Exploring the BRFSS data

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(tidyr)
library(stringr)
library(knitr)
library(kableExtra)
require(gridExtra)
```

Load data

```
load("brfss2013.RData")
```

Part 1: Data

The data provided by BRFSS of its ongoing surveillance system designed to measure behavioral risk factors for non-institutionalized adult population residing in the US is an example of **retrospective sampling model**.

Data is being collected from every state by dialing a randomly selected phone number which ensures **random sampling**. However the individual being approached has the power to decline to participate in the study thus introducing **voluntary response bias** into the data.

Moreover this being an observational study, the results can only be generalized & only association can be established between various parameters being studied.

Part 2: Research questions

Research question 1: Healthcare expenses have skyrocketed over the past decade and are beyond the reach of people in the low-income bracket. With little to no Healthcare cover, people are at risk of going bankrupt when faced with unexpected health expenses like that being witnessed during the current pandemic.

Question 1.1: Does being in the low-income bracket increase one's probability of being at-risk in relation to unexpected healthcare expenses.

What percent of the total population faces this risk is an important parameter that the government needs to keep an eye on.

Research question 2:

Being physically active is crucial for ones health. It is a general belief that the participation of an individual in physical activities goes down with ones age. But when evaluated within the people who regularly do some form of workout, does the association hold true? That is,

Question 2.1: Within the physically active group, does the amount of physical activity done depend on a person's age?

Question 2.2: Also, does the activity of choice bear an association with their age?

This information would be useful for new Gyms and Health Clubs in deciding the infrastructure (equipment type and volume, requirement of water spa or physiotherapist etc.) of their facility.

Research question 3: As per the website, Diabetes.co.uk (https://www.diabetes.co.uk/Diabetes-Risk-factors.html): obesity, living a sedentary lifestyle, unhealthy eating, high blood pressure, high cholesterol and ageing are the major factors that increase the chances of a person developing Diabetic condition.

Of the above-mentioned factors, we shall check for correlation between obesity, ageing and diabetes. This can be broken into 2 parts:

Question 3.1: Are obese people more prone to being diabetic.

Question 3.2: Does one's age increase the risk of developing diabetes.

This information can be used by the health agencies in deciding the minimum age after which individuals should be encouraged to get themselves screened periodically.

Part 3: Exploratory data analysis

Research question 1:

Healthcare expenses have skyrocketed over the past decade and are beyond the reach of people in the low-income bracket. With little to no Healthcare cover, people are at risk of going bankrupt when faced with unexpected health expenses like that being witnessed during the current pandemic.

Question 1.1: Does being in the low-income bracket increase one's probability of being at-risk in relation to unexpected healthcare expenses.

The underlying task in this analysis is to first arrive at the **"at-risk"** population which is the population susceptible to bankruptcy when faced with unexpected healthcare expense. This can be derived from two parameters:

1. **Availability of Healthcare coverage:** Having a healthcover in any form helps an individual deal with major chunk of unexpected medical expenses.
2. **Financially stability:** relates to the ability to cover for additional expenses not covered by healthcover.This would include ability to pay for medication, doctor fees e.g. dental, surgery, & timely payment of other related medical expenses.

Based on these aspects, the **"At-risk" classification table** is defined as described below:

```
# Creating a classification table

At_Risk_Table <- matrix(c("Low", "Moderate", "Moderate", "High"),ncol=2,byrow=TRUE)
colnames(At_Risk_Table) <- c("Financially stable","Financially unstable")
rownames(At_Risk_Table) <- c("With Healthcare coverage", "Without Healthcare coverage")
At_Risk_Table <- as.table(At_Risk_Table)

kable(At_Risk_Table) %>%
  add_header_above(c("AT-RISK CLASSIFICATION TABLE"=3))%>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

AT-RISK CLASSIFICATION TABLE

	Financially stable	**Financially unstable**
With Healthcare coverage	Low	Moderate
Without Healthcare coverage	Moderate	High

VARIABLES USED IN ANALYSIS:

A) DIRECT VARIABLES FROM BRFSS2013 SURVEY DATABASE

1. `hlthpln1`: indicates if one has any kind of healthcare coverage, including health insurance, prepaid plans such as HMOs, government plans such as Medicare, or Indian Health Service
2. `medcost`: indicates if one needed to see a doctor but could not, because of cost, in the past 12 months
3. `medscost`: indicates if there has been a time in the past 12 months when one did not take medication as prescribed because of cost
4. `medbills`: indicates ones inability to pay-off any overdue medical bills
5. `income2`: Annual household income from all sources

B) NEW VARIABLES CALCULATED USING EXISTING VARIABLES

1. `fin_stab` : indicates the Financial stability of an individual
2. `at-risk` : classifies the population based on the **"AT-RISK" CLASSIFICATION TABLE**

First, the direct variables of interest are selected from the `brfss2013` database:

```
# Selecting variables of interest

df1<-brfss2013 %>% select(hlthpln1, medcost, medscost, medbills, income2)
summary (df1)
```

```
##  hlthpln1      medcost                                medscost
##  Yes :434571   Yes : 60107   Yes                          : 28347
##  No  : 55300   No  :430447   No                           :292775
##  NA's:  1904   NA's:  1221   No medication was prescribed: 28028
##                              NA's                         :142625
##
##
##
##  medbills                  income2
##  Yes : 63896   $75,000 or more  :115902
##  No  :283582   Less than $75,000: 65231
##  NA's:144297   Less than $50,000: 61509
##                Less than $35,000: 48867
##                Less than $25,000: 41732
##                (Other)          : 87108
##                NA's             : 71426
```

Using these direct variables, the two new variables of interest, `fin_stab` , & `at-risk` are derived.

