Exploring the BRFSS data

Setup

Load packages

```
library(ggplot2)
library(dplyr)
```

Load data

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

Part 1: Data

The BRFSS data set is collected using Simple Random Sampling. This is observational study so we can only generalize the drawn inference to the population.

The correlation will not imply any causation between predictor and response variable as it is observational study not random assignment experimental study.

Part 2: Research questions

Research quesion 1: Is there any relationship between Body Mass Index and Diabetes.We can generalize drawn inference to population whether person of specific weight group suffers more from diabetes.

Research quesion 2: Analyze whether High Cholesterol is related to physical activity category and Heavy Alcohol consumptions or not. Also check if alcohol consumption is related to High Cholesterol across different Physical Activity Category.

Research quesion 3: Analyze how different age groups and different income groups are related to Depressive Disorder.Find out whether in Across the Income Category or Age group the variability in proportion of people having Depressive Disorder is more.

Part 3: Exploratory data analysis

Research quesion 1:

Create a new table named diab_bmi that shows proportion of people suffering from diabetes in each categorical BodyMassIndex(BMI). Diabetes during pregnency or any record with missing values are ellimineted.

```
diab_bmi<- brfss2013%>%
filter(!is.na(X_bmi5cat),!is.na(prediab1),prediab1!="Yes, during pregnancy")%>%
group_by(X_bmi5cat,prediab1)%>%
summarise(n=n())%>%
mutate(prop=round(n/sum(n),3))

brfss2013%>%
filter(!is.na(X_bmi5cat),!is.na(prediab1),prediab1!="Yes, during pregnancy")%>%
group_by(X_bmi5cat,prediab1)%>%
summarise(n=n())%>%
mutate(prop=round(n/sum(n),3))
```

```
## # A tibble: 8 x 4
## # Groups:   X_bmi5cat [4]
##   X_bmi5cat     prediab1     n  prop
##   <fct>         <fct>    <int> <dbl>
## 1 Underweight   Yes        201 0.046
## 2 Underweight   No        4140 0.954
## 3 Normal weight Yes       4059 0.052
## 4 Normal weight No       73870 0.948
## 5 Overweight    Yes       7446 0.093
## 6 Overweight    No       72599 0.907
## 7 Obese         Yes       9588 0.168
## 8 Obese         No       47614 0.832
```

We will use the above table to create a segmeneted bar graph showing proportions of diabetic condition in each MBI category.

```
ggplot(diab_bmi,aes(x=X_bmi5cat,fill=factor(prediab1,levels=c("No","Yes")),
                    y=prop))+
geom_bar(stat = "identity",position = "fill")+
geom_text(aes(label = prop),size = 3,position = position_stack(vjust = 0.5))+
labs(fill="Diabetic",x="BodyMassIndex",y="Prop")
```

As seen in the above graph that diabetic condition seems to increase with BMI. We will simulate the above study using randomization to analyze whether there truly exitsts any relationship between predictor and response variable or not.

First we calculate the probability of diabetes in our observational study.

```
brfss2013%>%
filter(!is.na(X_bmi5cat),!is.na(prediab1),prediab1!="Yes, during pregnancy")%>%
select(X_bmi5cat,prediab1)%>%
group_by(prediab1)%>%
summarise(n=n())%>%
mutate(prop=n/sum(n))
```

```
## # A tibble: 2 x 3
##   prediab1      n   prop
##   <fct>     <int>  <dbl>
## 1 Yes       21294 0.0970
## 2 No       198223 0.903
```

$$P(\text{Diabetes}) = 0.0.097$$

We will simulate the study in such a way that diabetic condition in each Categorical BodyMassIndex is purely due to chance i.e. random assignment. Code to generate random diabetic outcome.

```
diab_outcom<-c("Yes","No")
sim_diab_condition<-sample(diab_outcom,size = 219517,replace = TRUE,
                           prob = c(0.097,0.903))
```

Assigning random diabetic outcome two each of the four categorical BodyMassIndex.

