Final Project Introduction Statistics Fall 2020

Elias Houstis

Final Project Fall 2020

dataset: NHANES

The data we're going to work with comes from the National Health and Nutrition Examination Survey (NHANES) program at the CDC. You can read a lot more about NHANES on the CDC's website or Wikipedia. NHANES is a research program designed to assess the health and nutritional status of adults and children in the United States. The survey is one of the only to combine both survey questions and physical examinations. It began in the 1960s and since 1999 examines a nationally representative sample of about 5,000 people each year. The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. The physical exam includes medical, dental, and physiological measurements, as well as several standard laboratory tests. NHANES is used to determine the prevalence of major diseases and risk factors for those diseases. NHANES data are also the basis for national standards for measurements like height, weight, and blood pressure. Data from this survey is used in epidemiology studies and health sciences research, which help develop public health policy, direct and design health programs and services, and expand the health knowledge for the Nation.

We are using a small slice of this data. We're only using a handful of variables from the 2011-2012 survey years on about 5,000 individuals. The CDC uses a sampling strategy to purposefully oversample certain subpopulations like racial minorities. Naive analysis of the original NHANES data can lead to mistaken conclusions because the percentages of people from each racial group in the data are different from general population. The 5,000 individuals here are resampled from the larger NHANES study population to undo these oversampling effects, so you can treat this as if it were a simple random sample from the American population. # PartI - Descriptive Statistics

Load the data

```
##
## 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
```

```
## Warning: `tbl_df()` is deprecated as of dplyr 1.0.0.
## Please use `tibble::as_tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
##  [1] "ï..id"              "Gender"             "Age"
##  [4] "Race"               "Education"          "MaritalStatus"
##  [7] "RelationshipStatus" "Insured"            "Income"
## [10] "Poverty"            "HomeRooms"          "HomeOwn"
## [13] "Work"               "Weight"             "Height"
## [16] "BMI"                "Pulse"              "BPSys"
## [19] "BPDia"              "Testosterone"       "HDLChol"
## [22] "TotChol"            "Diabetes"           "DiabetesAge"
## [25] "nPregnancies"       "nBabies"            "SleepHrsNight"
## [28] "PhysActive"         "PhysActiveDays"     "AlcoholDay"
## [31] "AlcoholYear"        "SmokingStatus"
```

Alternative views of the data

Recall several built-in functions that are useful for working with data frames.

- Content:
  - head(): shows the first few rows
  - tail(): shows the last few rows
- Size:
  - dim(): returns a 2-element vector with the number of rows in the first element, and the number of columns as the second element (the dimensions of the object)
  - nrow(): returns the number of rows
  - ncol(): returns the number of columns
- Summary:
  - colnames() (or just names()): returns the column names
  - glimpse() (from dplyr): Returns a glimpse of your data, telling you the structure of the dataset and information about the class, length and content of each column

insert table of variables description in nhnanes dataset

Variable	Definition
id	A unique sample identifier
Gender	Gender (sex) of study participant coded as male or female
Age	Age in years at screening of study participant. Note: Subjects 80 years or older were recorded as 80.
Race	Reported race of study participant, including non-Hispanic Asian category: Mexican, Hispanic, White, Black, Asian, or Other. Not availale for 2009-10.
Education	Educational level of study participant Reported for participants aged 20 years or older. One of 8thGrade, 9-11thGrade, HighSchool, SomeCollege, or CollegeGrad.
MaritalStatus	Marital status of study participant. Reported for participants aged 20 years or older. One of Married, Widowed, Divorced, Separated, NeverMarried, or LivePartner (living with partner).

