FinalProject

Gurleen Kaur

12/18/2019

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)

##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

+++++++++++++++++++ ProjectStart ++++++++++++++++++++++++

The code below is used to read the original csv files, extract the columns to be used for analysis then save them to the new csv files.

```
psam_pusa <- read_csv("C:/Users/gurleen/Downloads/csv_pus/psam_pusa.csv")

psam_pusb <- read_csv("C:/Users/gurleen/Downloads/csv_pus/psam_pusb.csv")

psam_pusc <- read_csv("C:/Users/gurleen/Downloads/csv_pus/psam_pusc.csv")

psam_pusd <- read_csv("C:/Users/gurleen/Downloads/csv_pus/psam_pusd.csv")

psam_pusa <- psam_pusa %>%
  select(COW,PWGTP,AGEP,DDRS,DEAR,DEYE,DOUT,DREM,SEX,DIS,
       PRIVCOV,PUBCOV,OCCP)

psam_pusb <- psam_pusb %>%
  select(COW,PWGTP,AGEP,DDRS,DEAR,DEYE,DOUT,DREM,SEX,DIS,
       PRIVCOV,PUBCOV,OCCP)

psam_pusc <- psam_pusc %>%
  select(COW,PWGTP,AGEP,DDRS,DEAR,DEYE,DOUT,DREM,SEX,DIS,
       PRIVCOV,PUBCOV,OCCP)

psam_pusd <- psam_pusd %>%
  select(COW,PWGTP,AGEP,DDRS,DEAR,DEYE,DOUT,DREM,SEX,DIS,
       PRIVCOV,PUBCOV,OCCP)

write.csv(psam_pusa,'psam_pusa_final.csv')
write.csv(psam_pusb,'psam_pusb_final.csv')
write.csv(psam_pusc,'psam_pusc_final.csv')
write.csv(psam_pusd,'psam_pusd_final.csv')
```

The below chunk is used to read the final selected data from the csv files to be used for analysis.

```
psam_pusa_final <- read_csv("psam_pusa_final.csv")

## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   COW = col_double(),
##   PWGTP = col_character(),
##   AGEP = col_character(),
##   DDRS = col_double(),
##   DEAR = col_double(),
##   DEYE = col_double(),
##   DOUT = col_double(),
##   DREM = col_double(),
##   SEX = col_double(),
##   DIS = col_double(),
##   PRIVCOV = col_double(),
##   PUBCOV = col_double(),
##   OCCP = col_character()
## )

psam_pusb_final <- read_csv("psam_pusb_final.csv")

## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   COW = col_double(),
##   PWGTP = col_character(),
##   AGEP = col_character(),
##   DDRS = col_double(),
##   DEAR = col_double(),
##   DEYE = col_double(),
##   DOUT = col_double(),
##   DREM = col_double(),
##   SEX = col_double(),
##   DIS = col_double(),
##   PRIVCOV = col_double(),
##   PUBCOV = col_double(),
##   OCCP = col_character()
## )
```

```
psam_pusc_final <- read_csv("psam_pusc_final.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   COW = col_double(),
##   PWGTP = col_character(),
##   AGEP = col_character(),
##   DDRS = col_double(),
##   DEAR = col_double(),
##   DEYE = col_double(),
##   DOUT = col_double(),
##   DREM = col_double(),
##   SEX = col_double(),
##   DIS = col_double(),
##   PRIVCOV = col_double(),
##   PUBCOV = col_double(),
##   OCCP = col_character()
## )
```

```
psam_pusd_final <- read_csv("psam_pusd_final.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   COW = col_double(),
##   PWGTP = col_character(),
##   AGEP = col_character(),
##   DDRS = col_double(),
##   DEAR = col_double(),
##   DEYE = col_double(),
##   DOUT = col_double(),
##   DREM = col_double(),
##   SEX = col_double(),
##   DIS = col_double(),
##   PRIVCOV = col_double(),
##   PUBCOV = col_double(),
##   OCCP = col_character()
## )
```

The below chunk is used to combine all the datasets into one.

```
psam_pus_final <-
union(psam_pusa_final,psam_pusb_final,psam_pusc_final,psam_pusd_final)
```

Data Summary:

The data chosen for the analysis is the of the Person Record. It has the detailed information about a person living in the US. The person's origin, employment and health related attributes are present in this dataset.

