hw1

Noah Love

1/29/2022

R Markdown

Download the file 'tx_austin_2020_04_01.zip' from the Assignment folder on Courseworks, and extract the .csv file inside. These data regard police traffic stops in Austin, Texas from 2005-2016. The dataset is published by the Stanford Open Policing Project. The .csv file is a very small subset of the available data from this organization.

The questions for you to answer are indicated by bold text. Keep your answers short and succinct! Code here is highlighted in blue. You do not need to include any R code when you submit.

Read the data file into RStudio using your method of choice (try the 'Import Dataset' dropdown in the Environment pane, if you like). Give the dataframe a short but informative name, such as stops. How many rows are in the dataframe? How many columns? What does each row represent? (Check the About page online if you're unsure)

Number of rows:

```
nrow(tx_austin)
```

```
## [1] 483255
```

Number of columns:

```
ncol(tx_austin)
```

```
## [1] 29
```

Each row represents an individual police stop that took place in Austin Texas.

All data is represented in R as a "type", such as numeric, character, or logical. What are the data types contained in the following columns: subject_age, search_basis, vehicle_model, frisk_performed? There are multiple ways to get this information, but find a way to do it with a SINGLE line of code, or NO code, and tell me how you did it (Try searching Google for "r check type of dataframe column" if you get stuck).

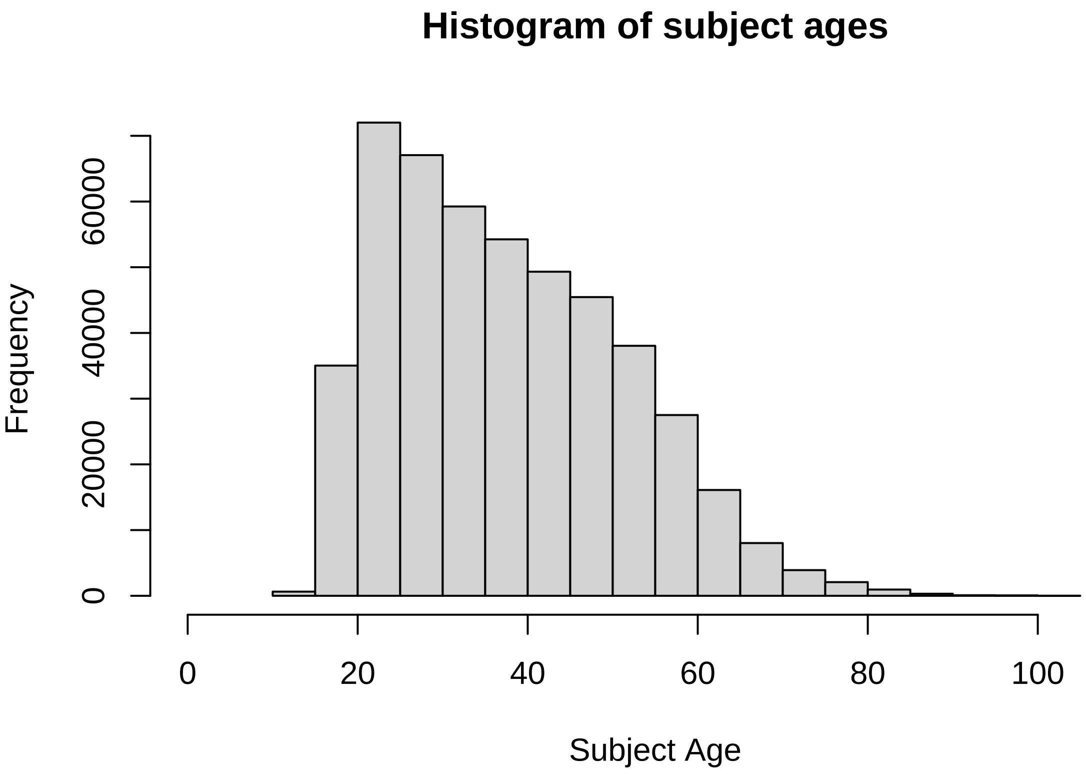
- subject age: double
- search_basis: character
- vehicle_model: character
- frisk_performed: logical

Just used spec(`df_name`) to list the type of each column in the df.

