Duncan McKinnon

West

HW 2

QA

Load Packages and Data

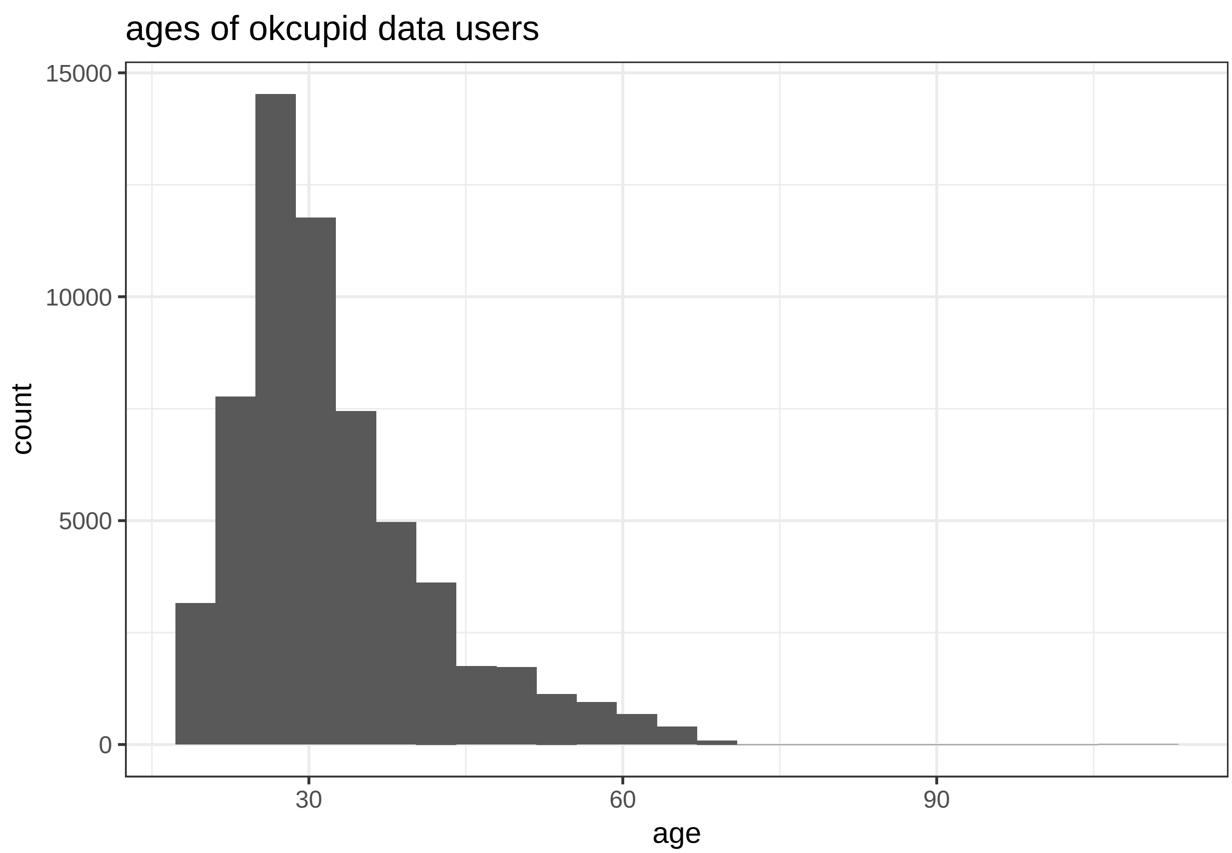
```
suppressPackageStartupMessages({
  library(tidyverse)
  library(okcupiddata)
})

# Load Houses dataset with messages suppressed
Houses <- read_csv("Houses.csv", col_names = T)
```

1).

Age in the okcupid dataset is relatively normally distributed with a mean of about 30. The distribution may be slightly right-skewed, which can be attributed to the fact that there is a hard left limit at 0, while the right is unbounded (although recorded human lifespan has been consistently finite).

