**Customer Segmentation of Capital Bikeshare Riders**

*Overview*

The Capital Bikeshare shared-ride system is an increasingly popular way for both locals and tourists to move around Washington, DC. This analysis looks at segments within the Capital Bikeshare data to understand how and when users are riding the bikes.

Why would we do that? From a business perspective, it will be interesting to see which casual customers “behave” like members, and to identify if there are ways to convert these casual riders to full-paying members.

*Data Access*

Our first step is to get our hands on Capital Bikeshare data. To encourage analysis, Capital Bikeshare freely publishes their data on their website [here](https://www.capitalbikeshare.com/system-data). As we see below, this data contains

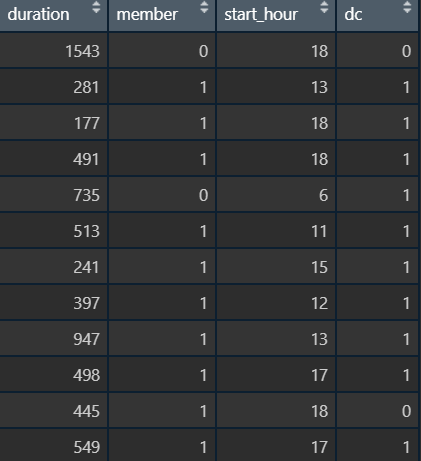
* Duration: the time, in seconds, of a trip. The mean (average) time of a trip is 18 minutes, while the median time for trips is 11 minutes.
* Start\_date: indicates the date and time a user started their trip
* End\_date: indicates the date and time a user ended their trip
* Start\_station\_number: the number ID of the station the user took the bike from
* Start\_station: the name of the station the user took the bike from
* End\_station\_number: the number ID of the station the user returned the bike to
* End\_station: the name of the station the user returned the bike to
* Bike\_number: the ID of the specific bike used
* Member\_type: “Member” or “Casual”. These member types are explained below
* File: This was my own creation so I could identify which file the info came from!



Not all of this information is useful, however. And much of it is categorical, which will not work very well for trying to understand segments. Additionally, there were over 7.6 million rows! My single machine can’t handle all that.

So, I broke out a random sample of 100,000 rows of data, and reformatted it to fit my purposes. My new data has any fewer columns, which are:

* Duration: the time, in seconds, of a trip (same as above). The mean and median times of the trip remain 18 minutes and 11 minutes, respectively
* Member: a binary variable indicating whether the user was a member or not
* Start\_hour: the hour a user began their trip (0 = 12 am, 20 = 8 pm, etc)
* DC: a binary variable indicating whether the user started their trip in DC



*Exploratory Analysis*

Now, we can conduct some exploratory analysis to find any simple patterns in the data. The first thing I’m interested in are the times that these bikes are being used, and how long they are being used for.

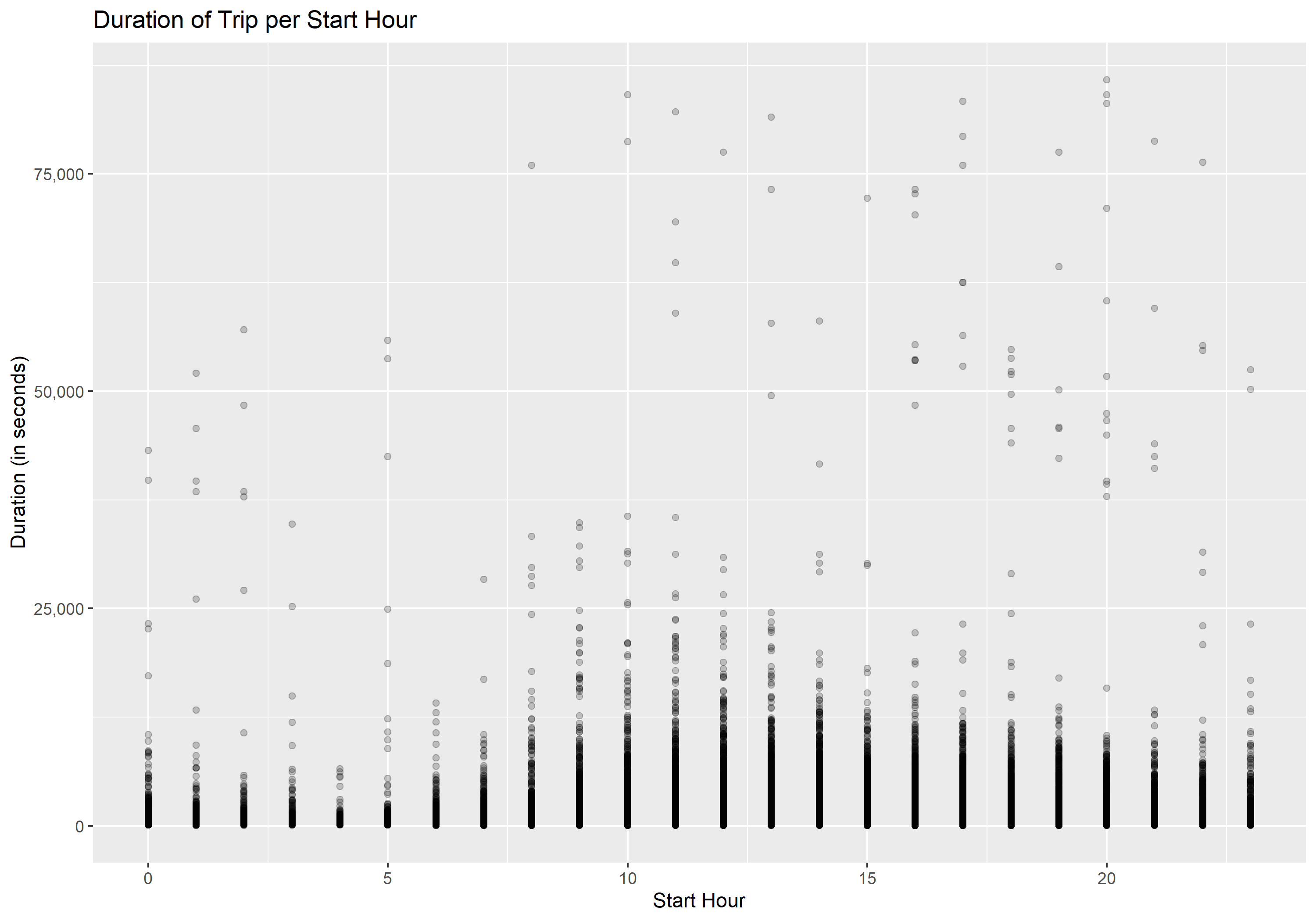
It is important before we begin to discuss the member types and the pricing structure of Capital Bikeshare. This will explain some of the trends we see below. So, there are two types of riders: member, and casual. In both membership types, once a user takes out a bike they have 30 minutes free and then must pay an additional fee for every minute they ride over 30 minutes.

