新闻设备案例参见churnclass.csv。一位媒体高管最近问了我以下问题（Q1）阅读设备如何影响用户留存，特别是（Q2）多渠道读者的流失率是否更低？术语“多通道”是指子通道Scriber在多个设备上读取数据。我给你的是一家出版公司的真实数据（与谷歌news这样的新闻聚合器相反），它提供五个频道：桌面（即阅读）电脑浏览器（包括笔记本电脑）、移动设备（使用移动设备在浏览器上阅读）、平板电脑、应用程序（使用公司的新闻应用程序阅读），其他（在GameBoy或Kindle等其他设备上阅读）。您有一个包含10,000个订阅者的数据集，每个月的周期（105,133个观测值）的数据设置运行逻辑回归模型。你知道每种类型的访问次数，用户是否在下个月流失（nextchurn）。变量t给出月份数在客户的生命中。我会让你先回答设备/多通道的问题然后使用SRM和离散时间模型对设备进行聚类。变量sessions等于手机、桌面、app、平板等的总和

IMC460: Data Science
Case 6: : K-means
Due: Thursday, March 20 by 23:59

Submit homework at the before class on Canvas. Work in your homework groups and submit one copy of the homework per group with the names of all your group members.

Email me a schedule a 15-minute time to present on March 14, 15, 17 or 18. Copy all memeber of your team on the email and tell me your team number, e.g., 20.2.

News device case see `churnclass.csv`. A media executive recently asked me the following questions, (Q1) how does the reading device affect subscriber retention, and in particular (Q2) whether multichannel readers have lower churn rates? The term *multichannel* means that a subscriber is reading on more than one device. I have given you real data from a publishing company (as opposed to a news aggregator like Google News) that offers five channels: desktop (i.e., reading on a computer browser including laptops), mobile (read on browser using mobile device), tablet, app (read using company's news app), other (read on another device such as GameBoy or Kindle). You have a data set of 10,000 subscribers with monthly periods (105,133 total observations). The data is set up to run a logistic regression model. You know the number of visits of each type and whether the subscriber churns in the next month (`nextchurn`). The variable `t` gives the month number in the customer's life. I will ask you to answer the device/multichannel questions by first clustering on devices then using the SRM and discrete-time models. The variable `sessions` equals the sum of mobile, desktop, app, tablet and other.

```
setwd("put your directory here")
dat = read.csv("churnclass.csv") %>%
  mutate(sessions = mobile + tablet + desktop+app+other)
```

1. 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?
   (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.
   (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.
   (d) Profile your clusters on `nextchurn` and `sessions`. Briefly discuss these statistics.
2. 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?
3. 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?
4. 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?