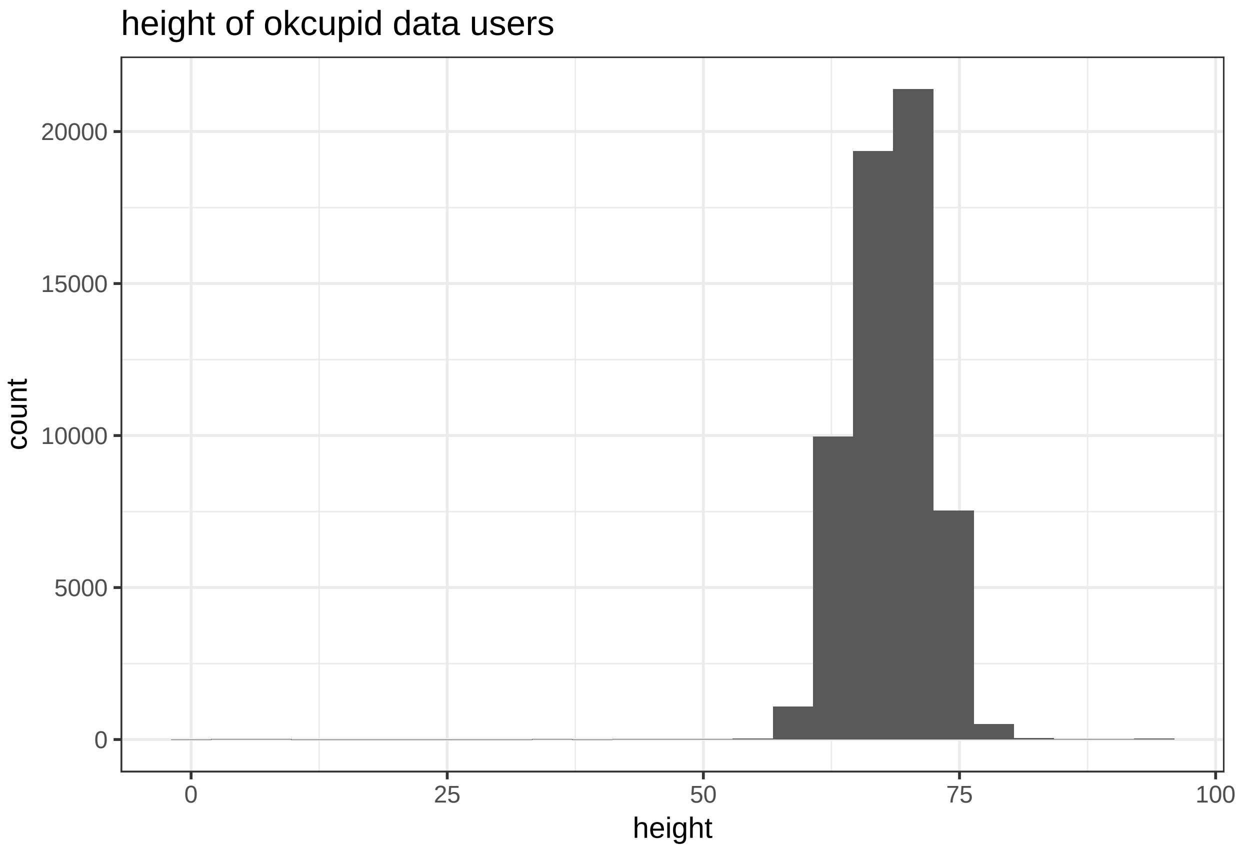
```
# Filter NA's and plot histogram of age data
ggplot(profiles %>% filter(!is.na(age))) +
  geom_histogram(aes(x = age), bins = 25) +
  labs(title = 'ages of okcupid data users',
       x = 'age',
       y = 'count') +
  theme_bw()
```



2).

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.

