University of Moratuwa
Department of Computer Science & Engineering
MBA in Information Technology - 2018

Name & Index :

- D.N.K. Medawatta - 189115D   - T. M. H. Thennakoon - 189126L

- W.D.T.D Senanayake - 189124E

Title of Assignment : Analytics Challenge - PSID

Assignment No : 01   Group■   Individual □

Subject Code : CS5122

Subject : Descriptive and Predictive Analytics

Lecturer : Dr. T. Uthayashanker

Student's Statement :

We certify that we have not plagiarized the work of others or participated in unauthorized collusion when preparing this assignment

Office use only :

On/ before deadline   Extension Given   Late Submission

Marks Given :

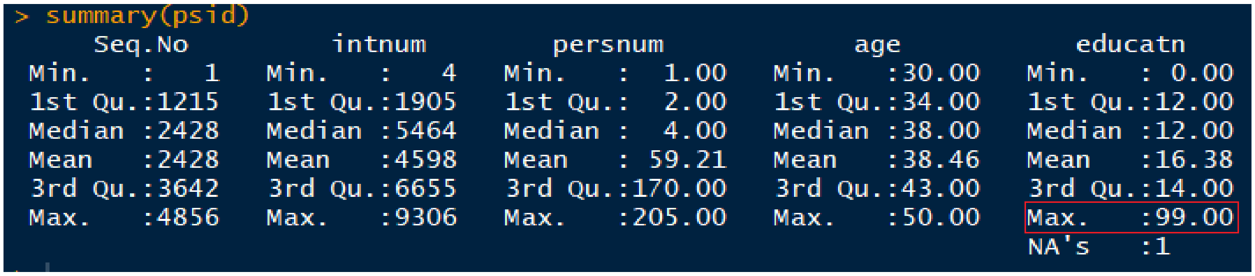
EXECUTIVE SUMMARY

1. Families who have higher education levels, can earn more
2. Families who have less education levels earn less
3. Some of the families who have higher education levels, still earn less
4. Families with more kids are can utilize a limited number of hours for working

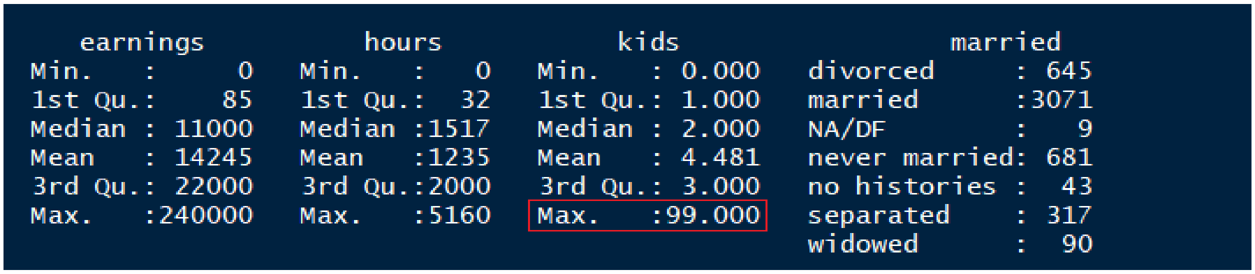
SUMMARY OF THE DATASET

According to the summary, Max (Education) = 99 & Max (Kids) = 99. These can be outliers. These outliers can cause for wrong analytics.

psid<-read.csv("PSID.csv")%>%
  distinct() %>%
summary(psid)

