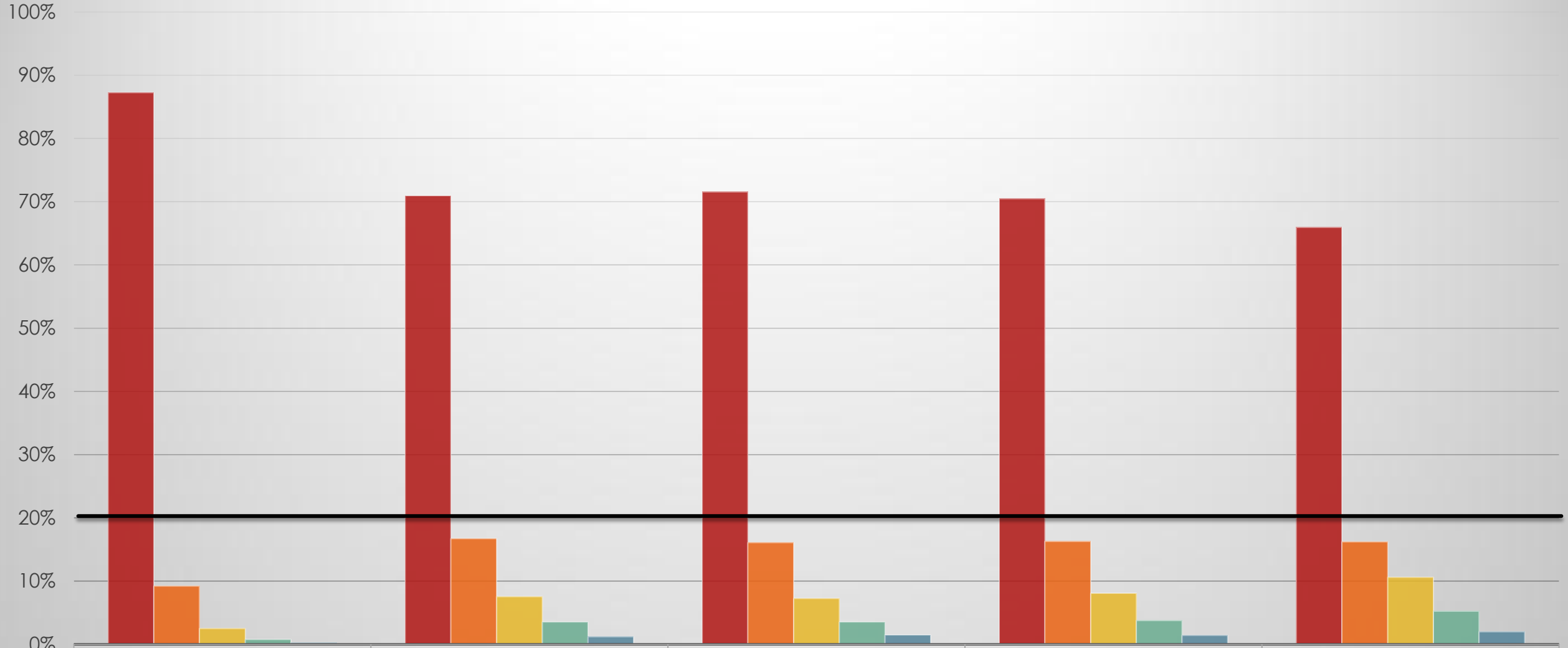

700000
600000
500000
400000
300000
200000
100000
0
AAA
Duke
ECU
Maryland
Miami
PVs
UPVs

The uneven distributions, as pictured in 5 sets of quintiles



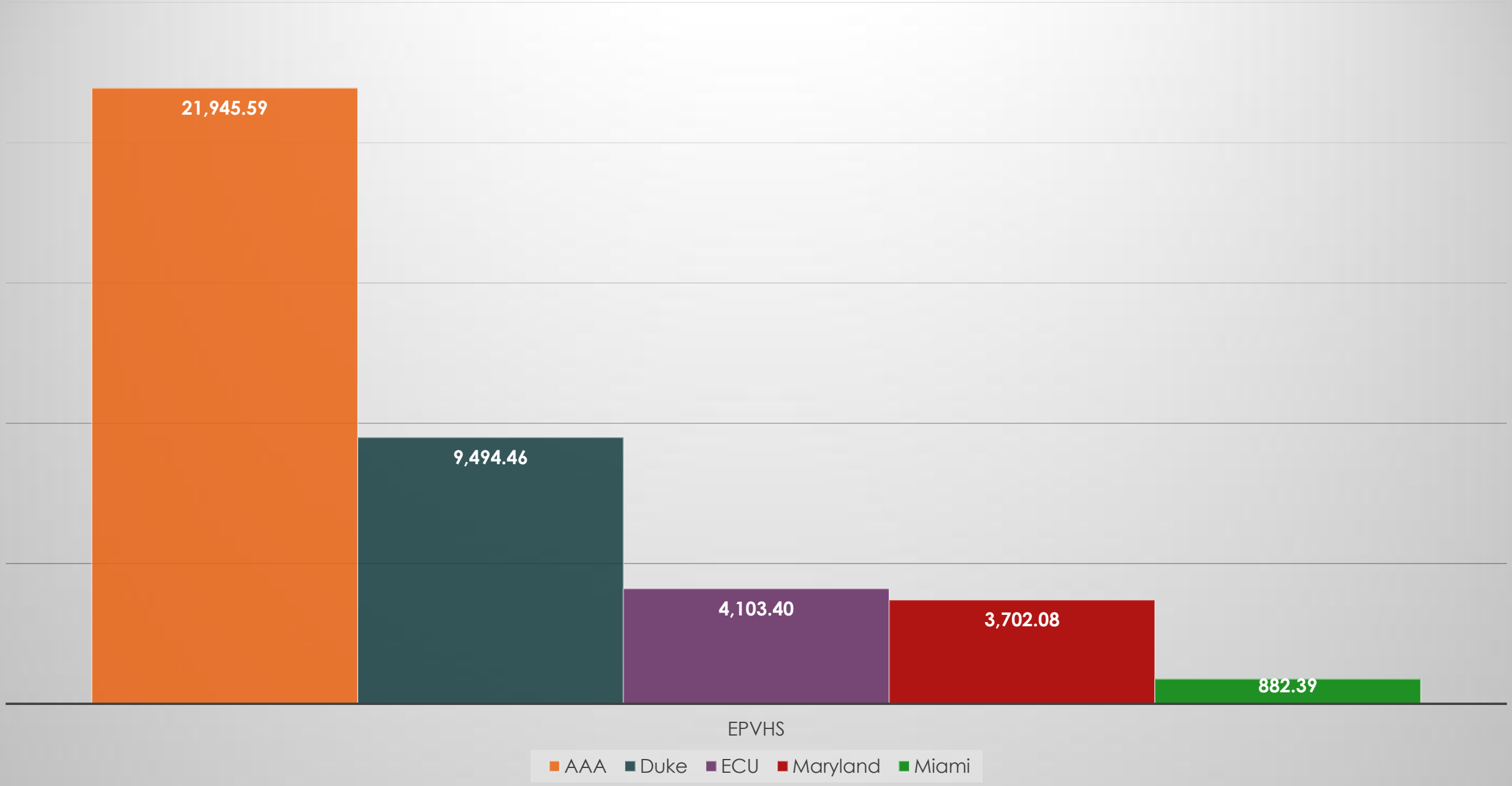
	AAA	Duke	ECU	Maryland	Miami
1st	87.27%	70.96%	71.57%	70.47%	65.93%
2nd	9.22%	16.71%	16.10%	16.28%	16.22%
3rd	2.51%	7.55%	7.26%	8.08%	10.61%
4th	0.76%	3.56%	3.56%	3.75%	5.24%
5th	0.24%	1.22%	1.51%	1.42%	2.00%