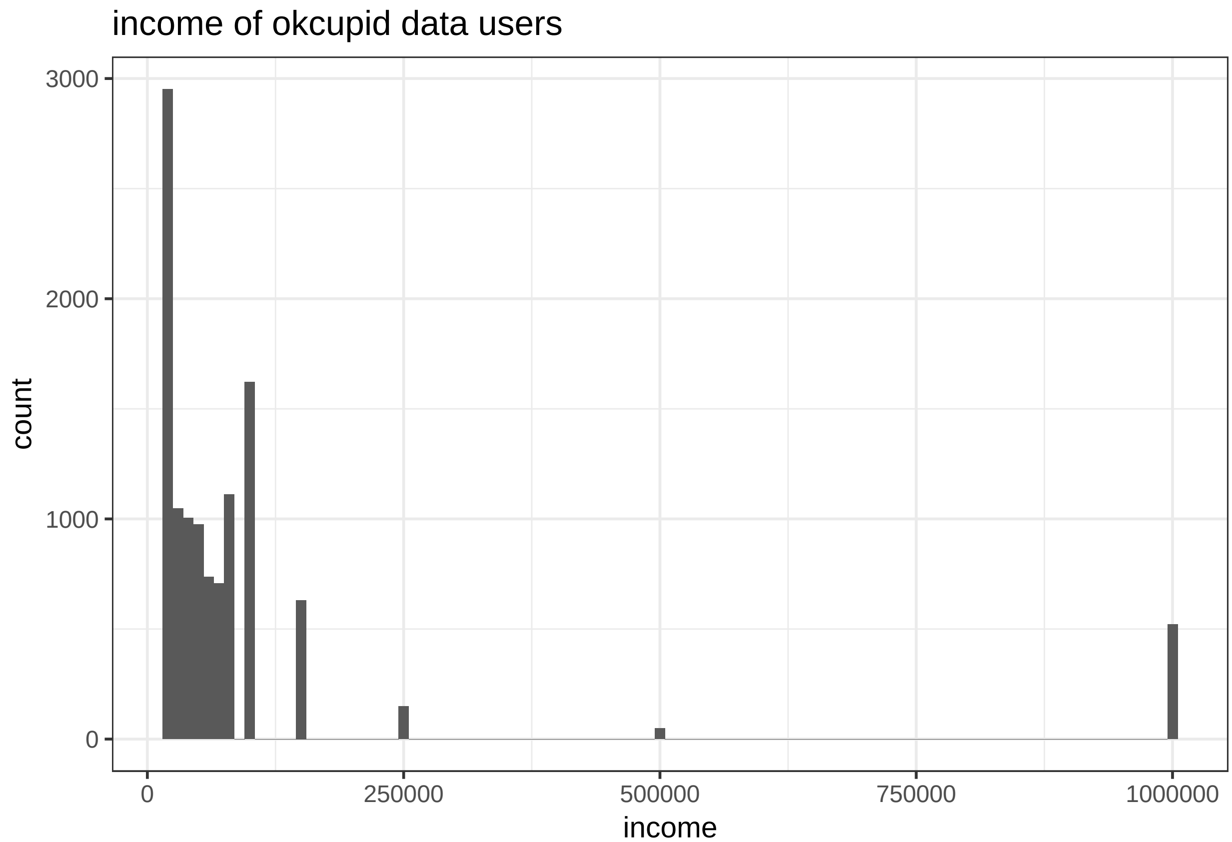
```r
# Filter NA's and plot histogram of height data
ggplot(profiles %>% filter(!is.na(height))) +
  geom_histogram(aes(x = height), bins = 25) +
  labs(title = 'height of okcupid data users',
       x = 'height',
       y = 'count') +
  theme_bw()
```



3).

