Name: Tim Lin

Title: Average Temperatures in US Cities

Data Summary:
The dataset is the average temperatures in January and July for selected US cities. The variables are CITY (name of the cities), JAN (average temperature of January), and JULY (average temperatures of July). There are 58 cities in the dataset. The summary of temperature for January and July are the following:

```
JAN              JULY
 Min.   : 8.20   Min.   :63.80
 1st Qu.:24.55   1st Qu.:71.90
 Median :31.30   Median :75.40
 Mean   :32.10   Mean   :75.61
 3rd Qu.:39.75   3rd Qu.:78.72
 Max.   :67.20   Max.   :91.20
```

Statement of Problem:

1.Examine the relationship between the average temperatures for cities.

2. Which cities have similar temperatures given average temperatures in January and July?

Analysis:

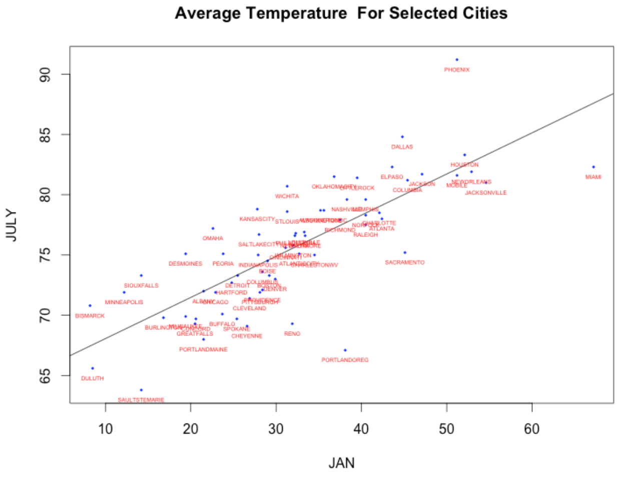
Linear Model: Use JAN as predictor and JULY as response.

lm(formula = JULY ~ JAN, data = citytemp)

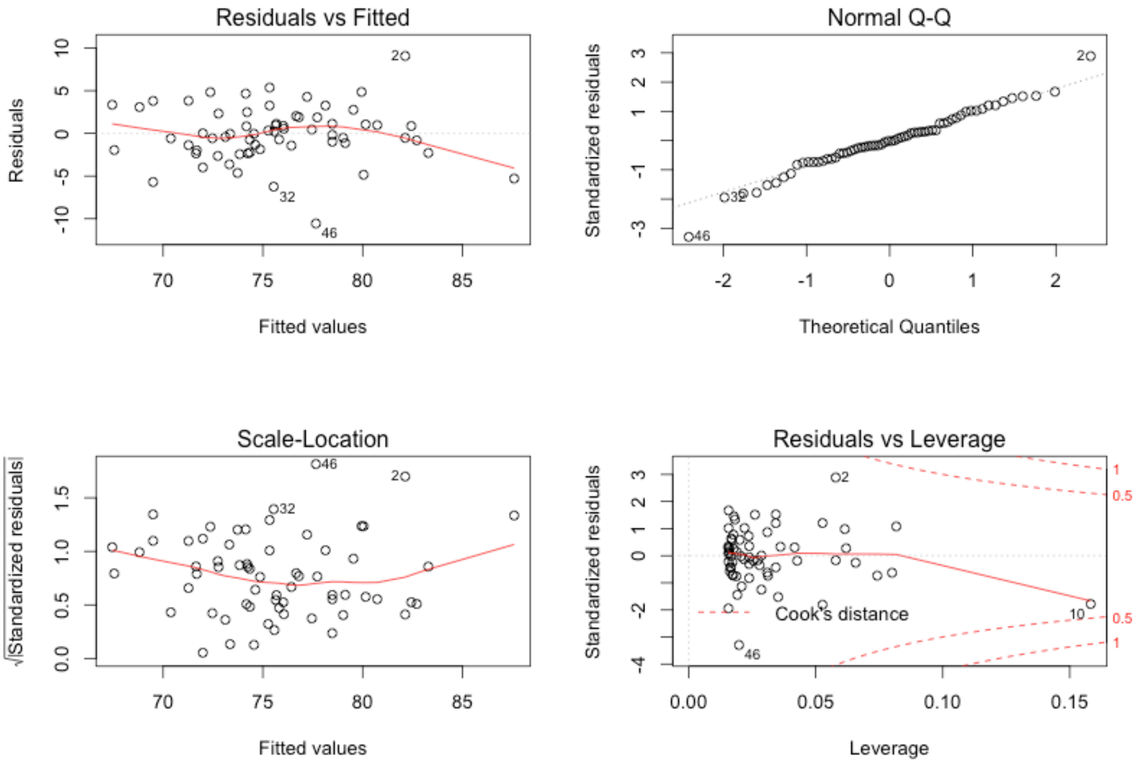
```
Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.65172    1.18827  54.408  < 2e-16 ***
JAN         0.34136    0.03481   9.806 3.16e-14 ***
```