For my analysis, I have choosed attributes related to a person's health and its employment. It seemeed interesting to me to see how the person health can be related to his/hers working class. So the columns chosen for analysis will tell us if there are some interesting insights regarding the person's class of work and their's health. The health columns in the dataset were given as the common difficulties faced by people. It includes: hearing difficulty (DEAR),visual difficulty (DEYE), self-care difficulty (DDRS),independant living difficulty (DOUT) and cognitive difficulty (DREM). These are the columns which tells us about the person's health. The age (AGEP) and the sex (SEX) of the person has also been included in the data.

To analyse the data the below chunk is used to convert the age (AGEP) of the person as numeric and the class of workers (COW) as a factor.

```
psam_pus_final$AGEP <- as.numeric(psam_pus_final$AGEP)
psam_pus_final <- psam_pus_final %>% filter(COW!=9)
psam_pus_final$COW <- as_factor(psam_pus_final$COW)
levels(psam_pus_final$COW)

## [1] "1" "2" "3" "4" "5" "6" "7" "8"
```

The class of workers (COW) variable has 1-9 digit values each of which represents the class under which the person's employment falls. The last category (9 value) shows if the person is unemployed.

To give more meaning values to the COW. The below chunk is executed.

```
levels(psam_pus_final$COW) <- c("Private-Profitable Company
Employee","Private Non-profitable Company Employee","Local Government
Employee","State Governemnt Employee",
"Federal Government Employee",
"Self Employed in own not incorporated business",
"Self Employed in own incorporated business",
"Working without pay in family business or farm")
```

This will make it easy to visualise the data. Below is the summary of the data.

```
summary(psam_pus_final)
```

```
##        X1
##  Min.   :      1
##  1st Qu.:1029577
##  Median :2045804
##  Mean   :2080880
##  3rd Qu.:3027601
##  Max.   :4691834
##
##                                               COW
##  Private-Profitable Company Employee           :3062973
##  Private Non-profitable Company Employee       : 392831
##  Local Government Employee                     : 342287
##  Self Employed in own not incorporated business: 326614
##  State Governemnt Employee                     : 223206
##  Self Employed in own incorporated business    : 181331
##  (Other)                                       : 169871
##     PWGTP                AGEP           DDRS             DEAR
##  Length:4699113     Min.   :16.0   Min.   :1.000   Min.   :1.000
##  Class  :character   1st Qu.:31.0   1st Qu.:2.000   1st Qu.:2.000
##  Mode   :character   Median :45.0   Median :2.000   Median :2.000
##                      Mean   :44.3   Mean   :1.989   Mean   :1.974
##                      3rd Qu.:57.0   3rd Qu.:2.000   3rd Qu.:2.000
##                      Max.   :97.0   Max.   :2.000   Max.   :2.000
##
##       DEYE            DOUT            DREM            SEX
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000
##  Median :2.000   Median :2.000   Median :2.000   Median :1.000
##  Mean   :1.984   Mean   :1.979   Mean   :1.972   Mean   :1.488
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000
##
##       DIS            PRIVCOV          PUBCOV           OCCP
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Length:4699113
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000   Class  :character
##  Median :2.000   Median :1.000   Median :2.000   Mode   :character
##  Mean   :1.911   Mean   :1.238   Mean   :1.782
##  3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:2.000
##  Max.   :2.000   Max.   :2.000   Max.   :2.000
##
```

The summary of this table shows us the sum of observations for each category in class of workers (COW). This information will be used later for the analysis. This also tells us the mean, median and quantiles for the variables.