What percent of this dataset contain occurrences where contraband goods were found? Use the column contraband_found for this, and recall that you access a column in a dataframe using the $ symbol, like so: stops$contraband_found. A powerful feature of logical data is that R treats TRUE as 1 and FALSE as 0. This means you can use the sum() function to count how many TRUE's are in a column. However, if missing values (NA: not available) are present, sum() will output something you may not expect (try it). Read the docs by running ?sum to find an optional argument to sum() that solves this problem. Another hint: the function nrow() may be helpful here.

9.87% of recoreded occurences in the database found contraband.

Let's visualize the ages of people who are pulled over. Use the hist() function to plot a histogram of the subject_age column, and include the plot here. Is there anything difficult to interpret about the axes? How old were the youngest and oldest people stopped? The max() and min() functions will help you, and if you get a strange output, see the hint in #3 above.



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10.00   26.00   36.00   37.98   48.00  103.00    3164
```

Minimum age was 10, and possibly even more surprising was the 103 year old at the maximum. This makes it slightly more confusing if the axes aren't adjusted appropriately. There are also more than 3000 NAs.

```
## # A tibble: 4 x 3
##   subject_age search_conducted reason_for_stop
##         <dbl> <lgl>            <chr>
## 1         101 FALSE            VIOLATION OF TRANSPORTATION/VEHICLE LAWS
## 2         103 FALSE            VIOLATION OF TRANSPORTATION/VEHICLE LAWS
## 3         102 FALSE            VIOLATION OF TRANSPORTATION/VEHICLE LAWS
## 4         102 FALSE            VIOLATION OF TRANSPORTATION/VEHICLE LAWS
```

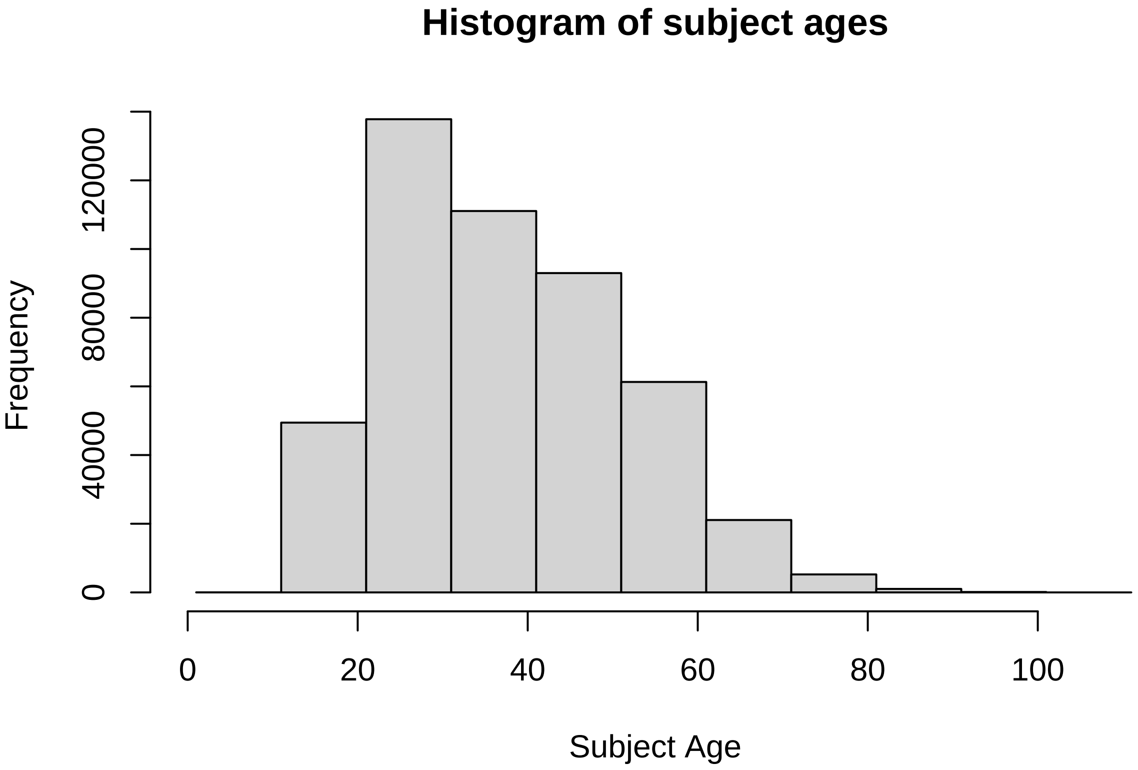
although none of them were searched.

The appearance of a histogram depends on the size of the bins (the width of the bars). We can change this parameter with a different value for the breaks argument in the hist() function. Read the docs by running ?hist and report the default value for breaks by looking in the Usage section.

Default is sturge's rule. i.e. 1 + 3.322 log (# of observations)

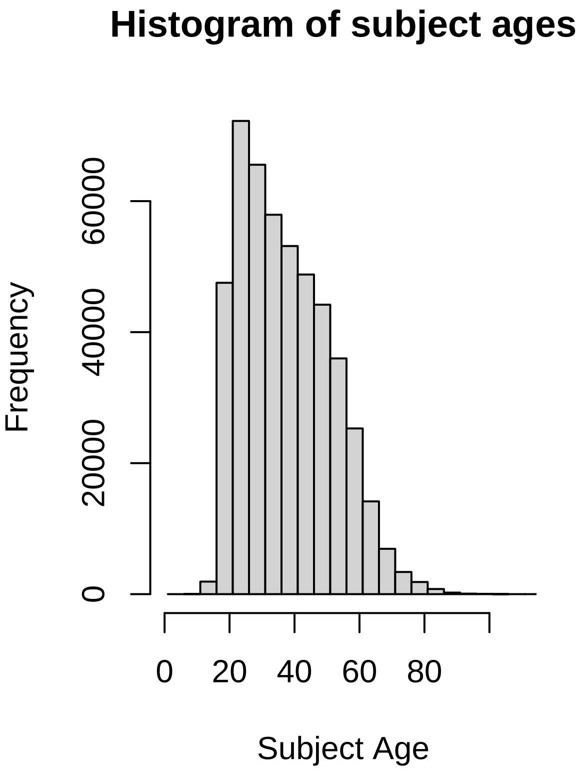