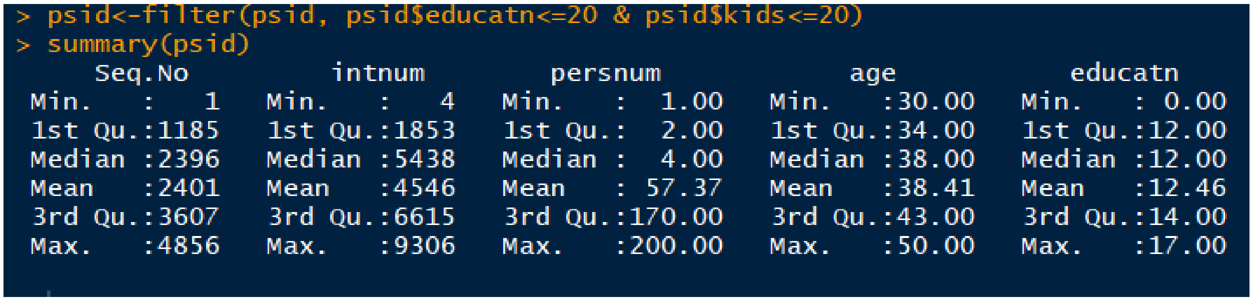


```
> summary(psid)
     Seq.No        intnum        persnum            age           educatn    
 Min.   :   1   Min.   :   4   Min.   :  1.00   Min.   :30.00   Min.   : 0.00  
 1st Qu.:1215   1st Qu.:1905   1st Qu.:  2.00   1st Qu.:34.00   1st Qu.:12.00  
 Median :2428   Median :5464   Median :  4.00   Median :38.00   Median :12.00  
 Mean   :2428   Mean   :4598   Mean   : 59.21   Mean   :38.46   Mean   :16.38  
 3rd Qu.:3642   3rd Qu.:6655   3rd Qu.:170.00   3rd Qu.:43.00   3rd Qu.:14.00  
 Max.   :4856   Max.   :9306   Max.   :205.00   Max.   :50.00   Max.   :99.00  
                                                                NA's   :1      
```

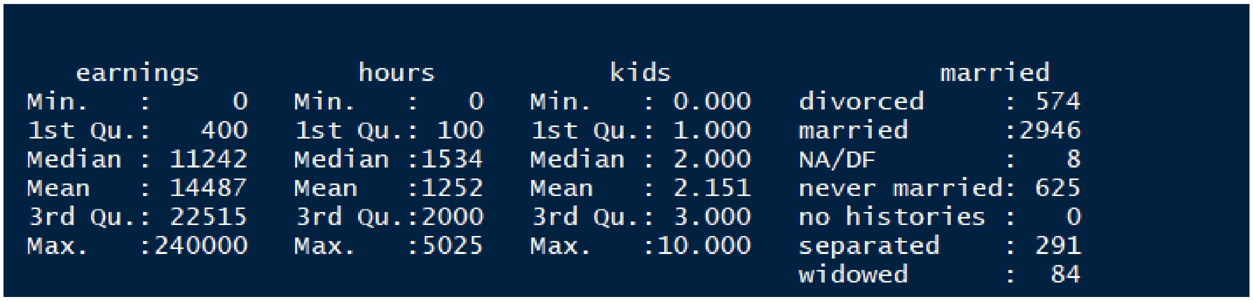


```
    earnings          hours           kids                married    
 Min.   :     0   Min.   :   0   Min.   : 0.000   divorced     : 645  
 1st Qu.:    85   1st Qu.:  32   1st Qu.: 1.000   married      :3071  
 Median : 11000   Median :1517   Median : 2.000   NA/DF        :   9  
 Mean   : 14245   Mean   :1235   Mean   : 4.481   never married: 681  
 3rd Qu.: 22000   3rd Qu.:2000   3rd Qu.: 3.000   no histories :  43  
 Max.   :240000   Max.   :5160   Max.   :99.000   separated    : 317  
                                                  widowed      :  90  
```

# Remove outliers
# Max Kids = 99 and Max Education = 99. These could be outliers
psid<-filter(psid, psid$educatn<=20 & psid$kids<=20)
summary(psid)



