Lab 8

George Rhodes

October 27, 2017

Using your own dataset (which may include more than one table) carry out the following data cleaning steps. Knit together the PDF document and commit both the Lab 8 RMD file and the PDF document to Git. Push the changes to GitHub so both documents are visible in your public GitHub repository.

Before you begin: as many of you have large datasets, you're going to want to select only the variables you're interested in utilizing for this project (ideally no more than twenty columns but perhaps much smaller) so you don't have R Studio's memory working on the entire dataset. The example code provided below can be modified to allow you to subset your data to only the variables you wish to use. First, read in your complete dataset and save it as `data`. Then, add the names of the variables you wish to use for your poster project to the `select` function, separated by commas. Run the two lines of code to save this new, smaller version of your data to `data_subset`. Use this smaller dataset to complete the rest of the lab

```
#library packages!!!
library("dplyr")
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
##
## 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("tidyr")
```

```
## Warning: package 'tidyr' was built under R version 3.4.2
```

```
# Read in your data with the appropriate function
load("/Users/george/Documents/School/UW/SOC321/honors_thesis/honors_thesis/NSDUH-2015-survey-data.rda")
  names(PUF2015_102016) <- tolower(names(PUF2015_102016))

subset_nsduh2015 <- PUF2015_102016 %>%
  select(txevrrcvd, alclottm, sexage, newrace2, sexrace, eduhighcat, ireduhighst2, al30est, alcus30d, al
  # replace with variable's you wish to add
```