When users are MEMBERS they pay a yearly fee to ride on bikes whenever they choose (only paying additional fees if they ride a bike for more than 30 minutes).

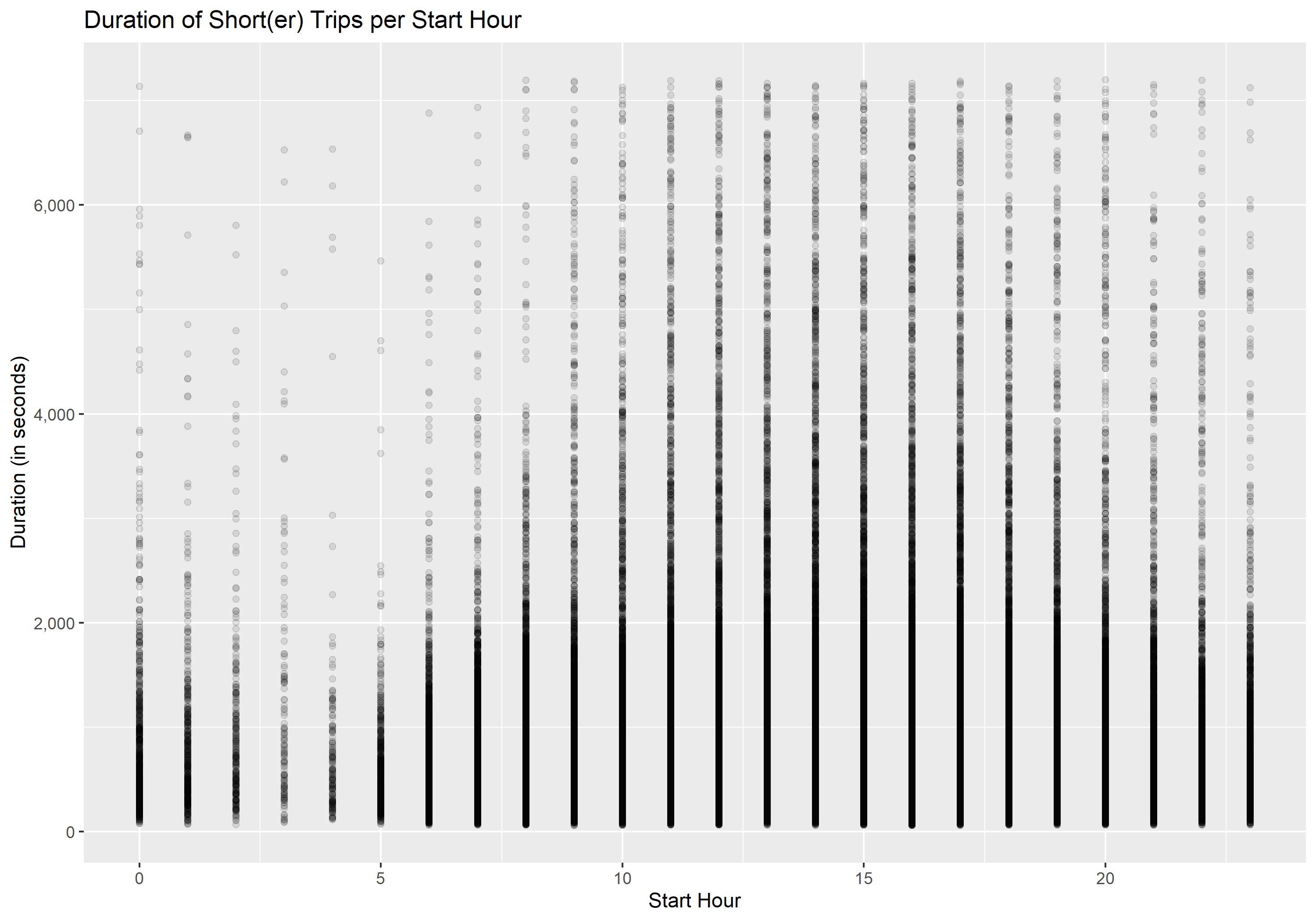
When users are CASUAL they have two options: pay a one-time fee to unlock a bike, or pay a day-use fee to take pre-paid 30 minute rides throughout a certain time period.