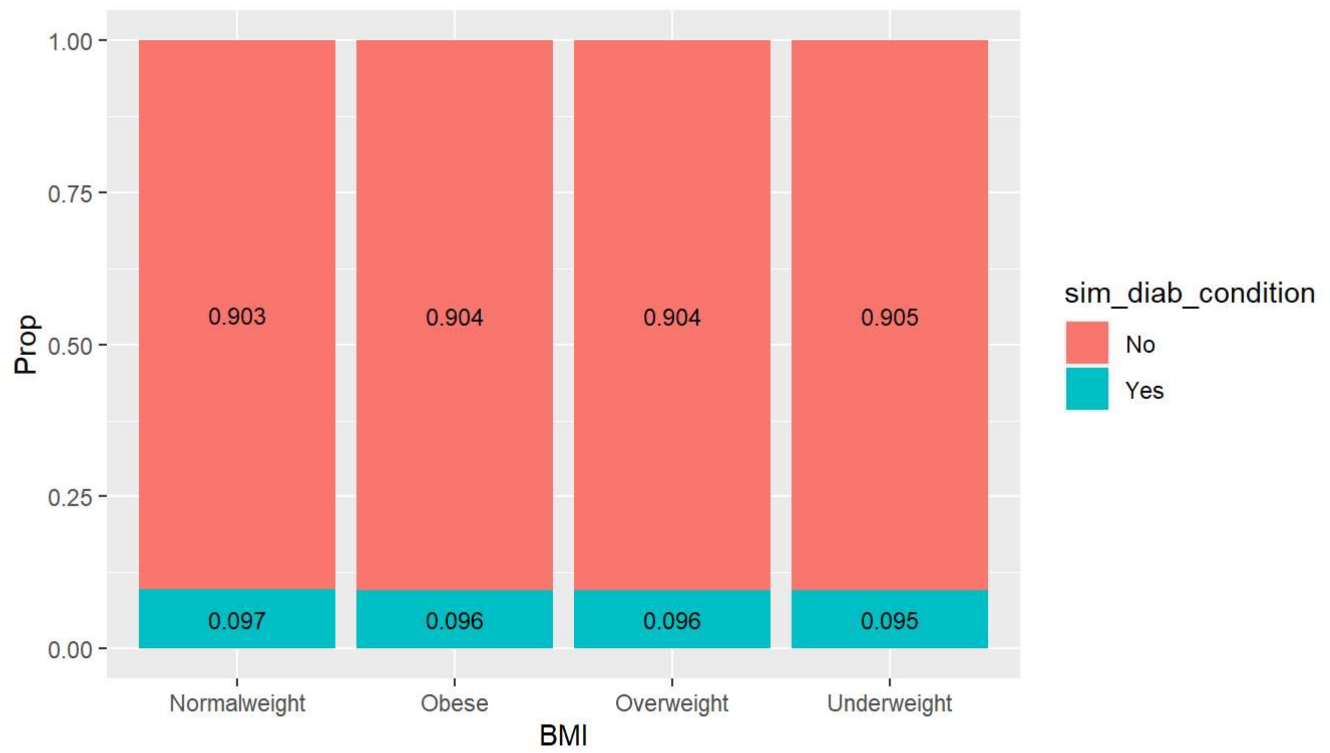
```
df1 <- data.frame(Sl_No = 1:4341, BMI = c("Underweight"))
df2 <- data.frame(Sl_No = 4342:82270, BMI = c("Normalweight"))
df3 <- data.frame(Sl_No = 82271:162315, BMI = c("Overweight"))
df4 <- data.frame(Sl_No = 162316:219517, BMI = c("Obese"))
simulation_table<- rbind(df1,df2,df3,df4)
simulation_table<- cbind(simulation_table,data.frame(sim_diab_condition))
```

Calculating proportion of Daibetes in randomized simulated study for each Categorical BodyMassIndex.

```
diab_bmi_simulation<- simulation_table%>%
group_by(BMI,sim_diab_condition)%>%
summarise(n=n())%>%
mutate(prop=round(n/sum(n),3))
```

Segmented bar graph to visulaise the proportion of diabtes in our simulated study.

```
ggplot(diab_bmi_simulation,aes(x=BMI,fill=factor(sim_diab_condition,
                                                  levels=c("No","Yes")),y=prop))+
geom_bar(stat = "identity",position = "fill")+
geom_text(aes(label = prop),size = 3,position = position_stack(vjust = 0.5))+
labs(fill="sim_diab_condition",x="BMI",y="Prop")
```



In the above plot the proportion of diabetes is almost same in all the BMI groups.So when assignment is random and chance diabetes is almost same across all groups of BMI i.e the chance of a person suffering from Diabetes is not related to his/her BMI.

CONCLUSION

From the above observation it is clear that diabetic condition and BodyMassIndex are correlated and hence we can generalize for the population that chnace of diabetes increase with increase in BodyMassIndex.

Research quesion 2:

We will create an additional variable t1:heavy alcohol consumer{Yes,No} in "brfss2013" data set. We are considering more than 60 drinks per month as Heavy Alcohol Consumption Group.

```
brfss2013<- brfss2013%>%
mutate(t1=ifelse(X_drnkmo4>60,"Yes","No"))
```

We are creating a new data frame named "cholesterol" where observations include Physical Activity Categories, Heavy Alcohol consumption and High cholesterol

```
cholesterol<- brfss2013%>%
select(X_pacat1,t1,X_rfchol)
cholesterol<- cholesterol%>%
filter(!is.na(X_pacat1),!is.na(t1),!is.na(X_rfchol))%>%
select_all()
```