`fin_stab` is calculated based on the following criteria:

`fin_stab` = **"Yes"**, An Individual is identified as financially stable if his reading for `medcost` , `medscost` & `medbills` is "No"

`fin_stab` = **"No"**, the Individual is considered to be at financial risk if his reading for any one or more of the variables `medcost` , `medscost` & `medbills` is "Yes"

```
# Creating new variables to define 'Financial Stability' of an individual
df1<-df1 %>%
  mutate(fs_1 = ifelse(medcost=="No" & medbills=="No" & is.na(medscost), "Yes", ifelse(medcost==
         "No"& medscost=="No"& is.na(medbills),"Yes",ifelse(medscost=="No" & medbills=="No"& is.
         na(medcost), "Yes", "No"))))

df1<-df1 %>%
  mutate(fs_2 = ifelse(medcost=="No" & is.na(medscost) & is.na(medbills), "Yes", "No" ))

df1<-df1 %>%
  mutate(fs_3 = ifelse(medbills=="No" & is.na(medscost) & is.na(medcost), "Yes", ifelse(medscost
         =="No" & is.na(medcost) & is.na(medbills), "Yes",ifelse(medscost=="No" & medcost=="No"
         & medbills=="No", "Yes","No" ))))

# Converting cells with NA's to 0
df1$fs_1[which(is.na(df1$fs_1))]<- 0
df1$fs_2[which(is.na(df1$fs_2))]<- 0
df1$fs_3[which(is.na(df1$fs_3))]<- 0

df1<-df1 %>%
  mutate(fin_stab = ifelse(fs_1=="Yes" | fs_2=="Yes" | fs_3=="Yes", "Yes", ifelse(fs_1=="0" & fs
         _2=="0" & fs_3 == "0", "NA", "No")))

df1%>%
  group_by(fin_stab)%>%
  summarise(count=n())
```

```
## # A tibble: 3 x 2
##   fin_stab  count
## * <chr>     <int>
## 1 NA          376
## 2 No       134313
## 3 Yes      357086
```

Next we derive the `at-risk` variable based on the **"AT-RISK" CLASSIFICATION TABLE**.

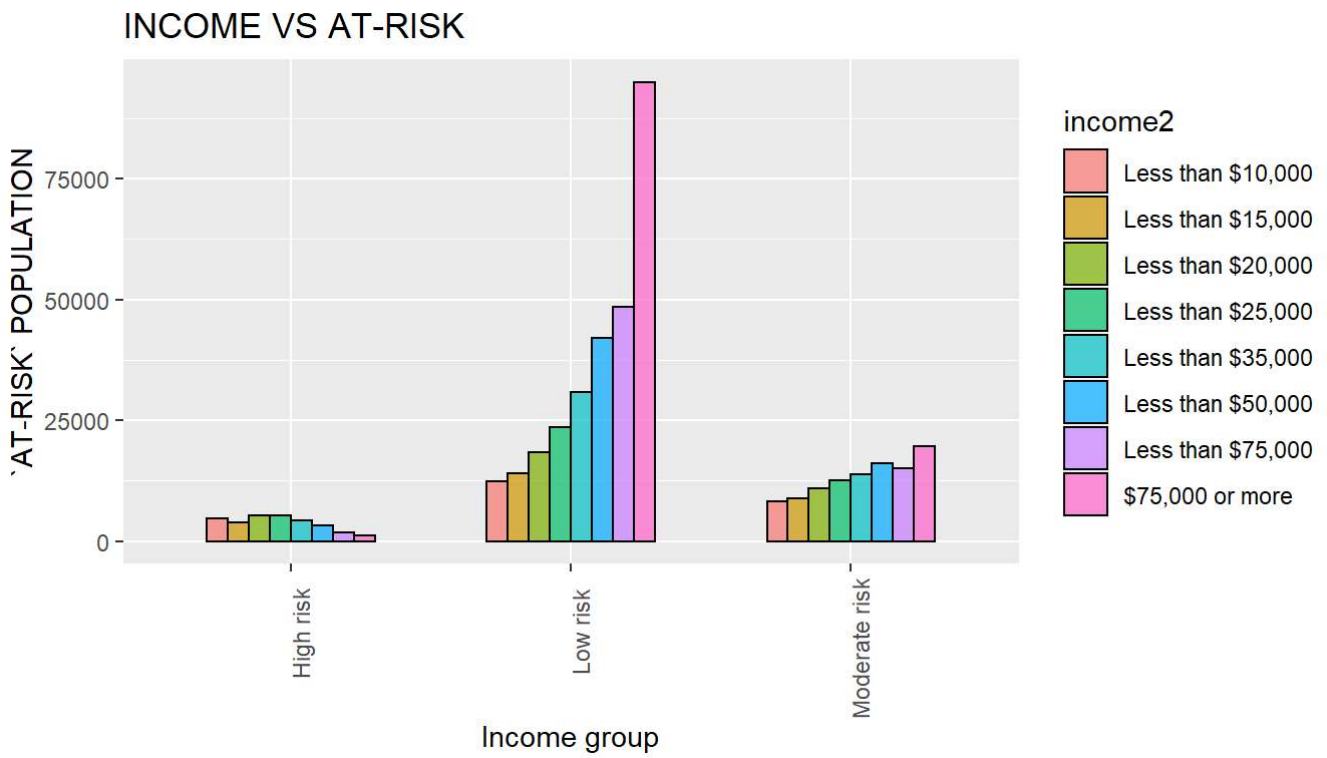
```
# Creating new variable to define 'At-risk' individual
df1<-df1 %>%
mutate(at_risk = ifelse( hlthpln1 =="Yes" & fin_stab =="Yes", "Low risk",ifelse(hlthpln1 == "Ye
         s" & fin_stab =="No", "Moderate risk", ifelse(hlthpln1 == "No" & fin_stab =="Yes", "Mod
         erate risk", ifelse(hlthpln1=="No" & fin_stab == "No", "High risk", "NA")))))

df1%>%
  group_by(at_risk)%>%
  summarise(count=n())
```

```
## # A tibble: 5 x 2
##   at_risk        count
## * <chr>          <int>
## 1 High risk      33848
## 2 Low risk      334441
## 3 Moderate risk 121233
## 4 NA               376
## 5 <NA>            1877
```

```
# Deleting NA cells from the data set
df1<-df1%>%
  filter(!is.na(at_risk),at_risk!="NA", !is.na(income2))
```

