Duncan McKinnon

West

HW 3

```
#load tidyverse and okaycupiddata
suppressPackageStartupMessages({
  library("tidyverse")
  library("lubridate")
  library("okcupiddata")
})
```

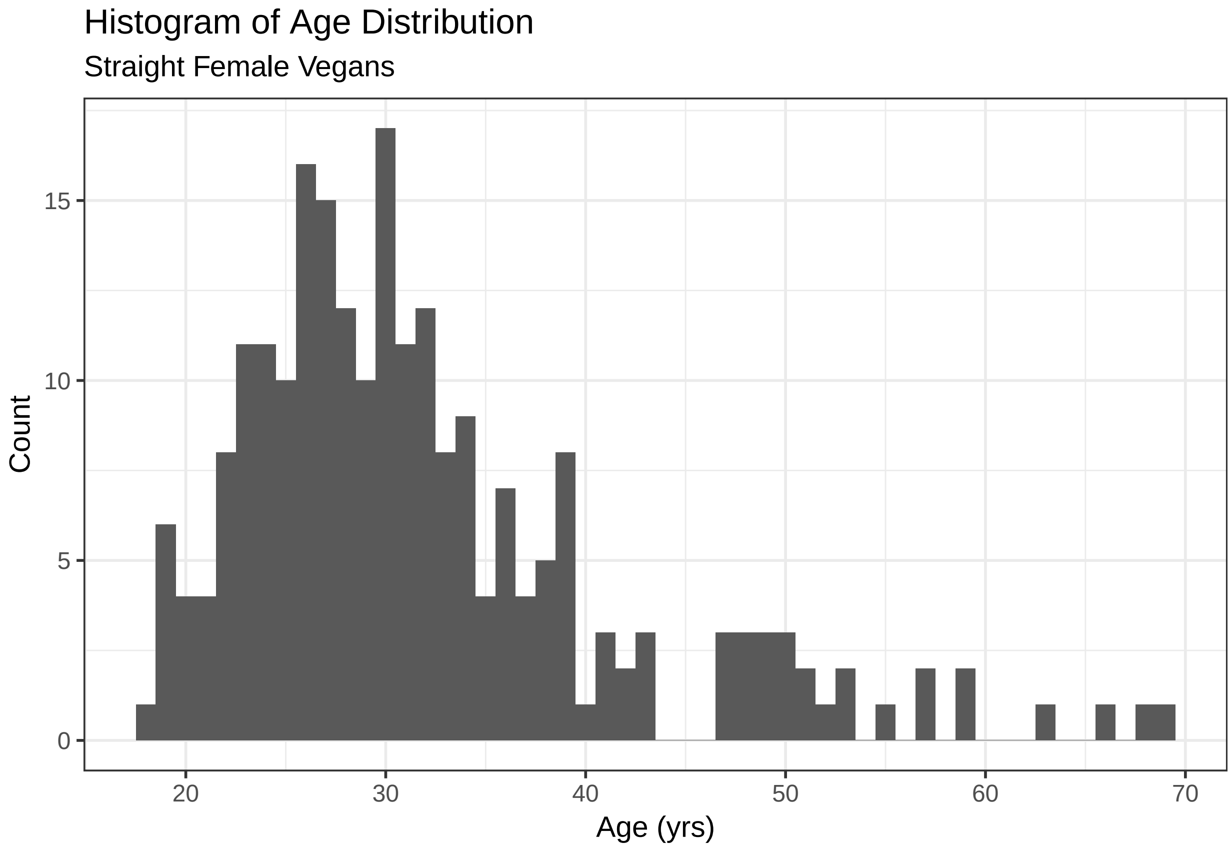
Q1 A).

```
# get data for Jason's preferences: straight women who are at least 'mostly' vegan
# filter okcupid profiles to get only straight female vegans
jason_data <- profiles %>%
  filter(sex == 'f',
         orientation == 'straight',
         diet %in% c('vegan', 'strictly vegan', 'mostly vegan'))
```

1).

