Classifying Riders

Given the results from last chapter, there is a clear need to understand what kinds of riders are in this data set. Ride Report, because they need to respect the privacy of their users, cannot identify individual riders in the data they provide to clients, yet our model results show that differentiating riders is crucial to getting good estimates in our models. However, there is one potential solution to this problem: If we can identify clusters of riders in the data set that give us nearly the same information as grouping by individuals did, these could be provided by Ride Report without nearly the same level of risks to user privacy as identifying individual riders.

In this chapter, we identify predictors that differentiate types of riders and then use these variables to identify clusters of riders. To tie these predictors into our model, we test to see how these predictors do as rider-level predictors for the intercepts and even some coefficients. We also compare random intercepts with riders to random intercept models done by cluster.

Characterizing riders

We characterized riders based on their rides. Ride Report does not collect data about their users besides their rides and email address. We limited our exploration to riders that had over 20 rated rides, because we wanted to focus on riders who had been using the app for some time and had an identifiable pattern of rides.

Cyclists patterns in their rides are complex, particularly their time patterns. Computing their mean ride length for weekends was useful, but mean time of day for their rides does not capture anything meaningful. So we took care in selecting features that distinguished different rider patterns we saw when exploring the data and that did not have high covariance.

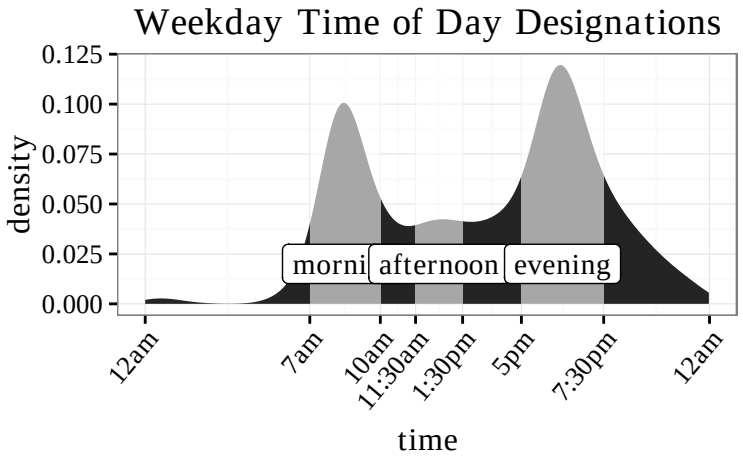

Figure 1: These intervals define the time designations we used in clustering. The proportion of each riders rides in each of these time intervals made up three of our features.

First, we define the collection of cyclist j's rides as $H_j = \{i | j[i] = j\}$. This index set can be partitioned into the rides that occured on the weekend, $H_j^{\text{weekend}} = \{i | i \in H_j, x_i^{\text{weekend}} = 1\}$, and those that occurred on weekdays, $H_j^{\text{weekday}} = H_j \setminus H_j^{\text{weekend}}$. Then we define the following for each rider j: frequency of rides[1] (u_j^{freq}), proportion of rides on weekdays (u_j^{weekend}), median length of rides on weekends ($u_j^{\text{med.len}}$) and weekends ($u_j^{\text{med.len.w}}$), variance of ride length on weekdays ($u_j^{\text{var.len}}$) and weekdays ($u_j^{\text{var.len.w}}$), and proportion of weekday

[1] We define frequency of a cyclist's rides as the number of rides divided by the difference between the time of the most recent ride and time of the first ride. (Units are arbitrary, because we standardized all of our rider-level variables.)

rides during morning rush (u_j^{morning}), lunch rush (u_j^{lunch}), and evening rush (u_j^{evening}). The time intervals that describe the morning, lunch, and evening rush are shown in Figure 1.

Selecting variables for a cluster analysis is difficult, for the reason that many choices about what to use are arbitrary. We have chosen these variables, but two important other choices remain: what scales and transformations should these variables have? We chose here to transform all variables to be approximately gaussian—eliminating the right skew that was present in most of these features with log and square root transformions—and standardizing them by subtracting their mean and dividing by their standard deviation. Future research may find more appropriate ways to select features for clustering, but in our approach here we stick to a naive and simple approach to see what we can learn.

With the rider-level predictors in hand, we clustered the riders using the k-means algorithm, using Euclidean distance as the metric. To choose the number of clusters, we assessed the total of the sum of squares within each cluster for different values of k, shown in Figure 2, and selected $k = 4$ as the point where we thought after which there was little value in more clusters.

