Analysis of Lottery Draws Between 2009 and 2017

Carlo Carandang, BSChE, MD

Data Scientist

May 22, 2017

Analysis of Lottery Draws Between 2009 and 2017

This project entails the analysis of a dataset of historical lottery draws between 2009 and 2017 inclusive, scraped from the website of a lottery by my colleague, Gregory Horne. We had a question whether the winning numbers could be predicted, based on past draws, but needed to know if the winning numbers clustered, or were randomly drawn.

In this lottery, ping-pong balls are labeled with one number, ranging from 1 to 49. One of each number is placed in a barrel. The barrel is spun to mix up all the balls, then one ball is drawn. This is repeated 5 more times for a winning number set of 6 winning numbers. In addition, there is a bonus draw, which gives 7 winning numbers.

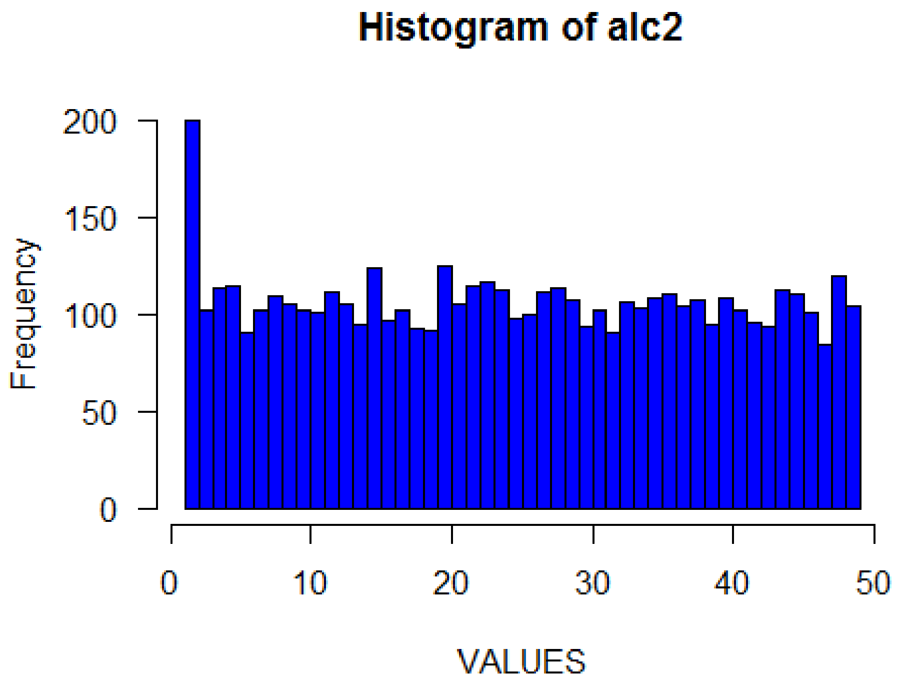
We will first analyze the winning numbers from 2009 to 2015, then add the winning numbers from 2016 to 2017, to see how the analysis is changed with new data. Thus, we will analyze two lottery datasets, one from 2009 to 2015, and the other from 2016 to 2017.

We propose to perform cluster analysis on this lottery dataset. We hypothesize that the cluster analysis should be random, and therefore the datapoints should plot in a uniform manner in the feature space. This hypothesis is based on the premise that this specific lottery draw is indeed random in nature. However, if our analysis leads to clustering that is significant, then this can lead to further analysis and speculation on the method of determining winners for this specific lottery.

Are the winning numbers randomly drawn?

We produced a histogram of the winning numbers 2009-2015 dataset to determine the frequencies of each winning number. Here is the script in R that was used, and the histogram:

```r
alc <- read.csv("C:/Users/carandangc/Documents/Winter 2017 BIA/2. Data Sets/alc_winning_numbers/data/alc_winning_numbers.csv", header=FALSE)
alc2 <- c(alc$V2,alc$V3,alc$V4,alc$V5,alc$V6,alc$V7,alc$V8)
hist(alc2, col="BLUE", ylab = "Frequency",las=1, xlim=c(1,49),xlab = "VALUES", breaks=49)
```



As you can see, the frequencies are not evenly distributed, and one number (#1) tend to occur in much greater frequency, and some other numbers (#14, #19, #48) seem to peak also. So this gives us a signal that this is not random, and can therefore proceed with the analysis to look for clustering.

What is the application of clustering for predictive analysis? How do you predict behaviors from clustering analysis?