Replace the value for breaks by adding breaks = seq(1,120,10) to your call to hist(). Be sure to separate all function arguments (the inputs inside the parentheses) with a comma followed by a space, and make sure to include the breaks = part. What happened to your histogram when you included this argument? What does the seq() function do? You can either look at the docs, or just type seq(1,120,10) into the console and look at the output.

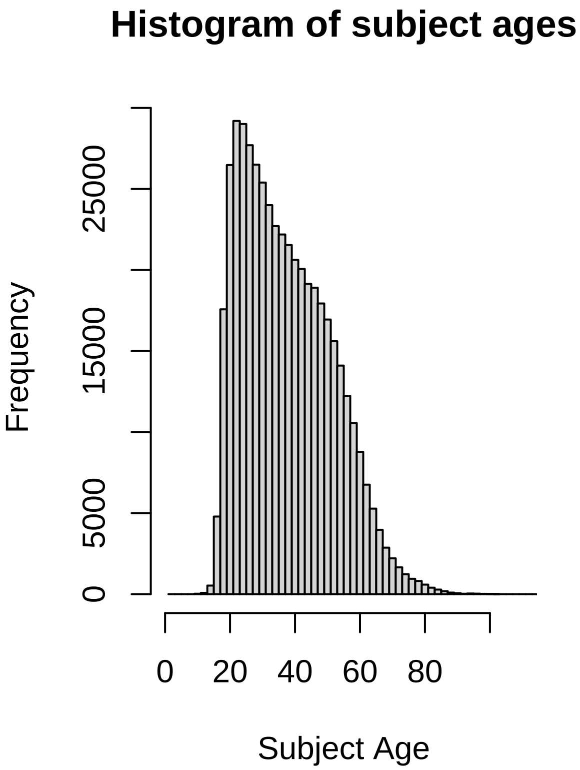
The axes are changes to teh sequence provided.



(1,120,10) created a sequence of numbers from 1, to 120, by increments of 10.

Play with the third argument to seq() to change the look of your histogram. Try breaks = seq(1,120,5), and breaks = seq(1,120,2). Which do you think is most informative? Are 16-year-olds pulled over more than 25-year-olds? Keyboard shortcut alert! Press the up arrow key when at the prompt in the console to quickly enter your last line of executed code. You can keep pressing up to move back in your command history. This is much better than typing a function over and over while you're testing – try it!

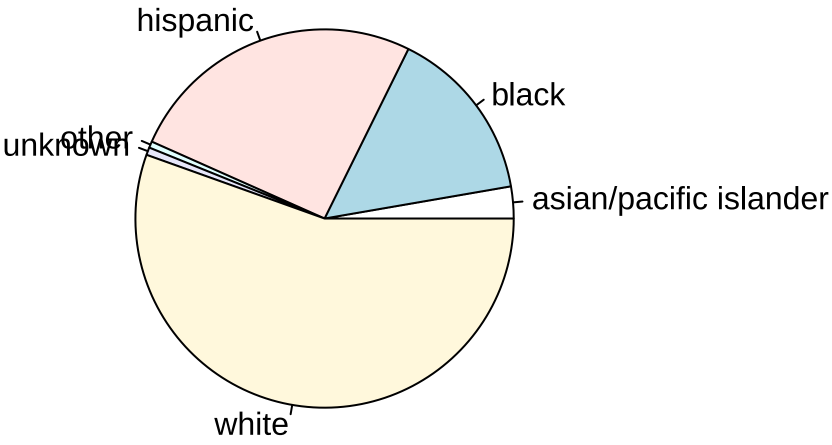




Hard to say. I like the 2 year graph for the teenagers to show the growth more clearly. However, the 2 year plot as a whole looks overwhelming and I prefer the 5 year one the most overall.

25 year olds were pulled over far more than 16 year olds.

Make a pie chart to quickly inspect the racial breakdown of traffic stops using the subject_race column, and include the plot here. Remember the table() function returns counts of a categorical variable, and the pie() function makes a plot, taking in as an argument the results of table(). See the R code from Lab 1 on Courseworks for a refresher.



The raw counts from the table() function aren't particularly informative. We can convert the counts to percentages using math operations, like we could with any vector (or matrix). Assign the results of table() to a new object counts. Next, divide counts by the sum of counts and multiply by 100 to convert the table into percentages. Finally, check out the federal census data for Austin, TX in 2010. Does there seem to be any racial disparity in who was stopped relative to their representation in the population? Why or why not (use numbers to justify)? (We would certainly do a more careful analysis before coming to any conclusions)

