Variable Manipulation

```
load("data/anes20.rda")
library(descr)
library(DescTools)
library(Hmisc)
```

```
##
## Attaching package: 'Hmisc'
```

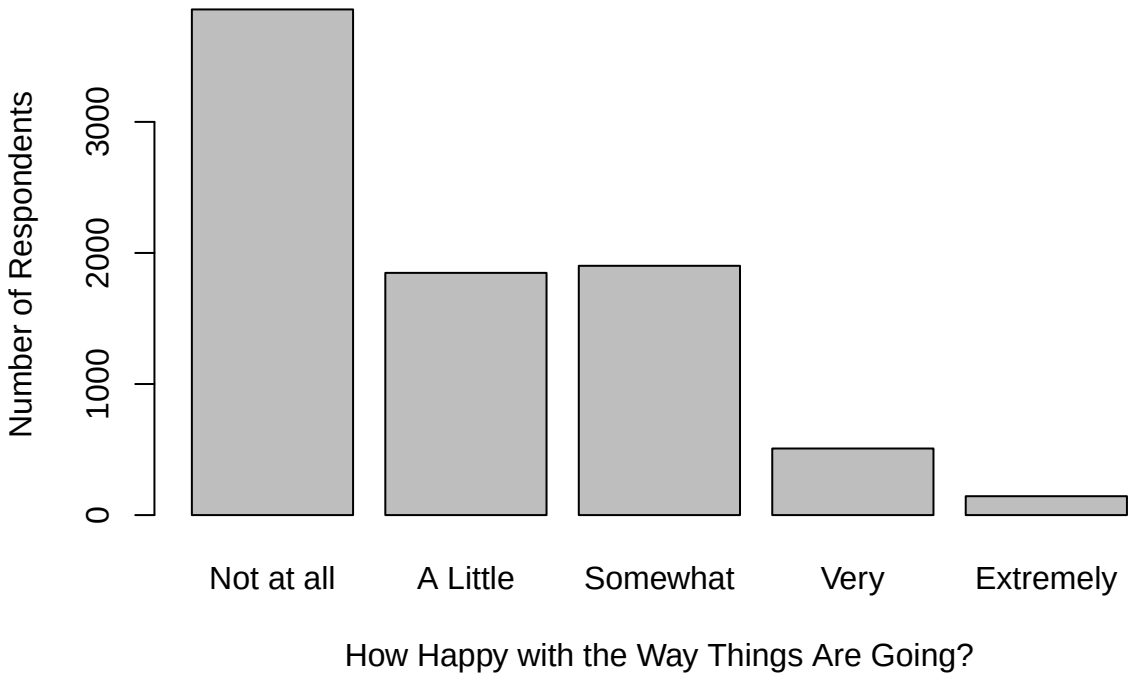
```
## The following objects are masked from 'package:DescTools':
##
##     %nin%, Label, Mean, Quantile
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

1

Fix the error that appears when creating a bar chart for `anes20$V201119`, a variable that measures how happy people are with the way things are going in the U.S.