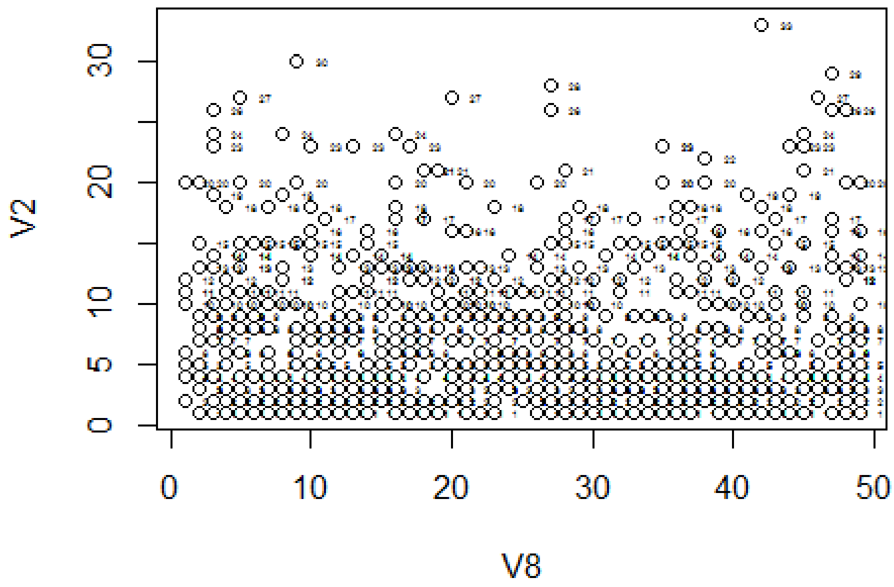
If you can figure out which data tends to cluster, then you can use the features that characterize that dataset, and use those same features to predict which data will fall into that same category/label. In this case study, we want to see if there is clustering for the winning numbers, so that we can identify those same features to predict future winning draws.

For the dataset to be analyzed, does the data tend to separate/cluster, or is the data random with uniform distribution?

We used a cluster dendrogram for the analysis. In the clustering analysis, we used the frequency of the winning numbers to convert categorical data into quantitative data. This is how we can transform categorical data into quantitative data, and the most common way is to use frequency of use. Then this numerical data (that was previously categorical) can be analyzed by the clustering algorithm used.

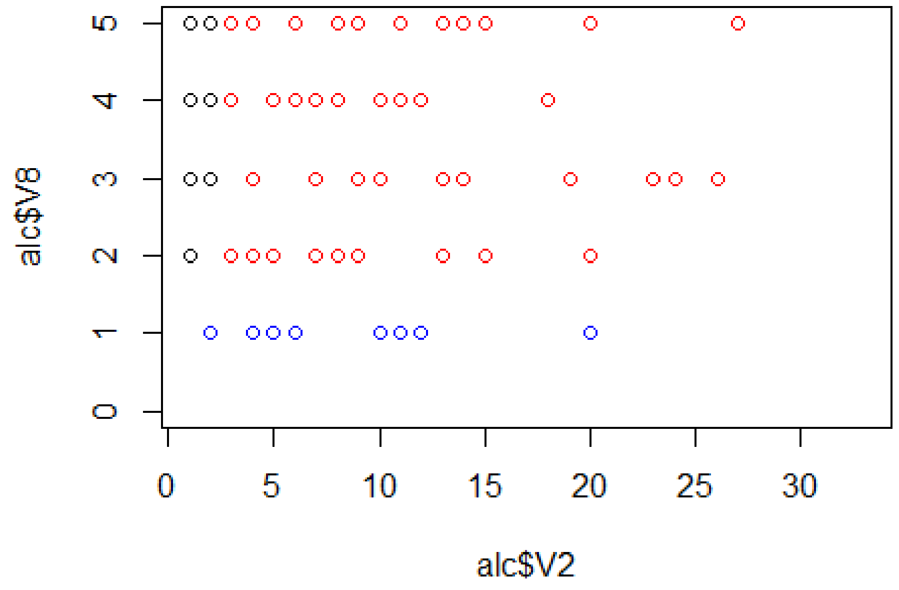
With the analysis of the lottery data from 2009 to 2015, the data does tend to cluster. The data is not random, and does not have a uniform distribution. The dendrogram below shows 2 separate and unequal groups when you cut it off at a height of 4.5, and shows 4 separate and unequal groups at a height of 3.7. Please see the R script and plots of the dendrograms showing the clustering below:

```
#scatter plot
plot(V2~V8, alc)
with(alc,text(V2~V8, labels=V2,pos=4,cex=.3))
```

