Introducing the data

EXPLORATORY DATA ANALYSIS IN R

Andrew Bray
Assistant Professor, Reed College

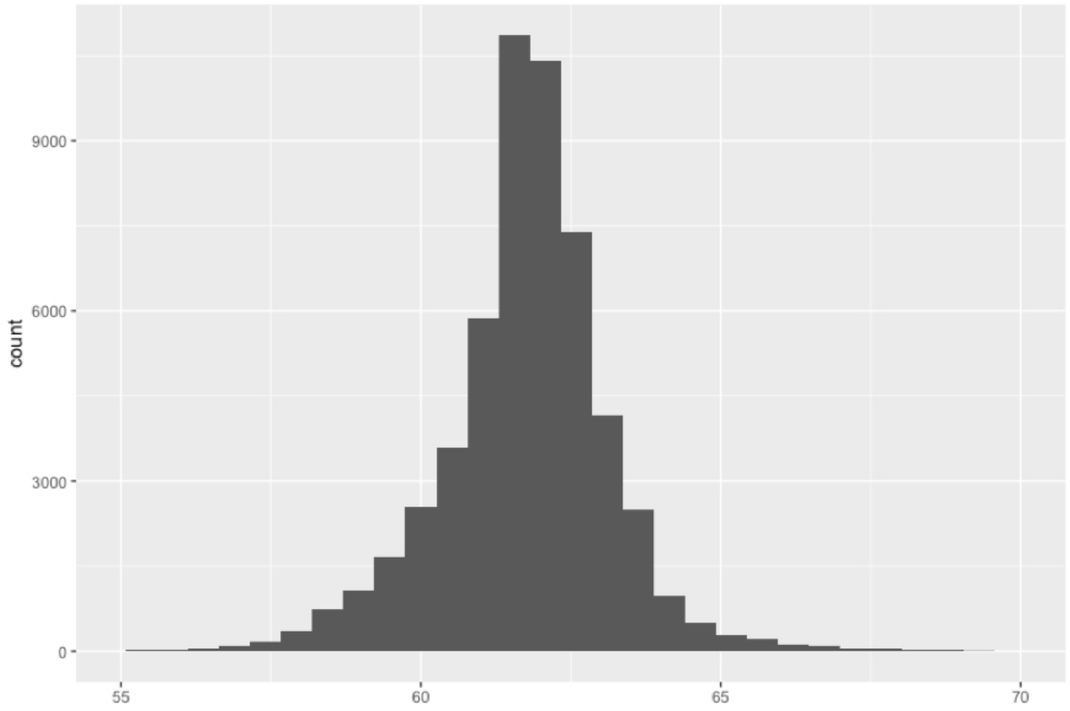
Email data set

```
email
```

```
# A tibble: 3,921 × 21
       spam to_multiple  from    cc sent_email                time image
     <fctr>       <dbl> <dbl> <int>      <dbl>              <dttm> <dbl>
1  not-spam           0     1     0          0 2012-01-01 01:16:41     0
2  not-spam           0     1     0          0 2012-01-01 02:03:59     0
3  not-spam           0     1     0          0 2012-01-01 11:00:32     0
4  not-spam           0     1     0          0 2012-01-01 04:09:49     0
5  not-spam           0     1     0          0 2012-01-01 05:00:01     0
6  not-spam           0     1     0          0 2012-01-01 05:04:46     0
7  not-spam           1     1     0          1 2012-01-01 12:55:06     0
8  not-spam           1     1     1          1 2012-01-01 13:45:21     1
9  not-spam           0     1     0          0 2012-01-01 16:08:59     0
10 not-spam           0     1     0          0 2012-01-01 13:12:00     0
# ... with 3,911 more rows, and 14 more variables: attach <dbl>,
#   dollar <dbl>, winner <fctr>, inherit <dbl>, viagra <dbl>,
#   password <dbl>, num_char <dbl>, line_breaks <int>, format <dbl>,
#   re_subj <dbl>, exclaim_subj <dbl>, urgent_subj <dbl>,
#   exclaim_mess <dbl>, number <fctr>
```

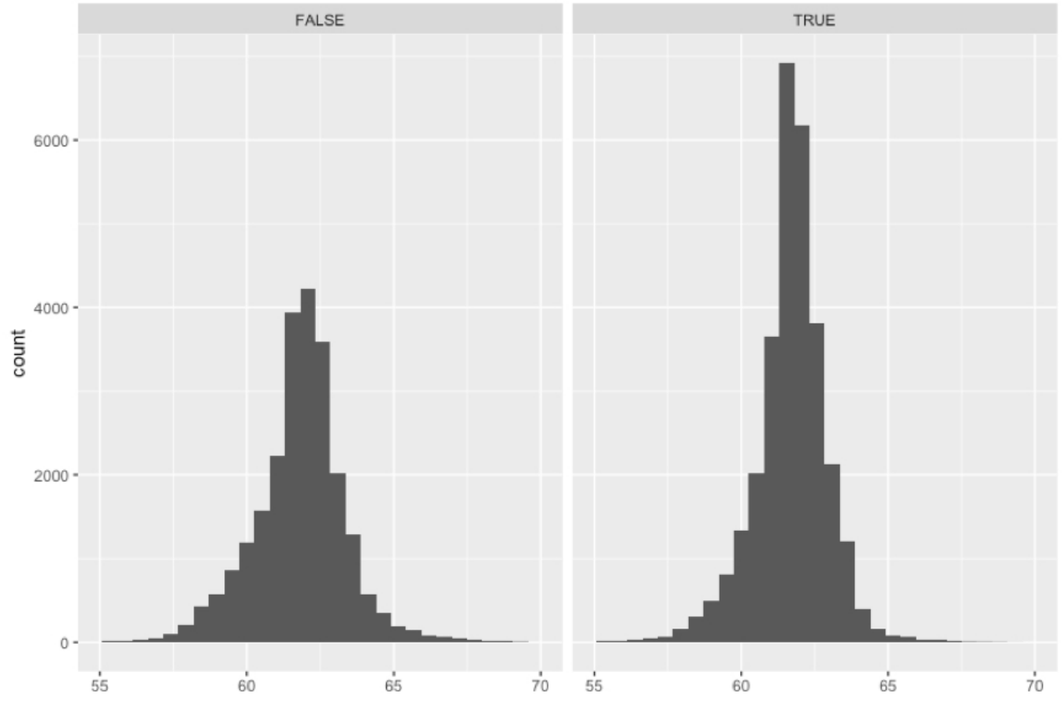
Histograms

```
ggplot(data, aes(x = var1)) +
  geom_histogram()
```



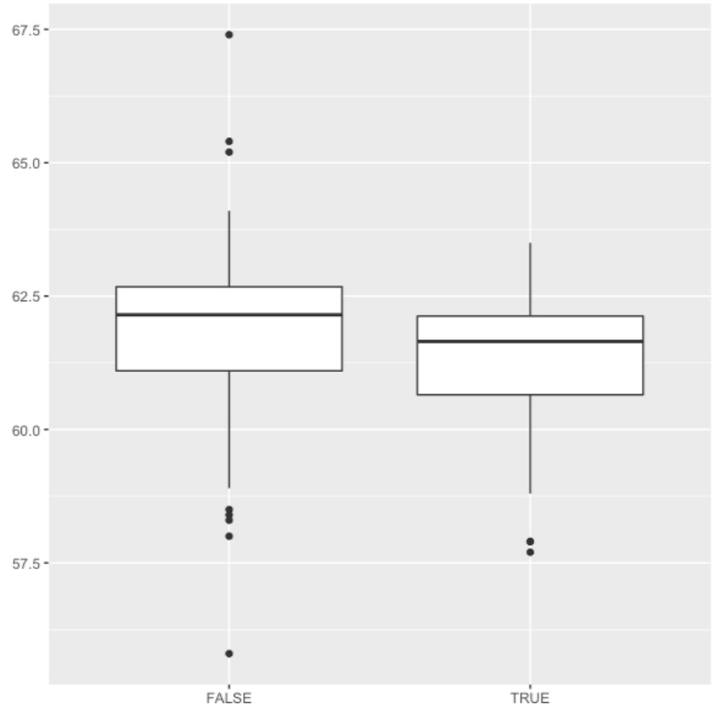
Histograms

```
ggplot(data, aes(x = var1)) +
  geom_histogram() +
  facet_wrap(~var2)
```



