

Exploring the BRFSS data

Setup

Load packages

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

Load data

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

Part 1: Data

What is this data about?

- In 1984 the Center for Disease control (CDC) started a *survey* which is state-based in the US.
- The *survey* is called **Behavioral Risk Factor Surveillance System** (BRFSS)
- The survey is conducted via landline and cellular telephones.
- The survey uses a standardized questionnaire.
- The survey is used to collect data among the US population regarding their: risk behaviour and preventive health practices that can affect their health status.

Structure of the survey?

- **Standard Core Questions:** Questions that are asked by every state across the US.
- **Rotating Core Questions:** Questions that are asked by all states on an alternate year basis.
- **Optional Questions:** Optional questions from a standardized set of questions that each state can choose from to include in its questionnaire.
- **State-added Questions:** State specific questions that are not common across all the states but specific to that particular state asking them.

How to ensure survey is uniform across all of the US?

In order to ensure uniformity of the survey across all of the US. BRFSS has certain standards that are followed:

- All states ask the **Standard Core Questions** without modification. This ensures uniform data observation from the survey. Addition of the **Optional Questions** and **State-added Questions** is left to the state.
- unobtrusive electronic monitoring of the interviewers is also done across all states. To ensure uniformity of survey taking across all the states.
- Clear definition of what constitutes a unit from which the survey is taken: Household.

What sampling approach is used in selecting candidates for the survey?

- The survey is conducted in all states across the US. In this way it covers the whole of the US population in all the states and territories. (DC and GUAM, Porto Rico including)
- Sampling style used is Stratified Sampling: Each state decides to stratify based on any of the following ways: county, public-health district or other sub-geography.
- Inside each Strata, further Stratified Sampling is done based on households with Landlines, household with Cell phones.
- Landlines: Disproportionate stratified Sampling (DSS) is applied by splitting data into regions with high density landlines vs those with low-density landlines. A ratio of 1:1.5 is followed to ensure equal representation from regions with mid-density landlines. Furthermore inside each of the sampled households, one of the occupants of the household is randomly selected, with each occupant of the household having an equally likely chance of getting selected for the survey.
- Therefore at the topmost level we have stratified sampling, with each strata further subjected to stratified sampling, finally simple random sampling is used at the lowermost level to select the subject to be surveyed.
- Cell phones: Every cell phone is considered an individual single owner household. And random samples are drawn from a list of cell numbers, where each cell number has an equal likelihood of being sampled.

Answer :

- **Generalizability** : Since the data is collected using a survey of randomly selected individuals from stratified sample across the US. We can safely say the survey is generalizable **only** across the US population.
- **Causality** : The survey is **NOT** useful for causality. As we have not randomly selected and randomly assigned individuals into experimental groups to draw any causal relationships among the variable of the data in which we seek to establish causality.
- **Potential biases**: There is a chance of bias, due to non-participation of randomly selected individuals. Volunteer bias due to certain segments of individuals in each strata more willing to participate than other individuals.

In conclusion the survey is **Generalizable** across the US. Any patterns we observe are **Associative** and we need further controlled experiments to establish **Causality**. There is a possibility of **Participation bias** as well as **Non-response** bias to be introduced into the survey.

Part 2: Research questions

Research question 1:

Which has a stronger association with General Health, Income or Exercise?

- I am interested in seeing if any associative relationship exists between income and health, versus exercise and health.
- This question is of interest to me because. Not everyone is rich and can have access to better healthcare.
- But everyone can exercise to stay fit.
- So it will be interesting to see which has a more stronger associative relationship leading to better health if any.

Research question 2:

Is there any association between BMI and States ?

- I am interested in finding out, if all States in the US have equal distribution of BMI or if there are any interesting patterns in the data to be observed.
- This is important because BMI is an excellent indicator of a State's health metric.
- States with people in the Obese and Overweight BMI range are overall in an unhealthy category.

Research question 3:

Is there an association between Income and Depression ?

- I am interested in seeing if better income has any association with Depression.
 - I find this question really interesting as many rich and famous people commit suicide.
 - So this has made me curious, was there any association with high income that leads to these suicides.
 - This EDA will seek to see if the above line of thinking has any association or no association
-

Part 3: Exploratory data analysis

Research question 1:

Which has a stronger association with General Health, Income or Exercise?