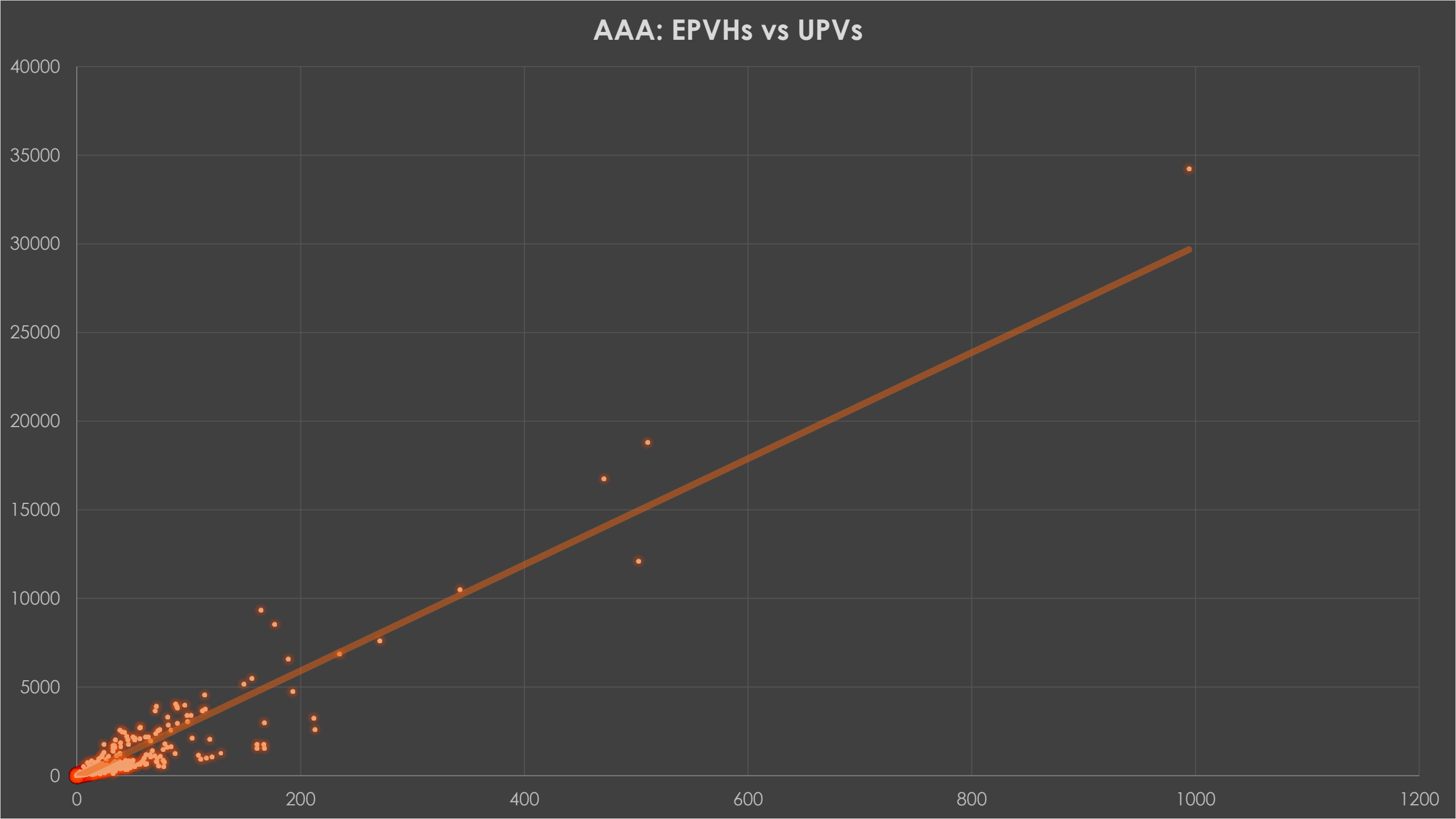
The uneven distributions, as pictured in 5 sets of quintiles

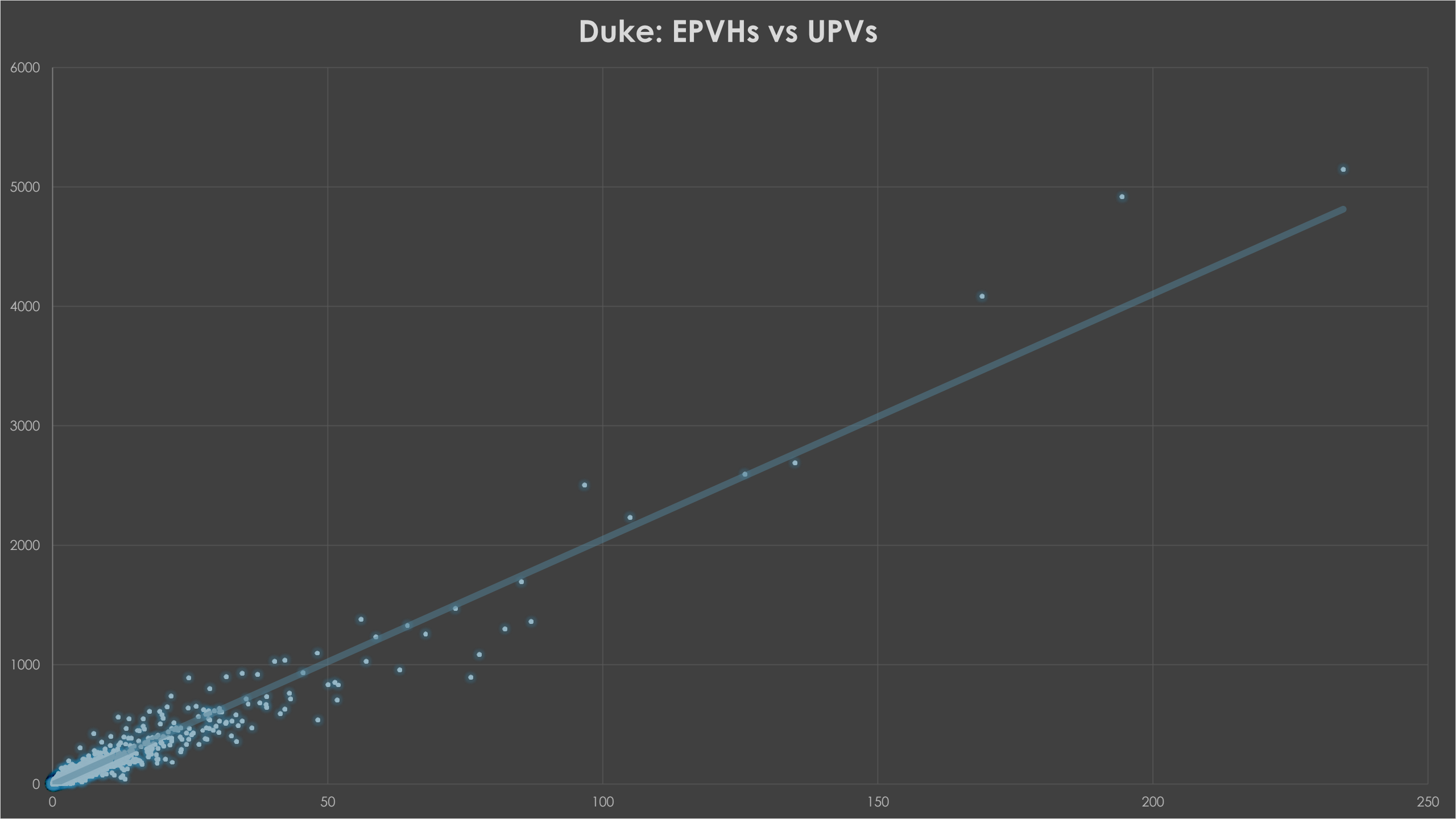


	AAA	Duke	ECU	Maryland	Miami
1st	87.27%	70.96%	71.57%	70.47%	65.93%
2nd	9.22%	16.71%	16.10%	16.28%	16.22%
3rd	2.51%	7.55%	7.26%	8.08%	10.61%
4th	0.76%	3.56%	3.56%	3.75%	5.24%
5th	0.24%	1.22%	1.51%	1.42%	2.00%

