DATA MINING HOMEWORK 1

GROUP I: Saurabh Annadate, Tian Fu, Max Holiber, Molly Srour, Anjali Verma

Data Dictionary

Field	**Definition**
fire_fly_id	Individual Customer ID
visit_num	Individual Visit ID
page_views	No of pages viewed for that visit
content_type	Describes the type of content a user viewed (article, photo gallery, etc)
section	Topic or general grouping of content such as sports or news
sub_section	Drill down of section
topic	Drill down of subsection
gup_anon_id	Global identifier to link all visitor behavior
browser_value	Type of browser used for visit
referrer_type_value	Referral source of topic
author	Journalist name
zip	Zip Code of the customer during visit
event_date	Date of record

Question 1. Exploratory Data Analysis

Total no. of records: 1,024,613

Customer Information:

- No. of unique customers: 5000
- Average number of visits per customer: 40
- 1781 subscribers only made one visit
- Average number of page views: 16.8

Missing Values:

- 540 observations missing sub_section and topic
- 56 observations missing gup_anon_id
- 22,051 observations missing content_type
- 520,313 observations missing author
- 566 observations missing zip

Event Dates:

- Most recent event date is on 2018-08-05
- Earliest event date is on 2016-06-30

Questions/Issues with the data:

1) Zip codes are mixed. There are zip codes as combinations of letters and numbers. There are also zip codes as 0 - we interpret is same as NA
2) visit_num for each subscriber is not consecutive
3) author column is mixed with dates, not only NA and names
4) There are cases that for the same visit_num of a particular person, the event_dates for different sections differ

Question 2. Feature Creation

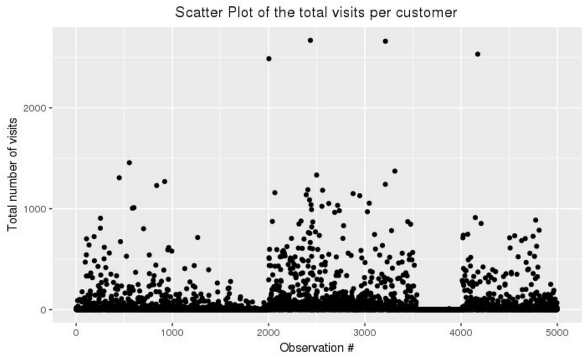
We utilized the features provided in the raw data set to create several engineered features that we deemed relevant to our clustering analysis. For many of these variables, we used visit day as one unit rather than visit number, since in this dataset visits sometimes spanned multiple days.

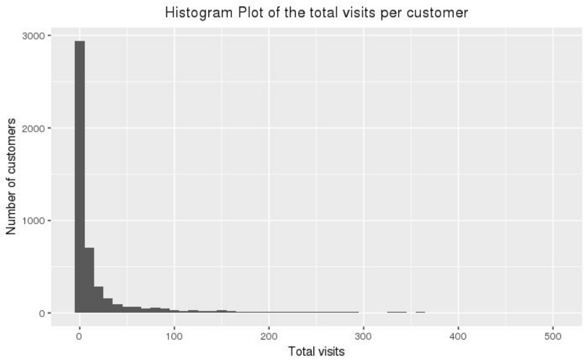
The variables we used in our analysis are detailed below:

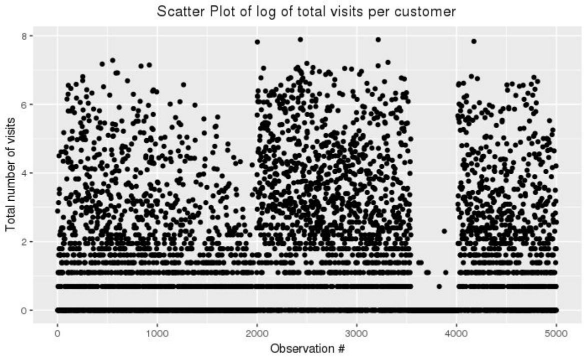
tot_visit (total number of visit days per customer)

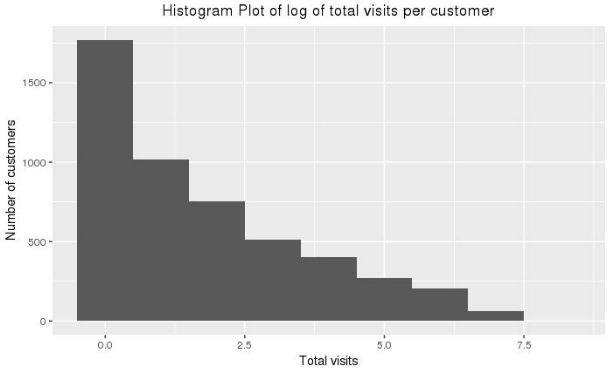
- Definition: Total number of visit days that a customer makes
- Transformation: Log
- Method of calculation: For each user, count the distinct number of visit days
- Summary:
    - Min. : 1.00
    - 1st Qu.: 1.00
    - Median : 3.00
    - Mean : 47.01
    - 3rd Qu.: 18.00
    - Max. : 2,669.00

- Plots:





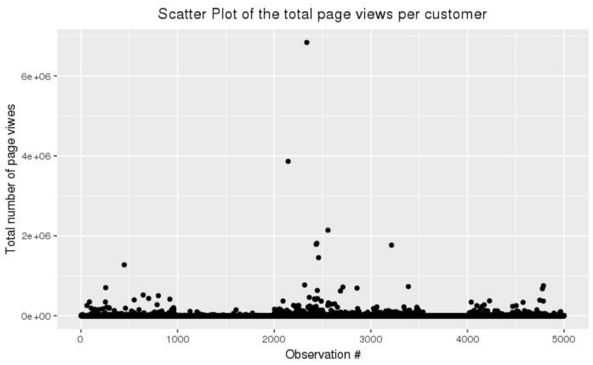


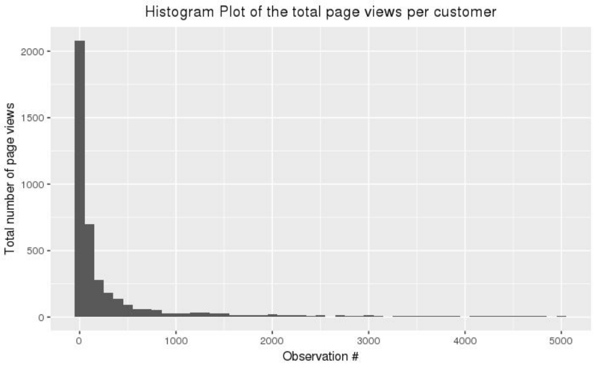


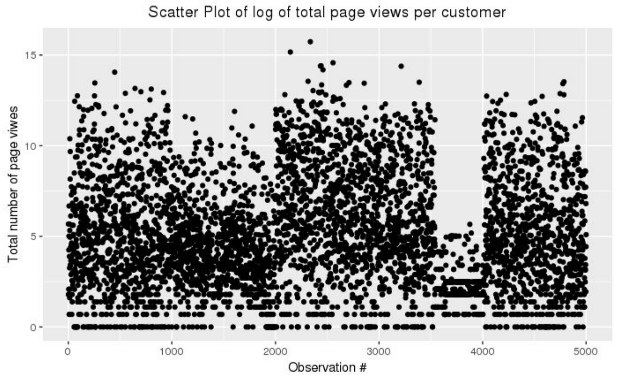
tot_page_views

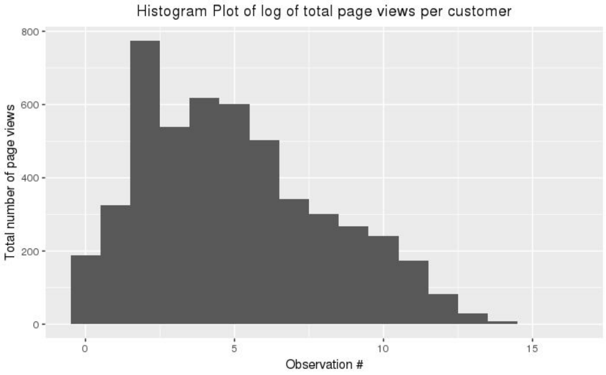
- Definition: Total number of pages viewed by a customer
- Transformation: Log
- Method of calculation: For each user, sum the number of page views
- Summary:
    - Min.   :   1
    - 1st Qu.:   12
    - Median :   98
    - Mean   : 13,848
    - 3rd Qu.:   1,177
    - Max.   : 6,836,554

- Plots:



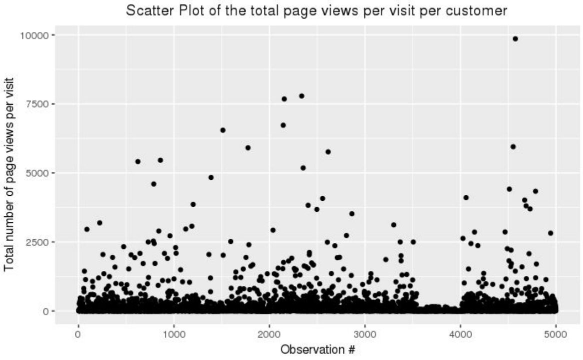


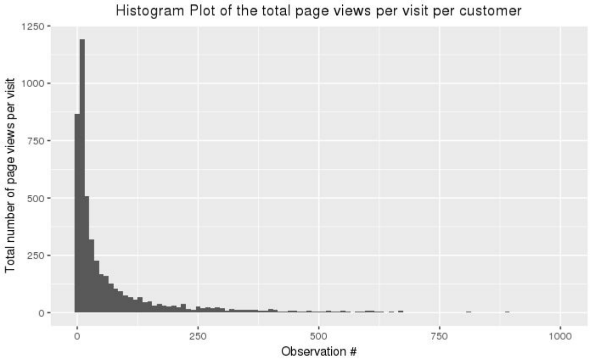


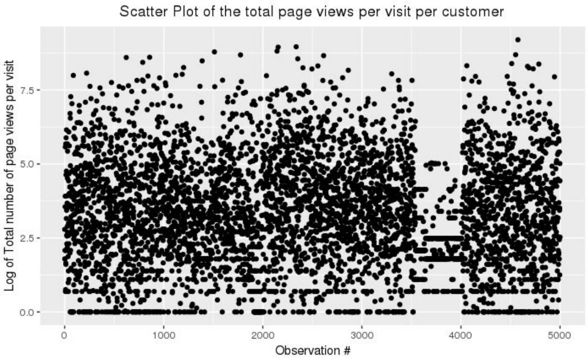


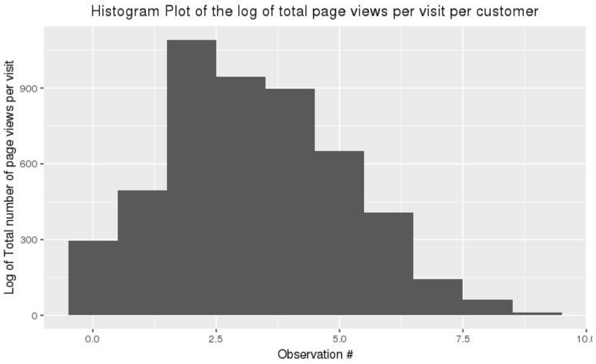
Page_views_per_visit

- Definition: Total number of pages views divided by number of visit days for a customer
- Transformation: Log
- Method of calculation: For each user, divide the total number of page views by the total number of visit days
- Summary:
  - Min.   :   1.000
  - 1st Qu.:   7.333
  - Median : 24.000
  - Mean   : 146.950
  - 3rd Qu.: 92.948
  - Max.   : 9,858.912
- Plots:



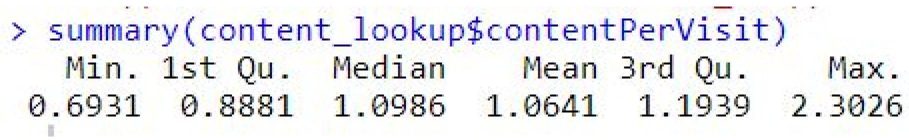






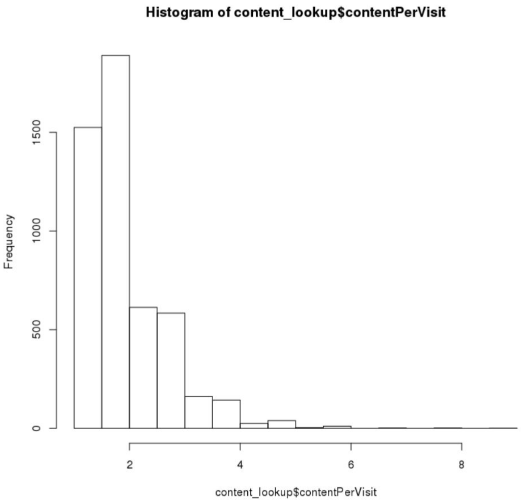
contentPerVisit

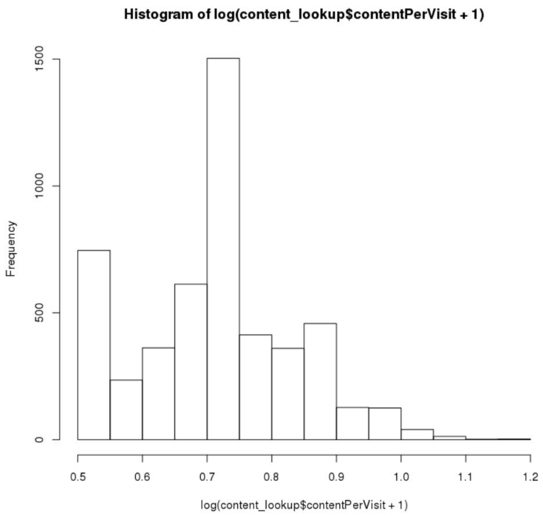
- Definition: The average number if distinct content types a user views on each visit day
- Transformation: Log
- Method of calculation:
    - For each user and visit day, count the distinct number of content types
    - For each user, sum the total count of distinct content types across all visit days and divide by the total number of visit days for that user
- Summary:



```
> summary(content_lookup$contentPerVisit)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.6931  0.8881  1.0986  1.0641  1.1939  2.3026 
```