Step1 : Load and Clean up Data for analysis

we begin by loading the required libraries; and loading the dataset into memory

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

```
##
## Attaching package: 'gridExtra'

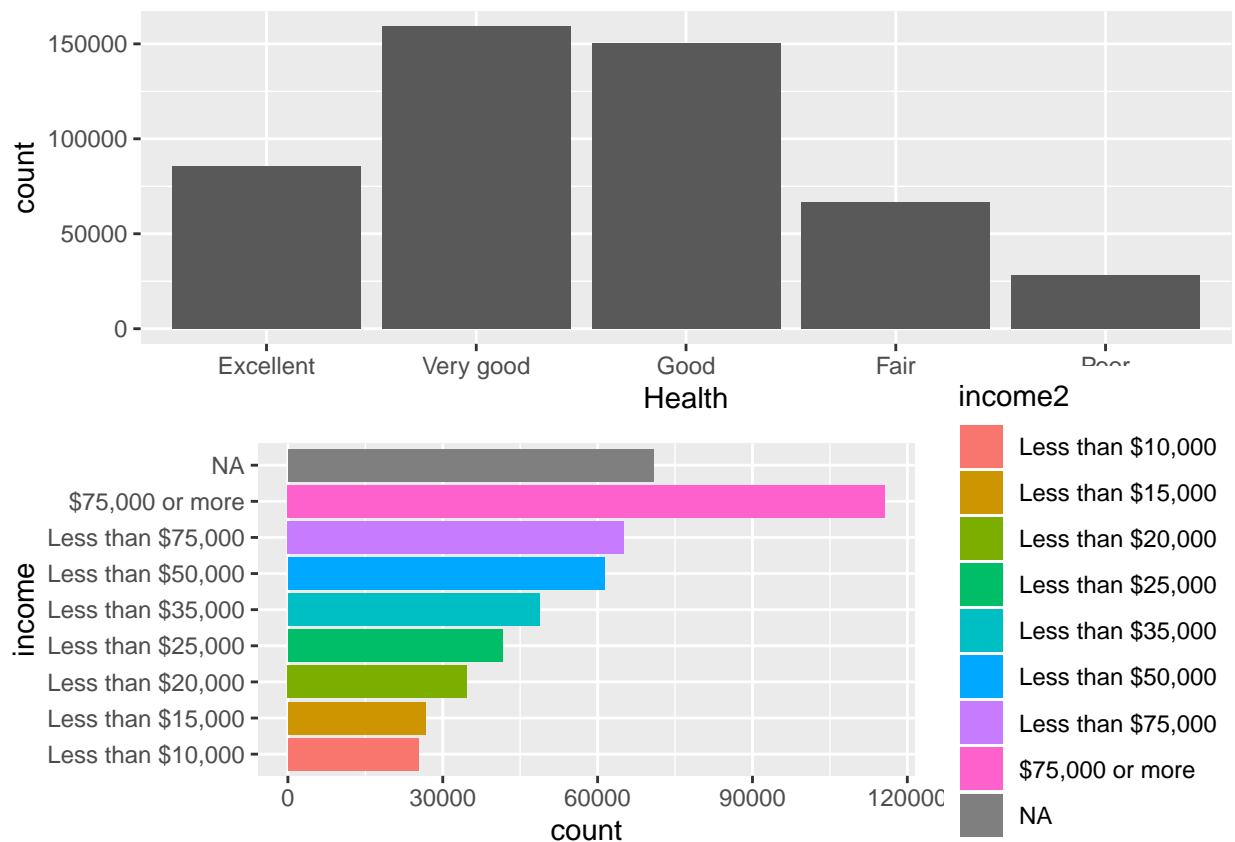
## The following object is masked from 'package:dplyr':
##
## combine
```

```
library(forcats)

load("brfss2013.RData")
```

We then select the variables of interest: genhlth(Ordinal), income2(Ordinal), exerany2(categorical)
Here we have: 2 Ordinal Variables and 1 categorical variable for doing our analysis.

```
analysis1 <- brfss2013%>%select(c("genhlth","income2","exerany2"))%>%filter(!is.na(genhlth) )
plothealth <- ggplot(analysis1, aes(x=genhlth))+geom_bar()+xlab("Health")
plotincome <- ggplot(analysis1, aes(fill = income2,x = income2))+geom_bar(position = "stack")+coord_flip()
grid.arrange(plothealth,plotincome,nrow = 2)
```



We see a quick plot of General health distribution and also income levels.

```
analysis1$income2 <- fct_explicit_na(analysis1$income2, na_level = "NotDisclosed")
analysis1%>%group_by(income2)%>%summarise(countgp = n(), total = length(analysis1$income2), percentage =
```

```
## # A tibble: 9 x 4
##   income2      countgp total percentage
##   <fct>      <int>   <int>     <dbl>
## 1 Less than $10,000  25252 489790     5.16
## 2 Less than $15,000  26633 489790     5.44
## 3 Less than $20,000  34705 489790     7.09
## 4 Less than $25,000  41563 489790     8.49
## 5 Less than $35,000  48687 489790     9.94
```

```
## 6 Less than $50,000    61319 489790    12.5
## 7 Less than $75,000   65102 489790    13.3
## 8 $75,000 or more     115659 489790    23.6
## 9 NotDisclosed        70870 489790    14.5
```

We see that there is a huge chunk of NotDisclosed **14.46%** , which is one of the levels in the income category. As a result we cannot drop it. But we will keep it as a separate category of it's own.