Side-by-side bar plots are then plotted to examine the association between variables.

```
 # Creating Bar plot: Income vs at-risk

ggplot(df1, aes(x=at_risk, fill = income2)) +
  geom_bar(position="dodge", width = 0.6, color="black", size=0.5, alpha = 0.7) + theme(legend.p
         osition = "right", axis.text.x = element_text(angle=90, hjust = 0.8))+ labs(x="Income g
         roup", y="`AT-RISK` POPULATION", title="INCOME VS AT-RISK")
```



Summary 1.1: It can be seen from the plot that the number of people in the **High-risk** category decrease, while an opposite trend in visible for people in the **Low-risk** category.

Thus, the plot does suggest that there is a possible assosciation between one's income group & his/her ability to bear unexpected healthcare expenses.

Research question 2:

Being physically active helps in maintenance of ones health. It is generally believed that the younger generation is much more active than the older generations. But when evaluated within the people who regularly do some form of workout, does this association hold true? That is,

Question 2.1: Within the physically active group, does the amount of physical activity done depend on a person's age?

Question 2.2: Also, does the activity of choice bear an association with their age?

VARIABLES USED IN ANALYSIS:

A) DIRECT VARIABLES FROM BRFSS2013 SURVEY DATABASE

1. `X_age_g`: indicates the age-group of the individual
2. `X_totinda`: indicates if one has or hasn't performed any kind of physical activity in the last 30 days
3. `exract11` & `exract21`: indicates the top 2 types of physical activities or exercises done by an individual during the past month
4. `exeroft1` & `exeroft2`: indicates how many days per week or per month did one take part in the activity during the past month?
5. `padur1_` & `padur2_`: indicates how many minutes or hours did one work out.

B) NEW VARIABLES CALCULATED USING EXISTING VARIABLES

1. `total_mins` : indicates the cumulative value of the weekly minutes spent in performing the top two types of physical activities.

First, the direct variables of interest are selected from the `brfss2013` database: database:

```
# Selecting variables of interest

df2<-brfss2013%>%
  select(X_age_g, X_totinda,exract11, exeroft1, padur1_,exract21, exeroft2, padur2_)

df2<-df2%>%
  filter(!is.na(X_totinda), X_totinda=="Had physical activity or exercise", !is.na(X_age_g))

summary(df2)
```

```
##             X_age_g
##  Age 18 to 24   : 20118
##  Age 25 to 34   : 36200
##  Age 35 to 44   : 42699
##  Age 45 to 54   : 57734
##  Age 55 to 64   : 74016
##  Age 65 or older:101693
##
##                                              X_totinda
##  Had physical activity or exercise               :332460
##  No physical activity or exercise in last 30 days:     0
##
##
##
##
##
##                                           exract11         exeroft1
##  Walking                                     :180049   Min.   :101.0
##  Running                                     : 23152   1st Qu.:103.0
##  Gardening (spading, weeding, digging, filling): 20024   Median :105.0
##  Other                                       : 14119   Mean   :135.8
##  Weight lifting                              : 10226   3rd Qu.:203.0
##  (Other)                                     : 83250   Max.   :299.0
##  NA's                                        :  1640   NA's   :4857
##     padur1_                                              exract21
##  Min.   :  1.00   No other activity                          :109524
##  1st Qu.: 30.00   Walking                                    : 44868
##  Median : 45.00   Weight lifting                             : 18414
##  Mean   : 63.34   Other                                      : 17959
##  3rd Qu.: 60.00   Gardening (spading, weeding, digging, filling): 16111
##  Max.   :599.00   (Other)                                    :118853
##  NA's   :9085     NA's                                       :  6731
##     exeroft2         padur2_
##  Min.   :  2.0    Min.   :  0.00
##  1st Qu.:102.0    1st Qu.: 30.00
##  Median :104.0    Median : 45.00
##  Mean   :139.2    Mean   : 71.97
##  3rd Qu.:203.0    3rd Qu.: 90.00
##  Max.   :299.0    Max.   :599.00
##  NA's   :118923   NA's   :124341
```