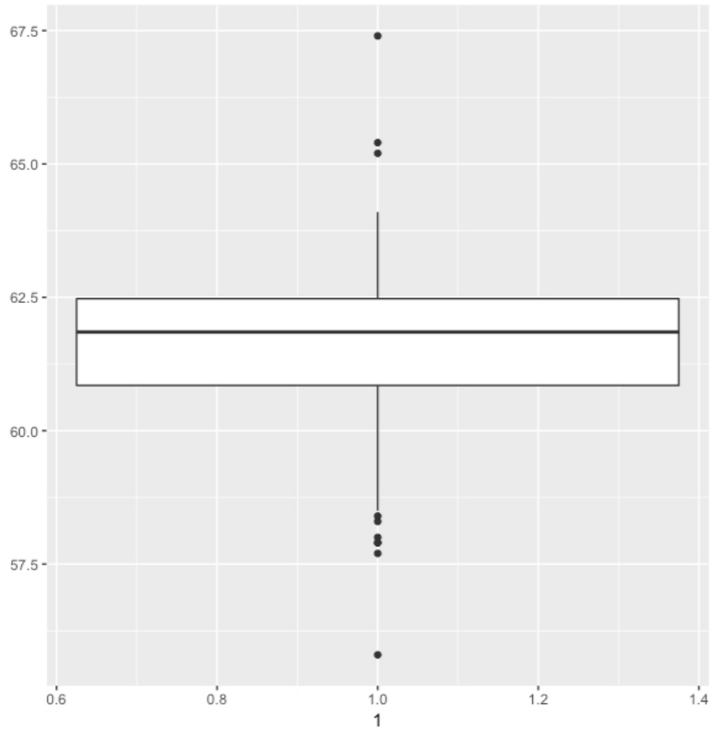
Boxplots

```
ggplot(data, aes(x = var2, y = var1)) +
  geom_boxplot()
```



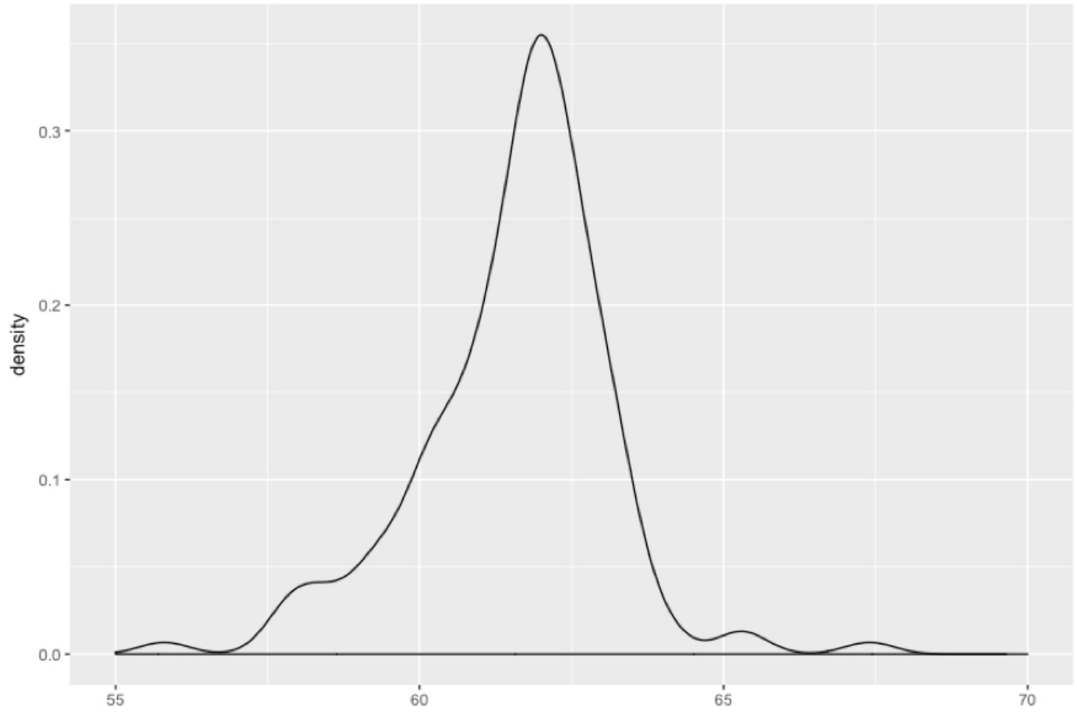
Boxplots

```
ggplot(data, aes(x = 1, y = var1)) +
  geom_boxplot()
```



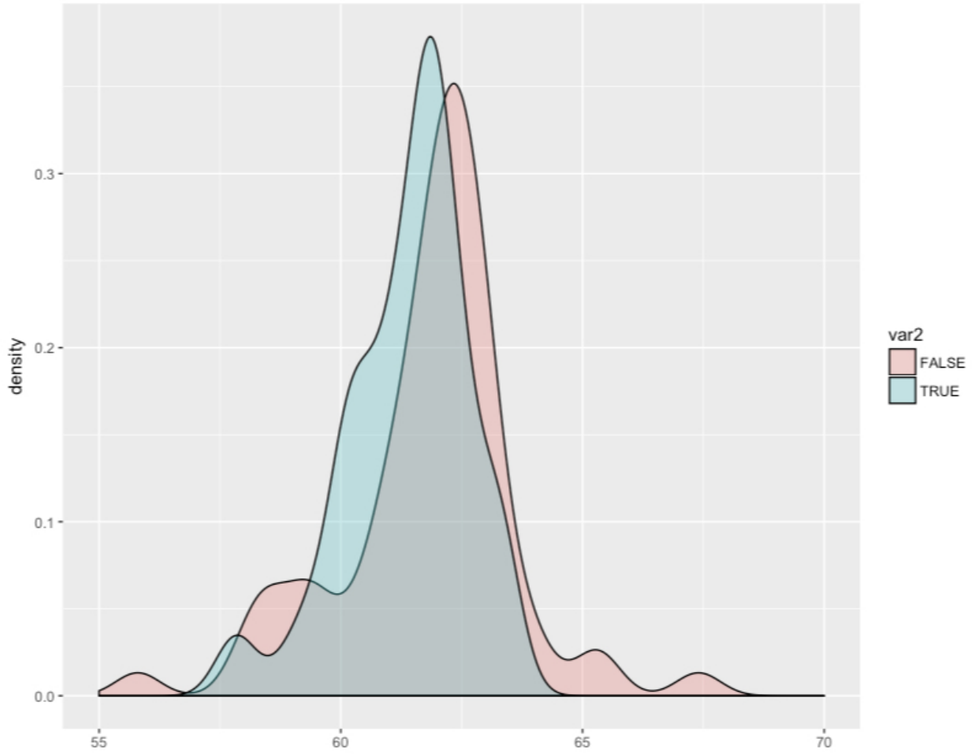
Density plots

```
ggplot(data, aes(x = var1)) +
  geom_density()
```



Density plots

```
ggplot(data, aes(x = var1, fill = var2)) +
  geom_density(alpha = .3)
```



Let's practice!

