Homework 1 - Group J - IndyStar Data

Team Members: Henry Park, Naomi Kaduwela, Eileen Zhang, Tony Colucci, Nathan Franklin

Part A : Data Exploration & Issues

The first issue we came across is how is how to interpret a row in the dataset. We initially believed each row represented a unique users' visit to the site, with each visit having multiple page views. Yet, there can be multiple rows for a specific visit number, if that visit involved multiple articles, and thus we found that each row is a visit by a unique user to a specific article.

Additionally, section tags seemed to be very useful in high-level categorization, but it is very inconsistent whether a given topic will have details at the subsection level. For example, the topics under the news section is predominately just "news", while a sports section may provide some detail on which team the article was focused on. Therefore, the specific tags seemed to provide limited flexibility to incorporate into the model, but we were able to use the higher level aggregations.

We ended up focusing on browser value (mobile vs non-mobile), zip code (distance from Indy city center), and event date (weekend vs. weekend access) as our other columns. No significant problems were encountered with these variable columns.

Part B : Feature Engineering

We came up with 7 dimensions with which we wanted to cluster our readers and between one and 4 features for each of these dimensions. These were content, location, timing, platform, recency, frequency, amount.

Content - Four Features

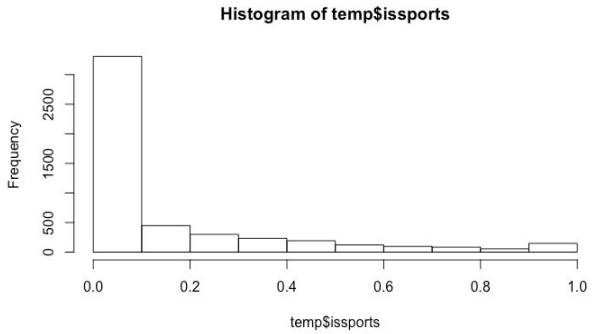
- *ishome*: Proportion of the number of home records to total number of records for each user
- *isspports*: Proportion of the number of sports records to total number of records for each user
- *isnews*: Proportion of the number of news records to total number of records for each user
- *others*: Proportion of all other records that are not sports, home, or news related to total number of records for each user

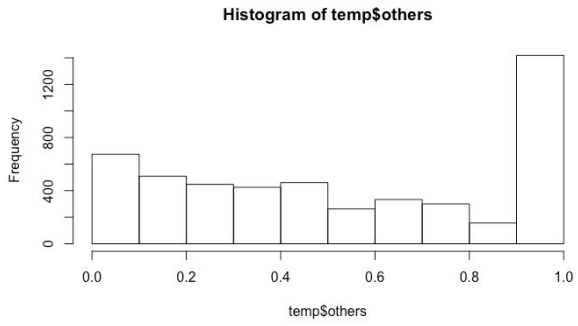
Content Type	# Of Rows Reading This Article Type	% of Rows Reading This Article Type

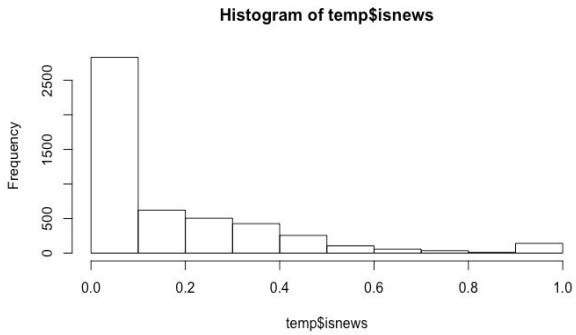
Home	270774	26.4
Sports	303921	29.7
News	248705	24.3
Others	201213	19.6
Total	1024613	100

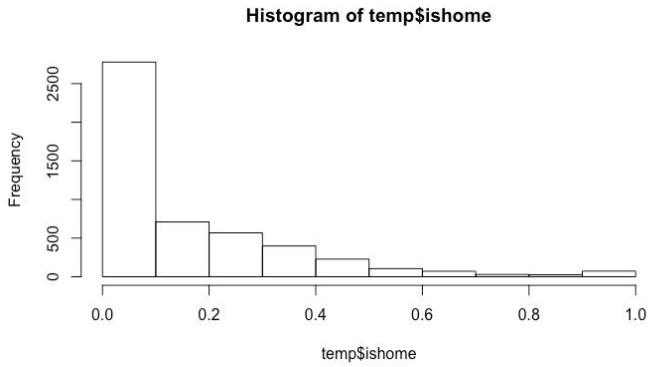
We see an even balance between these 4 sections. We considered including specific sports or types of sports content in our model, but because these entries only accounted for a very small proportion of the overall data, we did not think it would add value. We then transformed these raw counts into proportions, explained above, and named for whichever content type it tracked.

