Data Science Final Project - Phish

Ido Algom & Natalie Gilboa

1 December 2016

Figure 1:

Abstract

Phish is an American rock and roll band noted for their musical improvisation, extended jams, blending of musical genres, and dedicated fan base. The Band Formed at the University of Vermont in 1983 (with the current line-up solidifying in 1985).

In the following report we'll show the spreads of songs the bandhas played over 33 years. We'll see that there is consistensy about songs pick-ups. and most of the songs are played for few years and then syat behind.

Before reading the report it's important to mention a few facts about Phish shows: 1. A show never repeted itself more than once - Each show offers completely different setlist from another. 2. When the band hits a town it can be for 1 to 4 nights in a row. When it's 2 it's called "a Run". Inadditionto fact number 1 it's important to mention that a song will never repeat itself more than once in. But important to mention that in rare shows a song can repeat itself more than once at the same show. (sandwich song)

In the report we'll use song appearances counts and debut years to see how creative the band is. We'll use k-meand to cluster the songs. The reason we chose 3 clusters is that it's the same amount of eras. (description below)

Business Understanding

The Mockingbird Foundation is a charitable organization founded by fans of the jam band Phish in 1996 to support music education for children. The website Phish.net is a fan website that formally adopted by The Foundation as a project in 2005. The site had begun in 1991 and served Phish fans for nearly two decades as static HTML pages. It was re-launched as an integrated database - of setlists, song histories, reviews, and more - in fall 2009.

The website contains tables about each song that ever played in show. Every single show the band has been playing during the years has a unique setlist. This information about each song and each show is recorded at the website database for fans.

The report's targets are made for statisitcs only. As mentioned, The Mockingbird Foundation is a charitable organization and the band Phish is not related in any way to the orginization exept in it's careness and appriciation to thier fan base.

Dictinary

Hiatus - a temporal break-up. Phiah went on hiatus 2 times. The first time was on 2001. All the era between 1983 to 2001 is called 1.0 They cameback toghether about a year later in 2002 untill the summer of 2004. This era is called 2.0 and they cameback again to studio and stages in the spring of 2009 untill today. this era is 3.0. More words are described in Data understanding.

Data Understanding

We chose to explore the statistics about the songs. I copied the table about all the songs the band played during the years. The table contains the following columns:

- Song Name
- Original Artist (The band played a lot of cover songs)
- Time Played (At how many shows the song has played)
- Debut (Show date first time played)
- Last Played (Show date last time played)
- Gap (Amount of shows the song hasn't played)

The last show was played in 2016/10/31

Let's get some code - preparations

```
require(gtable)
require(ggplot2)
library(gtable)
library(ggplot2)
```

Let's read the file and show the table header

```
yem <- read.csv("phish.csv",stringsAsFactors = FALSE,colClasses = c(NA,NA,NA,"Date","Date",NA))
head(yem)
```

```
##                Song.Name Original.Artist Times.Played      Debut
## 1            Army of One           Phish           17 2003-07-12
## 2            Bathtub Gin           Phish          247 1989-05-26
## 3                    Bug        Amfibian           63 1999-06-24
## 4 Crosseyed and Painless   Talking Heads           37 1996-10-31
## 5      Down with Disease           Phish          261 1993-12-31
## 6              Farmhouse           Phish           62 1997-11-07
##    Last.Seen Current.Gap
## 1 2016-01-17           0
## 2 2016-01-17           0
## 3 2016-01-17           0
## 4 2016-01-17           0
## 5 2016-01-17           0
## 6 2016-01-17           0
```

Total amount of songs we have:

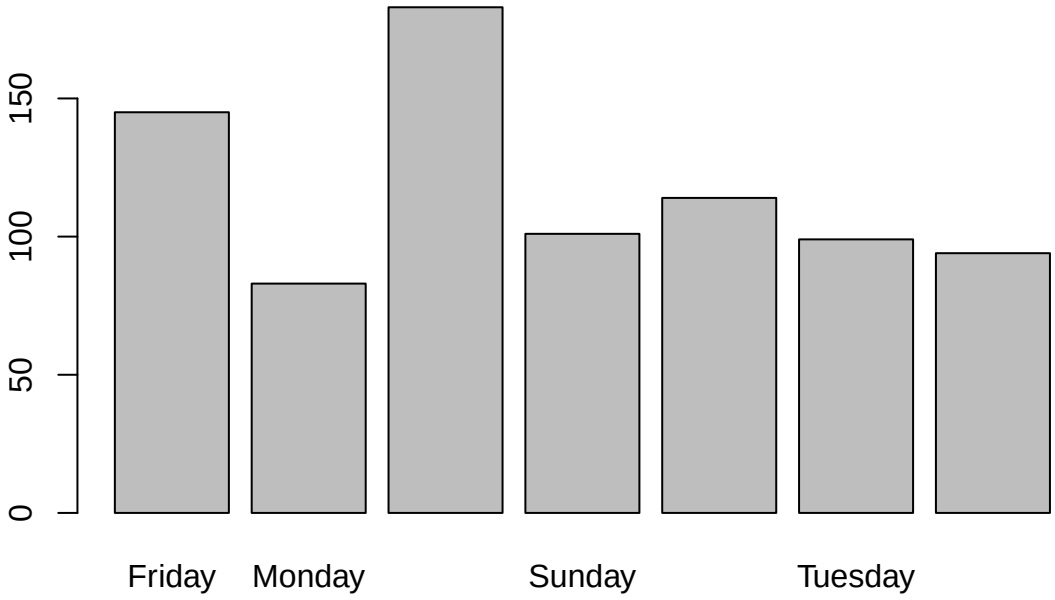
```r
nrow(yem)
```

```
## [1] 819
```

Basic Information

How many songs debuted at each day of the week. In this Graph we took the debut day and converted it to Date and to Weekdays with basic R functions.

```r
barplot(table(weekdays(as.Date(yem$Debut))))
```



