MBA 2018 – DPA – TEAM E – Assignment 01

Executive Summary

- Majority of persons are currently married. It is 63.2% form data set
- 24.5% persons surveyed are unemployed
- 1,949 persons work above the average no of working hours per year
- 19 persons work more than 4,000 hours per year which is above double the average
- Mode value for number of kids is 2
- Earnings gives a right skewed histogram with Standard Deviation of 15985.45
- Age seems like symmetric. But it is slightly right skewed. Standard deviation is 5.59

Git Link : https://github.com/Nethmini/MBA18-DPA-Challenges

No.	Group Member	Contribution
13	K.M.N. Dulanjalee - 189103N	Documentation and Analyzing
14	M.I. Nazly - 189117K	Analyzing and Reporting
15	K.S.H. Sarathchandra - 189123B	Analyzing and Reporting

Data Analysis

Initially we tried to study the structure of given data set which is survey results of "**Panel Survey of Income Dynamics**"

When we run the ***str()*** command to get the structure of data set, it has identified all variable types correctly.
The data frame consists of 9 variables in which 8 are integers and 1 factor.

```
> str(PSID)
'data.frame':   4856 obs. of  9 variables:
 $ Seq.No  : int  1 2 3 4 5 6 7 8 9 10 ...
 $ intnum  : int  4 4 4 4 5 6 6 7 7 7 ...
 $ persnum : int  4 6 7 173 2 4 172 4 170 171 ...
 $ age     : int  39 35 33 39 47 44 38 38 39 37 ...
 $ educatn : int  12 12 12 10 9 12 16 9 12 11 ...
 $ earnings: int  77250 12000 8000 15000 6500 6500 7000 5000 21000 0 ...
 $ hours   : int  2940 2040 693 1904 1683 2024 1144 2080 2575 0 ...
 $ kids    : int  2 2 1 2 5 2 3 4 3 5 ...
 $ married : Factor w/ 7 levels "divorced","married",..: 2 1 2 2 2 2 2 1 2 2 ...
```

Then we executed **summary()** on data frame.

```
> summary(PSID)
     Seq.No         intnum         persnum            age          educatn         earnings
 Min.   :   1   Min.   :   4   Min.   :  1.00   Min.   :30.00   Min.   : 0.00   Min.   :     0
 1st Qu.:1215   1st Qu.:1905   1st Qu.:  2.00   1st Qu.:34.00   1st Qu.:12.00   1st Qu.:    85
 Median :2428   Median :5464   Median :  4.00   Median :38.00   Median :12.00   Median : 11000
 Mean   :2428   Mean   :4598   Mean   : 59.21   Mean   :38.46   Mean   :16.38   Mean   : 14245
 3rd Qu.:3642   3rd Qu.:6655   3rd Qu.:170.00   3rd Qu.:43.00   3rd Qu.:14.00   3rd Qu.: 22000
 Max.   :4856   Max.   :9306   Max.   :205.00   Max.   :50.00   Max.   :99.00   Max.   :240000
                                                                NA's   :1
     hours            kids                married
 Min.   :   0   Min.   : 0.000   divorced     : 645
 1st Qu.:  32   1st Qu.: 1.000   married      :3071
 Median :1517   Median : 2.000   NA/DF        :   9
 Mean   :1235   Mean   : 4.481   never married: 681
 3rd Qu.:2000   3rd Qu.: 3.000   no histories :  43
 Max.   :5160   Max.   :99.000   separated    : 317
                                 widowed      :  90
```

Then we tried to analyze each variable in the data frame.

1. "MARRIED"

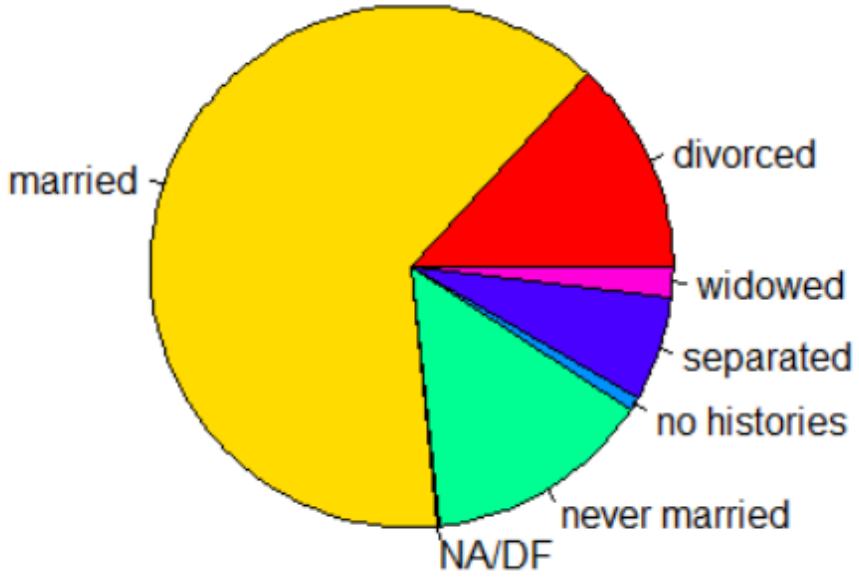
We started with "married" as it was the only categorical variable.