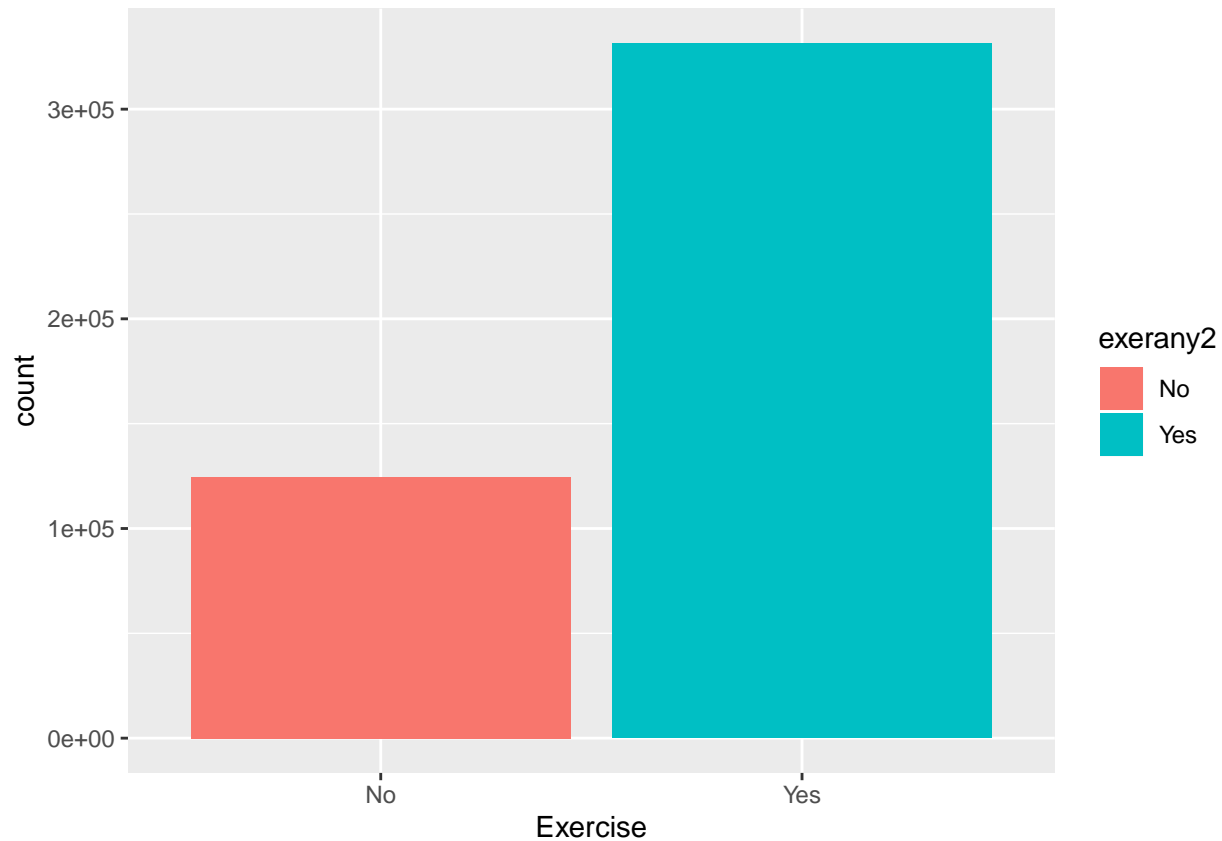
We will now analyze exercise to see if any NAs can be dropped

```
analysis1$exerany2 <- as.character(analysis1$exerany2)
analysis1$exerany2[is.na(analysis1$exerany2)] <- "NotDisclosed"
analysis1$exerany2 <- as.factor(analysis1$exerany2)
plotexercis <- ggplot(analysis1, aes(fill = exerany2,x=exerany2))+geom_bar()
analysis1%>%group_by(exerany2)%>%summarise(countgp = n(), total = length(analysis1$exerany2), percentage = countgp/total)
```

```
## # A tibble: 3 x 4
##   exerany2    countgp total percentage
##   <fct>         <int> <int>      <dbl>
## 1 No           124561 489790      25.4
## 2 NotDisclosed  33840 489790       6.91
## 3 Yes          331389 489790      67.7
```

6 percent is NotDisclosed for exercise this is a very small number, hence we can drop it from our analysis.

```
analysis1 <- analysis1%>%select(c("genhlth", "income2", "exerany2"))%>%filter(!(exerany2 == "NotDisclosed"))
plotexercis <- ggplot(analysis1, aes(fill = exerany2,x=exerany2))+geom_bar(position = "stack")+xlab("Exercise")
grid.arrange(plotexercis, nrow = 1)
```



We now have only Yes and No categories for those who exercise. As we have removed the 6% NotDisclosed category.