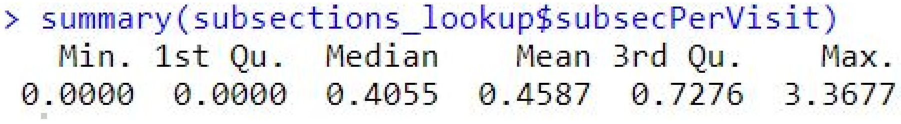
- Plots of original and log transformed variable:





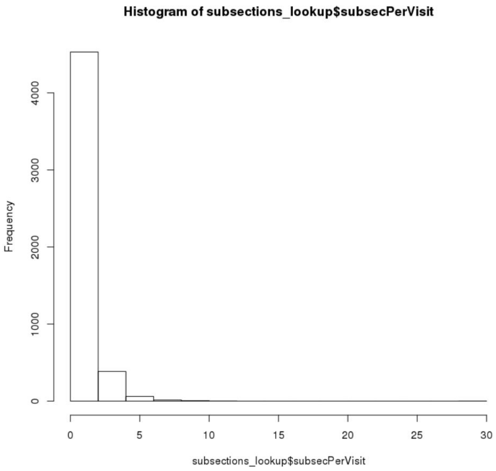
subsecPerVisit

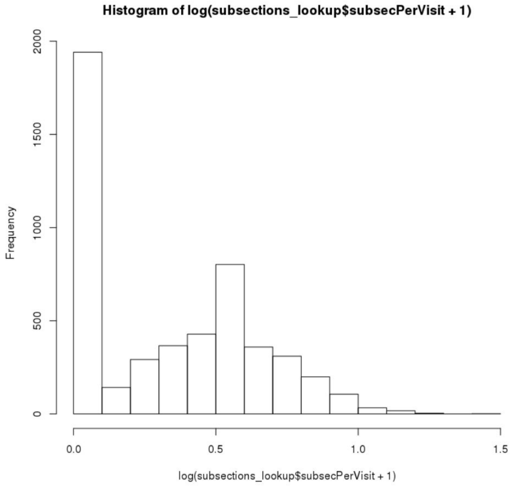
- Definition: the average number of subsections a user views on each visit day
- Transformation: Log
- Method of calculation:
    - For each user and visit day, count the number of subsections (this is equivalent to counting how many times the ':' appears in the subsection field for each visit day
    - For each user, sum the total count of subsections across all visit days and divide by the total number of visit days for that user
- Summary:



```
> summary(subsections_lookup$subsecPerVisit)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.4055  0.4587  0.7276  3.3677 
```

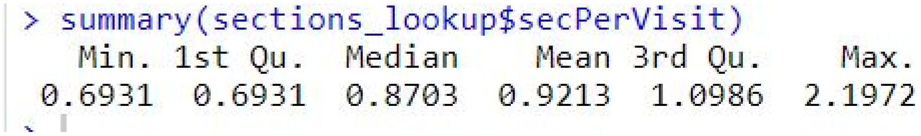
- Plots of original and log transformed variable:





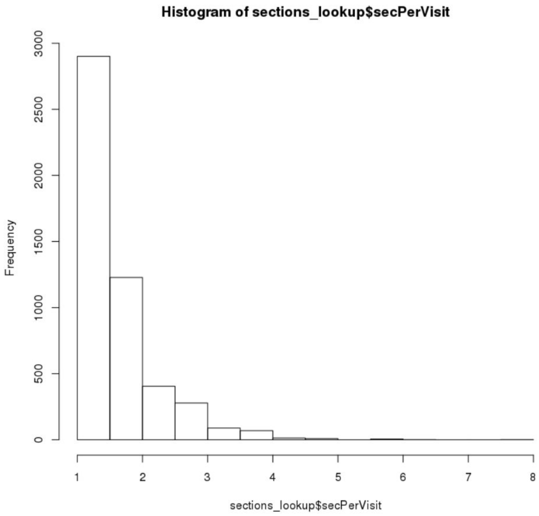
secPerVisit

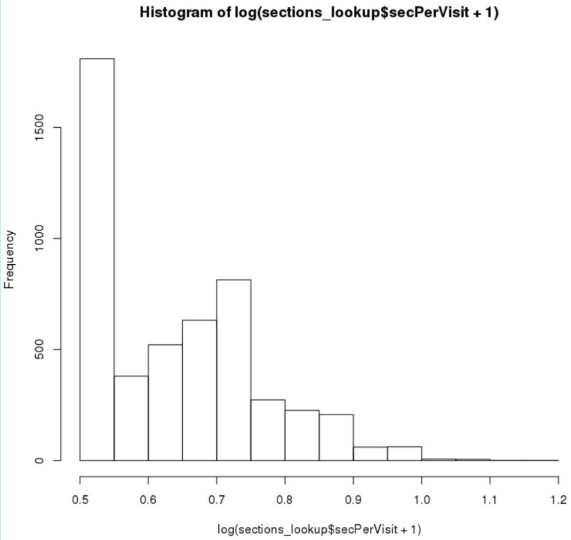
- Definition: the average number of distinct sections a user views on each visit day
- Transformation: Log
- Method of Calculation:
    - For each user and visit day, count the distinct number of sections
    - For each user, sum the total count of distinct sections across all visit days and divide by the total number of visit days for that user
- Summary:



```
> summary(sections_lookup$secPerVisit)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.6931  0.6931  0.8703  0.9213  1.0986  2.1972 
```

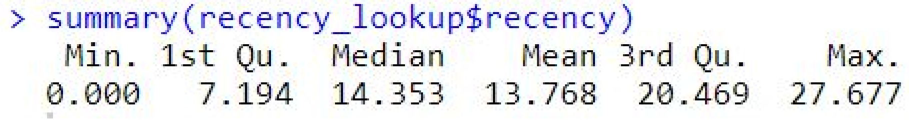
- Plots of original and log transformed variable:





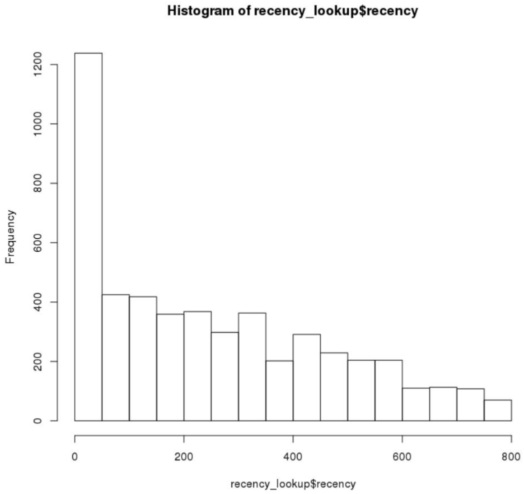
Recency

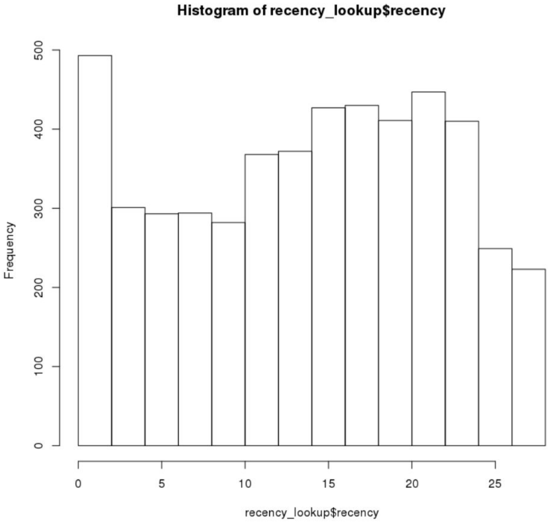
- Definition: the number of days between the user's most recent visit day and the maximum date in the dataset
- Transformation: Square Root
- Method of calculation:
    - For each user, select the most recent visiting date and find the difference in days between this date and the most recent date in the entire data set
- Summary:



```
> summary(recency_lookup$recency)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   7.194  14.353  13.768  20.469  27.677 
```

