Problem set #2

Yaoxi Shi

17 October 2019

Computation

1

```
p=c(1,2)
q=c(3,4)

#Manhattan distance
dist_manhattan=sum(abs(p-q))
dist_manhattan
```

```
## [1] 4
```

```
#Canberra distance
dist_canberra=sum(abs(p-q)/(abs(p)+abs(q)))
dist_canberra
```

```
## [1] 0.8333333
```

```
#Euclidean distances
dist_euclidean=sqrt(sum((p-q)^2))
dist_euclidean
```

```
## [1] 2.828427
```

2

```
x=rbind(p,q)

#Manhattan distance
dist(x, method = "manhattan")
```

```
##   p
## q 4
```

```
#Canberra distance
dist(x, method = "canberra")
```

```
##           p
## q 0.8333333
```

```
#Euclidean distances
dist(x, method = "euclidean")
```

```
##          p
## q 2.828427
```

The results are correct.

3

Euclidean distance is the straight line distance between two points.

The Manhattan distance between two points is the sum of the differences of their corresponding components, it is the distance between two points measured along axes at right angles.

Canberra distance is a weighted version of Manhattan distance, it is easily biased for measures around the origin and very sensitive for values close to 0, where it is more sensitive to proportional than to absolute differences.

Using different distance measurements, the results of clustering of these points with other data points would be very different.

4 **4.1.Numeric Description of the Data:**

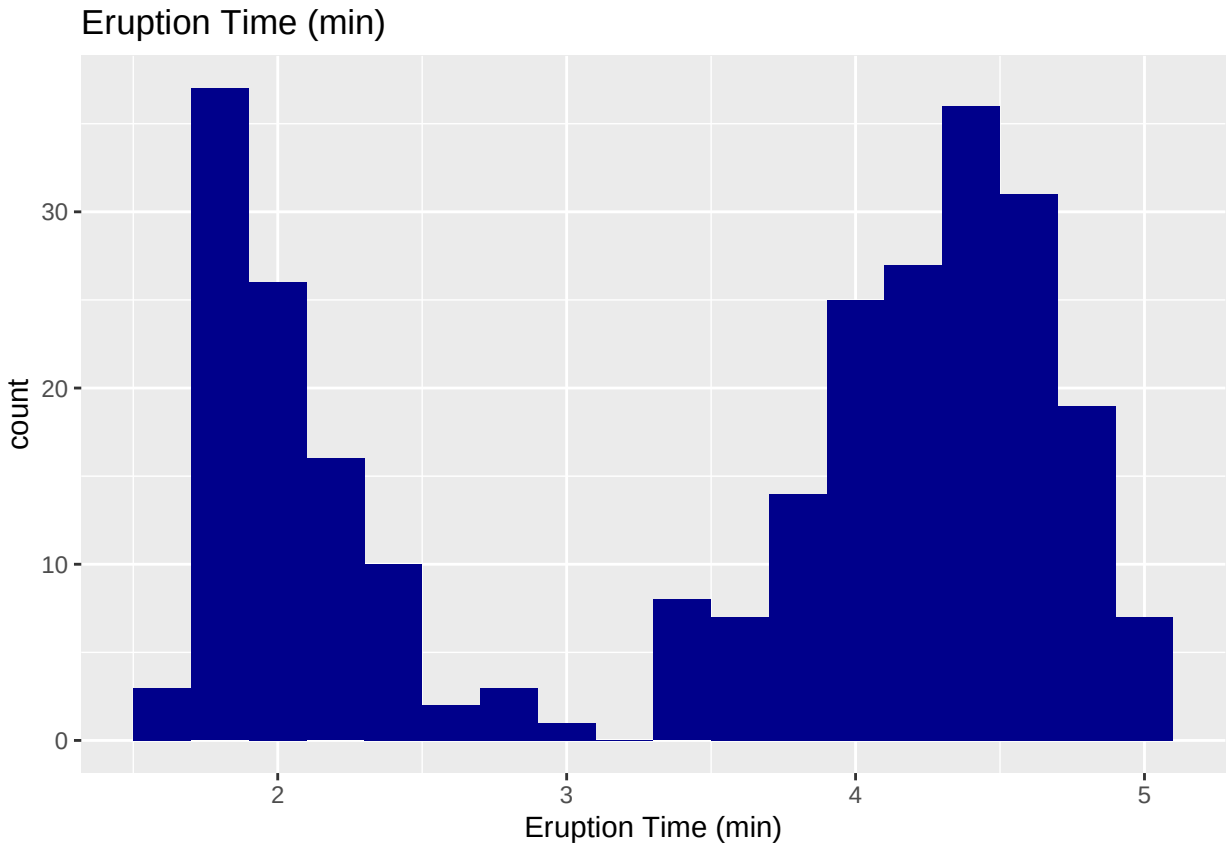
```
summary(faithful)
```

```
##    eruptions        waiting
##  Min.   :1.600   Min.   :43.0
##  1st Qu.:2.163   1st Qu.:58.0
##  Median :4.000   Median :76.0
##  Mean   :3.488   Mean   :70.9
##  3rd Qu.:4.454   3rd Qu.:82.0
##  Max.   :5.100   Max.   :96.0
```