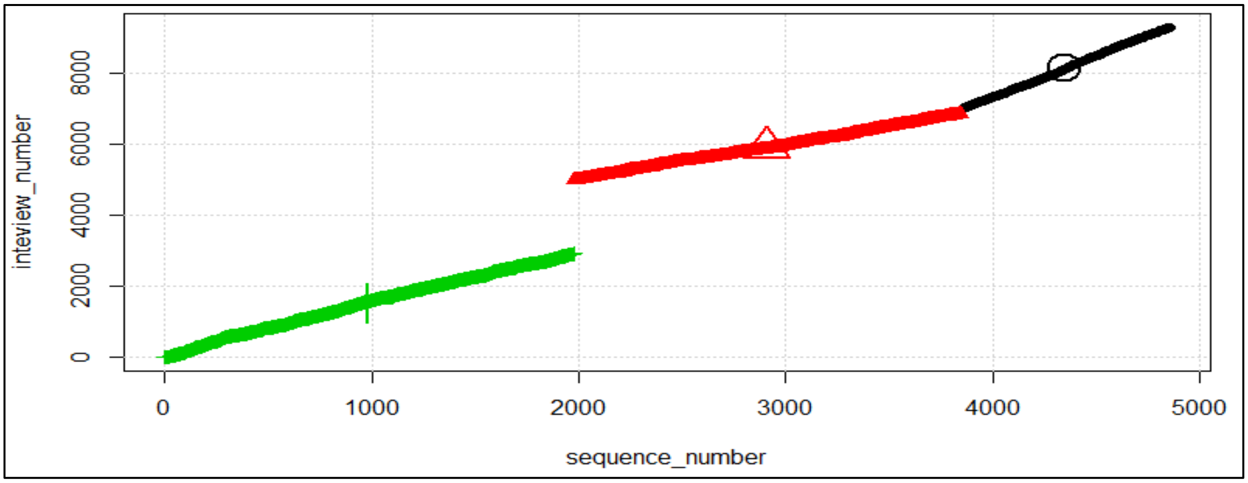
```
> psid<-filter(psid, psid$educatn<=20 & psid$kids<=20)
> summary(psid)
     Seq.No        intnum        persnum            age           educatn    
 Min.   :   1   Min.   :   4   Min.   :  1.00   Min.   :30.00   Min.   : 0.00  
 1st Qu.:1185   1st Qu.:1853   1st Qu.:  2.00   1st Qu.:34.00   1st Qu.:12.00  
 Median :2396   Median :5438   Median :  4.00   Median :38.00   Median :12.00  
 Mean   :2401   Mean   :4546   Mean   : 57.37   Mean   :38.41   Mean   :12.46  
 3rd Qu.:3607   3rd Qu.:6615   3rd Qu.:170.00   3rd Qu.:43.00   3rd Qu.:14.00  
 Max.   :4856   Max.   :9306   Max.   :200.00   Max.   :50.00   Max.   :17.00  
```



```
    earnings          hours           kids                married    
 Min.   :     0   Min.   :   0   Min.   : 0.000   divorced     : 574  
 1st Qu.:   400   1st Qu.: 100   1st Qu.: 1.000   married      :2946  
 Median : 11242   Median :1534   Median : 2.000   NA/DF        :   8  
 Mean   : 14487   Mean   :1252   Mean   : 2.151   never married: 625  
 3rd Qu.: 22515   3rd Qu.:2000   3rd Qu.: 3.000   no histories :   0  
 Max.   :240000   Max.   :5025   Max.   :10.000   separated    : 291  
                                                  widowed      :  84  
```