```
happy.tbl <- table(anes20$V201119)
barplot(happy.tbl,
        names.arg=c("Not at all", "A Little",
"Somewhat", "Very", "Extremely"),
        xlab="How Happy with the Way Things Are Going?",
        ylab="Number of Respondents")
```



To avoid getting an error, I first created a table 'happy.tbl' with the results from 'anes20$V201119' and then I used the new table to create the barplot. I had to change this because the barplot function only accepts numerical or matrix values, and our column contained ordinal responses (ex. "Not at all") that the function did not recognize.

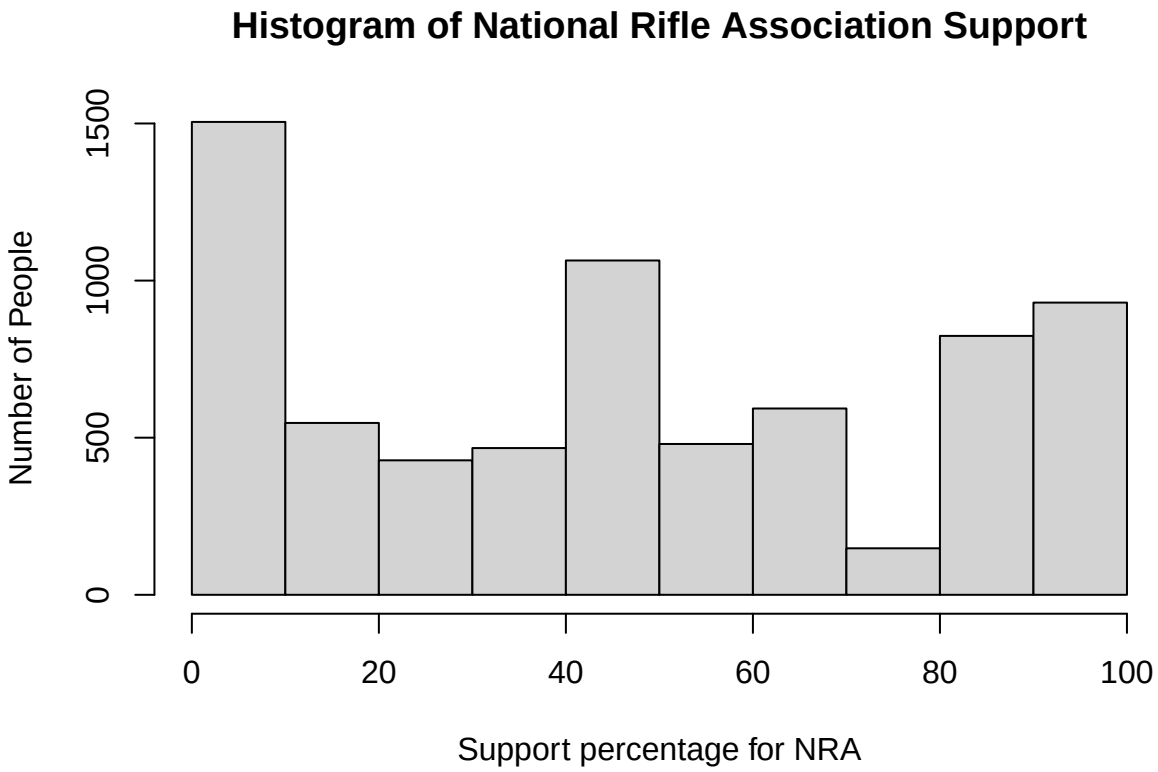
2

Create a frequency table to summarize the distribution of values for the two variables listed below. Variables: `anes20$V202178`, `anes20$V202384`

```
#variable anes20$V202178, POST: Feeling thermometer: National Rifle Association (NRA)
Freq(anes20$V202178)
```

```
##        level   freq   perc  cumfreq  cumperc
## 1     [0,10]  1'505  21.5%    1'505    21.5%
## 2    (10,20]    547   7.8%    2'052    29.4%
## 3    (20,30]    428   6.1%    2'480    35.5%
## 4    (30,40]    467   6.7%    2'947    42.2%
## 5    (40,50]  1'064  15.2%    4'011    57.4%
## 6    (50,60]    480   6.9%    4'491    64.3%
## 7    (60,70]    593   8.5%    5'084    72.8%
## 8    (70,80]    148   2.1%    5'232    74.9%
## 9    (80,90]    824  11.8%    6'056    86.7%
## 10  (90,100]    930  13.3%    6'986   100.0%
```

```r
hist(anes20$V202178,
xlab="Support percentage for NRA",
ylab="Number of People",
main="Histogram of National Rifle Association Support")
```



The variable 'anes20$V202178' shows the support of people for the National rifle Association (NRA) on a scale of 0-100 points, for each person. For this variable, I decided to use the function 'Freq' because it condensates the variables into levels, shows their frequency and their relative percentage, as well as their cumulative percentage. This is useful because it clearly shows how many people support or not support the NRA, and their relative percentage levels. In addition, since the variable was expressed in numerical data, a histogram was more useful in order to show their distribution.