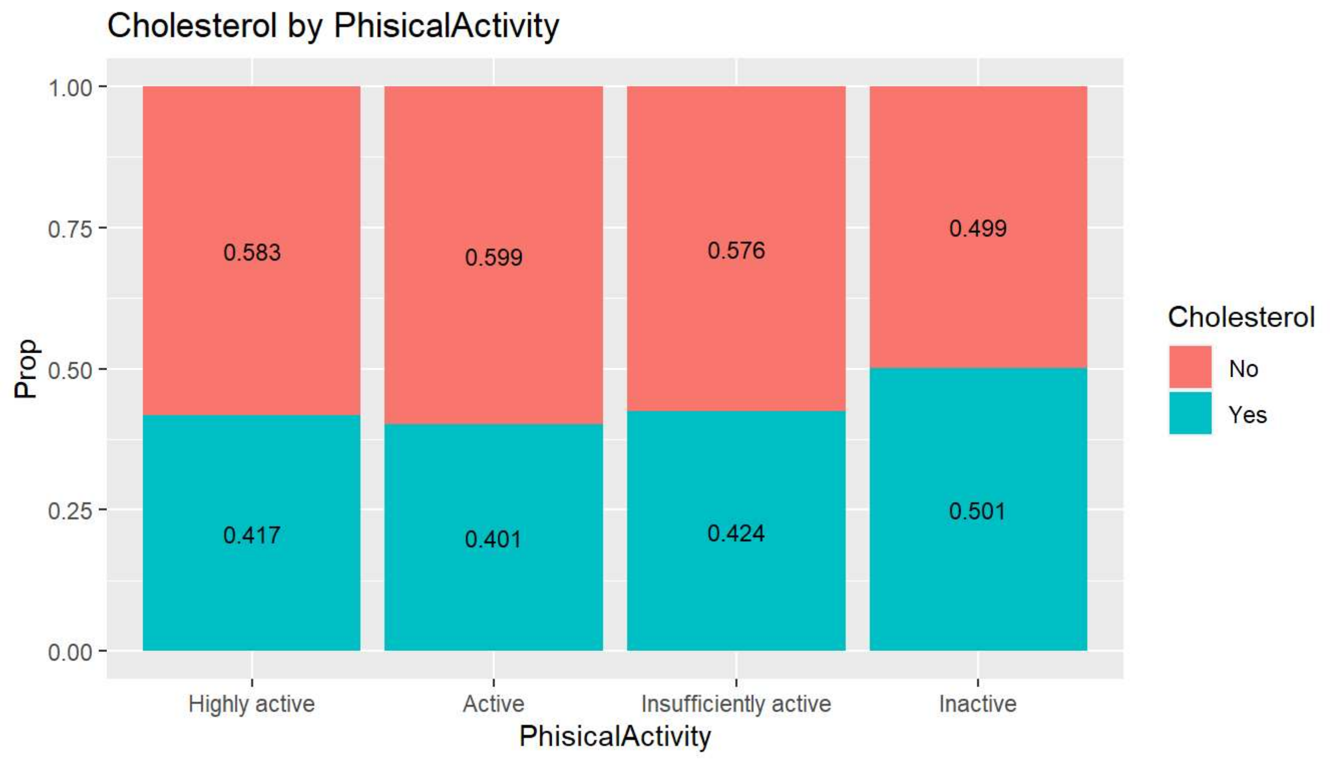
Let us see how does the cholesterol level varies across different Physical Activity Categories

```
chol_prop<- cholesterol%>%
group_by(X_pacat1,X_rfchol)%>%
summarise(n=n())%>%
mutate(prop=round(n/sum(n),3))
cholesterol%>%
group_by(X_pacat1,X_rfchol)%>%
summarise(n=n())%>%
mutate(prop=round(n/sum(n),3))
```

```
## # A tibble: 8 x 4
## # Groups:   X_pacat1 [4]
##   X_pacat1              X_rfchol     n  prop
##   <fct>                 <fct>    <int> <dbl>
## 1 Highly active         No       68218 0.583
## 2 Highly active         Yes      48769 0.417
## 3 Active                No       36634 0.599
## 4 Active                Yes      24519 0.401
## 5 Insufficiently active No       37213 0.576
## 6 Insufficiently active Yes      27351 0.424
## 7 Inactive              No       54150 0.499
## 8 Inactive              Yes      54331 0.501
```

The visualization for above data set.

```
ggplot(chol_prop,aes(x=X_pacat1,fill=factor(X_rfchol,levels=c("No","Yes")),
                     y=prop))+
geom_bar(stat = "identity",position = "fill")+
geom_text(aes(label = prop),size = 3,position = position_stack(vjust = 0.5))+
labs(fill="Cholesterol",x="PhisicalActivity",y="Prop",title = "Cholesterol by PhisicalActivity")
```



From the above graph it is seen that cholesterol level seems to increase as Physical Activity decreases. Now let us check if Cholesterol level is associated with heavy alcohol consumption.

