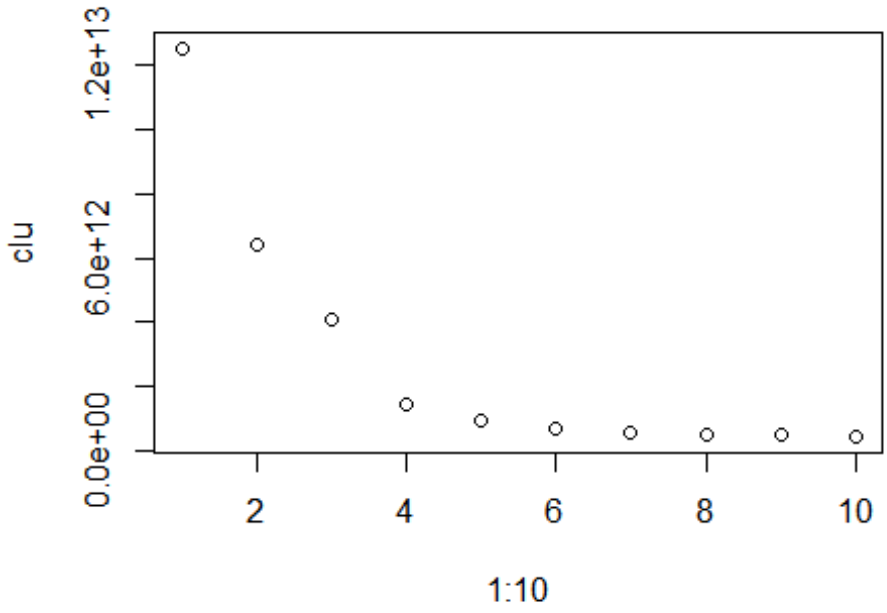
riskfactor.R

Lina Zhou

```
riskraw<-read.csv("RISKFACTORSANDACCESSTOCARE.csv",header = T)
#find the number of missing data

#data cleaning
riskraw[riskraw==-1111.1]<-0
riskraw[riskraw==-2222.2]<-NA
riskraw[riskraw==-2222]<-NA
riskraw$Dentist_Rate[250]<-NA

riskfactor<-data.frame(riskraw)
# remove NA
riskfactor<-subset(riskfactor,is.na(riskfactor$Elderly_Medicare)==F)
riskfactor<-subset(riskfactor,is.na(riskfactor$Disabled_Medicare)==F)
riskfactor<-subset(riskfactor,is.na(riskfactor$Uninsured)==F)

#select useful to run cluster
riskcluster<-data.frame(countycode=riskfactor$County_FIPS_Code,countyna
me=riskfactor$CHSI_County_Name,state=riskfactor$CHSI_State_Abbr,strate=
riskfactor$Strata_ID_Number,NO_Exercise=riskfactor$No_Exercise,
                        Few_F_V=riskfactor$Few_Fruit_Veg,obesity=riskfa
ctor$Obesity,HBP=riskfactor$High_Blood_Pres,smoker=riskfactor$Smoker,di
abetes=riskfactor$Diabetes,uninsured=riskfactor$Uninsured,
                        elderly_M=riskfactor$Elderly_Medicare,disabled_
M=riskfactor$Disabled_Medicare,primary_D=riskfactor$Prim_Care_Phys_Rate,
dentist=riskfactor$Dentist_Rate,CHC=riskfactor$Community_Health_Center_
Ind,HPSA=riskfactor$HPSA_Ind)
dim(riskcluster)

## [1] 3108   17

#cluster
##determine the center
clust<-riskcluster[,5:17]


clu<-NULL
for(i in 1:10)
{
  clu[i]<-sum(kmeans(clust,centers = i)$withinss)
}
##sum of the sum of square in each cluster
plot(1:10,clu)
```




```
clu[4]/clu[1]
```

```
## [1] 0.113057
```

```
clu[5]/clu[1]
```

```
## [1] 0.0750736
```