```r
#variable anes20$V202384, POST: Attention to sexual harassment has gone too far or not far enough
freq(anes20$V202384, plot=FALSE)
```

```
## POST: Attention to sexual harrassment has gone too far or not far enough
##                            Frequency Percent Valid Percent
## 1. Has gone too far             1603   19.36         21.86
## 2. Has not gone far enough      2604   31.45         35.52
## 3. Has been about right         3125   37.74         42.62
## NA's                             948   11.45
## Total                           8280  100.00        100.00
```

```r
barplot(table(anes20$V202384), names.arg=c("Has gone too far", "Has not gone far enough", "Has been abou
xlab="Has it gone too far or not far enough?",
```

```
ylab="Number of people",
main = "Perception of attention to sexual harassment")
```

The variable 'anes20$V202384' shows peple's perception of attention to sexual harassment, asking them if they thought it had gone too far or not far enough. People could choose to express their opinion through oen fo three options: has gone too far, has not gone far enough, has been about right. For this variable, I used the command 'freq' because I believe that it is also important to see how many people decided not to answer the question, as it could also tell us important information regarding people's perception. In addition, since the variables were expressed in ordinal form, I decided to use a barplot while showing the data's distribution.

3

I was trying to combine responses to two gun control questions into a single, three-category variable measuring gun control attitude, `anes20$gun_cntrl`. The original variables are `anes20$V202337` (should the federal government make it more difficult or easier to buy a gun?) and `anes20$V202342` (Favor or oppose banning 'assault-style' rifles).

Fix the code and produce a frequency table for the new index and report how you fixed it.

```
anes20$buy_gun<-as.numeric(anes20$V202337=="1. More difficult")
anes20$ARguns<-as.numeric(anes20$V202342=="1. Favor")
anes20$gun_cntrl<-anes20$buy_gun + anes20$ARguns
freq(anes20$gun_cntrl, plot=F)
```

```
## anes20$gun_cntrl
```

```
##       Frequency Percent Valid Percent
## 0          2597   31.36         35.23
## 1          1743   21.05         23.64
## 2          3032   36.62         41.13
## NA's        908   10.97
## Total      8280  100.00        100.00
```

First, the variables had their most liberal response swapped, so what should have been "1. More difficult" for variable 'should the federal government make it more difficult or easier to buy a gun?', was instead given to variable 'Favor or oppose banning 'assault-style' rifles' and vice versa, so I had to interchange those two. In addition, in the frequency index command, the title of the .rda table was misspelled, as in 'anes' without the '20', therefore I fixed it and then ran the code.

4

Use the current six-category variable measuring marital status (`anes20$V201508`) and collapse it into a new variable, `anes20$marital`, with three categories, "Married", "Never Married", and "Other".

```
#create new variable
anes20$marital<-(anes20$V201508)
#Then, write over existing six labels with three new labels
levels(anes20$marital)<- c("Married", "Married", "Other",
"Other", "Other", "Never Married")
freq(anes20$marital, plot=F)
```

```
## PRE: Marital status
##               Frequency  Percent Valid Percent
## Married            4322  52.1981         52.55
## Other              1951  23.5628         23.72
## Never Married      1951  23.5628         23.72
## NA's                 56   0.6763
## Total              8280 100.0000        100.00
```

While condensing the old variables, I decided to include "1. Married: spouse present" and "2. Married: spouse absent {VOL - video/phone only}" within the new category "Married". Then I decided to include " "3. Widowed", "4. Divorced", "5. Separated" under the new category "Other", since I thought that all three of these variables could not be included in the "Still Married" and also couldn't be categorized as "Never Married". And finally, "Never Married" just remained the same category.

5

Use `V201231x` of the `anes20` data set. This variable measures respondents' party identification, using a seven-point ordinal scale, ranging from Strong Democrat to Strong Republican. Create a new variable:`anes20$ptyID.3` that includes three categories: Democrat, Independent, and Republican. Provide a frequency table of the both `anes20$V201231x` and `anes20$ptyID.3`.