Methodology:

From the final dataframe (psam_pus_final), firstly I wanted to see the average age of a person working in the respective class. This would help me see later if age can be considered as a factor of number of health dificulties per class of workers. So to see this, the two columns were selected from the final dataframe (psam_pus_final) and then filtered for NA values and for age >=18 as I was interested to see only the data for the adults. The mean of the age column was determined using mean() function in summarise and the graph was plotted for average age vs class of workers

```
cow_age <- psam_pus_final %>% select(COW,AGEP) %>%
  filter(!is.na(COW), AGEP>=18) %>%
  group_by(COW) %>%
  summarise(average=mean(AGEP,na.rm=TRUE))

first_analysis <- cow_age %>%
  ggplot(aes(x=COW,y=average))+
  geom_point(aes(color=COW),size=4)+
  xlab("Class Of Workers")+
  ylab("Average Age")+
  labs(color="Class Of Workers")+
  theme(axis.text.x = element_text(face="bold",angle=90))
```

Secondly, the percentage of people with different 5 kinds of difficulties (hearing, visual, self-care,independant living and cognitive) in their respective class of work was analysed. To acheive this, a separate table having the the class of workers and all the difficulties was created. The NA values in the COW were filtered out and the data was gathered so count the total number of people in each of the COW and difficulty combination. This table will be used to calulate the percentage.

```
COW_difficulties_total_count <- psam_pus_final %>%
  select(COW,DDRS,DEAR,DEYE,DOUT,DREM) %>%
  filter(!is.na(COW)) %>%
  gather("DDRS","DEAR","DEYE","DOUT","DREM",
         key="Difficulties",
         value="Response") %>%
  group_by(COW,Difficulties) %>%
  summarise(Total=n())
```

Then, the table above is joined with the below table to see what percentage of people have each kind of difficulty in each class of work. The bar graph is used to plot the percentage of people having different difficulties vs each class of work.

```
cow_difficulties <- psam_pus_final %>%
select(COW,DDRS,DEAR,DEYE,DOUT,DREM)%>%
  filter(!is.na(COW)) %>%
  gather("DDRS","DEAR","DEYE","DOUT","DREM",
         key="Difficulties",
         value="Response") %>%
  filter(Response==1)%>%
  group_by(COW,Difficulties) %>%
  summarise(count=sum(Response)) %>%
  inner_join(COW_difficulties_total_count,by=c("COW",
        "Difficulties")) %>%
  mutate(Percentage=(count/Total)*100)
second_analysis <- cow_difficulties %>%
  ggplot()+
  geom_bar(aes(x=COW,y=Percentage,fill=Difficulties),
           stat="identity",position ="dodge")+
  xlab("Class Of Workers")+
  ylab("Percentage Of People")+
  scale_fill_manual(labels=c("Self care Difficulty",
                             "Hearing Difficulty",
                             "Vision Difficulty",
                             "Independent living
                              Difficulty",
                             "Cognitive Difficulty"),
                    values=c("#F8766D","#BB9D00","#00B81F",
                              "#00A5FF","#E76BF3"))+
  theme(axis.text.x = element_text(face="bold",angle=90))
```

Thirdly, after visually analysing the difficulties, I wished to check weather those difficulties could have been the cause of disability and how these disabilities are related with the class of workers. So, initially a seperate table was created to calculate the total number of people in each class of work. Then this table was joined with another table having disability (DIS) column and filtered for the people having disability.(DIS==1) The results were plotted using a bar graph.

