Analysis on Dataset “citytemp”

by Yimi Zhao

1 Introduction of Dataset

The “citytemp” dataset contains 60 observations across 3 variables. It records the temperature of 60 cities in January and July respectively.

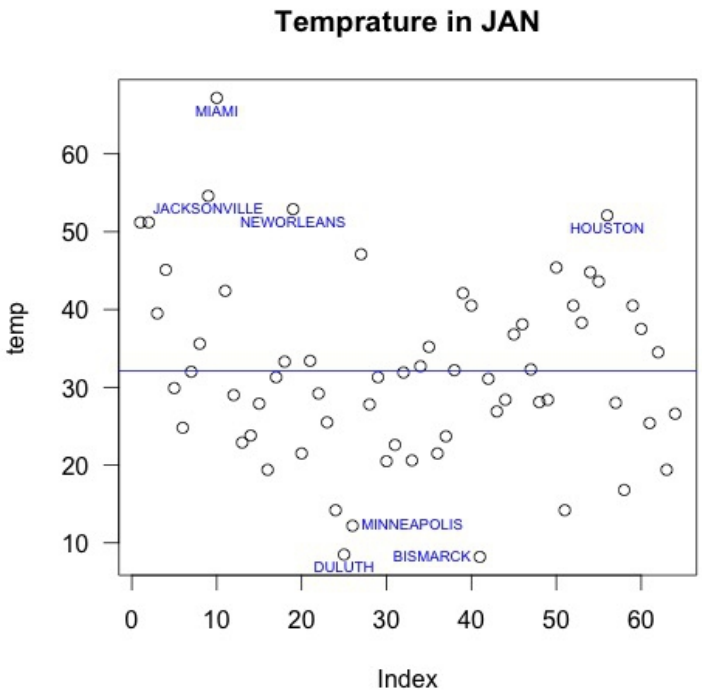
The data format is pretty simple: **citytemp** city, jan, july.

```
> head(citytemp)
        CITY      JAN      JULY
1     MOBILE      51.2     81.6
2    PHOENIX      51.2     91.2
3    LITTLEROCK   39.5     81.4
4    SACRAMENTO   45.1     75.2
5    DENVER       29.9     73.0
6    HARTFORD     24.8     72.7
```

2 Data Analysis

2.1 Data Exploration

First, we explore the dataset a little bit. It turns out that the mean temperature in January of these 60 cities is about 32 ºF and 75.6 ºF in July. And the standard variations indicate temperature in January(with sd 11.7) is more diverse than in July(with sd 5.1).



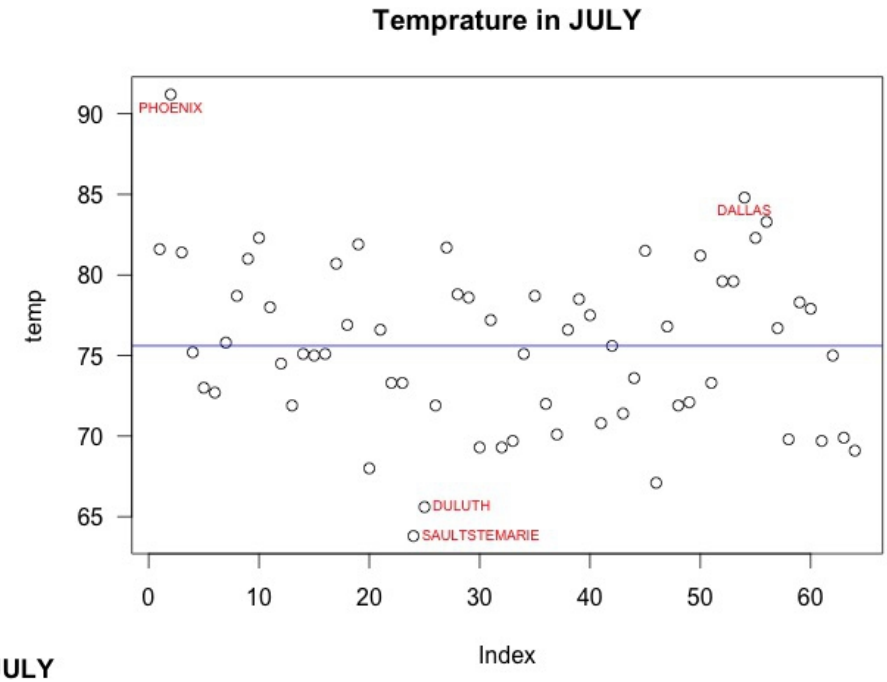

```
#temp mean and sd
> mean<-sapply(citytemp[,-1],mean)
> sd<-sapply(citytemp[,-1],sd)
> rbind(mean, sd)
          JAN       JULY
mean 32.09531  75.607812
sd   11.71243   5.127619
# plot temp in JAN
> plot(citytemp$JAN,main="Temprature in
JAN",ylab="temp",las=1)
> plot(citytemp$JAN,main="Temprature in
JAN",ylab="temp",las=1)
> abline(h=mean(citytemp$JAN),col="blue")
>identify(citytemp$JAN,
labels=citytemp$CITY,cex=0.6,col="blue")
[1]  9 10 19 25 26 41 56
```