1. To get a feel for its structure, look at the class, dimensions, column names, structure, and basic summary statistics of your data.

```
class(subset_nsduh2015)
```

```
## [1] "data.frame"
```

```
dim(subset_nsduh2015)
```

```
## [1] 57146    16
```

```r
names(subset_nsduh2015)
```

```
##  [1] "txevrrcvd"    "alclottm"     "sexage"       "newrace2"
##  [5] "sexrace"      "eduhighcat"   "ireduhighst2" "al30est"
##  [9] "alcus30d"     "alcbng30d"    "irpinc3"      "irfamin3"
## [13] "poverty3"     "coutyp2"      "alcwd2sx"     "alcemopb"
```

```r
str(subset_nsduh2015)
```

```
## 'data.frame':    57146 obs. of  16 variables:
##  $ txevrrcvd   : int  2 2 2 2 2 98 2 91 91 98 ...
##  $ alclottm    : int  93 2 2 93 2 98 2 91 91 91 ...
##  $ sexage      : int  1 5 5 2 4 3 3 2 1 2 ...
##  $ newrace2    : int  1 7 1 7 1 5 7 1 1 2 ...
##  $ sexrace     : int  1 5 2 6 2 7 5 2 1 4 ...
##  $ eduhighcat  : int  5 2 4 5 3 3 2 5 5 5 ...
##  $ ireduhighst2: int  7 8 11 4 9 9 8 5 3 1 ...
##  $ al30est     : int  99 99 93 93 93 98 99 91 91 91 ...
##  $ alcus30d    : int  7 975 993 993 993 998 6 991 991 991 ...
##  $ alcbng30d   : int  1 10 93 93 93 98 2 91 91 91 ...
##  $ irpinc3     : int  1 2 1 1 1 1 4 1 1 1 ...
##  $ irfamin3    : int  1 4 1 7 2 1 4 7 7 1 ...
##  $ poverty3    : int  1 2 1 3 1 1 3 3 3 1 ...
##  $ coutyp2     : int  3 2 3 2 3 1 2 1 2 1 ...
##  $ alcwd2sx    : int  93 99 99 93 2 98 2 91 91 91 ...
##  $ alcemopb    : int  93 2 2 93 2 98 2 91 91 91 ...
##  - attr(*, "val.labels")= chr  "" "" "vl_cigever" "vl_cigofrsm" ...
##  - attr(*, "var.labels")= chr  "RESPONDENT IDENTIFICATION" "CREATION DATE OF THE DATA FILE" "EVER SM
##  - attr(*, "label.table")=List of 2666
##   ..$ : NULL
##   ..$ : NULL
##   ..$ : Named num  1 2
##   .. ..- attr(*, "names")= chr  "1 - Yes" "2 - No"
##   ..$ : Named num  1 2 3 4 94 97 98 99
##   .. ..- attr(*, "names")= chr  "1 - Definitely Yes" "2 - Probably Yes" "3 - Probably Not" "4 - Defi
##   ..$ : Named num  1 2 3 4 94 97 98 99
##   .. ..- attr(*, "names")= chr  "1 - Definitely Yes" "2 - Probably Yes" "3 - Probably Not" "4 - Defi
##   ..$ : Named num  985 991 994 997
##   .. ..- attr(*, "names")= chr  "985 - BAD DATA Logically assigned" "991 - NEVER USED CIGARETTES" "9
##   ..$ : Named num  9985 9989 9991 9994 9997 ...
##   .. ..- attr(*, "names")= chr  "9985 - BAD DATA Logically assigned" "9989 - LEGITIMATE SKIP Logicall
##   ..$ : Named num  1 2 3 4 5 6 7 8 9 10 ...
##   .. ..- attr(*, "names")= chr  "1 - January" "2 - February" "3 - March" "4 - April" ...
##   ..$ : Named num  1 2 3 4 8 9 11 14 19 29 ...
##   .. ..- attr(*, "names")= chr  "1 - Within the past 30 days" "2 - More than 30 days ago but within
##   ..$ : Named num  91 93 94 97 98
##   .. ..- attr(*, "names")= chr  "91 - NEVER USED CIGARETTES" "93 - DID NOT USE CIGARETTES IN THE PAS
##   ..$ : Named num  1 2 3 4 5 6 91 93 94 97 ...
##   .. ..- attr(*, "names")= chr  "1 - 1 or 2 days" "2 - 3 to 5 days" "3 - 6 to 9 days" "4 - 10 to 19
##   ..$ : Named num  1 2 3 4 5 6 7 91 93 94 ...
##   .. ..- attr(*, "names")= chr  "1 - Less than one cigarette per day" "2 - 1 cigarette per day" "3 -
##   ..$ : Named num  101 102 104 105 107 109 110 111 112 113 ...
##   .. ..- attr(*, "names")= chr  "101 - Basic" "102 - Benson & Hedges" "104 - Camel" "105 - Capri" ..
##   ..$ : Named num  1 2 3 4 91 93 94 97 98
```

```
##   .. ..- attr(*, "names")= chr  "1 - Lights" "2 - Ultra Lights" "3 - Mediums" "4 - Full Flavor" ...
##   ..$ : Named num  1 2 91 93 94 98
##   .. ..- attr(*, "names")= chr  "1 - Yes" "2 - No" "91 - NEVER USED CIGARETTES" "93 - DID NOT USE CI
##   ..$ : Named num  1 2 3 91 93 94 98 99
##   .. ..- attr(*, "names")= chr  "1 - Shorts" "2 - Regulars or king-sized" "3 - 100s" "91 - NEVER USE
##   ..$ : Named num  1 2 91 93 94 97 98
##   .. ..- attr(*, "names")= chr  "1 - Yes" "2 - No" "91 - NEVER USED CIGARETTES" "93 - DID NOT USE CI
##   ..$ : Named num  1 2 5 91 94 97
##   .. ..- attr(*, "names")= chr  "1 - Yes" "2 - No" "5 - Yes LOGICALLY ASSIGNED" "91 - NEVER USED CIG
##   ..$ : Named num  985 991 994 997 998 999
##   .. ..- attr(*, "names")= chr  "985 - BAD DATA Logically assigned" "991 - NEVER USED CIGARETTES" "9
##   ..$ : Named num  9985 9989 9991 9994 9997 ...
##   .. ..- attr(*, "names")= chr  "9985 - BAD DATA Logically assigned" "9989 - LEGITIMATE SKIP Logicall
##   ..$ : Named num  1 2 3 4 5 6 7 8 9 10 ...
##   .. ..- attr(*, "names")= chr  "1 - January" "2 - February" "3 - March" "4 - April" ...
##   ..$ : Named num  1 2 5 91 94 97
##   .. ..- attr(*, "names")= chr  "1 - Yes" "2 - No" "5 - Yes LOGICALLY ASSIGNED" "91 - NEVER USED CIG
##   ..$ : Named num  1 2 94 97
##   .. ..- attr(*, "names")= chr  "1 - Yes" "2 - No" "94 - DON T KNOW" "97 - REFUSED"
##   ..$ : Named num  985 991 994 997 998
##   .. ..- attr(*, "names")= chr  "985 - BAD DATA Logically assigned" "991 - NEVER USED SMOKELESS TOBA
##   ..$ : Named num  9985 9989 9991 9994 9997 ...
##   .. ..- attr(*, "names")= chr  "9985 - BAD DATA Logically assigned" "9989 - LEGITIMATE SKIP Logicall
##   ..$ : Named num  1 2 3 4 5 6 7 8 9 10 ...
##   .. ..- attr(*, "names")= chr  "1 - January" "2 - February" "3 - March" "4 - April" ...
##   ..$ : Named num  1 2 3 4 8 9 11 14 19 29 ...
##   .. ..- attr(*, "names")= chr  "1 - Within the past 30 days" "2 - More than 30 days ago but within
##   ..$ : Named num  91 93 94 97 98
##   .. ..- attr(*, "names")= chr  "91 - NEVER USED SMOKELESS TOBACCO" "93 - DID NOT USE SMOKELESS TOBA
##   ..$ : Named num  1 2 3 4 5 6 91 93 94 97 ...
##   .. ..- attr(*, "names")= chr  "1 - 1 or 2 days" "2 - 3 to 5 days" "3 - 6 to 9 days" "4 - 10 to 19
##   ..$ : Named num  1 2 94 97
##   .. ..- attr(*, "names")= chr  "1 - Yes" "2 - No" "94 - DON T KNOW" "97 - REFUSED"
##   ..$ : Named num  985 991 994 997 998
##   .. ..- attr(*, "names")= chr  "985 - BAD DATA Logically assigned" "991 - NEVER USED CIGARS" "994 -
##   ..