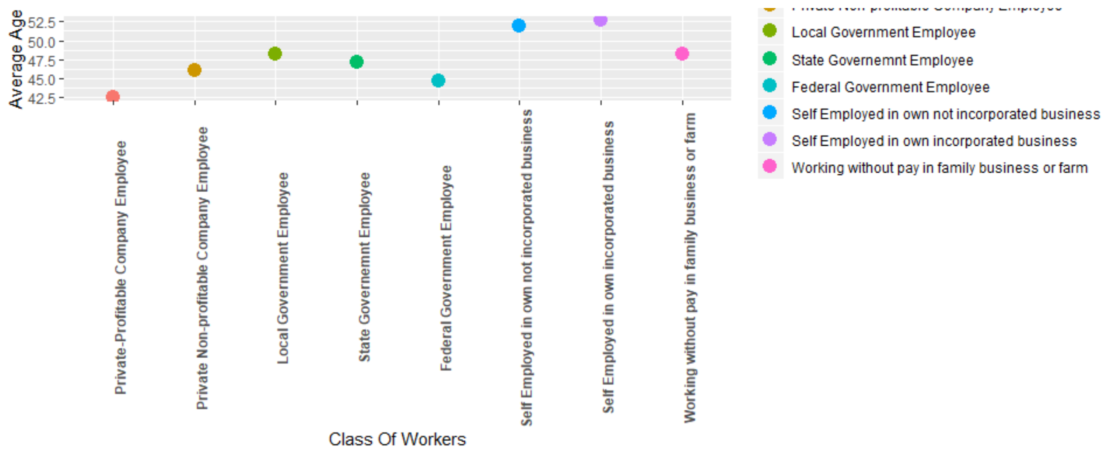
```
COW_total_count <- psam_pus_final %>% select(COW) %>%
  filter(!is.na(COW)) %>%
  group_by(COW) %>%
  summarise(Total=n())

cow_disability <- psam_pus_final %>% select(COW,DIS) %>%
  filter(!is.na(COW)) %>%
  group_by(COW) %>%
  filter(DIS==1) %>%
  summarise(count=sum(DIS)) %>%
  inner_join(COW_total_count,by="COW") %>%
  mutate(Percentage=(count/Total)*100)

third_analysis <- cow_disability %>%
  ggplot()+
  geom_bar(aes(x=COW,y=Percentage),stat="identity")+
  theme(axis.text.x = element_text(face="bold",angle=90))+
  xlab("Class Of Workers")+
  ylab("People with Disabilities (percentage)")
```