I expected that the distribution of ages for straight women who are vegans would be right skewed because OkCupid only allows users who are 18yrs or older which represents a lower bound. I also expected that the mean would be relatively young compared with the full distribution of straight women on okcupid, because veganism is trendy for younger people today. The histogram generated shows a mean of 32.04yrs, a median of 30yrs and a standard deviation of 9.73yrs. The statistics for the ages of vegan straight women are only slightly lower than those for non-vegan straight women (mean 33.32yrs, median 31yrs, sd 10.15yrs), showing that veganism does not exclusively correspond to younger straight women.

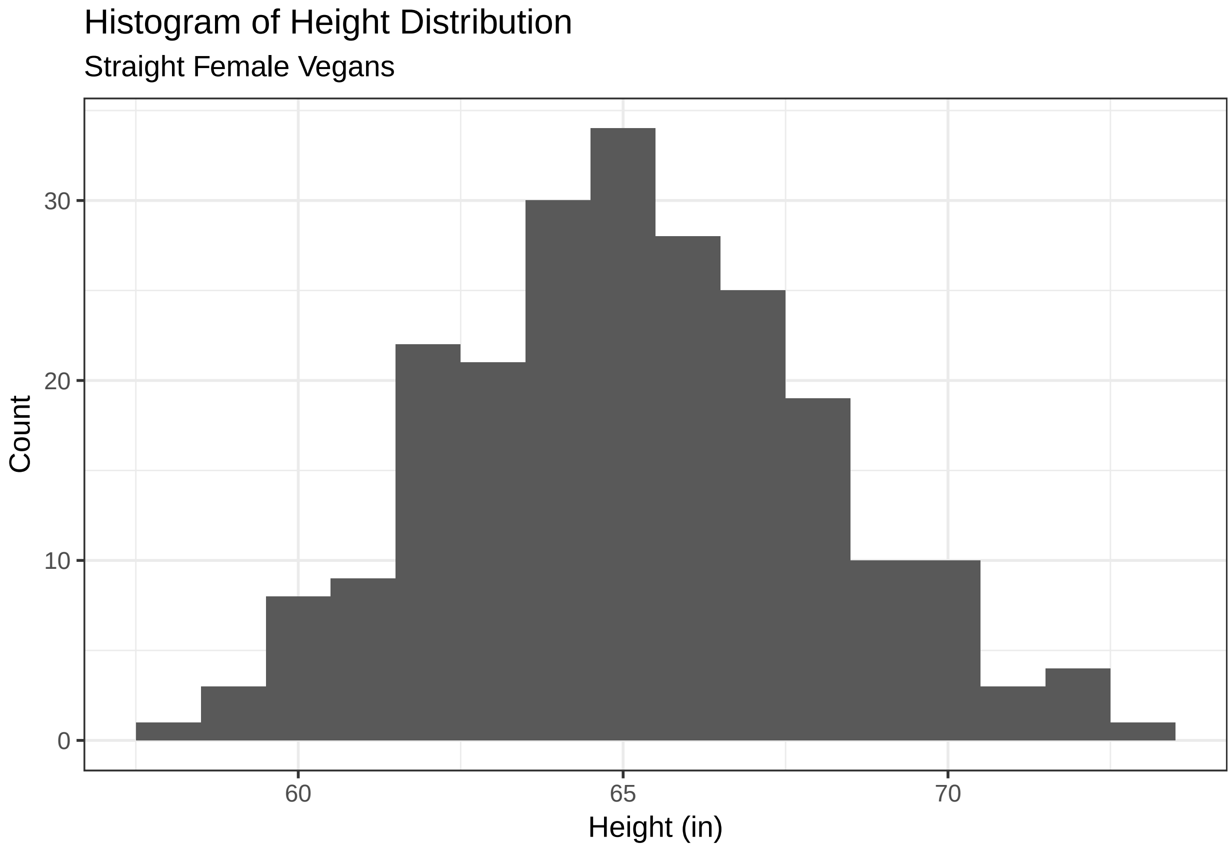
```
# histogram of ages for straight women who are vegans
# plot age distribution of Jason's demographic target

# filter na ages
ggplot(jason_data %>% filter(!is.na(age))) +
  geom_histogram(aes(x = age), binwidth = 1) +
  labs(title = 'Histogram of Age Distribution',
       subtitle = 'Straight Female Vegans',
       x = 'Age (yrs)',
       y = 'Count') +
  theme_bw()
```



2).

