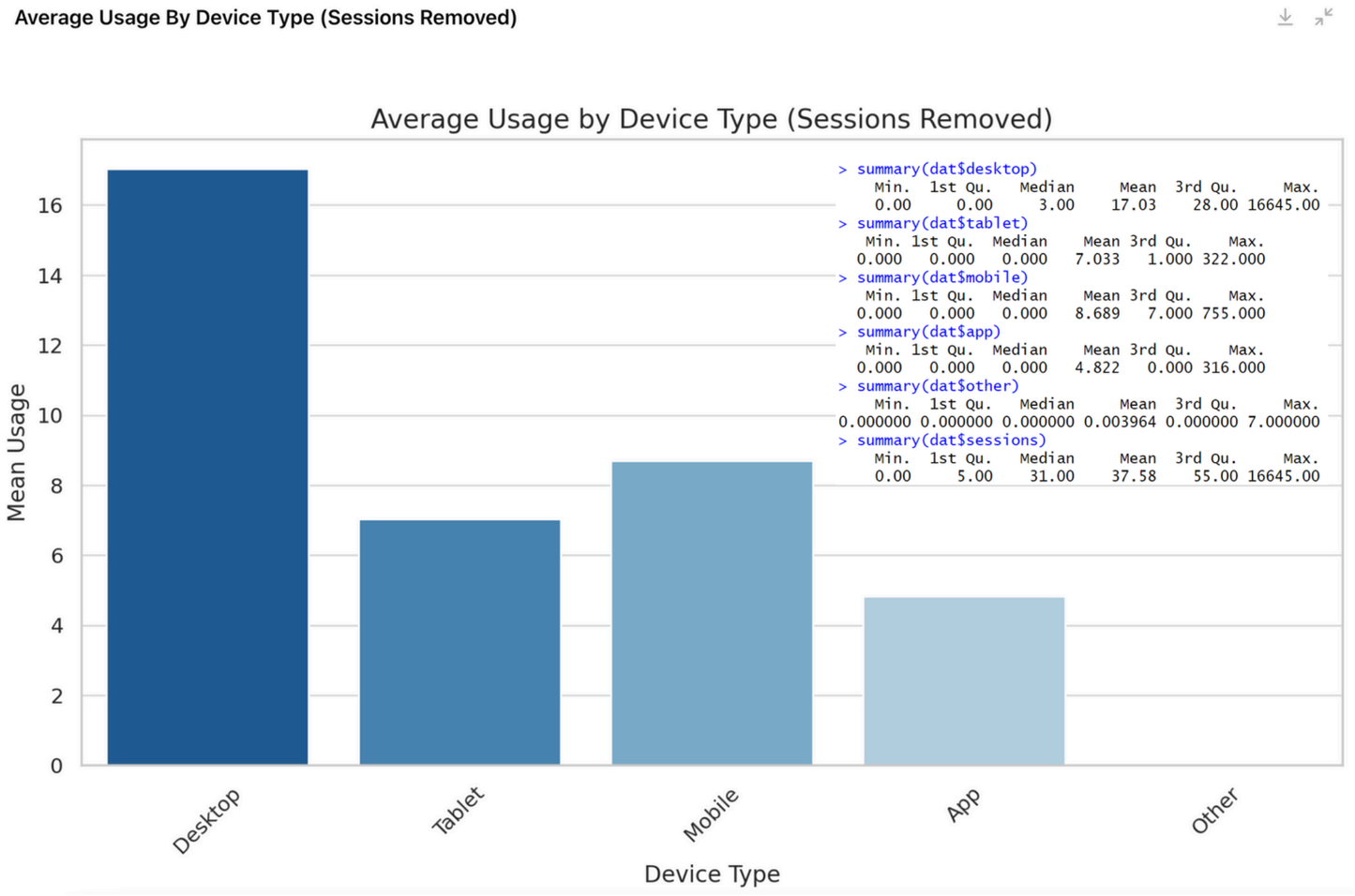
460 DATA SCIENCE FOUNDATIONS
FINAL CASE

Team 21.1
Nazarette Vallis, Priyal Rajnish Khanuja, Ruiyang Cheng, Sheryl Xu, Xinyi Wang