Variable	Definition
RelationshipStatus	Simplification of MaritalStatus, coded as Committed if MaritalStatus is Married or LivePartner, and Single otherwise.
Insured	Indicates whether the individual is covered by health insurance.
Income	Numerical version of HHIncome derived from the middle income in each category
Poverty	A ratio of family income to poverty guidelines. Smaller numbers indicate more poverty
HomeRooms	How many rooms are in home of study participant (counting kitchen but not bathroom). 13 rooms = 13 or more rooms.
HomeOwn	One of Home, Rent, or Other indicating whether the home of study participant or someone in their family is owned, rented or occupied by some other arrangement.
Work	Indicates whether the individual is current working or not.
Weight	Weight in kg
Height	Standing height in cm. Reported for participants aged 2 years or older.
BMI	Body mass index (weight/height2 in kg/m2). Reported for participants aged 2 years or older.
Pulse	60 second pulse rate
BPSys	Combined systolic blood pressure reading, following the procedure outlined for BPXSAR.
BPDia	Combined diastolic blood pressure reading, following the procedure outlined for BPXDAR.
Testosterone	Testerone total (ng/dL). Reported for participants aged 6 years or older. Not available for 2009-2010.
HDLChol	Direct HDL cholesterol in mmol/L. Reported for participants aged 6 years or older.
TotChol	Total HDL cholesterol in mmol/L. Reported for participants aged 6 years or older.
Diabetes	Study participant told by a doctor or health professional that they have diabetes. Reported for participants aged 1 year or older as Yes or No.
DiabetesAge	Age of study participant when first told they had diabetes. Reported for participants aged 1 year or older.
nPregnancies	How many times participant has been pregnant. Reported for female participants aged 20 years or older.
nBabies	How many of participants deliveries resulted in live births. Reported for female participants aged 20 years or older.
SleepHrsNight	Self-reported number of hours study participant usually gets at night on weekdays or workdays. Reported for participants aged 16 years and older.
PhysActive	Participant does moderate or vigorous-intensity sports, fitness or recreational activities (Yes or No). Reported for participants 12 years or older.
PhysActiveDays	Number of days in a typical week that participant does moderate or vigorous-intensity activity. Reported for participants 12 years or older.