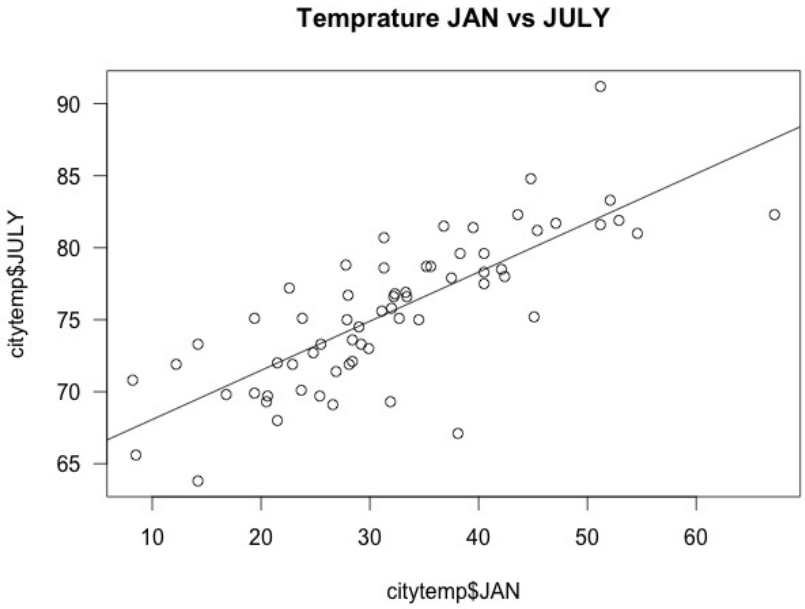
The plot shows MIAMI is the warmest city in January, with temperature 67.2 ºF, the other three cities less warmer are JACKSONVILLE, NEWORLEANS and HOUSTON(around 52ºF). On the contrary, the coldest city in January is BISMARCK, only has 8.2 ºF. With 8.5 ºF, DULUTH is almost as cold as BISMARCK, then comes MINNEAPOLIS, has 12.2 ºF as the third coldest city.

Similarly, we found out that PHOENIX is super hot in July(91.2 ºF), and DALLAS is the second hottest with 84.8 ºF in July. The coolest cities are DULUTH(65.6 ºF) and SAULTSTEMARIE(63.8 ºF)

```
#temp in JULY
> plot(citytemp$JULY, main="Temprature in JULY",ylab="temp",las=1)
> abline(h=mean(citytemp$JULY),col="blue")
> identify(citytemp$JULY, labels=citytemp$CITY, cex=0.6,col="red")
[1]  2 24 25 54
> citytemp[c(9,10,19,25,26, 41, 56,2,24),]
          CITY           JAN JULY
9   JACKSONVILLE  54.6 81.0
10         MIAMI          67.2 82.3
19    NEWORLEANS 52.9 81.9
25        DULUTH        8.5 65.6
26   MINNEAPOLIS 12.2 71.9
41      BISMARCK      8.2 70.8
56       HOUSTON     52.1 83.3
2        PHOENIX       51.2 91.2
24 SAULTSTEMARIE 14.2 63.8

# Jan vs July
> plot(citytemp$JAN, citytemp$JULY,
main="Temprature JAN vs JULY",las=1)
> abline(lm(JULY~JAN, citytemp))
> cor(citytemp$JAN, citytemp$JULY)
[1] 0.7797321
```





Moreover, it seems the temperatures in January and July have a significantly strong linear positive correlation(0.78). Therefore, we may be able to reach a conclusion like: the lower temperature a city has in Jan, the lower temperature it will has in July, and vice versa.