Running ***show(table(PSID$married)/length(PSID$married))*** gave us the percentages of each marital status.

```
    divorced      married          NA/DF never married  no histories
0.132825371   0.632413509   0.001853377   0.140238880   0.008855025   0
```

Maximum participation was from married which was 63.2%.

Below is the representation in a pie chart.



2. Non-Working (HOURS)

A subset of data created using ***psid_non_working = subset(PSID, hours == 0)***

```
> nrow(psid_non_working)
[1] 1190
> nrow(PSID)
[1] 4856
> nrow(psid_non_working)/nrow(PSID)
[1] 0.2450577
```

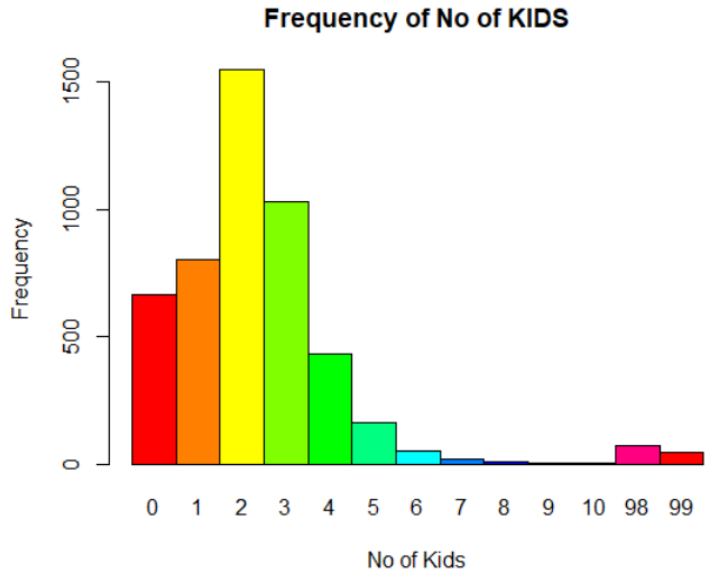
24.5% persons surveyed are unemployed.

Average number of working hours per year in US is 1,811. Considering above data, 1949 persons work above the average no of working hours per year which will be harmful for their health

```
> nrow(subset(PSID, hours > 1811))
[1] 1949
```

Around 19 persons work more that 4000 hours per year which is above double the average.

3. KIDS

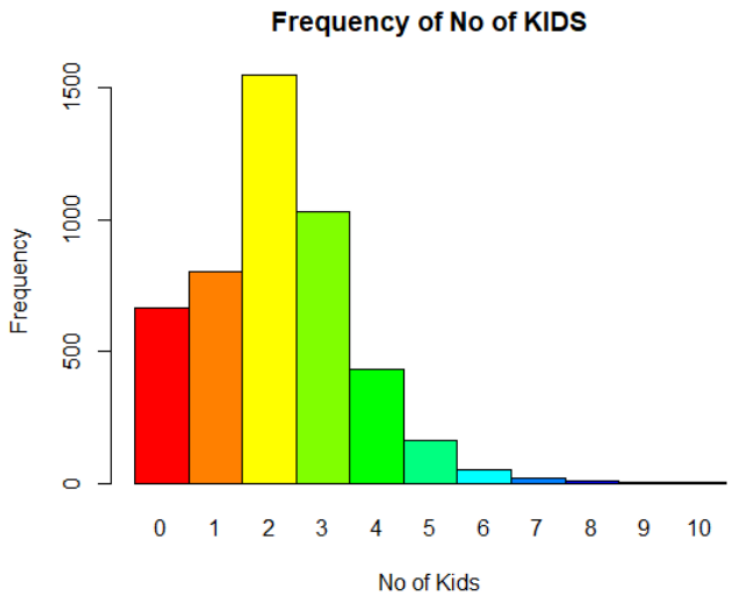


0	1	2	3	4	5	6	7	8	9	10	98	99
668	803	1549	1029	435	165	52	21	8	4	4	73	45

There were records with KIDS count as 98 (73) and 99 (45) as well. Which seems to be anomalies or outliers.