Variable	Definition
AlcoholDay	Average number of drinks consumed on days that participant drank alcoholic beverages. Reported for participants aged 18 years or older.
AlcoholYear	Estimated number of days over the past year that participant drank alcoholic beverages. Reported for participants aged 18 years or older.
SmokingStatus	Smoking status: Current Former or Never.

```
## # A tibble: 6 x 32
##   ï..id Gender   Age Race  Education MaritalStatus RelationshipSta~ Insured
##   <int> <chr>  <int> <chr> <chr>     <chr>         <chr>            <chr>  
## 1 62163 male      14 Asian <NA>      <NA>          <NA>             Yes    
## 2 62172 female    43 Black High Sch~ NeverMarried  Single           Yes    
## 3 62174 male      80 White College ~ Married       Committed        Yes    
## 4 62174 male      80 White College ~ Married       Committed        Yes    
## 5 62175 male       5 White <NA>      <NA>          <NA>             Yes    
## 6 62176 female    34 White College ~ Married       Committed        Yes    
## # ... with 24 more variables: Income <int>, Poverty <dbl>, HomeRooms <int>,
## #   HomeOwn <chr>, Work <chr>, Weight <dbl>, Height <dbl>, BMI <dbl>,
## #   Pulse <int>, BPSys <int>, BPDia <int>, Testosterone <dbl>, HDLChol <dbl>,
## #   TotChol <dbl>, Diabetes <chr>, DiabetesAge <int>, nPregnancies <int>,
## #   nBabies <int>, SleepHrsNight <int>, PhysActive <chr>, PhysActiveDays <int>,
## #   AlcoholDay <int>, AlcoholYear <int>, SmokingStatus <chr>
```

```
## # A tibble: 6 x 32
##   ï..id Gender   Age Race  Education MaritalStatus RelationshipSta~ Insured
##   <int> <chr>  <int> <chr> <chr>     <chr>         <chr>            <chr>  
## 1 71909 male      28 Mexi~ 9 - 11th~ NeverMarried  Single           No     
## 2 71909 male      28 Mexi~ 9 - 11th~ NeverMarried  Single           No     
## 3 71910 female     0 White <NA>      <NA>          <NA>             Yes    
## 4 71911 male      27 Mexi~ College ~ Married       Committed        Yes    
## 5 71915 male      60 White College ~ NeverMarried  Single           Yes    
## 6 71915 male      60 White College ~ NeverMarried  Single           Yes    
## # ... with 24 more variables: Income <int>, Poverty <dbl>, HomeRooms <int>,
## #   HomeOwn <chr>, Work <chr>, Weight <dbl>, Height <dbl>, BMI <dbl>,
## #   Pulse <int>, BPSys <int>, BPDia <int>, Testosterone <dbl>, HDLChol <dbl>,
## #   TotChol <dbl>, Diabetes <chr>, DiabetesAge <int>, nPregnancies <int>,
## #   nBabies <int>, SleepHrsNight <int>, PhysActive <chr>, PhysActiveDays <int>,
## #   AlcoholDay <int>, AlcoholYear <int>, SmokingStatus <chr>
```

```
## [1] 5000   32
##  [1] "ï..id"              "Gender"             "Age"               
##  [4] "Race"               "Education"          "MaritalStatus"     
##  [7] "RelationshipStatus" "Insured"            "Income"            
## [10] "Poverty"            "HomeRooms"          "HomeOwn"           
## [13] "Work"               "Weight"             "Height"            
## [16] "BMI"                "Pulse"              "BPSys"             
## [19] "BPDia"              "Testosterone"       "HDLChol"           
## [22] "TotChol"            "Diabetes"           "DiabetesAge"       
## [25] "nPregnancies"       "nBabies"            "SleepHrsNight"     
## [28] "PhysActive"         "PhysActiveDays"     "AlcoholDay"        
## [31] "AlcoholYear"        "SmokingStatus"
```

```
## Rows: 5,000
## Columns: 32
## $ ï..id              <int> 62163, 62172, 62174, 62174, 62175, 62176, 62178,...
## $ Gender             <chr> "male", "female", "male", "male", "male", "femal...
## $ Age                <int> 14, 43, 80, 80, 5, 34, 80, 35, 17, 15, 57, 57, 5...
## $ Race               <chr> "Asian", "Black", "White", "White", "White", "Wh...
## $ Education          <chr> NA, "High School", "College Grad", "College Grad...
## $ MaritalStatus      <chr> NA, "NeverMarried", "Married", "Married", NA, "M...
## $ RelationshipStatus <chr> NA, "Single", "Committed", "Committed", NA, "Com...
## $ Insured            <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes",...
## $ Income             <int> 100000, 22500, 70000, 70000, 12500, 100000, 2500...
## $ Poverty            <dbl> 4.07, 2.02, 4.30, 4.30, 0.39, 5.00, 0.05, 0.87, ...
## $ HomeRooms          <int> 6, 4, 7, 7, 7, 8, 6, 6, 6, 4, 4, 4, 4, 4, 12, 12...
## $ HomeOwn            <chr> "Rent", "Rent", "Own", "Own", "Rent", "Own", "Ow...
## $ Work               <chr> NA, "NotWorking", "NotWorking", "NotWorking", NA...
## $ Weight             <dbl> 49.4, 98.6, 95.8, 95.8, 23.9, 68.7, 85.9, 89.0, ...
## $ Height             <dbl> 168.9, 172.0, 168.1, 168.1, 119.8, 171.6, 173.5,...
## $ BMI                <dbl> 17.3, 33.3, 33.9, 33.9, 16.7, 23.3, 28.5, 27.9, ...
## $ Pulse              <int> 72, 80, 56, 56, NA, 92, 68, 66, 86, 76, 84, 84, ...
## $ BPSys              <int> 107, 103, 97, 97, NA, 107, 121, 107, 108, 113, 1...
## $ BPDia              <int> 37, 72, 39, 39, NA, 69, 72, 66, 64, 27, 65, 65, ...
## $ Testosterone       <dbl> 274.95, 47.53, 642.82, 642.82, NA, 21.11, 562.78...
## $ HDLChol            <dbl> 1.14, 1.89, 1.40, 1.40, NA, 1.42, 1.22, 0.85, 1....
## $ TotChol            <dbl> 3.98, 4.37, 5.25, 5.25, NA, 4.42, 5.20, 3.70, 3....
## $ Diabetes           <chr> "No", "No", "No", "No", "No", "No", "No", "No", ...
## $ DiabetesAge        <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ nPregnancies       <int> NA, 3, NA, NA, NA, 5, NA, NA, NA, NA, NA, NA, NA...
## $ nBabies            <int> NA, 2, NA, NA, NA, 2, NA, NA, NA, NA, NA, NA, NA...
## $ SleepHrsNight      <int> NA, 8, 9, 9, NA, 7, 6, 7, 7, NA, 8, 8, 8, 8, 6, ...
## $ PhysActive         <chr> "No", "No", "No", "No", NA, "Yes", "No", "No", "...
## $ PhysActiveDays     <int> 1, 2, 7, 5, 7, 5, NA, NA, 4, 7, 2, NA, 7, NA, NA...
## $ AlcoholDay         <int> NA, 3, NA, NA, NA, 2, NA, 1, NA, NA, 1, 1, 1, 1,...
## $ AlcoholYear        <int> NA, 104, 0, 0, NA, 104, NA, 2, NA, NA, 260, 260,...
## $ SmokingStatus      <chr> NA, "Current", "Never", "Never", NA, "Never", "N...
```