I expected the distribution of heights for straight women who are vegans would be pretty unimodal and normal, with a mean and median around 65in and a sd less than 5in. I expected that it would not differ from the distribution of women overall, as it seems unlikely that there would be any causal factor making straight or vegan women's heights significantly different from the overall population. The histogram fit very closely with the expectations, showing a very unimodal normal distribution with a mean of 65.08in, a median of 65in and a sd of 2.91in. This also closely matched the data for non-vegan straight women's heights (mean 65.17in, median 65in, sd 2.88in).

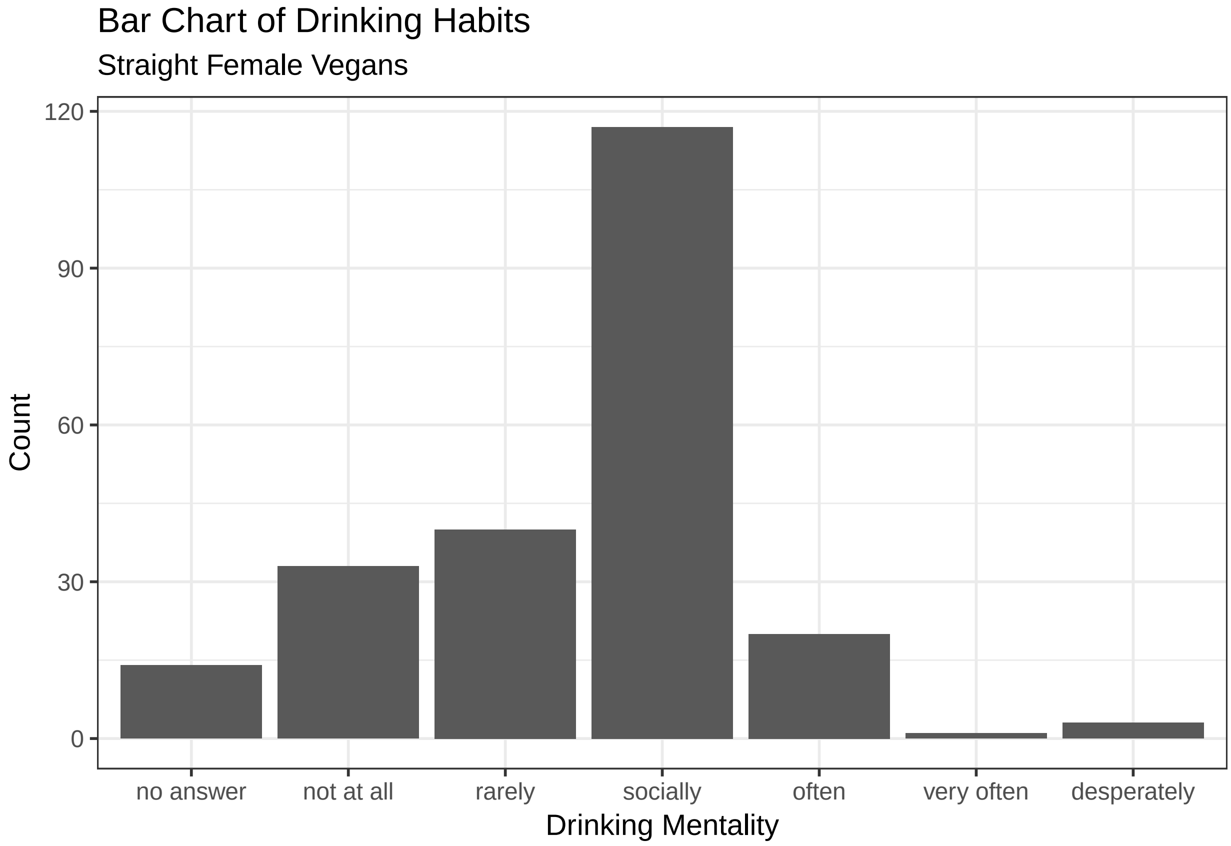
```r
# histogram of heights for straight women who are vegans
# plot height distribution of Jason's demographic target

# filter na heights
ggplot(jason_data %>% filter(!is.na(height))) +
  geom_histogram(aes(x = height), binwidth = 1) +
  labs(title = 'Histogram of Height Distribution',
       subtitle = 'Straight Female Vegans',
       x = 'Height (in)',
       y = 'Count') +
  theme_bw()
```



3).