Variable	Mean	sd
ishome	0.15	0.21
issports	0.15	0.25
isnews	0.16	0.23
others	0.54	0.36



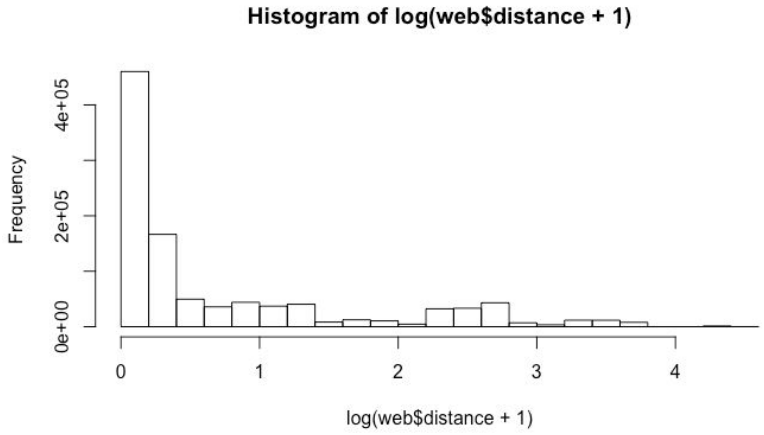






Location - One Feature

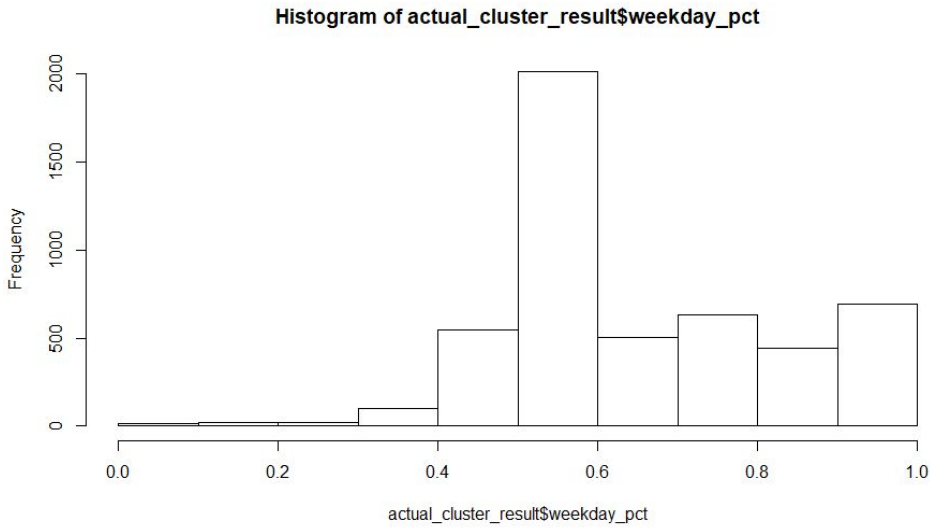
- *Distance:* Euclidean distance of latitude and longitude to the center of Indianapolis



The log distance histogram shows that even though most of the readers originate from Indiana, there are significant number of subscribers who visit from far away.