$ : Named num  9985 9989 9991 9994 9997 ...
##   .. ..- attr(*, "names")= chr  "9985 - BAD DATA Logically assigned" "9989 - LEGITIMATE SKIP Logicall
##   ..$ : Named num  1 2 3 4 5 6 7 8 9 10 ...
##   .. ..- attr(*, "names")= chr  "1 - January" "2 - February" "3 - March" "4 - April" ...
##   ..$ : Named num  1 2 3 4 8 9 11 14 19 29 ...
##   .. ..- attr(*, "names")= chr  "1 - Within the past 30 days" "2 - More than 30 days ago but within
##   ..$ : Named num  91 93 94 97 98
##   .. ..- attr(*, "names")= chr  "91 - NEVER USED CIGARS" "93 - DID NOT USE CIGARS IN THE PAST 30 DAY
##   ..$ : Named num  1 2 3 4 5 91 93 94 97 98 ...
##   .. ..- attr(*, "names")= chr  "1 - 1 or 2 days" "2 - 3 to 5 days" "3 - 6 to 9 days" "4 - 10 to 19
##   ..$ : Named num  112 118 401 402 404 405 408 409 411 412 ...
##   .. ..- attr(*, "names")= chr  "112 - Marlboro" "118 - Newport" "401 - Antonio y Cleopatra" "402 -
##   ..$ : Named num  1 2 94 97
##   .. ..- attr(*, "names")= chr  "1 - Yes" "2 - No" "94 - DON T KNOW" "97 - REFUSED"
##   ..$ : Named num  1 2 91 94 97 98
##   .. ..- attr(*, "names")= chr  "1 - Yes" "2 - No" "91 - NEVER USED PIPE TOBACCO" "94 - DON T KNOW"
##   ..$ : Named num  1 2 85 94 97
##   .. ..- attr(*, "names")= chr  "1 - Yes" "2 - No" "85 - BAD DATA Logically assigned" "94 - DON T KN
##   ..$ : Named num  985 991 994 997 998
```

```
##   .. ..- attr(*, "names")= chr  "985 - BAD DATA Logically assigned" "991 - NEVER USED ALCOHOL" "994 -
##   ..$ : Named num  9985 9989 9991 9994 9997 ...
##   .. ..- attr(*, "names")= chr  "9985 - BAD DATA Logically assigned" "9989 - LEGITIMATE SKIP Logicall
##   ..$ : Named num  1 2 3 4 5 6 7 8 9 10 ...
##   .. ..- attr(*, "names")= chr  "1 - January" "2 - February" "3 - March" "4 - April" ...
##   ..$ : Named num  1 2 3 8 9 11 85 91 97 98
##   .. ..- attr(*, "names")= chr  "1 - Within the past 30 days" "2 - More than 30 days ago but within
##   ..$ : Named num  985 991 993 994 997 998
##   .. ..- attr(*, "names")= chr  "985 - BAD DATA Logically assigned" "991 - NEVER USED ALCOHOL" "993 -
##   ..$ : Named num  1 2 98
##   .. ..- attr(*, "names")= chr  "1 - Trimmed to 365 days" "2 - Trimmed relative to the 30-day freq"
##   ..$ : Named num  1 98
##   .. ..- attr(*, "names")= chr  "1 - Trimmed to be consistent with mo/yr of 1st use" "98 - BLANK"
##   ..$ : Named num  1 2 3 11 12 13 85 91 93 94 ...
##   .. ..- attr(*, "names")= chr  "1 - Prefer to answer in days per week" "2 - Prefer to answer in days
##   ..$ : Named num  985 989 991 993 994 997 998 999
##   .. ..- attr(*, "names")= chr  "985 - BAD DATA Logically assigned" "989 - LEGITIMATE SKIP Logically
##   ..$ : Named num  85 89 91 93 94 97 98 99
##   .. ..- attr(*, "names")= chr  "85 - BAD DATA Logically assigned" "89 - LEGITIMATE SKIP Logically as
##   ..$ : Named num  85 91 93 94 97 98 99
##   .. ..- attr(*, "names")= chr  "85 - BAD DATA Logically assigned" "91 - NEVER USED ALCOHOL" "93 - DI
##   ..$ : Named num  85 91 93 94 97 98
##   .. ..- attr(*, "names")= chr  "85 - BAD DATA Logically assigned" "91 - NEVER USED ALCOHOL" "93 - DI
##   ..$ : Named num  1 2 3 4 5 6 85 91 93 94 ...
##   .. ..- attr(*, "names")= chr  "1 - 1 or 2 days" "2 - 3 to 5 days" "3 - 6 to 9 days" "4 - 10 to 19 d
##   ..$ : Named num  1 98
##   .. ..- attr(*, "names")= chr  "1 - Edited for consistency with ALCYRTOT or ALCBNG30D" "98 - BLANK"
##   ..$ : Named num  975 985 991 993 994 997 998
##   .. ..- attr(*, "names")= chr  "975 - AT LEAST 4 OR 5 Logically assigned" "985 - BAD DATA Logically
##   ..$ : Named num  80 85 91 93 94 97 98
##   .. ..- attr(*, "names")= chr  "80 - NO OCCAS OF 4+ or 5+ DRINKS PST 30 DAYS Log assn" "85 - BAD DAT
##   ..$ : Named num  1 2 94 97
##   .. ..- attr(*, "names")= chr  "1 - Yes" "2 - No" "94 - DON T KNOW" "97 - REFUSED"
##   ..$ : Named num  985 991 994 997 998
##   .. ..- attr(*, "names")= chr  "985 - BAD DATA Logically assigned" "991 - NEVER USED MARIJUANA" "994
##   ..$ : Named num  9985 9989 9991 9994 9997 ...
##   .. ..- attr(*, "names")= chr  "9985 - BAD DATA Logically assigned" "9989 - LEGITIMATE SKIP Logicall
##   ..$ : Named num  1 2 3 4 5 6 7 8 9 10 ...
##   .. ..- attr(*, "names")= chr  "1 - January" "2 - February" "3 - March" "4 - April" ...
##   ..$ : Named num  1 2 3 8 9 11 91 97 98
##   .. ..- attr(*, "names")= chr  "1 - Within the past 30 days" "2 - More than 30 days ago but within
##   ..$ : Named num  985 991 993 994 997 998
##   .. ..- attr(*, "names")= chr  "985 - BAD DATA Logically assigned" "991 - NEVER USED MARIJUANA" "993
##   ..$ : Named num  1 2 98
##   .. ..- attr(*, "names")= chr  "1 - Trimmed to 365 days" "2 - Trimmed relative to the 30-day freq"
##   ..$ : Named num  1 98
##   .. ..- attr(*, "names")= chr  "1 - Trimmed to be consistent with mo/yr of 1st use" "98 - BLANK"
##   ..$ : Named num  1 2 3 11 12 13 85 91 93 94 ...
##   .. ..- attr(*, "names")= chr  "1 - Prefer to answer in days per week" "2 - Prefer to answer in days
##   ..$ : Named num  985 989 991 993 994 997 998 999
##   .. ..- attr(*, "names")= chr  "985 - BAD DATA Logically assigned" "989 - LEGITIMATE SKIP Logically
##   ..$ : Named num  85 89 91 93 94 97 98 99
##   .. ..- attr(*, "names")= chr  "85 - BAD DATA Logically assigned" "89 - LEGITIMATE SKIP Logically as
##   ..$ : Named num  85 91 93 94 97 98 99
```

```