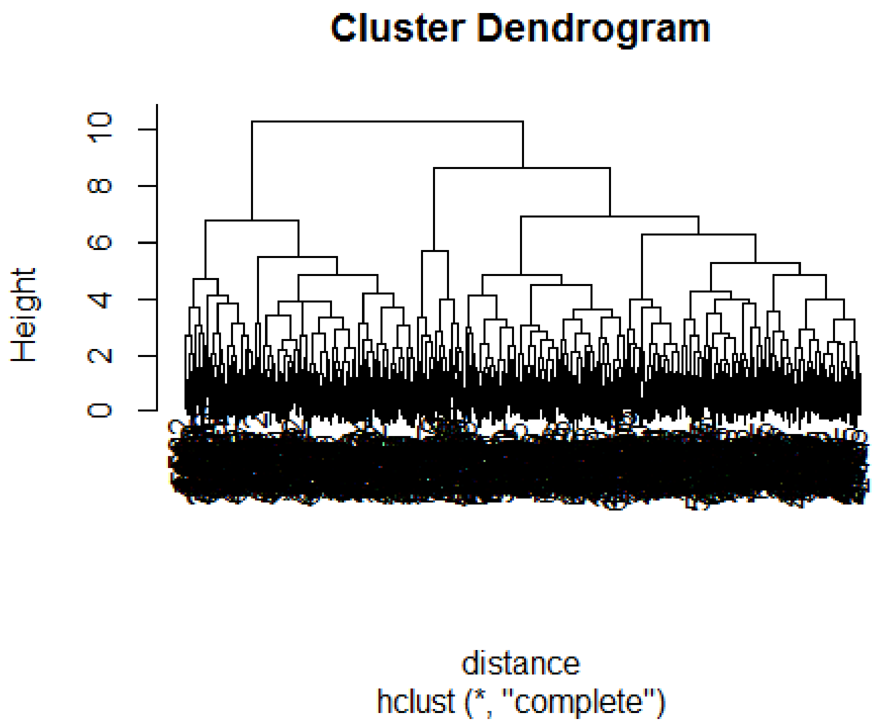


```
#Normalization will create a level plain field, the average for each variable
becomes 0 and std becomes 1
Z <-alc[, -c(1,1)]
#calculating mean for rows = 1 and columns= 2
M <-  apply(Z,2,mean)
SD <- apply(Z,2,sd)
Z <- scale(Z,M,SD)
```

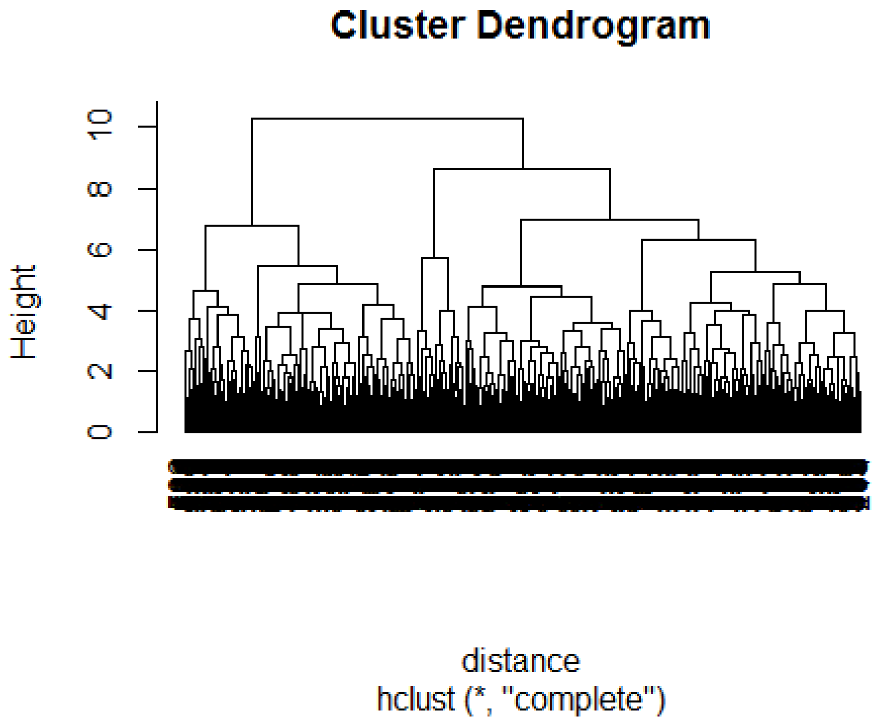
```
# Create new column filled with default colour
alc$Colour="black"
# Set new column values to appropriate colours
alc$Colour[alc$V2>=3]="red"
alc$Colour[alc$V8<=1]="blue"
# Plot all points at once, using newly generated colours
plot(alc$V2,alc$V8,  col=alc$Colour, ylim=c(0,5))
```