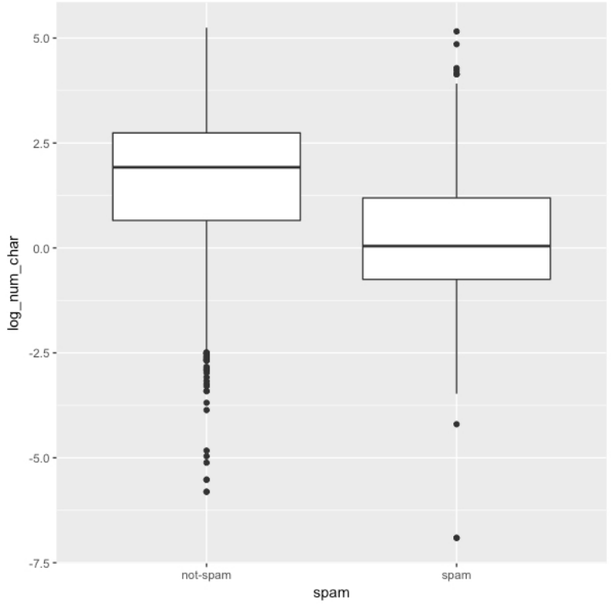
EXPLORATORY DATA ANALYSIS IN R

Check-in 1

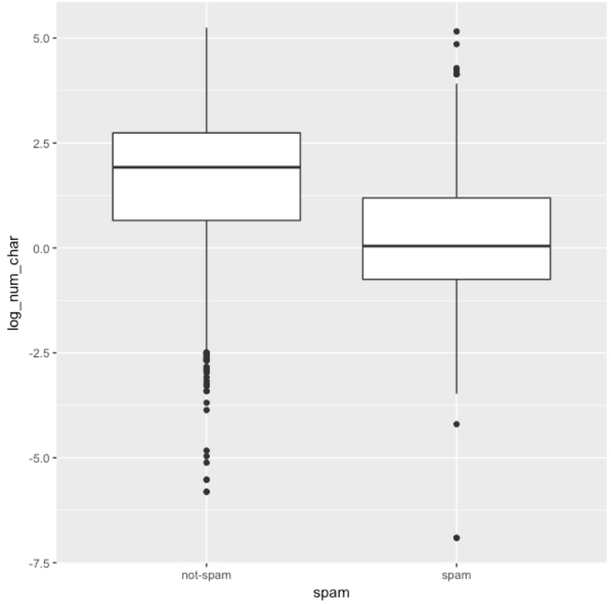
EXPLORATORY DATA ANALYSIS IN R

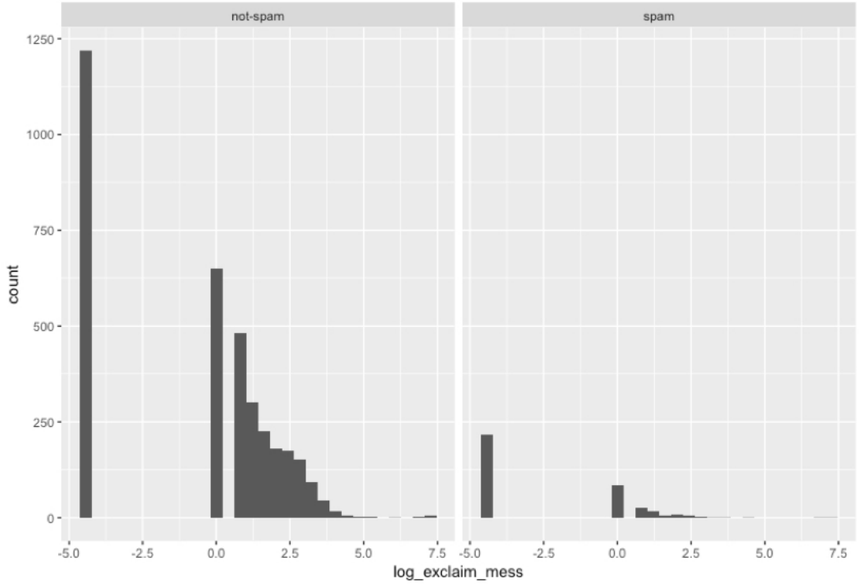
Andrew Bray
Assistant Professor, Reed College

Review



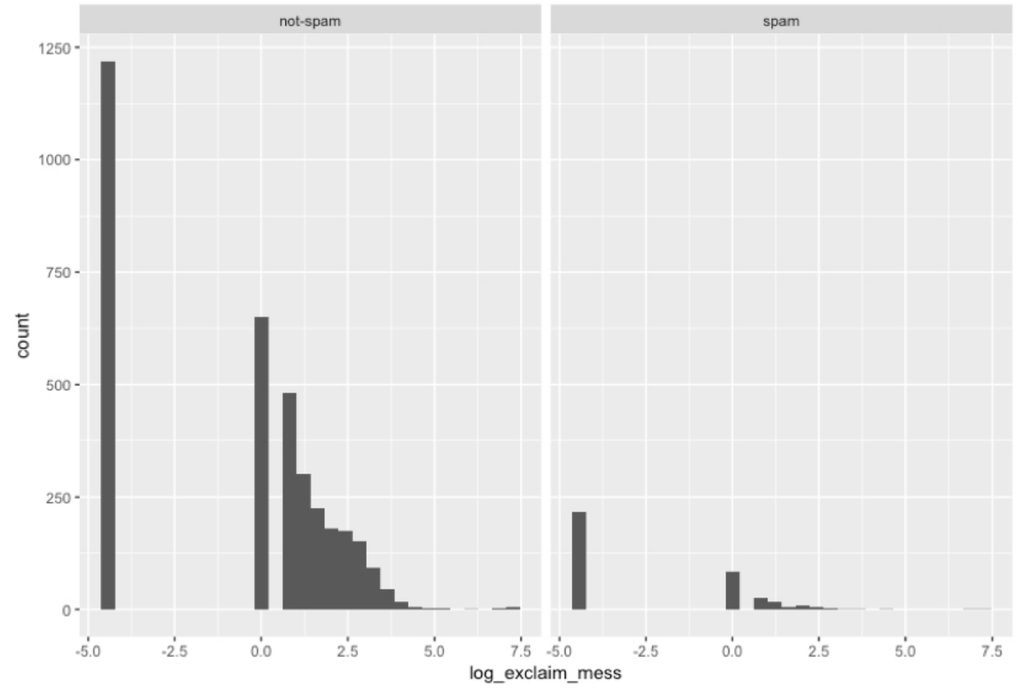
Review


5.0
2.5
0.0
-2.5
-5.0
-7.5
log_num_char
not-spam
spam
spam


not-spam
spam
1250
1000
750
500
250
0
count
-5.0
-2.5
0.0
2.5
5.0
7.5
log_exclaim_mess

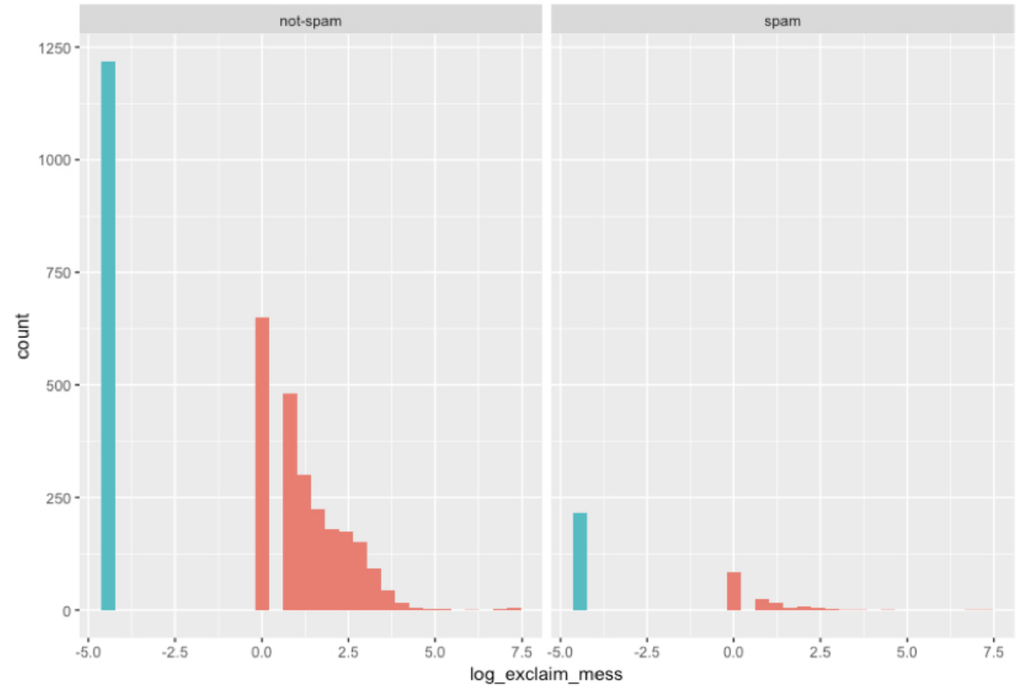
Zero inflation strategies

- Analyze the two components separately
- Collapse into two-level categorical variable