Step2 : Exploratory Data Analysis among the variables

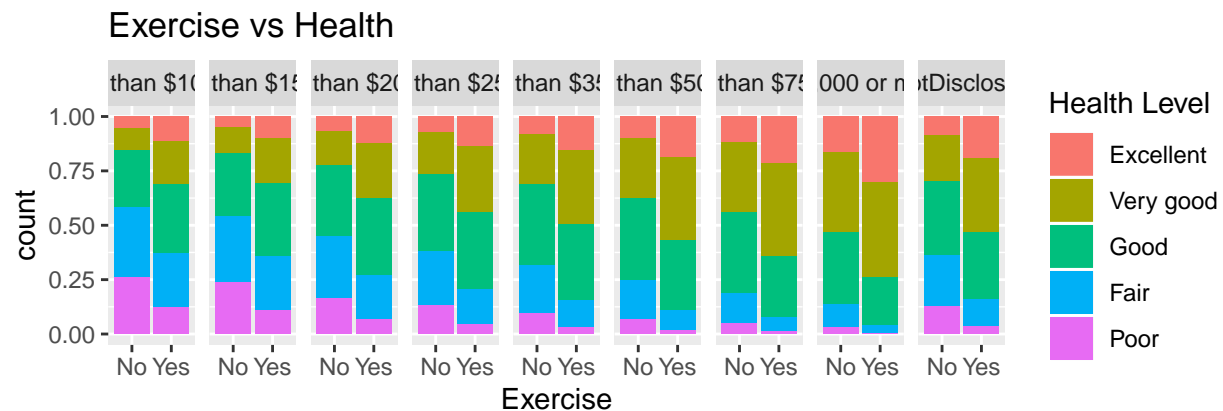
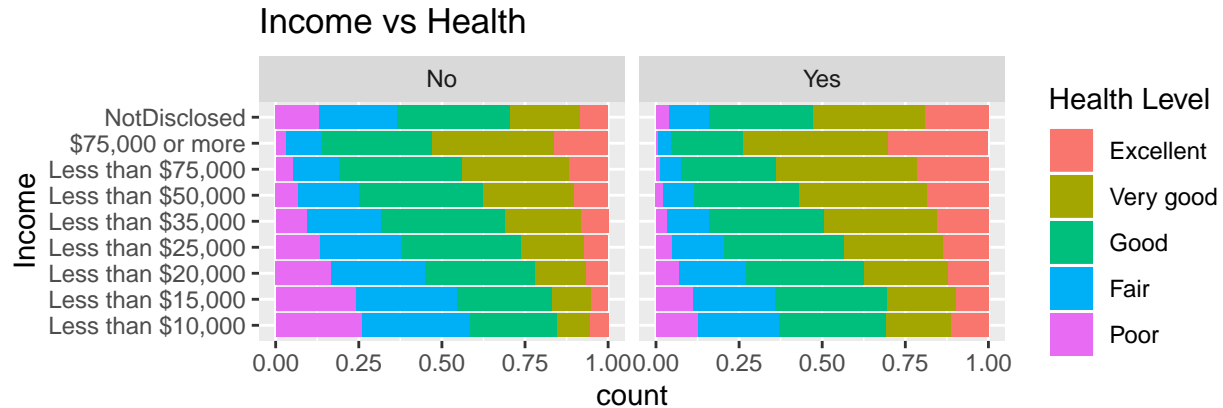
```
# There seems to be an associative relationship between income and health
plotincomevshealth <- ggplot(analysis1, aes(fill=genhlth, x = income2))+geom_bar(position = "fill")+xlab("Income")

# There is an associative relationship between exercise and health.
plotexercisvshealth <- ggplot(analysis1, aes(fill = genhlth,x=exerany2))+geom_bar(position = "fill")+xlab("Exercise")

# Relationship between income and exercise
plotincomevsexercise<- ggplot(analysis1, aes(fill = exerany2,x=income2))+geom_bar(position = "fill")

#Findings: There is an associative connection between: income&Health, exercise&health, income&exercise.
# Let us see what the relationship between these 3 would be:

p1 <- plotincomevshealth+facet_grid(~exerany2)+coord_flip()+ggtitle("Income vs Health")+guides(fill=guide_legend())
p2 <- plotexercisvshealth+facet_grid(~income2)+labs("Health")+ggtitle("Exercise vs Health")+guides(fill=guide_legend())
grid.arrange(p1,p2, nrow = 2)
```



Line1: Creates and stores of plot of Income Vs Health in the variable : `plotincomevshealth`
 Line2: Creates and stores a plot of Exercise Vs Health in the variable: `plotexercisvshealth`
 Line3: Creates and stores a plot of Income Vs Exercise in the variable: `plotincomevsexercise`
 Line 4: Creates breaks up the `plotincomevshealth` into 2 grids: Those who exercise vs Those who don't. This enables us to not only view any trends based on income but also trends based on exercise
 Line 5: Breaks up `plotexercisvshealth` on the basis of income. This enables us to see more clearly the impact of exercise on health, across all income categories.

Findings:

- **Warning** : Association is not Causation.
- From Plot: (*Income vs Health*) it can be seen there is an associative relationship between income and health and exercise.
- As income increases, there is a marked positive increase in health.
- A interesting observation is: With exercise, there is a greater increase in health benefits. This can be seen from Plot: (*Exercise vs Health*) by the steep difference between those who exercise(Yes) versus those who don't(No), even within same income categories.
- **Therefore, Exercise has a bigger impact on health, compared to income** , the relationship is purely correlational, we need further experiment to formally establish causality.

Research question 2:

Is there any association between BMI and States ?