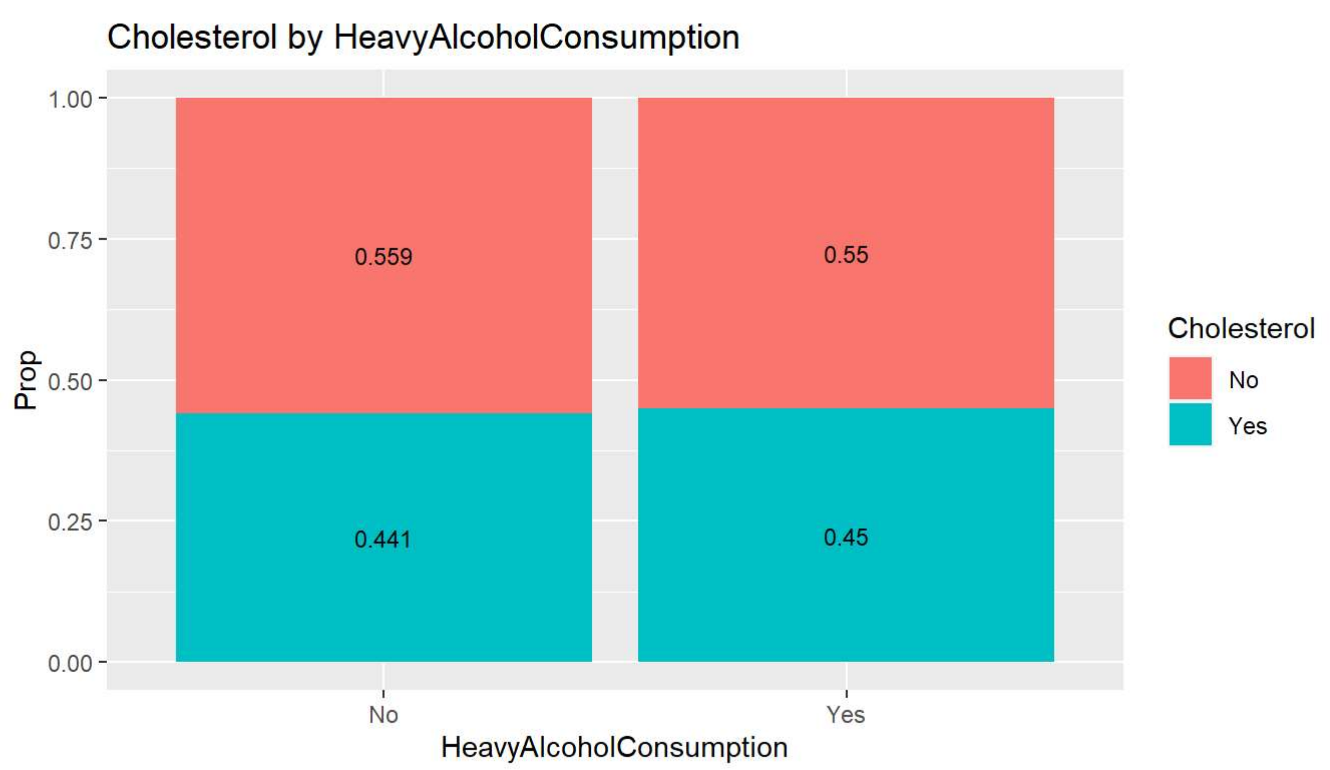
```
chol_prop1<- cholesterol%>%
group_by(t1,X_rfchol)%>%
summarise(n=n())%>%
mutate(prop=round(n/sum(n),3))

cholesterol%>%
group_by(t1,X_rfchol)%>%
summarise(n=n())%>%
mutate(prop=round(n/sum(n),3))
```

```
## # A tibble: 4 x 4
## # Groups:   t1 [2]
##   t1    X_rfchol      n  prop
##   <chr> <fct>     <int> <dbl>
## 1 No    No       190645 0.559
## 2 No    Yes      150421 0.441
## 3 Yes   No         5570 0.55
## 4 Yes   Yes        4549 0.45
```

The Visualization for above data.

```
ggplot(chol_prop1,aes(x=t1,fill=factor(X_rfchol,levels=c("No","Yes")),
                      y=prop))+
geom_bar(stat = "identity",position = "fill")+
geom_text(aes(label = prop),size = 3,position = position_stack(vjust = 0.5))+
labs(fill="Cholesterol",x="HeavyAlcoholConsumption",y="Prop",title = "Cholesterol by HeavyAlcoholConsumption")
```



From the above it is clear that cholesterol level is not associated with Alcohol Consumption as people from both the group are effected almost equally by high Cholesterol.

Since Heavy Alcohol Consumption alone is not effecting cholesterol level we will try to analyze if Heavy Alcohol Consumption is effecting Cholesterol level differently across the Physical Activity Category.

To do that we will first create tow table showing proportion of people suffering from Cholesterol across the Physical Activity Category in low Alcohol Consumption group and heavy Alcohol Consumption group.