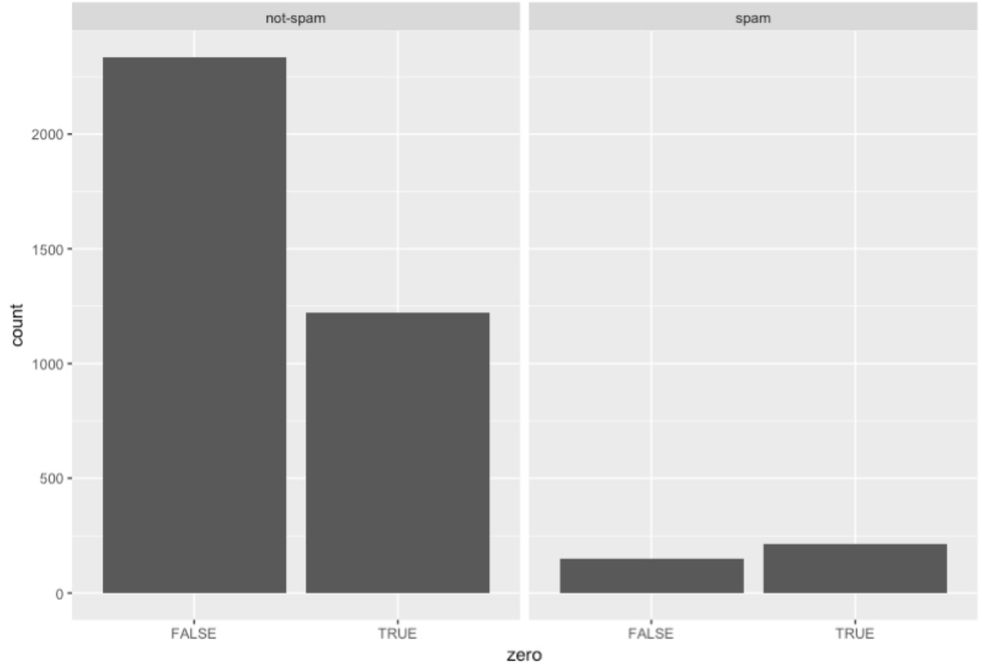
Zero inflation strategies

- Analyze the two components separately
- Collapse into two-level categorical variable



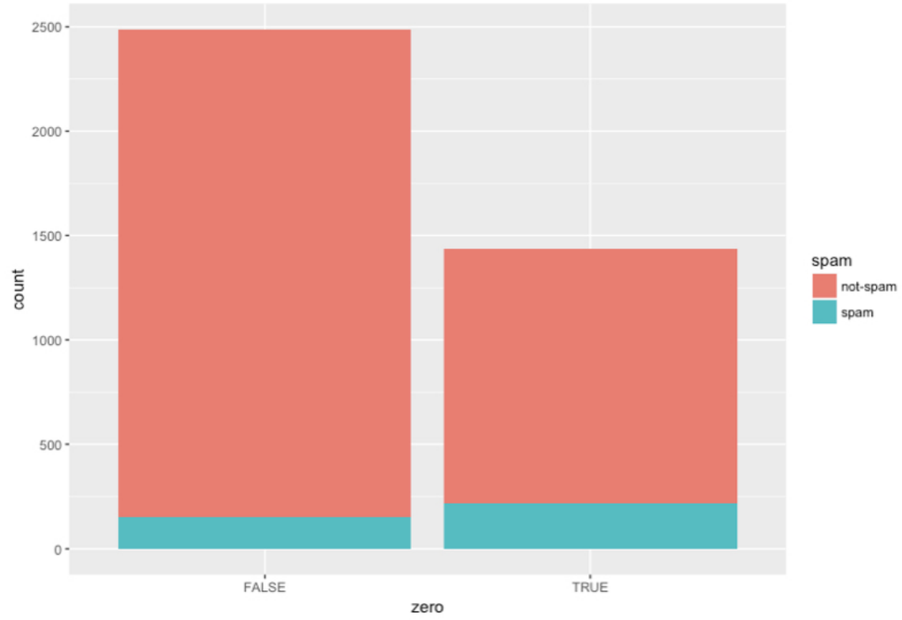
Zero inflation strategies

```
email %>%
  mutate(zero = exclaim_mess == 0) %>%
  ggplot(aes(x = zero)) +
  geom_bar() +
  facet_wrap(~spam)
```



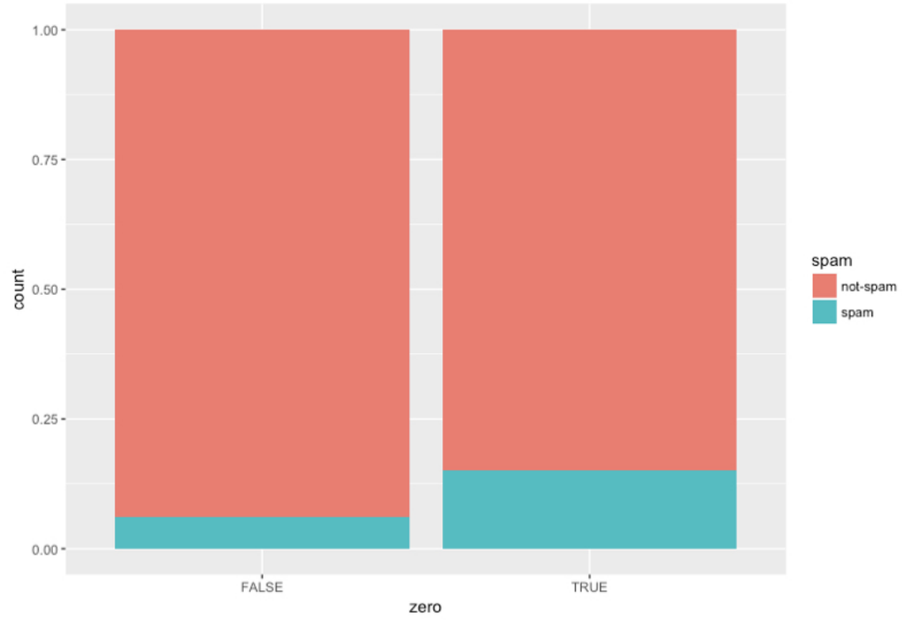
Barchart options

```
email %>%
  mutate(zero = exclaim_mess == 0) %>%
  ggplot(aes(x = zero, fill = spam)) +
  geom_bar()
```



Barchart options

```
email %>%
  mutate(zero = exclaim_mess == 0) %>%
  ggplot(aes(x = zero, fill = spam)) +
  geom_bar(position = "fill")
```



Let's practice!

