Exploratory Data Analysis

### Load packages:

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

### Load data

work<-read.csv("NIOSH\_QWL (4).csv")  
work

\*\* ##PART1: Research Questions:

*Research Question 1:* Is there any impact on health of the respondents due to stress at work?

*Research Question 2:*

What is the impact on the financial satisfaction of the respondents depending on the number of children owned by them?

*Research Question 3:*

Americans are more satisfied from job when they are self-employed or working under someone else?

# PART 2: EDA

*Research Question 1:*

Is there any impact on health of the respondents due to stress at work?

In this analysis, I am interested to know how the busy schedule, pressure and stress at work affects the health of citizens of US. In this competitive world people often neglect the health in the race of achieving success in life. My prediction for this analysis is, more stress at work will lead to poor health conditions.

Variables taken are:

1. Stressr- Categorical variable which describes the stress level of the respondents at work.

2. health- Categorical variable which describes the health of the respondents.

3. sex- Categorical variable describing the gender of the respondents.

Here, I am removing N/A from all the 3 columns:

stress<-work%>%  
   
filter(!is.na(stressr) & !is.na(health) & !is.na(sex))

Numerical summaries of the data chosen for this analysis:

table(stress$stressr,stress$health,stress$sex)

## , , = FEMALE  
##   
##   
## EXCELLENT FAIR GOOD POOR  
## Always 108 101 188 21  
## Hardly ever 216 78 321 6  
## Never 114 31 118 6  
## Often 319 141 492 14  
## Sometimes 582 238 862 29  
##   
## , , = MALE  
##   
##   
## EXCELLENT FAIR GOOD POOR  
## Always 122 92 204 16  
## Hardly ever 227 89 286 6  
## Never 85 38 116 5  
## Often 298 151 471 18  
## Sometimes 570 224 870 25

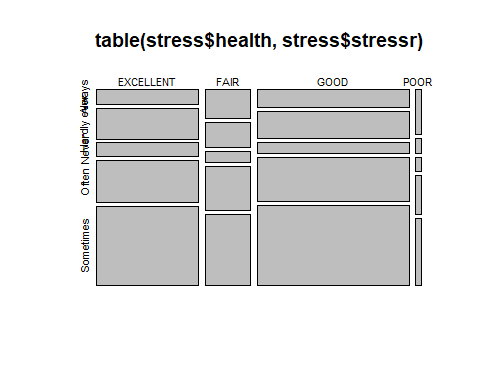
Proportion of the data used in the analysis:

prop.table(table(stress$stressr,stress$health,stress$sex))

## , , = FEMALE  
##   
##   
## EXCELLENT FAIR GOOD POOR  
## Always 0.0136743479 0.0127880476 0.0238034946 0.0026589010  
## Hardly ever 0.0273486959 0.0098759180 0.0406432008 0.0007596860  
## Never 0.0144340339 0.0039250443 0.0149404913 0.0007596860  
## Often 0.0403899721 0.0178526209 0.0622942517 0.0017726007  
## Sometimes 0.0736895417 0.0301342112 0.1091415548 0.0036718156  
##   
## , , = MALE  
##   
##   
## EXCELLENT FAIR GOOD POOR  
## Always 0.0154469486 0.0116485186 0.0258293239 0.0020258293  
## Hardly ever 0.0287414535 0.0112686756 0.0362116992 0.0007596860  
## Never 0.0107622183 0.0048113446 0.0146872626 0.0006330717  
## Often 0.0377310712 0.0191187642 0.0596353507 0.0022790580  
## Sometimes 0.0721701697 0.0283616105 0.1101544695 0.0031653583

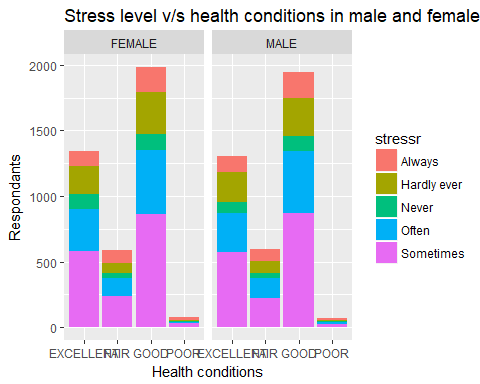
From the below mosaic graph, we can analyse the proportions of the variables better.

stress %>%  
 mutate(health = as.character(health)) -> stress  
   
  
plot(table(stress$health,stress$stressr))

 Interpretation: By looking at the poor health status column, we can see that "Always" stress level (Always stress at work) is highest when compared to people having excellent and good health. We can also see that proportion of "hardly ever" stress level is also least for people having poor health conditions. From this we can predict that people having poor health condition will mostly have stress at work.