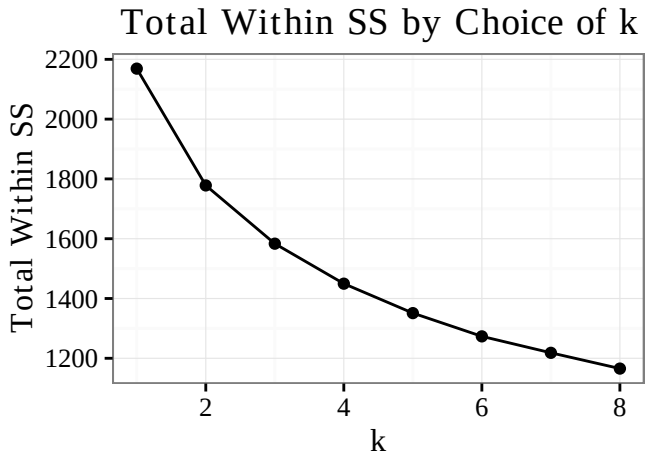

Figure 2: Total within sum of squares of each cluster, by number of clusters (k).

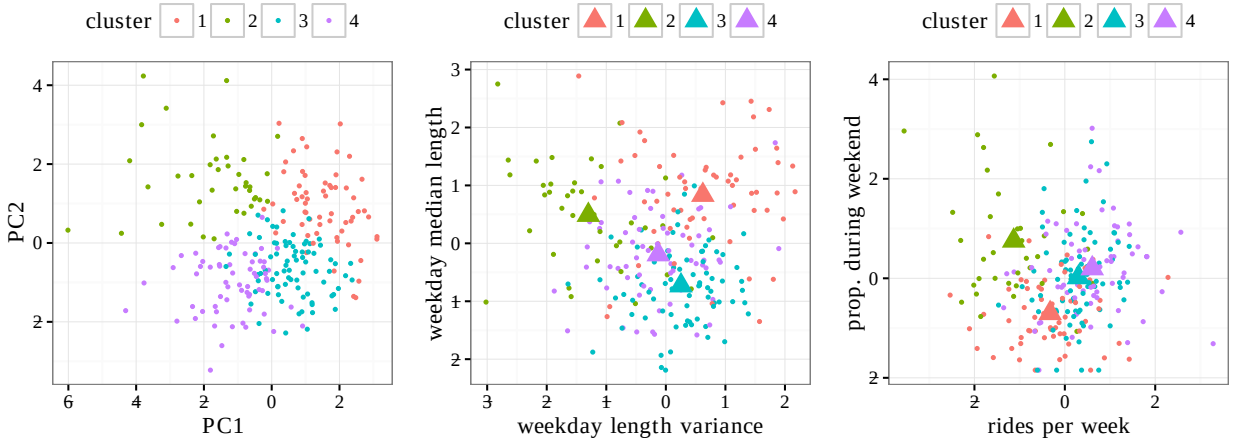

Figure 3: Rider clusters identified by k-means clustering. The triangles represent the centroids (computed as the mean) of the cluster members.

Looking at Figure 3, the clusters split the data into the four quadrants of the first two principal components of the rider data. This makes sense, given our scree plot only showed very differences in total within sum of

squares at k got larger. Despite having failed to identify distinct cluters, studying these different groups, can still off some valuable insight into how cylists differ in their ride patterns.

Figure 4 shows the complex patterns the different groups display. Clusters 1 and 3 seem to be the groups that are the most consistent cummuters, but are differtiated the the typical length of their weekend. Clusters 2 and 4 show much more variance in the timing of their weekdays rides, with cluster 2 having more consistent long lengths for their weekend rides.