There are two variables in this dataset, eruptions ranges from 1.6 mins to 5.1 mins with a mean equals to 3.49 mins and waiting time ranges from 43.0 mins to 96.0 mins with a mean equals to 70.9 mins.

4.2.Histogram of Eruption Time:

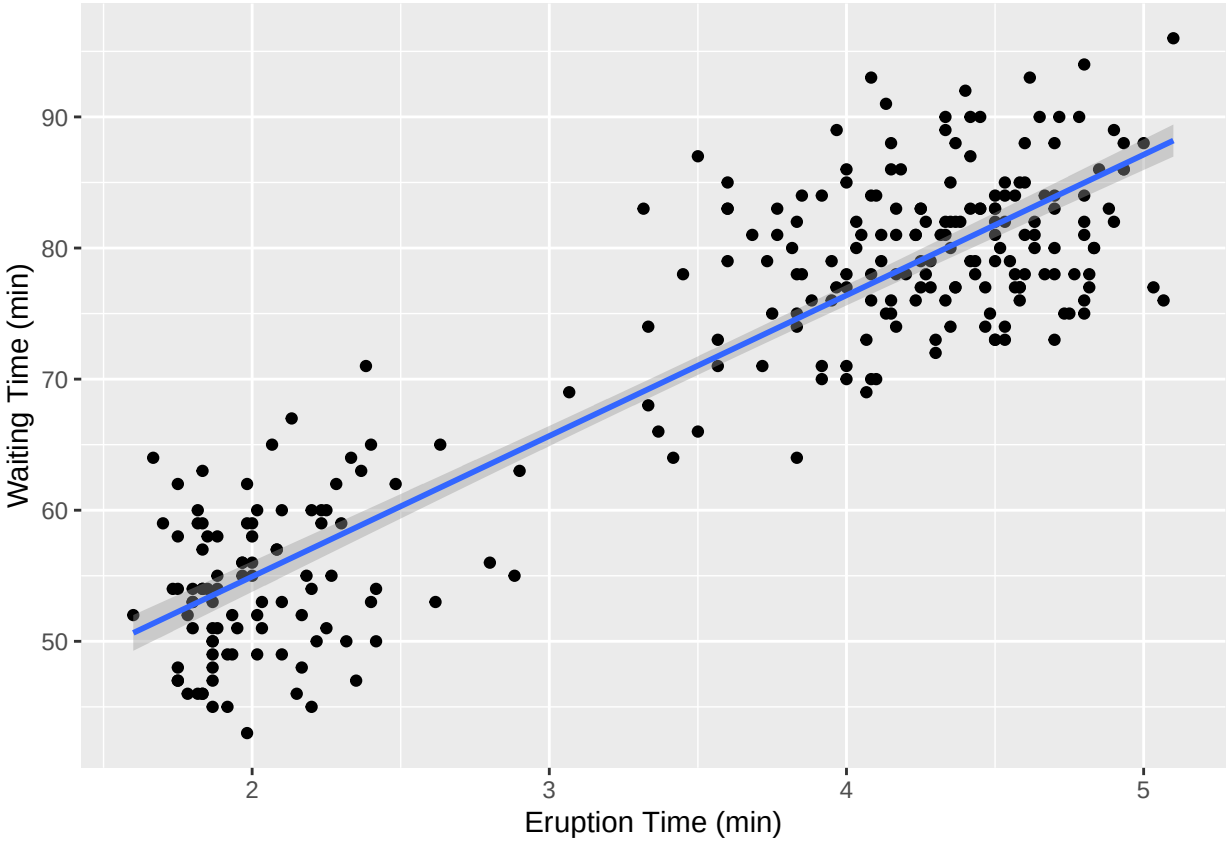
```
ggplot(faithful, aes(x = eruptions)) + geom_histogram(binwidth = 0.2, fill = 'darkblue') +
  ggtitle("Eruption Time (min)")+
  xlab("Eruption Time (min)")
```



