Lab 8

William Bernard

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(dplyr)
```

```
##
## 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
```

```
# Read in your data with the appropriate function
changing_lives <- load("C:\\Users\\William Bernard\\Desktop\\americans's_changing_lives_data_set\\ICPSR_



data_subset <- da04690.0001 %>%
  select(V104, V301) # 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.

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?

```
data_subset[1:15, ]
```

```
##    V104        V301
## 1    69 (1) COMPSAT
## 2    44 (2) VERYSAT
## 3    75 (2) VERYSAT
## 4    25 (2) VERYSAT
## 5    30 (2) VERYSAT
## 6    57 (2) VERYSAT
## 7    56 (3) SOMESAT
## 8    37 (1) COMPSAT
## 9    27 (3) SOMESAT
## 10   73 (2) VERYSAT
## 11   82 (2) VERYSAT
## 12   47 (2) VERYSAT
## 13   48 (2) VERYSAT
## 14   55 (1) COMPSAT
## 15   47 (1) COMPSAT
```

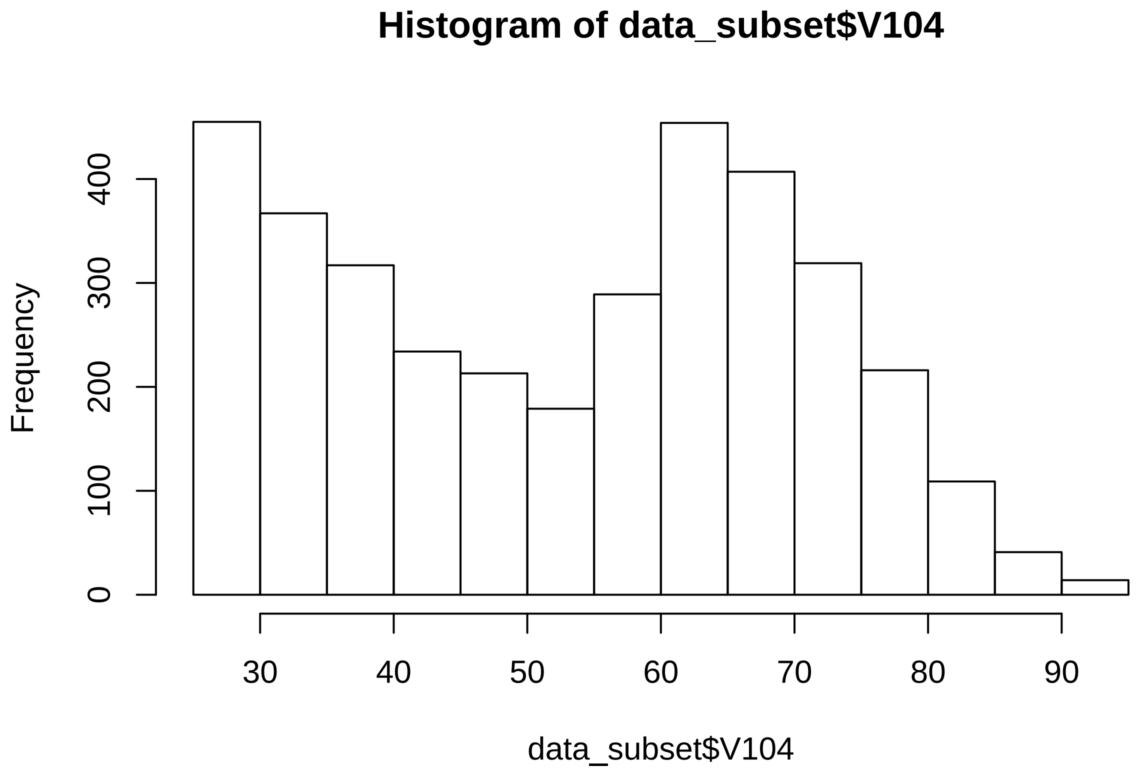
```
data_subset[3603:3617, ]
```

```
##      V104         V301
## 3603   31         <NA>
## 3604   29 (4) NVERYSAT
## 3605   81  (1) COMPSAT
## 3606   66  (2) VERYSAT
## 3607   67  (2) VERYSAT
## 3608   47  (1) COMPSAT
## 3609   40  (3) SOMESAT
## 3610   61 (4) NVERYSAT
## 3611   68  (1) COMPSAT
## 3612   43  (2) VERYSAT
## 3613   27  (3) SOMESAT
## 3614   47  (2) VERYSAT
## 3615   46  (3) SOMESAT
## 3616   59  (2) VERYSAT
## 3617   55  (2) VERYSAT
```

This data is tidy.