Let's first check for heavy Alcohol Consumption group

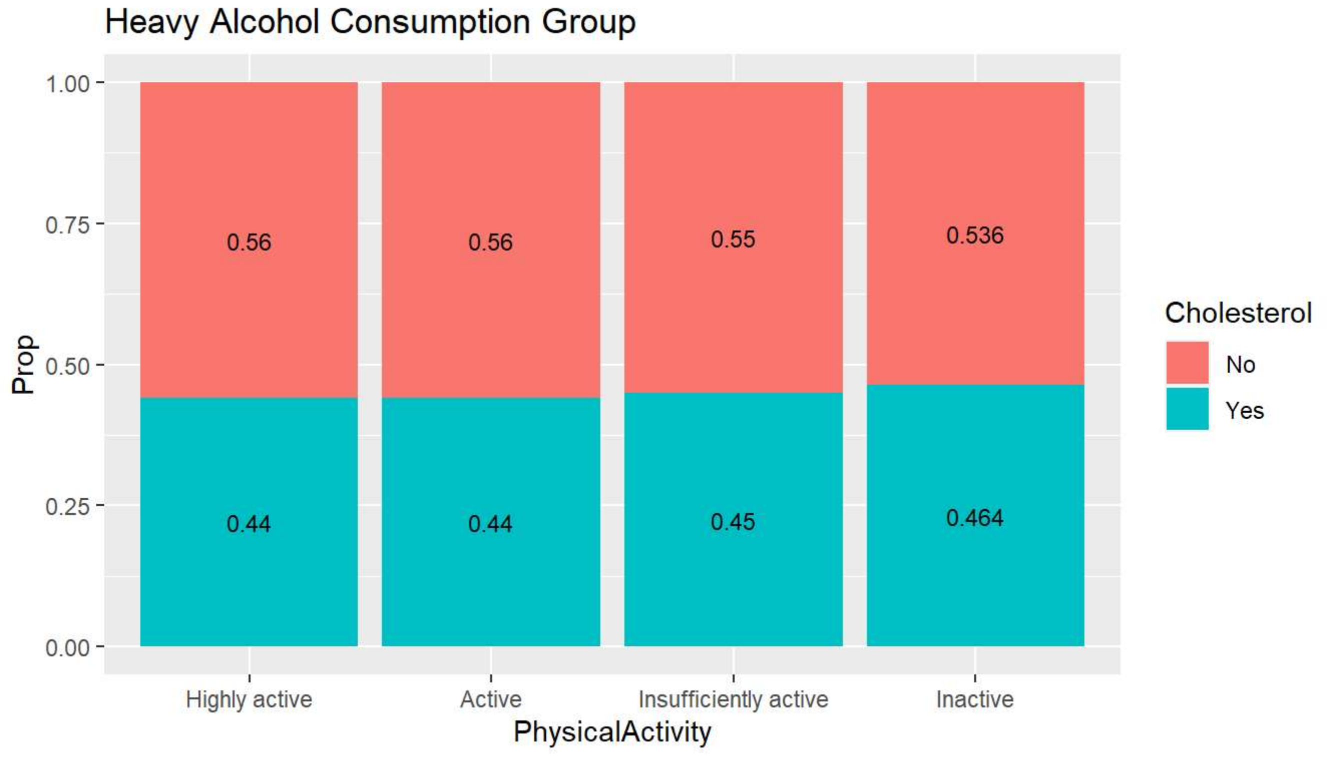
```
chol_hal_prop<- cholesterol%>%
filter(t1=="Yes")%>%
group_by(X_pacat1,X_rfchol)%>%
summarise(n=n())%>%
mutate(prop=round(n/sum(n),3))

cholesterol%>%
filter(t1=="Yes")%>%
group_by(X_pacat1,X_rfchol)%>%
summarise(n=n())%>%
mutate(prop=round(n/sum(n),3))
```

```
## # A tibble: 8 x 4
## # Groups:   X_pacat1 [4]
##   X_pacat1              X_rfchol     n  prop
##   <fct>                 <fct>    <int> <dbl>
## 1 Highly active         No        2071 0.56
## 2 Highly active         Yes       1630 0.44
## 3 Active                No         886 0.56
## 4 Active                Yes        696 0.44
## 5 Insufficiently active No         882 0.55
## 6 Insufficiently active Yes        722 0.45
## 7 Inactive              No        1731 0.536
## 8 Inactive              Yes       1501 0.464
```

Now we will visualize the above data.

```
ggplot(chol_hal_prop,aes(x=X_pacat1,fill=factor(X_rfchol,levels=c("No","Yes")),
                     y=prop))+
geom_bar(stat = "identity",position = "fill")+
geom_text(aes(label = prop),size = 3,position = position_stack(vjust = 0.5))+
labs(fill="Cholesterol",x="PhysicalActivity",y="Prop",title = "Heavy Alcohol Consumption Group")
```