EXPLORATORY DATA ANALYSIS IN R

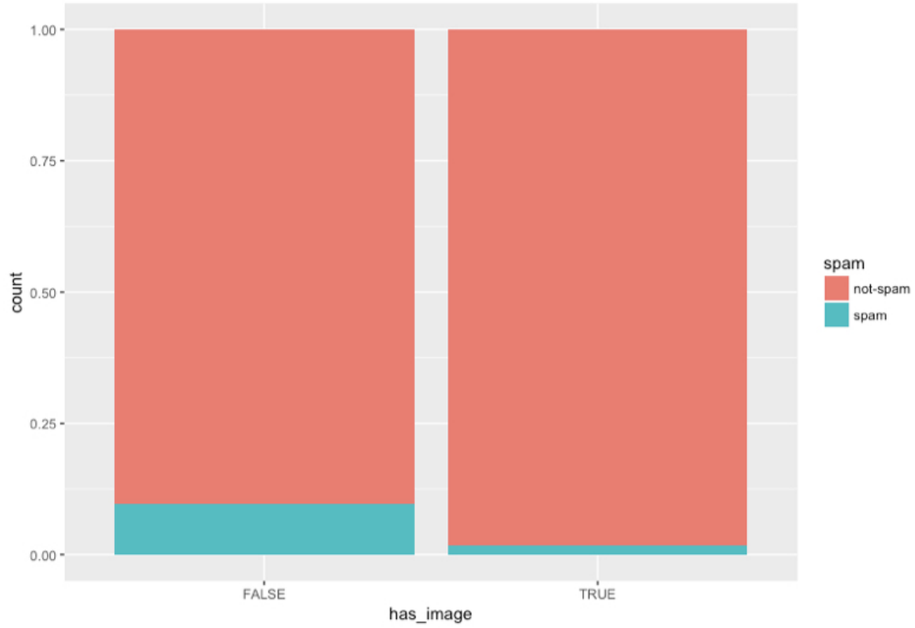
Check-in 2

EXPLORATORY DATA ANALYSIS IN R

Andrew Bray
Assistant Professor, Reed College

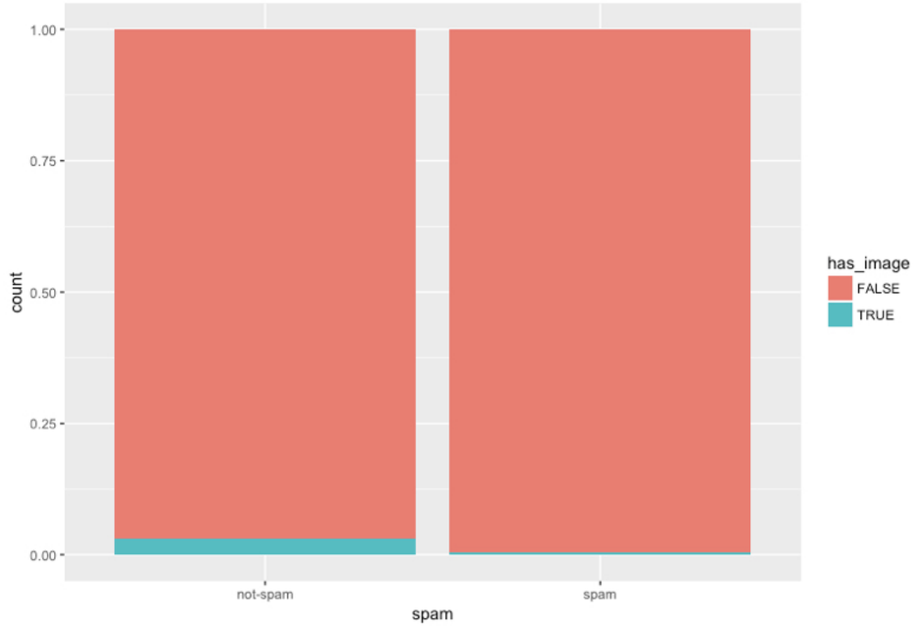
Spam and images

```
email %>%
  mutate(has_image = image 0) %>%
  ggplot(aes(x = as.factor(has_image), fill = spam)) +
  geom_bar(position = "fill")
```

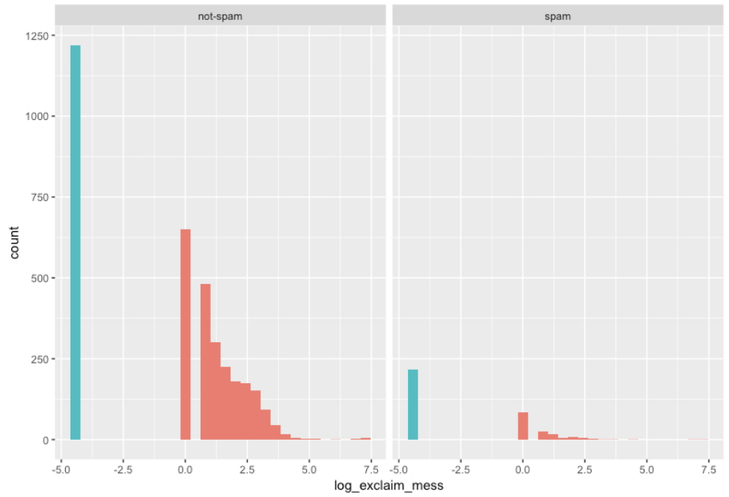


Spam and images

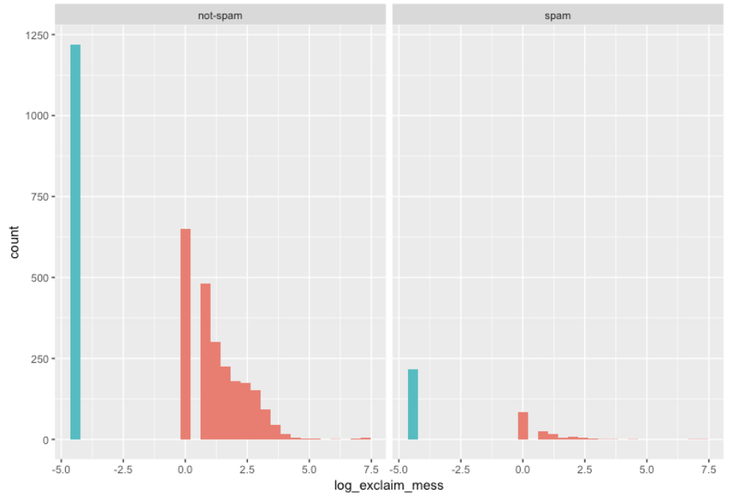
```
email %>%
  mutate(has_image = image 0) %>%
  ggplot(aes(x = spam, fill = has_image)) +
  geom_bar(position = "fill")
```

