workshop08

February 3, 2019

```
In [1]: library(tidyverse)

        # Plot size deppening on your screen resolution to 5 x 3
        options(repr.plot.width=4, repr.plot.height=4)
```

```
Attaching packages  tidyverse 1.2.1
ggplot2 2.2.1       purrr   0.2.5
tibble  1.4.2       dplyr   0.7.5
tidyr   0.8.1       stringr 1.3.1
readr   1.1.1       forcats 0.3.0
Conflicts  tidyverse_conflicts()
dplyr::filter() masks stats::filter()
dplyr::lag()    masks stats::lag()
```


1 Welcome to Workshop 8

1.0.1 Exercise 1: Replicate the slide example

Select the numeric columns and run kmeans: `km <- kmeans(ktcNumeric.df, centers=3, nstart=25)`

```
In [7]: ktc.df<-read.table("KTC.csv",header=TRUE,sep=",")
        head(ktc.df)

        ktcNumeric.df <- ktc.df %>%
            select(ID,Age,Income,Children)%>%
            column_to_rownames(var='ID')%>%
            scale
        head(ktcNumeric.df)

        km <- kmeans(ktcNumeric.df, centers=3, nstart=25)
        km
```

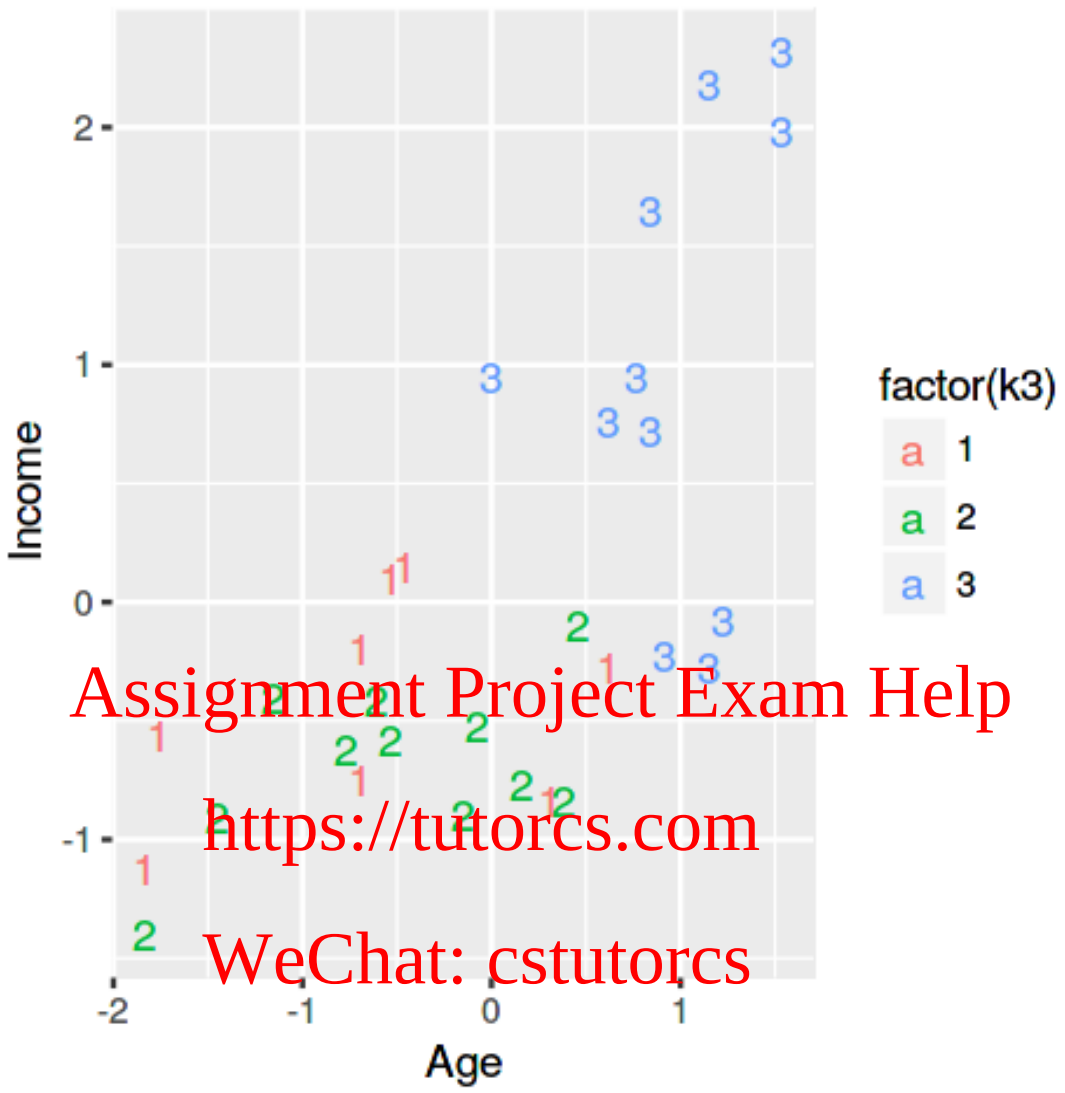
ID	Age	Female	Income	Married	Children	CarLoan	Mortgage
1	48	1	17546	0	1	0	0
2	40	0	30085	1	3	1	1
3	51	1	16575	1	0	1	0
4	23	1	20375	1	3	0	0
5	57	1	50576	1	0	0	0
6	57	1	37870	1	2	0	0