Using another graph to see if men face more stress at work or women.

ggplot(stress,aes(x=health,fill=stressr))+facet\_wrap(~sex)+geom\_bar()+labs(x="Health conditions", y="Respondants")+ggtitle("Stress level v/s health conditions in male and female")



Interpretation: From the above graph we can see that graph for female and male are almost same. Female and male face the same amount of stress at work.

*Research Question 2:* What is the impact on the financial satisfaction of the respondents depending on the number of children owned by them?

In this analysis, I am interested in knowing what is the financial satisfaction level of the respondents depending on the size of family. My prediction is that when the family size increases then the financial satisfaction decreases.

Variables used are:

1.satfin - how satisfied are respondents with their financial situation.

2.childs - total number of children owned by the respondent.

3.race - race of the respondent.

Numerical summaries of the analysis:

table(work$childs,work$satfin,work$race)

## , , = BLACK  
##   
##   
## MORE OR LESS NOT AT ALL SAT SATISFIED  
## 0 518 349 228  
## 1 397 343 143  
## 2 435 376 167  
## 3 281 275 120  
## 4 149 146 68  
## 5 71 80 50  
## 6 45 40 24  
## 7 23 29 13  
## EIGHT OR MORE 38 34 17  
##   
## , , = OTHER  
##   
##   
## MORE OR LESS NOT AT ALL SAT SATISFIED  
## 0 293 153 122  
## 1 130 80 74  
## 2 246 109 104  
## 3 118 68 61  
## 4 67 40 23  
## 5 24 15 10  
## 6 12 8 4  
## 7 6 4 4  
## EIGHT OR MORE 6 11 1  
##   
## , , = WHITE  
##   
##   
## MORE OR LESS NOT AT ALL SAT SATISFIED  
## 0 4099 2018 2649  
## 1 2115 1136 1131  
## 2 3366 1633 2155  
## 3 1845 959 1156  
## 4 867 439 533  
## 5 332 198 215  
## 6 160 91 90  
## 7 70 37 47  
## EIGHT OR MORE 68 47 43

Proportions of the variables used in the analysis:

prop.table(table(work$childs,work$satfin,work$race))

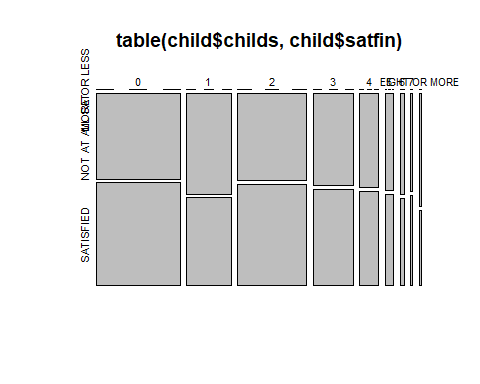
## , , = BLACK  
##   
##   
## MORE OR LESS NOT AT ALL SAT SATISFIED  
## 0 1.534769e-02 1.034043e-02 6.755355e-03  
## 1 1.176261e-02 1.016266e-02 4.236911e-03  
## 2 1.288851e-02 1.114041e-02 4.948002e-03  
## 3 8.325679e-03 8.147907e-03 3.555450e-03  
## 4 4.414684e-03 4.325798e-03 2.014755e-03  
## 5 2.103641e-03 2.370300e-03 1.481438e-03  
## 6 1.333294e-03 1.185150e-03 7.110900e-04  
## 7 6.814613e-04 8.592338e-04 3.851738e-04  
## EIGHT OR MORE 1.125893e-03 1.007378e-03 5.036888e-04  
##   
## , , = OTHER  
##   
##   
## MORE OR LESS NOT AT ALL SAT SATISFIED  
## 0 8.681224e-03 4.533199e-03 3.614708e-03  
## 1 3.851738e-03 2.370300e-03 2.192528e-03  
## 2 7.288673e-03 3.229534e-03 3.081390e-03  
## 3 3.496193e-03 2.014755e-03 1.807354e-03  
## 4 1.985126e-03 1.185150e-03 6.814613e-04  
## 5 7.110900e-04 4.444313e-04 2.962875e-04  
## 6 3.555450e-04 2.370300e-04 1.185150e-04  
## 7 1.777725e-04 1.185150e-04 1.185150e-04  
## EIGHT OR MORE 1.777725e-04 3.259163e-04 2.962875e-05  
##   
## , , = WHITE  
##   
##   
## MORE OR LESS NOT AT ALL SAT SATISFIED  
## 0 1.214483e-01 5.979082e-02 7.848656e-02  
## 1 6.266481e-02 3.365826e-02 3.351012e-02  
## 2 9.973038e-02 4.838375e-02 6.384996e-02  
## 3 5.466505e-02 2.841397e-02 3.425084e-02  
## 4 2.568813e-02 1.300702e-02 1.579212e-02  
## 5 9.836746e-03 5.866493e-03 6.370182e-03  
## 6 4.740600e-03 2.696216e-03 2.666588e-03  
## 7 2.074013e-03 1.096264e-03 1.392551e-03  
## EIGHT OR MORE 2.014755e-03 1.392551e-03 1.274036e-03

