Exploring Data in R

Kim Roth & Johnathan Clementi

2/16/2020

Library load

library(tidyverse) #loads ggplot and the others  
library(readxl)#for reading in excel files

Today we will import data and explore it to prepare for cleaning.

The advantage of cleaning data in R rather than excel is that you have a replicable process should you ever wonder what you did when cleaning or need to change it.

Most important tip when cleaning. Always keep the original data file. Always. Fortunately that’s the default in R, although you can export the edited data.

The data set is data from JC Blair Hospital, now known as Penn Highlands Huntingdon, based on data they collect yearly at the Huntingdon County Fair about blood pressure and demographics from voluntary participation by fair attendees. Upload the data file from moodle and then use File->Import Data Set-> from Excel. You’ll need to change my code for this to work. My reccomendation is to copy and paste the code you get from File->Import Dataset->From Excel. It is in the lower right hand corner and labeled Code Preview. You only need the middle line from the code, the one that resembles the code in the chunk nelow.

Pressure <- read\_excel("C:/Users/Johnathan/Google Drive/Grad School/Juniata\_DataScience/DS500/Week4/data/Bloodpressuredata.xlsx")

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting numeric in L2985 / R2985C12: got '62"'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting numeric in L3039 / R3039C12: got '56"'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting numeric in I3553 / R3553C9: got 'Too old'

Note that this loads the file as a data frame(technically a tibble) in R. Anything you do to the data in R will not effect this file unless you specifically overwrite it.

Note that the error message from R at the load indicates there may be some problems. In fact it indicates that R got rid of a value in line 3553 which was someone who listed their age as Too old

You could use View to look at the data frame, but this is a bit large of a file to do that effectively.

Variables: OVERALL-a number assoiated with the case, essentially a case ID YEAR-What year was the data taked REC- a number associated with case and year. Essentially which entry was this for this year. BP-blood pressure as a ratio, systolic/diastolic BP Ratio-The blood pressure ratio converted to a decimal Systolic-The top number in blood pressure Diastolic-The bottom number in blood pressure. BP Status-Whether the blood pressure is Low, Normal, Pre-Hypertension, Hypertension Age-reported age of person survey Weight- measured weight of person in pounds Height- height measured in feet and inches Height-Inches-height measured in inches BMI-a ratio of sorts between height and weight BMI Category-a category based on BMI Gender-reported gender of participant Diagnosed Diabetic-report of the participant if ever diagnosed diabetic Tobacco User-report of the participant if they use tobacco OnmedicationforBP-report of the participant if they are on medication for high blood pressure HuntingdonCountyResident-is the participant a resident of Huntingdon County Date-Date data taken

Pressure

## # A tibble: 3,670 x 20  
## OVERALL YEAR REC BP `BP Ratio` Systolic Diastolic `BP Status` Age  
## <dbl> <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <chr> <dbl>  
## 1 1 2007 1 147/~ 1.75 147 84 Hypertensi~ 51  
## 2 2 2007 2 133/~ 1.71 133 78 Pre-Hypert~ 49  
## 3 3 2007 3 143/~ 1.61 143 89 Hypertensi~ NA  
## 4 4 2007 4 156/~ 2.29 156 68 Hypertensi~ NA  
## 5 5 2007 5 147/~ 1.60 147 92 Hypertensi~ NA  
## 6 6 2007 6 130/~ 1.48 130 88 Pre-Hypert~ NA  
## 7 7 2007 7 122/~ 1.88 122 65 Normal NA  
## 8 8 2007 8 148/~ 1.90 148 78 Hypertensi~ 74  
## 9 9 2007 9 152/~ 1.77 152 86 Hypertensi~ 48  
## 10 10 2007 10 138/~ 2.82 138 49 Pre-Hypert~ 78  
## # ... with 3,660 more rows, and 11 more variables: Weight <dbl>, Height <chr>,  
## # `Height-Inches` <dbl>, BMI <dbl>, `BMI Category` <chr>, Gender <chr>,  
## # DiagnosedDiabetic <chr>, TobaccoUser <chr>, OnmedicationforBP <chr>,  
## # HuntingdonCountyResident <chr>, Date <chr>