```
##
## asian/pacific islander                  black               hispanic
##              2.7246598             14.9660737             25.6476421
##                  other                unknown                  white
##              0.5434007              0.6487285             55.4694953
```

Blacks were over-policied (15% of stops vs 7.8 percent of population). As a result, hispanics (33.9%) and whites-nonhispanic (48.3%) were slightly under-policied.

The table() function can also make so-called contingency tables, where counts are grouped based on more than one variable. Pass in two columns of the dataset as arguments to table(), first subject_race, then subject_sex. Assign the output to a new variable and convert to percentages, just as you did above. Does there seem to be any disparity in the sex recorded for people stopped? If so, is it the same for all races? Use numbers to (briefly) justify your answer.

```
##
##                               female        male
##   asian/pacific islander  0.94267727  1.78367296
##   black                   4.03454274 10.94184891
##   hispanic                5.65047217 20.01449636
##   other                   0.09650431  0.44565938
##   unknown                 0.16380881  0.41666667
##   white                  18.69760603 36.81204440
```

Across the board, men are arrested at 2 to 4 times the rate of females in the same category. It was particularly worse in the Hispanic community and the other category.

Finally, I want the lab assignments to be valuable for you – they should be a challenge, but not impossible, nor should they take an egregious amount of time. The questions below are optional, but the more data I have on how you're doing, the more useful I can make the assignments. Thanks in advance!

On a scale of 1-5, with 1 being "too easy", 5 being "too hard", and 3 being "a good level", how difficult was this assignment? 1, but that's okay

Approximately how much time did you spend working on this assignment? (i.e. actively solving the problem, not exploring the data independently) 45 minutes

Any thoughts on what you found useful, or what you found mundane, or confusing? (anything, big or small) Fun and very relevant dataset. Great choice.