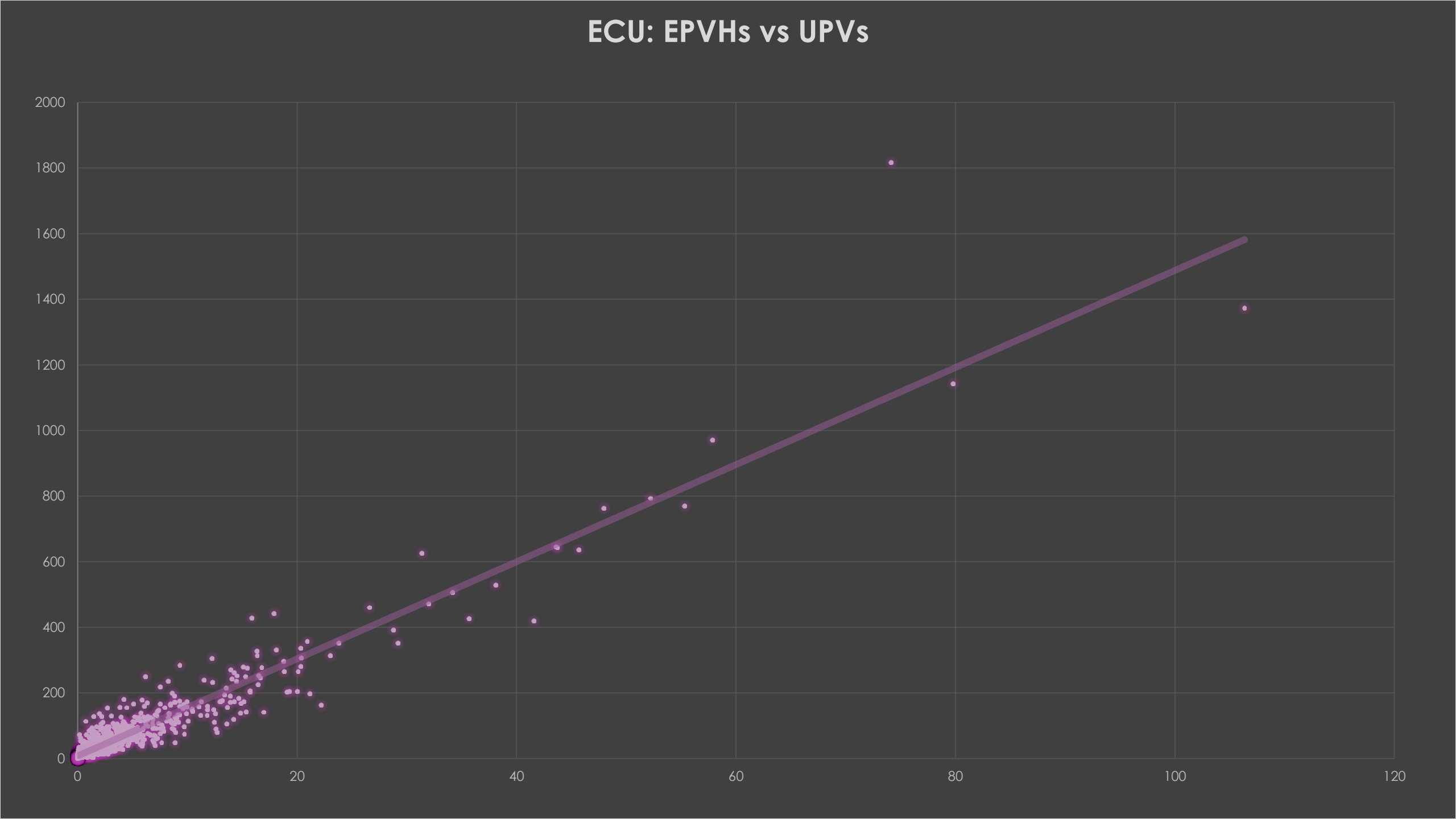
EPVHs in 2009-2010

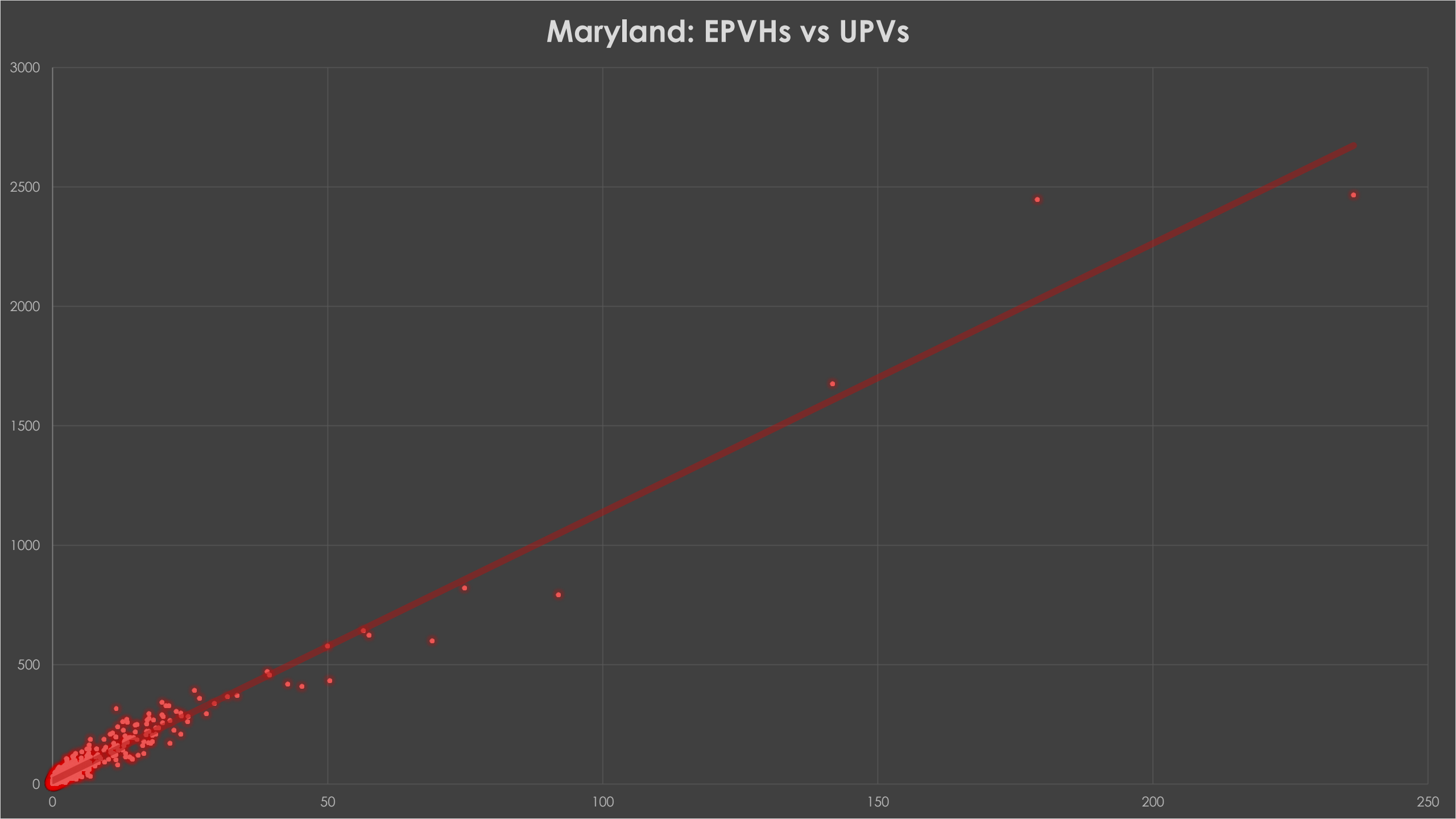


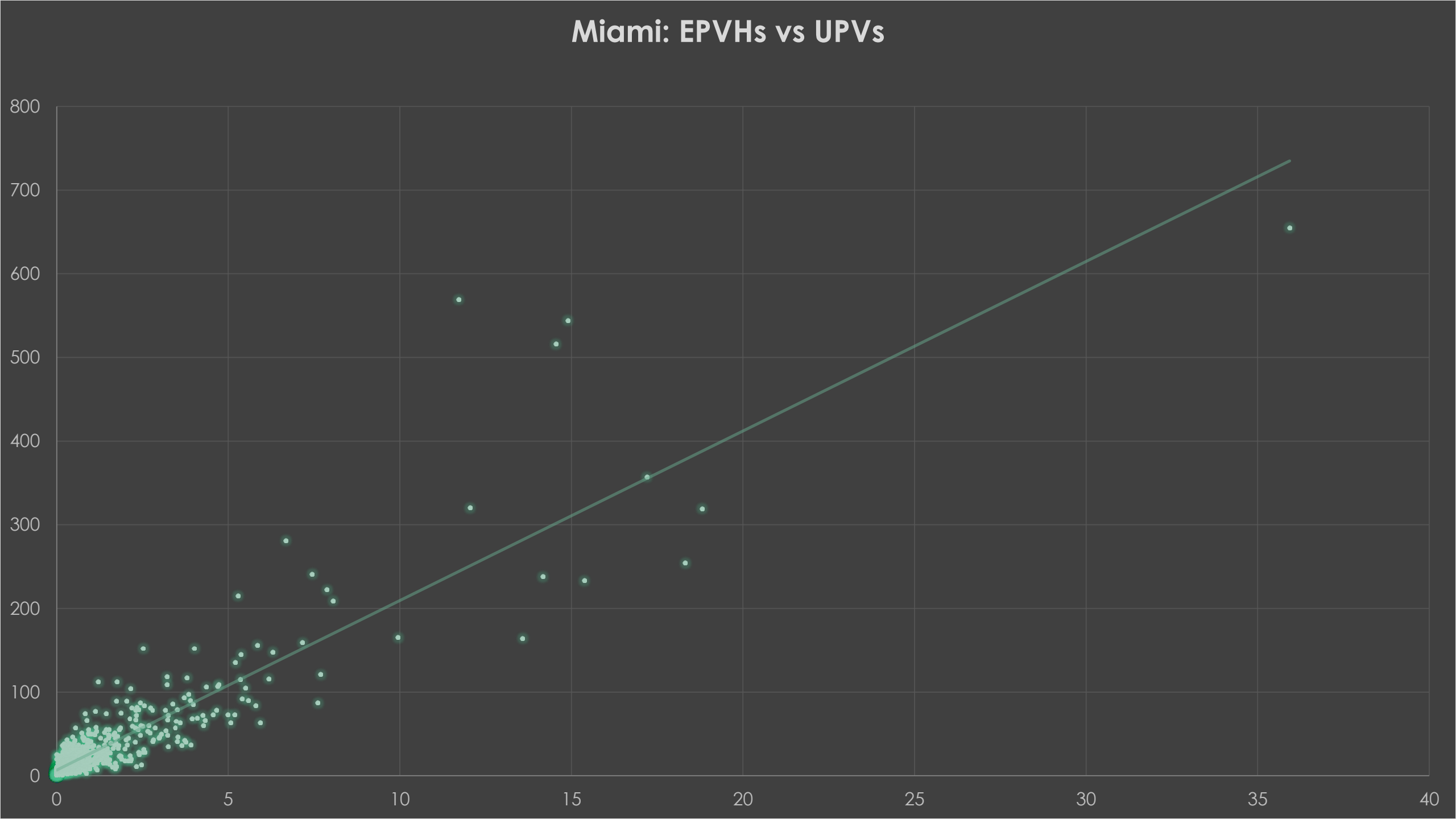

AAA: EPVHs vs UPVs
40000
35000
30000
25000
20000
15000
10000
5000
0
0
200
400
600
800
1000
1200