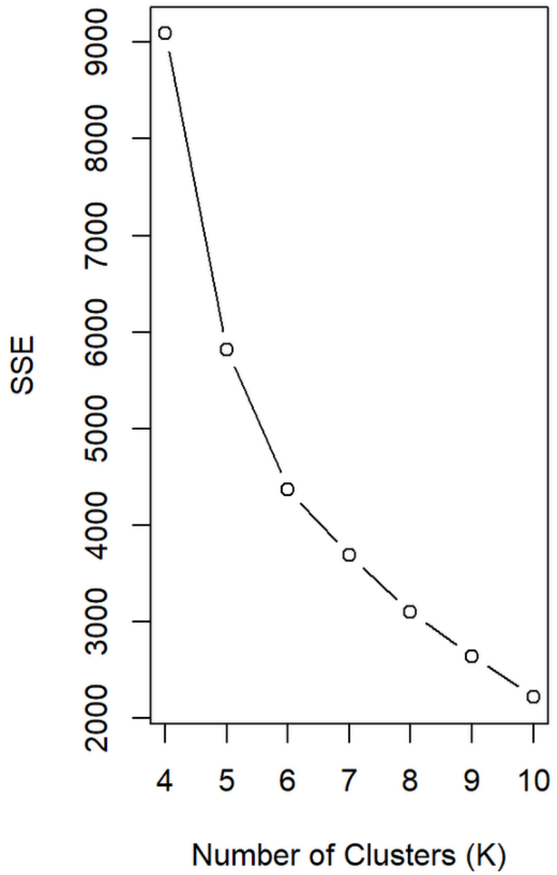
Find some clusters using the five device variables to identify types of device users.
(a) What pre-processing did you do to the data? Why?

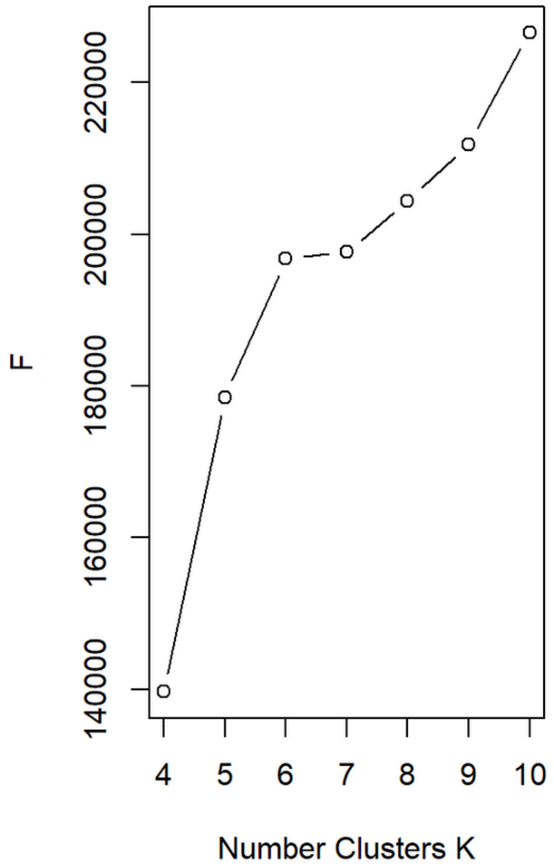


According to the data summary, the maximum and minimum values show that there is a significant difference in the use of the five devices by the users, and the difference in the total device use by each user is also more significant.

Therefore, the usage should be standardized by converting it to a percentage to highlight relative preferences and facilitate clustering.