Residual standard error: 3.236 on 62 degrees of freedom
Multiple R-squared: 0.608, Adjusted R-squared: 0.6017
F-statistic: 96.16 on 1 and 62 DF,
p-value: 3.164e-14

From the linear model, we can see that there is a relationship between the average temperature of January and July. This model does not fit well because the value of R^2 is low (.60). Then we exam if there any possible outliners cause this off fit.



Name: Tim Lin



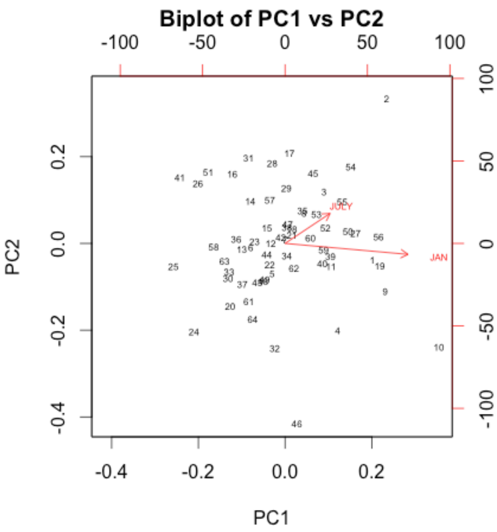
From the Residuals vs Fitted, we can see that the residual are normal and from the Normal Q-Q plot, we observed there might be two outliners (2, and 46). However, according to the leverage plot, these observations are not influential. Therefore, we want to exam the underlined relationship of the data using principal component.

```
> (pca = prcomp(~ JULY + JAN))
Standard deviations:
[1] 12.422182  3.027038

Rotation:
           PC1        PC2

JULY 0.3435323  0.9391409
JAN  0.9391409 -0.3435323
> summary(pca)
Importance of components:
                           PC1     PC2
Standard deviation     12.4222 3.02704
Proportion of Variance  0.9439 0.05605
Cumulative Proportion   0.9439 1.00000
```

The first principle component explained 94% of the variant of the data and the second principle component explained the rest.



Name: Tim Lin

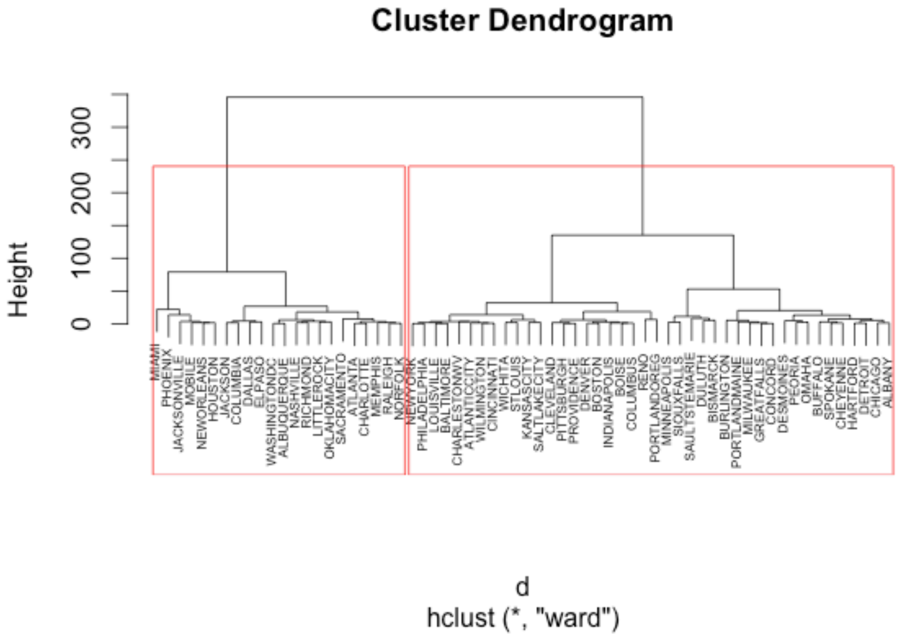
From the analysis we get:

PC1 = 0.326866 (JULY-75.92) + 0.945071 (JAN-32.55)
PC2 = 0.945071 (JULY-75.92) - 0.326866) (JAN-32.55)

The first principal component with both positive loadings suggests the measurements are from cities with warm climate since there is not much different in temperature between two months. The second principal component with a positive and negative loading suggest there is a different in two months which the measurement are taken from city with obvious temperature differences in winter and summer.

