Cause of Death in Worcester Report

Yangsu

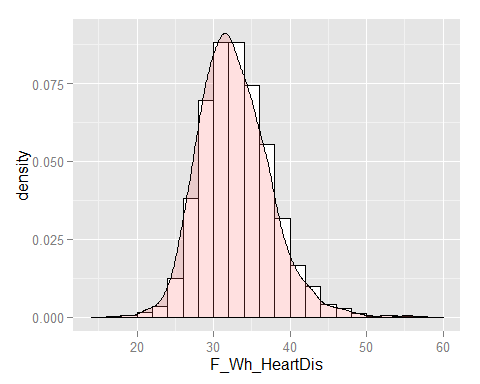
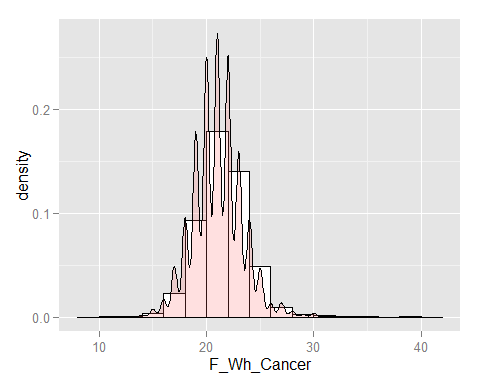
Sun Oct 18 10:35:24 2015

I choose Leading Cause of Death to be my dataset, from which I am trying to find patterns between different causes of deaths and regions.  
Since lots of variables inside are of great amount of missing values (equal -1111 or -2222), I picked 65+ years old White with cancer(F\_Wh\_Cancer), 65+ years old White with Heart Disease(F\_Wh\_HearDis),45-64 years old white with Heart Disease(E\_Wh\_HearDis), 45-64 years old white with Cancer(E\_Wh\_Cancer).

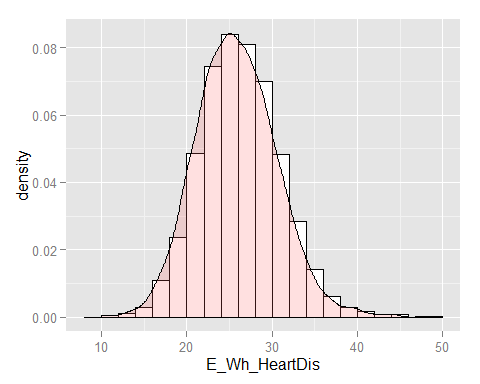
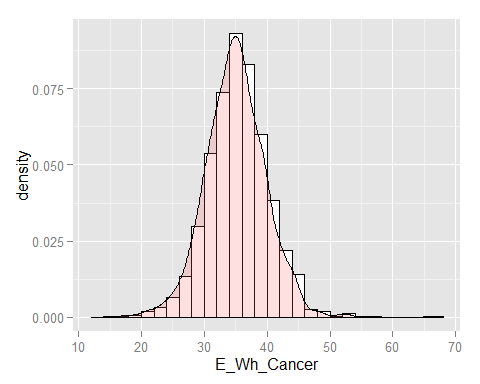
After read in data, I find the dimension to be 3141 in rows and 235 in columns.  
So firstly, I clean up the data set and make it useful.

In order to accomplish it, I build up a new data frame which includes variables I need to use:   
data<-**data.frame**(State\_FIPS\_Code,County\_FIPS\_Code,CHSI\_County\_Name,CHSI\_State\_Name,F\_Wh\_Cancer,F\_Wh\_HeartDis,E\_Wh\_HeartDis,E\_Wh\_Cancer). After that, dimensions are reduced to 3141 in rows and 8 in columns. Also, I find out that there are missing values for each variables, 18 missing values for F\_Wh\_Cancer, 18 missing values for F\_Wh\_HeartDis, 72 missing values for E\_Wh\_HeartDis, 71 missing values for E\_Wh\_Cancer. All variables above have quite low percentage of missing values so I assume they should work fine. By deleting all those missing values, dimension are reduced to 3068 in rows and 8 in columns.

so far we have no more extreme and missing values, so we can start do the analysis.  
Summary up the data, we find out that mean value for F\_Wh\_Cancer is 21.04, mean value for F\_Wh\_HeartDis is 32.7, mean value for E\_Wh\_HeartDis is 25.83, mean value for E\_Wh\_Cancer is 35.02.



Plot for F\_Wh\_Cancer Plot for F\_Wh\_HeartDis

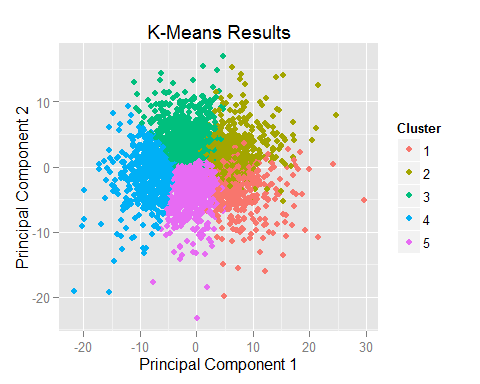


Plot for E\_Wh\_Cancer Plot for E\_Wh\_HeartDis

Afterwards, I compare Worcester data with other counties in US.

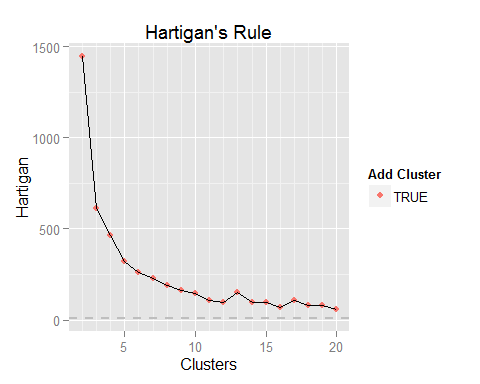
In Worcester, we have 22 records of F\_Wh\_Cancer, 29 records of F\_Wh\_HeartDis, 21 records of E\_Wh\_HeartDis, 36 records of E\_Wh\_Cancer.  
From above, we can see that for Worcester, two age groups whites are having higher death rate of cancer than mean of the whole country, but for heart disease, it shows lower rate. In order to prove that, we use k-means method.  
By setting cluster number to be 5, I got following results:

## K-means clustering with 5 clusters of sizes 375, 519, 791, 591, 792  
## Cluster means:  
## F\_Wh\_Cancer F\_Wh\_HeartDis E\_Wh\_HeartDis E\_Wh\_Cancer  
## 20.98632 29.10940 22.52308 31.71282  
## 19.67757 39.74065 28.40654 35.74766  
## 20.34706 34.86471 31.57647 29.56863  
## 21.42698 32.51951 26.52843 36.24638  
## 22.06424 29.50347 21.07986 40.67882



According to the clusters, Worcester belongs to the first cluster.  
It shows that Worcester do have higher rate in death of Cancer for middle age whites. But lower rate in death of Heart disease.

In order to find the best fit cluster numbers, I use FitKMeans method, and got the result as followed.



I think clusters of 10 works the best because after 10, Hartigan do not drop a lot.

As for conclusion, I would strongly suggest leader of Worcester in Massachusetts put more effort on caring about cancer risk for middle age whites and try to find out the reasons for causing the cancer. By doing so, I believe Worcester would be a better place to live in for middle age white people.

Complete version of R output is attached.