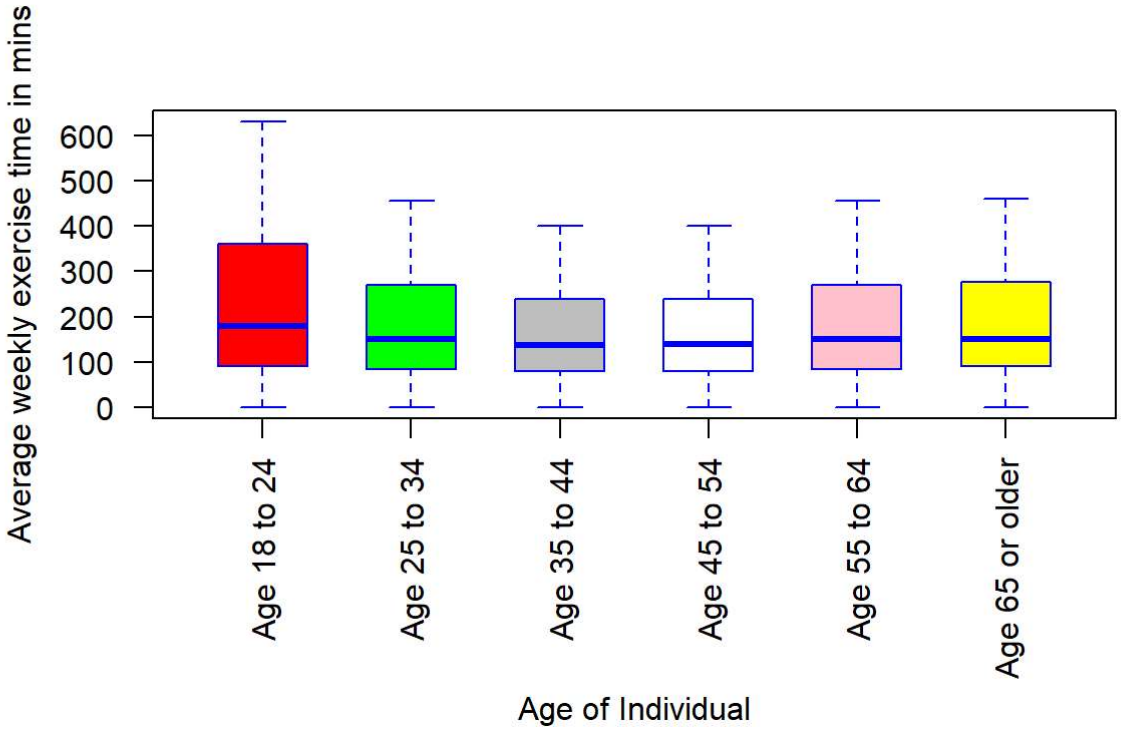
Next, the total weekly minutes are calculated by:

1. Standardizing the frequency of exercise, `exeroft1` & `exeroft2` , to activities/week.
2. Converting duration of exercise, `padur1_` & `padur2_` from hours to minutes, where necessary.
3. Combining the average time spent per week doing both the form of activities into a single variable `total_mins`

```
# Calculating total exercise time

df2 <- df2 %>%
mutate(days_w1 = ifelse(exeroft1<200,(exeroft1%%100),round({(exeroft1%%100)*12/52},1)))

df2 <- df2 %>%
mutate(days_w2 = ifelse(exeroft2<200,(exeroft2%%100),round({(exeroft2%%100)*12/52},1)))

df2 <- df2 %>%
mutate(total_mins1 = df2$days_w1*df2$padur1_)

df2 <- df2 %>%
mutate(total_mins2 = df2$days_w2*df2$padur2_)

df2 <- df2 %>%
  mutate(total_mins2 = ifelse(!is.na(total_mins2), df2$days_w2*df2$padur2_,0))

df2 <- df2 %>%
mutate(total_mins = df2$total_mins1 + df2$total_mins2)
```