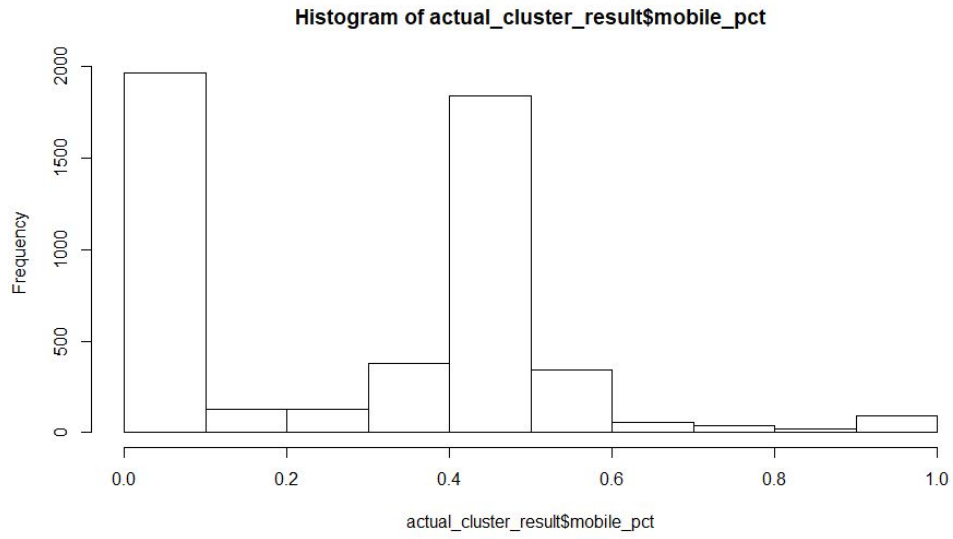
Timing - One Feature

- *Weekday_pct*
    - Percent of site accesses occurring on a weekday, weighted toward 50% for individuals with low numbers of site accesses (fewer than 10).
    - This feature provides a numeric indication of user habits regarding whether they access the newspaper's website on weekdays or weekends. This assumes subscribers will have weekly habits about how they check the website and may draw distinctions between users with interests or habits in different areas.
    - This measure is weighted toward 50% for users with low #s of accesses since we are less confident about habits formed by someone with a single weekend site access versus someone with 13 weekend accesses and 2 weekday accesses. Because of this, for those with fewer than 10 accesses, we calculate this feature by (((weekday accesses/total accesses) - 0.5) / (10 - total accesses) + 0.5).



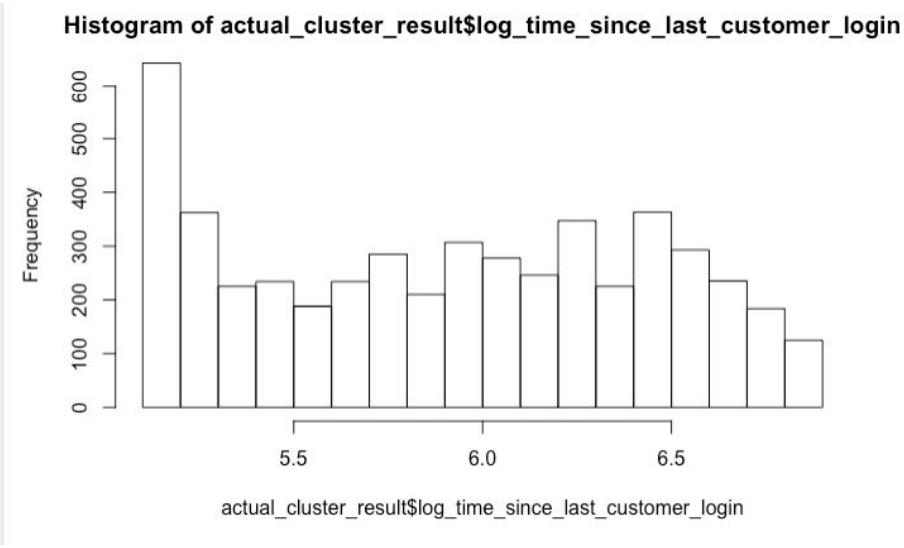
Platform - One Feature

- *Mobile_pct*
    - Percent of site accesses occurring from a mobile browser, weighted toward 50% for individuals with low numbers of site accesses (fewer than 10).
    - This feature provides a numeric indication of user habits regarding whether they access the newspaper's website through a mobile or web browser. This assumes that users primarily using mobile devices to access the site content may have different typical interactions or tendencies.
    - This measure is weighted toward 50% for users with low #s of accesses since we are less confident about habits formed by someone with a single mobile site access versus someone with 13 mobile accesses and 2 web accesses. Because of this, for those with fewer than 10 accesses, we calculate this feature by (((mobile accesses/total accesses) - 0.5) / (10 - total accesses) + 0.5).