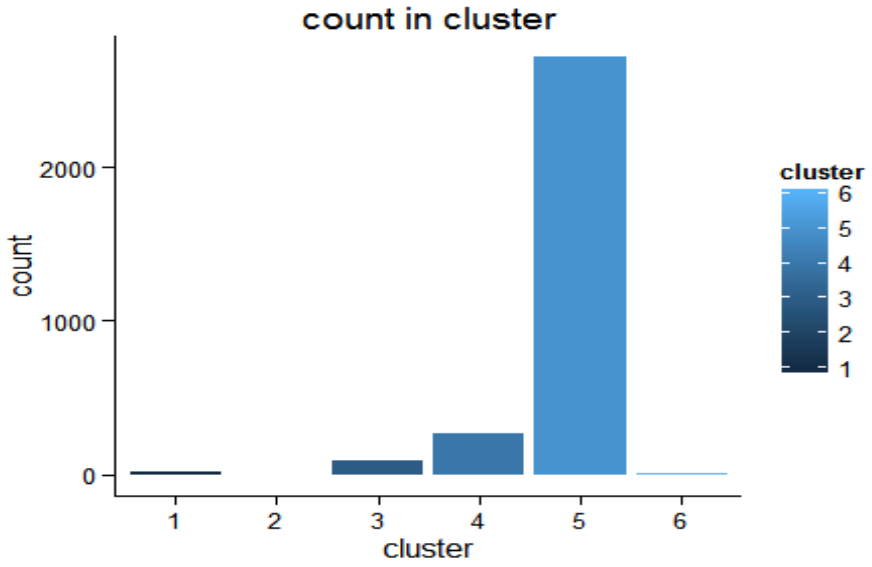
```
clu[6]/clu[1]
```

```
## [1] 0.05518323
```

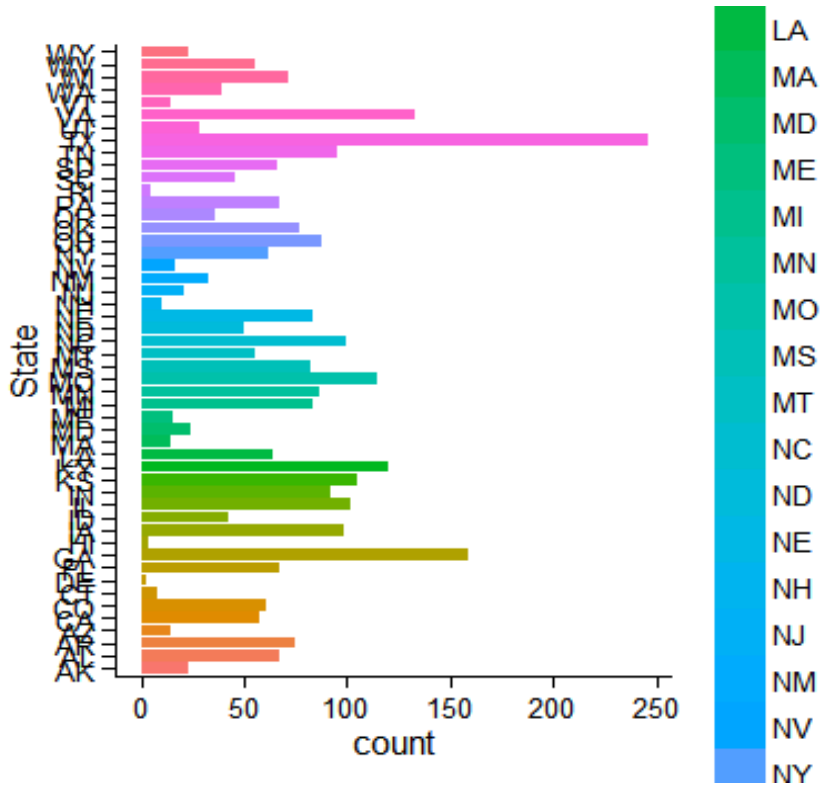
```
##confirm center=6
# make 6 cluster
clu6<-kmeans(clust,centers = 6)

#find Worcester
riskfactor$cluster<-clu6$cluster
clustworcester<-subset(riskfactor,riskfactor$CHSI_County_Name=="Worcest
er"&riskfactor$CHSI_State_Abbr=="MA")
cat("Worcester is in cluster :", clustworcester$cluster, "\n")
```

```
## Worcester is in cluster : 3
```

```
ggplot(riskfactor)+geom_histogram(aes(x=factor(cluster),fill=cluster))+
xlab("cluster")+theme_classic()+labs(title="count in cluster")+scale_co
lour_brewer()
```