(b) How did you select the number of clusters? Note that you will want to have enough clusters to answer both questions (Q1) and (Q2). You will want some types of multichannel subscribers, some (mostly) single-channel types, and a light reader cluster.





SSE:

According to the Elbow Method, the decrease of SSE slows down when there are 6 clusters.

F value:

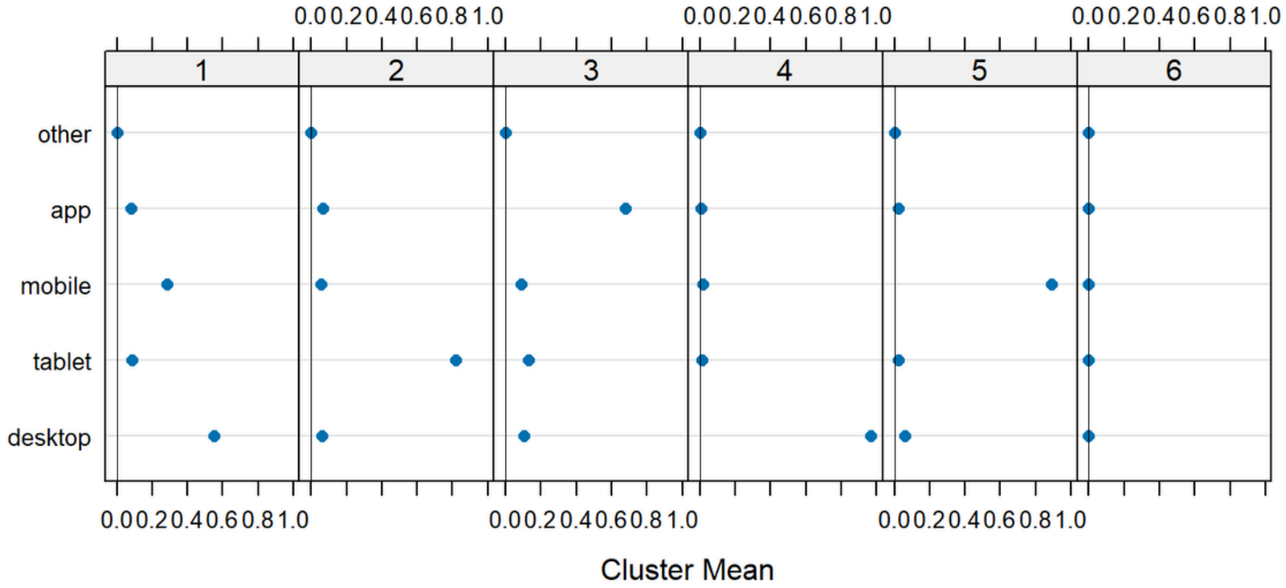
When the number of clusters changes from 6 to 7, the rate of F rise slows down significantly. Therefore the number of clusters that should be chosen is 6

Actional Marketing perspective:

For marketing effectiveness and realizability (to better answer the two questions), the number of clusters should not be too small (4 or even 5) or too big (9 or 10).

To summarize, we should go for six clusters.

(c) Describe your best clusters. At the minimum I would expect to see the size (percent of all cases), the cluster means, and names for the clusters.



Cluster	% of Subscribers	Desktop Users (%)	Tablet Users (%)	Mobile Users (%)	App Users (%)
1. Polyamorous Lovers	11	55	9	28	8
2. The Clingy	13	6	82	6	6
3. App Loyalists	10	10	13	9	68
4. Old Schoolers	31	97	1	2	1
5. Pocket News Junkies	17	6	2	89	2
6. The Ghost-ers	17	0	0	0	0
Total	100	40	14	21	9

(d) Profile your clusters on nextchurn and sessions. Briefly discuss these statistics.

Polyamorous Lovers

11%

Balanced use across desktop (**55%**), mobile (**28%**), and tablet (**9%**) but have their favourites. Heavy consumers of news who read at work and continue on mobile.