Step1 : Load and Clean up Data for analysis

We select the variables of interest: weight, height, marital, X_State and income

```
frameofinterest <- brfss2013%>%select(c("weight2","height3","marital","X_state","income2"))
frameofinterest$marital <- fct_explicit_na(frameofinterest$marital, na_level = "NotDisclosed")
frameofinterest%>%group_by(marital)%>%summarise(countgp = n(), total = length(frameofinterest$marital),
```

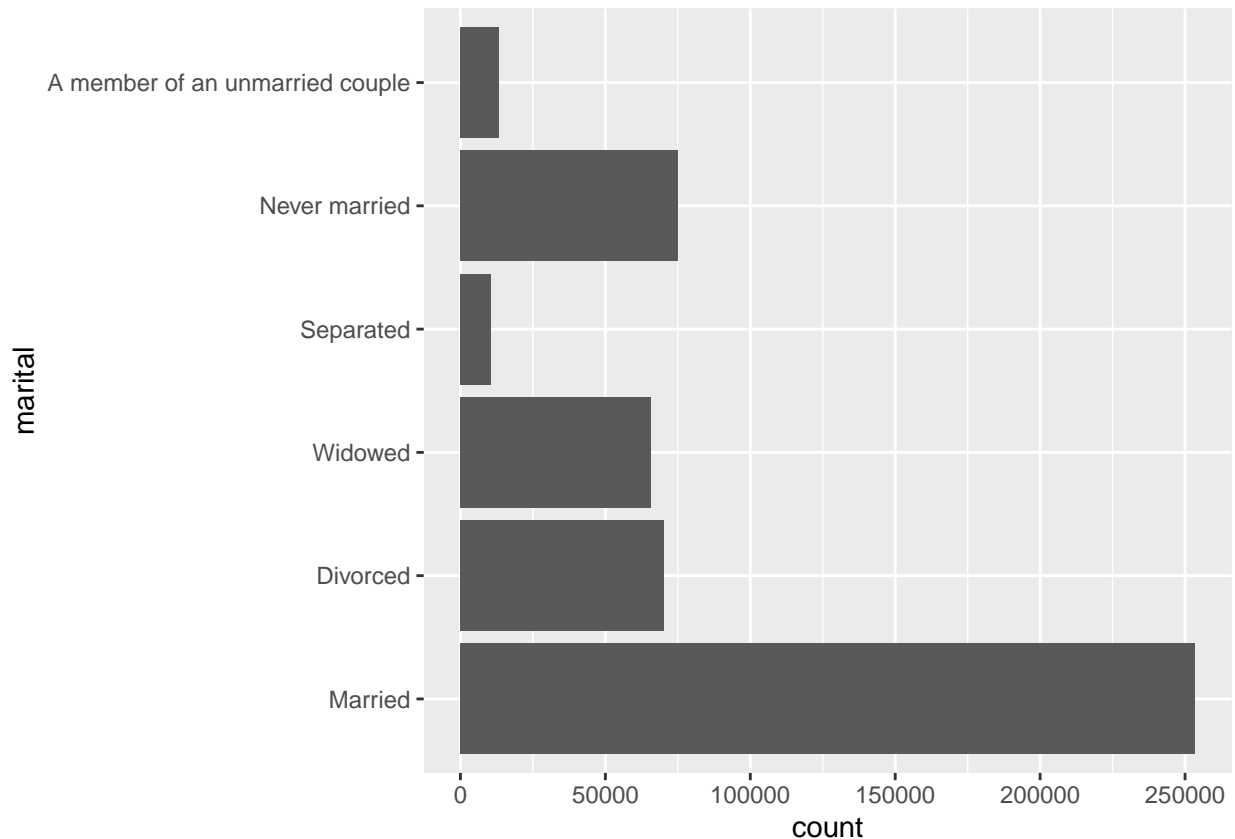
```
## # A tibble: 7 x 4
##   marital                countgp total percentage
##   <fct>                <int>   <int>      <dbl>
## 1 Married              253329 491775      51.5
## 2 Divorced              70376 491775      14.3
## 3 Widowed               65745 491775      13.4
## 4 Separated            10662 491775       2.17
## 5 Never married        75070 491775      15.3
## 6 A member of an unmarried couple 13173 491775       2.68
## 7 NotDisclosed          3420 491775       0.695
```

We then plot out a contingency table that shows the percentage distribution of marital status broken down by categories.

We look at the table to see if we can drop any NotDisclosed.

We find that there are very little NotDisclosed around **0.7 %** so we drop them from the marital, category.

```
frameofinterest <- frameofinterest%>%select(c("weight2","height3","marital","X_state","income2"))%>%filter(marital != "NotDisclosed")
ggplot(frameofinterest, aes(x=marital))+geom_bar()+coord_flip()
```

Step2 : Calculate BMI using: Height, Weight variables

We now proceed to calculate the BMI, by using the : height, and weight variables.