Descriptive statistics

We can access individual variables within a data frame using the $ operator, e.g., mydataframe$specificVariable. Let’s print out all the Race values in the data. Let’s then see what are the unique values of each. Then let’s calculate the mean, median, and range of the Age variable.

If you run the `summary()` function on a data frame, you get some very basic summary statistics on each variable in the data.

Exercise 1 a) display race values b) get unique values of Race c) length of Race d) Read the functions that dplyr (https://dplyr.tidyverse.org/ (https://dplyr.tidyverse.org/)) supports e) do the d) using dplyr way

Missing data

Let’s try taking the mean of a `income` variable.

```
## [1] NA
```

What happened there? NA indicates missing data. Take a look at the Income variable.

```
##  int [1:5000] 100000 22500 70000 70000 12500 100000 2500 22500 22500 30000 ...
```

Notice that there are lots of missing values for Income. Trying to get the mean a bunch of observations with some missing data returns a missing value by default. This is almost universally the case with all summary statistics – a single NA will cause the summary to return NA. Now look at the help for ?mean. Notice the na.rm argument. This is a logical (i.e., TRUE or FALSE) value indicating whether or not missing values should be removed prior to computing the mean. By default, it’s set to FALSE. Now try it again.

```
## [1] 57077.66
```

The `is.na()` function tells you if a value is missing. Get the `sum( `)` of that vector, which adds up all the TRUEs to tell you how many of the values are missing.

```
## [1] 377
```

##R is.na Function Example (remove, replace, count, if else, is not NA)

Before we can start, let’s create some example data in R (or R Studio).

```
##   x_num x_fac x_cha
## 1     8     2     p
## 2     0  <NA>     a
## 3    -4     2     j
## 4    NA     1     x
## 5    -6     1     s
## 6    -3  <NA>     k
```

Our data consists of three columns, each of them with a different class: numeric, factor, and character. This is how the first six lines of our data look like:

Let’s apply the is.na function to our **whole data set**:

```
##      x_num x_fac x_cha
## [1,] FALSE FALSE FALSE
## [2,] FALSE  TRUE FALSE
## [3,] FALSE FALSE FALSE
## [4,]  TRUE FALSE FALSE
## [5,] FALSE FALSE FALSE
## [6,] FALSE  TRUE FALSE
```

We are also able to check whether there is or is not an NA value in a column or vector:

```
## [1] FALSE FALSE FALSE  TRUE FALSE FALSE
```

```
## [1] FALSE  TRUE FALSE FALSE FALSE  TRUE
```

```
## [1] FALSE FALSE FALSE FALSE FALSE FALSE
```

```
## [1]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
```

```
## [1]  TRUE FALSE  TRUE  TRUE  TRUE FALSE
```

```
## [1] TRUE TRUE TRUE TRUE TRUE TRUE
```