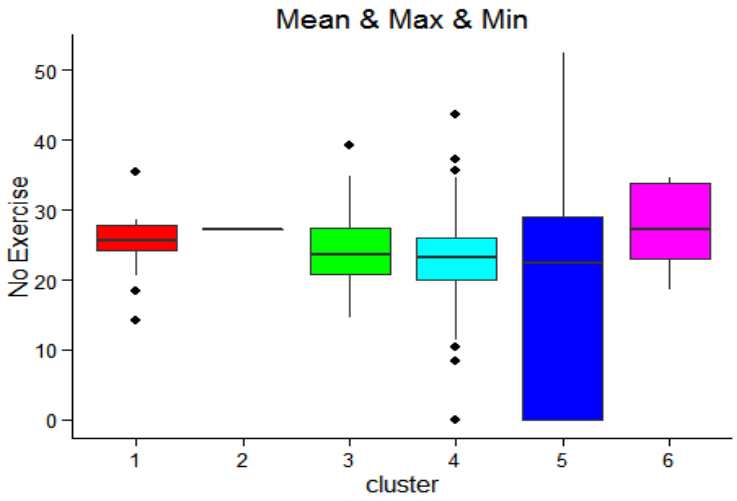



```
ggplot(riskfactor)+geom_histogram(aes(x=factor(CHSI_State_Abbr),fill=CH
SI_State_Abbr))+xlab("State")+theme_classic()+coord_flip()
```

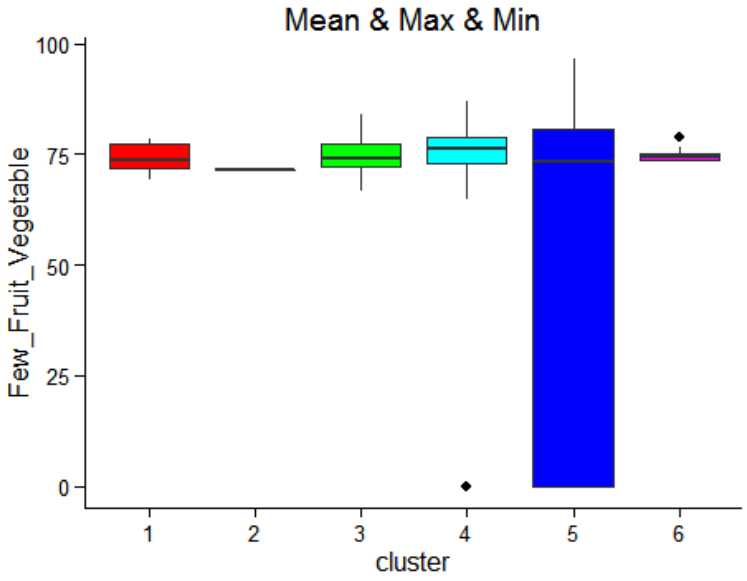



```
#compare Worcester nationwide
#compare for risk factor
# no exercise
ggplot(data=riskfactor)+geom_boxplot(aes(x=factor(riskfactor$cluster),y
```