str(Pressure) #view variables and types

## Classes 'tbl\_df', 'tbl' and 'data.frame': 3670 obs. of 20 variables:  
## $ OVERALL : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ YEAR : num 2007 2007 2007 2007 2007 ...  
## $ REC : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ BP : chr "147/84" "133/78" "143/89" "156/68" ...  
## $ BP Ratio : num 1.75 1.71 1.61 2.29 1.6 ...  
## $ Systolic : num 147 133 143 156 147 130 122 148 152 138 ...  
## $ Diastolic : num 84 78 89 68 92 88 65 78 86 49 ...  
## $ BP Status : chr "Hypertension" "Pre-Hypertension" "Hypertension" "Hypertension" ...  
## $ Age : num 51 49 NA NA NA NA NA 74 48 78 ...  
## $ Weight : num 201 155 NA NA NA NA NA 206 185 125 ...  
## $ Height : chr NA NA NA NA ...  
## $ Height-Inches : num 74 70 NA NA NA NA NA 64 64 64 ...  
## $ BMI : num 25.8 22.2 NA NA NA ...  
## $ BMI Category : chr "Overweight" "Normal" NA NA ...  
## $ Gender : chr "M" "M" "N/A" "N/A" ...  
## $ DiagnosedDiabetic : chr "N" "N" "N/A" "N/A" ...  
## $ TobaccoUser : chr "N" "Y" "N/A" "N/A" ...  
## $ OnmedicationforBP : chr "N" "N" "N/A" "N/A" ...  
## $ HuntingdonCountyResident: chr "Y" "Y" "N/A" "N/A" ...  
## $ Date : chr NA NA NA NA ...

1. Looking at the output of the str command, does it appear that the numerical data is all treated as numbers? Why or Why not? 1A. No, there are Null values that are marked by ‘NA’. The Height field is also treated as text, but should be numerical.

Let’s look at the cateorgical variables for what is wrong. The first places to start are graphing the variables and summarizing them. You really only need to do one of these, but it is fine to do both.

Here is one way to summarize a categorical variable a table of counts, called a one way table

count(Pressure, Gender)

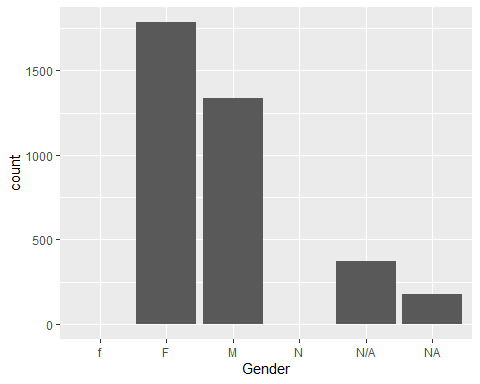
## # A tibble: 6 x 2  
## Gender n  
## <chr> <int>  
## 1 f 1  
## 2 F 1787  
## 3 M 1337  
## 4 N 1  
## 5 N/A 370  
## 6 <NA> 174

1. Graph Gender and look at both the graph and the table. What’s wrong that we need to fix. 2A. There are multiple values that represent the same category of case/record - i.e. f and F mark Female; N/A and NA

count(Pressure, Gender)

## # A tibble: 6 x 2  
## Gender n  
## <chr> <int>  
## 1 f 1  
## 2 F 1787  
## 3 M 1337  
## 4 N 1  
## 5 N/A 370  
## 6 <NA> 174

ggplot(data=Pressure) + geom\_bar(mapping = aes(x = Gender))

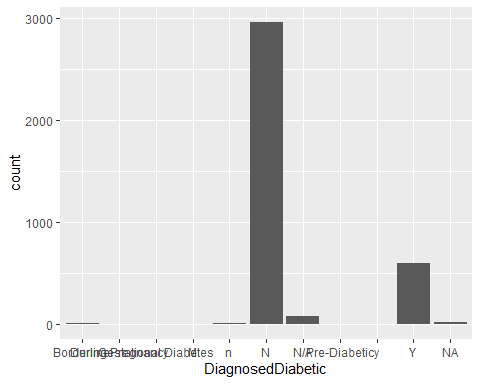


1. Pick another categorical variable (one without spaces in the variable name, as that causes problems). Repeat graphing and making a table. What seems to be wrong here? 3A. There are multiple values that represent the same category of case/record - i.e. f and F mark Female; N/A and NA.

count(Pressure, DiagnosedDiabetic)