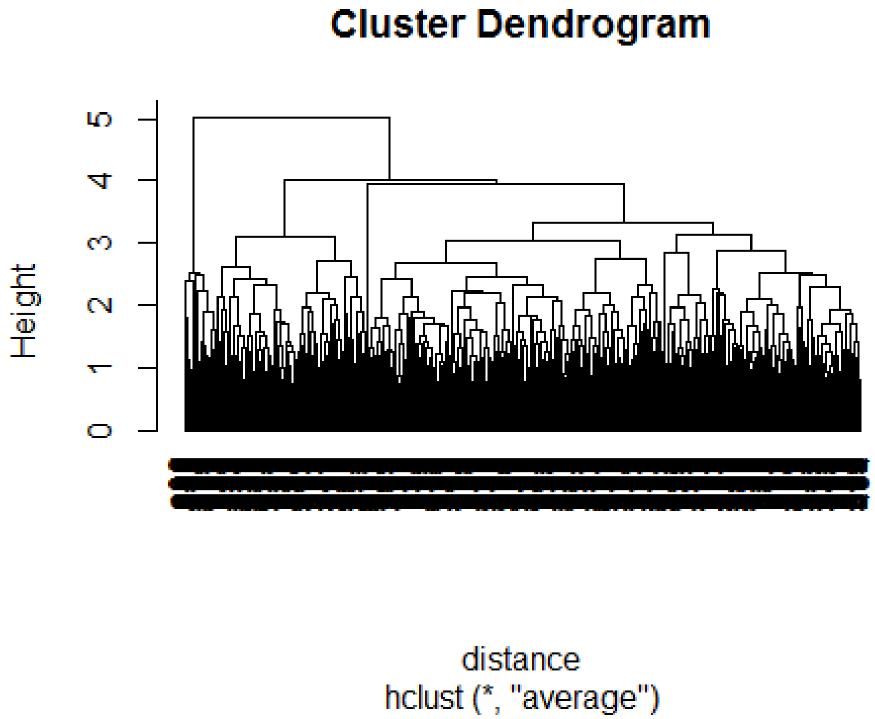
```
#cluster dendogram with complete linkage
hc.c <- hclust(distance)
plot(hc.c)
```



```
plot(hc.c, hang= -1)
```



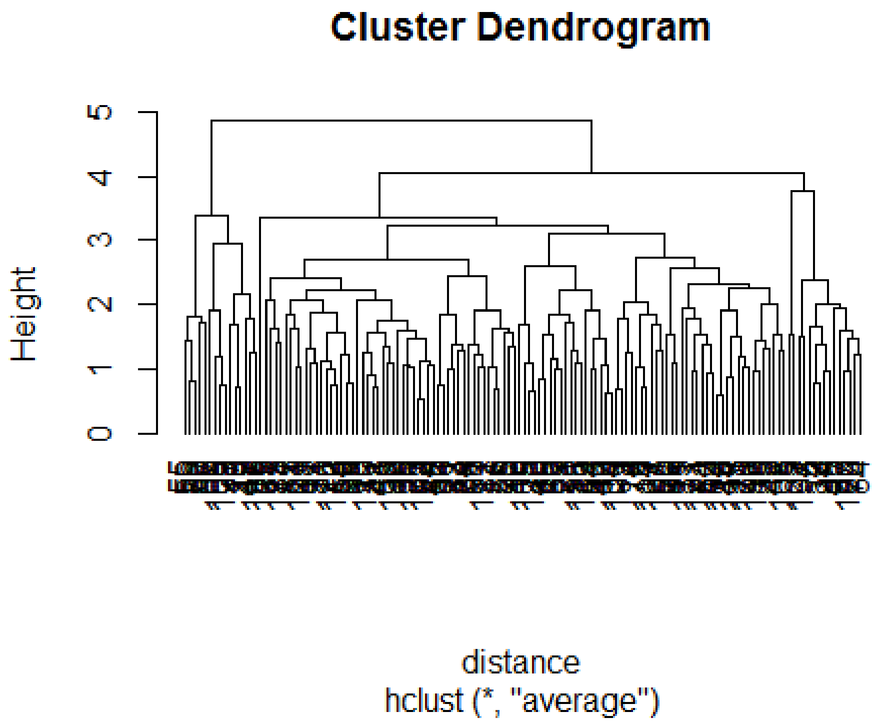
```
#Cluster dendogram average
hc.a <- hclust(distance,method="average")
plot(hc.a, hang=-1)
```



```r
#cluster member
lottery.c <- cutree(hc.c,3)
lottery.a <- cutree(hc.a,3)
```

