Duncan McKinnon

West

W 1

QA

Load Packages and Data

```
suppressPackageStartupMessages({
  library(tidyverse)
  library(okcupiddata)
})
```

1).

```
# get top ten cities by size

locations <- profiles %>%
  group_by(location) %>%
  summarize(num = n()) %>%
  separate(location, into = c('city', 'state'), sep = ', ', remove = TRUE) %>%
  arrange(desc(num))
```

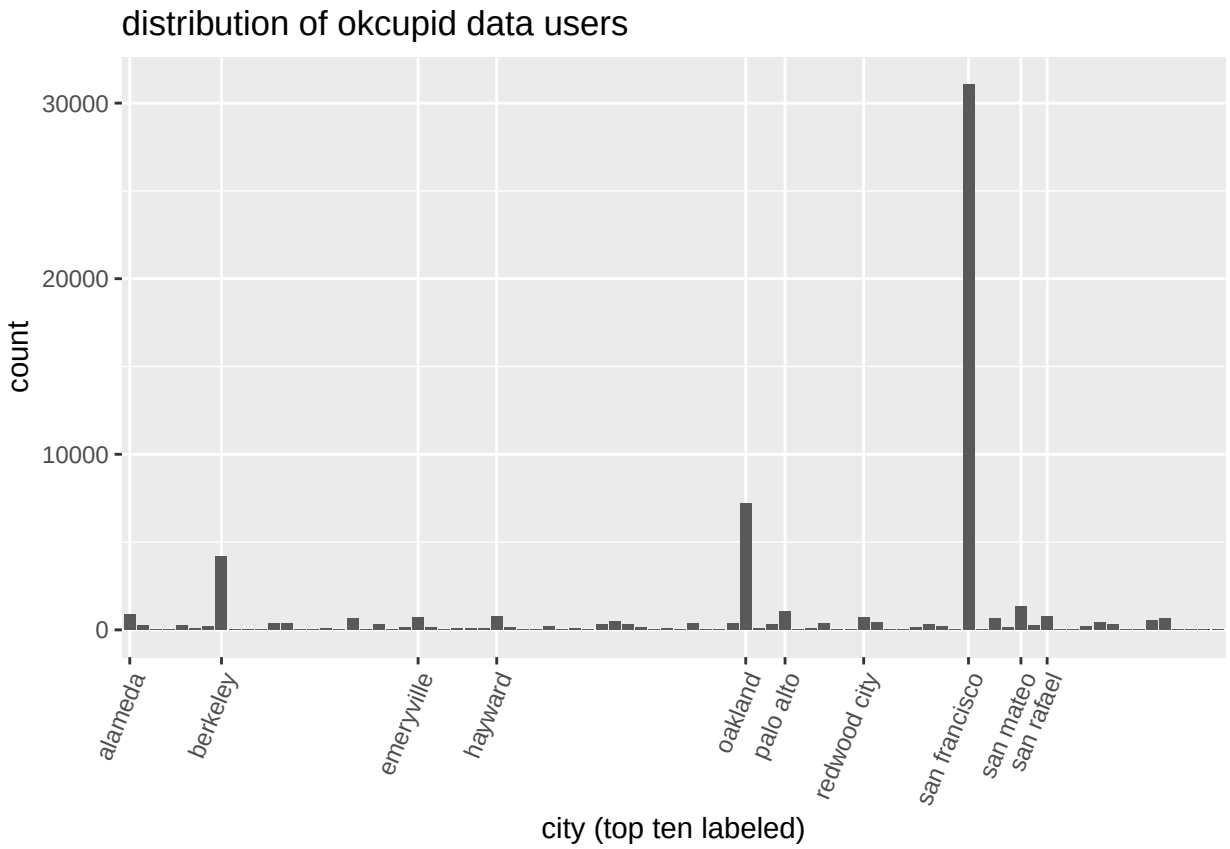
```
## Warning: Expected 2 pieces. Additional pieces discarded in 1 rows [191].
```

```
top_ten <- locations$city[1:10]
top_ten
```

```
##  [1] "san francisco" "oakland"       "berkeley"      "san mateo"
##  [5] "palo alto"     "alameda"       "san rafael"    "hayward"
##  [9] "emeryville"    "redwood city"
```