The data diverges into two groups on eruption time, and the group with longer eruption time has more data points.

3.Scatter plot of the data, eruptions versus waiting time, with a regression line:

```
ggplot(faithful, aes(x=eruptions, y=waiting))+
  geom_point()+
  geom_smooth(method=lm)+
  xlab("Eruption Time (min)")+
  ylab("Waiting Time (min)")
```



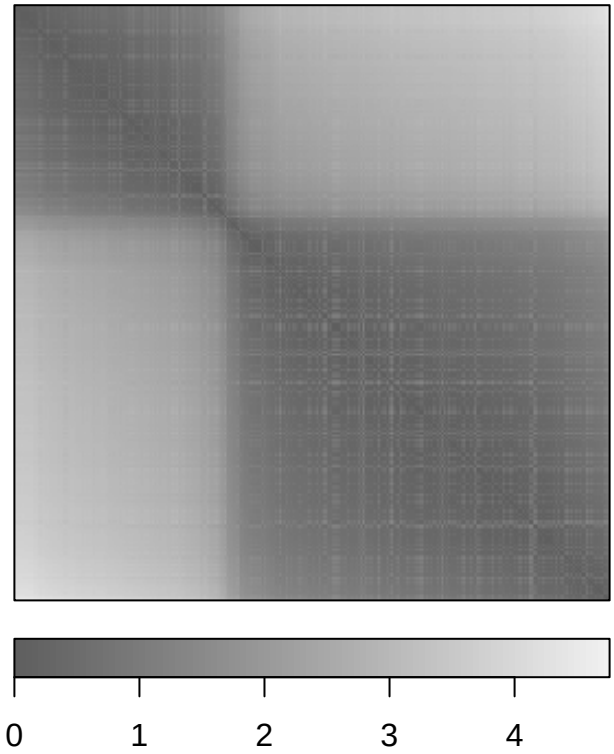
1) From the plot, we can easily see that the data points are naturally clustered into two groups. One group of data has relatively shorter eruption time which is centered around 2 mins, also has a shorter waiting time until next eruption centered around 55 mins. Another group of datapoints on the right up corner have longer eruption time around 4.5 mins and also have a longer waiting time around 80 mins.
2) We can also found a positive linear relationship between waiting time and eruption time. The eruption takes longer, the waiting time until next eruption is also likely to be longer.

5

```
faithful_scaled=scale(faithful)
faithful_dist=dist(faithful_scaled, method = "euclidean")
```