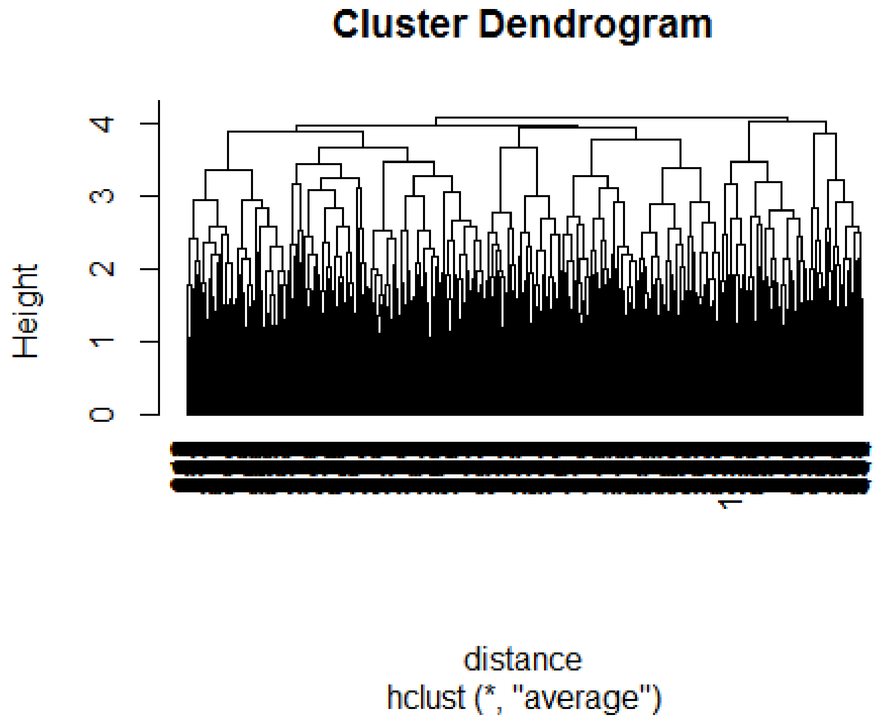
Does the cluster change over time with new data points being entered? If it does, are the changes due to things such as software updates, user behavior changes, etc.

We used a cluster dendrogram for the analysis. With the addition of new winning numbers lottery dataset from 2016 to 2017, added to the winning numbers dataset from 2009 to 2015, the data does change over time. Although the data continues to cluster and is not random, it actually begins to flatten out. The dendrogram below shows 2 separate and unequal groups when you cut it off at a height of 4.5, which is unchanged from the old data analysis. However, there is a change at a height of 3.7, where the new data separates into 3 groups, rather than 4 groups in the old dataset. But looking further down in height, the data appears to flatten out, and not cluster. We can assume that adding more datapoints will continue to flatten out the dendrogram, and hence become more random. The following shows the dendrogram for the new data points added (same R script above was used, and just adding the new data points):



How does a random number generator to generate a separate dataset compare?

We analyzed a random dataset, to see if the random number generator is indeed random, and to see if it clusters. With the clustering analysis, we found that the numbers are indeed random, as the dendrogram (below) has a uniform distribution at a height of 3.7 and below, and no clustering was found. The R script above was used, and utilized the random dataset generated by the script below. Please see the plot of the dendrogram showing this analysis, and the script used to generate the random dataset:



This script was used for randomly generating 1000 rows of 7-column numbers:

```
var generatedArr = [];
while (generatedArr.length < 10000) {
    var innerArr = [];
    while (innerArr.length < 7) {
        var generated = Math.floor(Math.random() * 49) + 1;
        if (innerArr.indexOf(generated) == -1) {
            innerArr.push(generated);
        }
    }
    generatedArr.push(innerArr)
}
for (i = 0; i < 1000; i++) {
    output += generatedArr[i].join(',') + "\r\n";
}
```

Summary

To conclude, the analysis of the winning numbers of a lottery draw initially appeared to cluster and not be randomly drawn. But with the addition of new datapoints, the initial clustering in the analysis became less pronounced, as the data tended to become more random. It is hypothesized that with the addition of more datapoints, it will ultimately make the analysis random, and should look like the analysis of a randomly-generated dataset as shown above.