Now that we have the `total_mins` variable, we can compare the amount of exercise performed by each age-group using a box plot:

```
# Creating Box-Plot - Age Vs minutes
myColors <- ifelse(levels(df2$X_age_g) =="Age 18 to 24", "Red",ifelse(levels(df2$X_age_g) =="Age
         25 to 34", "Green",ifelse(levels(df2$X_age_g) =="Age 35 to 44", "Gray",ifelse(levels(df
         2$X_age_g) =="Age 45 to 54", "White",ifelse(levels(df2$X_age_g) =="Age 55 to 64", "Pin
         k","Yellow")))))

boxplot(df2$total_mins1~df2$X_age_g, col= myColors, range=1.0, varwidth=FALSE, notch=FALSE, outl
         ine=FALSE, boxwex=0.6, border=c("blue"), xlab = "", ylab = "Average weekly exercise tim
         e in mins  ",las = 2, pars=list(par(mar=c(8,8,4,2))))
mtext("Age of Individual", side=1, line =7)
```



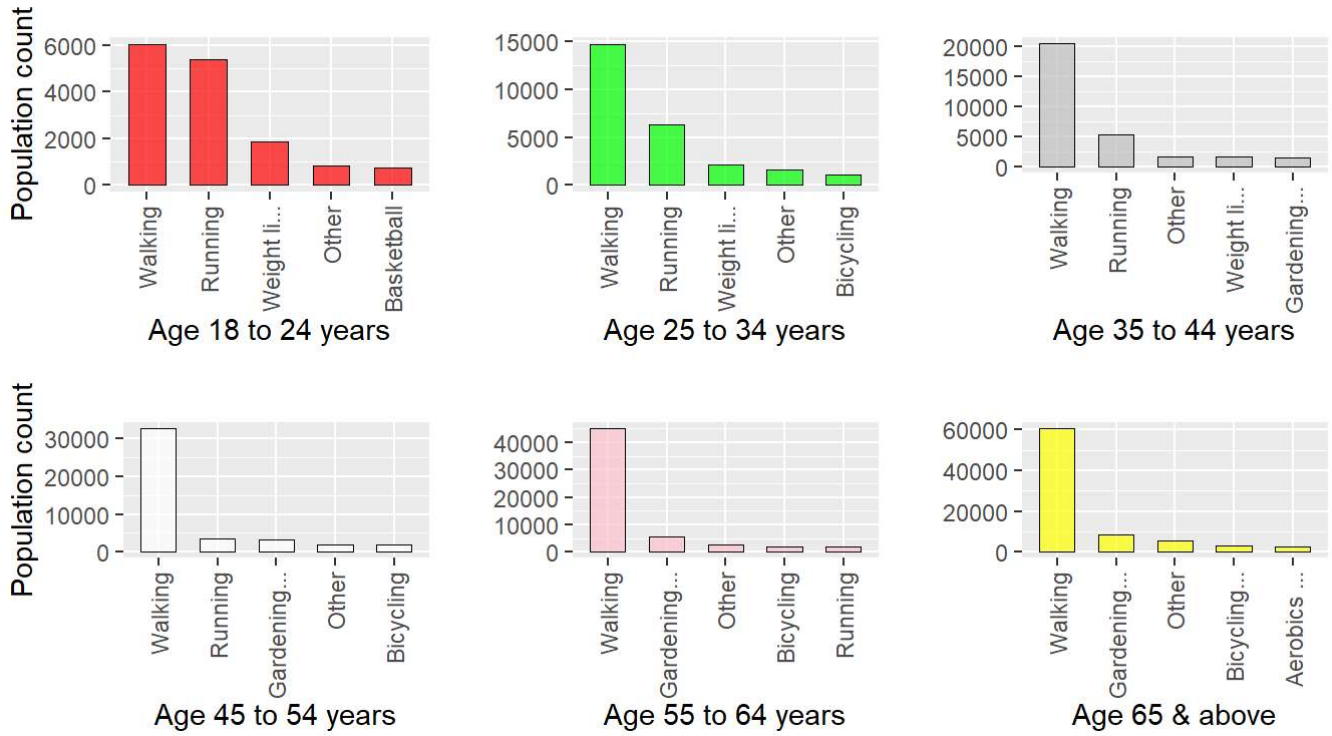
Summary 2.1:
From the Box plot following seem to be evident,

a. There is a correlation in the amount of time invested exercising and age; however, this correlation is not a linear one.

b. The trend appears to be highest for the youngest age group, then dips for middle age group (35-54 years) and then picks up with age.

```
# Creating side-by-side plot for top 5 activities per age group

# Code for Age group 18-24 years

v_1<-df2[df2$X_age_g=="Age 18 to 24",,]

v_1<-v_1$exract11

v_1.freq<-table(v_1)

v_1.freq<-sort(v_1.freq, decreasing = TRUE)

v_1.freq<-v_1.freq[1:5]

v_1.df<-as.data.frame(v_1.freq)

p1<-ggplot(v_1.df, aes(v_1, Freq)) + geom_bar(stat = "identity", width = 0.6, color="black", fil
         l="red", size=0.1, alpha = 0.7) + theme(legend.position = "none", axis.text.x = element
         _text(angle=90, hjust = 0.8))+ labs(x="Age 18 to 24 years", y="Population count", title
         ="") + scale_x_discrete(label = function(x) stringr::str_trunc(x, 12)) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```