6

```
dissplot(faithful_dist)
```



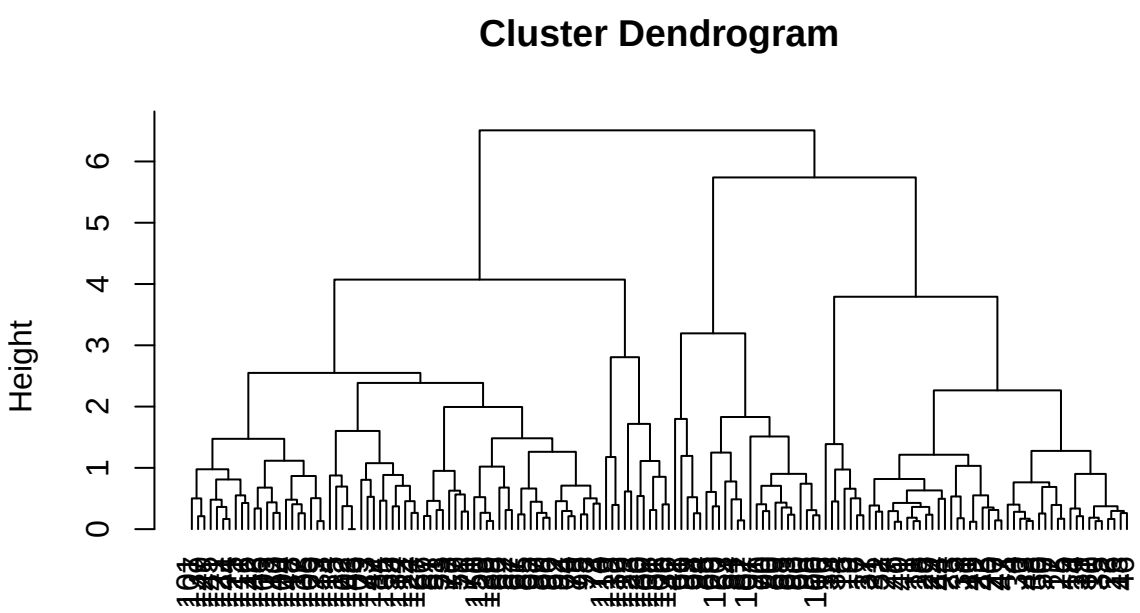
The ODI clearly shows two dark blocks, suggesting that the dataset could be cluseteres into two groups.

7

```
data(iris)
iris_sub_dist=iris[,-5] %>% scale() %>% dist()
```

8

```
iris_complete=hclust(iris_sub_dist, method = "complete")
plot(iris_complete, hang = -1, xlab ="Iris")
```



Iris
hclust (*, "complete")

All the data points are clusted bottom up, and the best choices for the number of clusters are 2 or 3.

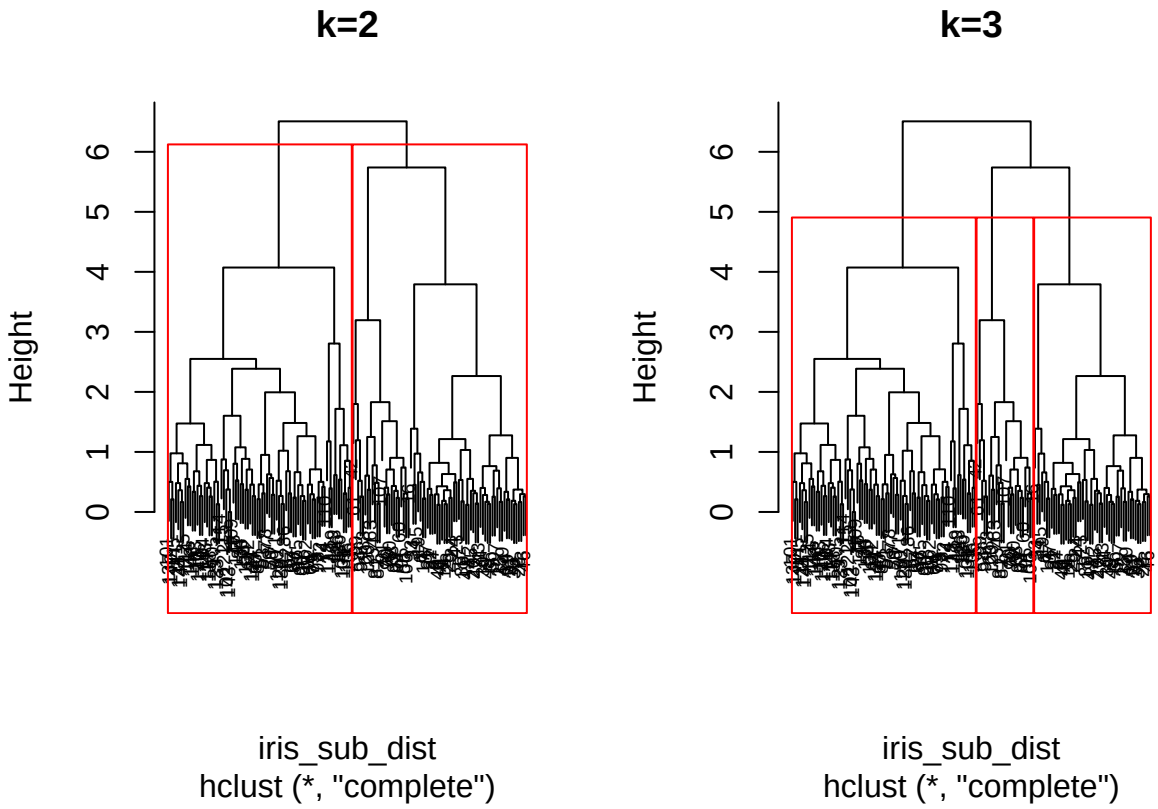
9

```
iris_cut=cutree(iris_complete, k=c(2,3))
table(iris_cut)
```

```
## iris_cut
##   1   2   3
## 122 101  77
```

```
par(mfrow=c(1,2))
{plot(iris_complete, cex=0.6, main="k=2")
rect.hclust(iris_complete, k=2)}

{plot(iris_complete, cex=0.6, main="k=3")
rect.hclust(iris_complete, k=3)}
```