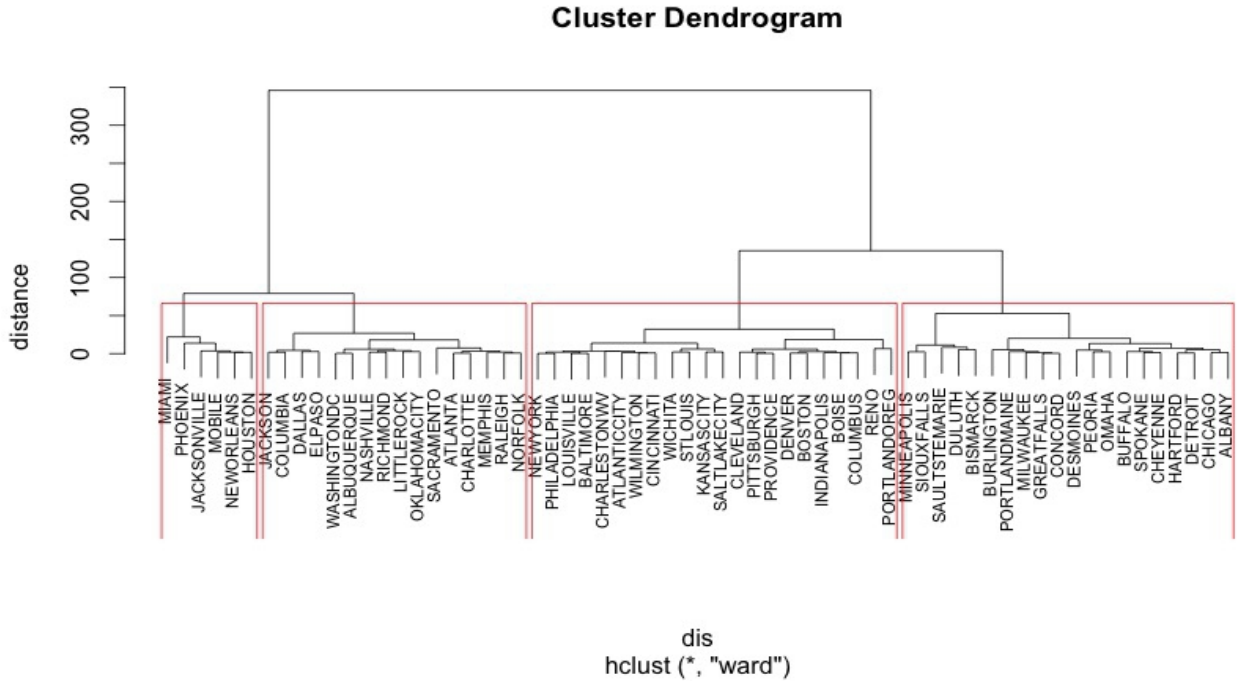
2.2 Data Analysis

From the previous data exploration, we found that some of the cities have quite similar temperatures. Thus, we want to check if these 60 cities can be divided into several groups according to their temperatures in January and July. If it does exist the groups, then we can get an idea which cities are similar form the temperature point of view.

2.2.1 Cluster Analysis

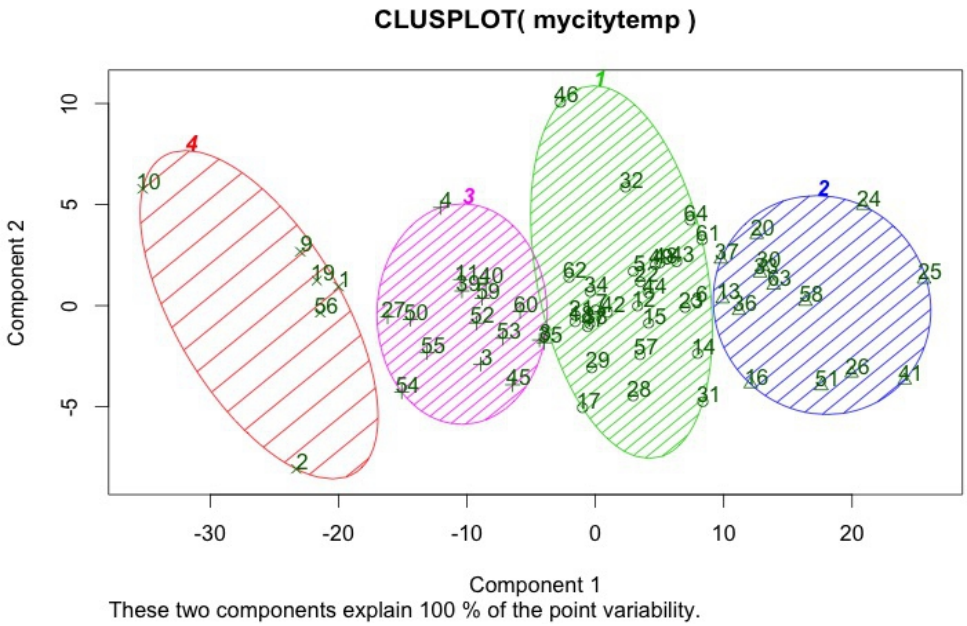
Next, we apply both hierarchical and k-means cluster analysis to explore the data. The dendrogram graph of "Ward" method hierarchical cluster analysis shows the cities are nicely divided into four clusters with size 6,16,22, and 20 .

```
# hierarchical cluster analysis
> mycitytemp<-citytemp[-1]
> dis<-dist(mycitytemp, method='euclidean')
> dis.matrix=as.matrix(dis)
> fit_ward<-hclust(dis, method="ward")
#plot the dendrogram cluster with city names at the bottom
> plot(fit_ward, labels=citytemp$CITY, ylab='distance',cex=0.7)
> groups<-cutree(fit_ward, k=4)
> rect.hclust(fit_ward, k=4, border="red")
> clusters<-cbind(citytemp,as.factor(groups))
> colnames(clusters)<-c("CITY","JAN","JULY","GROUP")
> table(clusters$GROUP)
     1  2  3  4
    6 16 22 20
> round(aggregate(citytemp[,2:3], by=list(groups),FUN=mean),3)
  Group.1        JAN     JULY
1       1       54.867  83.550
2       2       40.931  79.681
3       3       30.805  74.745
4       4      19.615   70.915
```