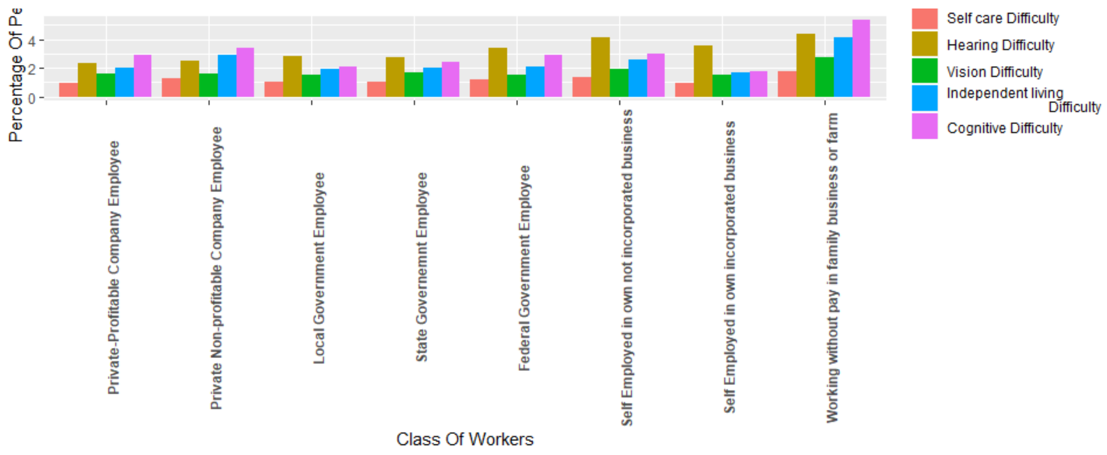
Findings:

```
first_analysis
```



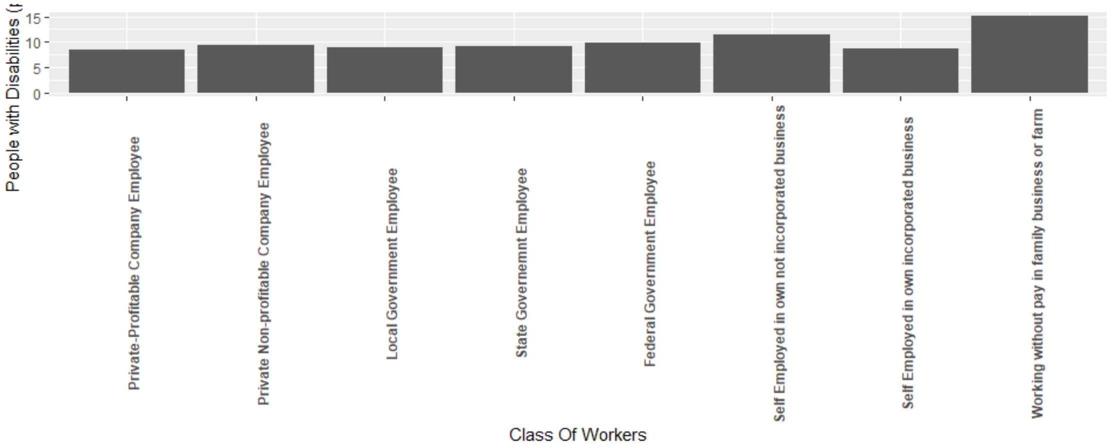
In this graph we can see how the average age of people varies with the class of work. It makes sense that the highest average age of people are those which are self employed in own incorporated business as at an young age people tend to work for someone or under someone. This explains why private profitable company employee has less average age of people.

```
second_analysis
```



After the second analysis, it can be seen that people who work without pay in family business or farm are the one which have higher percentage of diffculties. Within that category(working without pay in family), higher percentage of people suffer from cognitive difficulties such as trouble learning new things, concentrating, or making decisions that affect their everyday life. It make sense that higher percentage of people which are self employed in own business have hearing difficulty as we saw from our first analysis that these are the people with highest average age. Older people tend to have hearing difficulties. This category people have least percentage of other difficulties as compare to people working in other classes.