```
#create new variable
anes20$ptyID.3<-(anes20$V201231x)
#Write over existing labels with new labels
levels(anes20$ptyID.3)<- c("Democrat", "Democrat", "Democrat",
"Independent", "Republican", "Republican", "Republican")
freq(anes20$ptyID.3, plot=F)
```

```
## PRE: SUMMARY: Party ID
##             Frequency  Percent Valid Percent
## Democrat         3836  46.3285         46.53
## Independent       968  11.6908         11.74
## Republican       3441  41.5580         41.73
## NA's               35   0.4227
## Total            8280 100.0000        100.00
```

```
freq(anes20$V201231x, plot=F)
```

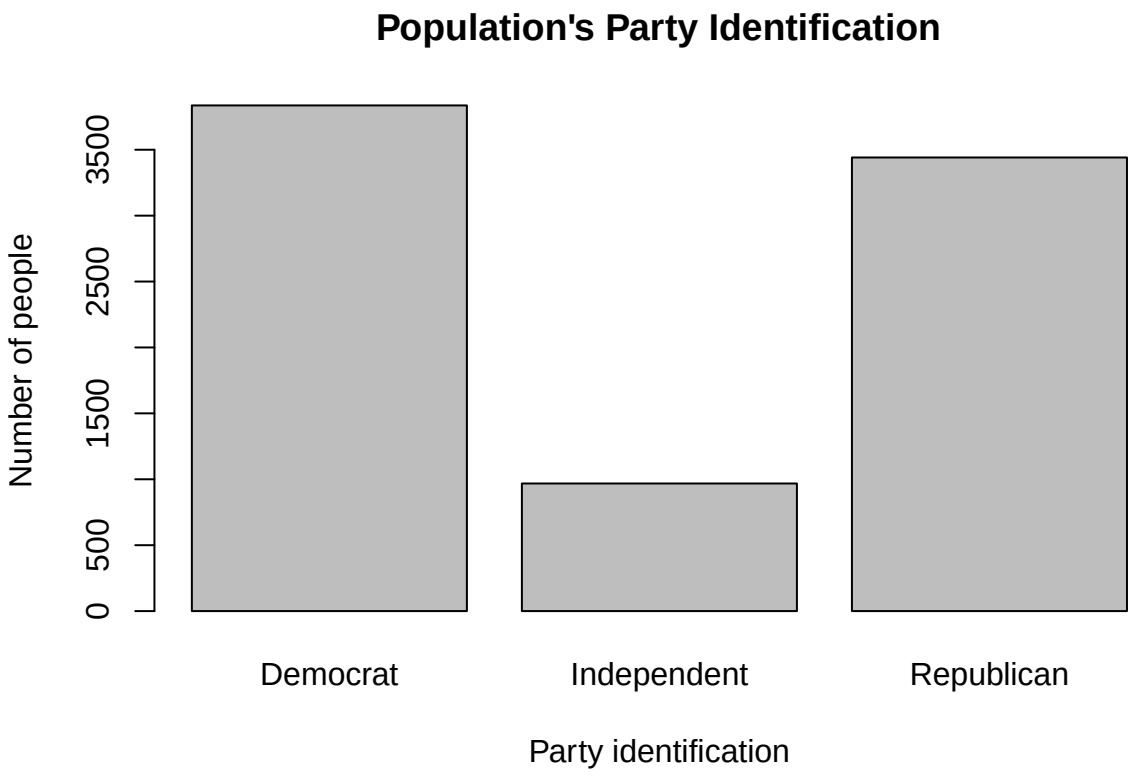
```
## PRE: SUMMARY: Party ID
##                                 Frequency  Percent Valid Percent
## 1. Strong Democrat                   1961  23.6836         23.78
## 2. Not very strong Democrat           900  10.8696         10.92
## 3. Independent-Democrat               975  11.7754         11.83
## 4. Independent                        968  11.6908         11.74
## 5. Independent-Republican             879  10.6159         10.66
## 6. Not very strong Republican         832  10.0483         10.09
## 7. Strong Republican                 1730  20.8937         20.98
## NA's                                   35   0.4227
## Total                                8280 100.0000        100.00
```