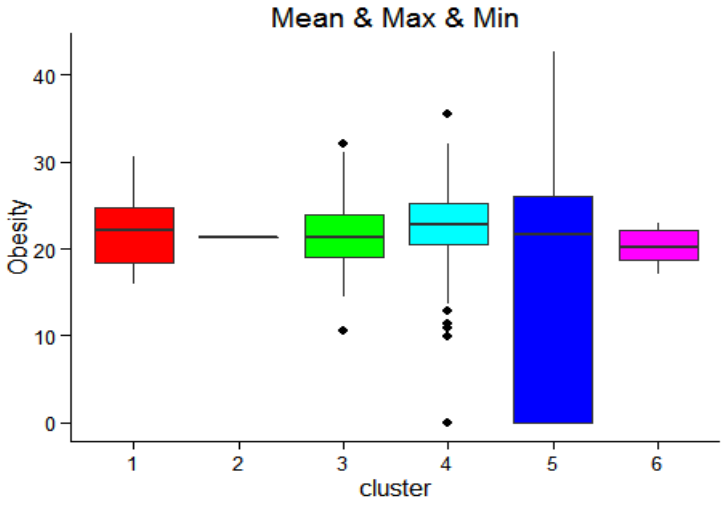
```
=riskfactor$No_Exercise),fill=rainbow(6))+xlab("cluster")+labs(y="No Ex
ercise",title="Mean & Max & Min")+theme_classic()
```



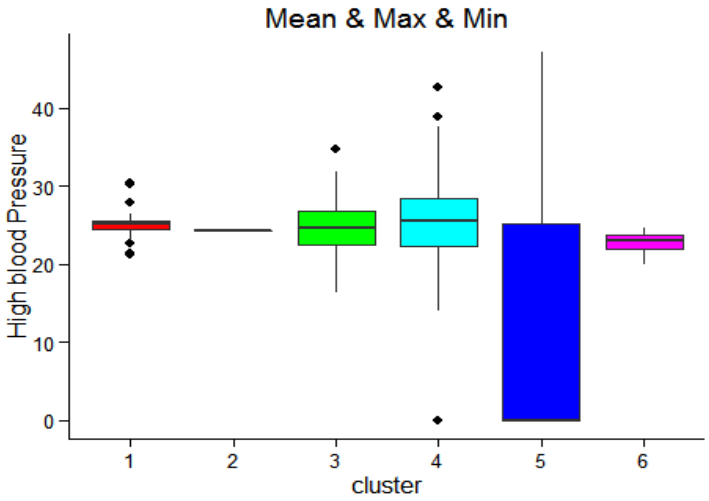

```
#few vegetables&fruits
ggplot(riskfactor)+geom_boxplot(aes(x=factor(riskfactor$cluster),y=risk
factor$Few_Fruit_Veg),fill=rainbow(6))+xlab("cluster")+theme_classic()+
labs(y="Few_Fruit_Vegetable",title="Mean & Max & Min")
```



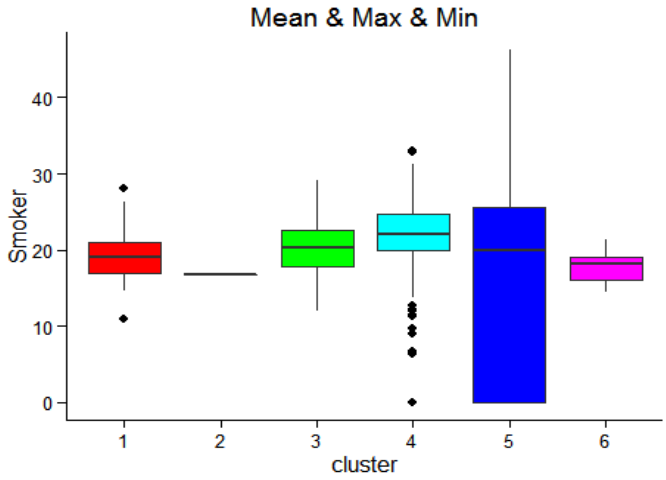

```
#obesity
ggplot(riskfactor)+geom_boxplot(aes(x=factor(riskfactor$cluster),y=risk
factor$Obesity),fill=rainbow(6))+xlab("cluster")+theme_classic()+labs(y
="Obesity",title="Mean & Max & Min")
```