The Clingy: I Just Need my Tablet Users

13%

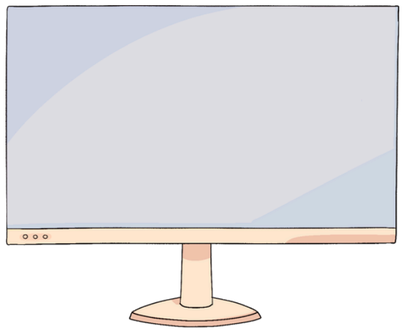
Heavy tablet users (**82%**) Little engagement on other desktop and mobile.
Likely casual readers, enjoying news in a relaxed setting.

App Loyalists

10%

Primarily use the news app (**68%** usage). Likely engaged users who prefer a dedicated experience rather than browser-based reading.

(d) Profile your clusters on nextchurn and sessions. Briefly discuss these statistics.



The Old Schoolers

31%

One woman guy who is mostly exclusive and commited to the Desktop but will flirt around occasionally with other devices

Pocket News Junkies

17%

Mobile-first readers (**89%** usage). People on the go, check the news on their mobiles.

The Ghost-ers

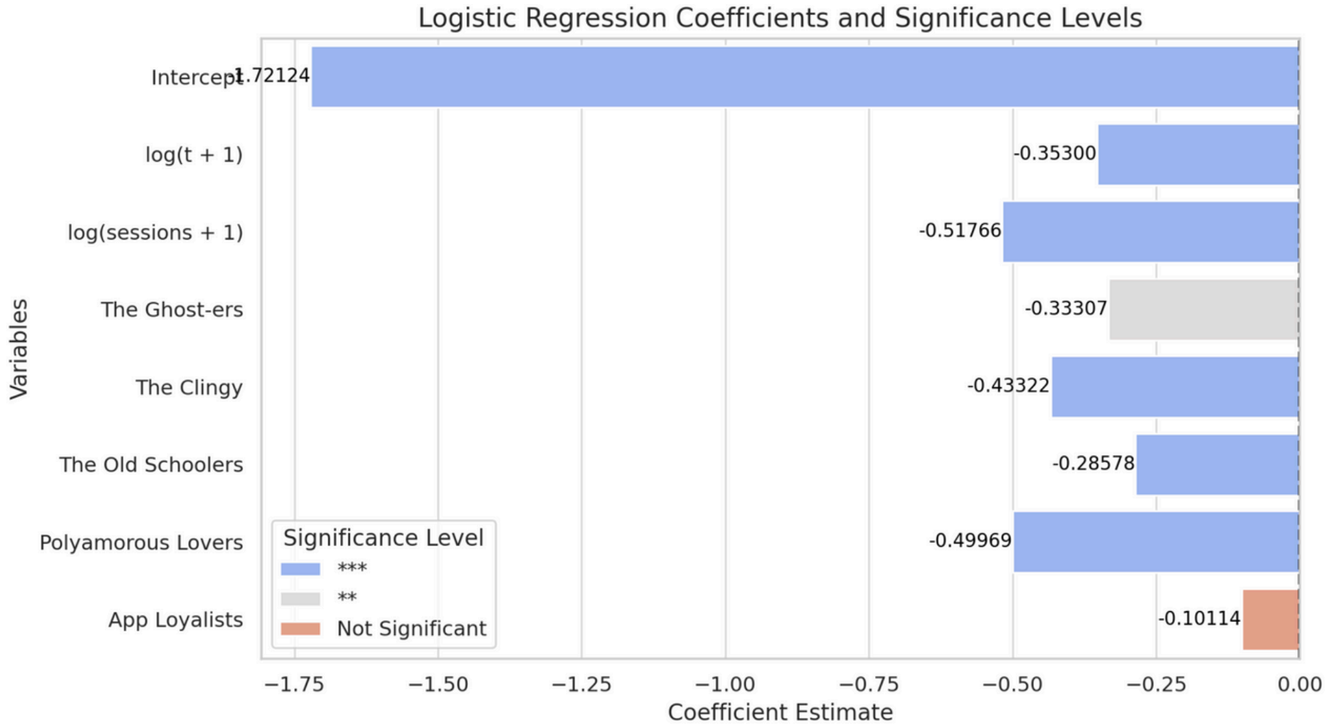
17%

The one’s who leave us on read - buy our subscription but never use it.