I expected the chart of drinking preference to have multiple modes corresponding to groups that never drink and who drink often. I didn't expect there to be any single category that was dominant, because the interpretation of the categories will effect where people end up placing themselves, which seems pretty subjective. The chart ended up being centered around the 'social drinking' category, with people who drink rarely to the left and people who drink often to the right. It was actually pretty unimodal in the end, although some users did not answer this question.

```r
# bar chart of drinking habits for straight women who are vegans
# plot drinking habits chart of Jason's demographic target

# add 'no answer' for na value of drinks field to data set with mutate
ggplot(jason_data %>% mutate(drinks = ifelse(is.na(drinks), 'no answer', drinks))) +
  geom_bar(aes(x = drinks)) +
  # set logic order to show categories (default is alphabetical)
  scale_x_discrete(limits = c('no answer',
                              'not at all',
                              'rarely',
                              'socially',
                              'often',
                              'very often',
                              'desperately')) +
  labs(title = 'Bar Chart of Drinking Habits',
       subtitle = 'Straight Female Vegans',
       x = 'Drinking Mentality',
       y = 'Count') +
  theme_bw()
```

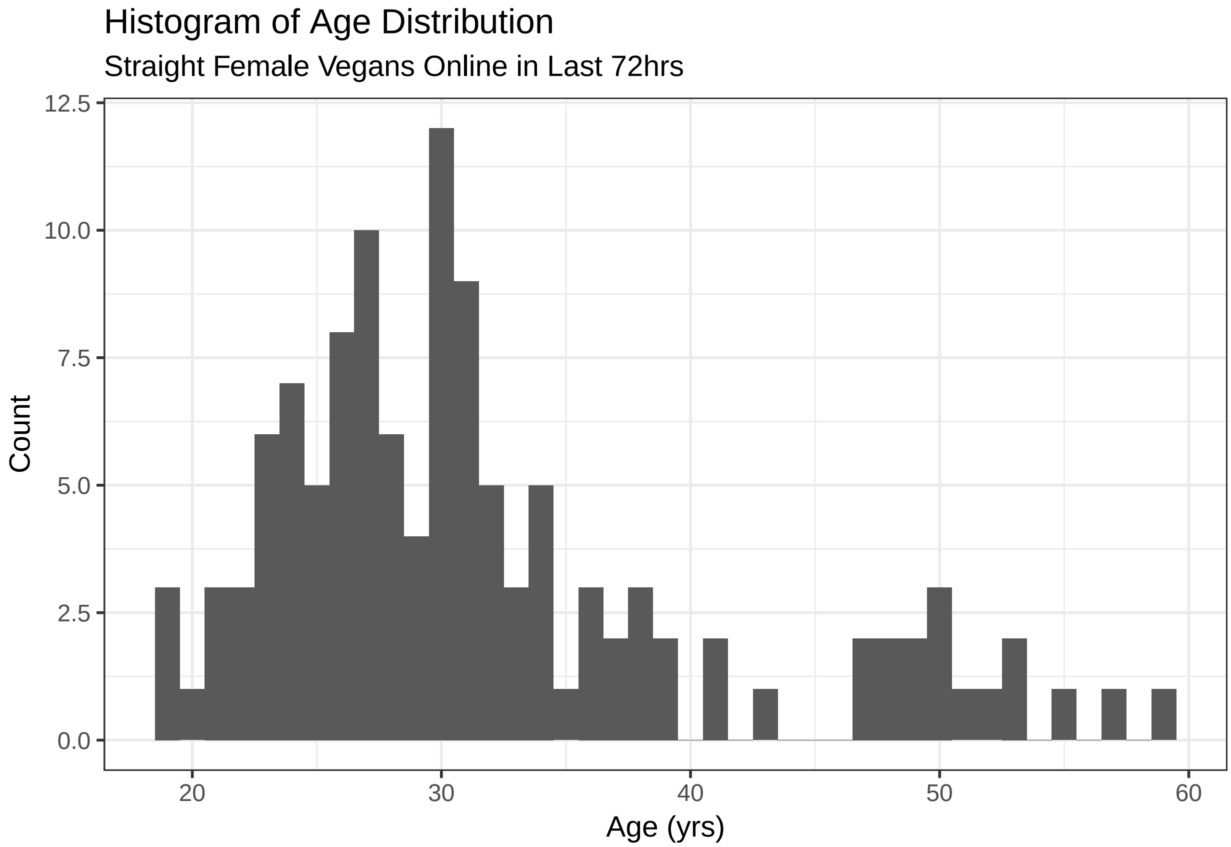


Q1 B).

```
# Filter Jason's demographic dataset to include only individuals who have signed on to okcupid within t

# create jason_filtered with profiles of straight female vegans
# who have been online over the 3 days prior to the last day in the dataset
jason_3_days <- profiles %>%
  filter(sex == 'f',
         orientation == 'straight',
         diet %in% c('vegan', 'strictly vegan', 'mostly vegan'),
         last_online > (max(last_online) - days(3)))
```

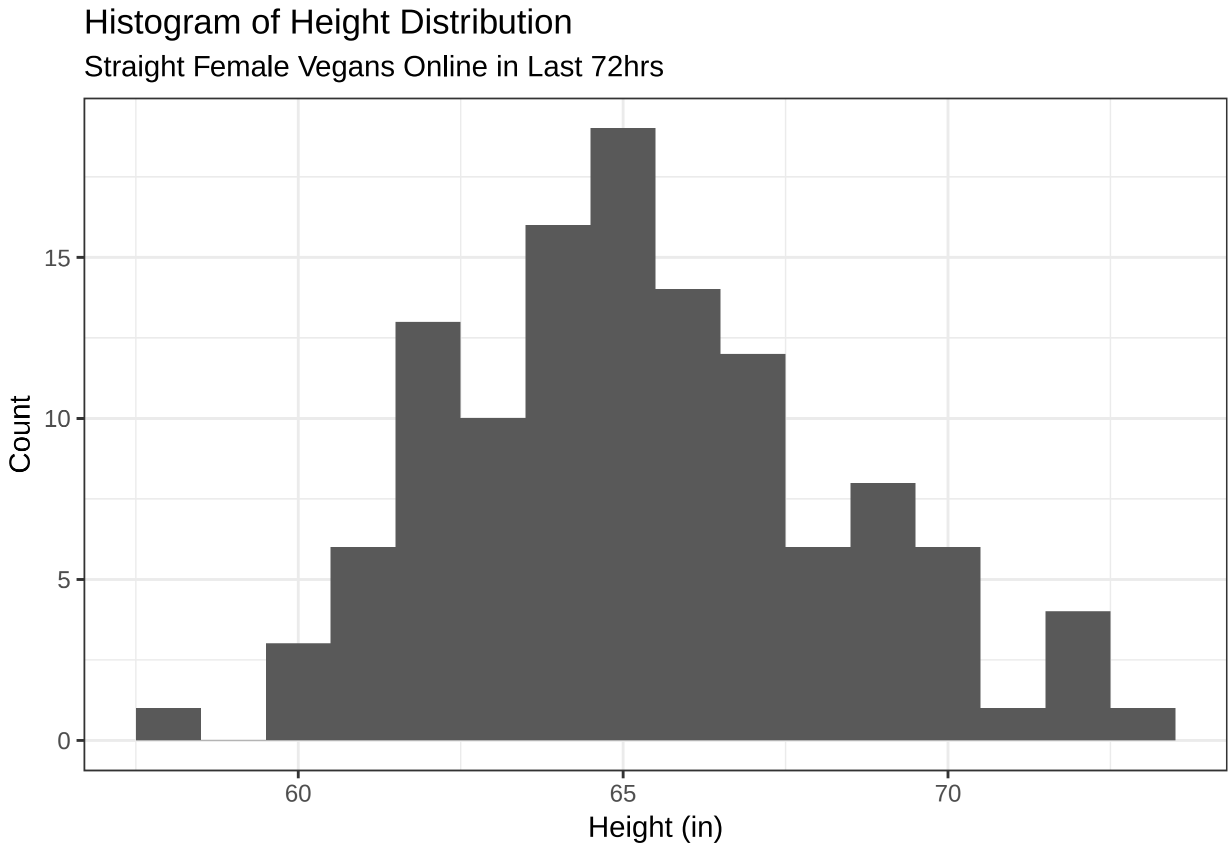
4).

```
# Histogram of ages for straight female vegans who have signed on in the last 3 days (of the dataset)
# plot age distribution of Jason's demographic target

# filter na ages
ggplot(jason_3_days %>% filter(!is.na(age))) +
  geom_histogram(aes(x = age), binwidth = 1) +
  labs(title = 'Histogram of Age Distribution',
       subtitle = 'Straight Female Vegans Online in Last 72hrs',
       x = 'Age (yrs)',
       y = 'Count') +
  theme_bw()
```