Very few people take rides over that 30 minute limit. In the sample of 100,000 rides, only 80 members and 98 casual riders exceeded 30 minutes on a given trip. Understanding Capital Bikeshare’s fee structure will help explain why *so many* rides are below 30 minutes – riders have already prepaid for these trips.

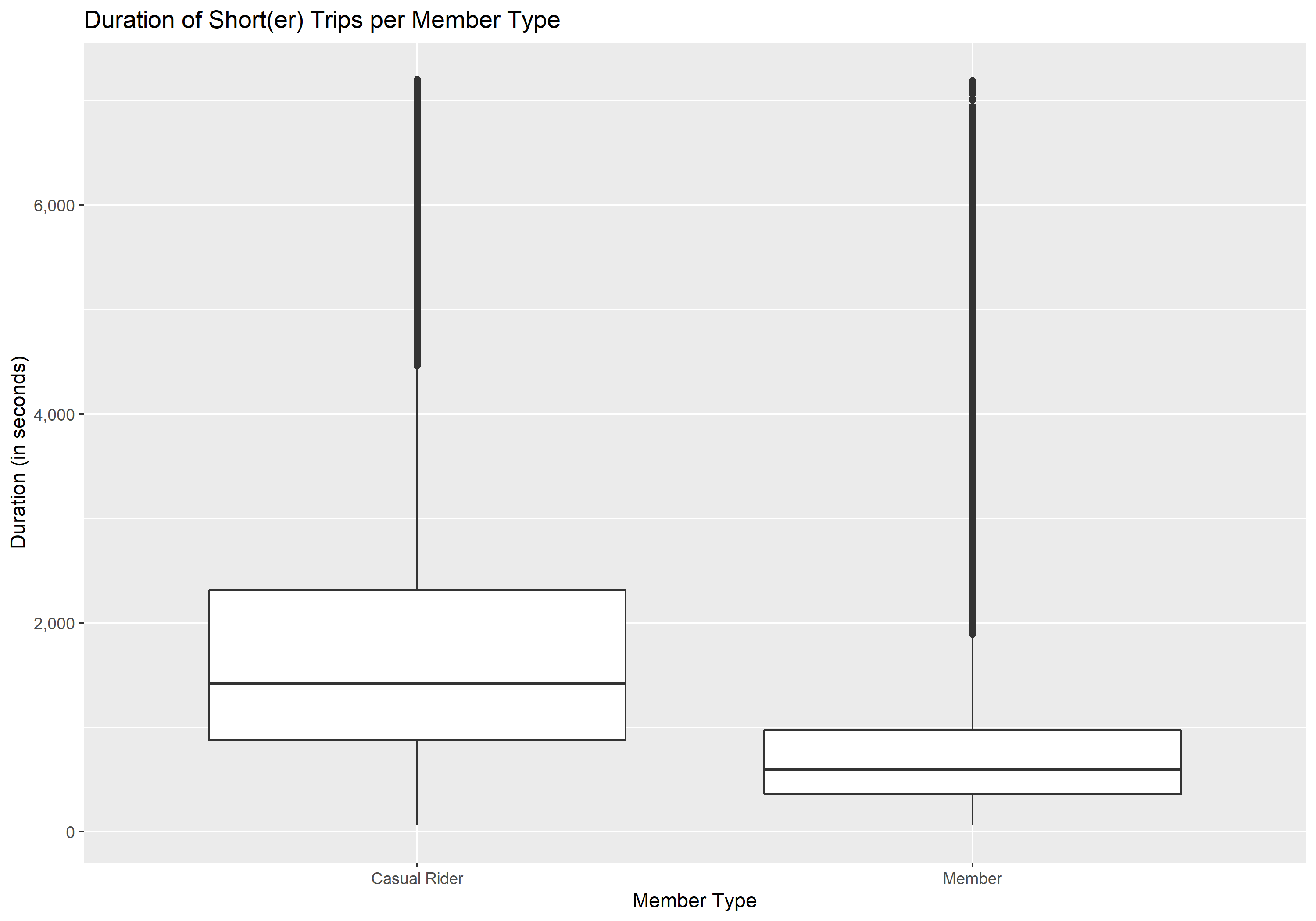
The graph below does show, however, that some users keep do bikes for hours on end (the longest recorded trip in the data was 23.8 hours!)