For this exercise, I decided to fit "1. Strong Democrat", "2. Not very strong Democrat", "3. Independent-Democrat" in the new variable "Democrat". Similarly, I put "Independent-Republican", "Not very strong Republican" and "Strong Republican" in the new variable "Republican." Lastly, I put "Independent" in its own original label. I thought that people who identified themselves as "Independent-other party" still somehow identified with policies and ideologies of that party, and therefore the term "Independent" would not have encompassed the effect that the adjacent party ideology would have had on the individual. Because of this reason, I decided to assign "Independent" its own variable. The table created from the original variable, 'anes20$V201231x' definitely shows a lot more nuance and is more detailed, therefore painting a more accurate picture of political affiliations. On the other hand, the first table shows the 'big picture' divisions among variables, painting a simplified version of the political spectrum. For example, only 23.68% of people consider themselves as "fully" Democrat in the original variable, however, that number goes up to 46.33% when looking at the new summary table, showing an impressive discrepancy resulting from condensing the variables.

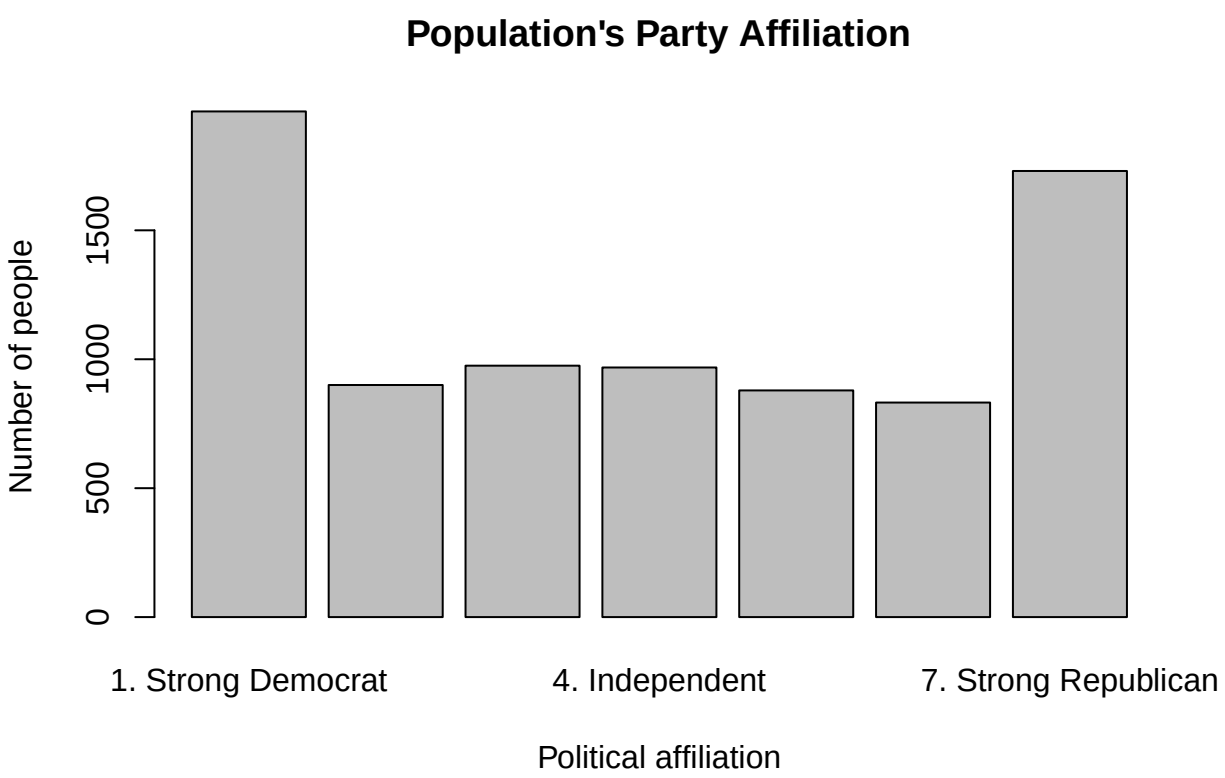
Now create bar charts for the two variables and comment on any differences in the impressions they make on you.

Do you prefer the frequency table or bar chart as a method for looking at these variables? Why?