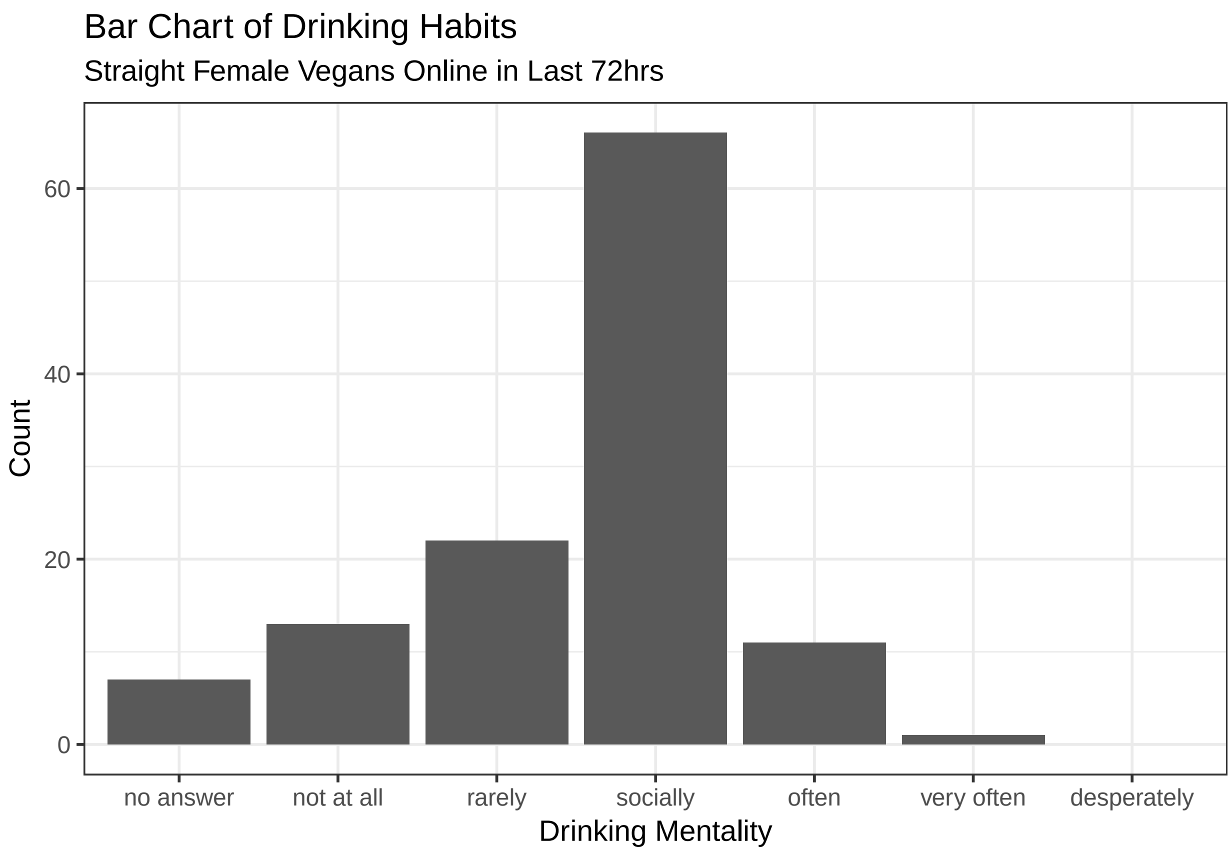
5).

```
# Histogram of heights for straight female vegans who have signed on in the last 3 days (of the dataset
# plot height distribution of Jason's demographic target

# filter na heights
ggplot(jason_3_days %>% filter(!is.na(height))) +
  geom_histogram(aes(x = height), binwidth = 1) +
  labs(title = 'Histogram of Height Distribution',
       subtitle = 'Straight Female Vegans Online in Last 72hrs',
       x = 'Height (in)',
       y = 'Count') +
  theme_bw()
```