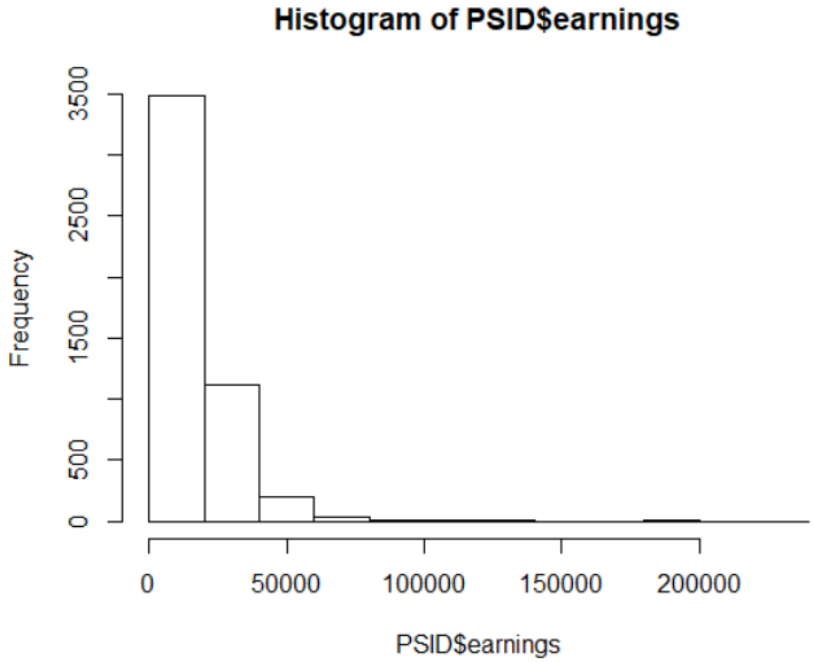
Therefore, a subset of data is created with number of kids up to 10.

0	1	2	3	4	5	6	7	8	9	10
668	803	1549	1029	435	165	52	21	8	4	4



4. EARNINGS

When earnings is considered, it gives a right skewed histogram.

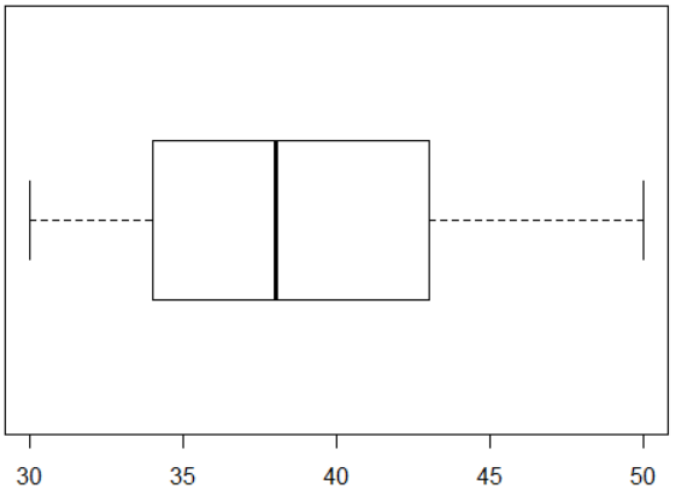


```
> #EARNING
> median(PSID$earnings)
[1] 11000
> sd(PSID$earnings)
[1] 15985.45
> #EARNING
> mean(PSID$earnings)
[1] 14244.51
> var(PSID$earnings)
[1] 255534530
```

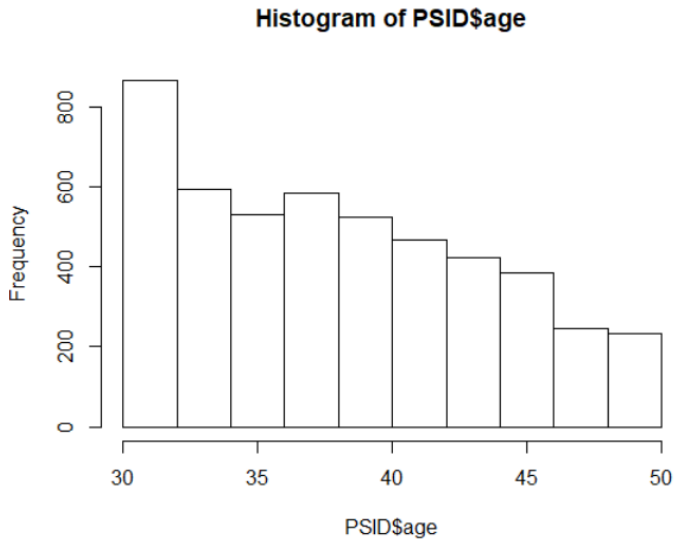
```
> summary(PSID$earnings)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0      85   11000   14245   22000  240000
```

5. AGE

Age seems like symmetric. But it is slightly right skewed



It is clearly visible when visualized into a histogram.