##   .. ..- attr(*, "names")= chr  "85 - BAD DATA Logically assigned" "91 - NEVER USED MARIJUANA" "93 -
##   ..$ : Named num  85 91 93 94 97 98
##   .. ..- attr(*, "names")= chr  "85 - BAD DATA Logically assigned" "91 - NEVER USED MARIJUANA" "93 -
##   ..$ : Named num  1 2 3 4 5 6 91 93 94 97 ...
##   .. ..- attr(*, "names")= chr  "1 - 1 or 2 days" "2 - 3 to 5 days" "3 - 6 to 9 days" "4 - 10 to 19 d
##   ..$ : Named num  1 2 94 97
##   .. ..- attr(*, "names")= chr  "1 - Yes" "2 - No" "94 - DON T KNOW" "97 - REFUSED"
##   ..$ : Named num  985 991 994 997 998
##   .. ..- attr(*, "names")= chr  "985 - BAD DATA Logically assigned" "991 - NEVER USED COCAINE" "994 -
##   ..$ : Named num  9985 9989 9991 9994 9997 ...
##   .. ..- attr(*, "names")= chr  "9985 - BAD DATA Logically assigned" "9989 - LEGITIMATE SKIP Logicall
##   ..$ : Named num  1 2 3 4 5 6 7 8 9 10 ...
##   .. ..- attr(*, "names")= chr  "1 - January" "2 - February" "3 - March" "4 - April" ...
##   ..$ : Named num  1 2 3 8 9 11 12 91 97 98
##   .. ..- attr(*, "names")= chr  "1 - Within the past 30 days" "2 - More than 30 days ago but within t
##   ..$ : Named num  985 991 993 994 997 998
##   .. ..- attr(*, "names")= chr  "985 - BAD DATA Logically assigned" "991 - NEVER USED COCAINE" "993 -
##   ..$ : Named num  1 2 98
##   .. ..- attr(*, "names")= chr  "1 - Trimmed to 365 days" "2 - Trimmed relative to the 30-day freq" "
##   ..$ : Named num  1 98
##   .. ..- attr(*, "names")= chr  "1 - Trimmed to be consistent with mo/yr of 1st use" "98 - BLANK"
##   ..$ : Named num  1 2 3 12 13 21 22 23 85 91 ...
##   .. ..- attr(*, "names")= chr  "1 - Prefer to answer in days per week" "2 - Prefer to answer in days
##   ..$ : Named num  985 989 991 993 994 997 998 999
##   .. ..- attr(*, "names")= chr  "985 - BAD DATA Logically assigned" "989 - LEGITIMATE SKIP Logically
##   ..$ : Named num  85 91 93 94 97 98 99
##   .. ..- attr(*, "names")= chr  "85 - BAD DATA Logically assigned" "91 - NEVER USED COCAINE" "93 - DI
##   ..$ : Named num  85 89 91 93 94 97 98 99
##   .. ..- attr(*, "names")= chr  "85 - BAD DATA Logically assigned" "89 - LEGITIMATE SKIP Logically as
##   ..$ : Named num  85 91 93 94 97 98
##   .. ..- attr(*, "names")= chr  "85 - BAD DATA Logically assigned" "91 - NEVER USED COCAINE" "93 - DI
##   ..$ : Named num  1 3 91 93 97 98 99
##   .. ..- attr(*, "names")= chr  "1 - 1 or 2 days" "3 - 6 to 9 days" "91 - NEVER USED COCAINE" "93 - N
##   ..$ : Named num  1 2 91 94 97 98
##   .. ..- attr(*, "names")= chr  "1 - Yes" "2 - No" "91 - NEVER USED COCAINE" "94 - DON T KNOW" ...
##   ..$ : Named num  991 994 997 998
##   .. ..- attr(*, "names")= chr  "991 - NEVER USED CRACK" "994 - DON T KNOW" "997 - REFUSED" "998 - BL
##   ..$ : Named num  9985 9989 9991 9994 9997 ...
##   .. ..- attr(*, "names")= chr  "9985 - BAD DATA Logically assigned" "9989 - LEGITIMATE SKIP Logicall
##   ..$ : Named num  1 2 3 4 5 6 7 8 9 10 ...
##   .. ..- attr(*, "names")= chr  "1 - January" "2 - February" "3 - March" "4 - April" ...
##   ..$ : Named num  1 2 3 8 9 91 97 98
##   .. ..- attr(*, "names")= chr  "1 - Within the past 30 days" "2 - More than 30 days ago but within t
##   ..$ : Named num  985 991 993 994 997 998
##   .. ..- attr(*, "names")= chr  "985 - BAD DATA Logically assigned" "991 - NEVER USED CRACK" "993 - N
##   ..$ : Named num  2 98
##   .. ..- attr(*, "names")= chr  "2 - Trimmed relative to the 30-day freq" "98 - BLANK"
##   ..$ : Named num  1 98
##   .. ..- attr(*, "names")= chr  "1 - Trimmed to be consistent with mo/yr of 1st use" "98 - BLANK"
##   ..$ : Named num  1 2 3 12 85 91 93 94 97 98
##   .. ..- attr(*, "names")= chr  "1 - Prefer to answer in days per week" "2 - Prefer to answer in days
##   ..$ : Named num  989 991 993 994 997 998 999
##   .. ..- attr(*, "names")= chr  "989 - LEGITIMATE SKIP Logically assigned" "991 - NEVER USED CRACK"
##   ..$ : Named num  85 91 93 94 97 98 99
```

```
##   .. ..- attr(*, "names")= chr  "85 - BAD DATA Logically assigned" "91 - NEVER USED CRACK" "93 - DID
##   ..$ : Named num  85 91 93 94 97 98 99
##   .. ..- attr(*, "names")= chr  "85 - BAD DATA Logically assigned" "91 - NEVER USED CRACK" "93 - DID
##   ..$ : Named num  85 91 93 97 98
##   .. ..- attr(*, "names")= chr  "85 - BAD DATA Logically assigned" "91 - NEVER USED CRACK" "93 - DID
##   ..$ : Named num  91 93 97 98 99
##   .. ..- attr(*, "names")= chr  "91 - NEVER USED CRACK" "93 - DID NOT USE CRACK IN THE PAST 30 DAYS"
##   ..$ : Named num  1 2 94 97
##   .. ..- attr(*, "names")= chr  "1 - Yes" "2 - No" "94 - DON T KNOW" "97 - REFUSED"
##   .. [list output truncated]
```

```
summary(subset_nsduh2015)
```

```
##    txevrrcvd        alclottm         sexage         newrace2
##  Min.   : 1.00   Min.   : 1.00   Min.   :1.000   Min.   :1.000
##  1st Qu.: 2.00   1st Qu.: 2.00   1st Qu.:3.000   1st Qu.:1.000
##  Median : 2.00   Median : 2.00   Median :5.000   Median :1.000
##  Mean   :23.58   Mean   :45.94   Mean   :3.793   Mean   :2.666
##  3rd Qu.: 2.00   3rd Qu.:91.00   3rd Qu.:5.000   3rd Qu.:5.000
##  Max.   :98.00   Max.   :98.00   Max.   :5.000   Max.   :7.000
##
##     sexrace        eduhighcat     ireduhighst2       al30est
##  Min.   :1.000   Min.   :1.000   Min.   : 1.000   Min.   : 1.00
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.: 6.000   1st Qu.:91.00
##  Median :2.000   Median :3.000   Median : 8.000   Median :93.00
##  Mean   :3.103   Mean   :3.241   Mean   : 7.738   Mean   :95.04
##  3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:10.000   3rd Qu.:99.00
##  Max.   :7.000   Max.   :5.000   Max.   :11.000   Max.   :99.00
##
##     alcus30d     alcbng30d        irpinc3         irfamin3
##  Min.   :  1   Min.   : 0.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:  2   1st Qu.: 1.00   1st Qu.:1.000   1st Qu.:3.000
##  Median :991   Median :91.00   Median :2.000   Median :5.000
##  Mean   :566   Mean   :51.94   Mean   :2.579   Mean   :4.748
##  3rd Qu.:993   3rd Qu.:93.00   3rd Qu.:4.000   3rd Qu.:7.000
##  Max.   :998   Max.   :98.00   Max.   :7.000   Max.   :7.000
##
##     poverty3        coutyp2         alcwd2sx        alcemopb
##  Min.   :1.000   Min.   :1.000   Min.   : 1.00   Min.   : 1.00
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.: 2.00   1st Qu.: 2.00
##  Median :3.000   Median :2.000   Median :91.00   Median : 2.00
##  Mean   :2.362   Mean   :1.764   Mean   :71.14   Mean   :46.02
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:99.00   3rd Qu.:91.00
##  Max.   :3.000   Max.   :3.000   Max.   :99.00   Max.   :98.00
##  NA's   :417
```