6).

```r
# chart of straight female vegan's drink preferences who have signed on in the last 3 days (of the data
# plot drinking habits chart of Jason's demographic target

# add 'no answer' for na value of drinks field to data set with mutate
ggplot(jason_3_days %>% mutate(drinks = ifelse(is.na(drinks), 'no answer', drinks))) +
  geom_bar(aes(x = drinks)) +
  # set logic order to show categories (default is alphabetical)
  scale_x_discrete(limits = c('no answer',
                              'not at all',
                              'rarely',
                              'socially',
                              'often',
                              'very often',
                              'desperately')) +
  labs(title = 'Bar Chart of Drinking Habits',
       subtitle = 'Straight Female Vegans Online in Last 72hrs',
       x = 'Drinking Mentality',
       y = 'Count') +
  theme_bw()
```



Q2).

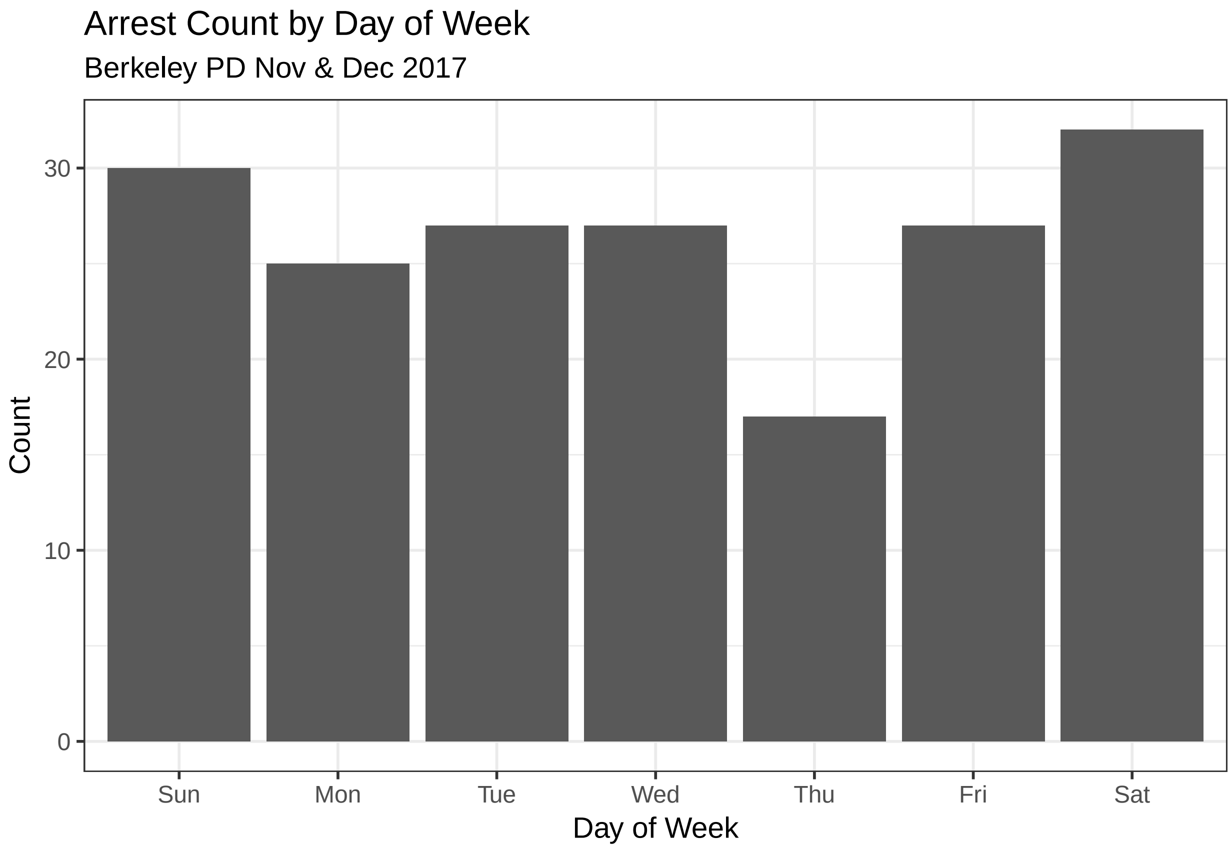
```r
#load data from the police department in the city of Berkeley and covert character fields DOB and time

# load berkeley PD log arrest data, convert character fields into dates
Berkeley_Arrests <- read_csv("Berkeley_PD_Log_-_Arrests.csv") %>%
  mutate(DOB = mdy(`Date of Birth`),
         DT = mdy_hm(`Date and Time`, tz = 'America/Los_Angeles')) %>%
  select(-`Date of Birth`, -`Date and Time`)
```