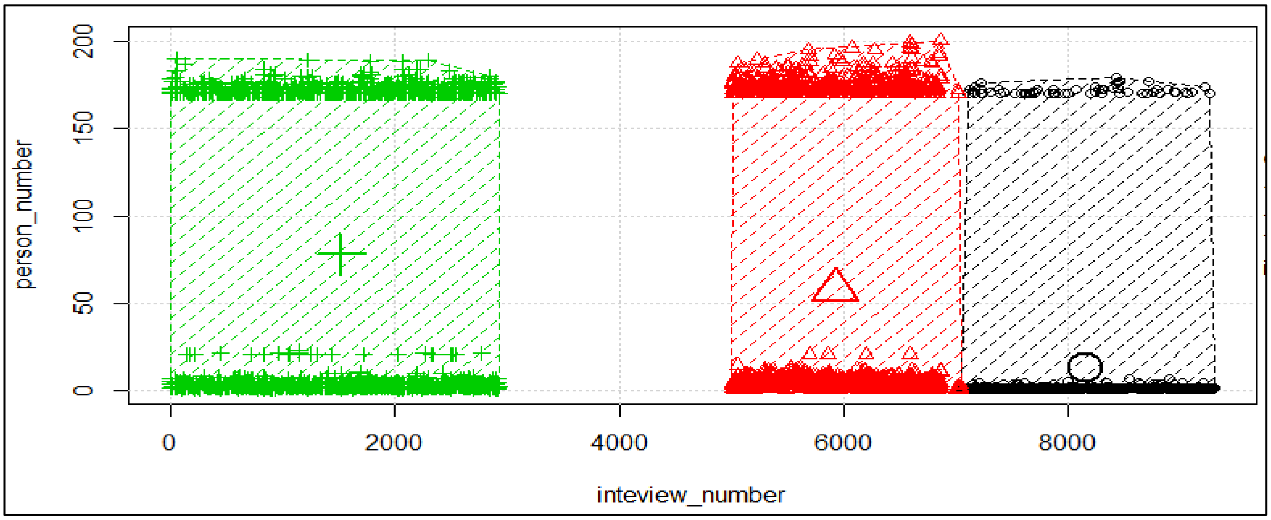
SAMPLE DATASET

```
psid<-psid %>%
 rename(
  sequence_number = Seq.No,
  interview_number = intnum,
  person_number = persnum,
  age = age,
  education = educatn,
  earnings = earnings,
  number_of_hours = hours,
  number_of_kids = kids,
  maritrial_status = married
 )
set.seed(12345)
kmeans.ani(psid[1:2], 3)
```



Following three clusters shows that there are six groups of people in the given sample.

```
set.seed(12345)
kmeans.ani(psid[2:3], 3)
```

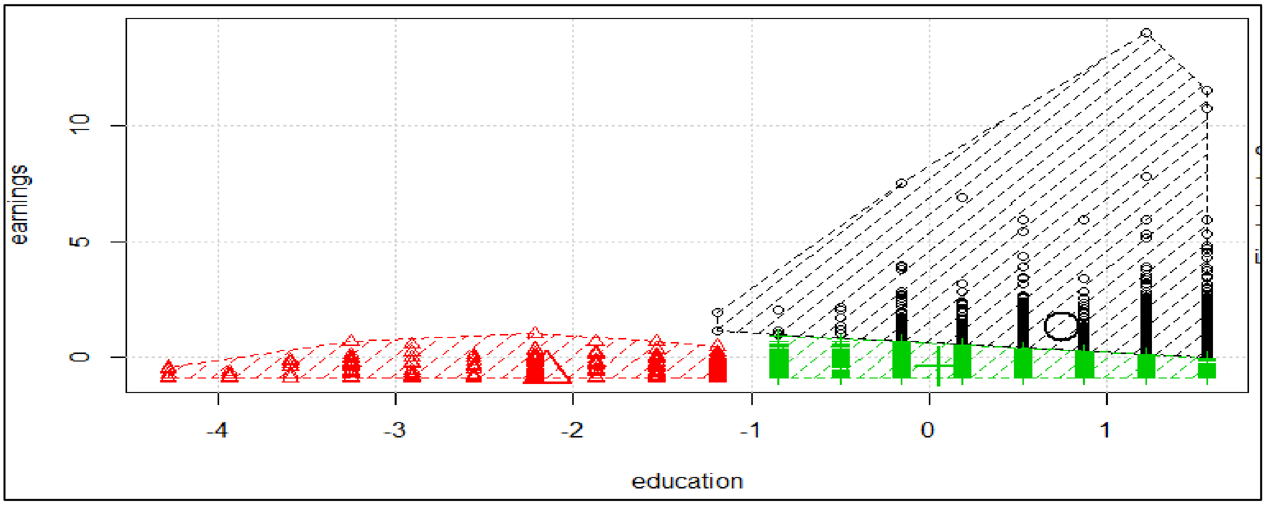


```
scaled_psid <- psid %>%
  mutate(age = scale (age),
       education = scale (education),
       earnings = scale(earnings),
       number_of_hours = scale (number_of_hours),
       number_of_kids = scale (number_of_kids)) %>%
  select(-c(sequence_number, interview_number, person_number, maritrial_status))
```