That’s nice, but the real power of `is.na` becomes visible in combination with other functions — And that’s exactly what I’m going to show you now.

is.na in Combination with Other R Functions

Remove NAs of Vector or Column

```
## [1] 1000
```

```
## [1] 799
```

Replace NAs with Other Values (i.e. mean)

In case of characters or factors, it is also possible in R to set NA to blank:

Count NAs via sum & colSums

Combined with the R function sum, we can count the amount of NAs in our columns. According to our previous data generation, it should be approximately 20% in x_num, 30% in x_fac, and 5% in x_cha.

```
## [1] 201
## [1] 305
```

```
## [1] 54
```

If we want to count NAs in multiple columns at the same time, we can use the function colSums:

```
## x_num x_fac x_cha
##   201   305    54
```

Detect if there are any NAs

We can also test, if there is at least 1 missing value in a column of our data. As we already know, it is TRUE that our columns have NAs.

```
## [1] TRUE
```

Locate NAs via which

In combination with the which function, is.na can be used to identify the positioning of NAs:

```
## [1]  4 10 13 15 17 18
```

Our first column has missing values at the positions 4, 5, 14, 17, 22, 23 and so forth.

if & ifelse

Missing values have to be considered in our programming routines, e.g. within the if statement or within for loops.

In the following example, I’m printing “Damn, it’s NA” to the R Studio console whenever a missing occurs; and “Wow, that’s awesome” in case of an observed value.

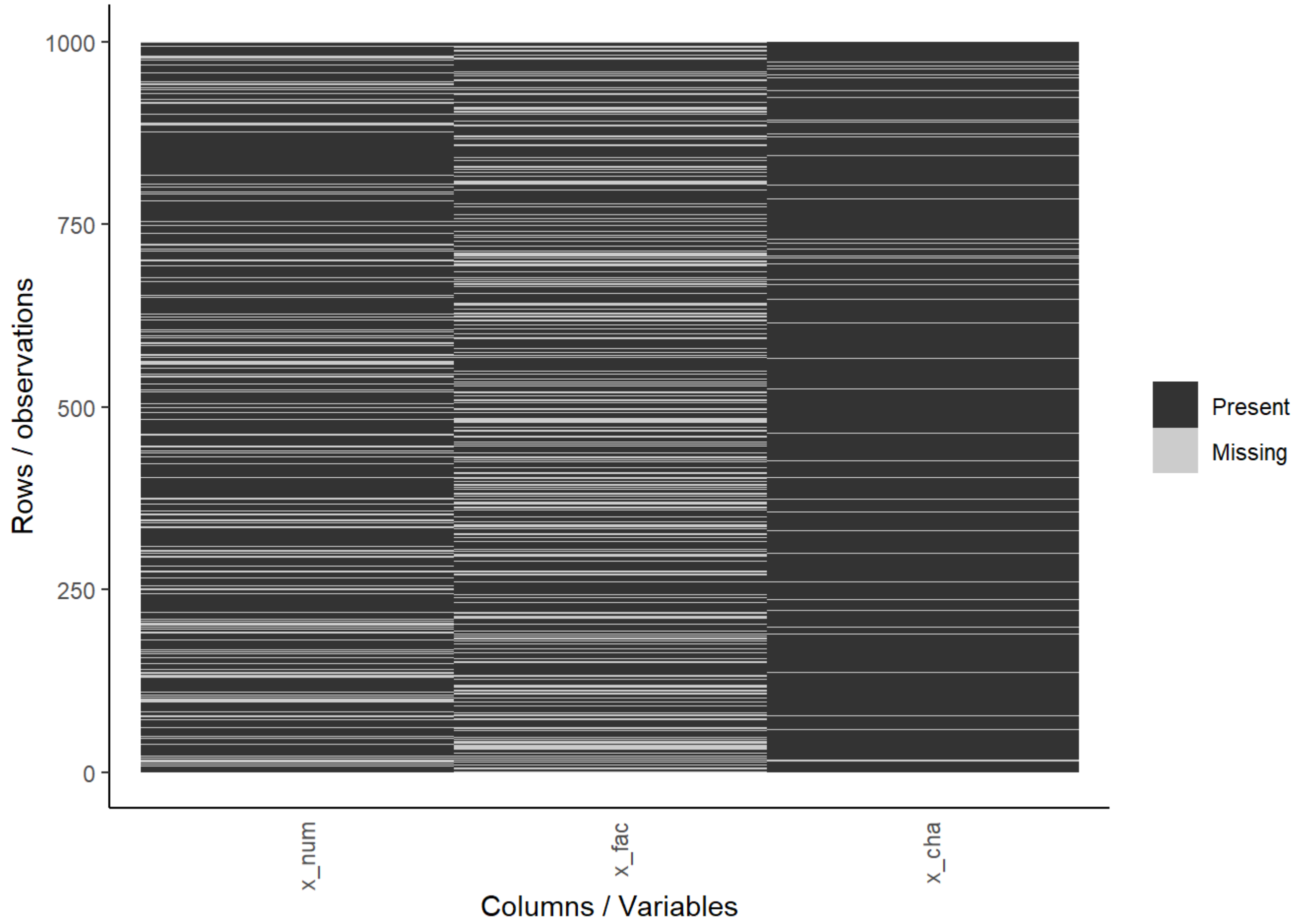
```
## [1] "Wow, that's awesome"
## [1] "Wow, that's awesome"
## [1] "Wow, that's awesome"
## [1] "Damn, it's NA"
## [1] "Wow, that's awesome"
## [1] "Wow, that's awesome"
## [1] "Wow, that's awesome"
## [1] "Wow, that's awesome"
## [1] "Wow, that's awesome"
## [1] "Damn, it's NA"
```