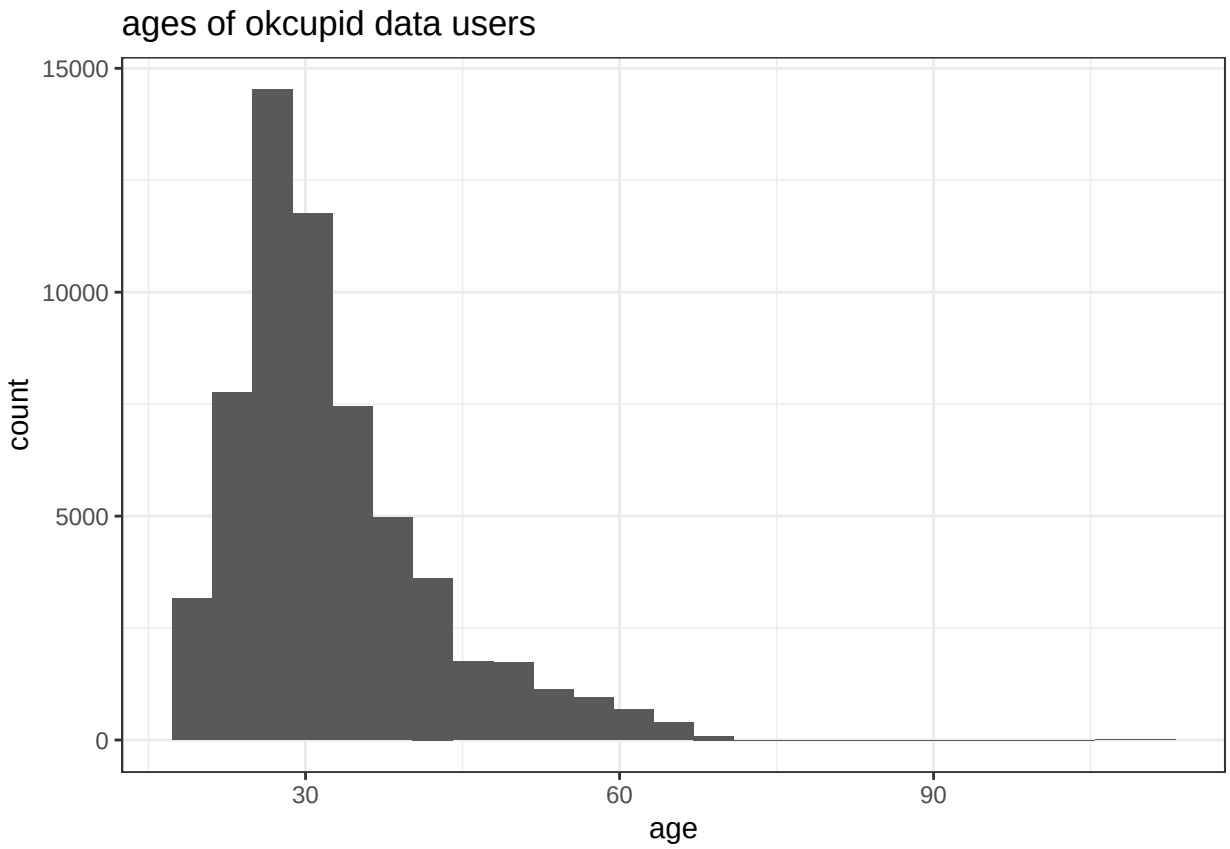
2).

```
ggplot(locations %>% filter(num > 2)) +
  geom_col(aes(x = city, y = num)) +
  scale_x_discrete(breaks = top_ten, labels = top_ten) +
  labs(title = 'distribution of okcupid data users',
       x = 'city (top ten labeled)',
       y = 'count') +
  theme(axis.text.x = element_text(angle = 67.5, hjust = 1))
```