```
barplot(table(anes20$ptyID.3), names.arg=c("Democrat", "Independent", "Republican"),
xlab="Party identification",
ylab="Number of people",
main = "Population's Party Identification")
```




```
barplot(table(anes20$V201231x),
xlab="Political affiliation",
ylab="Number of people",
main = "Population's Party Affiliation")
```



I personally don't have a strong preference between frequency table and bar chart, because I find that thay each have unique elements that paint slightly different pictures, and that therefore are both important. The frequency table, for example, is useful for looking at precise numbers of the population and for retrieving percentage numbers, including those of people that decided not to answer. On the other hand, the bar chart is more visually appealing, and with just one glance we can get a broad sense of what the division among parties is, and what the overarching trends are.

6

The table below summarizes information about four variables from the `anes20` data set that measure attitudes toward different immigration policies.

Use the information in the table to create four numeric indicator variables (one for each), and combine those variables into a new index of immigration attitudes named `anes20$immig_pol` (show all steps along the way).

Create a frequency table OR bar chart for `anes20immig_pol` and describe its distribution.

Name	Topic	Liberal Response
`V202234`	Accept Refugees	1. Favor
`V202240`	Path to Citizenship	1. Favor
`V202243`	Send back undocumented	2. Oppose
`V202246`	Separate undocumented parents/kids	2. Oppose

```r
#Create indicator variable for liberal category for "Accept Refugees"
anes20$immig_ref<-as.numeric(anes20$V202234 ==
"1. Favor")
#Create indicator variable for liberal category for "Path to Citizenship"
anes20$immig_citiz<-as.numeric(anes20$V202240 == "1. Favor")
#Create indicator variable for liberal category for "Send back undocumented"
anes20$immig_undoc<-as.numeric(anes20$V202243 == "2. Oppose")
#Create indicator variable for liberal category for "Separate undocumented parents/kids"
anes20$immig_separate<-as.numeric(anes20$V202246 == "2. Oppose")
#Create new index
anes20$immig_pol<- (anes20$immig_ref + anes20$immig_citiz + anes20$immig_undoc + anes20$immig_separate)
#Create frequency table
freq(anes20$immig_pol, plot=FALSE)
```

```
## anes20$immig_pol
##       Frequency Percent Valid Percent
## 0           759   9.167         10.32
## 1          1409  17.017         19.16
## 2          1538  18.575         20.91
## 3          1537  18.563         20.90
## 4          2111  25.495         28.71
## NA's        926  11.184
## Total      8280 100.000        100.00
```

What the frequency table of the new variable 'anes20$immig_pol' shows is that across all the four sub-variables that we incorporated to create this new one, only 9.17% of people answered with no liberal views, or, in other words, are not in favor of immigration policies, and have answered against them, through all four questions. 17.02% of people agree with one immigration policy or immigration view out of four. Respectively 18.57% and 18.56% of respondents agree with either two or three views pro-immigration. The majority, or 25.49% of respondents agreed with all four liberal pro-immigration views. Lastly, 11.19% of people in the sample did not answer an of the four questions on their views of immigration.