Note: Within the if statement we use is na instead of equal to — the approach we would usually use in case of observed values (e.g. if(x[i] == 5)).

Even easier to apply: the ifelse function.

```
## [1] "Wow, that's awesome" "Wow, that's awesome" "Wow, that's awesome"
## [4] "Damn, it's NA"       "Wow, that's awesome" "Wow, that's awesome"
```

There are a few handy functions in the `Tmisc` package for summarizing missingness in a data frame. You can graphically display missingness in a data frame as holes on a black canvas with `gg_na()` (use ggplot2 to plot NA values), or show a table of all the variables and the missingness level with propmiss().



```
## Warning: 'propmiss' is deprecated.
## Use 'Use summarize(across(everything(), ~sum(is.na(.))/n())' instead.
## See help("Deprecated")
```

```
## # A tibble: 3 x 4
##   var   nmiss     n propmiss
##   <chr> <dbl> <dbl>    <dbl>
## 1 x_num   201  1000    0.201
## 2 x_fac   305  1000    0.305
## 3 x_cha    54  1000    0.054
```

Exercise 2 Apply the above functions to other column of the dataset

Part II: Explatory Data Analysis (EDA)

It’s always worth examining your data visually before you start any statistical analysis or hypothesis testing. The big ones are histograms and scatterplots.