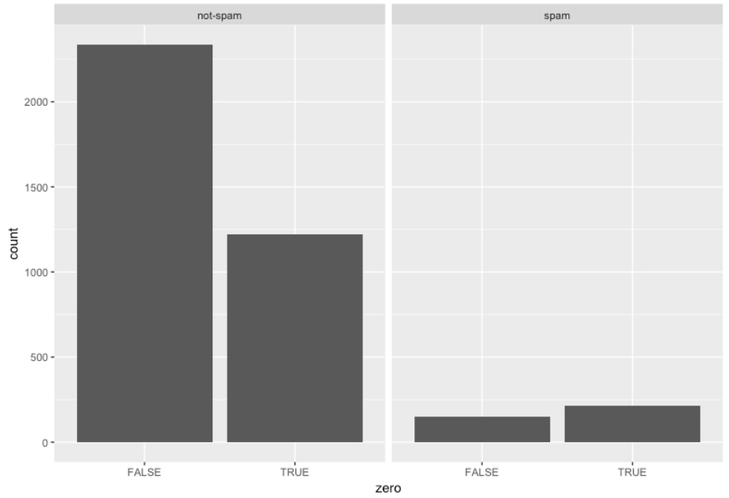


Ordering bars

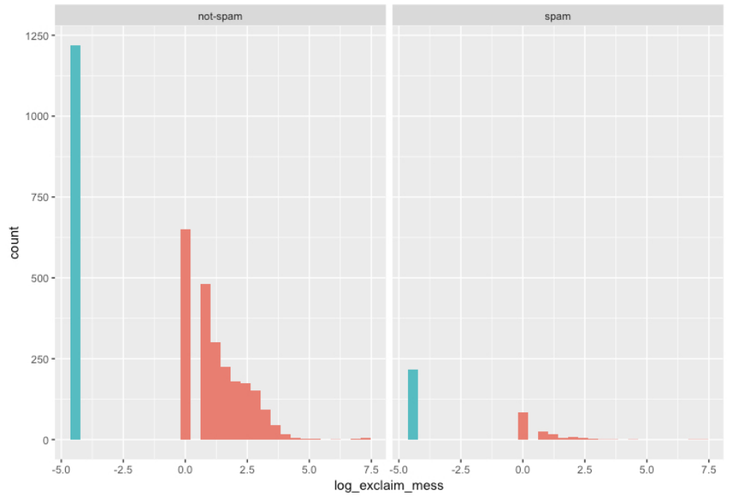


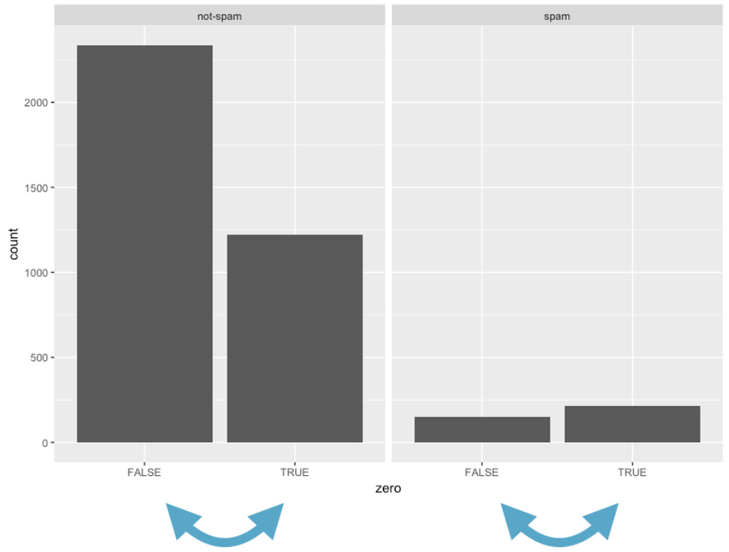
Ordering bars





Ordering bars





Ordering bars

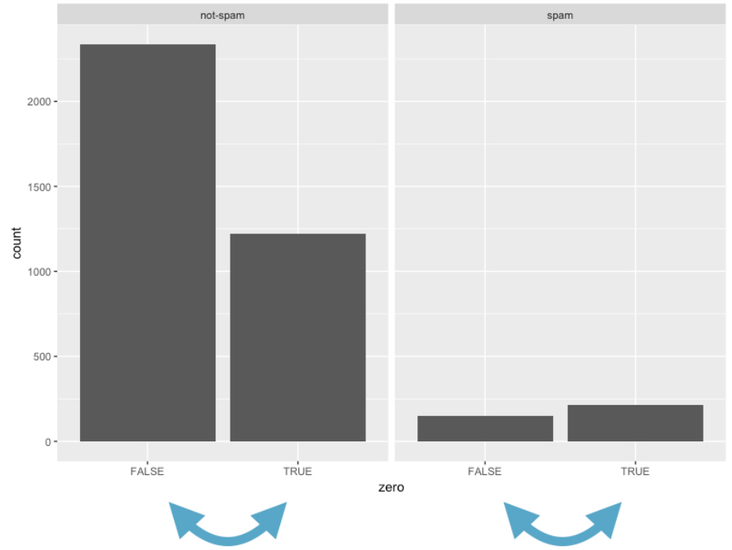
```
email <- email %>%
  mutate(zero = exclaim_mess == 0)
levels(email$zero)
```

```
NULL
```

```
email$zero <- factor(email$zero,
  levels = c("TRUE", "FALSE"))
```

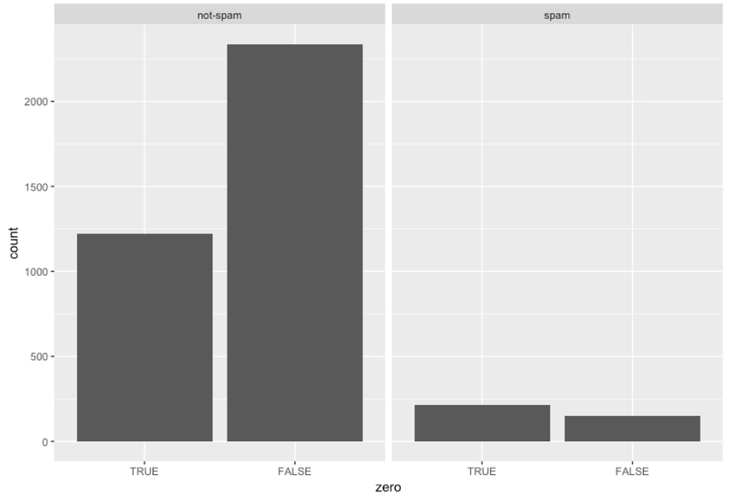
Ordering bars

```
email %>%
  ggplot(aes(x = zero)) +
  geom_bar() +
  facet_wrap(~spam)
```



Ordering bars..

```
email %>%
  ggplot(aes(x = zero)) +
  geom_bar() +
  facet_wrap(~spam)
```



Let's practice!