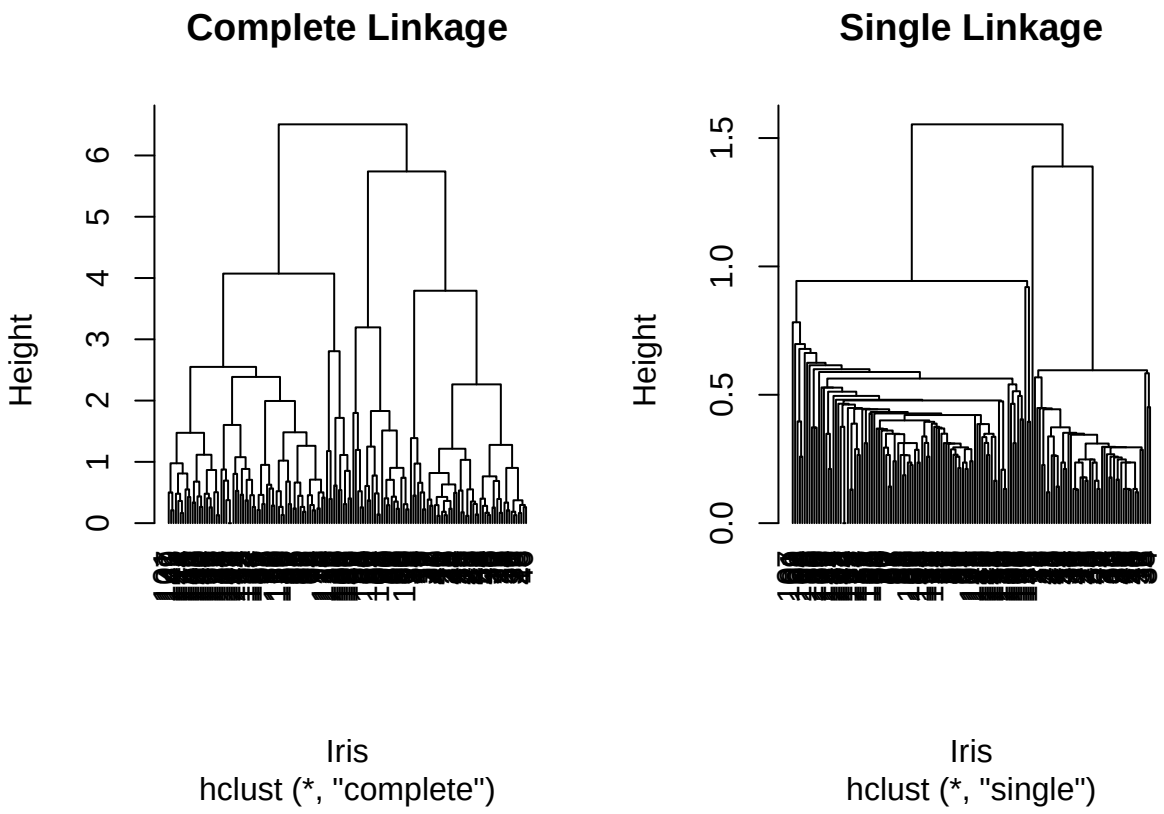


The right cluster dendrogram is cutting the right-side tree in the left cluster dendrogram into two clusters.