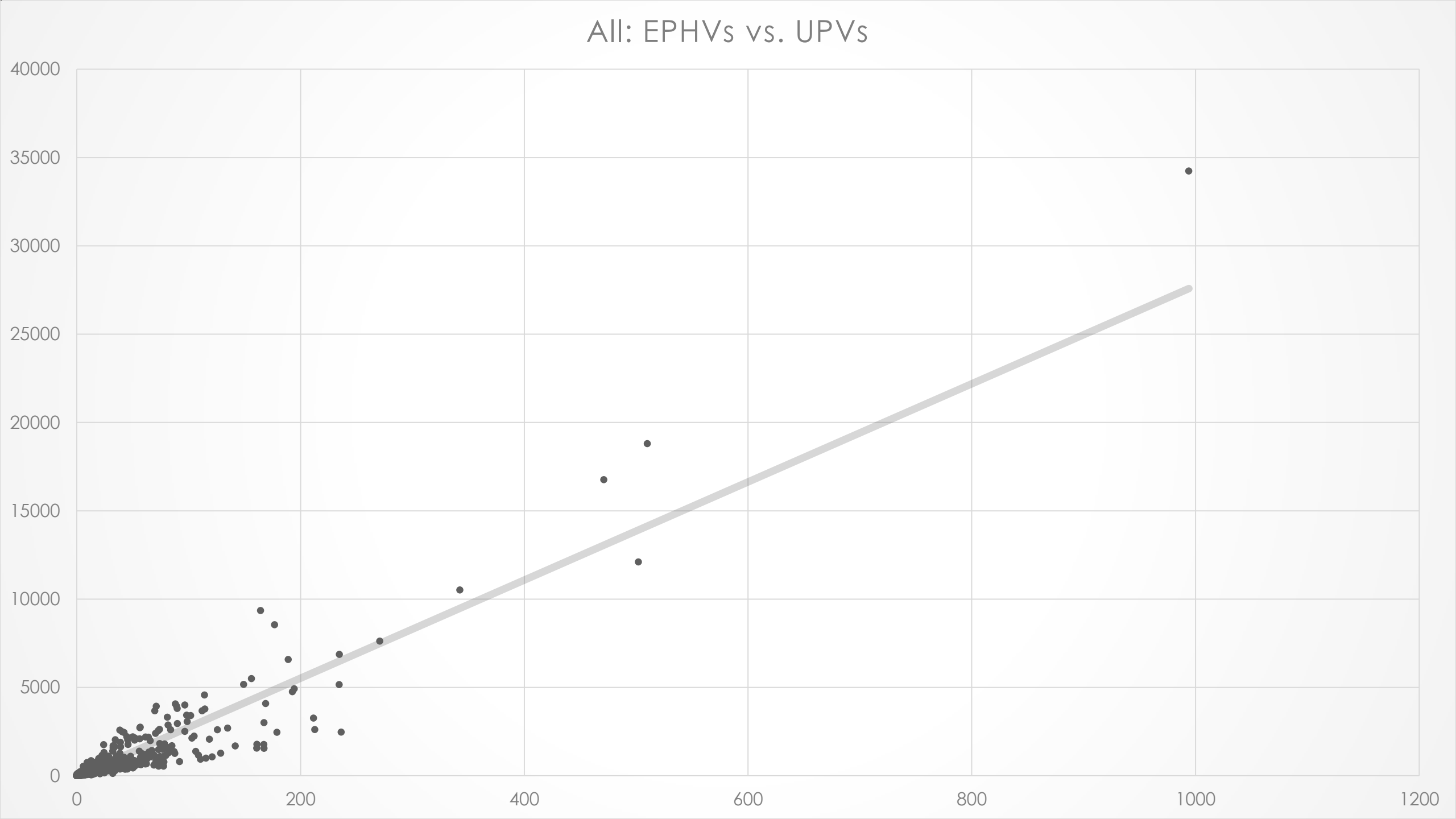

Duke: EPVHs vs UPVs
6000
5000
4000
3000
2000
1000
0
0
50
100
150
200
250

ECU: EPVHs vs UPVs

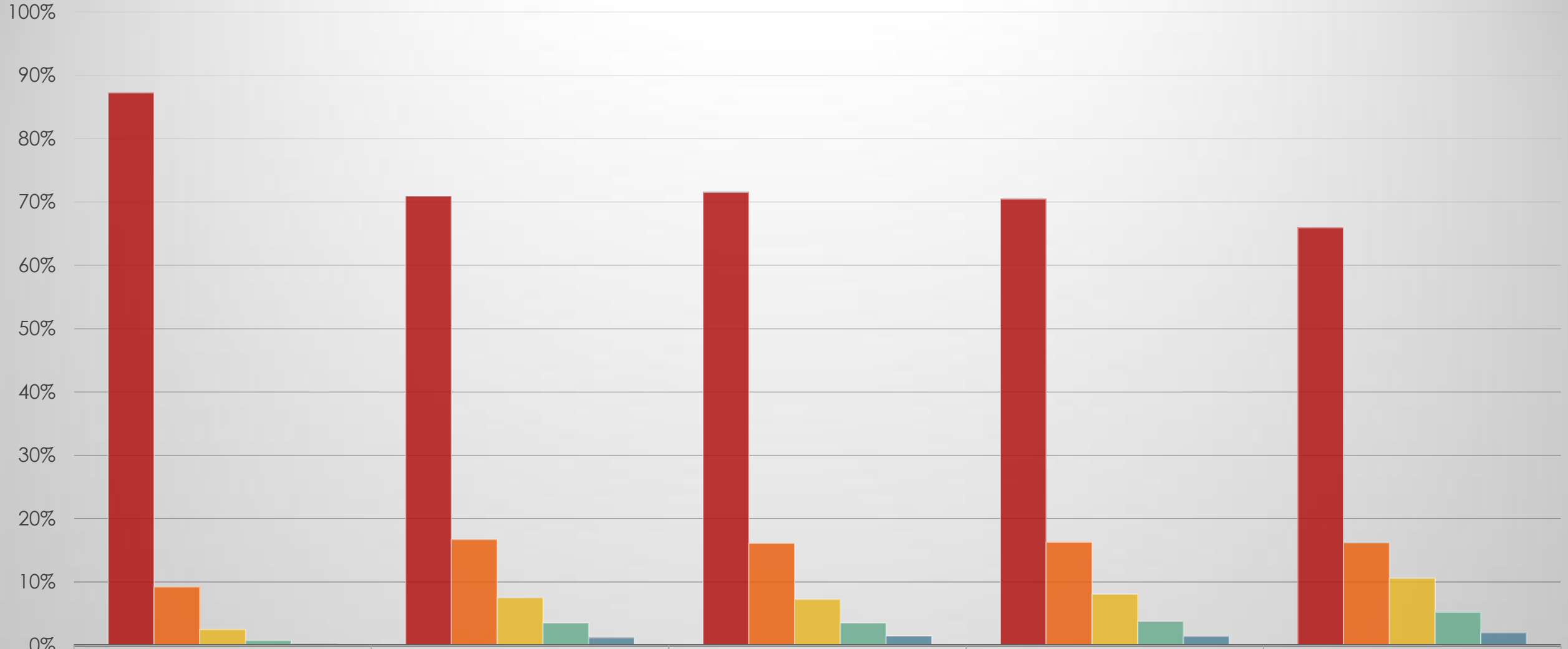



Maryland: EPVHs vs UPVs
3000
2500
2000
1500
1000
500
0
0
50
100
150
200
250

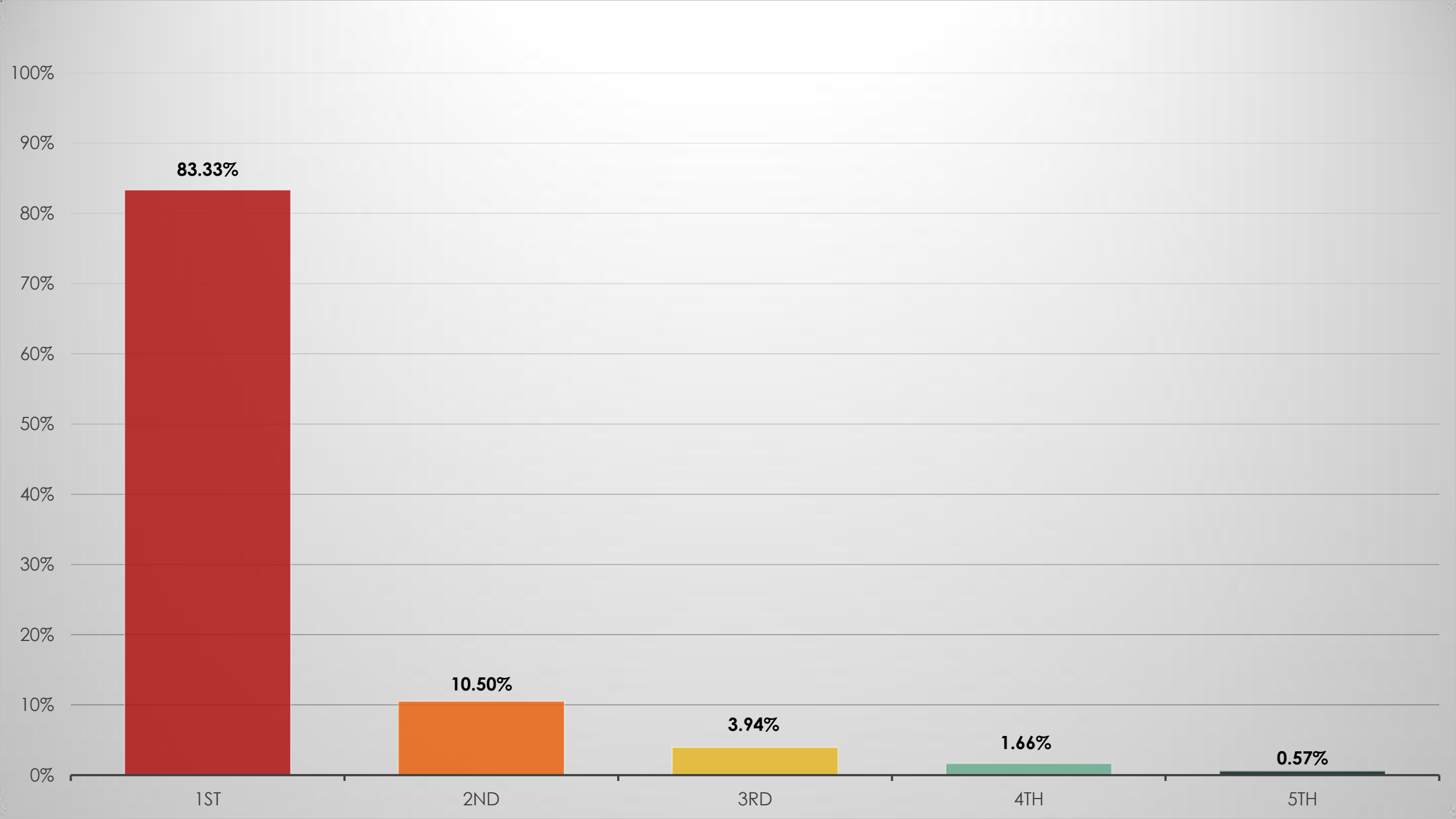

Miami: EPVHs vs UPVs
800
700
600
500
400
300
200
100
0
0
5
10
15
20
25
30
35
40


All: EPHVs vs. UPVs
40000
35000
30000
25000
20000
15000
10000
5000
0
0
200
400
600
800
1000
1200