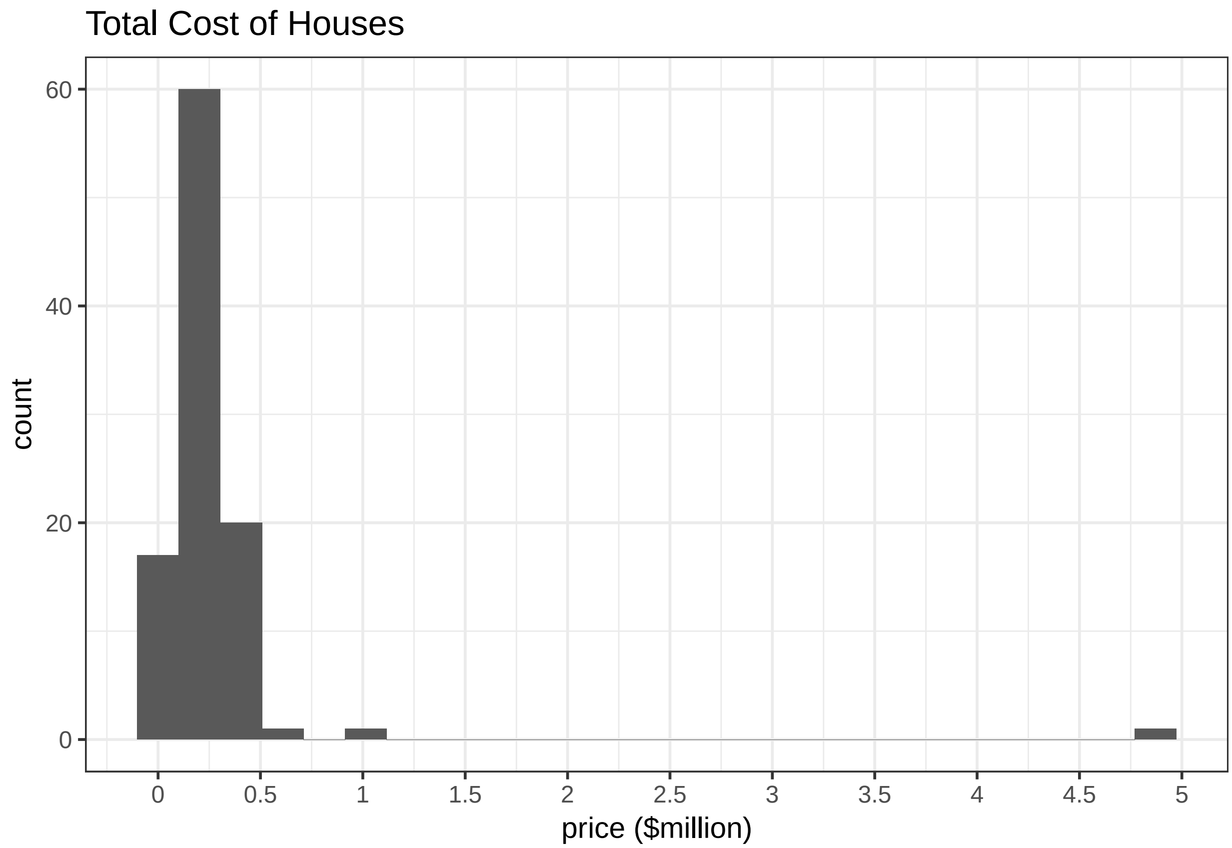
Income was pretty sparse in this dataset, with the majority of participants likely opting out. The distribution of incomes is very right skewed, with a high concentration between 20k and 100k and then smaller peaks around 250k, 500k. The right skew is again effected by the hard lower limit of $0 income. 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.

```
# Filter NA's and plot histogram of income data
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()
```



QB

```
# Remove NA's and plot histogram of Total
ggplot(Houses %>% filter(!is.na(Total))) +
  geom_histogram(aes(x = Total), bins = 25) +
  labs(title = 'Total Cost of Houses',
       x = 'price ($million)',
       y = 'count') +
  scale_x_continuous(breaks = seq(-5e5, 5.5e6, 5e5), labels=paste(seq(-0.5,5.5,0.5))) +
  theme_bw()
```



1).