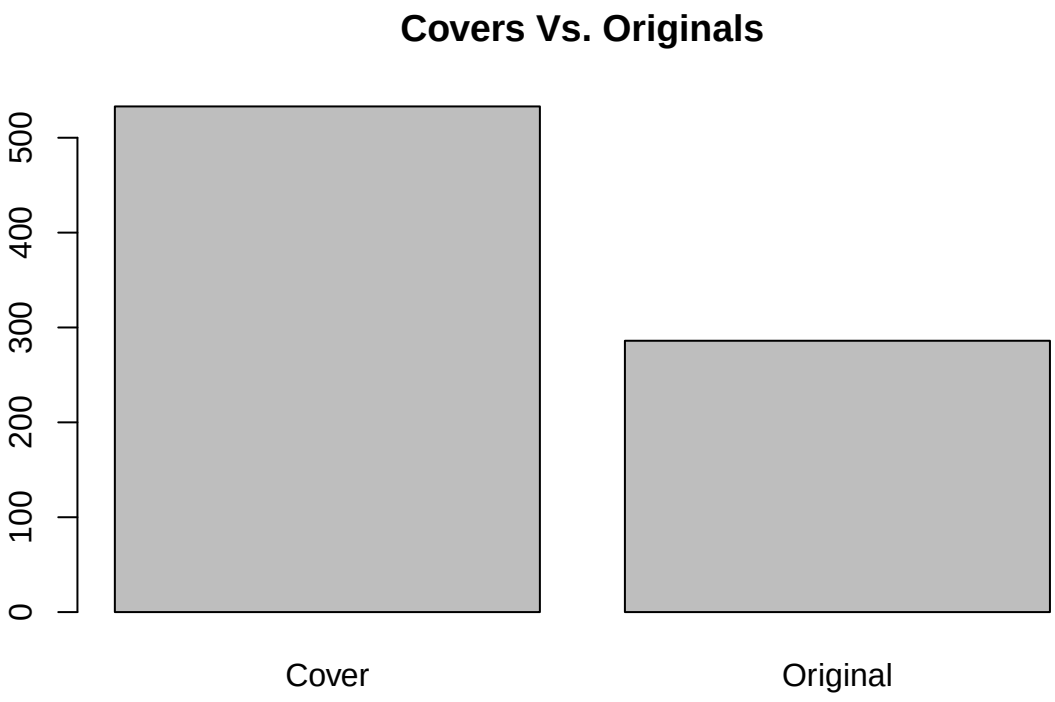
Songs Left Behind

In this Scatterplot we extracted the year from every song, when debuted and last played. The color shows if more songs this year was a cover or original. As stronger the color is, more frequent the group repeats in that point. From this graph we learn that a lot of covers are not repeating themselves. The band play it during it's debut year and than leave it behind. Another thing we can learn is that most of the original songs are not left behind. There're some songs that did left in the past before 1990, but the rest are keeping pooping out at shows.

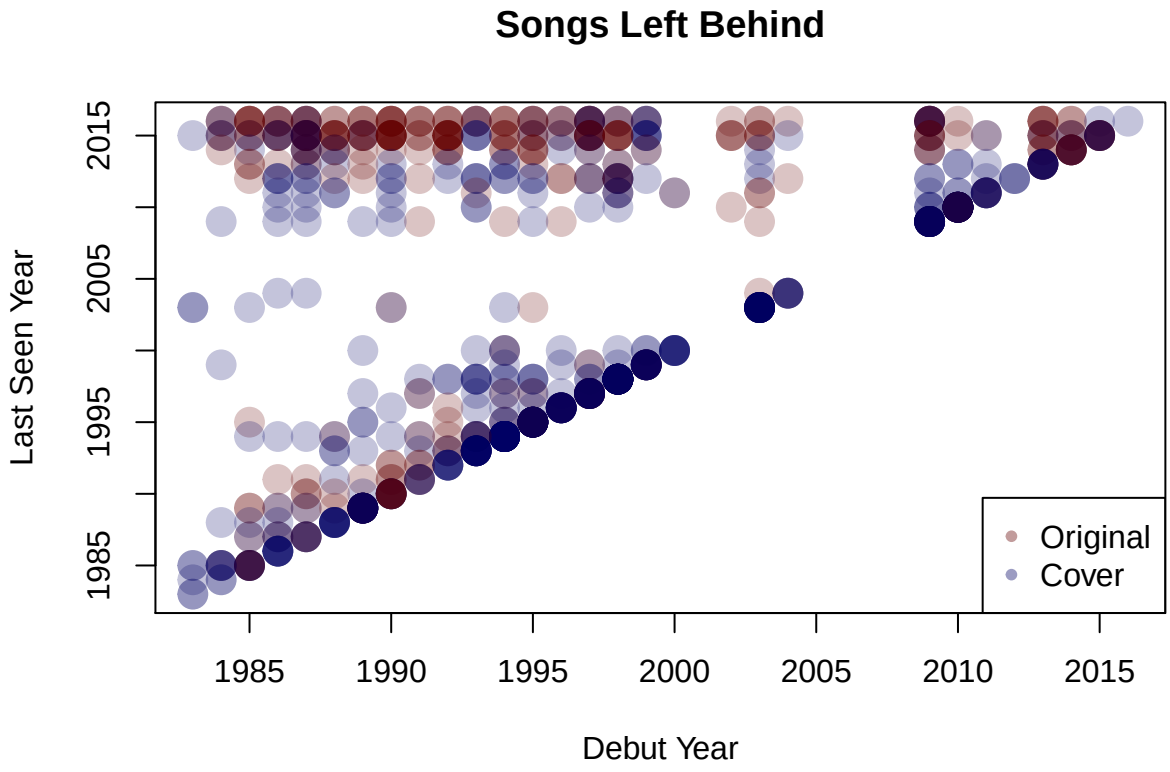
Some preparations

```r
cover <- format(yem$Original.Artist)
cover2 <- rep("Cover",length(cover))
cover2[grepl("Phish",cover)] <- "Original"
```

```r
barplot(table(cover2),main = "Covers Vs. Originals")
```