2. Preview the first and last 15 rows of your data. Is you dataset tidy? If not, what principles of tidy data does it seem to be violating?

```
head(subset_nsduh2015, n = 15)
```

```
##    txevrrcvd alclottm sexage newrace2 sexrace eduhighcat ireduhighst2
## 1          2       93      1        1       1          5            7
## 2          2        2      5        7       5          2            8
## 3          2        2      5        1       2          4           11
## 4          2       93      2        7       6          5            4
```

```
## 5          2        2      4        1       2          3            9
## 6         98       98      3        5       7          3            9
## 7          2        2      3        7       5          2            8
## 8         91       91      2        1       2          5            5
## 9         91       91      1        1       1          5            3
## 10        98       91      2        2       4          5            1
## 11         2       91      2        7       6          5            6
## 12        91       91      1        7       5          5            4
## 13         2        2      5        1       2          3           10
## 14         2        2      4        4       7          2            8
## 15         2        2      5        7       5          3            9
##    al30est alcus30d alcbng30d irpinc3 irfamin3 poverty3 coutyp2 alcwd2sx
## 1       99        7         1       1        1        1       3       93
## 2       99      975        10       2        4        2       2       99
## 3       93      993        93       1        1        1       3       99
## 4       93      993        93       1        7        3       2       93
## 5       93      993        93       1        2        1       3        2
## 6       98      998        98       1        1        1       1       98
## 7       99        6         2       4        4        3       2        2
## 8       91      991        91       1        7        3       1       91
## 9       91      991        91       1        7        3       2       91
## 10      91      991        91       1        1        1       1       91
## 11      91      991        91       1        2        1       1       91
## 12      91      991        91       1        6        3       1       91
## 13      99        1         0       4        7        3       1        2
## 14      93      993        93       1        1        1       3       97
## 15      99        2         0       4        7        3       1       99
##    alcemopb
## 1        93
## 2         2
## 3         2
## 4        93
## 5         2
## 6        98
## 7         2
## 8        91
## 9        91
## 10       91
## 11       91
## 12       91
## 13        2
## 14       97
## 15        2
```