```
# Creating Side-by-Side bar plots

grid.arrange(p1,p2,p3,p4,p5,p6, nrow=2, ncol=3, layout_matrix = rbind(c(1,2,3), c(4,5,6)))
```



Summary 2.2: From the side-by-side plot age does appear to have an influence on the choice of physical activity, with the exception of Walking. Running one could deduce is second most popular form of exercise up to 54 years of age; thereafter gardening appears to be more popular.

Research question 3:

As per the website, Diabetes.co.uk (https://www.diabetes.co.uk/Diabetes-Risk-factors.html): obesity, living a sedentary lifestyle, unhealthy eating, high blood pressure, high cholesterol and ageing are the major factors that increase the chances of a person developing Diabetic condition.

Of the above-mentioned factors, we shall check for correlation between obesity, ageing and diabetes. This can be broken into 2 parts:

Question 3.1: Are obese people more prone to being diabetic.

Question 3.2: Does one's age increase the risk of developing diabetes.

To answer this research question, we would:

P1. First explore the data & identify the variables of interest.

P2. Analyse the data to check the correlation between obesity & age with diabetes.

P1: Data study & conditioning:

Obesity (https://www.medicinenet.com/script/main/art.asp?articlekey=11760) has been defined by the National Institutes of Health (the NIH) as a BMI (Body Mass Index) of 30 and above. (A BMI of 30 is about 30 pounds overweight.) The BMI, a key index for relating body weight to height, is a person's weight in kilograms (kg) divided by their height in meters (m) squared.

In order to conclude whether a person is obese or not, we shall first create a new variable, **BMI** and then create a second variable, **Obese** to categorize them appropriately.

Following is the list of variables used in this analysis:

DIRECT VARIABLES FROM BRFSS2013 SURVEY DATABASE

weight2: gives weight of the person in pounds or kilograms

height3: gives height of the person in ft-inches, meters or centimeters

diabete3: provides information regarding the person's diabetic condition as "Yes"; "Yes, but female told only during pergnancy", No, prediabetic or borderline diabetic","No"

X_ageg5yr: gives age of the person in years

NEW VARIABLES CALCULATED USING EXISTING VARIABLES

ht_mtrs: gives height of the person in meters

wt_kgs: gives weight of the person in kilograms (kgs)

BMI: gives BMI of the person based on his height & weight

diabetic: classifies the person's diabetic condition as "Yes" or "No"

obese: classifies the person's obesity condition based on his BMI as "Yes" or "No"

First, the variables of interest are selected from the `brfss2013` database and cleaned to remove the cells with "NA":

```
# Selecting variables of interest

df3<-brfss2013[,c("weight2", "height3", "X_ageg5yr", "diabete3")]
df3 <- df3 %>%
filter(!is.na(weight2), !is.na(height3), !is.na(diabete3) , !is.na(X_ageg5yr))

summary (df3)
```

```
##     weight2          height3                 X_ageg5yr
##  180    : 21997   Min.   : 206.0   Age 60 to 64   : 53036
##  200    : 21808   1st Qu.: 504.0   Age 55 to 59   : 51878
##  150    : 21309   Median : 506.0   Age 65 to 69   : 49419
##  160    : 21083   Mean   : 551.1   Age 50 to 54   : 46509
##  170    : 18447   3rd Qu.: 510.0   Age 80 or older: 39605
##  140    : 16397   Max.   :9509.0   Age 70 to 74   : 39504
##  (Other):358580                    (Other)        :199670
##                                        diabete3
##  Yes                                       : 60925
##  Yes, but female told only during pregnancy:  4422
##  No                                        :405828
##  No, pre-diabetes or borderline diabetes   :  8446
##
##
##
```

From the summary table, it is evident that:

1. The `weight2` variable is a categorical variable which needs to be converted into a numeric variable.
2. The `diabete3` variable needs to be reduced from the existing 4-levels to 2-levels. (new variable: `diabetic`)

```
# Converting weight variable to a numeric variable

df3<-df3 %>%
filter(str_detect(weight2,"1")|str_detect(weight2,"2")|str_detect(weight2,"3")|str_detect(weight
         2,"4")|str_detect(weight2,"5")|str_detect(weight2,"6")|str_detect(weight2,"7")|str_dete
         ct(weight2,"8")|str_detect(weight2,"9")|str_detect(weight2,"0"))

df3$weight2<-df3$weight2%>%
as.numeric(as.character(df3$weight2))

df3<-df3%>%
mutate(diabetic = ifelse(diabete3 =="Yes","Yes","No"))

str(df3)
```

```
## 'data.frame':    464553 obs. of  5 variables:
##  $ weight2  : num  154 30 63 31 169 128 9 139 73 128 ...
##  $ height3  : int  507 510 504 504 600 503 500 602 505 601 ...
##  $ X_ageg5yr: Factor w/ 13 levels "Age 18 to 24",..: 9 7 8 9 10 6 4 7 10 5 ...
##  $ diabete3 : Factor w/ 4 levels "Yes","Yes, but female told only during pregnancy",..: 3 3 3
3 3 3 3 3 3 3 ...
##  $ diabetic : chr  "No" "No" "No" "No" ...
```

Next we convert:

1. `weight2`, currently available in pounds or kgs into a new variable, `wt_kgs` which gives the weight in kgs.
2. `height3`, currently available in ft-inches or meters into a new variable, `ht_mtrs` which gives the height in meters.