The uneven distributions, as pictured in 5 sets of quintiles

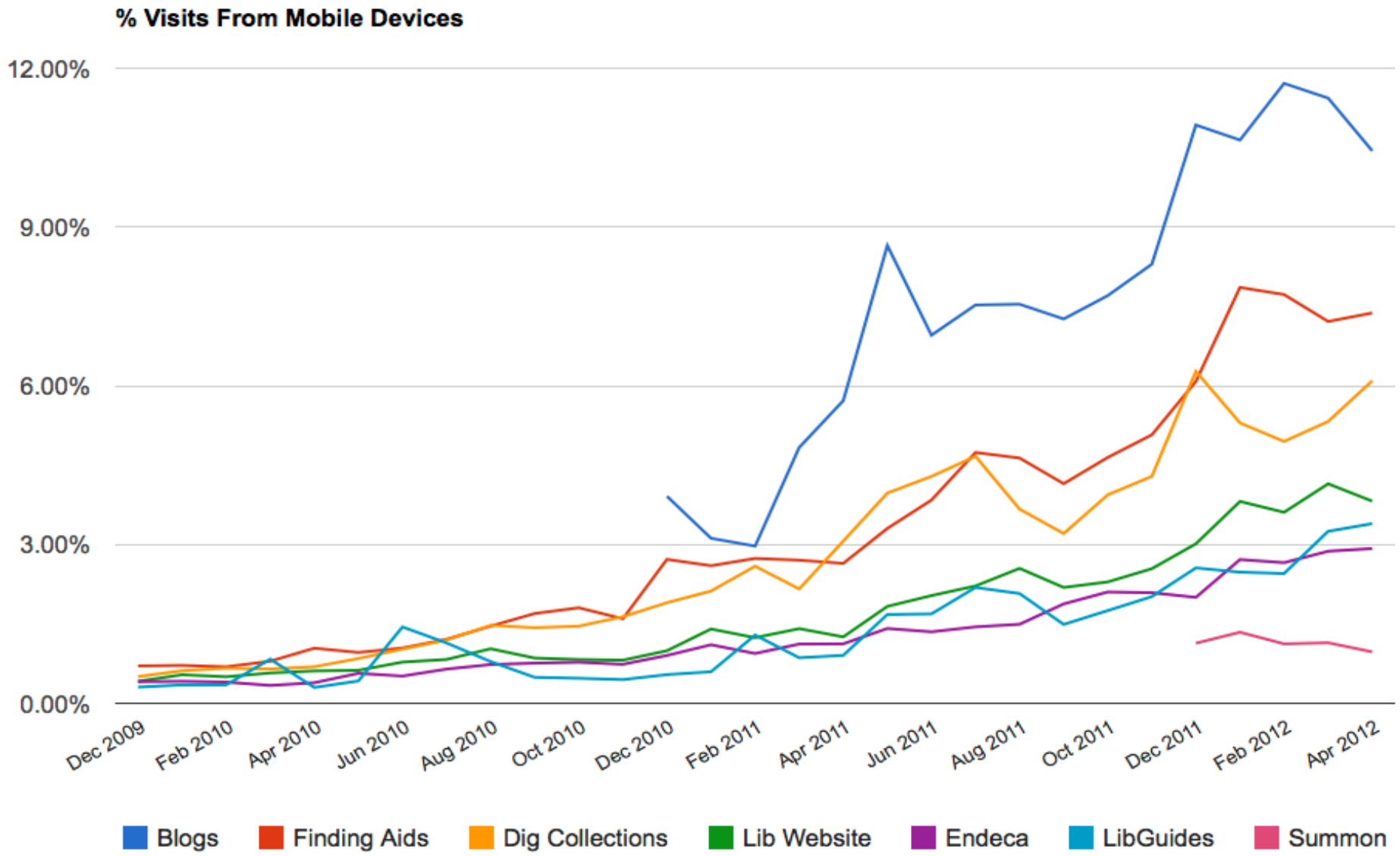


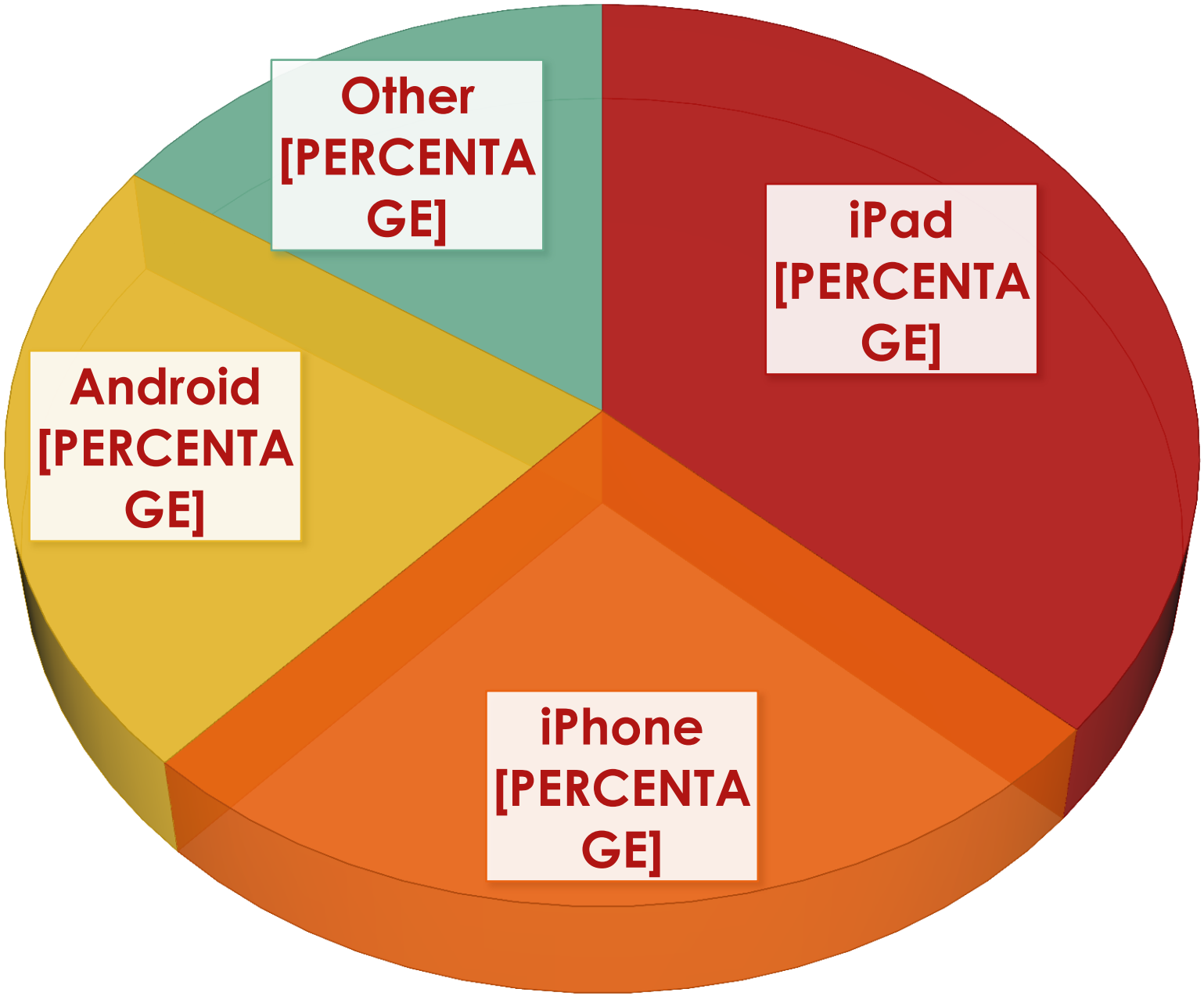
	AAA	Duke	ECU	Maryland	Miami
1st	87.27%	70.96%	71.57%	70.47%	65.93%
2nd	9.22%	16.71%	16.10%	16.28%	16.22%
3rd	2.51%	7.55%	7.26%	8.08%	10.61%
4th	0.76%	3.56%	3.56%	3.75%	5.24%
5th	0.24%	1.22%	1.51%	1.42%	2.00%


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
83.33%
10.50%
3.94%
1.66%
0.57%
1ST
2ND
3RD
4TH
5TH

5: Mobile

Mobile Traffic to Duke Library Resources
December 2009-April 2012




Other [PERCENTAGE]
iPad [PERCENTAGE]
Android [PERCENTAGE]
iPhone [PERCENTAGE]