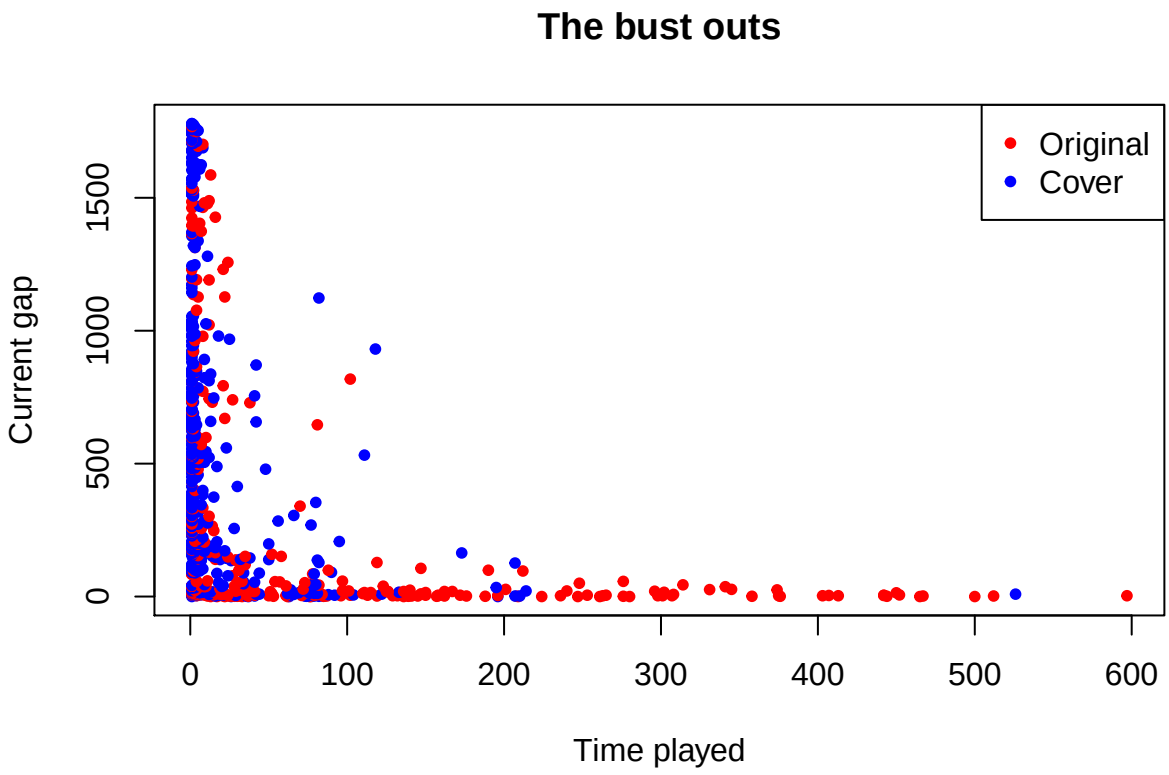

```
plot(as.numeric(format(as.Date(yem$Debut), "%Y")),as.numeric(format(as.Date(yem$Last.Seen), "%Y")), xlab
legend("bottomright",legend =c("Original","Cover"),col=c(rgb(100,0,0,100,maxColorValue=255),rgb(0,0,100
```



Bust out

A bust out is a sudden song appearance in a show with very big gap. From this plot we can learn most of the originals songs are played very often. Most of the covers don't make more than few appearance and there are some songs that used to be played often and now they're gone from the setlist. We can say that as close as a song to p(0,0) the chances to get played at the coming Summer Tour 2016 are higher. As far a song is on the Gap axis as bigger the bust out will be.

```
plot(yem$Times.Played,yem$Current.Gap,col=ifelse(cover2=="Original","red","blue"),pch=20,cex=1,main="Th
legend("topright",legend =c("Original","Cover"),col=c("red","blue"),pch=20)
```