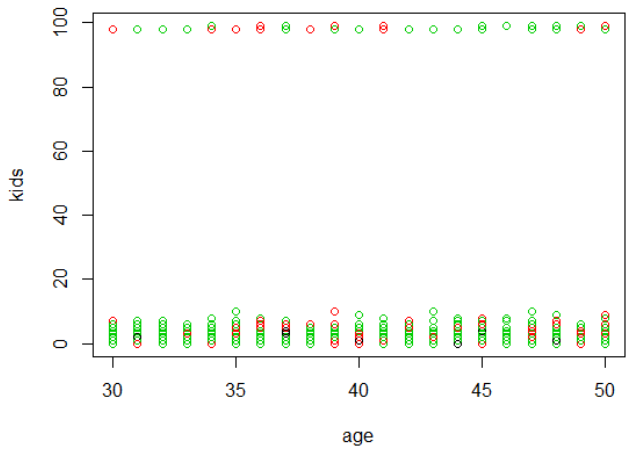

Histogram of PSID$age
Frequency
0
200
400
600
800
30
35
40
45
50
PSID$age

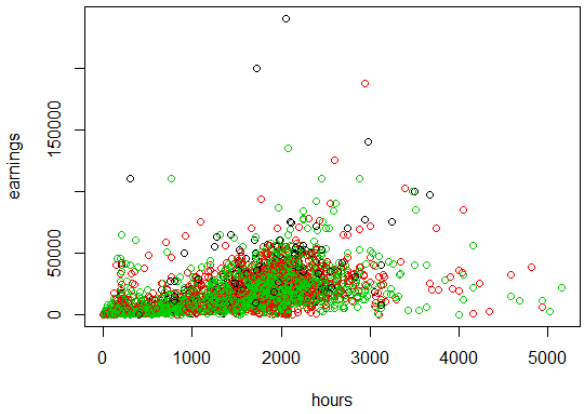
Application of K-Means Clustering

```
> kmeans_results = kmeans(na.omit(PSID.features), 7)
> kmeans_results
K-means clustering with 7 clusters of sizes 681, 14, 152, 1334, 591, 1147, 936

Cluster means:
    intnum  persnum      age  educatn   earnings     hours     kids
1 1562.816 79.53451 38.28194 16.86490   1823.587  433.5712 4.810573
2 1810.857 27.64286 38.71429 15.42857 133321.429 2486.5000 1.357143
3 3222.651 75.96711 40.21053 18.64474  58220.039 2103.2632 1.723684
4 6886.935 45.00825 38.06372 15.77286   1022.565  320.7211 6.849325
5 3920.492 60.41455 39.42301 17.79695  35142.335 1993.5161 2.478849
6 4599.357 56.46295 38.22406 15.31473  12292.439 1694.3208 3.861378
7 4240.405 65.09509 38.56624 16.93483  22415.452 1922.6720 3.389957
```

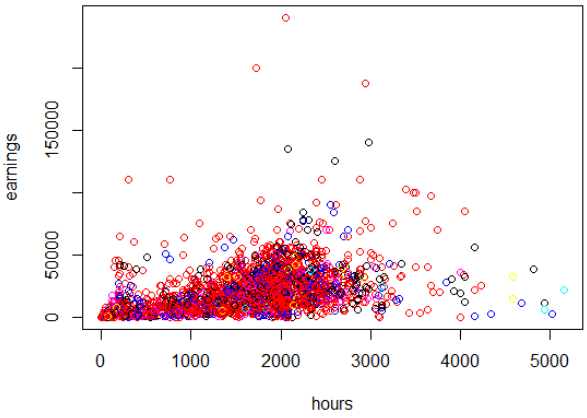
Number of clusters reduced to 3.

```
> kmeans_results = kmeans(na.omit(PSID.features), 3)
> kmeans_results
K-means clustering with 3 clusters of sizes 349, 1789, 2717

Cluster means:
    intnum  persnum      age  educatn earnings     hours     kids
1 3505.599 68.61605 40.08023 17.92837 52203.57 2064.7937 2.143266
2 4250.319 61.57854 38.59419 16.78256 22974.23 1910.5176 3.174958
3 4969.032 56.46816 38.17004 15.91093  3625.83  684.6732 5.643357
```