	Age	Income	Children
1	0.1559026	-0.7637480	0.06169096
2	-0.4574848	0.1512843	1.91241969
3	0.3859229	-0.8346067	-0.86367341
4	-1.7609332	-0.5573020	1.91241969
5	0.8459635	1.6466130	-0.86367341
6	0.8459635	0.7193939	0.98705533

```
K-means clustering with 3 clusters of sizes 11, 11, 8

Cluster means:
         Age     Income   Children
1 -0.5062770 -0.6762445 -0.6113013
2  0.9644588  0.9939719 -0.3589292
3 -0.6300000 -0.4368752  1.3340670

Clustering vector:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
 1  3  1  3  2  2  1  2  3  3  2  1  1  2  1  1  3  2  2  1  2  3  2  1  3  2
27 28 29 30
 1  1  3  2

Within cluster sum of squares by cluster:
[1]  8.683382 16.616678  8.282644
 (between_SS / total_SS =  61.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```



```
In [5]: ktcNumeric.df %>%
            as_tibble()%>%
            mutate(k3=km$cluster)%>%
            ggplot(data=.,
                   aes(x = Age,
                       y = Income,
                       colour = factor(k3),
                       label = k3))+
        geom_text()
```

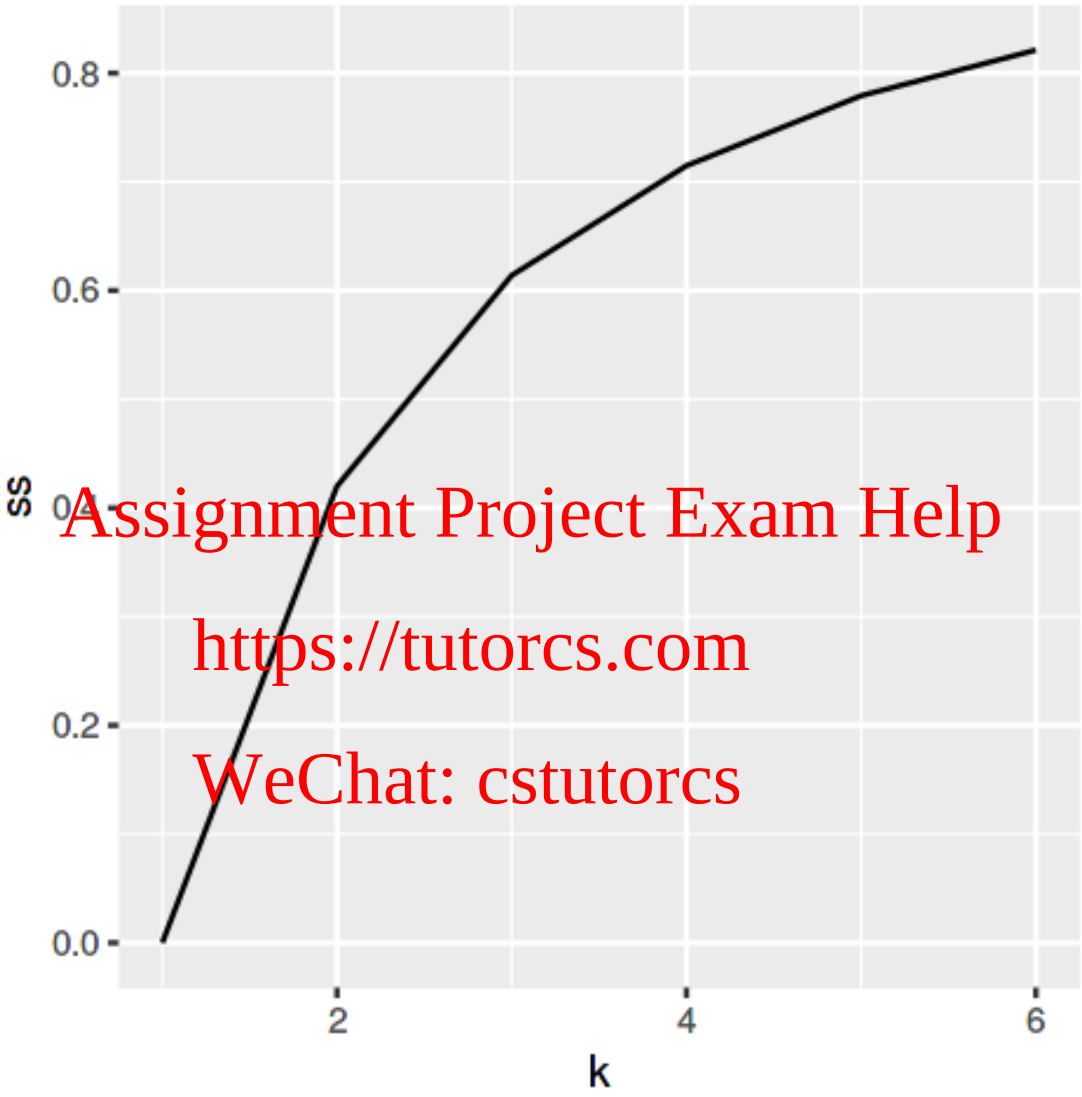


Plot the Cluster as color and `Age`, `Income` and `Children` as scatter plot with the cluster number. Write a for loop to calculate the within cluster sum of squares for 1 to 6 clusters using kmeans:

```
In [53]: n <- 6
         ss.df <- data.frame(k = numeric(n),
                             ss = numeric(n))
         for(i in 1: n){
             km            <- kmeans(ktcNumeric.df, centers=i, nstart=25)
             ss.df$k[i]    <- i
             ss.df$ss[i]   <- km$betweenss/km$totss
          }
```

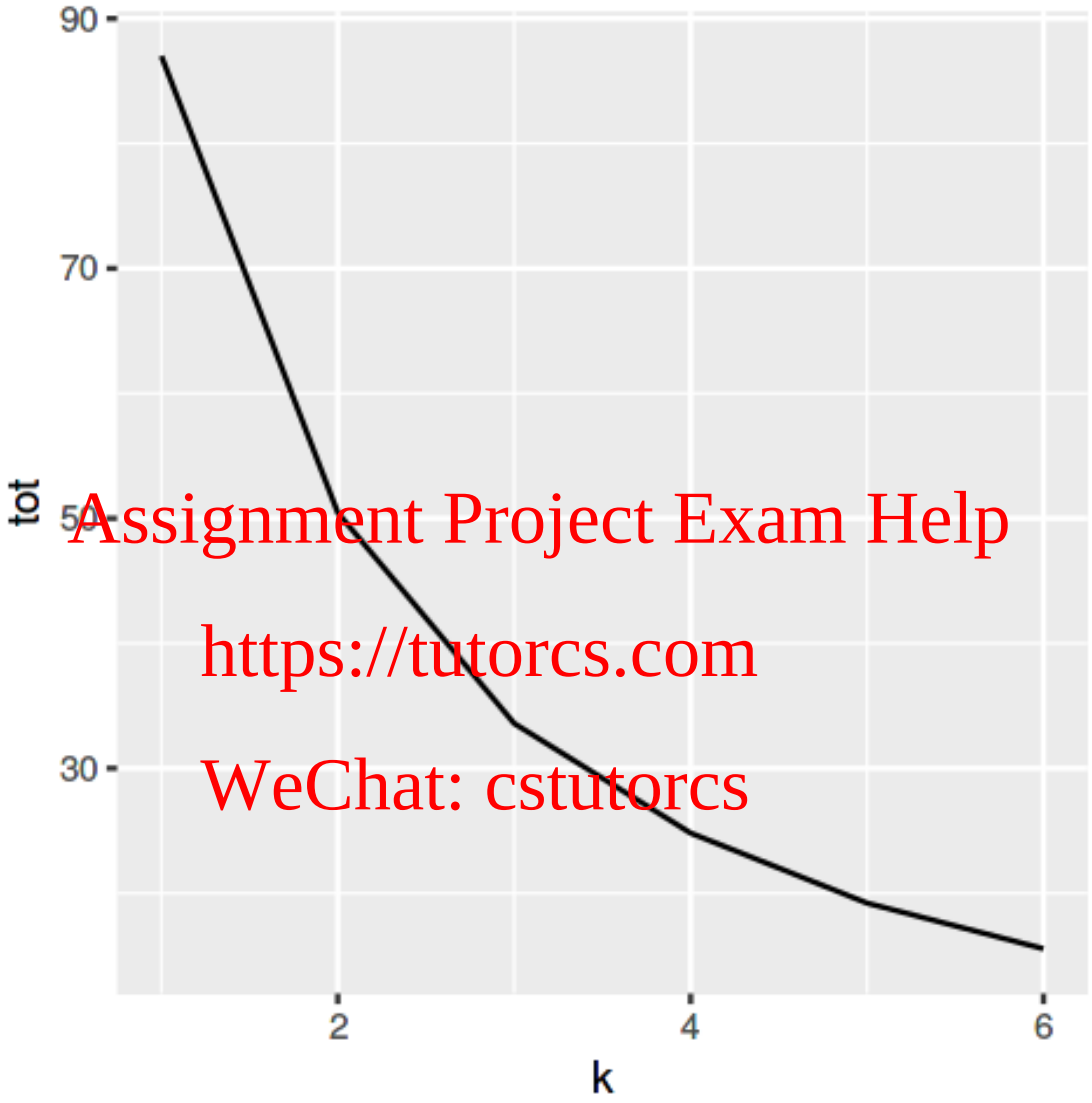
Create a line plot of `k` against `ss` and `tot`:

In [54]:
```
ggplot(data=ss.df,
       aes(x      = k,
           y      = ss))+
    geom_line()
```



In [68]:
```
tot.df <- data.frame(k = numeric(n),
                     tot = numeric(n))
for(i in 1: n){
    km            <- kmeans(ktcNumeric.df, centers=i, nstart=25)
    tot.df$k[i]    <- i
    tot.df$tot[i]   <- km$tot.withinss
 }
```

```
In [69]: ggplot(data=tot.df,
                aes(x      = k,
                    y      =tot))+
             geom_line()
```



1.0.2 Exercise 2: Analyse the clustering results

As per what we learned last class, compare the means across clusters.

```
In [ ]: # your code here
        fail() # No Answer - remove if you provide an answer
```

1.0.3 Exercise 3:

Apply clustering to the data-set `europe.csv` and compare kmeans and hierarchical cluster method. Are the groups similar to those obtained by hierarchical clustering?

```
In [ ]: # your code here
        fail() # No Answer - remove if you provide an answer
```

1.0.4 Review Questions

Q: What is meant by standardising variables and why is it done in cluster analysis?
Q: In k-means cluster analysis how should the number of clusters be determined?
Q: Describe the process by which clusters are chosen in k-means cluster analysis.