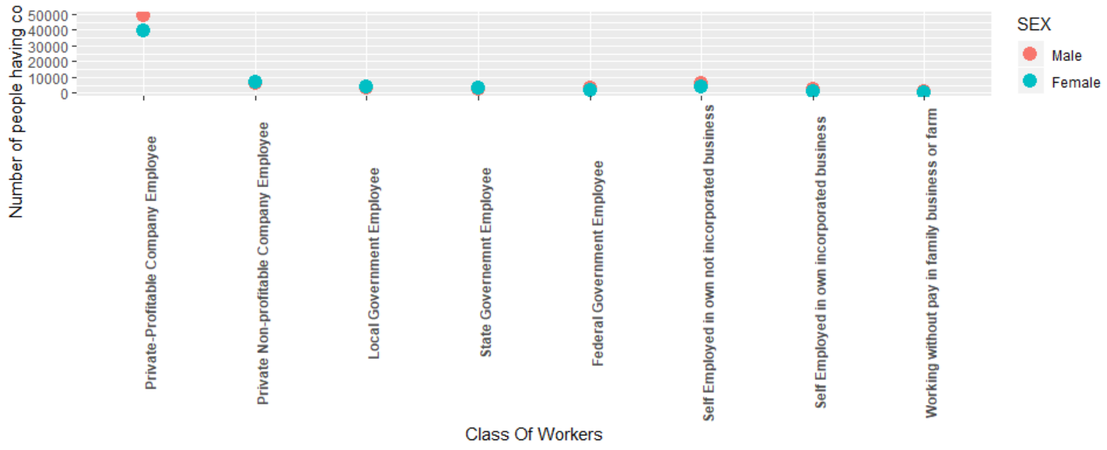
```
third_analysis
```



From my third analysis, we can see that the highest percentage of people with disabilities are working without pay in family business or farm. This verifies our second analysis as it also showed that these categor people have highest percentage of difficulties. People with least percentage of people having disabilities are the ones which work for the provate profitable company or are self employed in own incorporated business.

```
fourth_analysis <- psam_pus_final %>% select(COW,SEX,DREM)

fourth_analysis$SEX <- as.factor(fourth_analysis$SEX)
levels(fourth_analysis$SEX) <- c("Male","Female")
fourth_analysis %>%
  filter(!is.na(COW)) %>%
  group_by(COW,SEX) %>%
  filter(DREM==1) %>%
  summarise(count=n()) %>%
  ggplot(aes(x=COW,y=count))+
  geom_point(aes(color=SEX),size=4)+
  xlab("Class Of Workers")+
  ylab("Number of people having cognitive difficulty")+
  theme(axis.text.x = element_text(face="bold",angle=90))
```



In my fourth analysis, I focused on the cognitive difficulty as it was highe percentage of difficulty faced by people in almost every class of work. This shows how this difficulty varies according to the sex of a person. We can see form the graph that, among four class of work,number of males going through cognitive difficulties are more than females, specially in Private-profitable sector.And in other categories the number of males and females suffering from this difficulty is almost the same.

```
chisq.test(psam_pus_final$SEX,psam_pus_final$DREM,correct=FALSE)

##
##  Pearson's Chi-squared test
##
## data:  psam_pus_final$SEX and psam_pus_final$DREM
## X-squared = 612.19, df = 1, p-value < 2.2e-16
```

To test weather the two variables (SEX,DREM) sex and cognitive difficulty are independent of each other, chi square test is used. From the result we can see that p value is smaller than the 0.05, resulting in rejection of null hypothesis and accepting that the two variables are dependant.

```
cor_test <- cow_difficulties %>% select(COW,Percentage) %>%
inner_join(cow_disability,by="COW") %>% select(COW,Percentage.x,Percentage.y)
%>%
rename(Percentage.difficulties=Percentage.x,Percentage.disability=Percentage.
y)

cor(cor_test$Percentage.difficulties,cor_test$Percentage.disability)

## [1] 0.5495372
```

The above correlation test between percentage of people having disabilities and percentage of people having difficulties shows that there is a positive relation between them. It makes sense too as people with certain disabilities can cause them certain difficulties. But we can see that it's not extremely strong co-relation. That means that difficulties faced by people are not completely due to the disabilities of a person.

```
t.test(psam_pus_final$AGEP,alternative = "two.sided",mu=40)
```

```
##
##  One Sample t-test
##
## data:  psam_pus_final$AGEP
## t = 589.36, df = 4699112, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 40
## 95 percent confidence interval:
##  44.28209 44.31067
## sample estimates:
## mean of x
##  44.29638
```

The above one sample t test is applied on the table: cow_age having the average age of people working . It is used to check weather the average age of people equal 40. The result of the t test shows that the p value is less than 0.05 which means we reject null hypothesis that the average age is equal to 40.

Discussion:

So, after analysing some graphs and tests, we can say that the person's disabilty has caused them to face some difficulties but not all of them are due to their disabilities. The difficulties faced by the person varies from its class of work. People with more difficulties and disabilities are seen to be working without pay and in family businesses. People with less difficulties and disabilities are seen to be self employed in own incorporated businesses in the average age of 53.

We have tested the average age of people from our data, which is not equal to 40. We have also seen how sex of the person is related to the cognitive difficulty faced by a person. Males tend to have more difficulty in making decisions that affect their daily life.