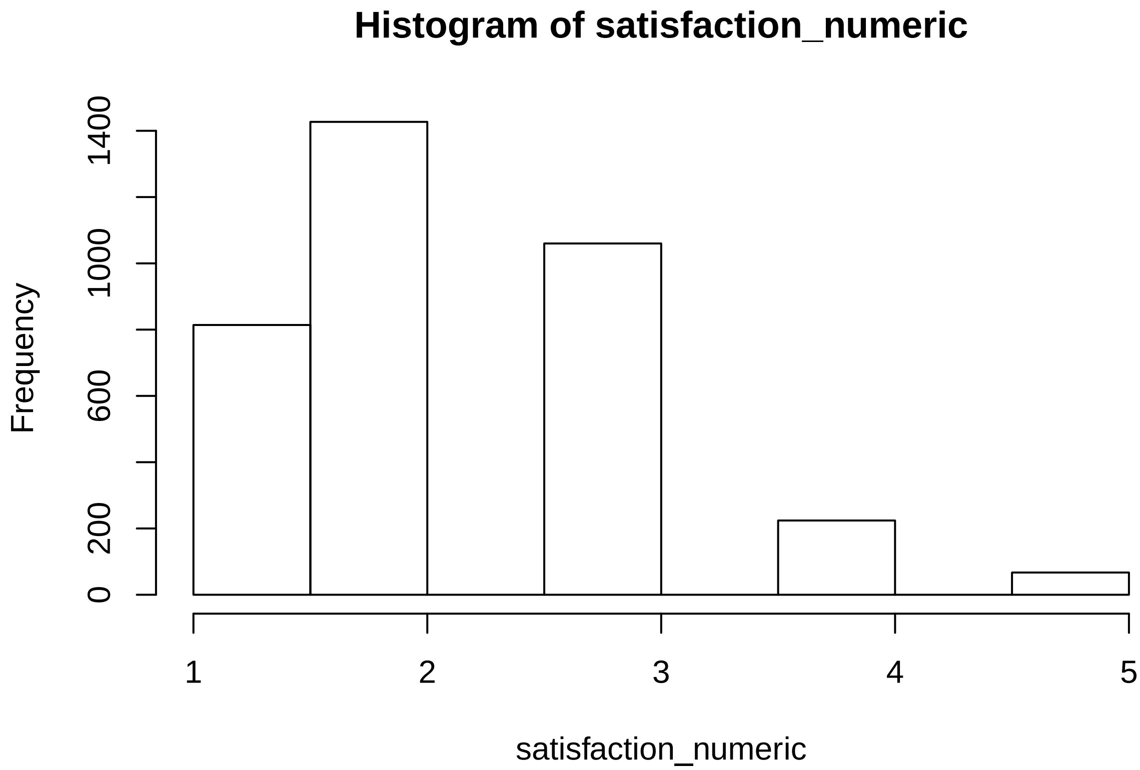
3. Create a histogram for at least two variables you plan to focus on for your study. Describe what these plots show you about these variables.

```
hist(data_subset$V104)
```



This plot tells me that a majority of the respondants are around 30 years old or around 60 years old. we have very few respondants past 85 years of age. and that mid 50 are somewhat underrepresented.

