Lab 1: HipHop Lyrics

YOUR NAME HERE

Instructions

Please submit an HTML document created using R Markdown, but you are more than welcome to test your code out in an R script first. **Even if a question does not say "write code," you should write code for your answer!**

Recall the `hiphop` dataset from the Day 2 In-Class Activity. The in-depth description of the datset is here:

http://conservancy.umn.edu/bitstream/handle/11299/116327/5/explanationAAEHiphopChesley.txt

BE SURE TO SAVE YOUR WORK REGULARLY!!!

Copy the following code into an R chunk, to load the data and gain access to the tidyverse package.

```
hiphop <- read.csv("https://www.dropbox.com/s/1qqyshx5ikt9zoc/hiphop.csv?dl=1")

library(tidyverse)

# If you get an error on library(tidyverse), run the line below in your console.  Do NOT uncomment this

# install.packages("tidyverse")
```

Introduction (5 points)

1. Provide a brief overview (2-4 sentences) of the dataset. You may simply reference your work on the Day 2 activity to make this summary. We are repeating this because it is always good practice to start an analysis by getting a feel for the data and providing a quick summary for readers.

2. How many unique AAVE words were studied in this dataset?

```
length(unique(hiphop$word))
```

```
## [1] 64
```

3. Make a new variable that recategorizes `ethnic` into only two groups, "white" and "non-white", to simplify your data.

Potentially helpful functions: `mutate()`, `case_when()`

```
hiphop <- hiphop %>%
  mutate(
    ethnic_group = case_when(
      ethnic == "white" ~ "white",
      TRUE ~ "non-white"
    )
  )
```

4. What are the demographics of the people in this study? Investigate the variables `sex`, `age`, and `ethnic` and summarize your findings in 1-3 complete sentences.

Functions: `select()`, `unique()` or `distinct(, .keep_all = TRUE)`, `count()`, `summary()`

```
subjects <- hiphop %>%
  select(subj, age, sex, ethnic_group) %>%
  distinct(subj, .keep_all = TRUE)


subjects %>% count(sex, ethnic_group)
```

```
## # A tibble: 4 x 3
##   sex    ethnic_group     n
##   <fct>  <chr>        <int>
## 1 Female non-white       26
## 2 Female white           91
## 3 Male   non-white        7
## 4 Male   white           44
```

```
summary(subjects$age)
```