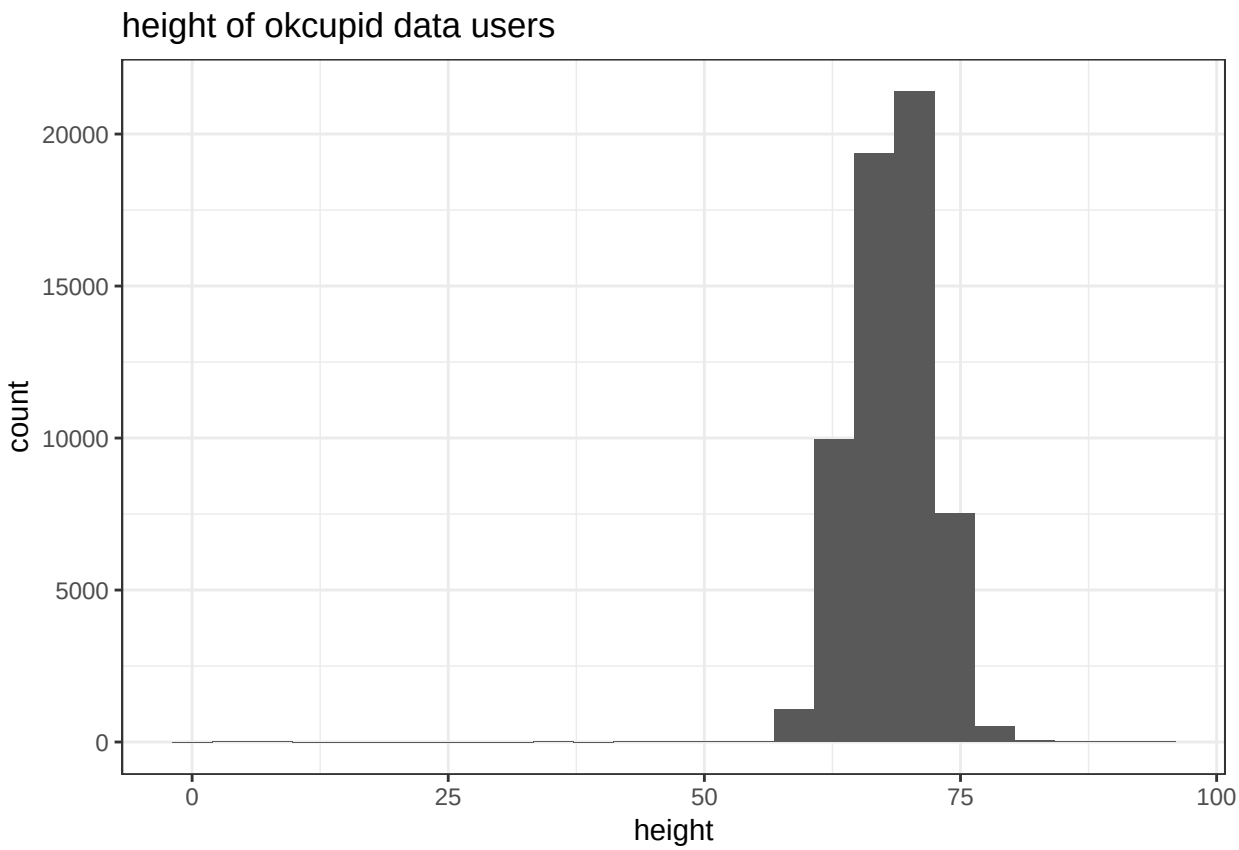
3).

```
ggplot(profiles %>% filter(!is.na(age))) +
  geom_histogram(aes(x = age), bins = 25) +
  labs(title = 'ages of okcupid data users',
       x = 'age',
       y = 'count') +
  theme_bw()
```



4).