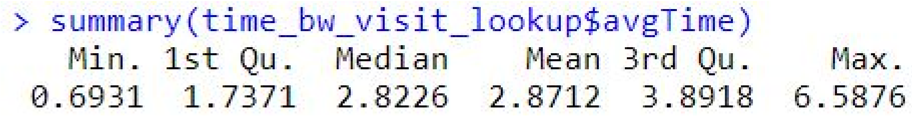
- Plots of original and log transformed variable:





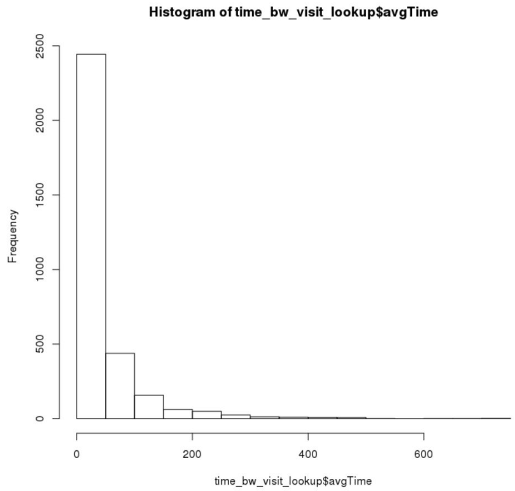
avgTime

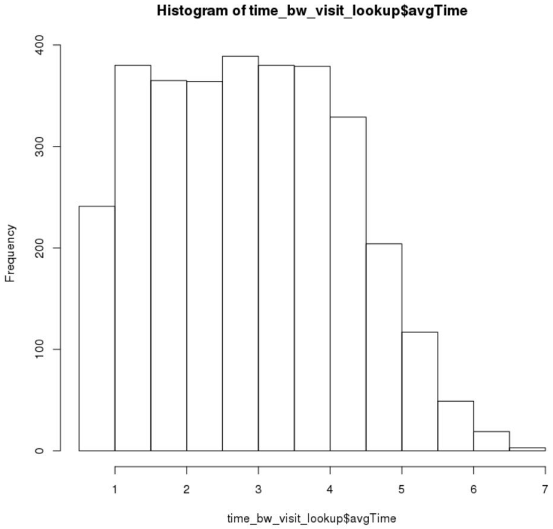
- Definition: the average number of days between visits for the user
- Transformation: Log
- Method of Calculation:
    - Create a data frame sorted descending by date within each user
    - Calculate the number of days between each row and the preceding row, grouped by user
    - Eliminate the “0” records (users with one 1 visit day, or the first visit day for each user)
    - Find the average number of days between visits for each user
- Summary:



```
> summary(time_bw_visit_lookup$avgTime)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.6931  1.7371  2.8226  2.8712  3.8918  6.5876 
```

- Plots of original and log transformed variables:





pct_gallery pct_homefront pct_section.front pct_story pct_sub.section

- Some content types are grouped to a broader content and content type that is missing is categorized as 'Other'.
- After grouping, top 5 content_type by pageviews are: "gallery", "homefront", "section front" "story", "sub-section"
- Definition: for each subscriber, number of pageviews made in particular content_type divided his/her total number of pageviews

statistics	pct_gallery	pct_homefront	pct_section.front	pct_story	pct_sub.section
min	0	0	0	0	0
max	1	1	1	1	1
median	0	0.004205	0	0.02263	0
mean	0.1772	0.107082	0.04800	0.13920	0.0175603

pct_home pct_sports pct_news pct_life
pct_entertainment pct_tech pct_weather pct_opinion
pct_travel pct_money pct_search

- Top 5 sections with most number of total pageviews are found. The topic 5 sections include home and sports where Indiana people care about. We also add additional general trending topic into the list
- Definition: for each subscriber, number of pageviews made in particular section divided his/her total number of pageviews
- The below is a summary for top 5 sections based on total pageviews

statistics	**pct_home**	**pct_sports**	**pct_news**	**pct_life**	**pct_entertainment**
min	0	0	0	0	0
max	1	1	1	1	1
median	0.004873	0	0.003854	0	0
mean	0.110456	0.1394	0.147604	0.061610	0.027710

Inside.Your.Site No.JavaScript Other.Web.Sites Search.Engines Social.Networks Typed.Bookmarked

- Definition: for each subscriber, number of dates accessed using particular referrer_type_value divided by his/her total number of dates accessed
- Method of calculation:
    - For each user, sum the total count of a particular referrer type across all dates and divide by the total number of dates for that user
- Statistics:

statistics	Inside.Your.Site	No.JavaScript	Other.Web.Sites	Search.Engines	Social Networks	Typed.Bookmarked
min	0	0	0	0	0	0
max	1	1	1	1	1	1
median	0.5	0	0	0	0	0.8113089
mean	0.5478397	3.365659e-05	0.04050988	0.1150347	0.08121292	0.6436684

CLUSTER ANALYSIS

With our total variable count at 30 (not including fire_fly_id), we realized that performing a cluster analysis using k means over all these variables would prove messy, yielding unclear results. Therefore, we decided to categorize our variables based on what aspect of customer behavior they relate to, and came up with 6 categories: breadth, depth, temporal, section, content and referral. The breadth variables describe variation of content viewed by customers, while the depth variables considered frequency of visits. The temporal variables are recency and time between visits. Section and content variables contain information about percentages of visit days which included browsing of specific section and content types, respectively. Finally, referral type variables contain information about percentages of visit days where the article page was accessed by a certain referral method.