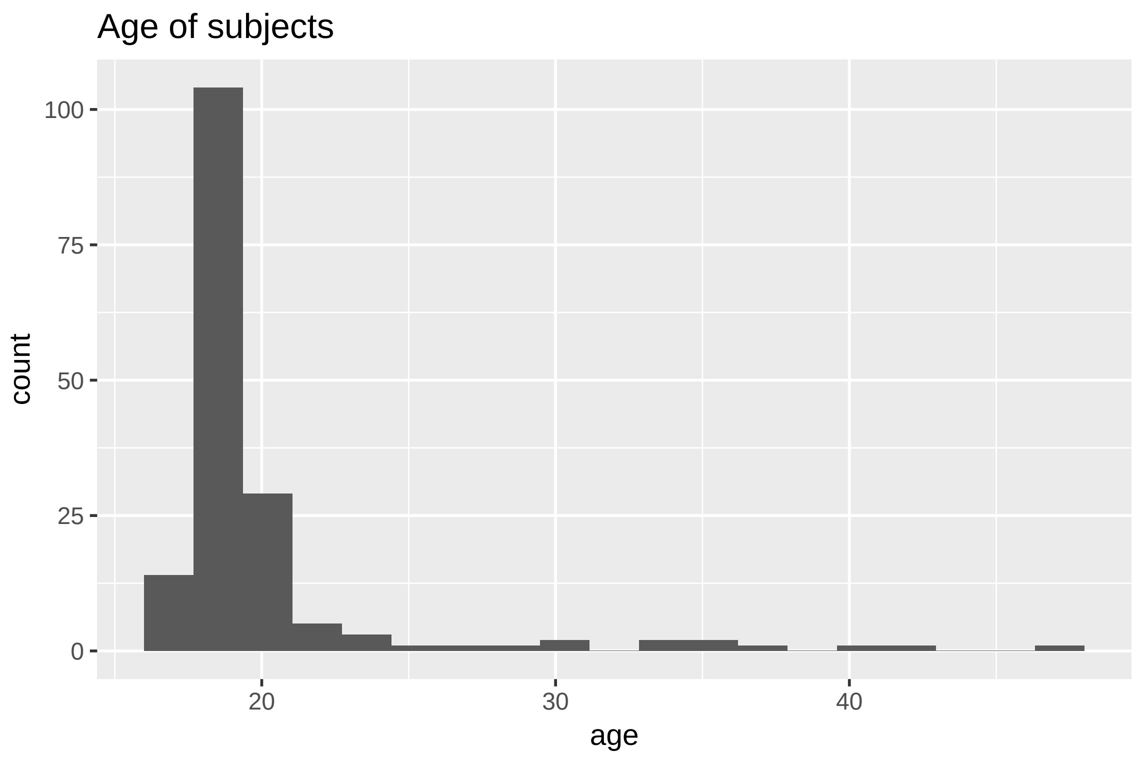
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   18.00   19.00   20.02   20.00   48.00
```

5. Make at least two plots to display the demographic information of the subjects in this study. You do not need to discuss these plots, but make sure they are appropriate to the data types and have informative titles and axis labels.

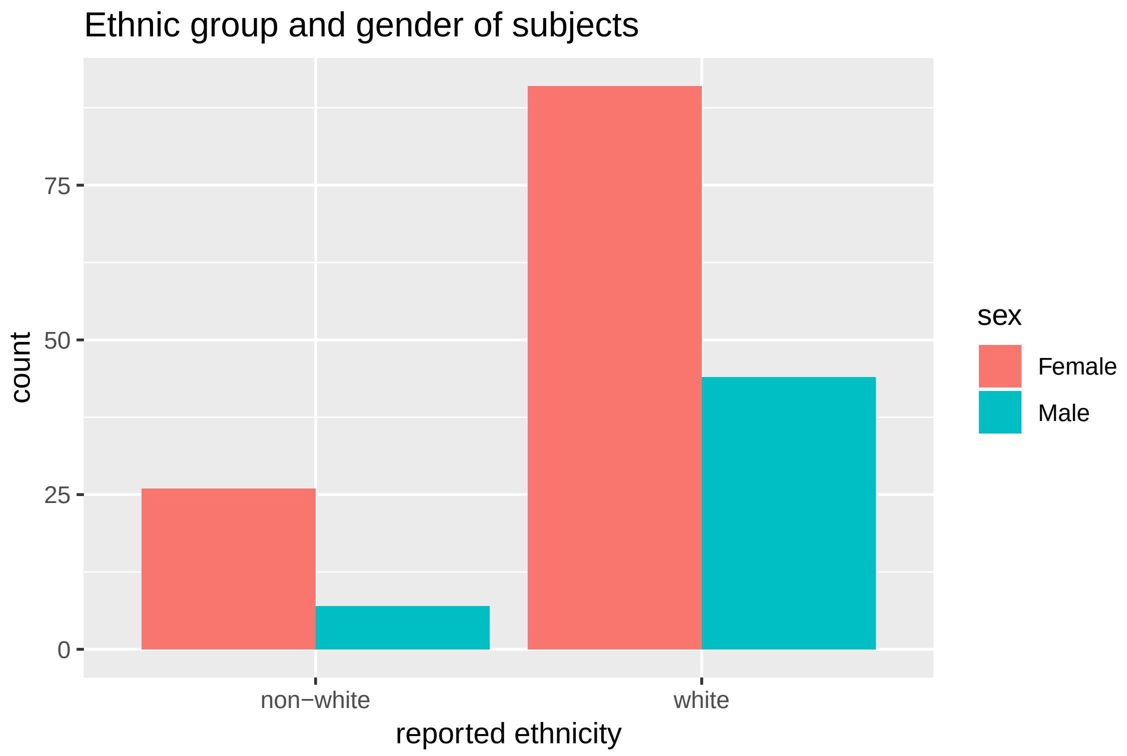
Functions: `ggplot()`, `geom_histogram()`, `geom_boxplot()`, `geom_bar()`, `ggtitle()`, `xlab()`, `ylab()`

```
# a few options exist, but I like these

subjects %>%
  ggplot(aes(x = age)) + geom_histogram(bins = 20) +
  ggtitle("Age of subjects")
```



```
subjects %>%
  ggplot(aes(x = ethnic_group, fill = sex)) + geom_bar(position = "dodge") +
  ggtitle("Ethnic group and gender of subjects") + xlab("reported ethnicity")
```



Familiar words

1. For each demographic group listed below, determine which word(s) in this study was the most and least familiar on average.

a. People below the age of 20
b. Non-white women
c. White men above the age of 30

Functions: `filter()`, `group_by()`, `summarize_at()`, `arrange()`, `desc()`