Histograms

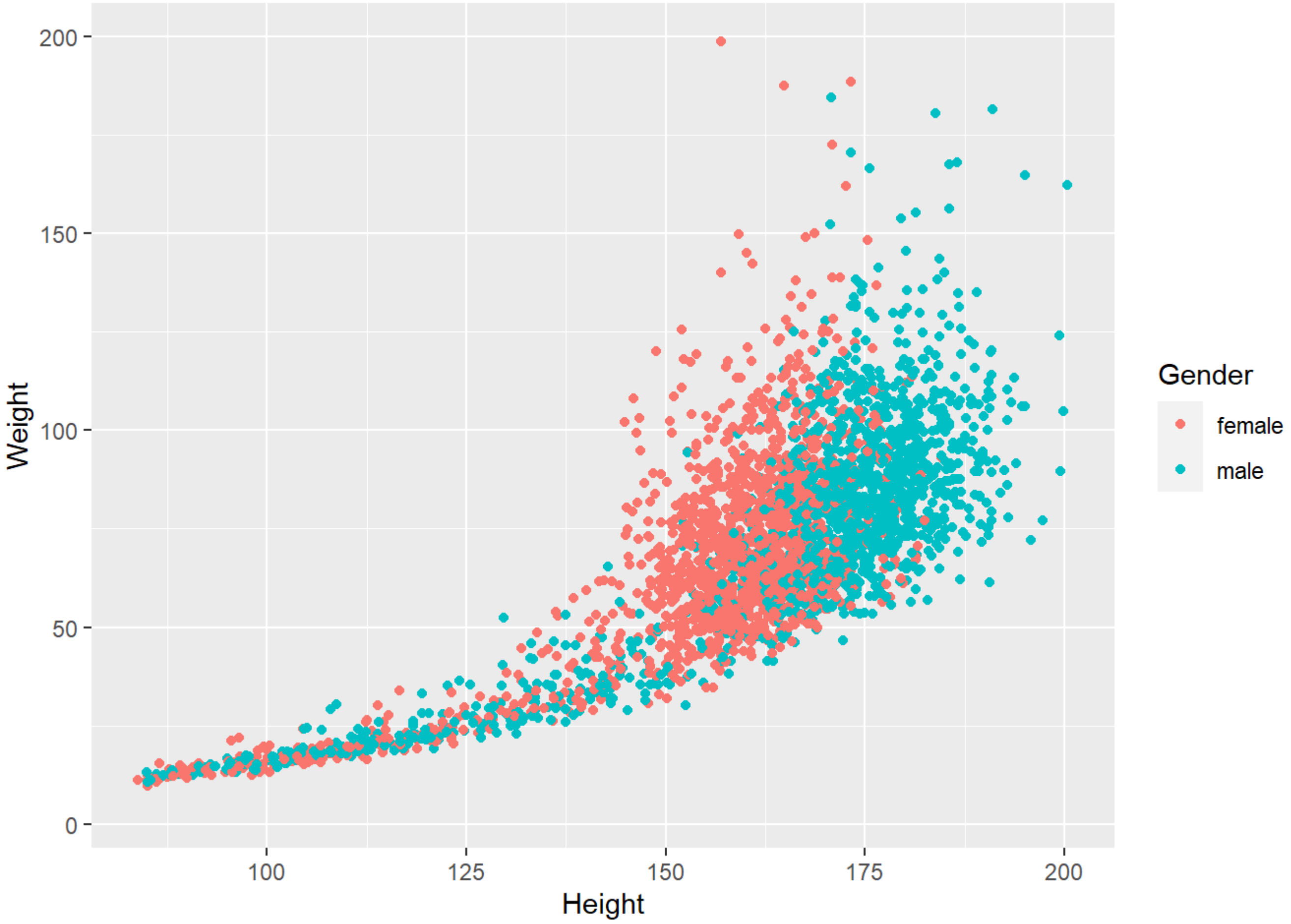
Exercise 3

a. Make some histograms with ggplot2 of 2 variables.
b. Look at BMI and indicate whether there outliers.
c. Look at weight. What their distribution looks like?
d. Check the age distribution. Are there kids in this data? Explain

Scatterplots

Let’s look at how a few different variables relate to each other. E.g., height and weight:

```
## Warning: Removed 166 rows containing missing values (geom_point).
```