After dividing the variables into these six bins, we performed 6 separate cluster analyses, each over a specific category of variables. The results of our clustering analyses are described below:

Breadth

Variables considered: secPerVisit, subsecPerVisit, contentPerVisit

Cluster 1 - “High Breadth All Around”

- These users show high breadth in terms of all attributes: high sections, subsections, and content types per visit

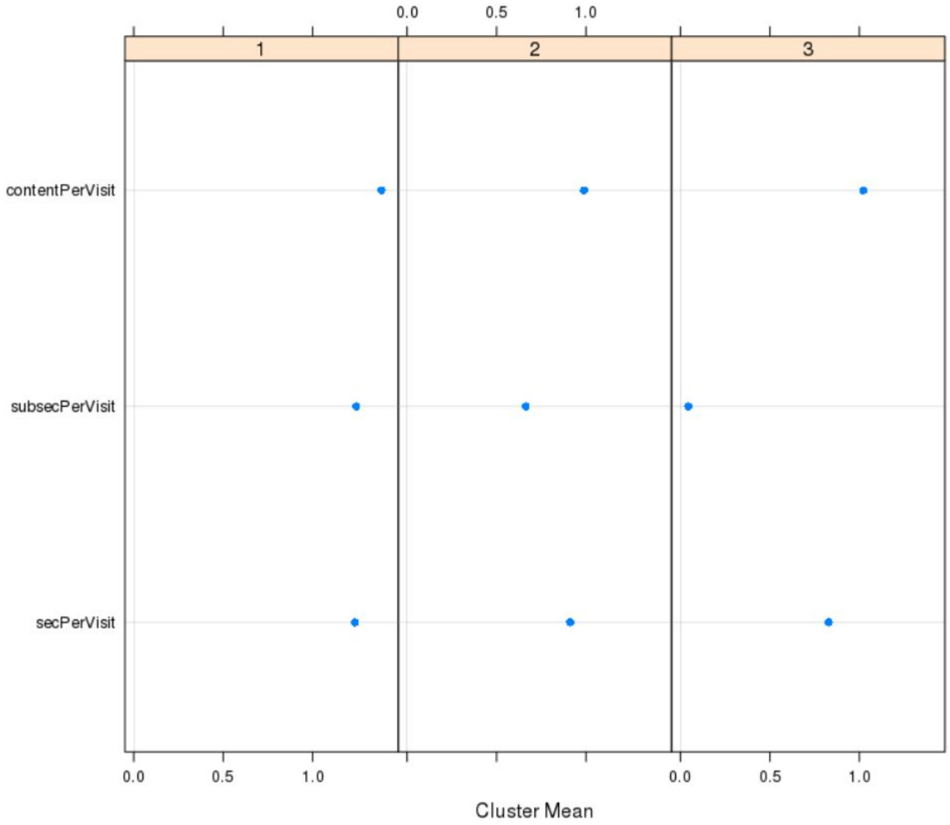
Cluster 2 - “Average Breadth”

- These users show a medium level breadth in terms of all attributes: sections, subsections, and content types per visit

Cluster 3 - “Low Subsections”

- These users have a well below average number of subsections per visit. Their sections and content types are about average

```
> summary.kmeans(fitBreadth)
     n  Pct secPerVisit subsecPerVisit contentPerVisit   RMSE
1  755 0.15        1.24           1.24            1.38 0.2648
2 1885 0.38        0.91           0.66            0.99 0.1806
3 2360 0.47        0.83           0.04            1.02 0.1833
  5000 1.00        0.92           0.46            1.06 0.1969
SSE =  581.0451 ; SSB =  1135.314
R-Squared =  0.6614665
Pseudo F =  4881.863
```



Temporal

Variables considered: Average Time, Recency

Cluster 1 - “Newer but Infrequent Visits”

- These users do not visit the paper frequently (high number of days between visits) but have viewed the paper more recently (low recency)

Cluster 2 - “Older but Frequent Visits”

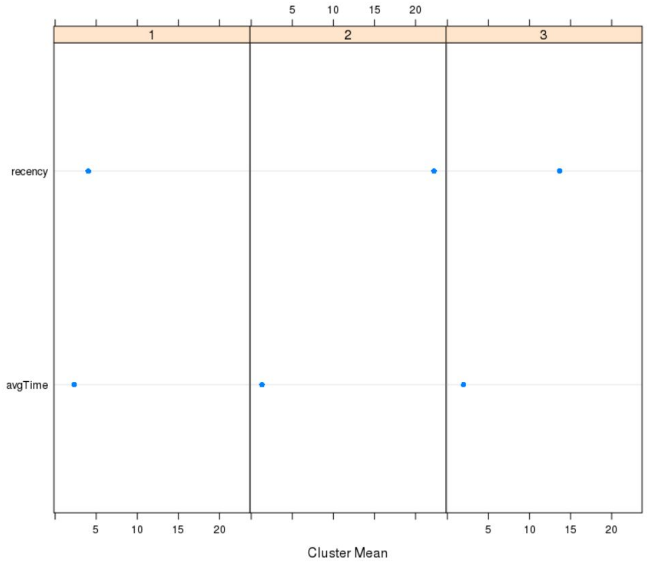
- These users visit the paper relatively frequently (low number of days between visits) but have not viewed the paper recently (high recency)
-

Cluster 3 - “Average Temporal”

- These users have average frequency and recency habits

```
> summary.kmeans(fitTemporal)
     n  Pct avgTime recency   RMSE
1 1526 0.31    2.33    4.04 2.1788
2 1774 0.35    1.31   22.24 2.2877
3 1700 0.34    1.98   13.66 2.2116
  5000 1.00    1.85   13.77 2.2294
SSE =  49672.43 ; SSB =  272936.1
```

```
R-Squared =  0.8460288
Pseudo F =  13728.56
```



Section

Variables considered: pct_home, pct_sports, pct_news, pct_life, pct_entertainment

Cluster 1 - “Entertainment Lover”

- These users love entertainment sections more than any other section.

Cluster 2 - “Homepage Lover”

- These users mainly browse homepages/headlines on the website.

Cluster 3 - “Other Categories Lover”

- These users are more interested in other news topics, not those trending topics in Indiana.

Cluster 4 - “News Lover”

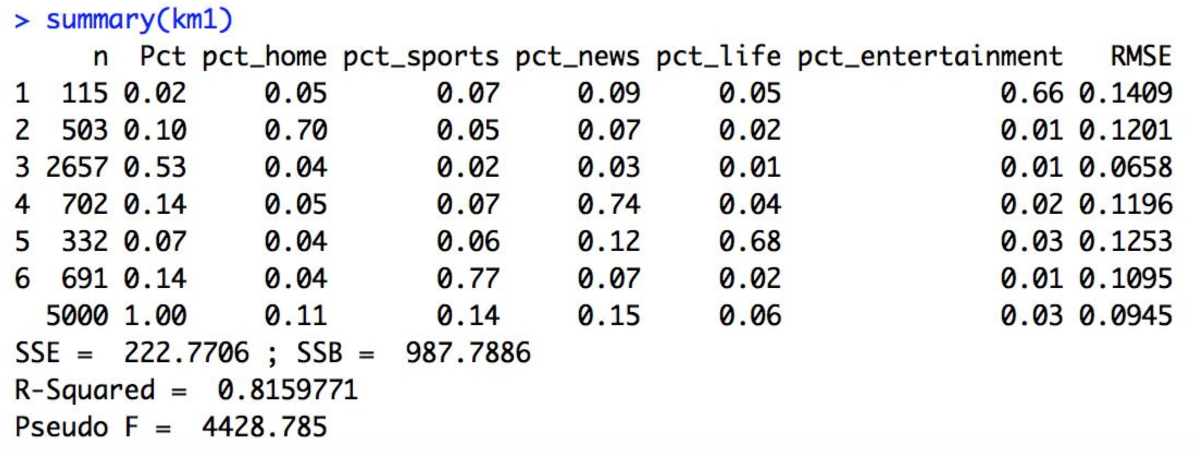
- These users are interested in reading news as opposed to other sections.