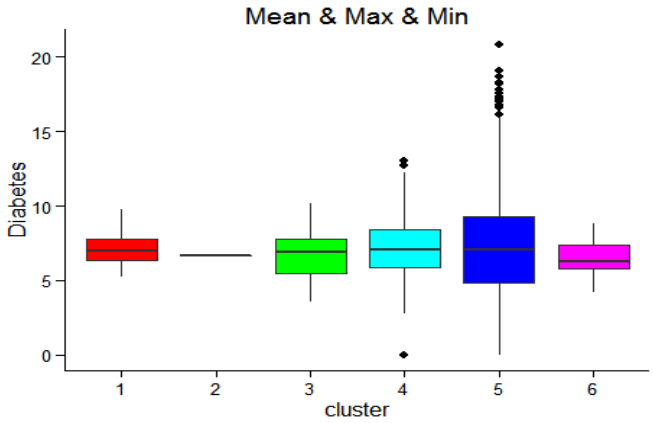

```
#high blood pressure
ggplot(riskfactor)+geom_boxplot(aes(x=factor(riskfactor$cluster),y=risk
factor$High_Blood_Pres),fill=rainbow(6))+xlab("cluster")+theme_classic()
+labs(y="High blood Pressure",title="Mean & Max & Min")
```



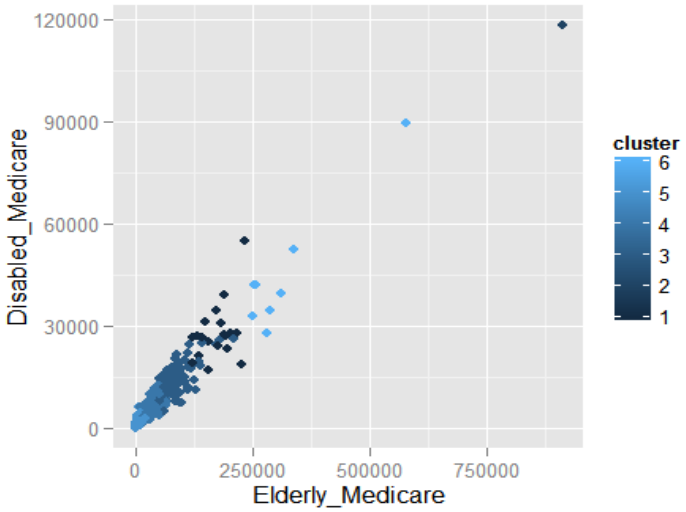

```
#smoker
ggplot(riskfactor)+geom_boxplot(aes(x=factor(riskfactor$cluster),y=risk
factor$Smoker),fill=rainbow(6))+xlab("cluster")+theme_classic()+labs(y=
"Smoker",title="Mean & Max & Min")
```



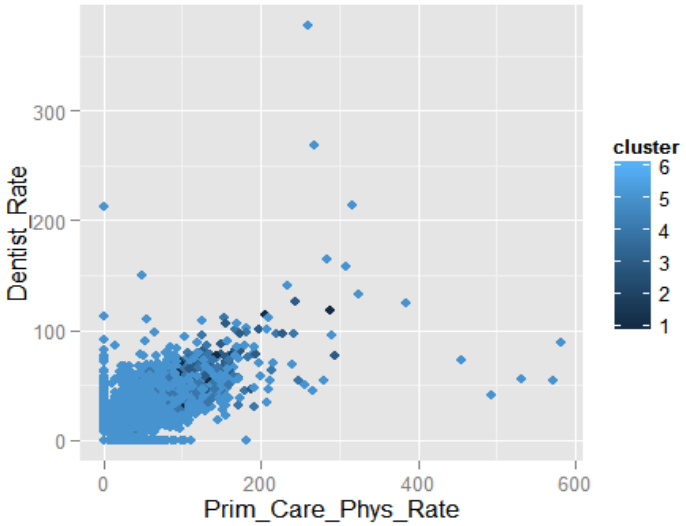

```
#diabetes
ggplot(riskfactor)+geom_boxplot(aes(x=factor(riskfactor$cluster),y=risk
factor$Diabetes),fill=rainbow(6))+xlab("cluster")+theme_classic()+labs
(y="Diabetes",title="Mean & Max & Min")
```