Similarly, we attain a pretty nice result from k-means cluster analysis.

```
#k-means cluster
> library(cluster)
> fit.k<-kmeans(mycitytemp, centers=4)
> fit.k$size
[1] 28 14 16  6
> round(aggregate(mycitytemp, by=list(fit.k$cluster),FUN=mean),3)
  Group.1       JAN      JULY
1       1       29.514   74.339
2       2       17.400   70.086
3       3       40.931   79.681
4       4       54.867   83.550
> clusplot(mycitytemp, fit.k$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```



2.2.1 Summary and Conclusion

Based on the previous cluster analysis, we get the summary of the groups:

Summary of City Clusters

Cluster	Cluster size	Mean temp(JAN)	Mean temp(JULY)
1	6	54.867	83.550
2	16	40.931	79.681
3	22	30.805	74.745
4	20	19.615	70.915

According to the summary of clusters displayed in the above table, we come up with the following conclusions:

- Sixty cities are divided into four categories with size 6,16,22,20;
- The six cities(MOBILE, PHOENIX, JACKONVILLE, MIAMI, NEWORLEANS, HOUSTON) in Cluster 1 have highest temperature in both January(54.87ºF) and July(83.55ºF). They have hot summer season and wild winter days.

```
> clusters[which(clusters$GROUP=="1")
           CITY  JAN JULY  GROUP
1        MOBILE 51.2 81.6     1
2       PHOENIX 51.2 91.2     1
9  JACKSONVILLE 54.6 81.0     1
10        MIAMI 67.2 82.3     1
19   NEWORLEANS 52.9 81.9     1
56      HOUSTON 52.1 83.3     1
```

- On the opposite of Cluster 1, the twenty cities in Custer 4 have lowest temperature in January (19.61ºF)and July(70.92ºF). Therefore, cities like CHICAGO, SAULTSTEMARIE, ALBANY would have really cold winters and cool summer days.

```
> head(clusters[which(clusters$GROUP=="4"),])
            CITY  JAN JULY GROUP
6       HARTFORD 24.8 72.7     4
13       CHICAGO 22.9 71.9     4
14        PEORIA 23.8 75.1     4
16     DESMOINES 19.4 75.1     4
20 PORTLANDMAINE 21.5 68.0     4
23       DETROIT 25.5 73.3     4
```

- Cities in Cluster 2 and Clusters have similar temperatures in July(79.68ºF vs 74.75ºF), but have different temperatures in January(40.93ºF vs 30.81ºF).

```
> head(clusters[which(clusters$GROUP=="2"),])
           CITY  JAN JULY GROUP
3    LITTLEROCK 39.5 81.4     2
4    SACRAMENTO 45.1 75.2     2
8  WASHINGTONDC 35.6 78.7     2
11      ATLANTA 42.4 78.0     2
27      JACKSON 47.1 81.7     2
35  ALBUQUERQUE 35.2 78.7     2
> head(clusters[which(clusters$GROUP=="3"),])
           CITY  JAN JULY GROUP
5        DENVER 29.9 73.0     3
7    WILMINGTON 32.0 75.8     3
12        BOISE 29.0 74.5     3
15 INDIANAPOLIS 27.9 75.0     3
17      WICHITA 31.3 80.7     3
18   LOUISVILLE 33.3 76.9     3
```