When we zoom in a bit to rides shorter than 2 hours, we see a trend more clearly: people take shorter rides in the early morning (between 1am and 6 am) and the longest rides in the late morning (between 10am and 1pm).



But who is making these trips under 2 hours? If we break down the duration of trips by member type instead of starting hour, we see that, on average, members tend to take shorter trips than casual riders. In fact, 99.6% of members take trips within 30 minutes while only 94.9% of casual riders take trips of that length or less.



From Capital Bikeshare’s perspective, these casual riders are actually the money-makers. Remember, only rides over 30 minutes incur extra fees – the others are covered by prepaid fees (whether yearly membership, day-use, or one-time-use fees).

|  |  |  |
| --- | --- | --- |
|  | **Member** | **Casual** |
| **Under 30 Mins** | 82,749 | 16,091 |
| **Over 30 Mins** | 306 | 854 |

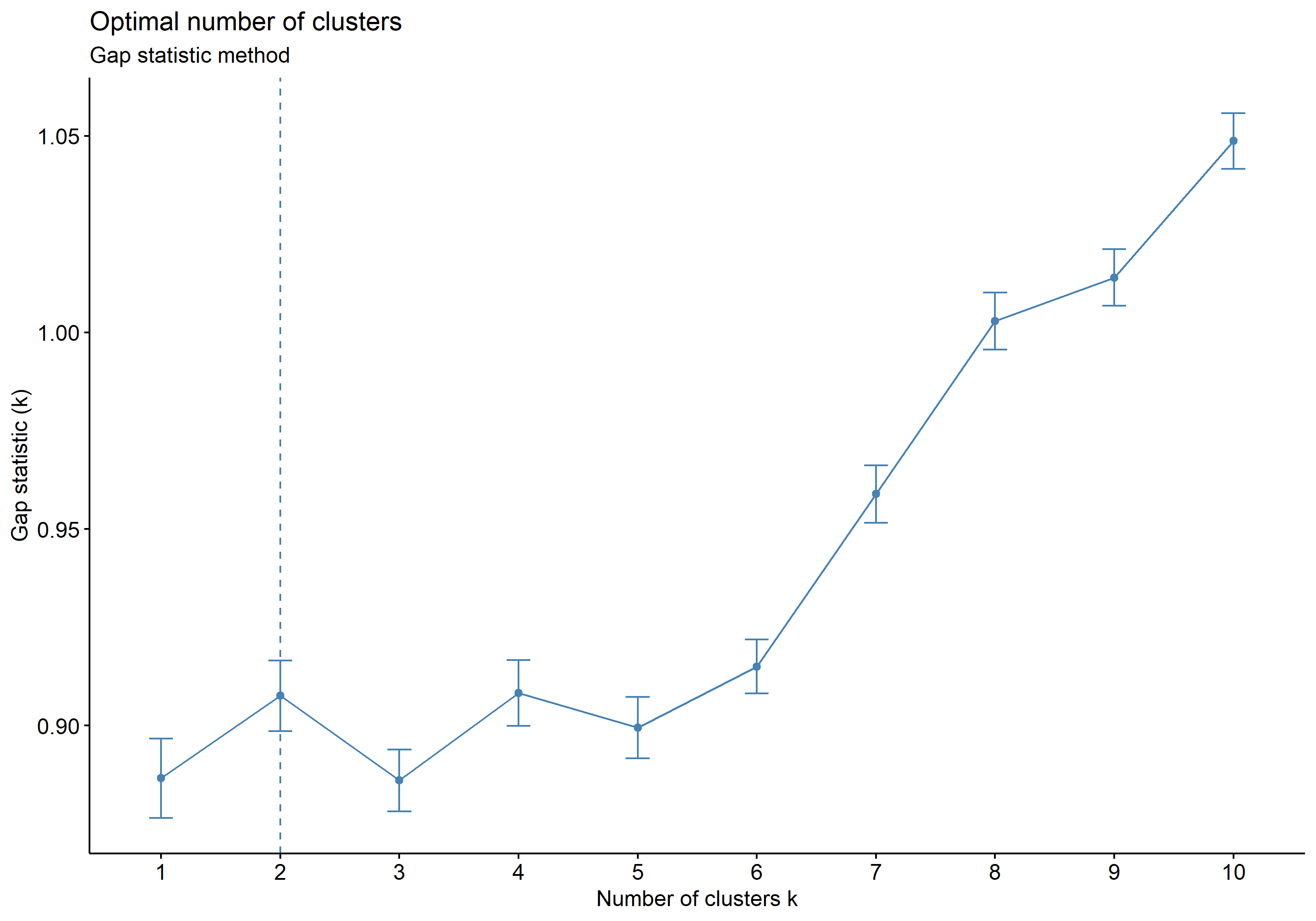
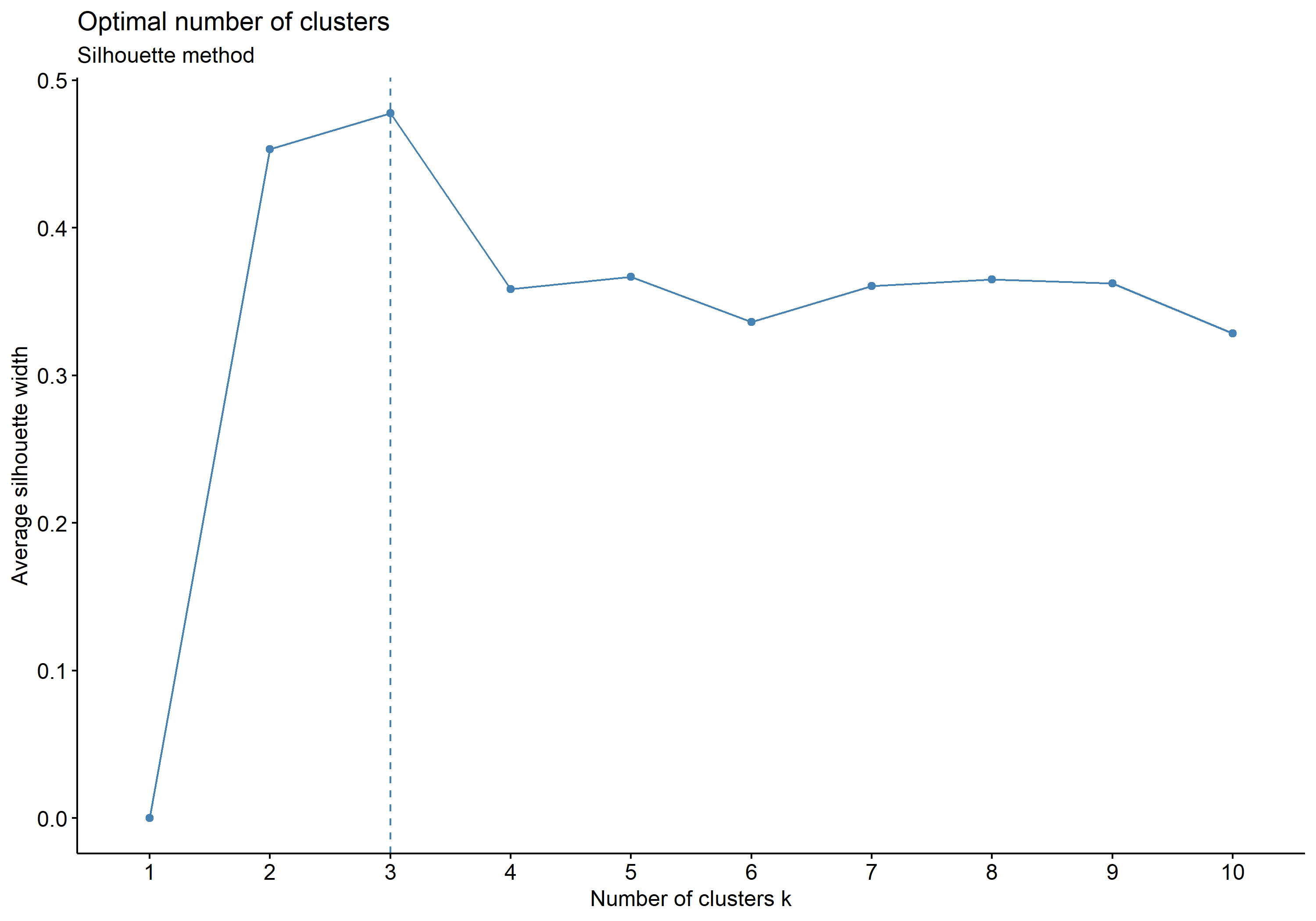
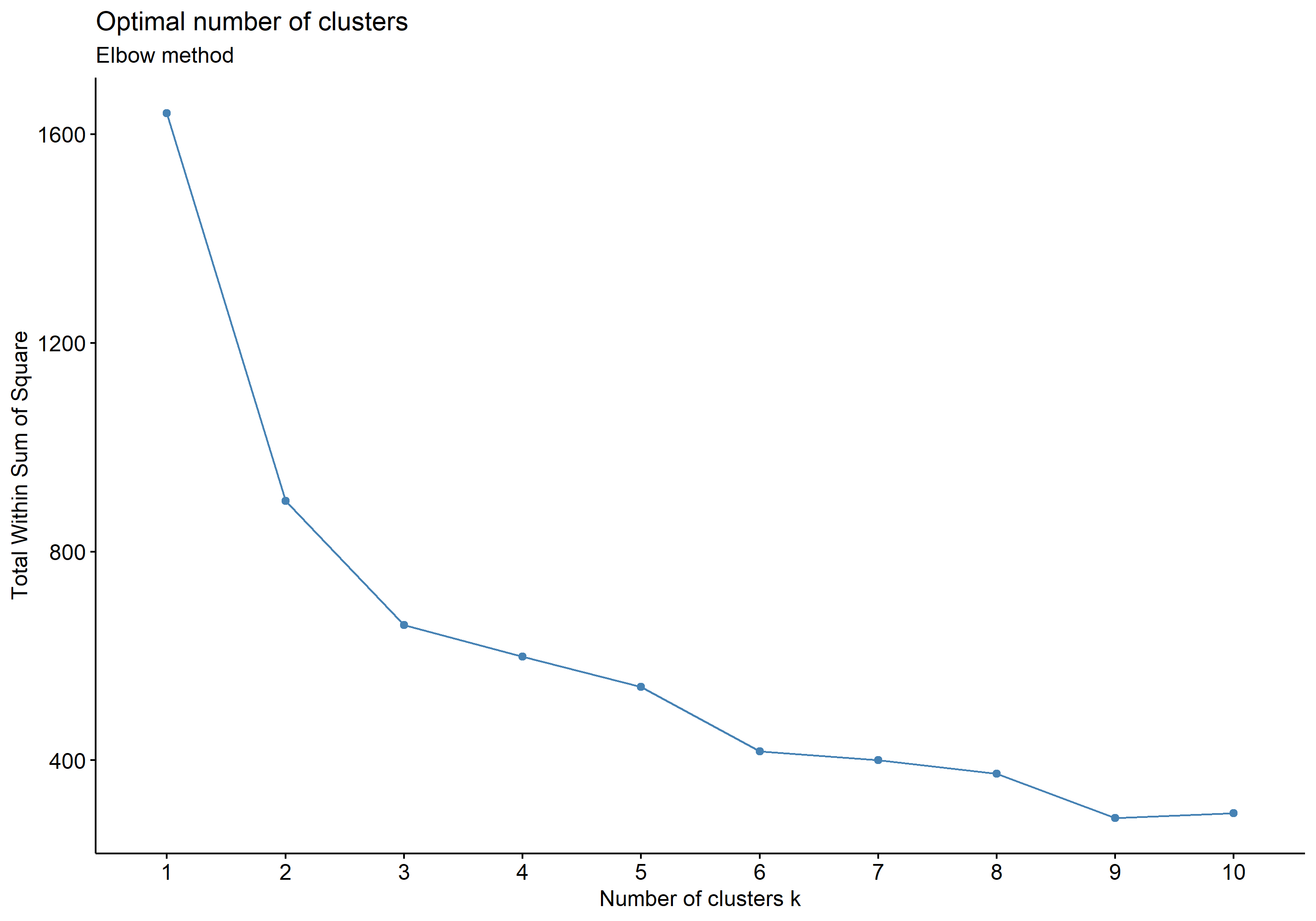
*Clustering (Segmentation)*