The addresses of the two outlier homes with Total prices greater than $1 million are 2029 Alston Ave. Cary, NC 27519 and 2211 Byrd St. Raleigh, NC 27608. Both homes are located in North Carolina, in the grater Raleigh area.

```
# filter homes that cost more that $1million and select their addresses
Outliers_Total <- Houses %>% filter(Total > 1e6) %>% select(Address, Zip, Total)
Outliers_Total
```

```
## # A tibble: 2 x 3
##   Address           Zip   Total
##   <chr>           <dbl>   <dbl>
## 1 2029 Alston Ave 27519 4904102
## 2 2211 Byrd St    27608 1113750
```

2).

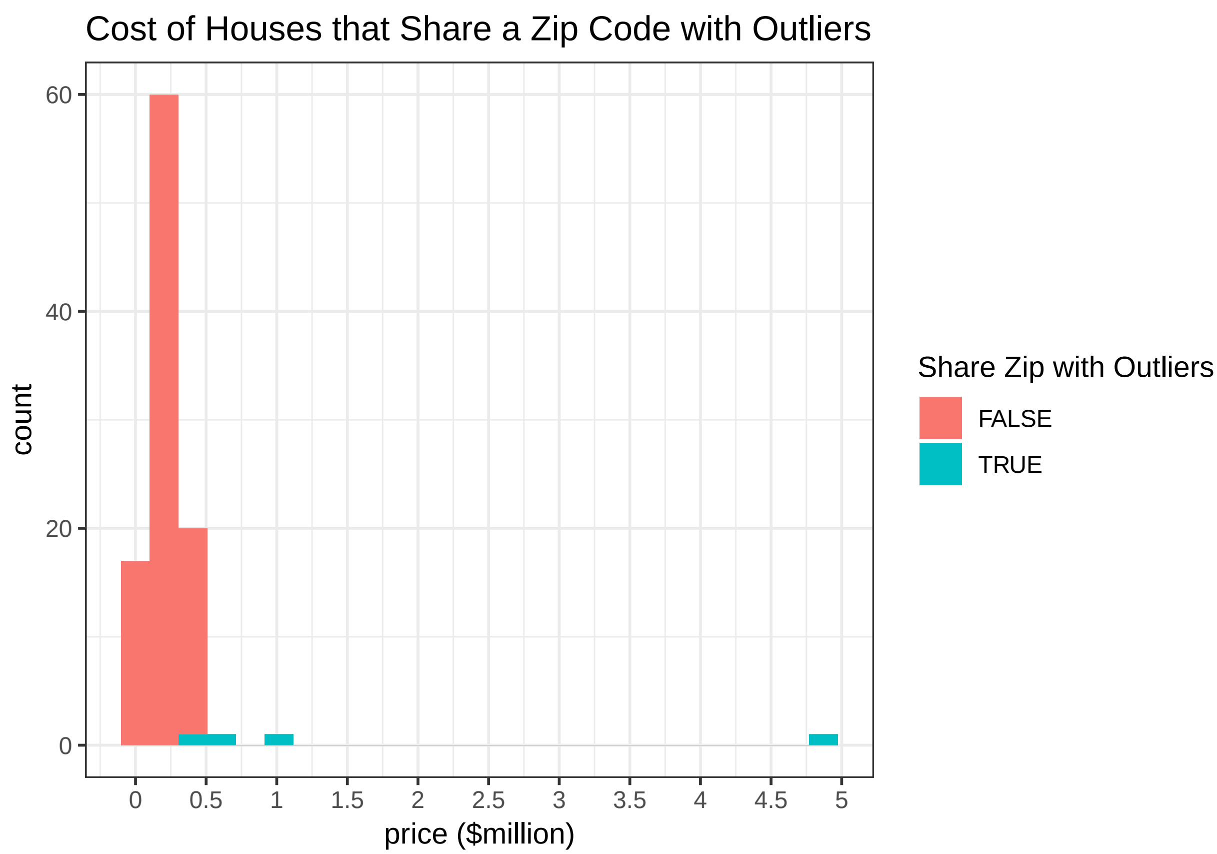
Exploratory Analysis of Outliers in Houses dataset

a).