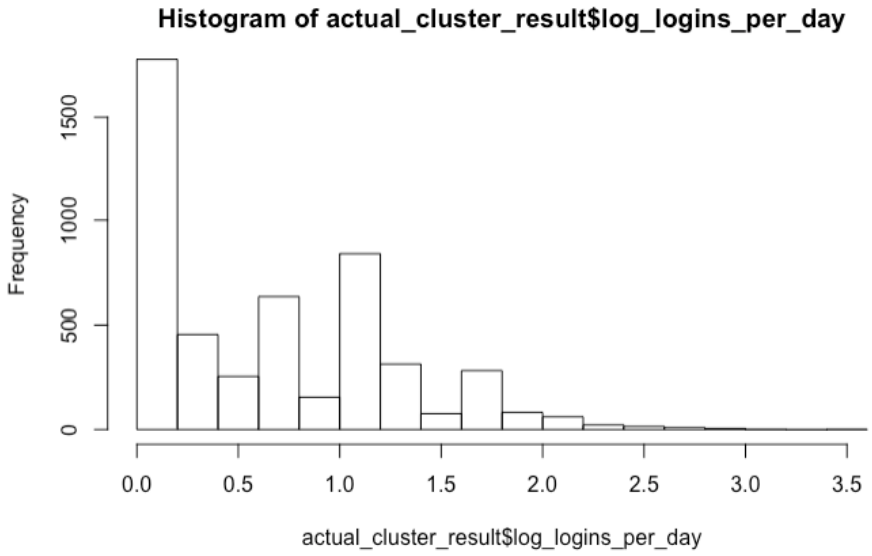
Recency - One Feature

- *Time_since_last_customer_login*
    - This variable measures the recency of customers' reading behavior. It indicates how long it has been since each individual customer's last login. More recent users might be the target customers of the newsletter.
    - Log of this variable was taken to standardize the data and the range of the log_time_since_last_customer_login is from 5.147 to 6.844 with an mean of 5.912



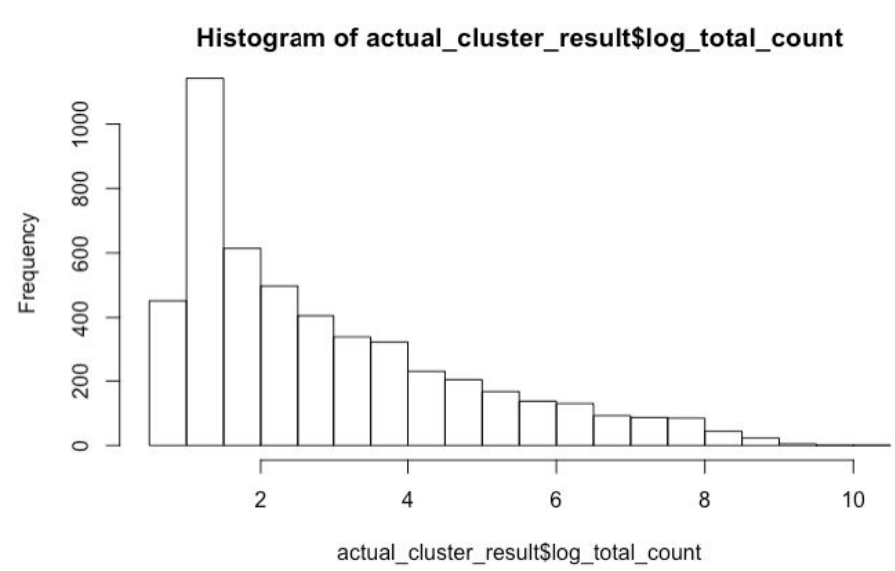
Frequency - One Feature

- *Logins_per_day*
    - This variable measures the average logins each users have every day during their visiting time, which is the frequency of their visits.
    - It is hypothesized that frequency is related to the topics as well as time of logins. (weekday vs. weekend)
    - This measure is highly skewed since there are very high numbers of logins, therefore log is taken to standardize the data. The log_logins_per_day ranges from 0.004016 to 3.583 with a mean of 0.667