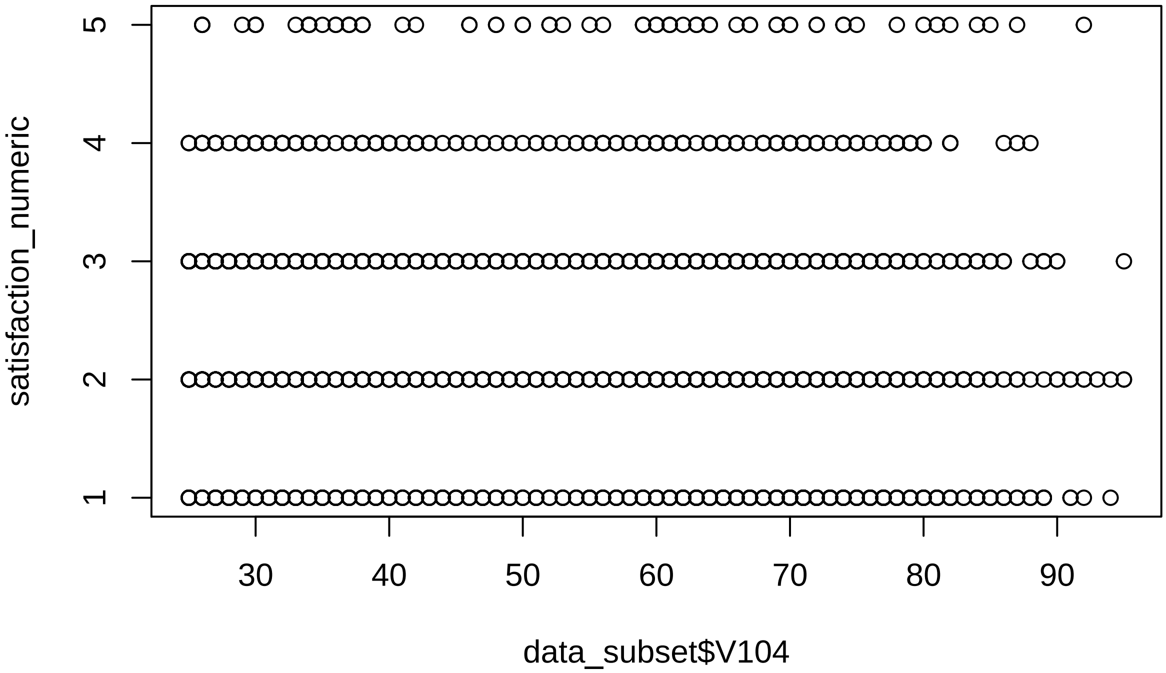
```
satisfaction_numeric <- as.numeric(data_subset$V301)
hist(satisfaction_numeric)
```



This plot gives me information on the number of people that said they were completely satisfied, very satisfied, somewhat satisfied, not very satisfied, and not satisfied (1-5 respectivly)

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(x = data_subset$V104, y = satisfaction_numeric)
```



we can tell there is a higher likelyhood of satisfaction at all ages. In addition there are cluster of unsatisfied individuals at certain age groups ~40 ~60. this could be something to invesitgate, but at this stage of analysis I cannot draw any conclusions or make any assumptions about this data.

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.