Mobile Visit Behavior

Avg. Pages/Visit

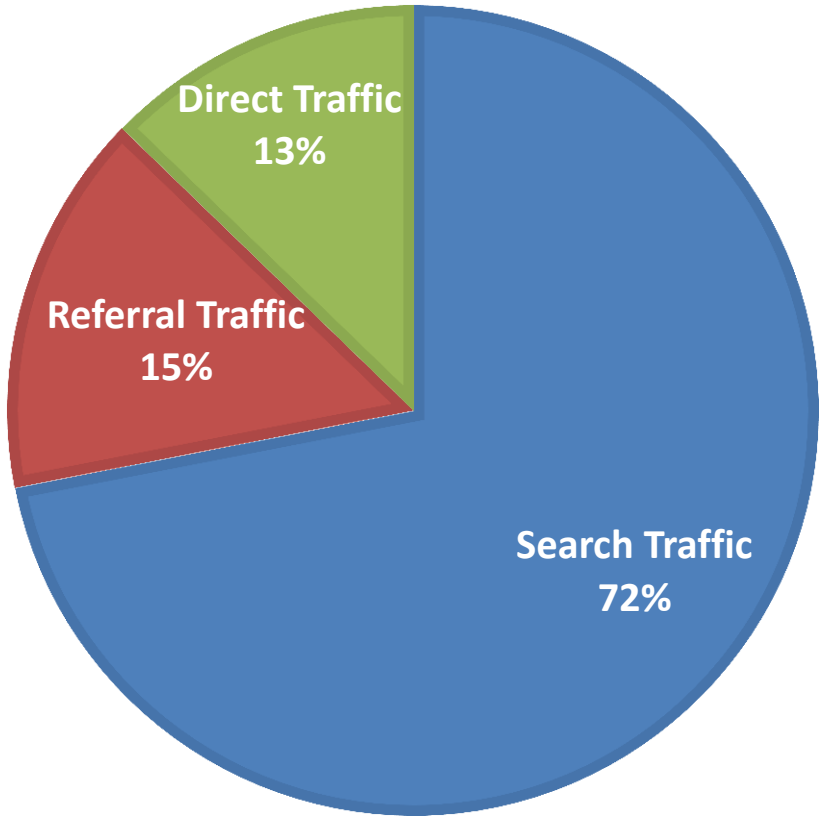
- Mobile visits - 1.83
- All visits – 3.04

Avg. Time on Site

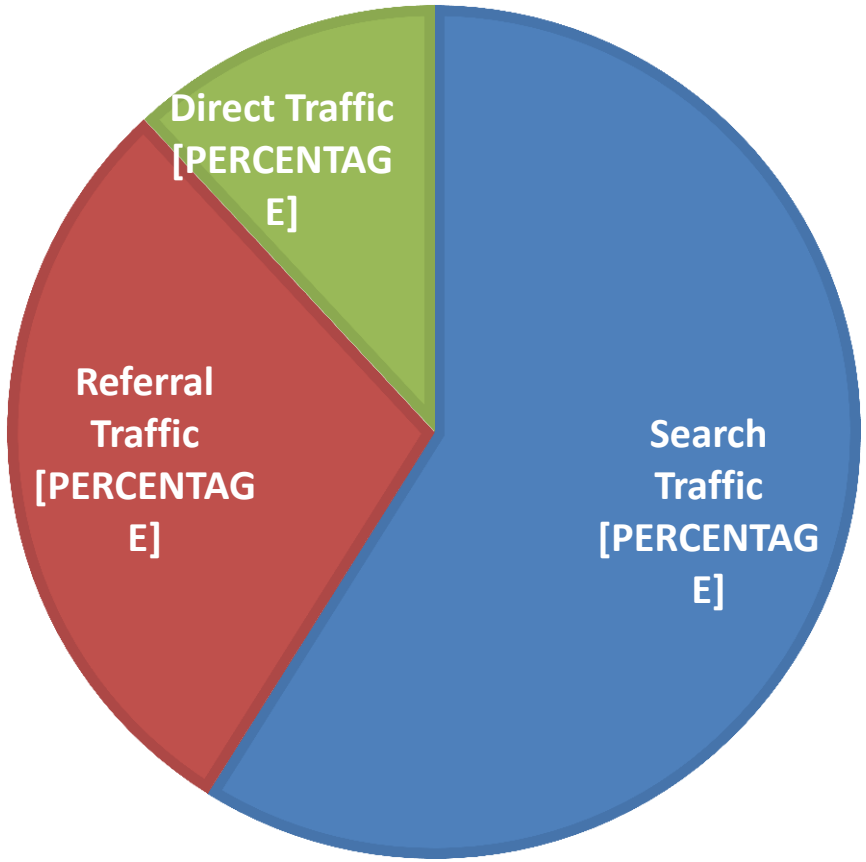
- Mobile visits - 1:07
- All visits - 2:31

From Google Analytics Data (July 1, 2011-June 30, 2012) from:
AAA, Duke University, East Carolina University, University of Maryland, and University of Miami

Traffic Sources: Mobile Visits vs. All Visits



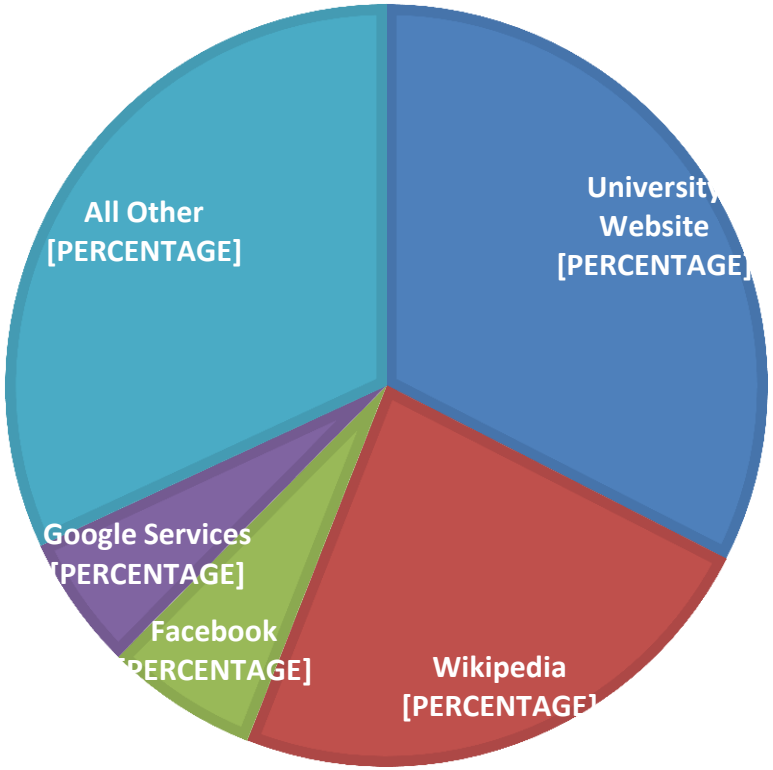
Mobile Visits



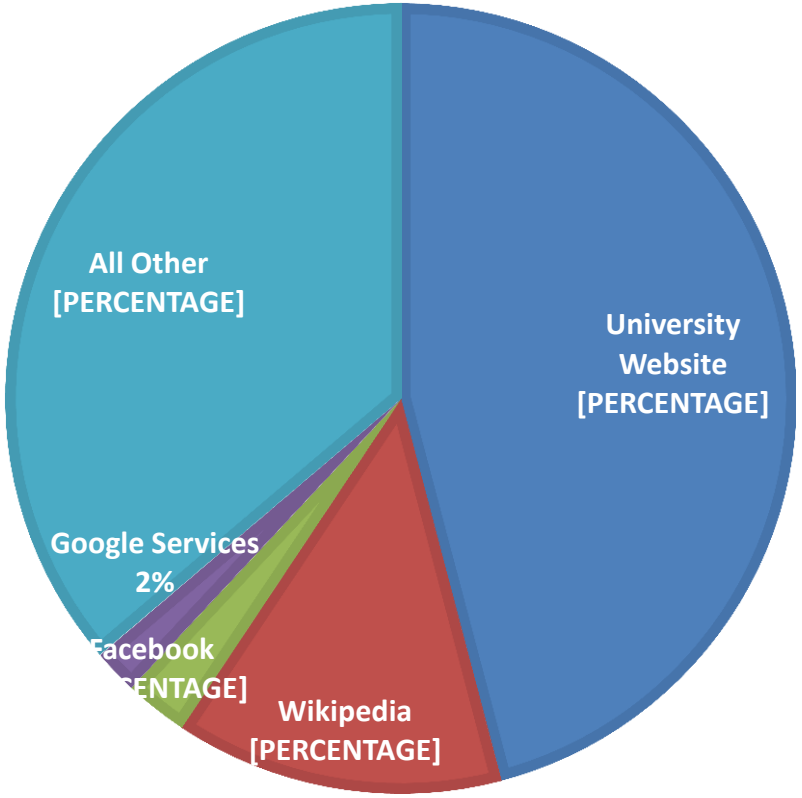
All Visits

From Google Analytics Data (July 1, 2011-June 30, 2012) from:
AAA, Duke University, East Carolina University, University of Maryland, and University of Miami