The histogram below shows the total costs of homes in this dataset, with homes in either of the zip codes that contain the outliers colored in blue and homes that do not share a zip code with the outlier in pink. From this chart we can see that there are only 2 homes that share a zip code with the outliers. Interestingly, both these homes are above the median home prices. In fact they represent the 3rd and 6th most expensive properties in this dataset.

```r
## First lets look at the distribution of homes that are in the same area code as the outliers
# Color code bars by whether the homes share a zip code with the outliers

# Plot histogram of total price with different colors for houses in the Zips that contain outliers
ggplot(Houses) +
  geom_histogram(aes(x = Total, fill = Zip %in% Outliers_Total$Zip), bins = 25) +
  labs(title = 'Cost of Houses that Share a Zip Code with Outliers',
       x = 'price ($million)',
       y = 'count',
       fill = 'Share Zip with Outliers') +
  scale_x_continuous(breaks = seq(-5e5, 5.5e6, 5e5), labels=paste(seq(-0.5,5.5,0.5))) +
  theme_bw()
```




```r
## From the histogram we can see that 4 of the 6 most expensive houses are all from the same 2 zipcodes

top_ten <- Houses %>% arrange(desc(Total)) %>% top_n(10, Total)
top_ten
```

```
## # A tibble: 10 x 12
##     `ID#`  Year  SQFT Story Acres Baths Fireplaces  Total   land building
##     <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>      <dbl>  <dbl>  <dbl>    <dbl>
##  1  78570  2000  1404  1    39.4    2            0 4.90e6 4.80e6   106352
##  2     35  1989  4650  1.5   0.49   3            1 1.11e6 4.80e5   634230
##  3  24250  1952  2044  1     0.42   1.5          1 5.23e5 3.45e5   178826
```

```
## 4 44951 1924 1829 1.5 0.25 3 1 4.96e5 2.92e5 203925
## 5 187858 1992 4071 1.75 0.32 3 1 4.84e5 1.10e5 373528
## 6 198863 1994 3483 2.5 0.36 3 1 4.33e5 9.80e4 334516
## 7 174478 1994 2847 1.5 0.36 2 1 4.32e5 1.38e5 294023
## 8 151211 1986 2535 2 1.26 2.5 1 4.31e5 1.80e5 251468
## 9 176210 1989 2895 1.75 0.98 2 1 4.24e5 1.30e5 294333
## 10 128630 1984 3103 2 0.95 2.5 1 4.10e5 1.80e5 230025
## # ... with 2 more variables: Zip <dbl>, Address <chr>
```

b).

This plot compares three variables: the size in acres of the land on which the property is built, the size in square feet of the building(s) on the property, and the total price of the property. The color coding is used to easily spot the properties with costs that we consider outliers. By looking at this plot we can quickly see that one of the outliers represents the largest property by acreage (by a lot), and the other represents the largest property in square feet.

```
## Next lets compare the sqftage and acreage of the outlier homes with others in the dataset

# Create a scatterplot with the size in acres of the land on the x axis, the size in sq ft of the build
ggplot(Houses %>% filter(!is.na(SQFT) & !is.na(Acres) & !is.na(Total)))+
  geom_point(aes(x = Acres, y = SQFT, size = Total, color = Total %in% Outliers_Total$Total)) +
  labs(title = 'Size of Land and Sq Ft of Space by Total Price',
       x = 'no. acres',
       y = 'building sq ft.',
       size = 'price ($)',
       color = 'Outliers') +
  theme_bw()
```

c).

From this chart, we can see that the high prices of these outliers actually correlates to their size in either acres or square feet. Even though they are significantly more expensive than other entries, the prices of these two outliers reflect the value of square footage and acreage of properties in this dataset, and removing them would significantly limit what a model of this information would be able to fit/predict.