```
library(tidyr)
```

from what I can tell all the columns respond to individual variables, that being the respondant. That being said I could be mistaken as I am not looking at all 3000 plus individual varaibles. for the purposes of this exercise I will write a sample line of code.

gather(data = changing_lives, key = V104, value = m_value, … = -col)

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.

the same can be said about this question.

separate(data = changing_lives, col = V402, intro = c("1", "2"))

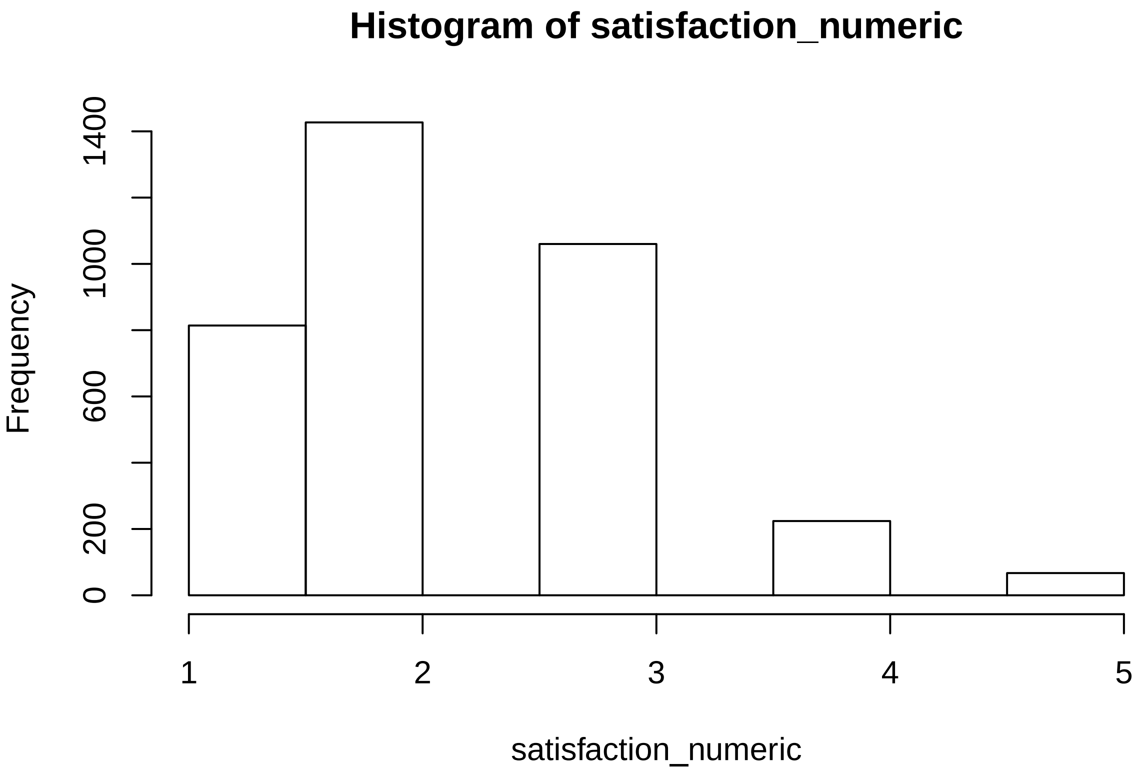
At this stage each row in your data should represent one observation, each column should be a variable, and each table should be observational unit.

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.

There are a number of different classes in the data I am trying to capture, from factors to numerics and even logicals. Some work for the data analysis I am trying to do and some do not. As seen previously in the assignment I had to assign one of the variables to a numeric in order to display it as a histogram. This worked with that subset of data because the numbers coresponded with a specific response to a multiple choice question but not all data will be as easily converted.

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)