Fit a logistic regression model predicting nextchurn from control variables log(sessions) and log(t), and your clusters (as a categorical/factor variable). What does it tell you?

Cluster Name	Churn Rate (%)
The Ghost-ers	4.4%
Pocket News Junkies	1.5%
The Old Schoolers	1.0%
App Loyalists	0.9%
The Clingy: Tablet Users	0.7%
Polyamorous Lovers	0.6%

Fit a logistic regression model predicting nextchurn from control variables log(sessions) and log(t), and your clusters (as a categorical/factor variable). What does it tell you?

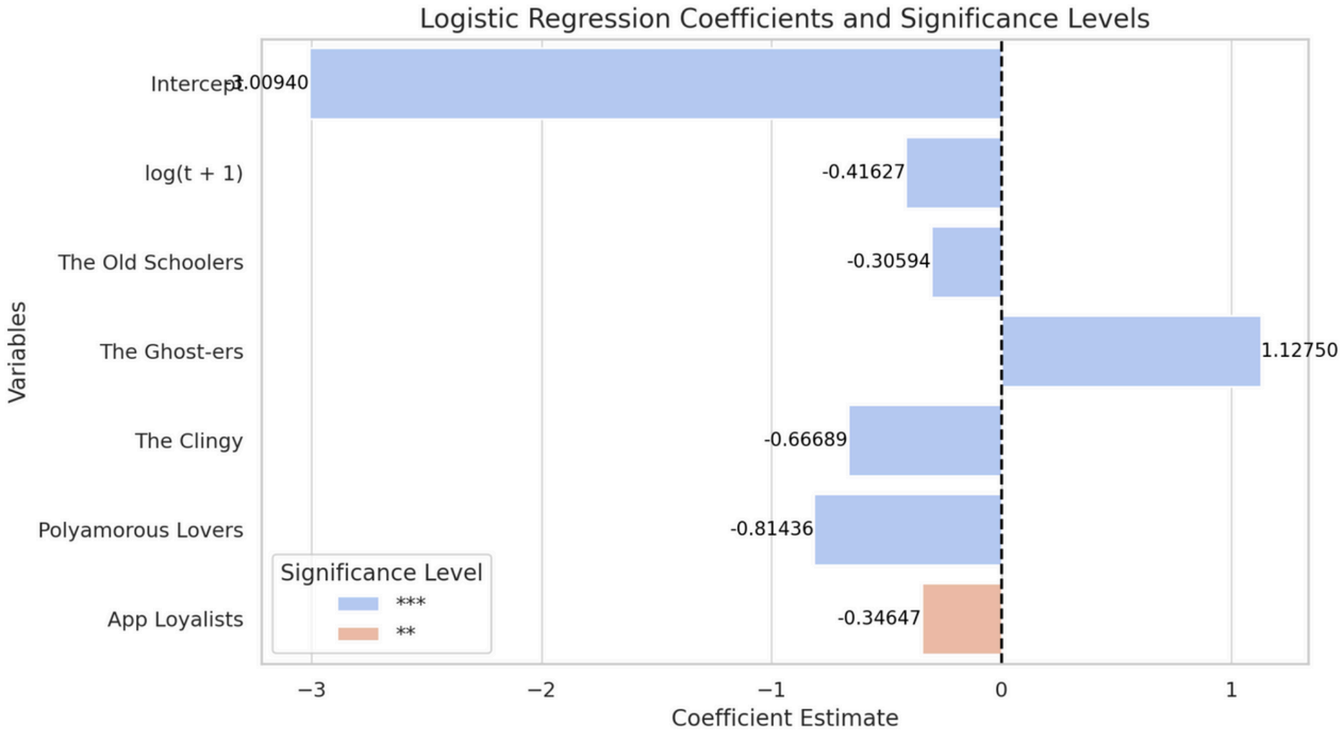


When t and sessions are included & with "Pocket News Junkies" as the reference group