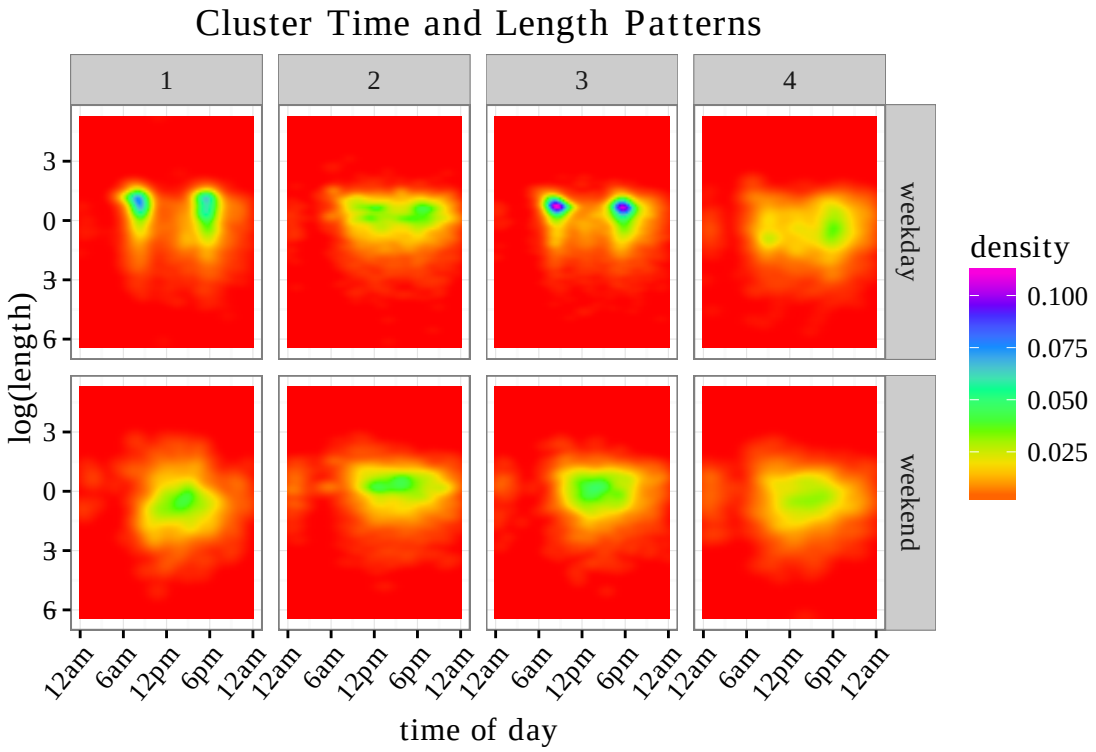

Figure 4: Patterns of ride length and ride time of day for each cluster..

Models with Rider-level predictors

We now have several variables that differentiate riders. How well do they predict our rider intercepts?

Now let U_j be the vector of rider-level variables. Then our model will be

$$Y_i \sim \text{Bernoulli}\left(\text{logit}^{-1}(\alpha_{j[i]} + X_i\beta)\right), \tag{1}$$

where,

$$\alpha_j \sim N(\gamma_0 + U_j\gamma) \tag{2}$$

This model should be comparable to Model 2.[2]

[2]Though we would prefer to use a model similar to Model 4 from the previous chapter (the one with smoothing splines for time of day) the current additive mixed models package `gamm4` (which uses `lme4` to fit the mixed models part) does not support estimating the variability in group-level estimates. Instead, we fit these models in Stan, a probabilitic modeling language that does full Bayesian statistical inference with Markov-chain Monte Carlo sampling. Unfortunately, in this package, smoothing splines would have to be coded by hand and we lacked the expertise to write the functions to fit smoothing splines ourselves.

Table 1: Estimates of rider level predictors.

Parameter	Estimate	2.5% percentile	97.5% percentile
γ^{freq}	0.08	-0.19	0.35
γ^{weekend}	-0.13	-0.50	0.35
γ^{morning}	0.06	-0.22	0.34
$\gamma^{\text{afternoon}}$	0.13	-0.20	0.44
γ^{evening}	-0.02	-0.31	0.27
$\gamma^{\text{med.len}}$	0.01	-0.25	0.29
$\gamma^{\text{med.len.w}}$	0.08	-0.19	0.36
$\gamma^{\text{var.len}}$	0.07	-0.15	0.31
$\gamma^{\text{var.len.w}}$	-0.15	-0.47	0.17
γ_0	-2.99	-3.29	-2.69
σ_α	1.47	1.27	1.69

The ride-level predictor coefficients from the fitted model, are unimpressive. The variance in the rider intecepts not captured by the predictors, quantified with σ_α, is high.

Cluster Intercepts Versus Rider Intercepts

Do these clusters provide similarly useful information that we got from introducing rider intercepts? Naturally, we expect that a model with cluster intercepts will do worse than rider intercepts simply by virtue of less flexibility. Having only four different intercepts rather than hundreds doesn't sound like a recipe for success. But perhaps there is still a significant benefit.

There isn't. We computed Model 7, which is identical to Model 4 from the previous chapter, but has random intercepts by cluster rather than rider. Model 7 performed slightly better than Model 6—which only had a fixed intercept—but nowhere near as well as Model 4. The separation plots, $\log(\mathcal{L})$, AIC, and AUC measures, shown in **??**, all demonstrate this clearly.