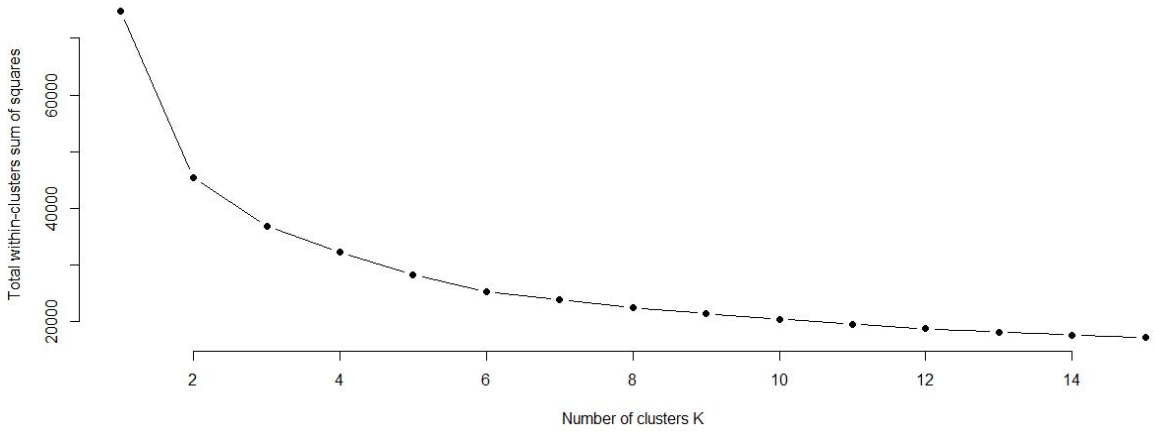


Amount - One Features

- *Total_count*
    - Total_count measures the total amount of logins each user has. Combined with other features like contents, insights can be drawn to see if people who are interested in certain type of contents reads more articles.
    - This variable ranges from 0.6931 to 10.1699 with mean at 2.8632.



- Choose Optimal Cluster Size



From the elbow curve plotted above we can see that k=7 seems to be the optimal choice. As k increases after 7, the decrease in the total within clusters sse is not very significant.

Part C: Cluster Analysis

We have clustered the data set based on the variable created in part b. 7 clusters seem to be ideal, and the reason why we thought that is because the sum of squares to the centroid seem to asymptote horizontally. The clustering was done from a data set where all counts were log-transformed and each feature was normalized to have a mean of 0 and standard deviation of 1. These transformations ensured equal weights for each feature and high interpretability of the model through the guitar plot.

We have decided to name each cluster:

1. Low-engagement readers - local (n = 1265)
    a. Among the lowest in total times logging on, high proportion of views for "other" content, but low engagements with news, sports or home.
    b. The lowest average distance from Indianapolis, these readers are probably here for updates once in a while
    c. These readers were very high in their mobile usage as well
2. News readers (n = 309)
    a. Very high proportion of content consumed is about news
    b. Highest average time since last login, so these may be subscribers that have moved to other publications, or news content may not be as sticky
    c. High mobile percentage and average distance, so these users may be looking for more real-time updates and have a wider than local interest
3. Low-engagement readers - out-of-towners (n = 649)
    a. These readers don't have a high proportion of articles focused one of the big topics (sports, home, news).
    b. Very similar to cluster 1, except that this group has the highest average distance from Indianapolis. This indicates that the two could potentially be part of a larger cluster if we lowered the importance of distance in our algorithm.
4. Daily readers (n = 1415)
    a. Subscribers who have a high number of accesses, but a low mean number of accesses per day, pointing to more regular site accesses
    b. Weighted more toward weekday access, from a desktop browser
    c. Match very close to mean preferences in terms of content, i.e. they seem to browse the entire site
5. Sports readers (n = 429)
    a. Very high proportion of content consumed is about sports
    b. Slightly above average log_mode_distance indicating that more of the people tend to be farther from Indianapolis, so this content may have more relevance to people outside of the immediate area
    c. Slightly below average weekday access, so accesses are weighted toward the weekend, which may correlate with the timing of sporting events of interest
6. Home readers (n = 329)
    a. Very high proportion of content is under the Home tag
    b. This group has close to the highest average time since last login and a below average total number of engagements indicating that these readers have had less recent engagement overall than other groups.
7. Power users (n = 588)

a. Highest number of total site accesses and lowest time since last login - these are the users that are most connected to the newspaper
b. Very low mobile percentage - people are accessing via desktop browser
c. High weekday percentage - they are accessing the site as part of their daily routine, not just for certain events