EDUCATION VS EARNINGS

```
set.seed(12345)
kmeans.ani(scaled_psid[1:2], 3)
```

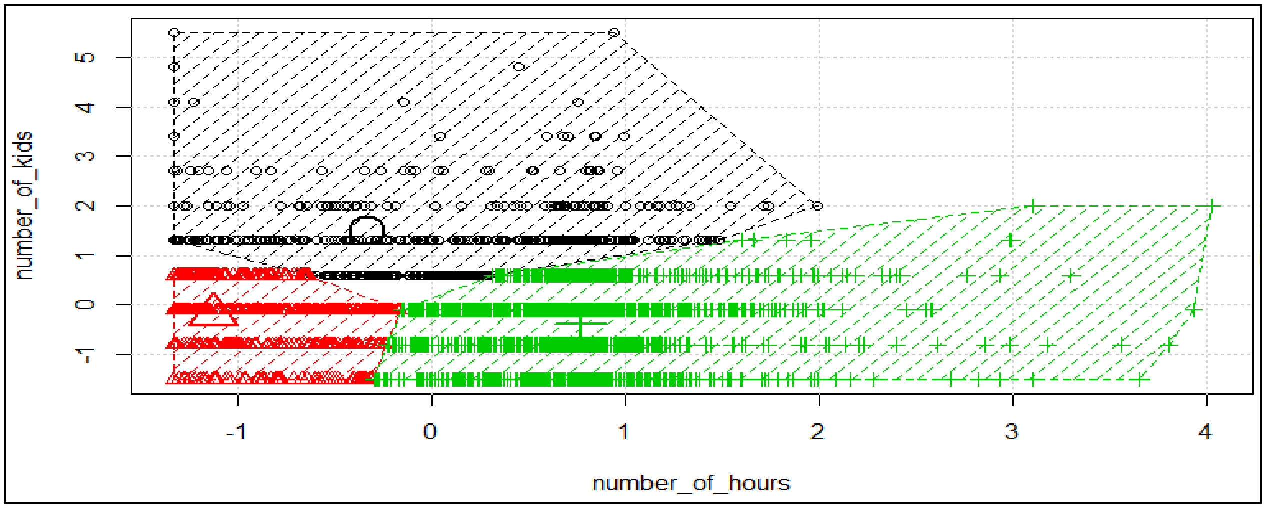
Following analysis reveals that, as the education level increased the earning level of the people is also increased.



NUMBER OF HOURS VS NUMBER OF KIDS

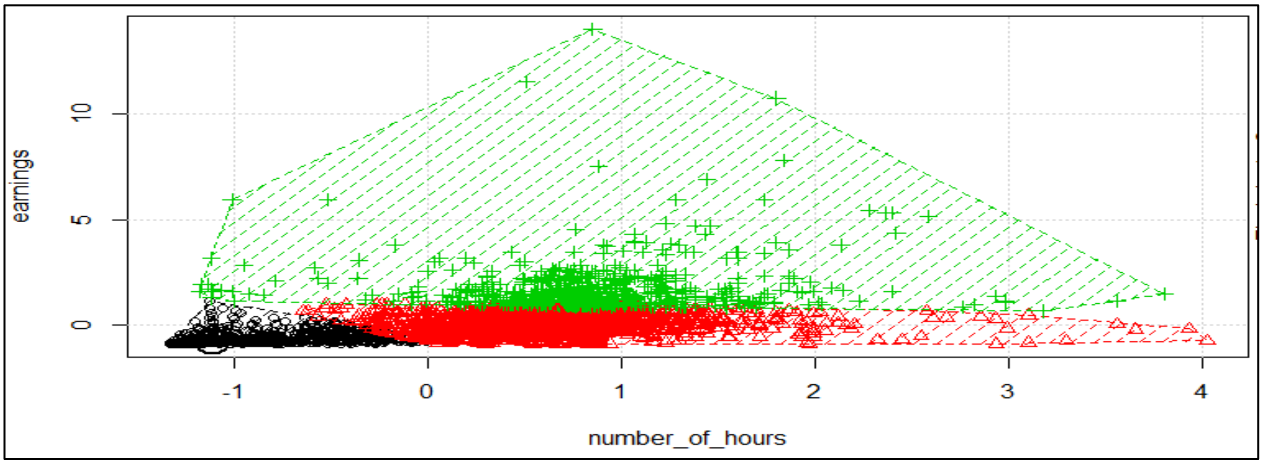
```
set.seed(12345)
kmeans.ani(scaled_psid[1:2], 3)
```