```
# Converting weights to kgs & heights to meters

df3 <- df3 %>%
mutate(wt_kgs = round((weight2/2.2),1))

df3 <- df3 %>%
mutate(ht_mtrs = ifelse(height3 < 1000,round({(height3%/%100)*0.3048+(height3%%100)*.0254},3),ro
         und({(height3%%1000)/100},3)))
```

Using the variables `wt_kgs` & `ht_mtrs`, we calculate the `BMI` for each reading and review the output data.

```
# Calculating BMI

df3 <- df3 %>%
mutate(BMI = round((wt_kgs/ht_mtrs^2),1))
```

Next we select the valid range of BMI values based on the data provided in CDC's,Anthropometric reference data for children and adults; United States, 2011-2014 (https://stacks.cdc.gov/view/cdc/40572). Based on this report we conclude that the valid range for `BMI` is (19.3, 50) & all values out of this range might be due to an error in data entry.

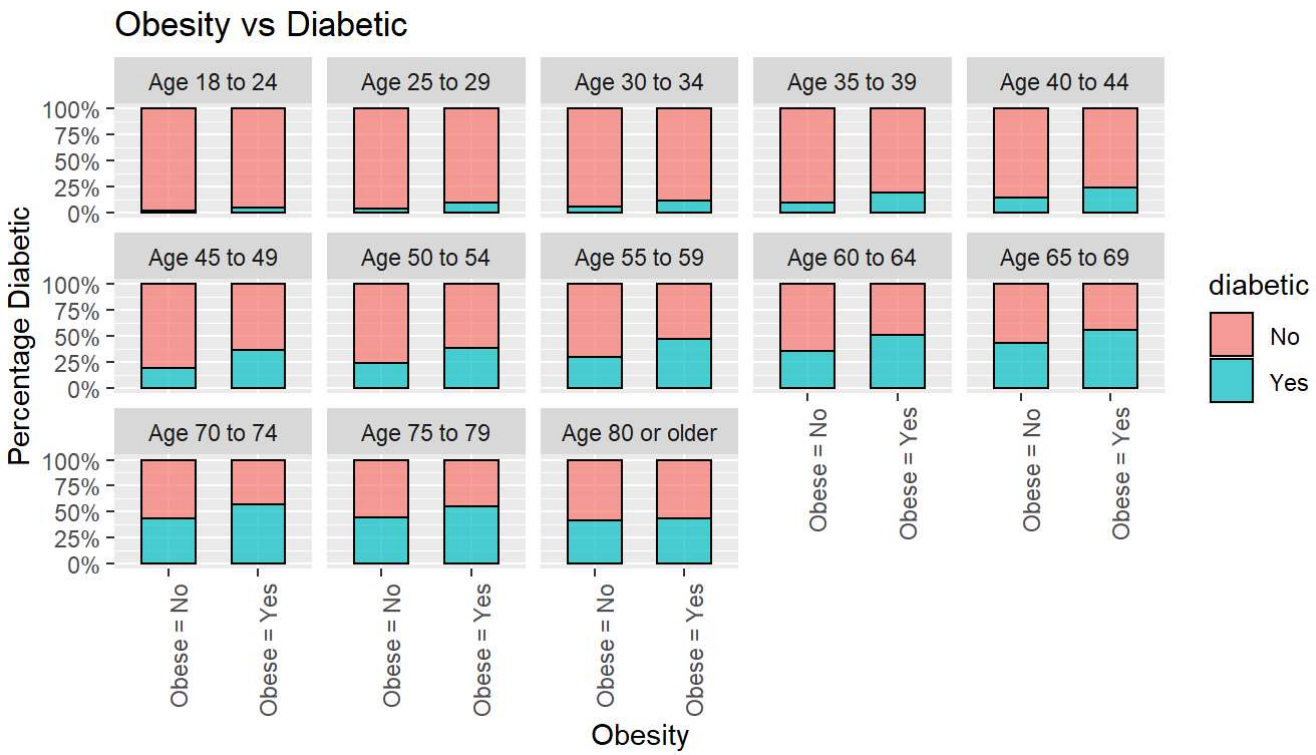
Further based on the `BMI` values, we label the reading with `BMI` > 30 as `obese` = "Yes" and `BMI` < = 30 as `obese` = "No"