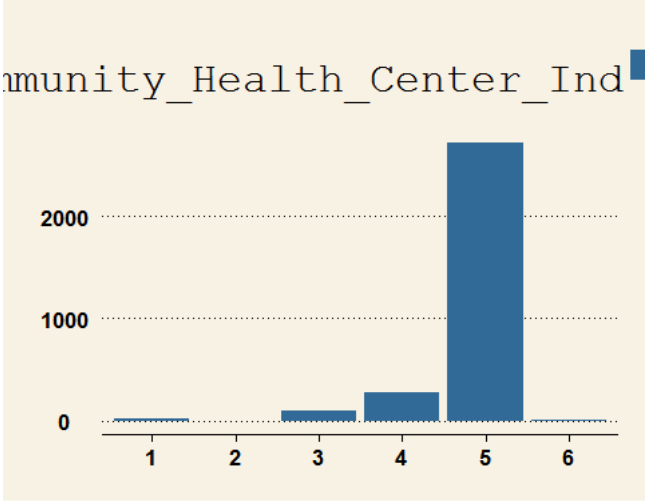

```
#access to health care
#elderly_medicare
#disable medicare
ggplot(data=riskfactor, mapping=aes(x=Elderly_Medicare, y=Disabled_Medi
care,  colour=cluster))+geom_point()
```



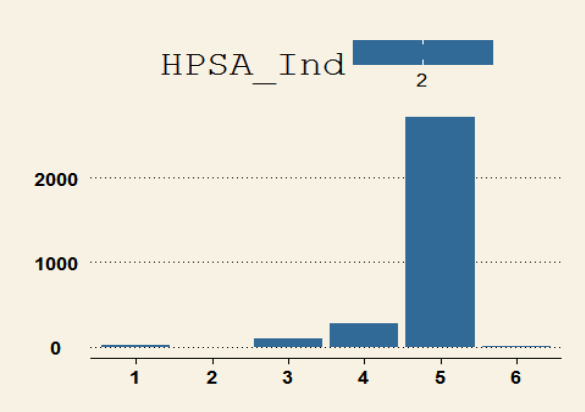

```
#primary care physician rate
#dentist rate
ggplot(data=riskfactor, mapping=aes(x=Prim_Care_Phys_Rate, y=Dentist_Ra
te,  colour=cluster))+geom_point()
```



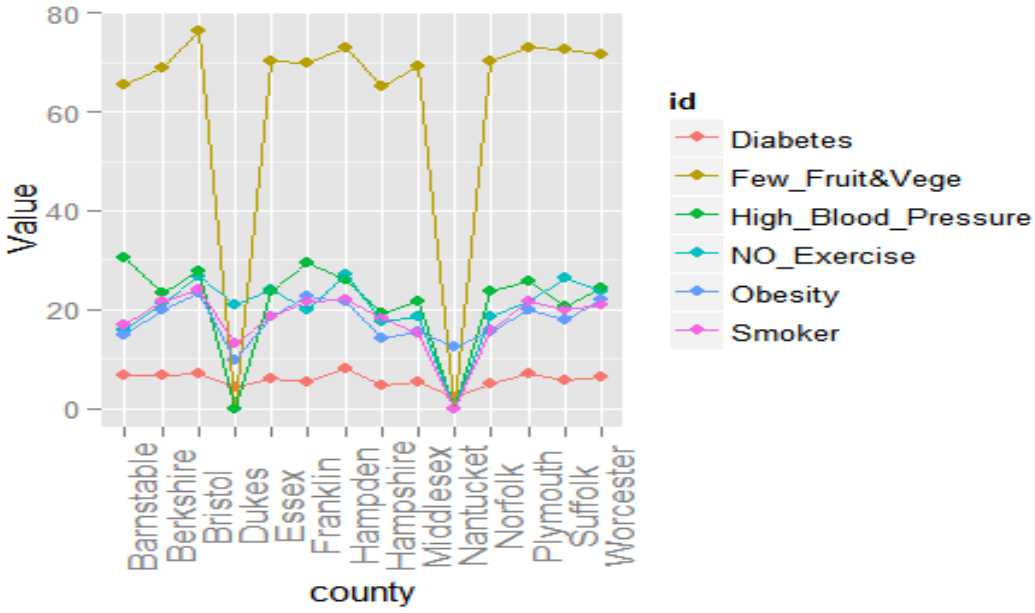

```
#community health center
#health professional shortage area
ggplot(riskfactor)+geom_histogram(aes(x=factor(cluster), fill=Community
_Health_Center_Ind))+theme_wsj()
```