Cluster 5 - “Life Lover”

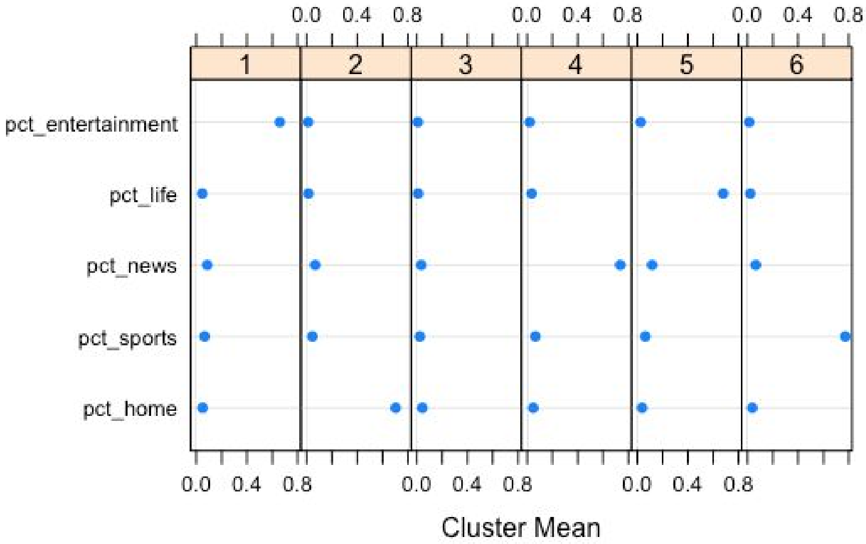
- These users are interested in reading life related articles as opposed to other sections.

Cluster 6 - “Sports Lover”

- These users are interested in sports and spend most of time viewing sports articles.



```
> summary(km1)
     n  Pct pct_home pct_sports pct_news pct_life pct_entertainment   RMSE
1  115 0.02     0.05       0.07     0.09     0.05              0.66 0.1409
2  503 0.10     0.70       0.05     0.07     0.02              0.01 0.1201
3 2657 0.53     0.04       0.02     0.03     0.01              0.01 0.0658
4  702 0.14     0.05       0.07     0.74     0.04              0.02 0.1196
5  332 0.07     0.04       0.06     0.12     0.68              0.03 0.1253
6  691 0.14     0.04       0.77     0.07     0.02              0.01 0.1095
  5000 1.00     0.11       0.14     0.15     0.06              0.03 0.0945
SSE =  222.7706 ; SSB =  987.7886
R-Squared =  0.8159771
Pseudo F =  4428.785
```



Content

Cluster 1- “All on the Home Front”

- These customers almost exclusively view content from the home front page

Cluster 2- “The Gallery Crowd”

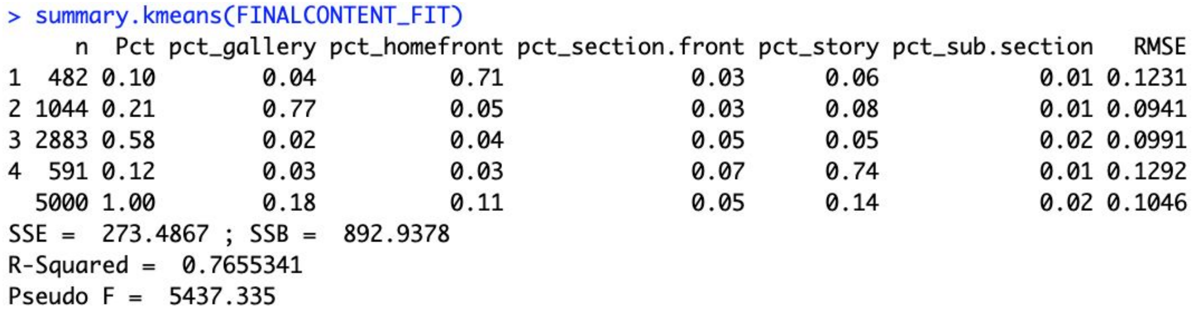
- These customers almost exclusively view content from the gallery

Cluster 3 - “The Nobodies”

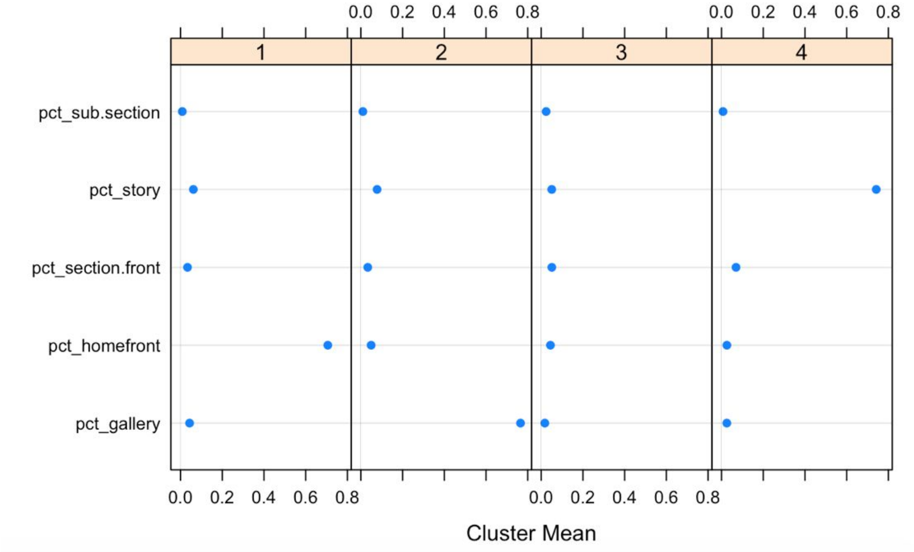
- These customers don’t seem to view much content at all

Cluster 4 - “All about the Story”

- These customers almost exclusively view story content



```
> summary.kmeans(FINALCONTENT_FIT)
     n  Pct pct_gallery pct_homefront pct_section.front pct_story pct_sub.section   RMSE
1  482 0.10        0.04          0.71              0.03      0.06            0.01 0.1231
2 1044 0.21        0.77          0.05              0.03      0.08            0.01 0.0941
3 2883 0.58        0.02          0.04              0.05      0.05            0.02 0.0991
4  591 0.12        0.03          0.03              0.07      0.74            0.01 0.1292
  5000 1.00        0.18          0.11              0.05      0.14            0.02 0.1046
SSE =  273.4867 ; SSB =  892.9378
R-Squared =  0.7655341
Pseudo F =  5437.335
```



Depth

Variables considered: total_visits, tot_page_views, page_views_per_visit

Cluster 1- “Least interested”