Here, I am removing all the entries with NA values for childs and satfin variable and also I am interested in only "satisfied " and "not at all satisfied" categories so removing "more or less" category in satfin variable.

child<-work%>%  
   
filter(!is.na(childs) & !is.na(satfin) & satfin!="MORE OR LESS" )

Below masaic graph best helps in understanding relationship between satfin and childs variables.

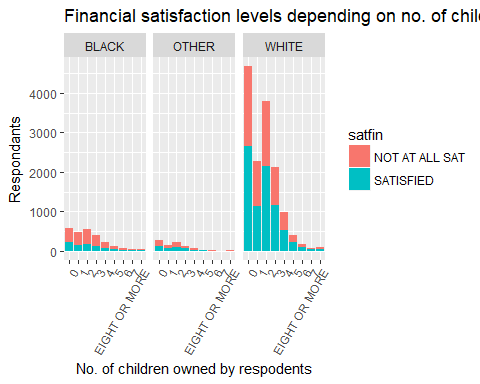
child %>%  
 mutate(childs = as.character(childs)) -> child  
  
plot(table(child$childs,child$satfin))

 Interpretation: 1.Initially, the proportions of satisfied and not at all satisfied categories are quite equal for the respondents having 5 children.

2. But when the number of children increases (8 or more) then the proportion of "not at all satisfied"" is more than "satisfied" category. From this we can predict that respondents are not financially satisfied when they have more number of kids.

In this below graph we can analyze how satisfaction with the financial situations depending on the number of kids owned by the kid differ between different race.

ggplot(data=child,aes(childs,fill=satfin))+facet\_wrap(~race)+geom\_bar()+theme(axis.text.x = element\_text(angle=60,hjust = 1))+labs(x="No. of children owned by respodents", y="Respondants")+ggtitle("Financial satisfaction levels depending on no. of children among different race")

 Interpretation: 1.From the first block we can interpret that, among black race, respondents are mostly not satisfied at all no matter how many children they own. 2. While when we look at the third block, most Americans are satisfied if they have up to 4 kids and above 5 kids the proportions are almost equal between satisfied and not at all satisfied categories.

*Research Question 3:* Americans are more satisfied from job when they are self-employed or working under someone?

It is often better to be the boss rather than working under someone for the sake of job satisfaction. So, in this analysis I am interested in knowing whether Americans are more satisfied with job when they are self-employed or when they are working under someone else.

Variables used are:

1. wrkslf : Categorical variable represents if the respondent is a self-employed or working for someone else.

2. satjob : Categorical variable representing how satisfied are the respondents with their job.

3. wrkstat : categorical variable represents whether the respondent is working part-time or full-time.

Below table gives the numerical summaries of the variables chosen:

table(work$wrkslf,work$satjob,work$wrkstat)

## , , = TEMP NOT WORKING  
##   
##   
## A LITTLE DISSAT MOD. SATISFIED VERY DISSATISFIED  
## SELF-EMPLOYED 14 54 4  
## SOMEONE ELSE 96 343 46  
##   
## VERY SATISFIED  
## SELF-EMPLOYED 108  
## SOMEONE ELSE 444  
##   
## , , = WORKING FULLTIME  
##   
##   
## A LITTLE DISSAT MOD. SATISFIED VERY DISSATISFIED  
## SELF-EMPLOYED 167 913 49  
## SOMEONE ELSE 2327 9504 821  
##   
## VERY SATISFIED  
## SELF-EMPLOYED 2107  
## SOMEONE ELSE 11255  
##   
## , , = WORKING PARTTIME  
##   
##   
## A LITTLE DISSAT MOD. SATISFIED VERY DISSATISFIED  
## SELF-EMPLOYED 70 370 33  
## SOMEONE ELSE 515 1831 239  
##   
## VERY SATISFIED  
## SELF-EMPLOYED 625  
## SOMEONE ELSE 1920