Ok, so now we have a pretty good understanding of what is going on in the data. Now it’s time to try and segment it into clusters we can analyze a bit further.

We do this using two techniques: k-means and PCA. K-means is one of the most commonly used clustering algorithms because of its speed and relative accuracy. PCA (Principal Components Analysis) is also widely used because of its reliability. With PCA, you get independent features that you can choose to explain variability.

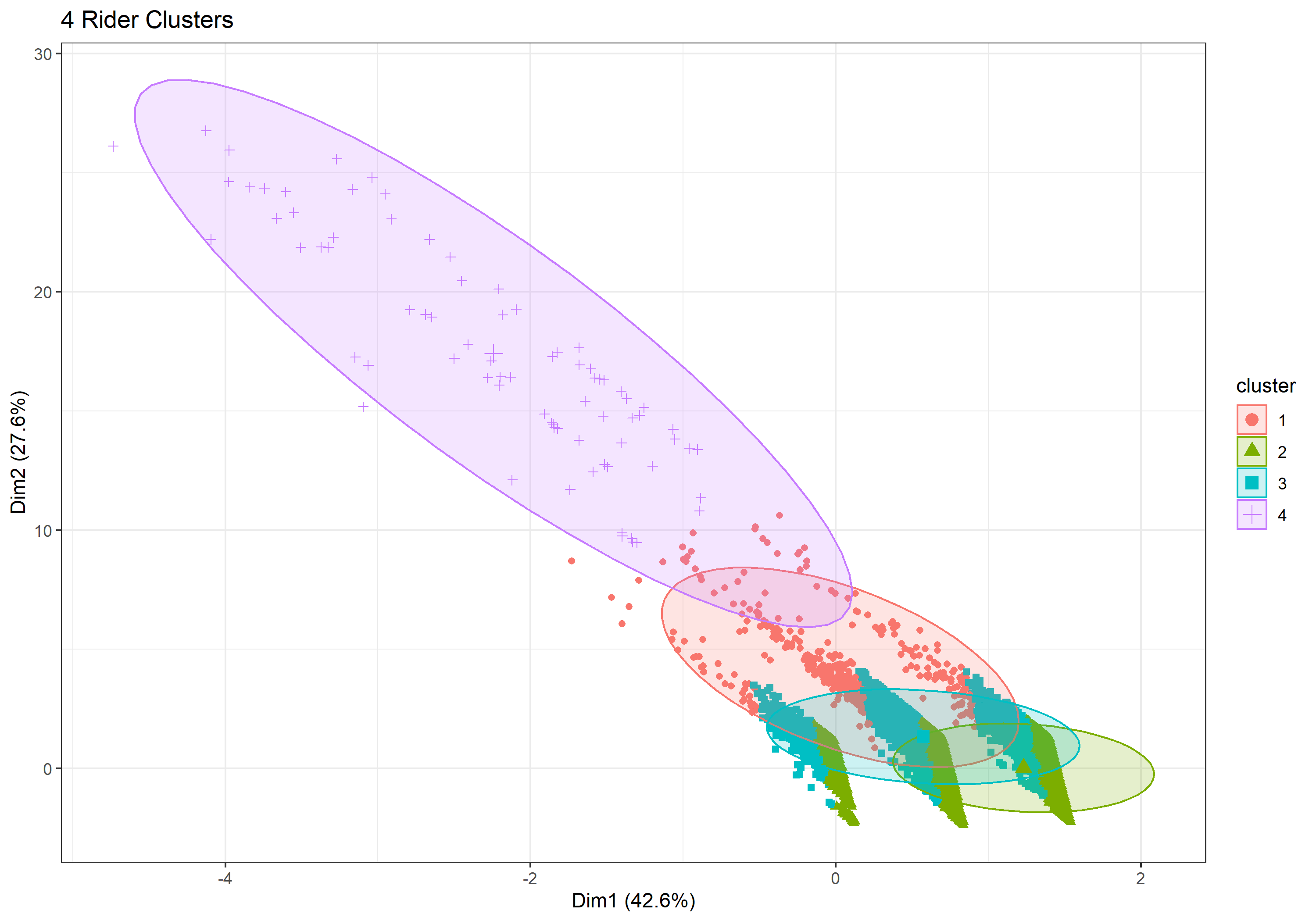
By using both algorithms, we can do a sanity-check off of them. As always with clustering algorithms, we use our common sense and check against our knowledge of the data going in.

To get started with K-Means, we need to choose the number of clusters. This is the most challenging part of k-means clustering, but we make use of three methods to help: the elbow method, the silhouette method, and the gap statistic method. Below, you can see the three graphs representing each method’s choice for k.