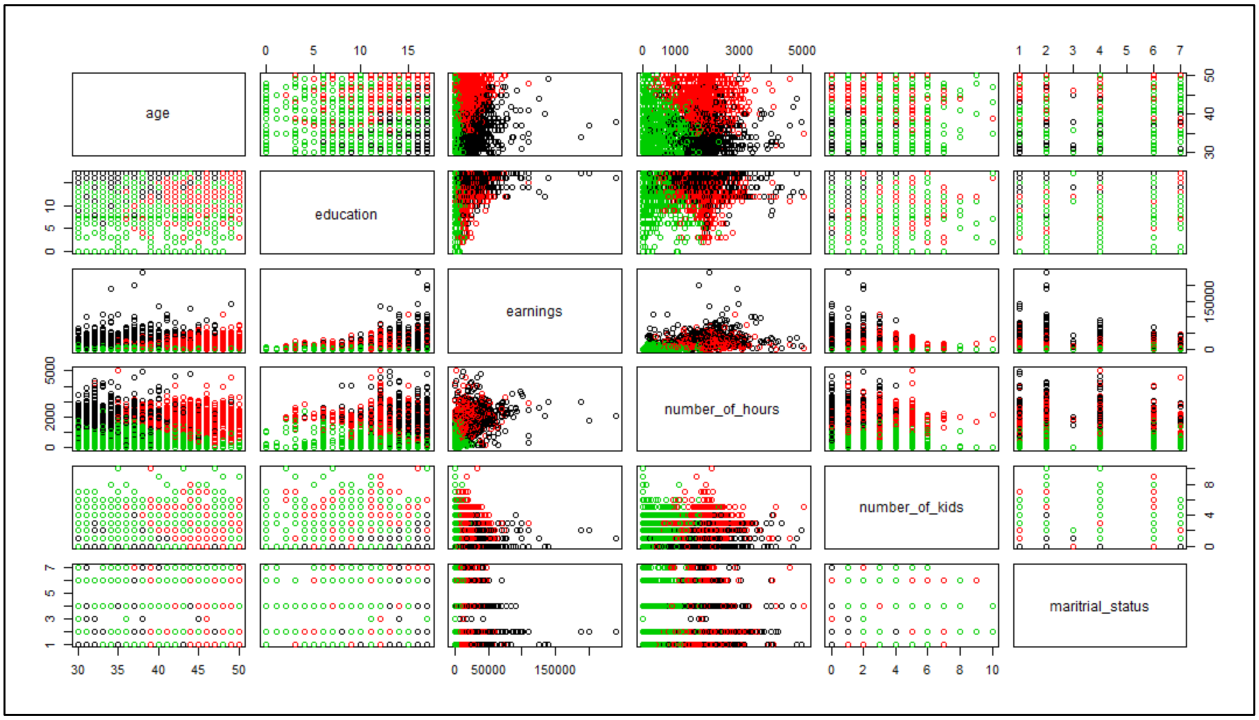
Following analysis reveals families with less number of kids work good number of hours. However, families with high number of kids are unable to work for high number of hours



NUMBER OF HOURS VS EARNINGS

```
set.seed(12345)
kmeans.ani(scaled_psid[4:3], 3)
```


earnings
number_of_hours


age
education
earnings
number_of_hours
number_of_kids
maritrial_status