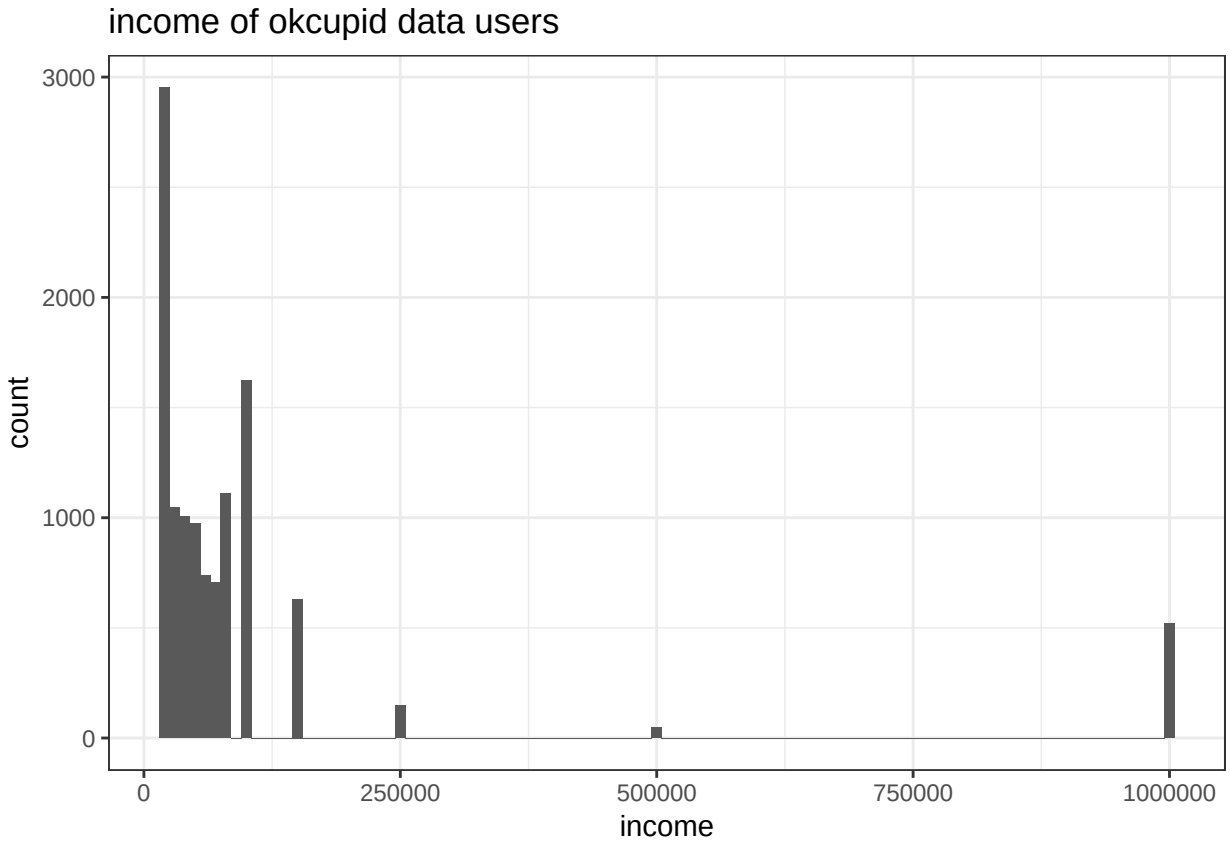
```
ggplot(profiles %>% filter(!is.na(height))) +
  geom_histogram(aes(x = height), bins = 25) +
  labs(title = 'height of okcupid data users',
       x = 'height',
       y = 'count') +
  theme_bw()
```



5).

Height in this dataset is very normally distributed with a mean around 72 inches (6 ft). While this mean is above what would be expected for a mixed population of men and women we could probably expect people to give optimistic heights up to at least 6ft.

```
ggplot(profiles %>% filter(!is.na(income))) +
  geom_histogram(aes(x = income), binwidth = 10000) +
  labs(title = 'income of okcupid data users',
       x = 'income',
       y = 'count') +
  theme_bw()
```



6).

Income was pretty sparse in this dataset, with the majority of participants likely opting out. The distribution of incomes is very right skewed, with a high concentration between 20k and 100k and then smaller peaks around 250k, 500k. The right skew is again effected by the hard lower limit of $0 income. The maximum entry of 1m is a lot more frequent than 250k or 500k, so it seems probable that there was also a hard upper limit in the question, or even a set of categories to choose from.