7).

```r
# Identify the day of the week with the highest arrest count

# Query to get day of week with most arrests
Arrest_Days <- Berkeley_Arrests %>%
  mutate(Arrest_Day_of_Week = wday(DT, label = T)) %>%
  group_by(Arrest_Day_of_Week) %>%
  summarize(Arrest_Count = n()) %>%
  # arrange days by number of arrest to show day with most arrests at the top
  arrange(desc(Arrest_Count))

Arrest_Days
```

```
## # A tibble: 7 x 2
##   Arrest_Day_of_Week Arrest_Count
##   <ord>                     <int>
## 1 Sat                          32
## 2 Sun                          30
## 3 Tue                          27
## 4 Wed                          27
## 5 Fri                          27
## 6 Mon                          25
## 7 Thu                          17
```

```
# Create chart of arrests by day of week

# Plot bar chart using Arrest_Days
ggplot(Arrest_Days) +
  geom_col(aes(x = Arrest_Day_of_Week, y = Arrest_Count)) +
  labs(title = 'Arrest Count by Day of Week',
       subtitle = 'Berkeley PD Nov & Dec 2017',
       x = 'Day of Week',
       y = 'Count') +
  scale_x_discrete(limits = c('Sun', 'Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat')) +
  theme_bw()
```



8).

```
# Age of individuals at the time of their arrest

# Add a column for the age of individuals at time of arrest using the difference between DT and DOB

Berkeley_Arrests <- Berkeley_Arrests %>%
  # divide the days between birth and arrest by the length of 1 year and round down
  mutate('Age_At_Arrest' = round(difftime(DT, DOB, units="days") / 365.25))
```

9).

```
# List of individuals whose recorded ages do not match their age at the time of arrest

# Create a list of names of individuals for whome 'Age' does not equal 'Age_At_Arrest'

# Filter individuals with arrest ages and recorded ages that differ by more than 1 year and select thei
Wrong_Age <- Berkeley_Arrests %>%
  filter(abs(Age - Age_At_Arrest) > 1) %>%
  select(Subject)

as.list(Wrong_Age)
```

```
## $Subject
##  [1] "Herbert Stephen Blue"          "Michael Joseph May"
##  [3] "Bowen Chen"                    "Anthony Wilfred Kerman"
##  [5] "Damon Lamont Jones"            "Gerald Arcos"
##  [7] "Sara Sofija Antunovich"        "Andrew Francis Supple"
##  [9] "Scotty Emmanuel Guess"         "CHRISTOPHER RANDOLPH TORRENCE"
## [11] "Edward Rae Mitchell"           "Jesse Vander Weele"
## [13] "Adan Mora Morfin"              "Mehdi Cherfaoui"
## [15] "LUIZA RENATA MOTTER"           "Christopher Cole Tabor"
## [17] "Daniel James Blackbear"        "Fredrick Arzell Chisom"
## [19] "Adam Kenneth Jones"            "PRICE WHEELER"
## [21] "Louis Joseph Lawyer"           "Nicholas M Shelby"
## [23] "JUAN BAUTISTA CHAVEZ WOLFE"
```