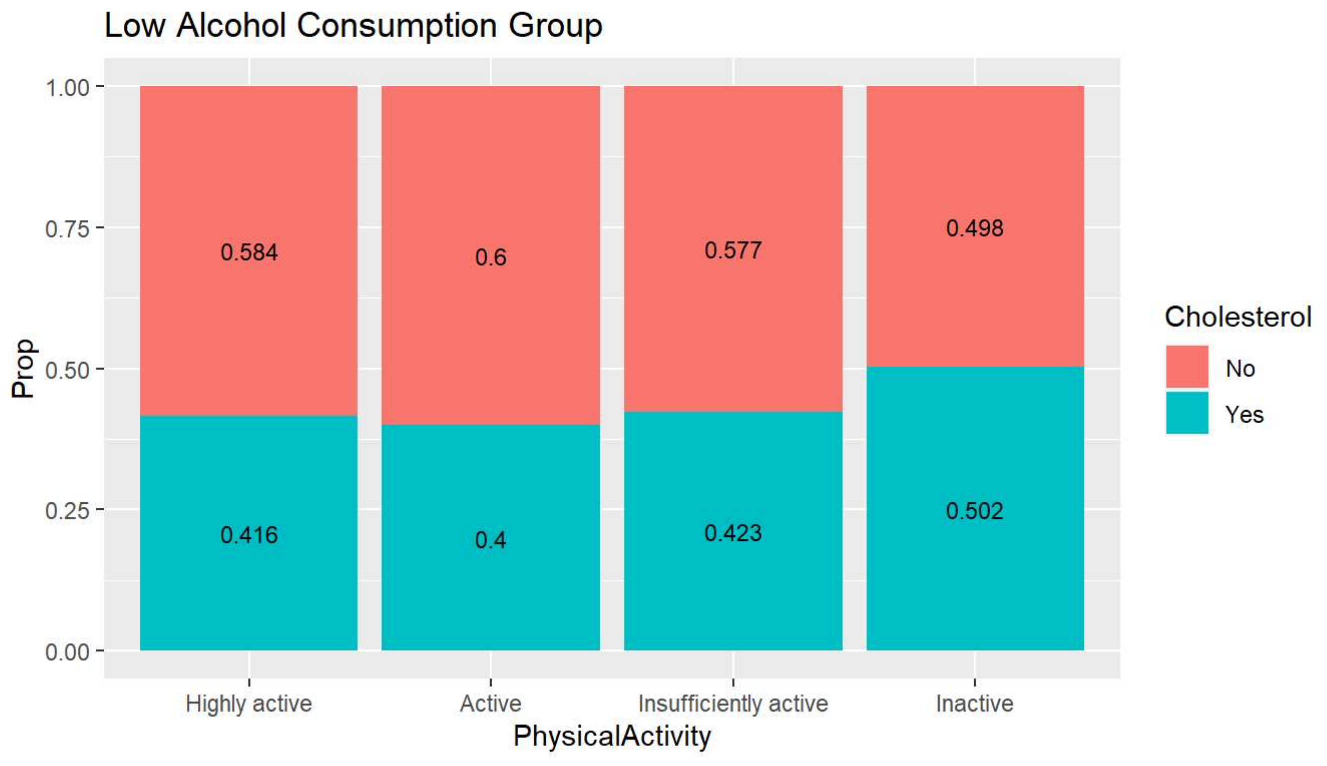
Now we will calculate the proportion of Cholesterol across Physical Activity Categories for Low Alcohol Consumption Group.

```
chol_lal_prop<- cholesterol%>%
filter(t1=="No")%>%
group_by(X_pacat1,X_rfchol)%>%
summarise(n=n())%>%
mutate(prop=round(n/sum(n),3))

cholesterol%>%
filter(t1=="No")%>%
group_by(X_pacat1,X_rfchol)%>%
summarise(n=n())%>%
mutate(prop=round(n/sum(n),3))
```

```
## # A tibble: 8 x 4
## # Groups:   X_pacat1 [4]
##   X_pacat1              X_rfchol      n  prop
##   <fct>                 <fct>     <int> <dbl>
## 1 Highly active         No        66147 0.584
## 2 Highly active         Yes       47139 0.416
## 3 Active                No        35748 0.6
## 4 Active                Yes       23823 0.4
## 5 Insufficiently active No        36331 0.577
## 6 Insufficiently active Yes       26629 0.423
## 7 Inactive              No        52419 0.498
## 8 Inactive              Yes       52830 0.502
```

```
ggplot(chol_lal_prop,aes(x=X_pacat1,fill=factor(X_rfchol,levels=c("No","Yes")),
                         y=prop))+
geom_bar(stat = "identity",position = "fill")+
geom_text(aes(label = prop),size = 3,position = position_stack(vjust = 0.5))+
labs(fill="Cholesterol",x="PhysicalActivity",y="Prop",title = "Low Alcohol Consumption Group")
```



OBSERVATION

Comparing the data from the above observations we can see that though Hevay Alcohol consumption is not related to cholesterol by itself but Cholesterol is effeted differenlt across the Physical Acitvity Category for the TWO ACOHOL CONSUMPTION GROUP.

1. In heavy alcohol consumption group the cholsterol effets tends to remain same across all the Physical Activity Category.
2. In low alcohol consumption group the cholsterol effets tends to increase with decrese in Physical Activity as seen without the effect of alcohol previously.
3. In heav alcohol consumption group the proportion of people suffering from cholesteol across the diffrent Physical Activity Category is slightly more than low alcohol consumption group except for the Inactive Category.

Research quesion 3:

We will create a new data frame named “dep_disorder” that contain the age_group,income_group and depressive_disoder information.

```
dep_disorder<- brfss2013%>%
select(X_incomg,X_ageg5yr,addepev2)
dep_disorder<- dep_disorder%>%
filter(!is.na(X_incomg),!is.na(X_ageg5yr),!is.na(addepev2))
```