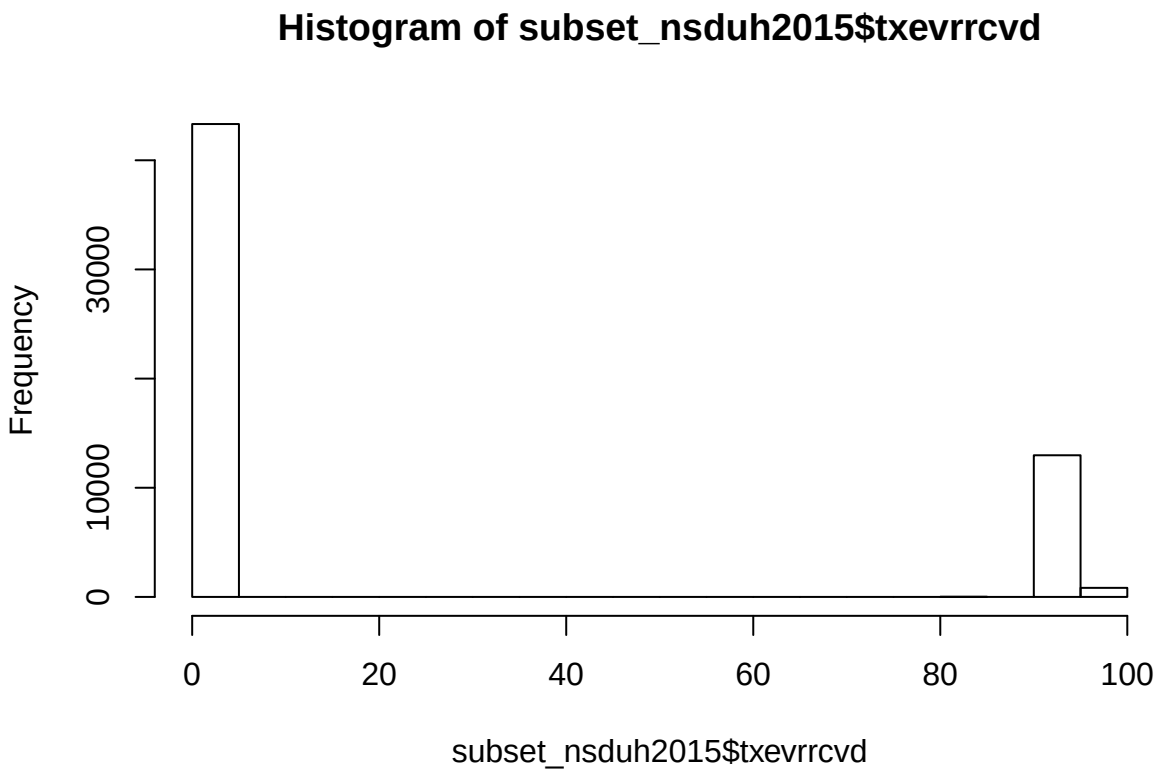
```
tail(subset_nsduh2015, n = 15)
```

```
##       txevrrcvd alclottm sexage newrace2 sexrace eduhighcat ireduhighst2
## 57132         2        2      5        1       2          3            9
## 57133         2        2      4        1       2          3            9
## 57134        91       91      1        1       1          5            9
## 57135         2        2      5        1       1          1            1
## 57136         1        1      5        1       2          2            8
## 57137         2       93      2        7       6          5            6
## 57138         2       91      2        7       6          5            4
## 57139         2        1      4        1       2          3            9
```