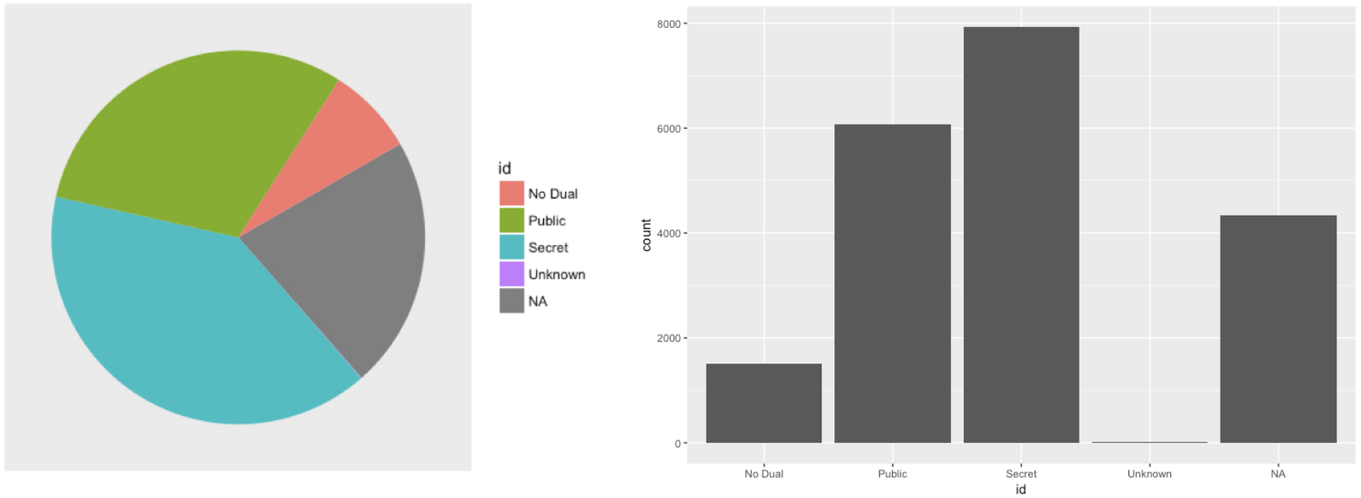
EXPLORATORY DATA ANALYSIS IN R

Conclusion

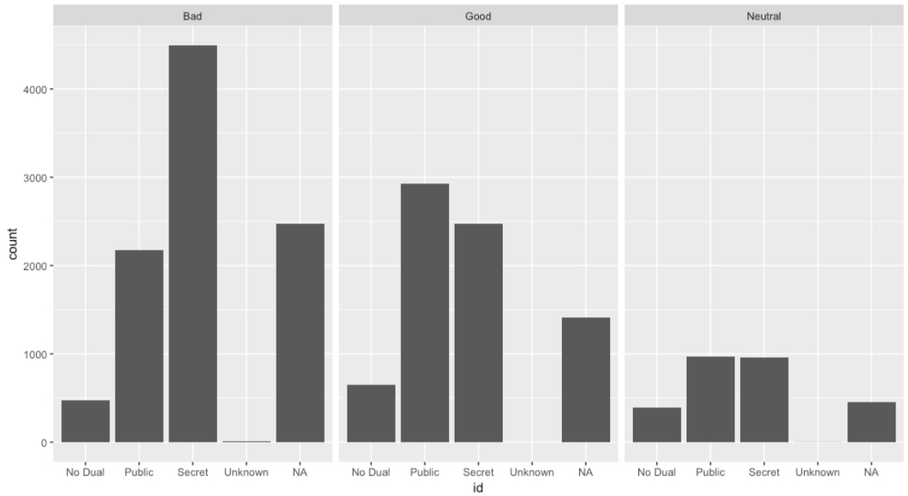
EXPLORATORY DATA ANALYSIS IN R

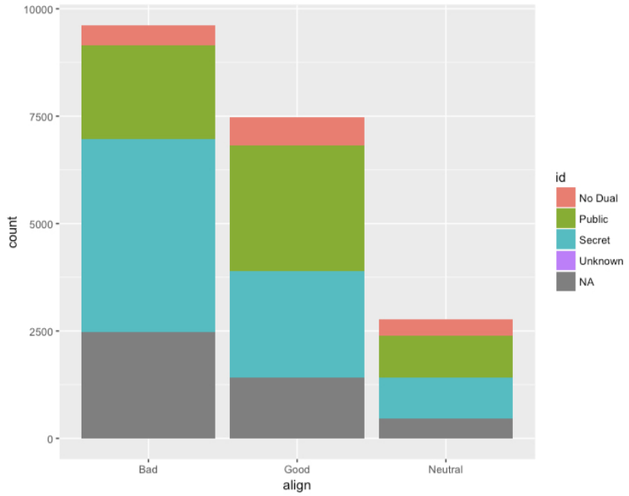
Andrew Bray
Assistant Professor, Reed College

Pie chart vs. bar chart



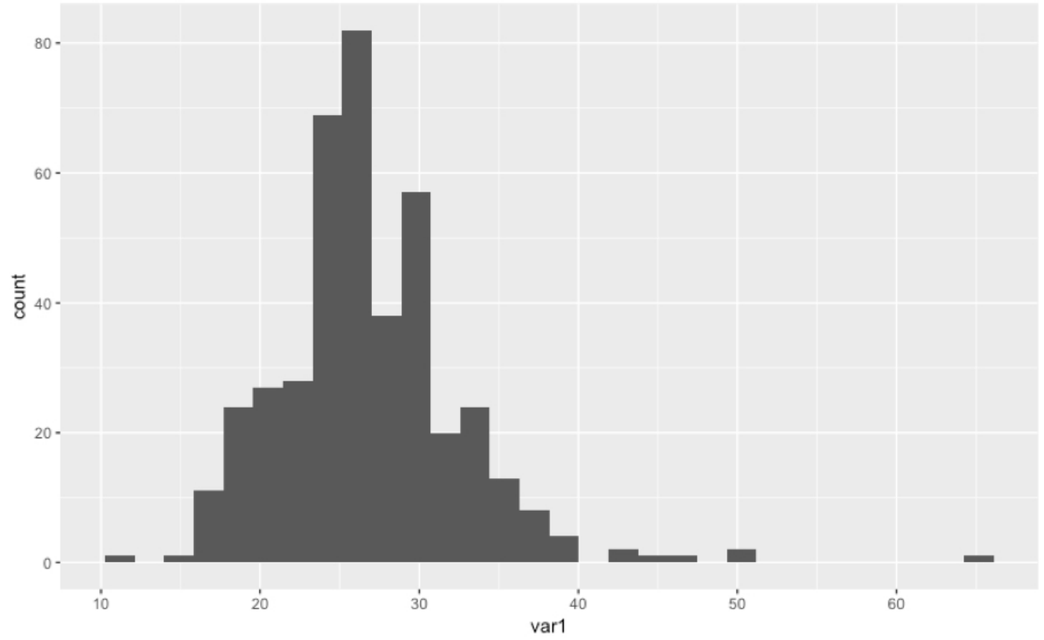
Faceting vs. stacking





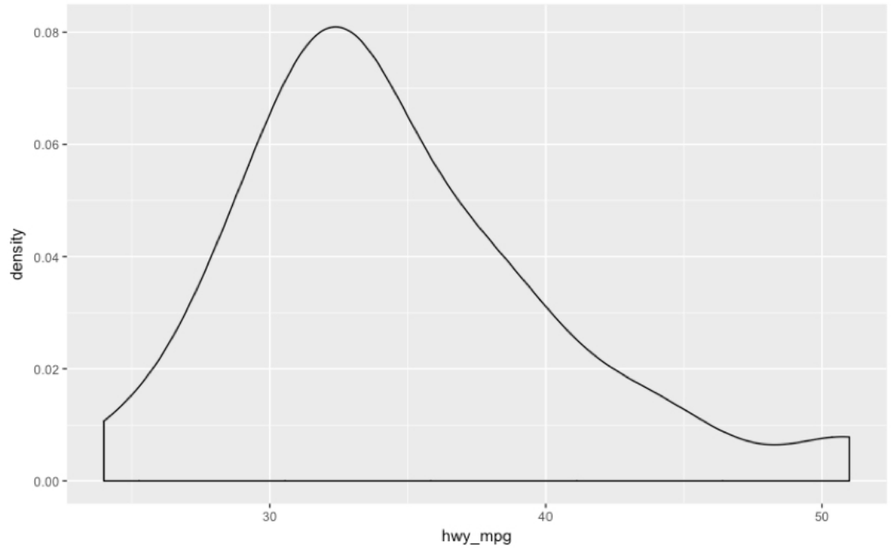
Histogram

```
ggplot(data, aes(x = var1)) +
    geom_histogram()
```



Density plot

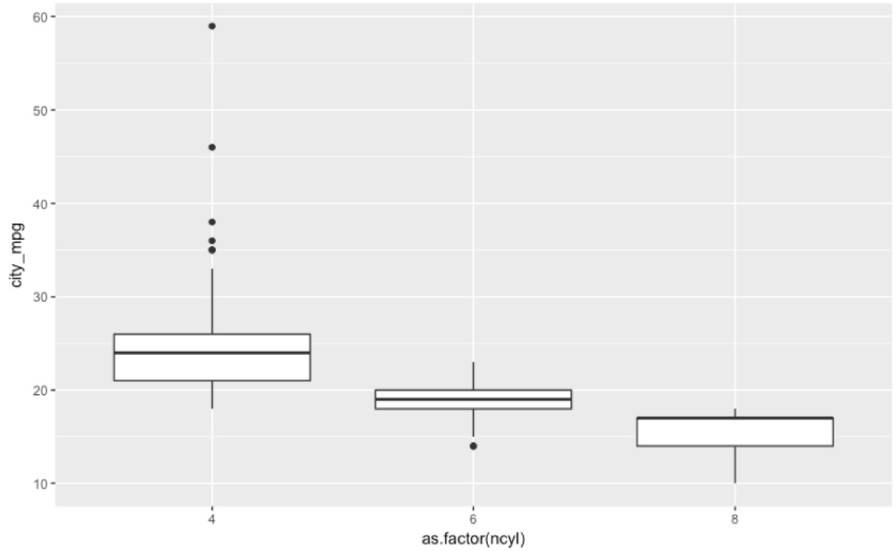
```
cars %>%
  filter(eng_size < 2.0) %>%
  ggplot(aes(x = hwy_mpg)) +
  geom_density()
```



Side-by-side box plots

```
ggplot(common_cyl, aes(x = as.factor(ncyl), y = city_mpg)) +
  geom_boxplot()
```

```
Warning message:
Removed 11 rows containing non-finite values (stat_boxplot).
```



