Neandertal Cluster Analysis

James Bain

9/15/2017

Data

The data are comprised of 84 different activities compiled from a network of about 100 hospitals across the United States. This data are published in the form of Excel files and made available through the NEISS.

NEISS collects information about product and activity related injuries. Details about the anatomical region, type of injury and other demographic variables are collected. Data about injuries are then transformed to be activity specific and aggregated to match the anatomical categories defined in Berger & Trinkaus (1995). Once body parts injuries are tabulated, the injury patterns are normalized into proportions for each activity. Below is an example:

```
##                          head_neck shoulder_arm      hand      pelvis
## acrobatics/gymnastics 0.01941748     0.526699 0.2079288 0.001618123
##                               leg      foot       trunk
## acrobatics/gymnastics 0.05582524 0.1820388 0.006472492
```

Partitioning Around Medoids (PAM)

Clustering algorithms can be used to uncover patterns that are often too subtle for humans to find. They do so by calculating the distance between observation and "clustering" observations that are close to one another. Partitioning algorithms involve explicitly supplying the number of clusters k. Therefore, we have to find the optimal number of groups k to supply our clustering algorithm.

Partitioning around Medoids (PAM) is a partitioning algorithm much like *k-means* but instead of using a centroid as the center point of each cluster, a medoid is used instead. Medoids are attibutes of the data which represent the center most observation in a cluster of observations, whereas a centroid is an artifially created point used to represent the center of a cluster. Therefore, a centroid is more sensitive to outliers as opposed to medoids.

In order to cluster the points, PAM minimizes the following function

$$F(x) = minimize \sum_{i=1}^{n} \sum_{j=1}^{n} d(i,j) z_{ij}$$

$d(i,j)$ is the disimilarity between entities i and j and the term z_{ij} is a variable meant to ensure that only the dissimilarity of items with in the same cluster are computed toward the main function. Observations are then assigned to a cluster given that presence in that minimizes the main function $F(x)$.

Finding the Optimal Number of Clusters

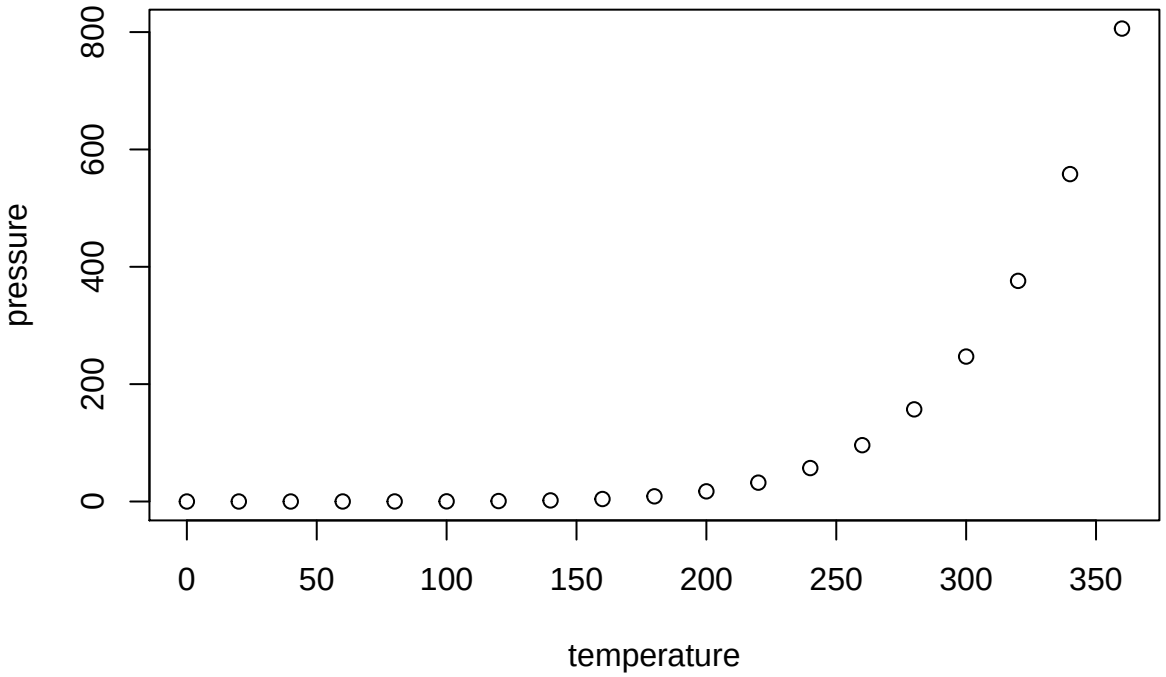
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Berger, Thomas D, and Erik Trinkaus. 1995. "Patterns of Trauma Among the Neandertals." *Journal of Archaeological Science* 22 (6). Elsevier: 841–52.