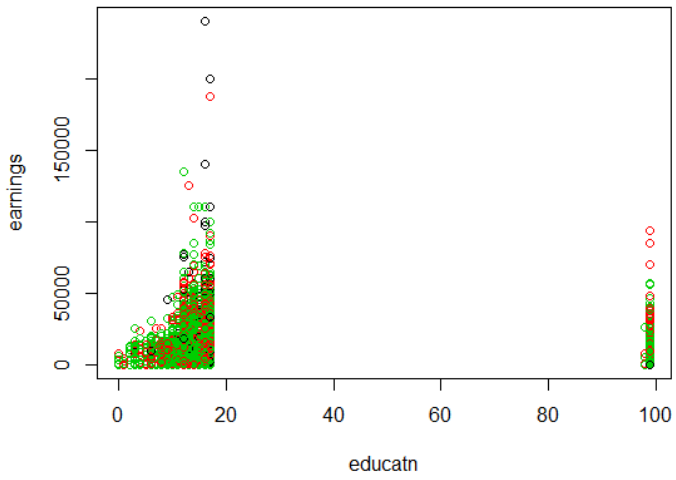




Plot with clustered data set: **plot(PSID[c("educatn", "earnings")], col = kmeans_results$cluster)**



Plot with original data set: **plot(PSID[c("earnings", "hours")], col = PSID$married)**


earnings
0
50000
150000
0
20
40
60
80
100
educatn

Appendix - R Code

```
PSID = read.csv("PSID.csv", header = TRUE)

View(PSID)

barplot(table(PSID$married))

levels(PSID$married)

table(PSID$married)

barplot(xtabs(~PSID$married), space = F, col = rainbow(7), ylab = "Frequency")

barplot(xtabs(~PSID$married), space = F, col = rainbow(7),legend.text = T, main = "RECORDS BY
MARRIED" , ylab = "Frequency")

barplot(xtabs(~PSID$married), space = F, col = rainbow(length(levels(PSID$married))), main = "RECORDS
BY MARRIED" , ylab = "Frequency")

PSID_temp <- read.csv("PSID.csv")

head(PSID_temp)



married = subset(PSID, married %in% "married")

View(married)

str(PSID)

pie(table(PSID$married) , col = rainbow(7))

pie(table(PSID$married)/length(PSID$married) , col = rainbow(7))

show(table(PSID$married)/length(PSID$married))

str(PSID)


View(summary(PSID))

show(summary(PSID$age))

show(table(PSID$married)/length(PSID$married))

pie(table(PSID$married)/length(PSID$married) , col = rainbow(7))
```

```
#WORK

psid_non_working = subset(PSID, hours == 0)

nrow(subset(PSID, hours > 1811))

nrow(subset(PSID, hours > 4000))

show(psid_non_working)


head(psid_non_working)


psid_non_working[, c("age", "kids", "married")]

nrow(psid_non_working)

nrow(PSID)

nrow(psid_non_working)/nrow(PSID)


#KIDS

barplot(xtabs(~PSID$kids), space = F, col = rainbow(12), main = "Frequency of No of KIDS" , ylab =
"Frequency", xlab = "No of Kids")

show(table(PSID$kids))



psid_kids = subset(PSID, kids < 11)

summary(psid_kids)

barplot(xtabs(~psid_kids$kids), space = F, col = rainbow(12), main = "Frequency of No of KIDS" , ylab =
"Frequency", xlab = "No of Kids")

show(table(psid_kids$kids))



#EARNING

mean(PSID$earnings)

median(PSID$earnings)
```

```
var(PSID$earnings)
sd(PSID$earnings)
hist(PSID$earnings)
boxplot(PSID$earnings)
summary(PSID$earnings)

#AGE
sd(PSID$age)
hist(PSID$age)
boxplot(PSID$age, horizontal = T)

#KMEANS
View(PSID)
PSID.features = PSID
PSID.features$married <- NULL
PSID.features$Seq.No <- NULL
View(PSID.features)
head(PSID.features)

kmeans_results = kmeans(na.omit(PSID.features), 3)
kmeans_results
kmeans_results$cluster
kmeans_results$size

table(PSID$married, kmeans_results$cluster)

plot(PSID)
plot(PSID[c("earnings", "hours")], col = kmeans_results$cluster)
plot(PSID[c("age", "kids")], col = kmeans_results$cluster)
```

```
plot(PSID[c("educatn", "earnings")], col = kmeans_results$cluster)

plot(PSID[c("hours", "earnings")], col = PSID$married)
```