```
# Determining whether an individual is Obese

df3<- df3%>%
filter(BMI>19.3, BMI<50)

df3<- df3%>%
mutate(obese = ifelse(BMI > 30,"Obese = Yes","Obese = No"))
```

P2: Data Analysis:

Based on the final subset of data thus obtained, we plot "Stacked bar-plots":

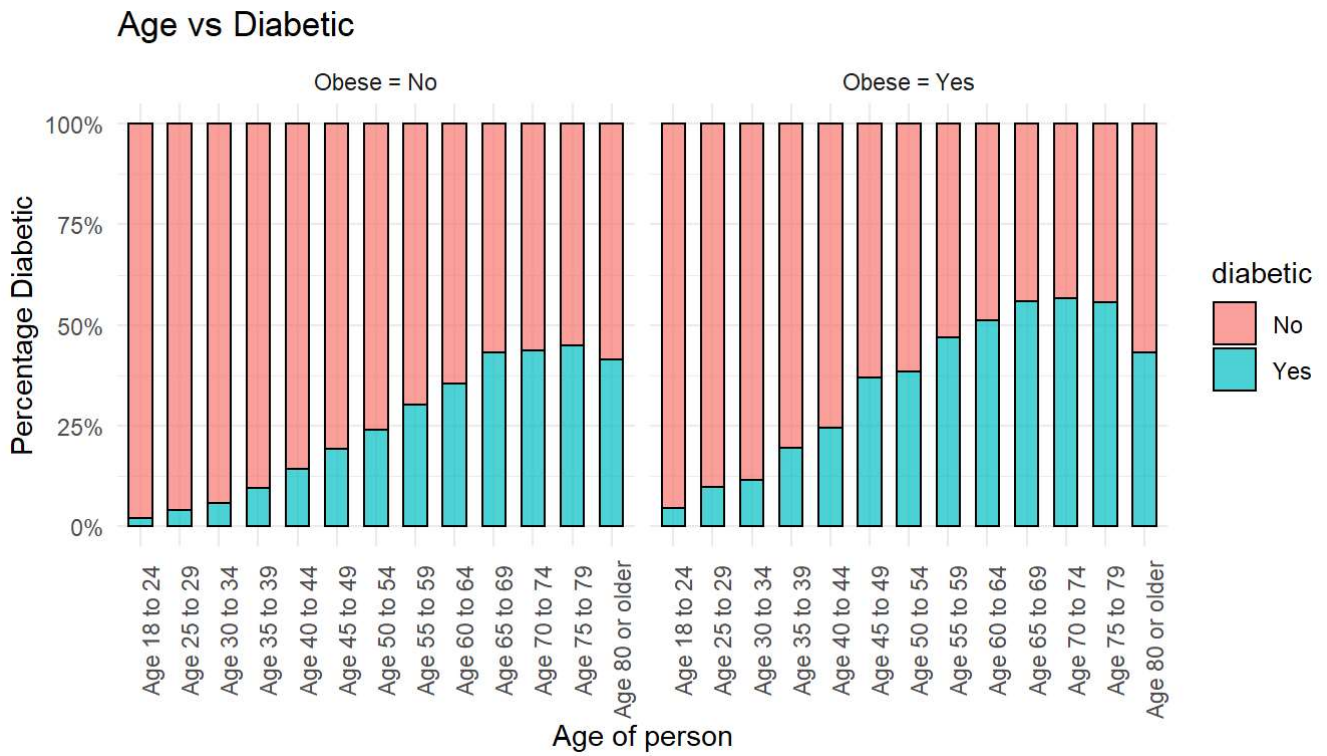
```
# Side-by-side Bar plot, Obesity Vs Diabetic

ggplot(df3, aes(x=obese, fill = diabetic)) +
geom_bar(position="fill", width = 0.6, color="black", size=0.5, alpha = 0.7) + facet_wrap( ~ X_a
         geg5yr, ncol=5)+ theme(legend.position = "right", axis.text.x = element_text(angle=90,
         hjust = 0.8))+ labs(x="Obesity", y="Percentage Diabetic", title="Obesity vs Diabetic")+
         scale_y_continuous(labels = scales::percent)
```



Plot 1: Obesity vs Diabetic:

Summary 3.1: Analyzing the plot we can conclude that the percentage of diabetic people increases if the population is obese, thus suggesting that **obese people are more prone to being diabetic.**

```
# Side-by-side Bar plot, Age Vs Diabetic

ggplot(df3, aes(x=X_ageg5yr, fill = diabetic)) + geom_bar(position="fill", width = 0.6, color="b
         lack", size=0.5, alpha = 0.7) + facet_wrap( ~ obese, ncol=5)+theme_minimal()+ theme(leg
         end.position = "right", axis.text.x = element_text(angle=90, hjust = 0.8))+ labs(x="Age
         of person", y="Percentage Diabetic", title="Age vs Diabetic")+ scale_y_continuous(label
         s = scales::percent)
```



Plot 2: Age vs Diabetic:

Summary 3.2: It is evident from the plot that irrespective of whether a person is obese or not, the risk of developing diabetes increases with the person's age. Thus suggesting, **one's age increases the risk of developing diabetes.**