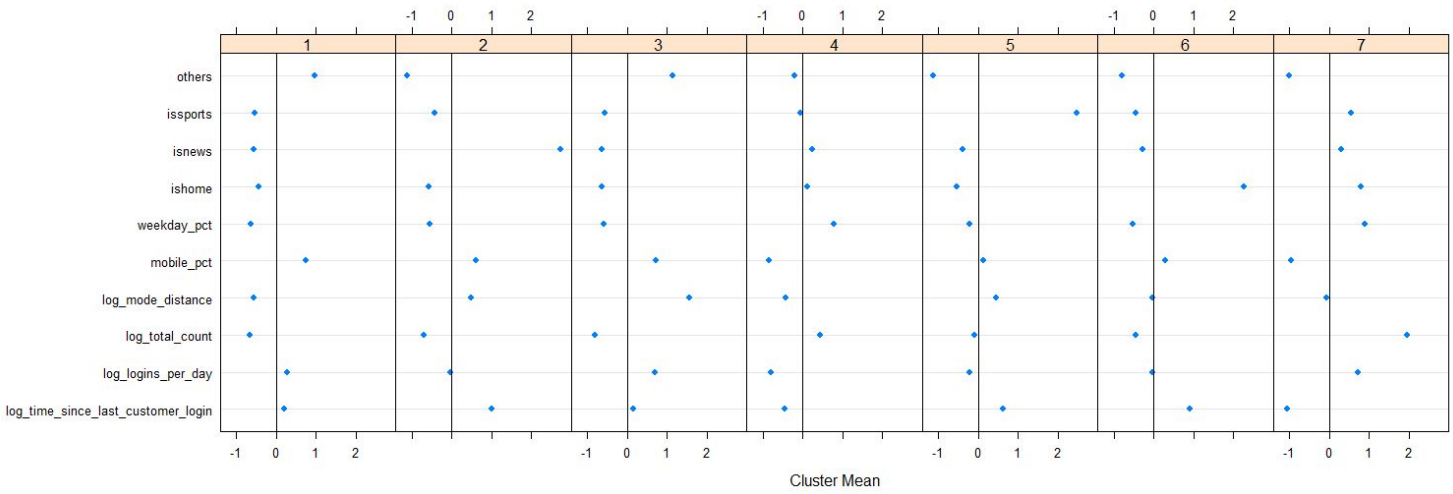


Table of means for each cluster:

cluster_number	Log_time_since_last_login	log_logins_per_day	log_total_count	log_mode_distance	mobile_pct	weekday_pct	ishome	isnews	issports	others
1	0.205	0.293	-0.665	-0.556	0.759	-0.64	-0.44	-0.551	-0.527	0.959
2	1.005	-0.021	-0.686	0.495	0.623	-0.538	-0.576	2.728	-0.413	-1.105
3	0.158	0.703	-0.815	1.571	0.713	-0.589	-0.634	-0.637	-0.571	1.155
4	-0.440	-0.801	0.432	-0.426	-0.855	0.774	0.124	0.249	-0.050	-0.193
5	0.623	-0.216	-0.085	0.438	0.134	-0.206	-0.54	-0.4	2.473	-1.136
6	0.916	-0.036	-0.447	-0.035	0.291	-0.534	2.258	-0.279	-0.453	-0.804
7	-1.051	0.712	1.962	-0.073	-0.951	0.896	0.781	0.304	0.549	-1.014

The EDA and clustering helped us discover that certain variables that we have considered to be potentially very important such as news related to Pacers did not turn out to be effective in clustering.

With further time, it could be interesting to experiment more with weighting certain dimensions to put more emphasis on total number of accesses or time since last login to increase the importance of those stratifications when making clusters.

Part D: Business Recommendations

The goal of clustering was to segment our users into behavioral user-groups which could be targeted with new products based on their characteristics. Group one's characteristics were high mobile usage and non-specific content consumers, as well as a low total count of visits. It is

possible that more mobile focused content could grab their attention, and analysis on videos vs. text consumption could further inform curated content for them.

Subscribers in home, news, and sports readers clusters (groups 2, 5, & 6) could be targeted with home, news, and sports newsletters respectively. Perhaps a weekly summation of these topics to catch them up on things users may have missed. Power user cluster group could be targeted for more frequent newsletter summary of bits of everything from the news. If we could see the time of day these users accessed the site we could pick even more optimal times for each user to send them curated content.

Lastly, low engagement local readers (group 1) could be sent more localized newsletter with bits of everything occasionally, and similarity for the low engagement readers that live far away (group 3) could be targeted with newsletter with bits of everything excluding local news occasionally.