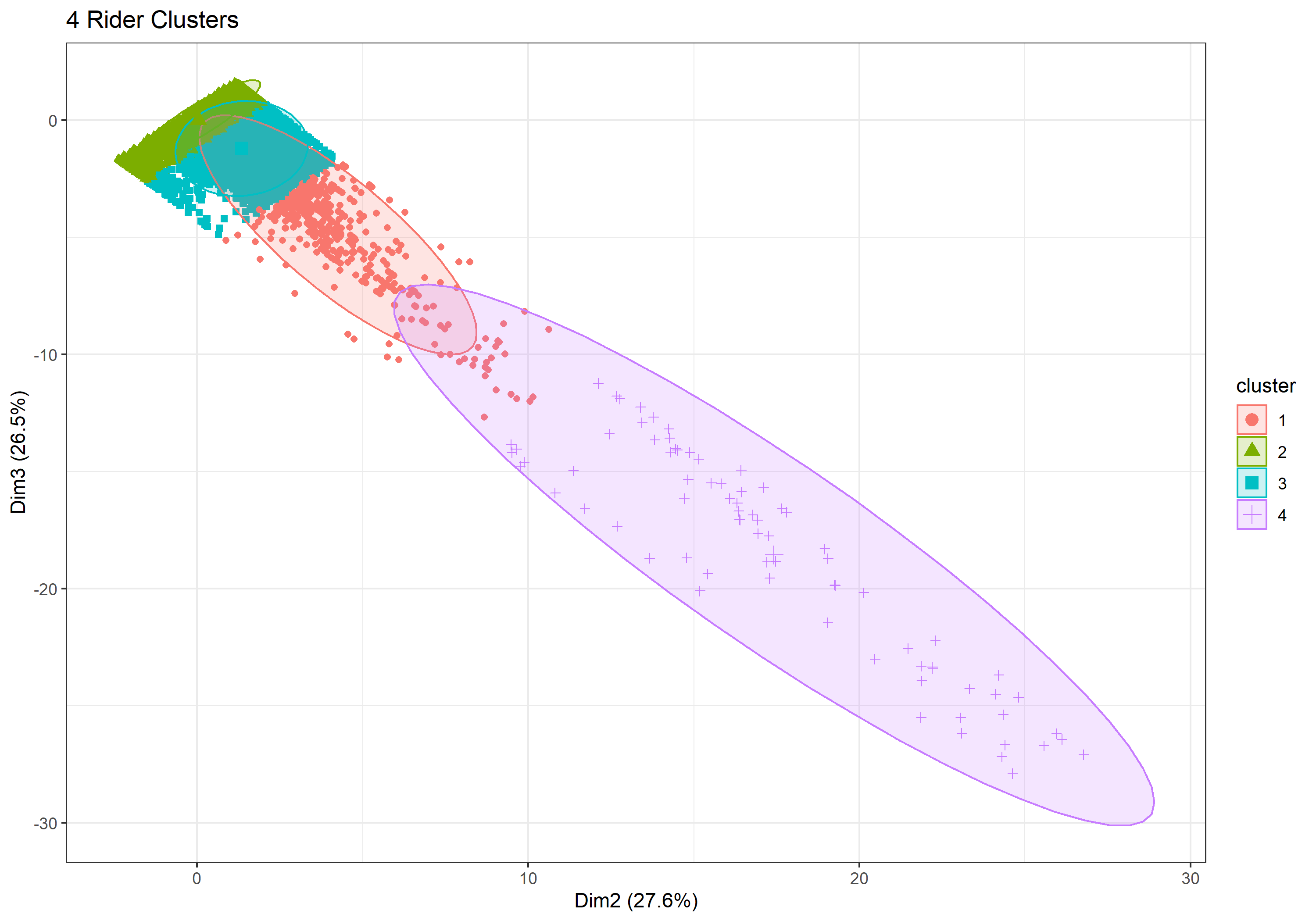


From these, I surmise that I want either 3 or 4 clusters. Because I had the advantage of visualizing both, I know that 4 clusters is more natural for this data, so I move on with k = 4!

If we cluster our data into 4 groups using k-means and visualize it using the traditional axes (1 and 2), we see the following:



These do not look like a natural cluster, however – if I were to do it by hand I would keep cluster 4 but break out the other chunks into their own cluster. Which shows how important axes and visualization are for clustering analysis in greater than 2 dimensions. Look at the same data visualized on the axes 2 and 3.



This is a much more natural visualization to understand the 4 clusters we’ve broken our Capital Bikeshare riders into.

*Results*

So now that we have our clusters, what can we say about them? To make sense of it all, we look at the mean of each of the 4 groups identified.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Duration | Member | Start Hour | DC |
| **Group 1** | 6.41 | 0.29 | -0.22 | 0.87 |
| **Group 2** | -0.14 | 0.86 | -0.006 | 0.86 |
| **Group 3** | 1.67 | 0.28 | 0.13 | 0.83 |
| **Group 4** | 25.59 | 0.5 | 0.19 | 0.87 |

I’ll name each one of these segments and describe their characteristics.

*Group 1: Early morning tourists*

* Rides in this group are more likely to be taken in the morning by casual members. They are of relatively long duration, indicating that they are maybe rides being taken by tourists to sightsee.

*Group 2: Commuters*

* Rides in this group are short, early, and taken by members with year-long passes. These are likely to be commuters using the Capital Bikeshare to get to their offices.

*Group 3: Late afternoon tourists*

* Rides in this group are likely to be taken by casual riders in the afternoon. Interestingly, rides in this group are most likely to be started outside of Washington, DC (in the surrounding suburbs). Due to the prevalence of casual riders but the relatively short duration of the trips, these could be rides taken after lunch to get to a new tourist destination.

*Group 4: Pleasure riders*

* Rides in this group are by far the longest and tend to start the latest in the day. Although I was expecting long rides to skew toward casual riders, rides in this group are equally as likely to be taken by year-long members. Due to the length of the rides in this group, these could be pleasure-seekers out for an afternoon ride.

My initial effort was to identify casual riders that behaved like members so that Capital Bikeshare could expend effort converting them into full-time riders. After conducting my analysis, I no longer think this should even be an effort by Capital Bikeshare, as so many more casual riders incur extra fees! Instead, I found four distinct groups of riders using the Capital Bikeshare system and drew out the characteristics that make them unique.