```
satisfaction_numeric <- as.numeric(data_subset$V301)
hist(satisfaction_numeric)
```



a demonstration of coercion methods used previously

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

N/A

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 missing data in this data set is coded as NA and there are a lot in the data set, especially for certain questions reguarding how many children and individual has. Ex. sex of 9th child. The NAs are coded as NAs because I imported an R data set into R studio.

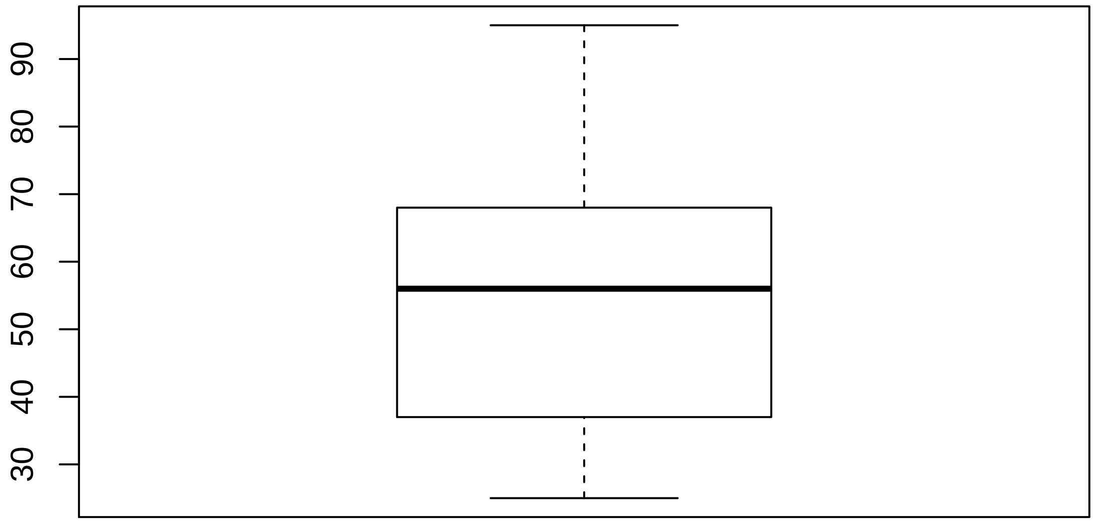
11. Are there any special values in your dataset? If so, what are they and how do you think they got there? *The presence of special values is less likely if you haven't performed any data manipulation yet so you*

should remember to return to this step each time you carry out a mathematical transformation of any values in your dataset.

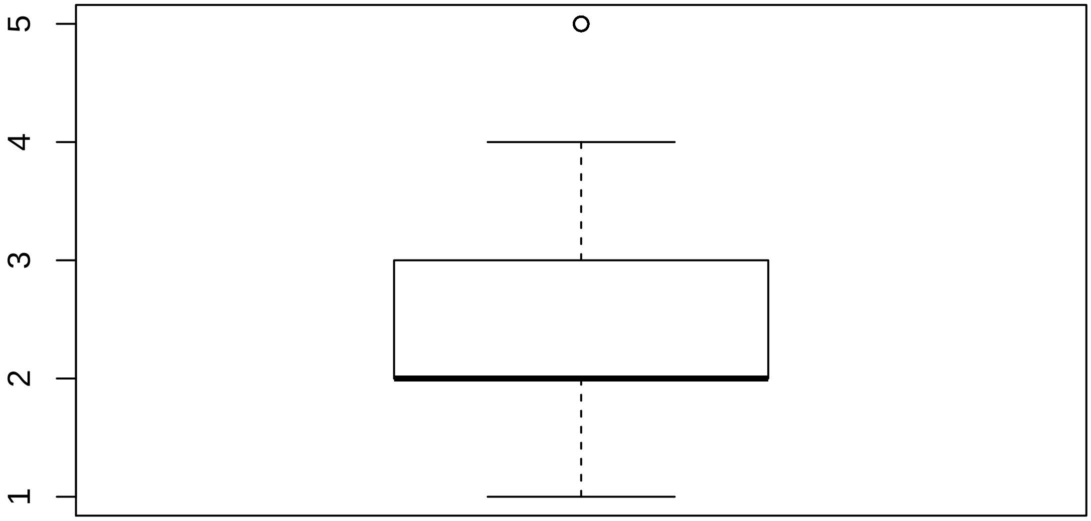
N/A or at least not discovered yet.

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?

```r
boxplot(data_subset$V104)
```



```r
boxplot(satisfaction_numeric)
```



there are over 3000 variables, so it seems unreasonable for me to run all of them for the purposes of this assignment. Instead I will run the two relevant ones I used previously in the assignment. the age of respondants box plot is relativly standard with no outliers.

This is most likely because of the consistantly diverse age range between respondants.

the second plot which looks at the respondants satisfaction with life, does contain an outlier as considered by the box plot. one of the few instances of no satisfaction is marked on the box plot as an outlier. However, considering this is a completely valid response, and only considered an outlier for it spearation from the rest of the data I will still be considering this data point in my analysis.

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?).

the second plot which looks at the respondants satisfaction with life, does contain an outlier as considered by the box plot. one of the few instances of no satisfaction is marked on the box plot as an outlier. However, considering this is a completely valid response, and only considered an outlier for it spearation from the rest of the data I will still be considering this data point in my analysis.

other outliers may be detrimental to the overal analysis of the data set. In those cases I would most likely run analyses including and excluding them in order to be thurough, then make an executive decision on if they should remain in the data set.