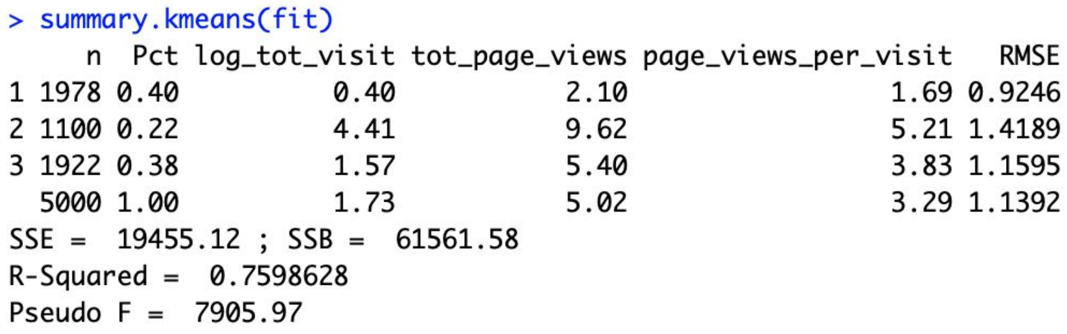
- These customers have the lowest number of total page views i.e. they hardly read during their rare visits to the paper

Cluster 2- “Most Interested”

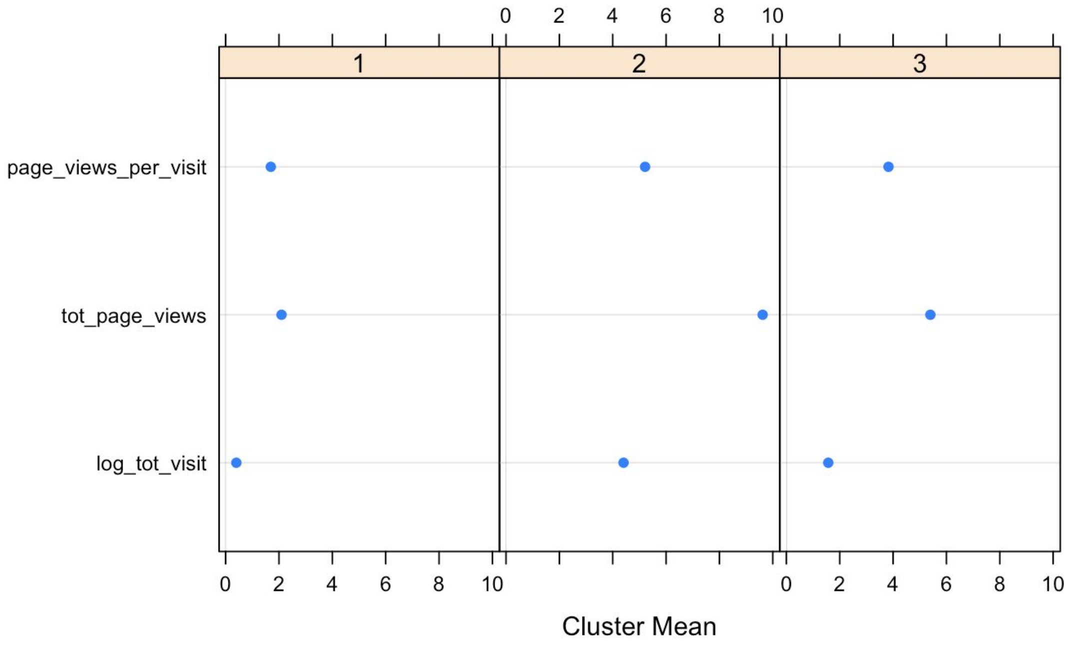
- These customers have the highest number of total page views and are enthusiastic readers when they visit the paper

Cluster 3 - “Moderately interested”

- These customers are moderately interested in what the paper has to offer, don’t visit the paper very often but have a considerable number of page views



```
> summary.kmeans(fit)
     n  Pct log_tot_visit tot_page_views page_views_per_visit   RMSE
1 1978 0.40          0.40           2.10                 1.69 0.9246
2 1100 0.22          4.41           9.62                 5.21 1.4189
3 1922 0.38          1.57           5.40                 3.83 1.1595
  5000 1.00          1.73           5.02                 3.29 1.1392
SSE =  19455.12 ; SSB =  61561.58
R-Squared =  0.7598628
Pseudo F =  7905.97
```



RefType

Variables considered: Inside.Your.Site, No.JavaScript, Other.Web.Sites,   Search.Engines, Social.Networks, Typed.Bookmarked

Cluster 1 - “Inside Your Site Only”

- These users almost exclusively are referred to the newspaper from inside its site

Cluster 2 - “Bookmarkers and Site Browsers”

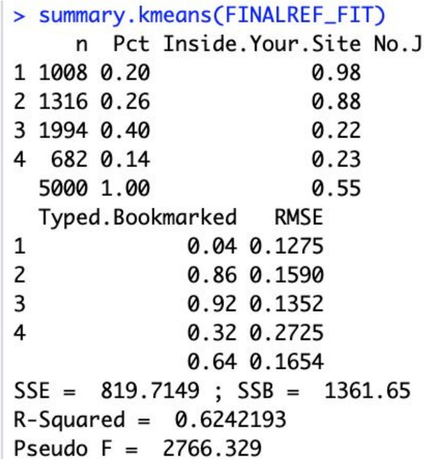
- These users equally are referred to the newspaper from bookmarks and from inside its site

Cluster 3- “Bookmarkers for the Most Part”

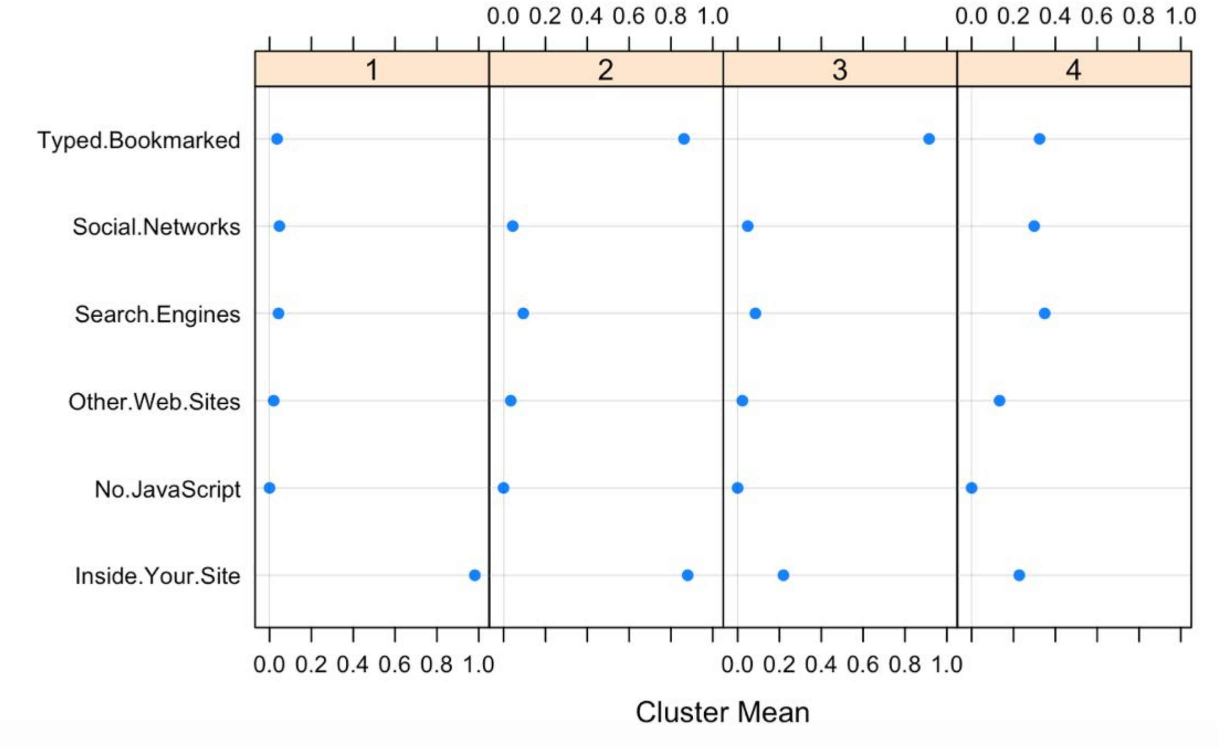
- These users are mostly referred to from bookmarks, but sometimes are referred to from inside its site