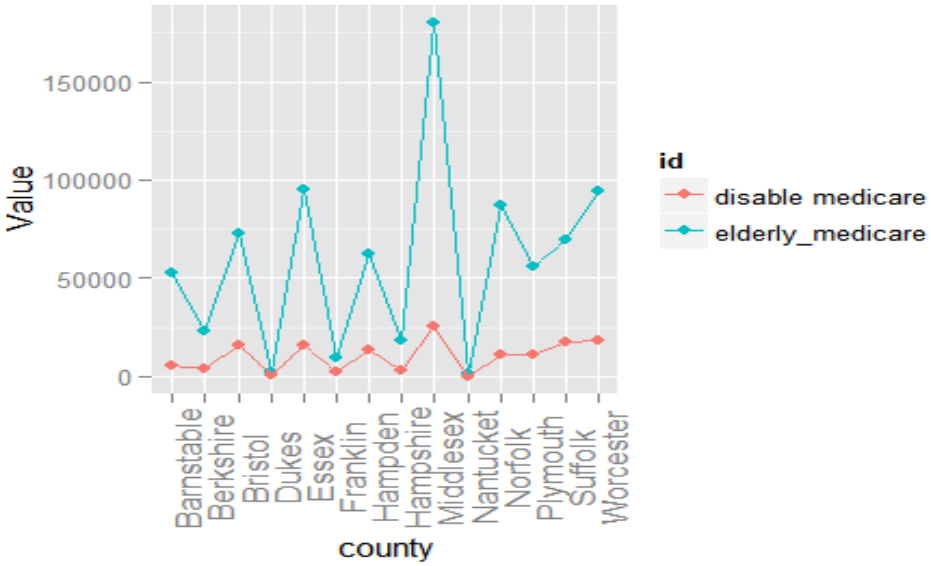

```
ggplot(riskfactor)+geom_histogram(aes(x=factor(cluster), fill=HPSA_Ind))
+theme_wsj()
```




```
#compare Worcester in the MA
MA<-subset(riskfactor,riskfactor$CHSI_State_Abbr=="MA")
id1<-rep("NO_Exercise",14)
id2<-rep("Few_Fruit&Vege",14)
id3<-rep("Obesity",14)
id4<-rep("High_Blood_Pressure",14)
id5<-rep("Smoker",14)
id6<-rep("Diabetes",14)
y<-MA$CHSI_County_Name
daf<-data.frame(county=rep(y,6),value=c(MA$No_Exercise,MA$Few_Fruit_Veg,
MA$Obesity,MA$High_Blood_Pres,MA$Smoker,MA$Diabetes),id=c(id1,id2,id3,i
d4,id5,id6))
#compare value of riskfactor in each county
qplot(county, value, data = daf,geom =c("point","line"),group=id, id =
id, stat = "identity", colour = id, ylab = "Value")+theme(axis.text.x=e
lement_text(angle=90,size=12))
```




```
id7<-rep("elderly_medicare",14)
id8<-rep("disable medicare",14)
id9<-rep("primary care physician rate",14)
id10<-rep("dentist rate",14)
df<-data.frame(county=rep(y,2),value=c(MA$Elderly_Medicare,MA$Disabled_
Medicare),id=c(id7,id8))
#compare number of medicare for elderly $ disable in each county
qplot(county, value, data = df,geom =c("point","line"),group=id, id = i
d, stat = "identity", colour = id, ylab = "Value")+theme(axis.text.x=el
ement_text(angle=90,size=12))
```



```
df2<-data.frame(county=rep(y,2),value=c(MA$Prim_Care_Phys_Rate,MA$Denti
st_Rate),id=c(id9,id10))
#compare ratio of primary care physician &dentist in each county
qplot(county, value, data = df2,geom =c("point","line"),group=id, id =
id, stat = "identity", colour = id, ylab = "Value")+theme(axis.text.x=e
lement_text(angle=90,size=12))
```