- The formula for BMI is : $(708 \times \text{weight in lbs}) / (\text{height in inches}^2)$
- Lines 1- 4 are for calculating the number of NAs. We found that only 1.3% of heights were NAs
- Line 5 selects weight, height, marital, State and income, it excludes, those 1.3% of the rows 6670/488355 that contain NAs for heights. It also only filters heights between 300 and 711. As other heights that have suffixes of 9xxx have units in cms.
- To make analysis easier we choose to ignore those, and only focus on heights between 300 and 711 that are in units of feet.
- To convert the heights, we divide by 100 and then grep and remove 0's.
- for example: 505 becomes 5.05 (multiplying by 100) transforming to 5.5(after removing the zero) representing 5 feet 5 inches by these sequence of steps.
- The final step is to multiply this value by 12 to convert it from feet to inches.
- 5.5 feet becomes 66 inches.
- we then calculate the BMI using the formula mentioned above.
- Finally we take the resulting BMIs, and group them into: underweight, NormalWeight, Overweight and Obese categories.

```
sum(is.na(frameofinterest$weight2)) # no NAs
```

```
## [1] 0
```

```
sum(is.na(frameofinterest$height3)) # 6670 / 488355 NAs is roughly only 1.3% so we can drop those rows.
```

```
## [1] 6670
```

```
unique(frameofinterest$height3)
```

```
## [1] 507 510 504 600 503 500 505 602 601 506 502 508 501 511 NA
## [16] 509 606 410 603 604 406 411 408 402 409 605 407 607 400 401
## [31] 403 608 610 9160 609 306 9205 9125 9171 9105 9150 9167 9162 9120 9173
## [46] 9164 700 9158 9140 9183 9166 405 9157 9152 9165 9185 9153 9186 9159 9175
## [61] 9110 304 300 9168 9156 309 404 311 9155 9174 611 9178 9169 9180 9170
## [76] 9172 9145 9190 9163 9184 9187 9179 9176 9154 9132 9104 9148 9177 9116 9151
## [91] 9107 702 9130 9135 9103 703 9200 9114 9182 9189 9181 9074 9195 9071 9161
## [106] 9146 9250 9507 308 9188 9509 9506 701 9017 9194 9504 9502 705 9100 704
## [121] 206 9199 9149 9143 9108 9192 9191 303 310 9106 706 9193 803 9057 9198
## [136] 9134 9117 210 301 302 708 9144 305 9102 9147 9124 9201 9211 9206 9210
## [151] 707 9208 709
```

```
length(frameofinterest$height3)
```

```
## [1] 488355
```

```
# We see the strange values starting with 9 are values entered in cms, with 9 to indicate that fact: 9.
```

```
# we find these values are only 0.3 percent of our reading, so we drop them, instead of spending time p
```

```
#Convert Feet to inches
```

```
frameofinterest <- frameofinterest%>%select(c("weight2","height3","marital","X_state","income2"))%>%fil
mutate(inches = height3/100)%>%mutate(inches = gsub("0","",inches))%>%mutate(inches = as.double(inches
```

```
# Now we have values strictly in feet and inches, we proceed to deal with these now.
```

```
frameofinterest <- frameofinterest%>%select(c("weight2","height3","marital","inches","X_state","income2
mutate(bmi = (703*weight2)/(inches*inches))%>%filter(!is.na(bmi))
```

```
## Warning: NAs introduced by coercion
```

```
#dropping NAs as they constitute only 3.2 percent of the data.
```

```
frameofinterest$income2 <- fct_explicit_na(frameofinterest$income2, na_level = "NotDisclosed")
```

```
frameofinterest$bmicat <- frameofinterest$bmi
```

```
frameofinterest$bmicat[frameofinterest$bmicat < 18.5] <- 0
```

```
frameofinterest$bmicat[frameofinterest$bmicat >= 18.5 & frameofinterest$bmicat < 25] <- 1
```

```
frameofinterest$bmicat[frameofinterest$bmicat >= 25 & frameofinterest$bmicat < 30] <- 2
```

```
frameofinterest$bmicat[frameofinterest$bmicat >= 30] <- 3
```

```
frameofinterest$bmicat <- as.character(frameofinterest$bmicat)
```

```
frameofinterest$bmicat[frameofinterest$bmicat == "0"] <- "Underweight"
```

```
frameofinterest$bmicat[frameofinterest$bmicat == "1"] <- "Normalweight"
```

```
frameofinterest$bmicat[frameofinterest$bmicat == "2"] <- "Overweight"
```

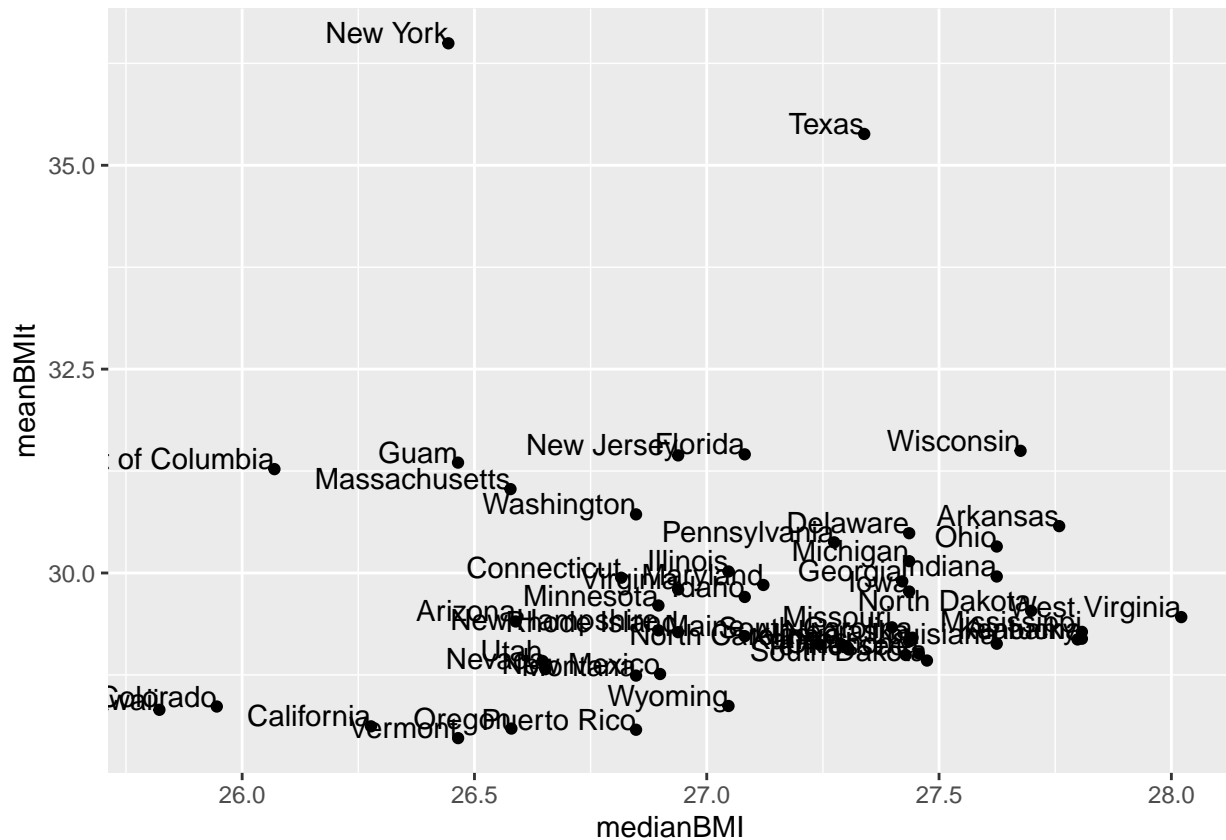
```
frameofinterest$bmicat[frameofinterest$bmicat == "3"] <- "Obese"
```