Cluster 4 - “The Variety Box”

- These users use numerous methods to get to the newspaper - possibly avid readers or just people who like to change things up



```
> summary.kmeans(FINALREF_FIT)
     n  Pct Inside.Your.Site No.JavaScript Other.Web.Sites Search.Engines Social.Networks
1 1008 0.20             0.98             0            0.02           0.04            0.05
2 1316 0.26             0.88             0            0.03           0.09            0.04
3 1994 0.40             0.22             0            0.02           0.08            0.05
4  682 0.14             0.23             0            0.13           0.35            0.30
  5000 1.00             0.55             0            0.04           0.12            0.08
  Typed.Bookmarked   RMSE
1             0.04 0.1275
2             0.86 0.1590
3             0.92 0.1352
4             0.32 0.2725
              0.64 0.1654
SSE =  819.7149 ; SSB =  1361.65
R-Squared =  0.6242193
Pseudo F =  2766.329
```



Final Clustering Analysis

We crossed the clusters created by the 6 clustering analyses to see how they created intersecting clusters. This yielded a whopping 609 clusters, many of which only had a single individual in them. We knew that consolidation would be key to create an accurate representation of the true clusters, so we searched for patterns in the crossed clusters that could create aggregate clusters using Tableau. We assessed which combined clusters seemed appropriate given their size and relevance to the newspaper's goal of a targeted newsletter. Each cluster below was initially assigned independently and then based on order of importance, those clusters were assigned sequentially to each subscriber (by filtering out certain subscribers that were already assigned previously). The order depends on what we thought the newspaper should target first for personalization (where there could be maximum opportunity).

Final Cluster Assignments

Cluster	Size	% of Total
Lost Souls	1,030	20.60%
Sports Lovers	351	7.02%
One-hit-wonders	1,054	21.08%
The Purists	367	7.34%
Photo Aficionados	846	16.92%
Specialists	251	5.02%
Headline Hunters	335	6.70%
Loyalists	498	9.96%
Lost Causes	163	3.26%
Wildcards	376	7.52%

Cluster Profiling

Cluster 1: Lost Souls

- Size = 1030
- Medium depth, bookmarkers, average temporal and older but frequent visitors(high recency and small average time between visits)
- Those subscribers are older customers with a small average time between visits but a higher value in the recency variable. High recency correlates to a longer length of time since these individuals have last frequented the newspaper's site. These individuals

used to visit the site often, but have stopped. Therefore, they have the highest potential for churn.

Recommendation:

Lost Souls is a group that this newspaper would want to win back since they used to be customers who bookmarked the website but no longer frequent the website. The newspaper could send out email reminders with recommended content similar to content they bookmarked, in addition to new content. Using marketing strategies, it is possible for the newspaper to retain customers.

Cluster 2: Sports Lovers

- Size = 351
- Bookmark for the most part, sports section, high and average breadth
- Those people bookmarked sports pages so it will be easy for them to track sports clubs. They are those with Indiana spirit - a passion for sports.

Recommendation:

These customers love sports so the newspaper could send out sports-related personalization emails and more specifically, news or stories related to their favorite teams.

Cluster 3: **One-hit-wonders**

- Size = 1,054
- Low breadth, low depth, interested in other (unpopular) sections
- Those people are one-hit-wonders and they look at less popular sections with little depth and breadth. They are not typical Indiana subscribers.

Recommendation:

Those people do not tend to view the most popular sections, so based on their browsing history, the newspaper should send personalized emails with content relating to the specific sections they are interested in.

Cluster 4: The Purists

- Size = 367

Story content, average or less recent but frequent visitors

- These visitors are more interested in the elaborated content of stories. We hypothesize that the adoption of more visual content on the newspaper's website contributes to these formerly active users' recent hiatus from visiting.

Recommendation:

Since these people prefer written content rather than picture-filled articles, the newspaper can personalize content by sending out a preview of the story/article (e.g. the first two paragraphs) instead of just headlines.

Cluster 5: Photo Aficionados

- Size = 846
- Gallery (pictures) content, sports, news, and life sections

- These people like looking at pictures instead of reading through long articles, and they are interested in some of the more popular sections.

Recommendation:

Sending out emails with photos indicative of new content or fresh releases and continuing to maintain the level of visually attractive content would help the newspaper in retaining these customers.

Cluster 6: Specialists

- Size = 251
- Inside your site as referrer type, low subsection visits
- Those people explore one specific subsection and also browse other articles/content in this specific section.

Recommendation:

These specialists enjoy a particular section of the newspaper and do not deviate from this. The newspaper should send interesting stories contained in the particular section the individual most often reads. Since these people do not bookmark, they might be more interested in current events. Therefore, the newspaper could also add interesting current event links.

Cluster 7: Headline Hunters

- Size = 335
- Home section, homefront content
- Assuming the home section is the homepage, those people are more interested in reading headlines of newspaper homepages.

Recommendation:

The newspaper could send out emails with headlines to target these customers.

Cluster 8: Loyalists

- Size = 498
- High and medium depth, high and average breath, no content preference
- Those people read a wide variety of content in the newspaper, and spend a considerable amount of time browsing different pages/sections during each visit.

Recommendation:

These customers displayed high volume in both visits and number of sections visited, and do not require extra personalization. The action could be to send popular articles from the most common sections such as sports, life, entertainment, etc.

Cluster 9: Lost Causes

- Size = 163
- Older but frequent visitors, low depth (number of pageviews or visits is low)
- Those people have a deflated average time between visits (they might make just one or two visits instead of visiting the site frequently)

Recommendation:

These users have not visited recently, and when they did, the depth/volume of these visits was low. The recommendation is to not spend extra attention on these customers and to prioritize other groups.

Cluster 10: Wildcards

- Size = 376
- Those people do not fall into above clusters.

Recommendation:

These users do not have unique characteristics that give the potential for personalization, and should be the last priority.