```
#a
young <- hiphop %>%
  filter(age < 20) %>%
  group_by(word) %>%
  summarize_at(vars(familiarity), funs(mean))

young %>% top_n(1, desc(familiarity))
```

```
## # A tibble: 1 x 2
##   word             familiarity
##   <fct>                  <dbl>
## 1 catch the vapors        1.03
```

```
young %>% top_n(1, familiarity)
```

```
## # A tibble: 1 x 2
##   word         familiarity
##   <fct>              <dbl>
## 1 off the hook        4.42
```

```
#b

nonwhite_women <- hiphop %>%
  filter(sex == "Female", ethnic_group == "non-white") %>%
  group_by(word) %>%
  summarize_at(vars(familiarity), funs(mean))

nonwhite_women %>% top_n(1, desc(familiarity))
```

```
## # A tibble: 4 x 2
##   word              familiarity
##   <fct>                   <dbl>
## 1 break someone out           1
## 2 dukey rope                  1
## 3 plex                        1
## 4 rollie                      1
```

```
nonwhite_women %>% top_n(1, familiarity)
```

```
## # A tibble: 1 x 2
##   word    familiarity
##   <fct>         <dbl>
## 1 feel me        4.19
```

```
#c
old_white_men <- hiphop %>%
  filter(sex == "Male", ethnic_group == "white", age > 30) %>%
  group_by(word) %>%
  summarize_at(vars(familiarity), funs(mean))

old_white_men %>% top_n(1, desc(familiarity))
```

```
## # A tibble: 25 x 2
##    word              familiarity
##    <fct>                   <dbl>
##  1 ay yo trip                  1
##  2 beezy                       1
##  3 break someone out           1
##  4 catch the vapors            1
##  5 crossroads                  1
##  6 crump                       1
##  7 dap                         1
##  8 dollar cab                  1
##  9 domino                      1
## 10 duckets                     1
## # ... with 15 more rows
```

```
old_white_men %>% top_n(1, familiarity)
```

```
## # A tibble: 1 x 2
##   word  familiarity
##   <fct>       <dbl>
## 1 5-0           4.2
```

Use the data

A former Canadian child TV star named Aubrey Graham is interested in switching careers to become a rapper. Aubrey hires you to consult the `hiphop` dataset to help compose his new songs.

Note: There is no single right answer to these questions. You will need to think about how you want to address the question, and do the appropriate variable adjustments and calculations to come up with a reasonable answer.

1. Aubrey hopes that his songs will be percieved as authentically hiphop. He hopes his lyrics will be recognizeable to those who describe themselves as hiphop fans, but less recognizeable to those who do not consider themselves fans. Suggest some words or phrases that Aubrey should try to use, and some words he should avoid.

Hint: Do separate calculations for hiphop fans and not, then `full_join()` the data.

```
fans <- hiphop %>%
  filter(hiphop > 7) %>%
  group_by(word) %>%
  summarize(avg_fam = mean(familiarity))

not_fans <- hiphop %>%
  filter(hiphop < 4) %>%
  group_by(word) %>%
  summarize(avg_fam = mean(familiarity))

new <- full_join(fans, not_fans, by = "word") %>%
  mutate(
    diff = avg_fam.x - avg_fam.y
  )

new %>% top_n(5, diff)
```

```
## # A tibble: 5 x 4
##   word            avg_fam.x avg_fam.y  diff
##   <fct>               <dbl>     <dbl> <dbl>
## 1 ashy                  4.2      2.60  1.60
## 2 cuddie                3.4      1.32  2.08
## 3 dead presidents       3.6      1.94  1.66
## 4 what it do            4.2      2.79  1.41
## 5 wile out              3.2      1.56  1.64
```