Step3 : Exploratory Data Analysis BMI versus States

- To see if there are any states with BMI getting skewed. We plot the Mean BMI versus the Median BMI.
- This shows us of any potential exceptional outliers that skew the Mean BMI of a population away from it's Median BMI.

```
#mean bmi versus median BMi per state
sBMI <- frameofinterest%>%group_by(X_state)%>%summarise(meanBMIt = mean(bmi),medianBMI = median(bmi))
#ggplot(sBMI,aes(x= medianBMI,y=meanBMIt, col=X_state))+geom_point()+facet_wrap(~bmicat)

ggplot(sBMI,aes(x= medianBMI,y=meanBMIt, label=X_state))+geom_point()+geom_text(aes(label=X_state),hjust=
```



Findings:

- **Warning** : Association is not Causation.
- From the scatter plot of median(BMI) versus the mean(BMI), it can be seen that there are 2 noticable outlier states: Texas and NewYork.
- For both these states, the Average/Mean BMI is much much larger than the Median BMI value.
- This implies a right skew in the distribution, and there must be some really super obese individuals, in both these states skewing the data to the right.
- We will have to do greater investigation, or track down the really obese individuals and study their circumstances, to see if this is a real problem in Texas and Newyork.

Research question 3:

Is there any association between Income and Depression(MentalHealth) ?

Step1 : Load and Clean up Data for analysis

We load the data , and clean it up, handling any NAs, by refactoring them

```
analysis3 <- brfss2013%>%select(c("income2","menthlth"))
analysis3$income2 <- fct_explicit_na(analysis3$income2, na_level = "NotDisclosed")
analysis3$menthlth <- as.factor(analysis3$menthlth)
analysis3$menthlth <- fct_explicit_na(analysis3$menthlth, na_level = "Not Shared")
```

We print out a consicse summary table, of what we plan to plot out. We have grouped the days based on income, and this is what will be plotted out