Below table shows the proportions of the variables chosen:

prop.table(table(work$wrkslf,work$satjob,work$wrkstat))

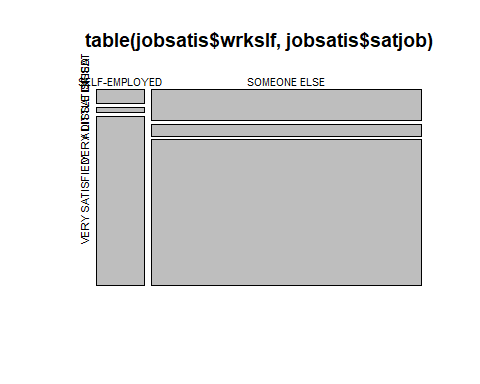
## , , = TEMP NOT WORKING  
##   
##   
## A LITTLE DISSAT MOD. SATISFIED VERY DISSATISFIED  
## SELF-EMPLOYED 0.0004135283 0.0015950377 0.0001181509  
## SOMEONE ELSE 0.0028356225 0.0101314429 0.0013587358  
##   
## VERY SATISFIED  
## SELF-EMPLOYED 0.0031900753  
## SOMEONE ELSE 0.0131147541  
##   
## , , = WORKING FULLTIME  
##   
##   
## A LITTLE DISSAT MOD. SATISFIED VERY DISSATISFIED  
## SELF-EMPLOYED 0.0049328017 0.0269679516 0.0014473490  
## SOMEONE ELSE 0.0687343081 0.2807266283 0.0242504800  
##   
## VERY SATISFIED  
## SELF-EMPLOYED 0.0622360065  
## SOMEONE ELSE 0.3324472013  
##   
## , , = WORKING PARTTIME  
##   
##   
## A LITTLE DISSAT MOD. SATISFIED VERY DISSATISFIED  
## SELF-EMPLOYED 0.0020676414 0.0109289617 0.0009747452  
## SOMEONE ELSE 0.0152119332 0.0540835918 0.0070595185  
##   
## VERY SATISFIED  
## SELF-EMPLOYED 0.0184610840  
## SOMEONE ELSE 0.0567124502

Here, I am removing the N/A data from the wrkslf and satjob variables and also removing the "MOD.SATISFIED" category in satjob variable and "TEMP NOT WORKING" category from wrkstat variable because we are only interested in satisfied and not satisfied categories.

jobsatis<-work%>%  
   
filter(!is.na(wrkslf) & !is.na(satjob) & !is.na(wrkstat) & satjob!="MOD. SATISFIED" & wrkstat!="TEMP NOT WORKING" )

Below mosaic plot represents the proportions of the categories in the variables satjob and wrkslf.

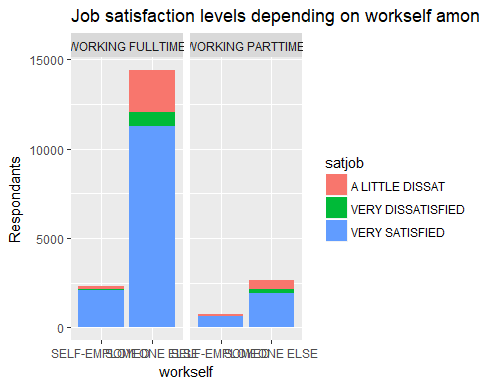
jobsatis%>%  
 mutate(satjob = as.character(satjob)) -> jobsatis  
   
  
plot(table(jobsatis$wrkslf,jobsatis$satjob))



Interpretation: 1. From the above graph we can see that people who are self-employed, the ratio of satisfaction is more than dissatisfaction ("very dissatisfied" and "a little dissatisfied"). 2. When we compare the dissatisfaction level among the people who are self-employed and working for someone else, then the people who are working for someone are more dissatisfied.

From the below bar graph I am interested in knowing the how satisfied are people with their jobs when they are either self-employed or working for someone else while doing parttime or fulltime job(Work status).

ggplot(jobsatis, aes(x=wrkslf,fill=satjob))+facet\_wrap(~wrkstat) +geom\_bar()+labs(x="workself", y="Respondants")+ggtitle("Job satisfaction levels depending on workself among different work status")



Intrepretation:

1. From the above graph we can interpret that people are more satisfied when they are self-employed first of all.
2. Second thing to notice is that people who are self-employed and doing part-time job are more satisfied.