Referring Sites: Mobile Visits vs. All Visits



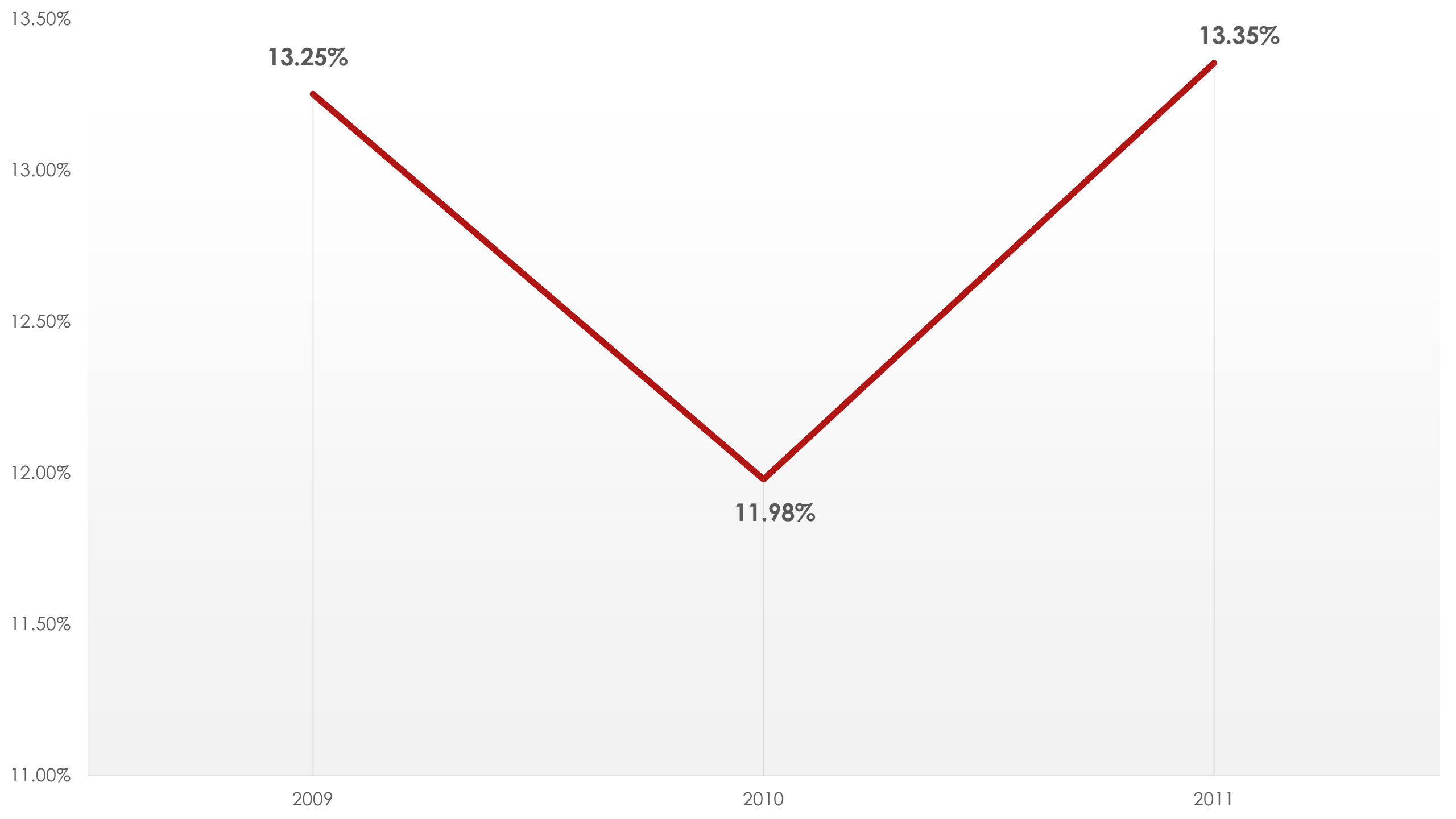
Mobile Visits – Top Referrers

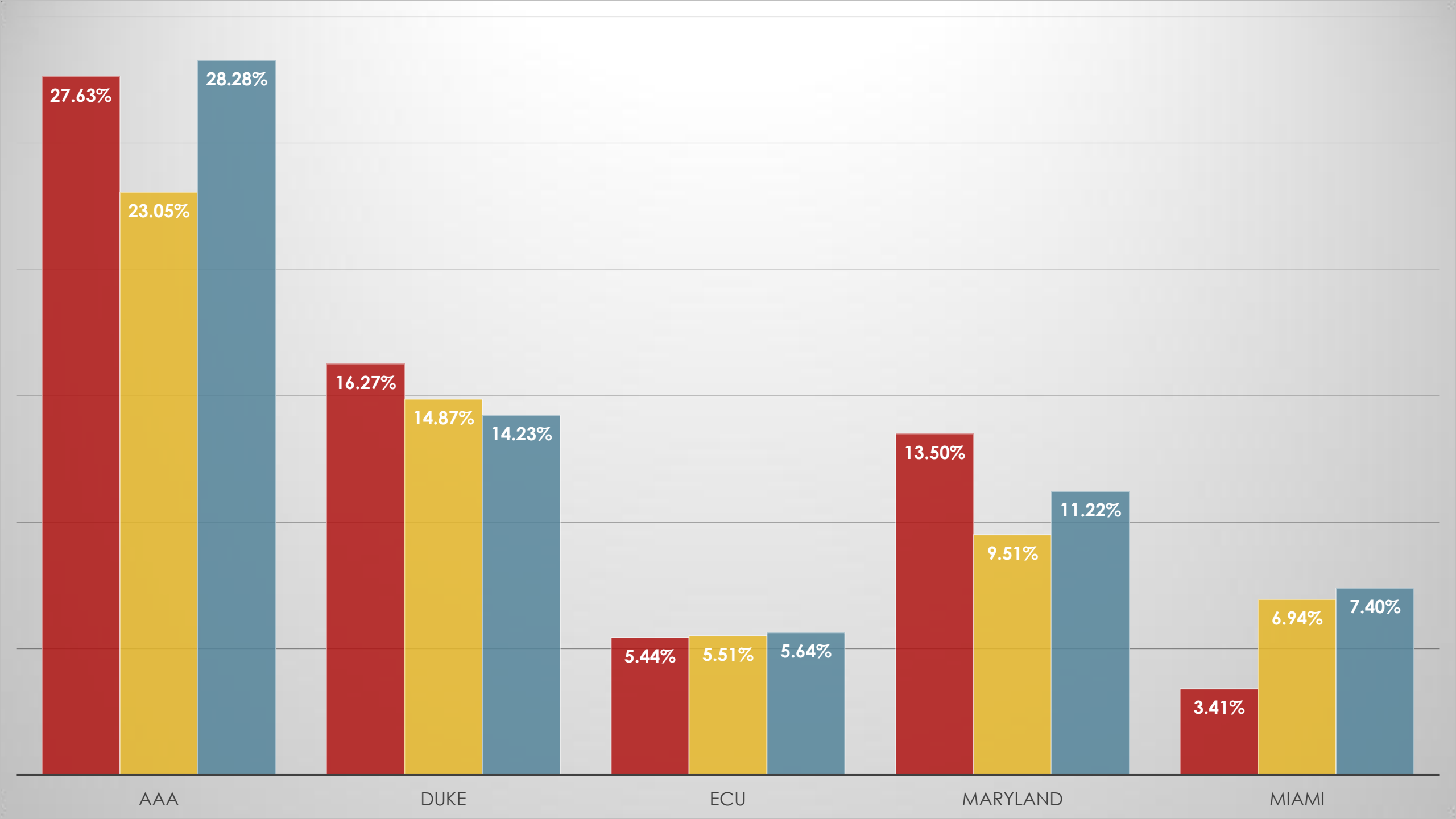


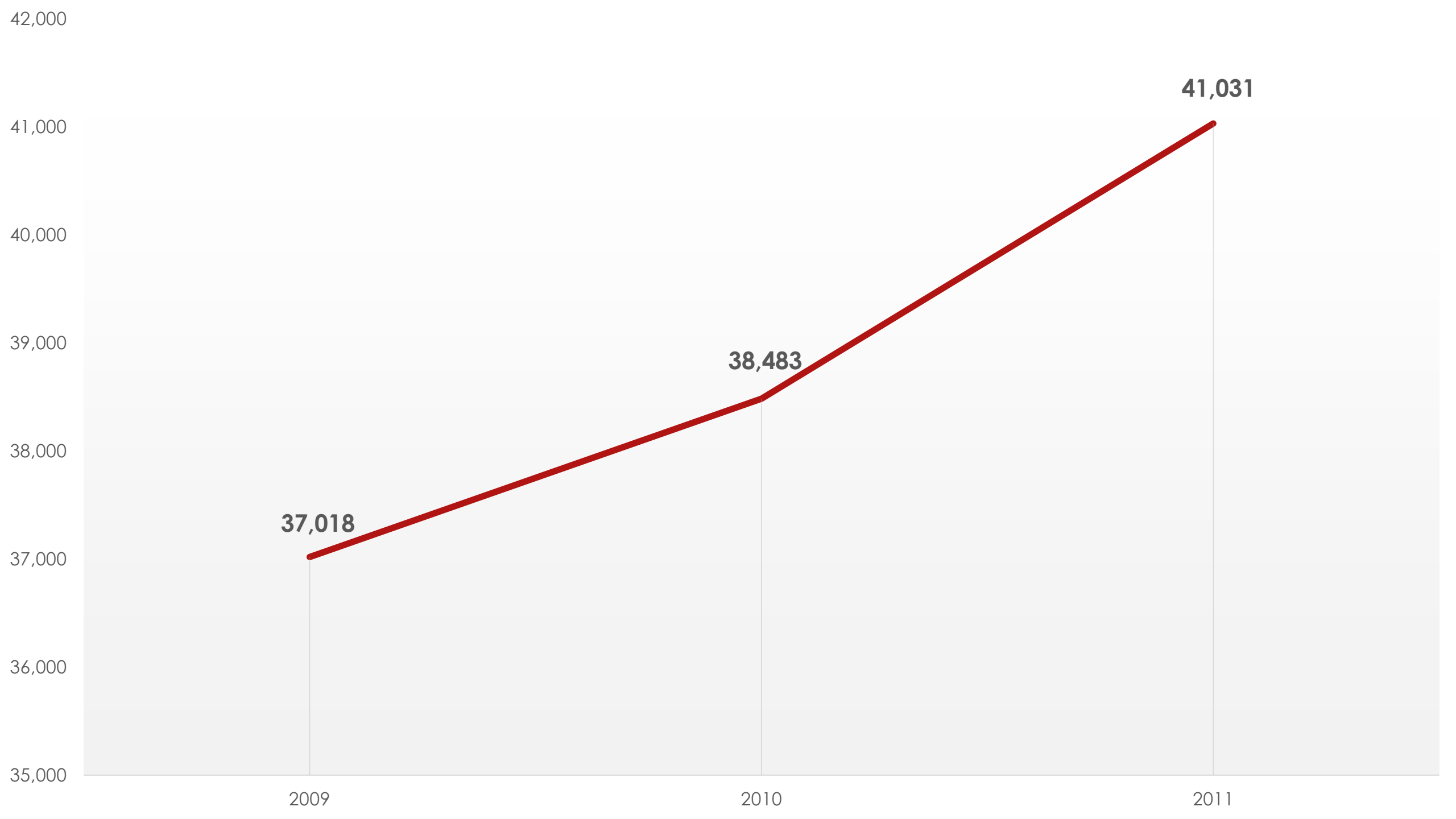
All Visits – Top Referrers

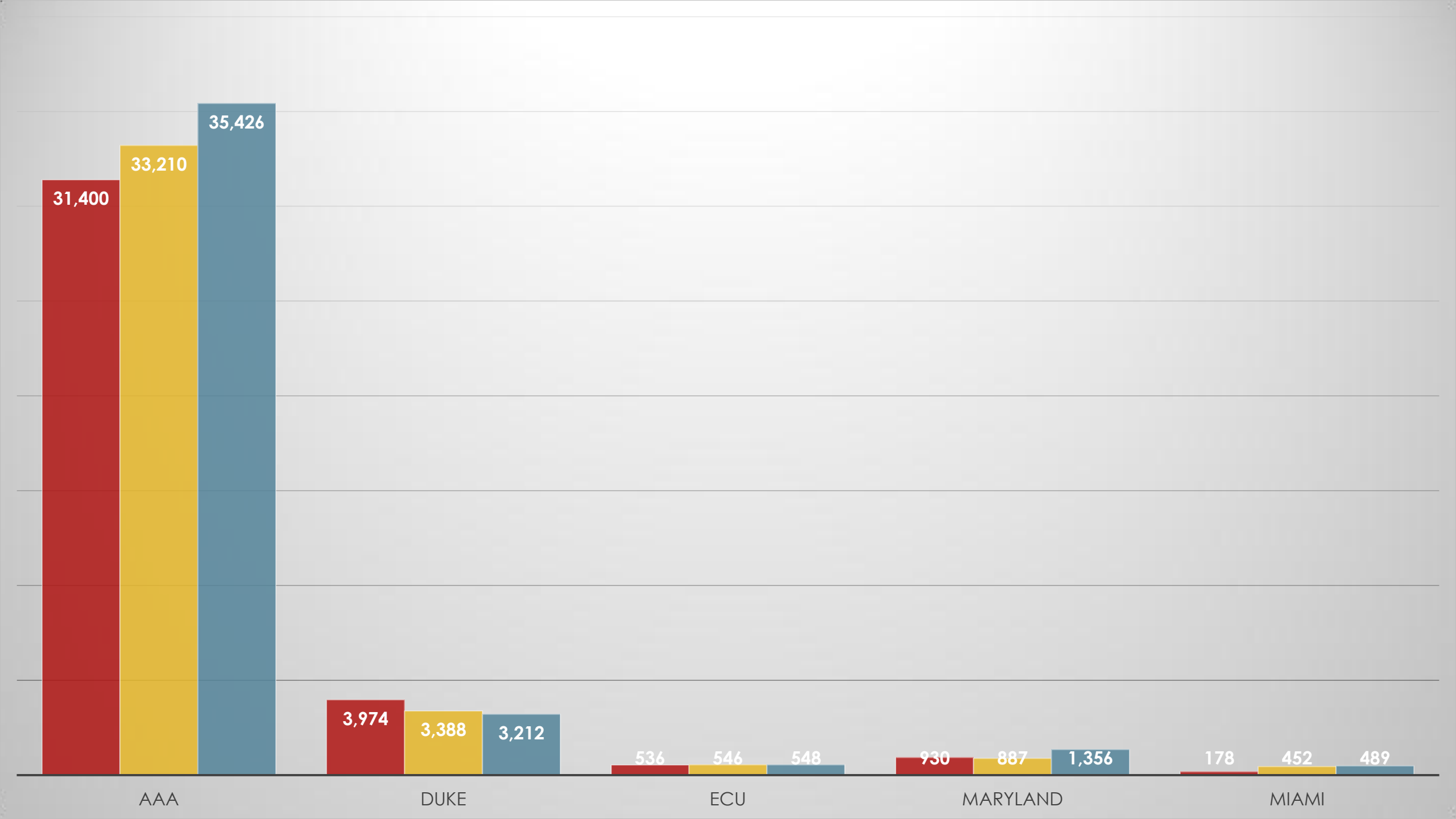
From Google Analytics Data (July 1, 2011-June 30, 2012) from:
AAA, Duke University, East Carolina University, University of Maryland, and University of Miami

6: Wikipedia


13.50%
13.00%
12.50%
12.00%
11.50%
11.00%
13.25%
11.98%
13.35%
2009
2010
2011


27.63%
23.05%
28.28%
16.27%
14.87%
14.23%
5.44%
5.51%
5.64%
13.50%
9.51%
11.22%
3.41%
6.94%
7.40%
AAA
DUKE
ECU
MARYLAND
MIAMI


42,000
41,000
40,000
39,000
38,000
37,000
36,000
35,000
37,018
38,483
41,031
2009
2010
2011


31,400
33,210
35,426
3,974
3,388
3,212
536
546
548
930
887
1,356
178
452
489
AAA
DUKE
ECU
MARYLAND
MIAMI

8: Conclusion → Next Steps

- How best to define a collection-level page? Should we?
- Which metrics are most useful for archivists, researchers, etc.?
- Beyond the collection, how can we analyze these data sets by subject / topic?
- How best to share this data?
- How else can it be analyzed?

8: Conclusion → Next Steps

- How best to define a collection-level page? Should we?
- Which metrics are most useful for archivists, researchers, etc.?

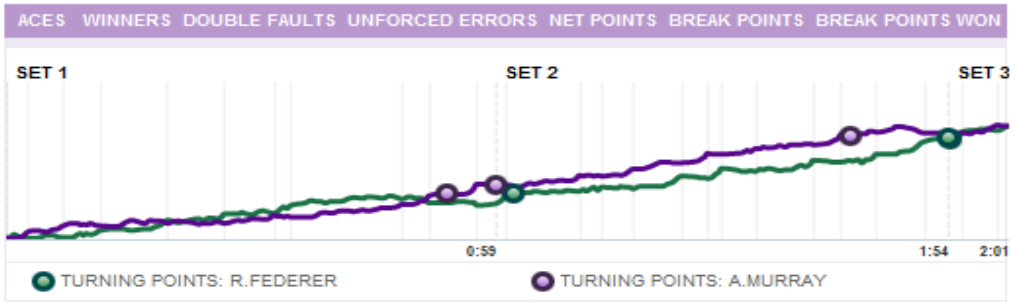
  My hunch:
  - UPVs
  - EPVHs
  - Reading Room Hours