Data Preparation

Right before we're going to explore our information we'll create a data frame to include all the information we need. Since we gonna cluster the songs we gonna need to use numerical values. Thus, we gonna take the var yem and change it a little. The data frame fluffhead is gonna include the following vectors:

- Time played - How many times a song has been played live
- Current Gap - How many shows a song hasn't been played live
- Years since debuted - How many years since this song was debuted live
- Years since last played - How many years the song hasn't been played live
- is Original- Either the song is originally written by the band or is a cover.

```
fluffhead = data.frame(
  "time_played" = yem$Times.Played,
  "current_gap" = yem$Current.Gap,
  "years_since_debuted" = floor(as.numeric(as.Date("2017", format="%Y") - yem$Debut)/365.25),
```

```
  "years_since_last_played" = floor(as.numeric(as.Date("2017", format="%Y") - yem$Last.Seen)/365.25),
  "is_original" = cover2
)
```

Modeling

The model we gonna use is clustering model. We're gonna try to estimate and see groups of songs that left in early era (1.0) and songs that keep popping up at shows. The model will show those those facts from different points of view using K-means algorithm. Each model will be separated by the fact if the song is an original song written by the band or a cover.

Let's create some k-means models:

The first model we'll create will try to show how 2 hiatuses (2001,2004) created the 3 eras the band has: 1.0 (1983-2001), 2.0(2002-2004), 3.0 (2009-current):

```
forbin = kmeans(fluffhead[c("time_played","current_gap","years_since_debuted","years_since_last_played"
```

Let's compare the clusters with the songs

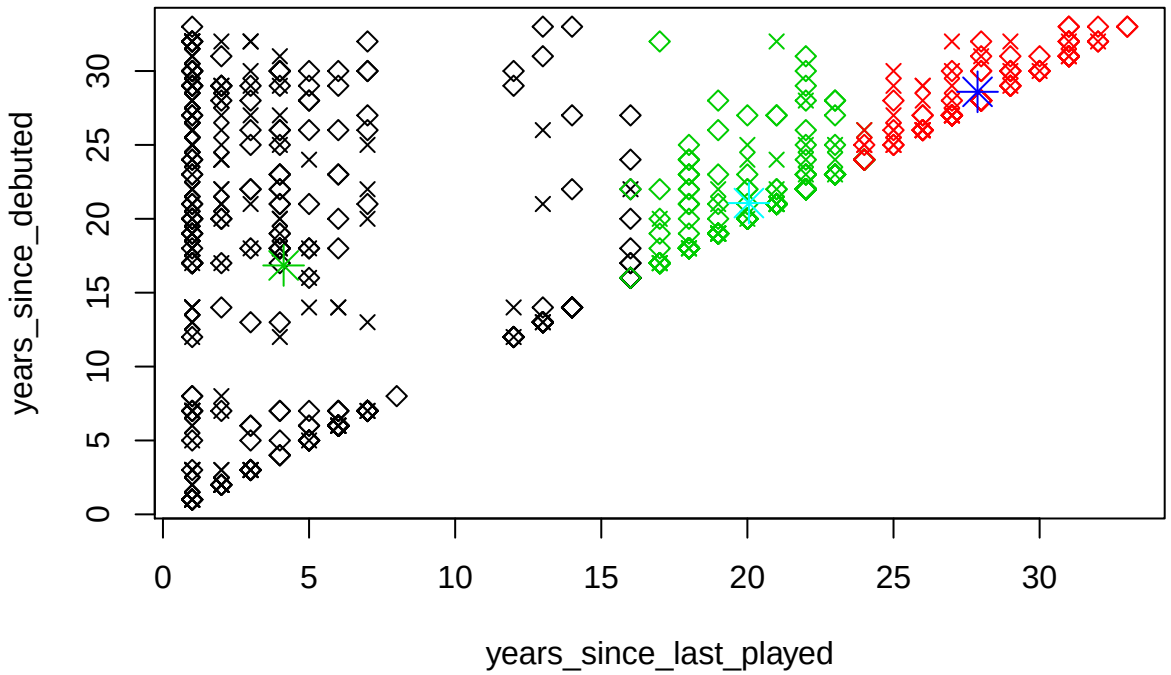
```
table(forbin$cluster, fluffhead$is_original)
```

```
##
##     Cover Original
##   1   243      214
##   2    76       37
##   3   214       35
```