- **Nearly all variables in the model are statistically significant.**
- ***Users with larger month numbers (t) and device usage (sessions) are significantly less likely to churn.***
- ***Pocket News Junkies (reference group) have the highest churn likelihood compared to other clusters, which contradicts with the previous conclusion.***

Fit a logistic regression model predicting nextchurn from control variables log(sessions) and log(t), and your clusters (as a categorical/factor variable). What does it tell you?

	GVIF	Df	GVIF^(1/(2*Df))
log(t + 1)	1.032104	1	1.015925
log(sessions + 1)	3.426590	1	1.851105
clus	3.437421	5	1.131419



When sessions is excluded & with "Pocket News Junkies" as the reference group

- **The whole model is significant.**
- ***Users with larger month numbers (t) are significantly less likely to churn.***
- ***Pocket News Junkies (reference group) have the second highest churn likelihood.***
- ***The Ghost-ers are more likely to churn compared with Pocket News Junkies (1.1).***
- ***Polyamorous Lovers have the lowest churn rate of -0.8, which aligns with the previous conclusion.***

Fit a logistic regression model predicting nextchurn from control variables log(sessions) and log(t), and your clusters (as a categorical/factor variable). What does it tell you?

Compare the means of *sessions* and *t* among the six clusters, The results are statistically significant

Cluster Name	Avg. Log(Sessions)	Avg. Log(t)
The Ghost-ers	0.002	3.01
Pocket News Junkies	3.63	3.29
The Old Schoolers	3.26	3.14
App Loyalists	3.21	3.04
The Clingy: Tablet Users	3.76	3.22
Polyamorous Lovers	3.63	3.25

The Ghost-ers have the lowest log-transformed average session count and tenure

Polyamorous Lovers (multi-device readers) have the highest log-transformed average session count

Write a short paragraph summarizing your findings, as if you were summarizing them for senior management at the company, including their VP of data science and CEO. Be sure to answer their questions (1) and (2). Feel free to run other analyses you think are necessary. Are there ways to answer the questions better without clustering first?

- The customer cluster analysis and profiling reveal six distinct clusters, each exhibiting varying churn tendencies. The churn rate ranks from highest to lowest as follows: The Ghost-ers (4.4%) > Pocket News Junkies (1.5%) > The Old Schoolers (1.0%) > App Loyalists (0.9%) > The Clingy: Tablet Users (0.7%) > Polyamorous Lovers (0.6%).
- The regression model shows that multi-device readers have lower churn rates. Our findings match the ranking from the prop table. It also shows that users with longer tenure (t) and higher total device usage (sessions) are significantly less likely to churn.
- Our compare means analysis shows that multi-device readers tend to have higher session counts (Polyamorous Lovers>The Clingy>The App Loyalists>The Old Schoolers>The Pocket News Junkies>The Ghosters). Multi-device readers may not have the longest tenure, while The Ghost-ers do have the shortest tenure

Why do you think you are seeing these results (i.e., the cause)? You would want to discuss this with management. With this understanding, what do you suggest they do in the future in terms of encouraging or discouraging certain behaviors?

- **Encourage and Enhance Multichannel Usage :** Improve Cross-Device Integration, Promote Cross-Platform Engagement
- **Target Ghost-ers :** tailored re-engagement through personalized outreach, incentives, content adjustments
- **Old Schoolers** : enhance desktop experience
- **Pocket News Junkies** : bite-sized content
- **Track engagement and lifetime :** Offer loyalty programs


THANK YOU