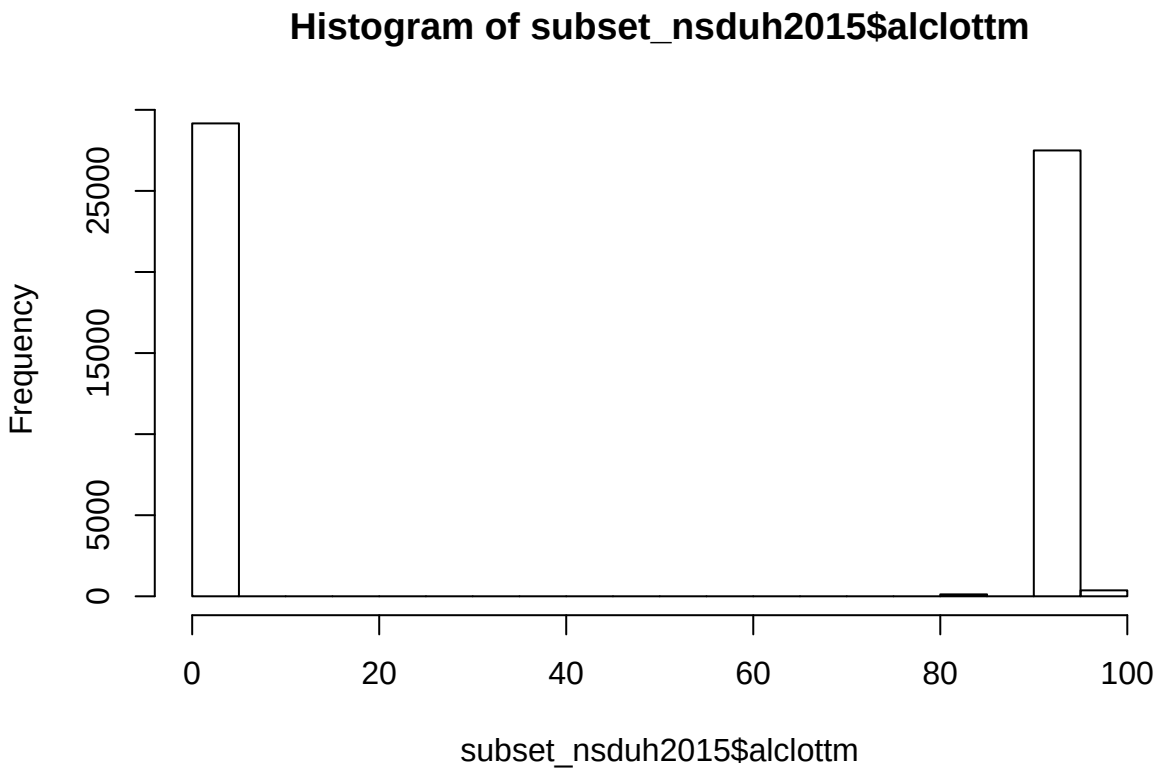
```
## 57140         2        2      5        7       6          3            9
## 57141        91       91      2        6       7          5            1
## 57142         2        2      5        1       1          4           11
## 57143         2       93      3        7       5          2            8
## 57144         2        2      4        5       7          3            9
## 57145         2        2      1        6       7          5            4
## 57146        91       91      2        1       2          5            5
##       al30est alcus30d alcbng30d irpinc3 irfamin3 poverty3 coutyp2
## 57132      99        2         0       2        4        2       3
## 57133      93      993        93       1        5        3       2
## 57134      91      991        91       1        7        3       1
## 57135      99        5         5       2        3        2       3
## 57136      93      993        93       3        3        2       3
## 57137      93      993        93       1        3        2       1
## 57138      91      991        91       1        2        1       2
## 57139      93      993        93       1        1        1       1
## 57140      93      993        93       2        3        1       1
## 57141      91      991        91       1        2        1       1
## 57142      99        1         0       6        6        3       2
## 57143      99        1         0       2        6        3       1
## 57144      99        2         2       1        1        1       2
## 57145      99        2         0       1        4        2       2
## 57146      91      991        91       1        7        3       2
##       alcwd2sx alcemopb
## 57132       99        2
## 57133        2        2
## 57134       91       91
## 57135        2        2
## 57136        2        2
## 57137       93       93
## 57138       91       91
## 57139        1        2
## 57140        2        2
## 57141       91       91
## 57142       99        2
## 57143       93       93
## 57144        2        2
## 57145        1        2
## 57146       91       91
```