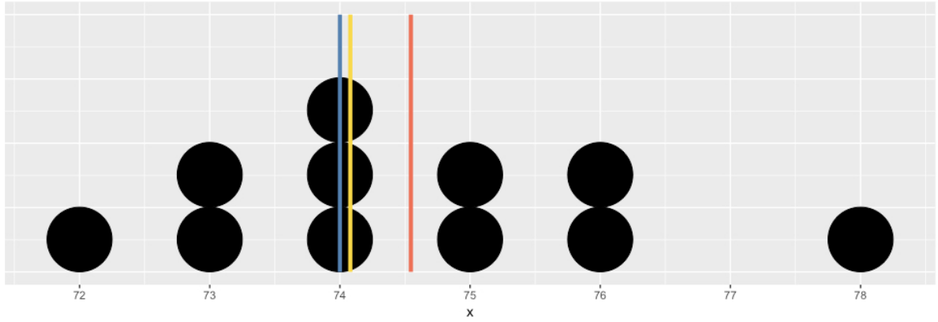
Center: mean, median, mode

```
x
```

```
76 78 75 74 76 72 74 73 73 75 74
```

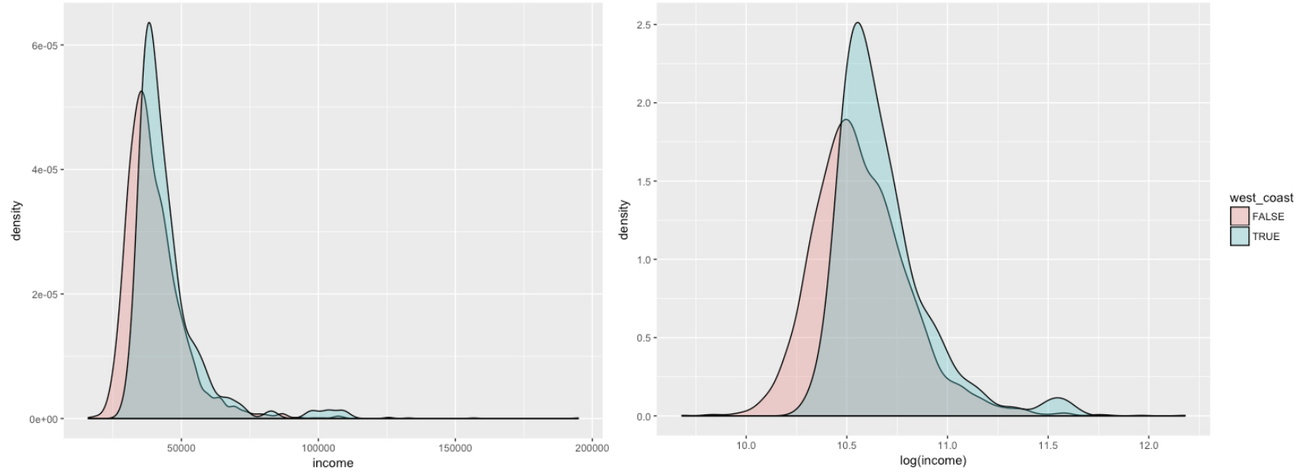
```
table(x)
x
```

```
72 73 74 75 76 78
 1  2  3  2  2  1
```



Shape of income

```
ggplot(life, aes(x = income, fill = west_coast)) +
  geom_density(alpha = .3)
ggplot(life, aes(x = log(income), fill = west_coast)) +
  geom_density(alpha = .3)
```



With group_by()

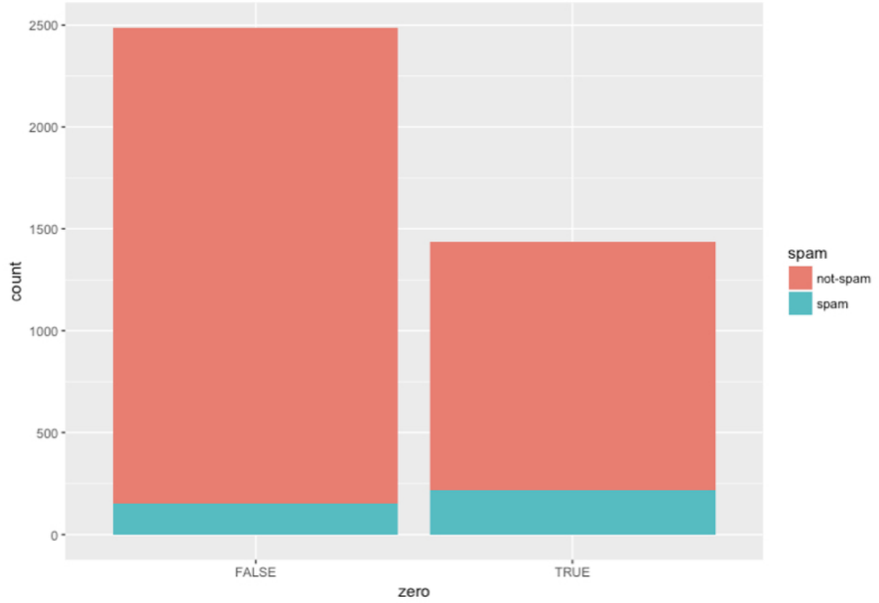
```
life %>%
  slice(240:247) %>%
  group_by(west_coast) %>%
  summarize(mean(expectancy))
```

```
# A tibble: 2 x 2
  west_coast mean(expectancy)
       <lgl>            <dbl>
1      FALSE          79.26125
2       TRUE          79.29375
```

state	county	expectancy	income	west_coast
California	Tuolumne	79.6	41770	TRUE
California	Ventura	81.1	54155	TRUE
California	Yolo	80.0	49063	TRUE
California	Yuba	76.3	37535	TRUE
Colorado	Adams	80.1	36962	FALSE
Colorado	Alamosa	77.4	34088	FALSE
Colorado	Arapahoe	80.3	52545	FALSE
Colorado	Archuleta	79.1	40307	FALSE

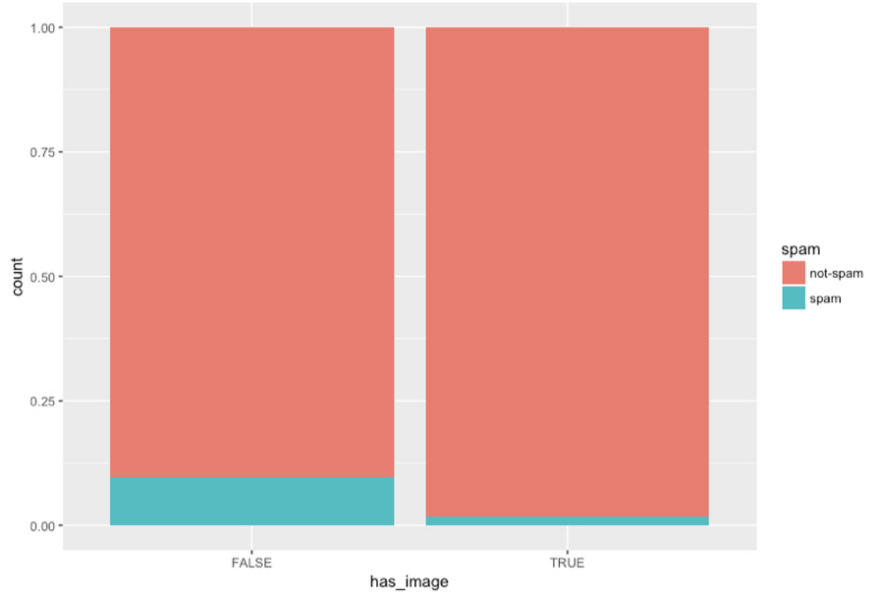
Spam and exclamation points

```
email %>%
  mutate(zero = exclaim_mess == 0) %>%
  ggplot(aes(x = zero, fill = spam)) +
  geom_bar()
```



Spam and images

```
email %>%
  mutate(has_image = image 0) %>%
  ggplot(aes(x = as.factor(has_image), fill = spam)) +
  geom_bar(position = "fill")
```



Let's practice!

EXPLORATORY DATA ANALYSIS IN R