Evaluation

The first evaluation we'll do is to see clustering the songs by amount of years since debuted and amaount of years since last played.

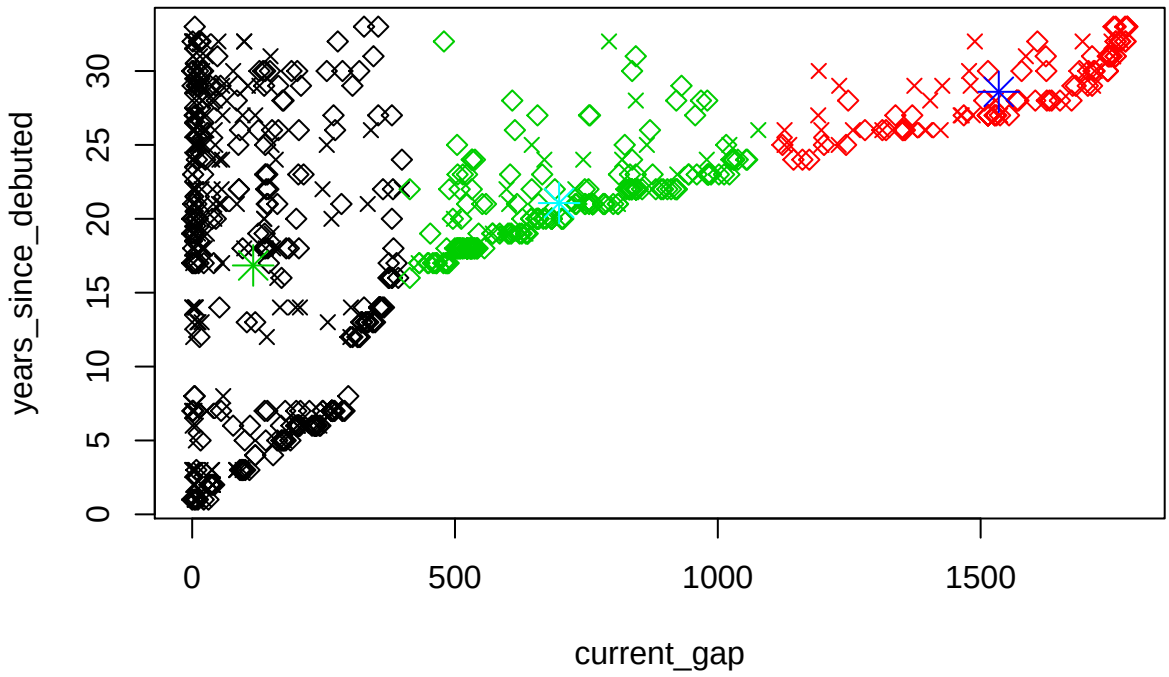
```
plot(fluffhead[c("years_since_last_played","years_since_debuted")], col= forbin$cluster, pch=ifelse(flu
points(forbin$centers[,c("years_since_last_played","years_since_debuted")], col=3:5,pch=8,cex=2)
```



During the firsts years, the band didn't have many originals songs. Actually the first studio album wentout on 1989. What means that for 6 years most of the songs the band has played are covers. We can see it in the plot where x stands for original song the diamond stands for cover. At the top rows we can see that ther're a lot of covers and only one cover which is debuted in that era is still getting played (top left diamond). On the other hand we can see that most of the songs debuted and oftenly played during the last years are originals. and as the years going forward ther'e are less covers. A single Ineresting poing in the plot showing a year between 15 to 20 on the x axis where the band left there a lot of covers she played years befire.