3. Create a histogram for at least two variables you plan to focus on for your study. Describe what these plots show you about these variables.

```
hist(subset_nsduh2015$txevrrcvd)
```




```
hist(subset_nsduh2015$alclottm)
```



hist(subset_nsduh2015$ireduhighst2)hist(subset_nsduh2015poverty3) hist(subset_nsduh2015$irfamin3)

These plots don't show much because of the coding on the survey–other than the need for more cleaning.

4. Create at least one bivariate plot showing the relationship between two variables of interest. What does/do the(se) plot(s) tell you about the association between these two variables?

plot(subset_nsduh2015$ireduhighst2, subset_nsduh2015$irfamin3)

Hard to tell given the coding of the data, as it just looks like a uniform spread. However, it is a very even distribution, and this may eventually reveal a positive correlation between level of education and family income.

5. Load the `tidyr` package. Do all of your columns correspond to variables? Do any columns represent multiple variables? If your answer is yes to either question, carry out the appropriate `tidyr` function (`gather()` or `spread()` respectively) to tidy your data.

install.packages("tidyr") library("tidyr")

I believe they all correspond to a single variable (a question on a survey).

6. Do any columns need to be separated into two or more? Do any columns need to be combined into one? If so, carry out the appropriate the appropriate `tidyr` function (`separate()` or `unite()` respectively) to tidy your data.

I would like to combine certain answers from 2 columns, but I don't think this is quite where I should do that.

At this stage each row in your data should represent one observation, each column should be a variable, and each table should be observational unit.

7. What is the class of each of the variables in your analysis? Are these classes appropriate for the type of measurement they purport to capture? Explain your reasoning.

They are integars. This makes sense for coding survey data, however, the current format is not representative of how I would like to view and analyze the data.

8. Do any of your variables need to be coerced into a different data type? If so, carry out the appropriate coercion methods below. (This includes transformation of any date objects using the `lubridate` package)

I don't think so.

9. Are there any strings you need to manipulate for your analysis? If so, use the appropriate function from the `stringr` package.

I don't think so.

10. Do you have any missing values in your dataset? How many and how are they coded? **Be sure to look out for specific codebook values for missing values (i.e. `-1` for `NA`) as well as empty strings or other software-specific values for `NA`.** Don't worry about removing NAs yet - we'll tackle this question later once discern whether they're random or systematically distributed.

The code book addresses this. There are option for "refused" "intentionally left blank" and "don't know." While all of these need to be addresses, to interpret the data, they have already coded and acounted for missing values.

11. Are there any special values in your dataset? If so, what are they and how do you think they got there? *The presence of special values is less likely if you haven't performed any data manipulation yet so you should remember to return to this step each time you carry out a mathematical transformation of any values in your dataset.*

12. Create a boxplot of your data (you can create an individual boxplot for each variable if there are too many variables in your dataset to meaningfully visualize them all in one plot). Are there any outliers? If so, what are they and to which variable do they correspond? Do any of these outliers seem like obvious errors? If so, why?

I don't think this will be helpful at this stage. An example of the coded survey data has values of 1,2,3,12,85,91,93,94,97,98 for a variable.

13. For any outliers and/or obvious errors, what do you think is the best way to handle them (i.e. remove them entirely, run analyses including and excluding them and compare the results, manually change them to an appropriate measure of center, or something else?).

Maybe I can change the values from INTs back to strings so that a histograms and plots will show the relative number of answers to each with the appropriate label. I didn't realize this until now. It's a little overwhelming to think about.