  - Reference Consultations
  - And?
- Beyond the collection, how can we analyze these data sets by subject / topic?
- How best to share this data?
- How else can it be analyzed?

Roger Federer (SUI)		Andy Murray (GBR)
6	Aces	7
2	Double faults	0
58 of 84 = 69 %	1st serves in	55 of 87 = 63 %
43 of 58 = 74 %	1st serve points won	41 of 55 = 75 %
11 of 26 = 42 %	2nd serve points won	14 of 32 = 44 %
130 MPH	Fastest serve	133 MPH
115 MPH	Average 1st serve speed	120 MPH
98 MPH	Average 2nd serve speed	87 MPH
36 of 47 = 77 %	Net points won	18 of 26 = 69 %
2 of 5 = 40 %	Break points won	2 of 6 = 33 %
32 of 87 = 37 %	Receiving points won	30 of 84 = 36 %
34	Winners	26
25	Unforced errors	10
86	Total points won	85

KEYS TO THE MATCH

Roger Federer	Andy Murray
1. Win more than 55% of 3 to 8 shot rallies. ✓ 55%	1. Have an Aggressive Ratio of more than 1.6. ✓ 1.6
2. Win more than 57% of points after a rally of 9 or more strokes. ✗ 57%	2. Average more than 7.2 points per game returning. ✗ 7.2
3. Win more than 26% of first serve return points. ✗ 26%	3. Keep first serve percentage more than 68%. ✗ 68%

MOMENTUM



8: Conclusion → Next Steps

- How best to define a collection-level page? Should we?
- Which metrics are most useful for archivists, researchers, etc.?
- Beyond the collection group, how can we analyze these data sets by subject / topic?
- How best to share this data?
- How else can it be analyzed?