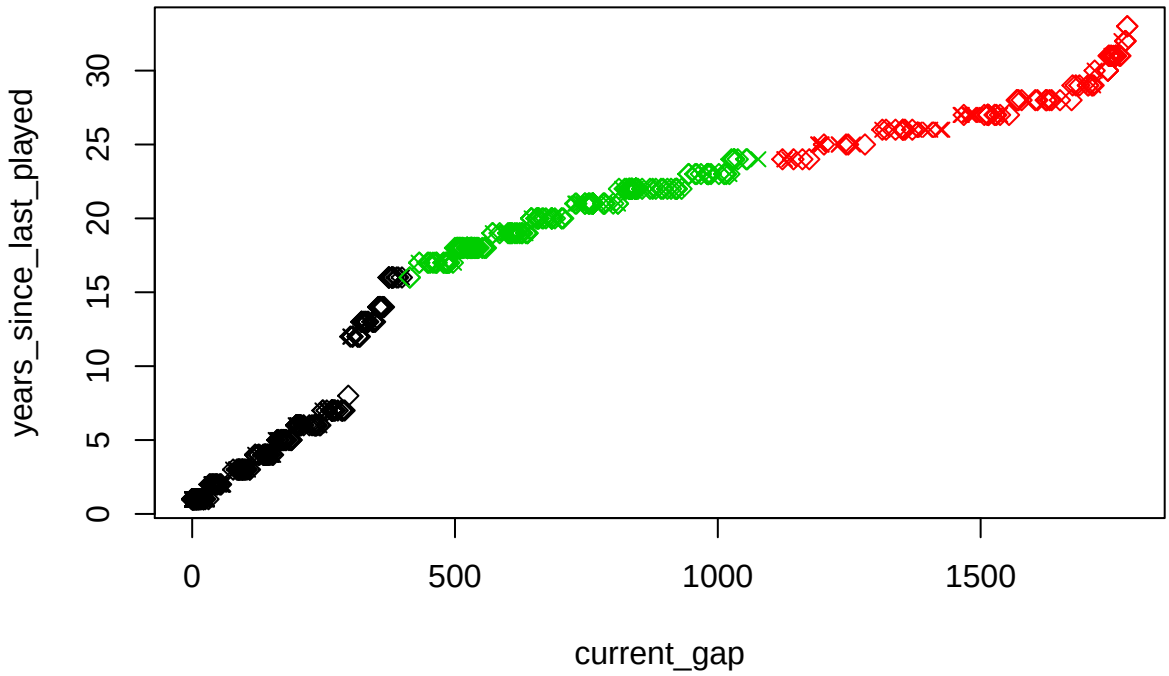
The clustering itself is little suprising. We can see that the black cluster where alot of songs abonded at the early era of 1.0 and another center is around 1996. It's important to mention that only in 1994 the band played over 250 shows, a fact the explains the green cluster centrum. During that a year there was a big variety of songs.

```
plot(fluffhead[c("current_gap","years_since_debuted")], col= forbin$cluster, pch=ifelse(fluffhead$is_or
points(forbin$centers[,c("current_gap","years_since_debuted")], col=3:5,pch=8,cex=2)
```



In the above plot we replaced the x axis to current_gap. we can see that the point are not going linear anymore. This fact is validating our data as the amount of shows is not linear to years. In order to see this curve we can just show all the songs in on the axis of current gap and years since last played:

```
plot(fluffhead[c("current_gap","years_since_last_played")], col= forbin$cluster, pch=ifelse(fluffhead$i
```



Deployment

Summary

Bibliography

- The database - Phish.net
- Wikipedia
- My knowledge and dozens of shows attendences.

Self-Note - Phish is the greatest band for me. A band that changed my entire life and made the person I am. Ido