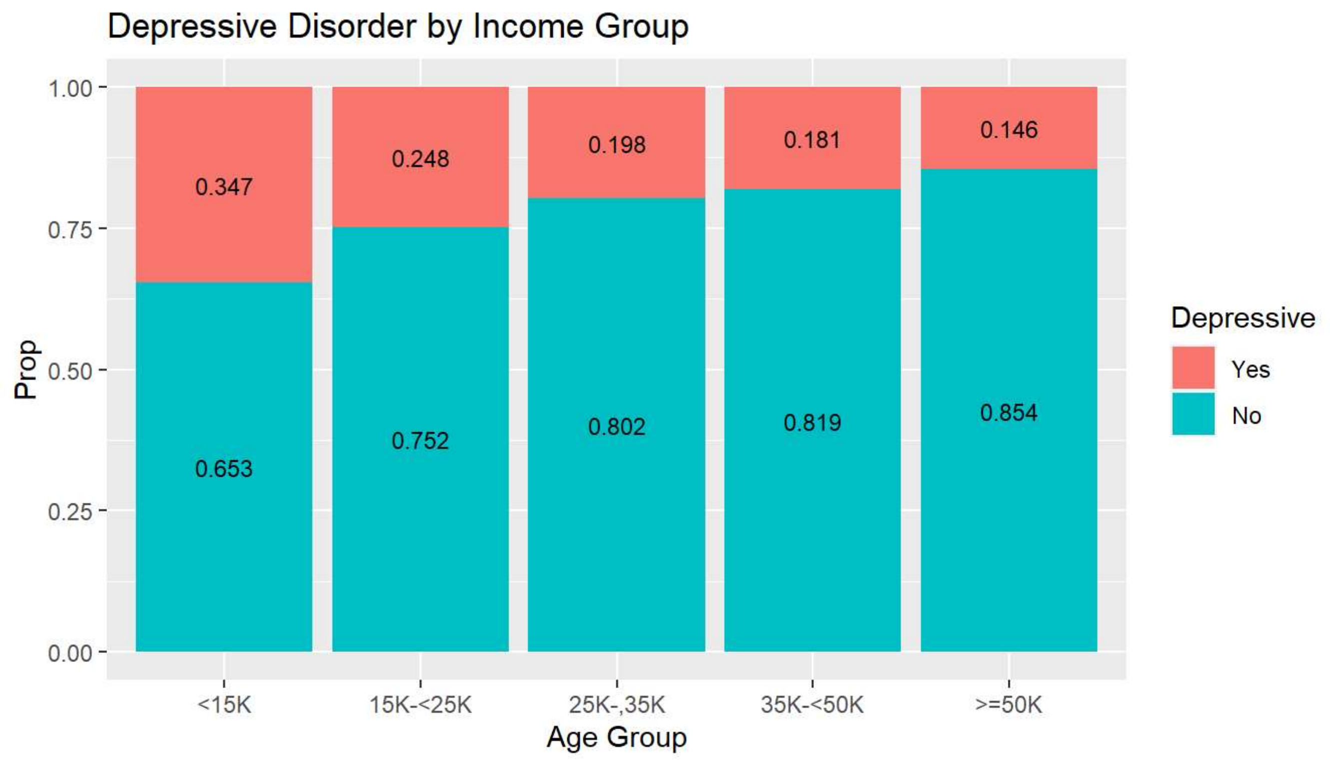
Now we will obsreve how Depressive Disorder is distributed across the different Income Category

```
dep_incomegp<- dep_disorder%>%
group_by(X_incomg,addepev2)%>%
summarise(n=n())%>%
mutate(prop=round(n/sum(n),3))
dep_disorder%>%
group_by(X_incomg,addepev2)%>%
summarise(n=n())%>%
mutate(prop=round(n/sum(n),3))
```

```
## # A tibble: 10 x 4
## # Groups:   X_incomg [5]
##    X_incomg                    addepev2      n  prop
##    <fct>                       <fct>     <int> <dbl>
##  1 Less than $15,000           Yes       17904 0.347
##  2 Less than $15,000           No        33698 0.653
##  3 $15,000 to less than $25,000 Yes      18801 0.248
##  4 $15,000 to less than $25,000 No       57050 0.752
##  5 $25,000 to less than $35,000 Yes       9580 0.198
##  6 $25,000 to less than $35,000 No       38875 0.802
##  7 $35,000 to less than $50,000 Yes      11022 0.181
##  8 $35,000 to less than $50,000 No       50014 0.819
##  9 $50,000 or more             Yes       26284 0.146
## 10 $50,000 or more             No       153558 0.854
```

Let’s check the graph for the above data set.

```
ggplot(dep_incomegp,aes(x=factor(X_incomg,
                              labels=c("<15K","15K-<25K","25K-,35K","35K-<50K",">=50K")),
                        fill=factor(addepev2,levels=c("Yes","No")),y=prop))+
geom_bar(stat = "identity",position = "fill")+
geom_text(aes(label = prop),size = 3,position = position_stack(vjust = 0.5))+
labs(fill="Depressive",x="Age Group",y="Prop",title = "Depressive Disorder by Income Group")
```



Form the above graph we observe that Depressive Disorder decreases as we move up to the Higher Income Group.

Now we will see the rlation of Depressive Disorder to the Age Groups.