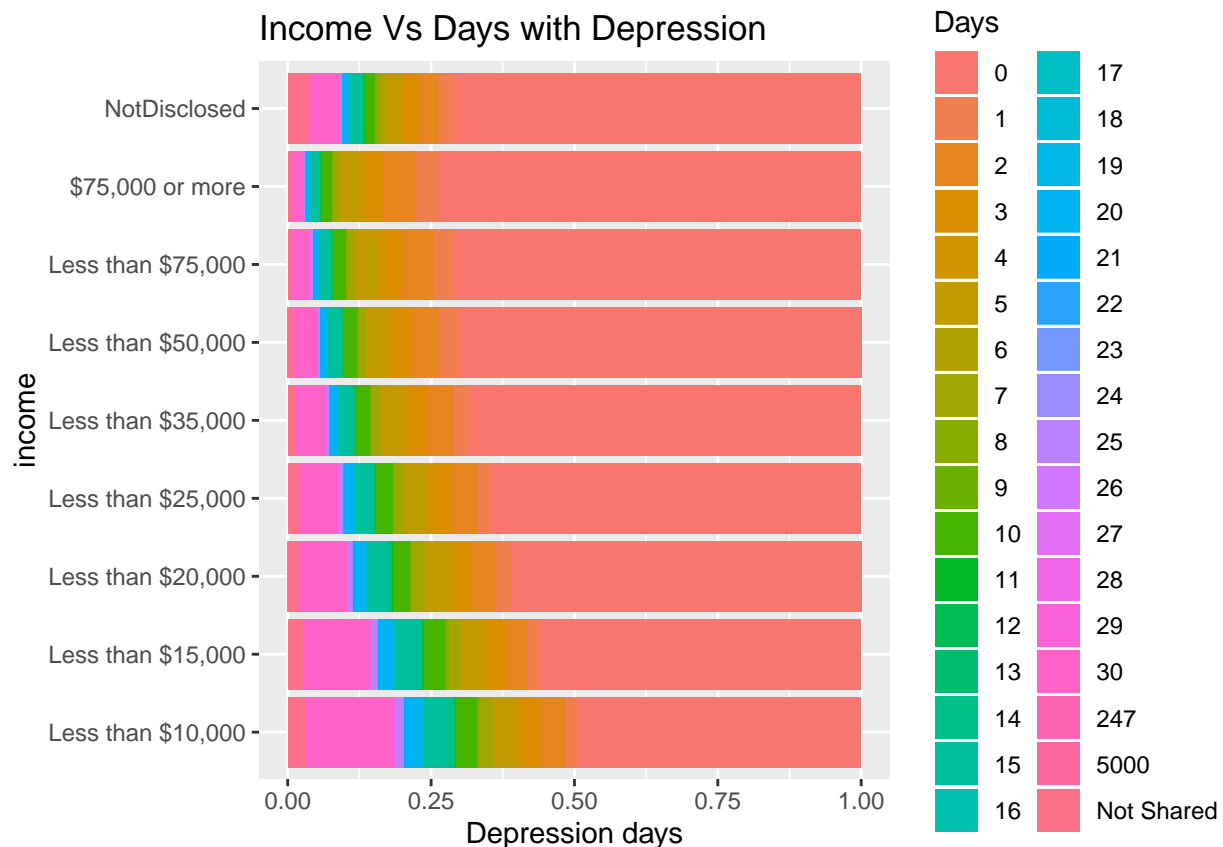
```
table(analysis3)
```

```
##               menthlth
## income2           0     1     2     3     4     5     6     7     8     9
## Less than $10,000 12587  528 1034  780  423  858  138  556  110  30
## Less than $15,000 14980  564 1074  759  376  932  139  449  118  24
## Less than $20,000 21300  858 1535  963  543 1309  152  622  117  16
## Less than $25,000 26890 1057 1894 1180  590 1450  202  667  107  25
## Less than $35,000 33295 1407 2355 1360  686 1690  201  663  114  17
## Less than $50,000 43147 1969 3048 1793  802 2169  186  713  161  21
## Less than $75,000 46230 2411 3542 1943  921 2386  217  754  138  18
## $75,000 or more   85280 4640 6408 3316 1513 4064  378 1117  209  25
## NotDisclosed      50752 1772 2630 1499  806 1796  248  812  170  22
##               menthlth
## income2           10    11    12    13    14    15    16    17    18    19
## Less than $10,000   902    7    78    8   289 1072    18    18    22    7
## Less than $15,000   988    7    87   10  229 1048    18    4    24    5
## Less than $20,000 1070    7    89   11  284 1140    17    17    17    6
## Less than $25,000 1190    9    97    7  252 1199    16    9    23    6
## Less than $35,000 1220   10    87   13  278 1094    12   14    26    4
## Less than $50,000 1520    7    79   11  249 1260    14    7    18    6
## Less than $75,000 1507   10    97    4  235 1161    17    8    13    3
## $75,000 or more    2210    5   107   13  350 1574    22   15    20    4
## NotDisclosed       1310   12    91   20  350 1362    27   13    18    4
##               menthlth
## income2           20    21    22    23    24    25    26    27    28    29
## Less than $10,000   737   61   21    7   20  280    14   29    85   53
## Less than $15,000   662   60    9    6    7  275     8   16    67   49
## Less than $20,000   753   50   13    8   15  240     9   21    61   43
## Less than $25,000   723   67    9    3    8  260    18   21    88   41
## Less than $35,000   673   54   10    6    5  251    11   19    63   49
## Less than $50,000   742   54    9    5   14  236     7   15    80   39
## Less than $75,000   671   38   13    7    7  223     6   14    68   43
## $75,000 or more     910   50   16    6    3  271     6   10    76   58
## NotDisclosed       762   63   13   13    8  282    10   19    92   40
##               menthlth
## income2           30   247  5000 Not Shared
```

##	Less than \$10,000	3836	0	0	833
##	Less than \$15,000	3070	1	0	729
##	Less than \$20,000	2833	0	0	754
##	Less than \$25,000	2837	0	0	787
##	Less than \$35,000	2429	0	0	751
##	Less than \$50,000	2431	0	0	697
##	Less than \$75,000	1996	0	0	530
##	\$75,000 or more	2421	0	0	805
##	NotDisclosed	3668	0	1	2741

Step2 : Exploratory Data Analysis between income and Depression

```
ggplot(analysis3, aes(fill = menthlth,x=income2))+geom_bar(position = "fill")+coord_flip()+ggtitle("Income Vs Days with Depression")
```



We plot the income slabs versus number of days people suffer from poor mental health/Depression. Our findings indicate a trend.

Findings :

- **Warning** : Association is not Causation.
- It can be seen that as income and depression have a reverse association. The richer a person, the less depression.
- We were trying to answer questions as to why the rich and famous commit suicide, but it looks like the general trend suggests that more money means a better state of mental/depression free life.

- Celebrities maybe outliers. Who cannot be accounted for by this plot.