## # A tibble: 11 x 2  
## DiagnosedDiabetic n  
## <chr> <int>  
## 1 Borderline 4  
## 2 During Pregnancy 1  
## 3 Gestational Diabetes 1  
## 4 M 1  
## 5 n 7  
## 6 N 2961  
## 7 N/A 75  
## 8 Pre-Diabetic 2  
## 9 y 1  
## 10 Y 599  
## 11 <NA> 18

ggplot(data=Pressure) + geom\_bar(mapping = aes(x = DiagnosedDiabetic))



For quantitative variables, we can summarize with the summary command. Here it is for age. It is an old command and we will learn about using the more modern summarize later.

summary(Pressure$Age)

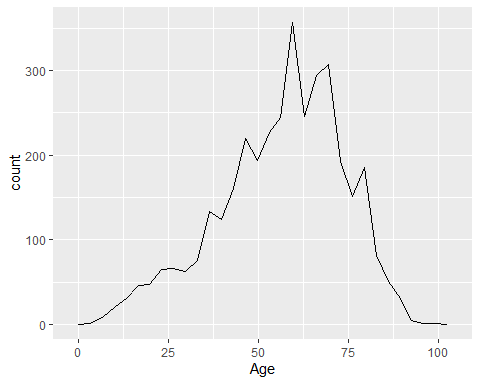
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 3.00 46.00 59.00 56.37 69.00 99.00 43

1. Make a graph of Age. Look at the graph and the summary statistics above. What stands out? Do you think something is wrong. 4A. There are 43 NA values that are not graphed. Its also a little strange to be taking the blood pressure of a 3 year old.

ggplot(data = Pressure) + geom\_freqpoly(mapping = aes(x = Age))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 43 rows containing non-finite values (stat\_bin).



1. Pick another quantitative variable (one without spaces in the variable name, as that causes problems). Repeat graphing and summarizing. What seems to be wrong here? 5A. THere are 404 NA’s, that’s a lot of people to not record weight data for.

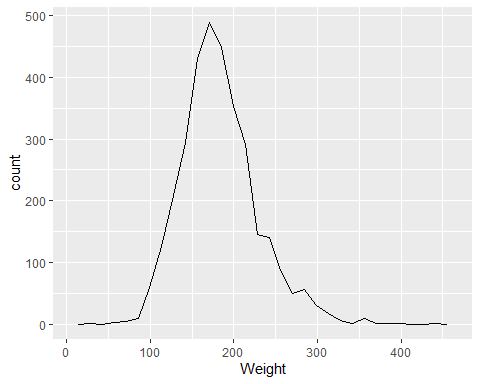
summary(Pressure$Weight)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 24 151 180 183 210 437 404

ggplot(data = Pressure) + geom\_freqpoly(mapping = aes(x = Weight))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

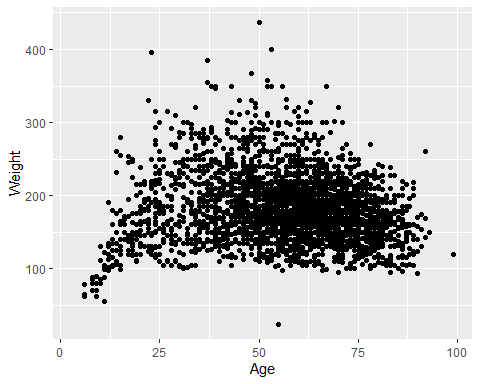
## Warning: Removed 404 rows containing non-finite values (stat\_bin).



1. Graph Age vs. Weight. Look for unusual values, called outliers. Do they seem okay to you? 6A. There is one outlier age around 54 weight is about 25 lbs. This is a value that should be removed from analysis if not enough context is available to fix it.

ggplot(data = Pressure) + geom\_point(mapping = aes(x = Age, y = Weight))

## Warning: Removed 411 rows containing missing values (geom\_point).



1. Pick another pair of quantitative variables (one without spaces in the variable name, as that causes problems). Repeat graphing. What seems to be wrong here? 7A. There is an incorrect value in the systolic field - outlier >800. BMI is calculated using height and weight. Therefore, if either of those fields does not have a value, we cannot calculate the BMI for that individual. It is worth checking that BMI has been calculated for all records because there are so man NA values (n=1466).

summary(Pressure$BMI)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 13.74 24.96 28.34 29.07 32.26 70.53 1466

summary(Pressure$Systolic)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0 118.0 130.0 130.5 142.0 814.0 44

ggplot(data = Pressure) + geom\_point(mapping = aes(x = Systolic, y = BMI))

## Warning: Removed 1502 rows containing missing values (geom\_point).