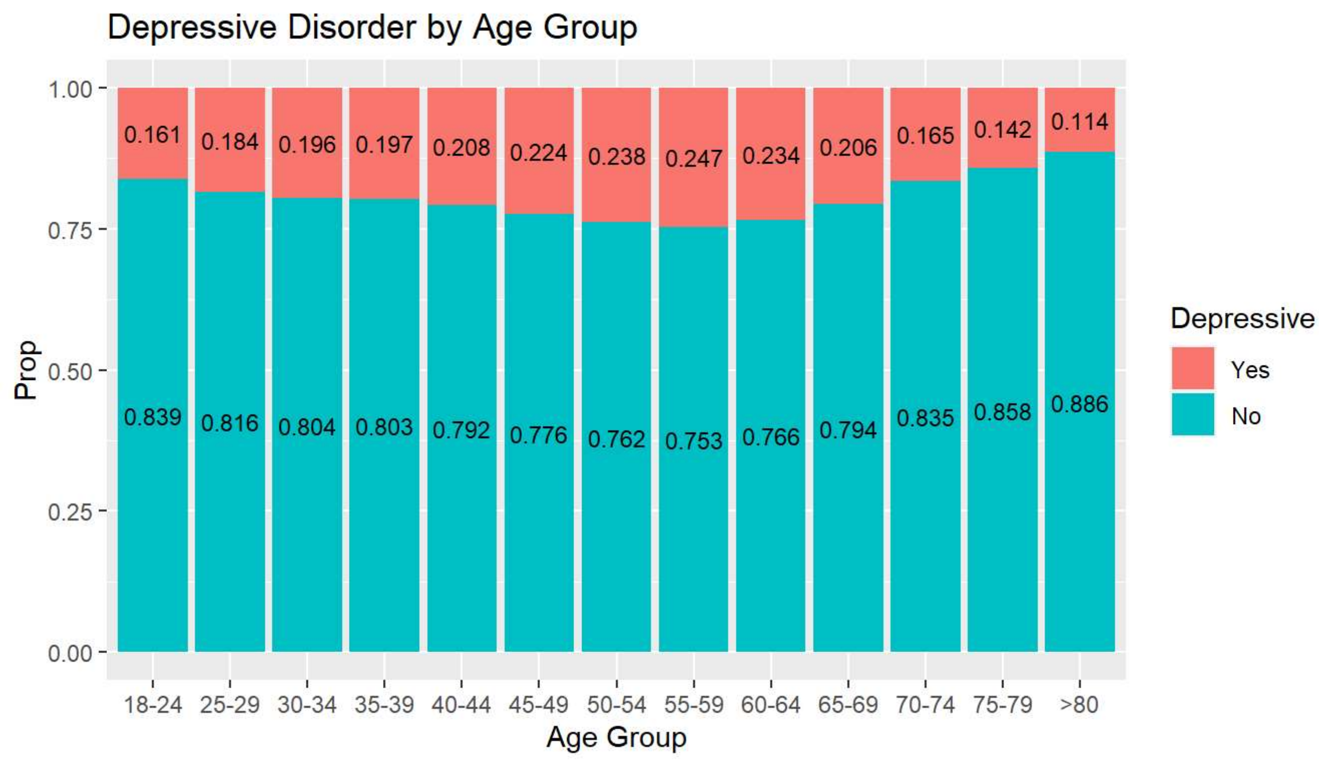
```
dep_agegp<- dep_disorder%>%
group_by(X_ageg5yr,addepev2)%>%
summarise(n=n())%>%
mutate(prop=round(n/sum(n),3))

dep_disorder%>%
group_by(X_ageg5yr,addepev2)%>%
summarise(n=n())%>%
mutate(prop=round(n/sum(n),3))
```

```
## # A tibble: 26 x 4
## # Groups:   X_ageg5yr [13]
##    X_ageg5yr    addepev2     n  prop
##    <fct>        <fct>    <int> <dbl>
##  1 Age 18 to 24 Yes       3323 0.161
##  2 Age 18 to 24 No       17376 0.839
##  3 Age 25 to 29 Yes       3760 0.184
##  4 Age 25 to 29 No       16723 0.816
##  5 Age 30 to 34 Yes       4861 0.196
##  6 Age 30 to 34 No       20001 0.804
##  7 Age 35 to 39 Yes       5061 0.197
##  8 Age 35 to 39 No       20638 0.803
##  9 Age 40 to 44 Yes       5989 0.208
## 10 Age 40 to 44 No       22762 0.792
## # ... with 16 more rows
```

Let’s visualize the distribution in graph

```
ggplot(dep_agegp,aes(x=factor(X_ageg5yr,                                    labels=c("18-24","25-29","30-34","35-39","40-44",
         "45-49","50-54","55-59","60-64","65-69","70-74","75-79",">80")),
          fill=factor(addepev2,levels=c("Yes","No")),y=prop))+
geom_bar(stat = "identity",position = "fill")+
geom_text(aes(label = prop),size = 3,position = position_stack(vjust = 0.5))+
labs(fill="Depressive",x="Age Group",y="Prop",title = "Depressive Disorder by Age Group")
```



From the above graph we see a peculiar trend that Depressive Disorder Inceaaes from lower age group to higher age group up to Age group 55 to 59 the decrese from next higher age group onwards. Now we will check for variability in proportion of Depressive Disorder Across Age Group and across Income Group.

First checking for variability across Age Group

```
dep_agegp%>%
group_by(addepev2)%>%
summarise(SD=sd(prop))%>%
filter(addepev2=="Yes")
```

```
## # A tibble: 1 x 2
##   addepev2     SD
##   <fct>     <dbl>
## 1 Yes      0.0395
```

Now checking for variability across Income Group

```
dep_incomegp%>%
group_by(addepev2)%>%
summarise(SD=sd(prop))%>%
filter(addepev2=="Yes")
```

```
## # A tibble: 1 x 2
##   addepev2     SD
##   <fct>     <dbl>
## 1 Yes      0.0780
```

OBSERVATION

From the above data it is clear that Variability of Depressive Disorder is more across the Income Group(SD=0.0780) than the Age Group(SD=0.0395)