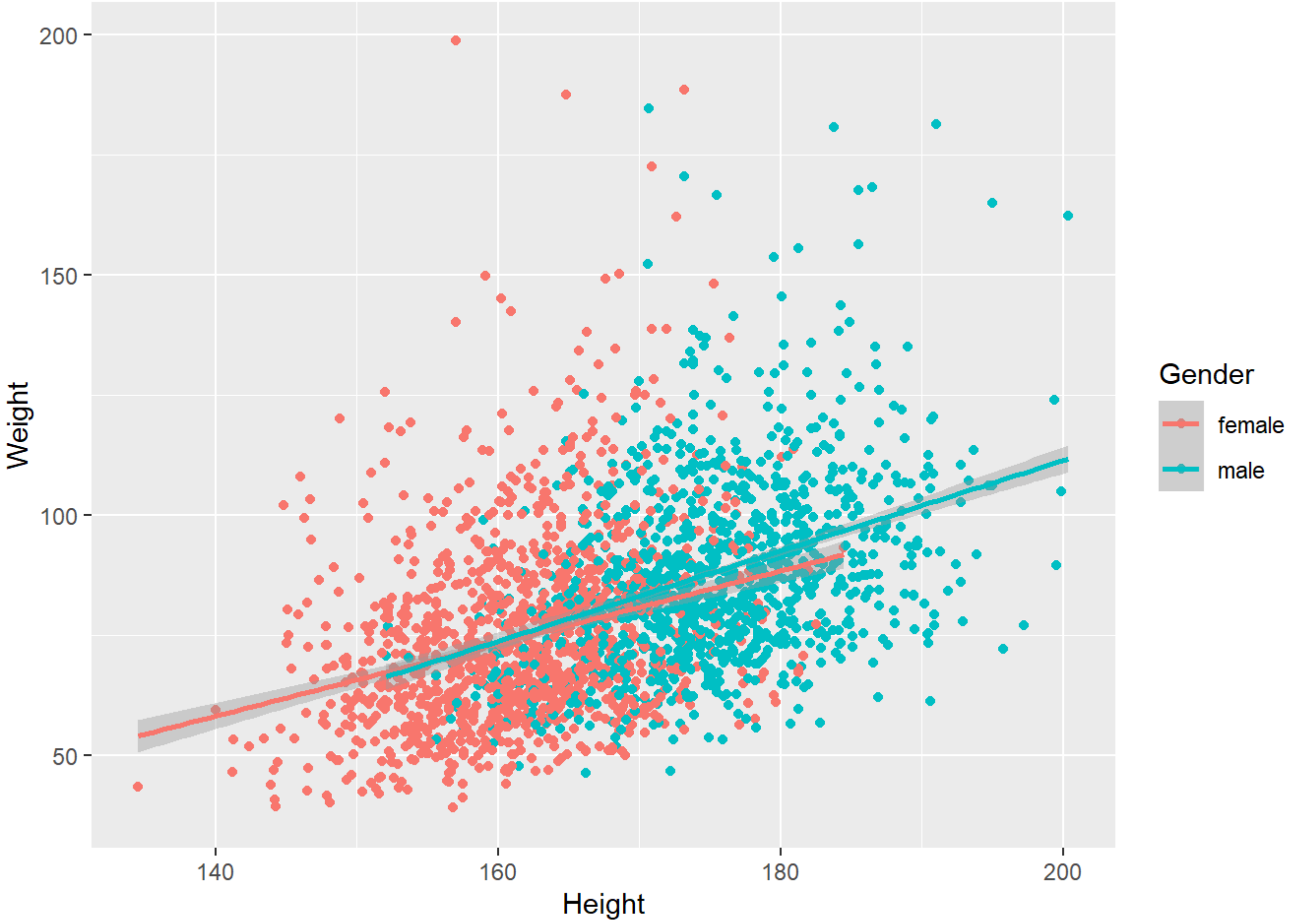
Let’s

filter out all the kids, draw trend lines using a linear model:

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 31 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 31 rows containing missing values (geom_point).
```



Exercise 4 1. What’s the mean 60-second pulse rate for all participants in the data?

2. What’s the range of values for diastolic blood pressure in all participants? (Hint: see help for min(), max(), and range() functions, e.g., enter ?range without the parentheses to get help).

3. What are the median, lower, and upper quartiles for the age of all participants? (Hint: see help for median, or better yet, quantile).

4. What’s the variance and standard deviation for income among all participants?

Part III: Point Estimates and Confidence Intervals

Parameter Estimation

Confidence Intervals

Exercise set 2

Part IV: Hypothesis Testing Continuous Variables

Hypothesis testing

Normality assumptions

T-tests

Wilcoxon test

Linear models

ANOVA

Linear regression

Multiple regression

Interpreting model summaries

Exercise set 3

Part V: Discrete variables

Contingency tables

Logistic regression

Interpreting model summaries

Exercise set 4