10

```
par(mfrow=c(1,2))
iris_complete=hclust(iris_sub_dist, method = "complete")
plot(iris_complete, hang = -1, xlab ="Iris", main = "Complete Linkage")

iris_single=hclust(iris_sub_dist, method = "single")
plot(iris_single, hang = -1, xlab ="Iris", main = "Single Linkage")
```



Different linkage methods applies different clustering principles, complete linkage uses the maximal inter-cluster dissimilarity, while Single linkage uses the minimal inter-cluster dissimilarity. The results of single linkage is more elongated, stringy-type clusters, while the levels in complete linkage methods looks much more clear. The results of the clustering using the two methods are also different.

Critical Thinking

1.a

Firstly, by informally plotting the distributions of the data, then use visulization tools to whether there are any clear grouping patterns, I could also further use mathmatical tools such as Hopkins statistic to test the randomness of the data using a sparse sampling test to determine the clusterability of the data.

1.b

1)Simple distribution plots.
Using scatter plots to visualize the relationsihp among features, to see whether there is any clear grouping patterns of the data points, if there is, than the data is clusterable.

2)VAT plots.
Caluculte the dissimilarity matrix and plot it, if there are darker blocks along the diagonal, it suggests that some groups of data has greater spatial similarity than others, so it's clusterable.

3)Hopkins statistic.
Calculate the pairwise dissimilarity across all observations in the actual data and compare to a set of simulated data drawn from some random distribution with the same standard deviation, if H >0.5, then the clustering would make sense to the data.

1.c

These techniques could provide information of the clusterability of the data from different perspectives. Plotting simple distribution is an informal and easiest way, should be used as a first step, we could get a general sence of the distribution of the variables and the relationships. VAT plots are more accurate, by seeing whether there are clear darker blocks along the diagonal, we can get a sense of the randomness of the data. The Hopkins statistic is the most accurate method to determine the randomness of the data, by calculating the H value and testing the hypothesis, we could determine the clusterability of the data mathematically. These three methods could be combined together to determine the clustering.

1.d

I would not proceed clustering if there is little to no support for clusterability, because if the data fails the clusterability diagnosis, it suggests that the data are randomly distributed, can not be effectively clustered and hard to interpret by clustering methods.

2.1
Paper: Beckstead, Jason W. "Using hierarchical cluster analysis in nursing research." Western Journal of Nursing Research 24, no. 3 (2002): 307-319.

Dateset: a set of 24 responses (11-point rating on impression) for each individual nurse.

The process: the author firstly scaled all the data, calculated proximities between all pairs of individuals, then selected Ward's sums of squares method and ran the algorithm for clustering, visualized the results, finally performed factorial ANOVA with several other dimensions of previous determined data to validate and help to interprete the clustering results.

2.2
The clustering analysis conducted in this paper went through all the steps we covered in the class except for assessing clusterability. In this project, I think the missing of the assessing step doesn't impact the findings. Because based on their results, there are clear five clusters emerged from the data, and each of them is well-intepretable, suggesting that the data is not randomly distributed. Though assessing clusterability before runnnig the algorithm is neccessary and important, but in this case, it doesn't affect the results.

2.3
This research found that cluster analysis are useful in discovering substantive differences in the way nurses formed impressions of impaired coworkers. Based on this, I think nurse researchers also could use hierarchical clustering methods to analyze the type of relationships between nurses and patient using data collected from their interactions, such as how many times a nurse interact with the patient, the time duration of the interactions, how many patients does a nurse serve and so on. This clustering analysis could help nurse researchers to identify different nursing styles.