```
new %>% top_n(5, desc(diff))
```

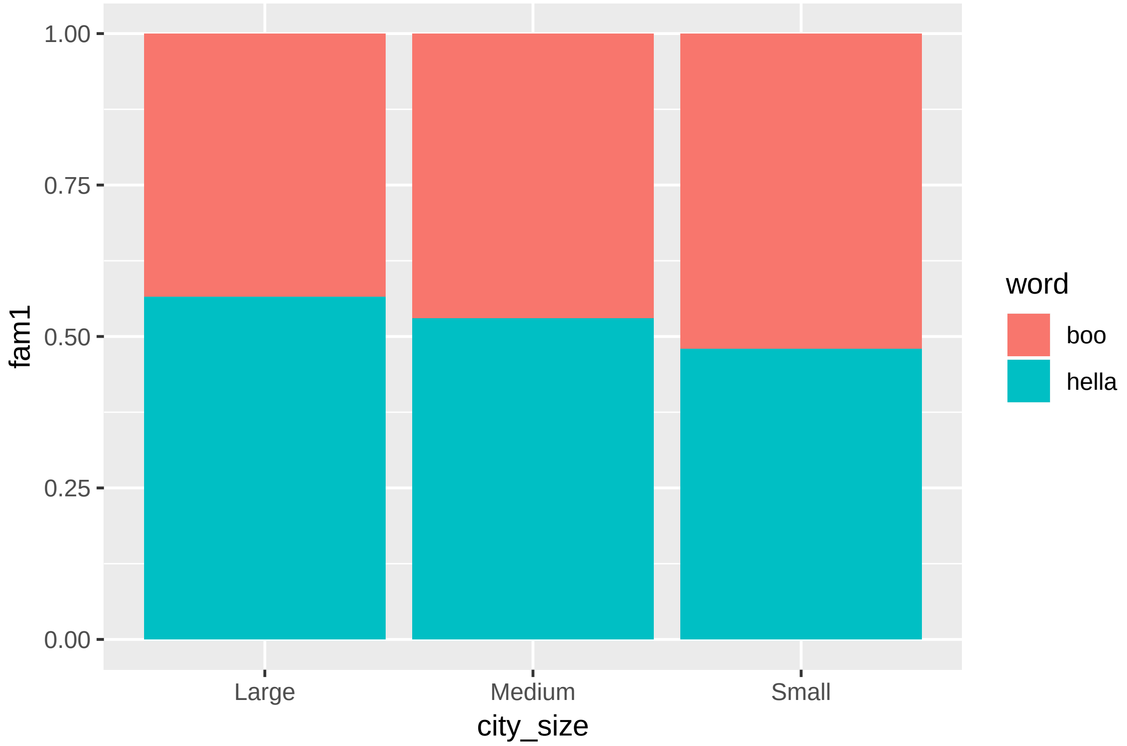
```
## # A tibble: 5 x 4
##   word         avg_fam.x avg_fam.y   diff
##   <fct>            <dbl>     <dbl>  <dbl>
## 1 bones              1        1.75 -0.75
## 2 boo                2.8      3.42 -0.619
## 3 ghostride          1        1.58 -0.581
## 4 hard               1.2      2.14 -0.940
## 5 off the hook       3.2      4.32 -1.12
```

2. Although Aubrey wants to be authentic, he also hopes to sell records, of course. Two titles have been suggested for his first album: “Hotline Boo” or “Hella Bling”. Based on the dataset, which will appeal more to the higher population areas? Make at least one plot to support your answer.

Hint: Consider first converting the population variable(s) to categories, such as “large”, “medium”, and “small”. You may also want to use the “fam1” variable instead of “familiarity”

```
titles <- hiphop %>%
  filter(word %in% c("boo", "hella")) %>%
  mutate(
    city_size = case_when(
      city > 200000 ~ "Large",
      city > 50000 ~ "Medium",
      TRUE ~ "Small"
    )
  )

ggplot(titles, aes(x = city_size, y = fam1, fill = word)) + geom_col(position = "fill")
```



3. Aubrey's true life dream is to collaborate with his fellow Canadian musician Justin Bieber. Luckily, he knows that Bieber himself was one of the subjects in this study! You know that Bieber is a white male, aged 17-23 at the time of the study, from a relatively small town (10,000-60,000 people) in Ontario.

Determine which subject is secretly Bieber, and justify your answer.

Then suggest a track listing (11 song titles) for Aubrey's song collaboration with the Biebs.

Hint: Refer again to the dataset description. There is another clue about Bieber's identity.

```
hiphop %>%
  distinct(subj, .keep_all = TRUE) %>%
  filter(sex == "Male", age <= 23, age >= 17, city < 60000, city > 10000) %>%
  top_n(1, bieber) %>%
  select(subj)
```

```
##   subj
## 1  p17
```

```
#biebs is subj17

hiphop %>%
  filter(subj == "p17") %>%
  select(word, familiarity) %>%
  top_n(1, familiarity)
```

```
##            word familiarity
## 1         A-town           5
## 2            boo           5
## 3         chedda           5
## 4        duckets           5
## 5        feel me           5
## 6           hard           5
## 7          hella           5
## 8   off the hook           5
## 9   player hater           5
## 10        toe up           5
## 11   What it is?           5
```