Then, we want to find out cities that suggested by the component analysis. Therefore, we cluster cities into two groups using Ward Hierarchical Clustering.

```
> d = dist(citytemp[2:3], method = "euclidean")
> fit = hclust(d, method="ward")
> plot(fit, labels = CITY, cex = .5)
> groups = cutree(fit,k=2)
> rect.hclust(fit, k=2, border="red")
```



Name: Tim Lin

```
> GROUP = as.factor(groups)
> clus = cbind(citytemp, GROUP)
> subset(clus, GROUP == 1)
           CITY  JAN JULY GROUP
1        MOBILE 51.2 81.6     1
2       PHOENIX 51.2 91.2     1
3    LITTLEROCK 39.5 81.4     1
4    SACRAMENTO 45.1 75.2     1
8  WASHINGTONDC 35.6 78.7     1
9  JACKSONVILLE 54.6 81.0     1
10        MIAMI 67.2 82.3     1
11      ATLANTA 42.4 78.0     1
19   NEWORLEANS 52.9 81.9     1
27      JACKSON 47.1 81.7     1
35  ALBUQUERQUE 35.2 78.7     1
39    CHARLOTTE 42.1 78.5     1
40      RALEIGH 40.5 77.5     1
45 OKLAHOMACITY 36.8 81.5     1
50     COLUMBIA 45.4 81.2     1
52      MEMPHIS 40.5 79.6     1
53    NASHVILLE 38.3 79.6     1
54       DALLAS 44.8 84.8     1
55       ELPASO 43.6 82.3     1
56      HOUSTON 52.1 83.3     1
59      NORFOLK 40.5 78.3     1
60     RICHMOND 37.5 77.9     1
```

```
> subset(clus, GROUP == 2)
            CITY  JAN JULY GROUP
5         DENVER 29.9 73.0     2
6       HARTFORD 24.8 72.7     2
7     WILMINGTON 32.0 75.8     2
12         BOISE 29.0 74.5     2
13       CHICAGO 22.9 71.9     2
14        PEORIA 23.8 75.1     2
15  INDIANAPOLIS 27.9 75.0     2
16     DESMOINES 19.4 75.1     2
17       WICHITA 31.3 80.7     2
18    LOUISVILLE 33.3 76.9     2
20 PORTLANDMAINE 21.5 68.0     2
21     BALTIMORE 33.4 76.6     2
22        BOSTON 29.2 73.3     2
23       DETROIT 25.5 73.3     2
24 SAULTSTEMARIE 14.2 63.8     2
25        DULUTH  8.5 65.6     2
26   MINNEAPOLIS 12.2 71.9     2
28    KANSASCITY 27.8 78.8     2
29       STLOUIS 31.3 78.6     2
30    GREATFALLS 20.5 69.3     2
31         OMAHA 22.6 77.2     2
32          RENO 31.9 69.3     2
33       CONCORD 20.6 69.7     2
34  ATLANTICCITY 32.7 75.1     2
36        ALBANY 21.5 72.0     2
37       BUFFALO 23.7 70.1     2
38       NEWYORK 32.2 76.6     2
41      BISMARCK  8.2 70.8     2
42    CINCINNATI 31.1 75.6     2
43     CLEVELAND 26.9 71.4     2
44      COLUMBUS 28.4 73.6     2
46  PORTLANDOREG 38.1 67.1     2
47  PHILADELPHIA 32.3 76.8     2
48    PITTSBURGH 28.1 71.9     2
49    PROVIDENCE 28.4 72.1     2
51    SIOUXFALLS 14.2 73.3     2
57  SALTLAKECITY 28.0 76.7     2
58    BURLINGTON 16.8 69.8     2
61       SPOKANE 25.4 69.7     2
62  CHARLESTONWV 34.5 75.0     2
63     MILWAUKEE 19.4 69.9     2
64      CHEYENNE 26.6 69.1     2
```

Conclusion:

The first cluster is cities with warm temperatures with range from 35.2 to 91.2 degree, and the second clusters are with cities with great difference in temperature range from 8.2 to 80.7.