Exploring the Enron dataset

Tyler Byers

EDA for Udacity Intro to Machine Learning Final Project

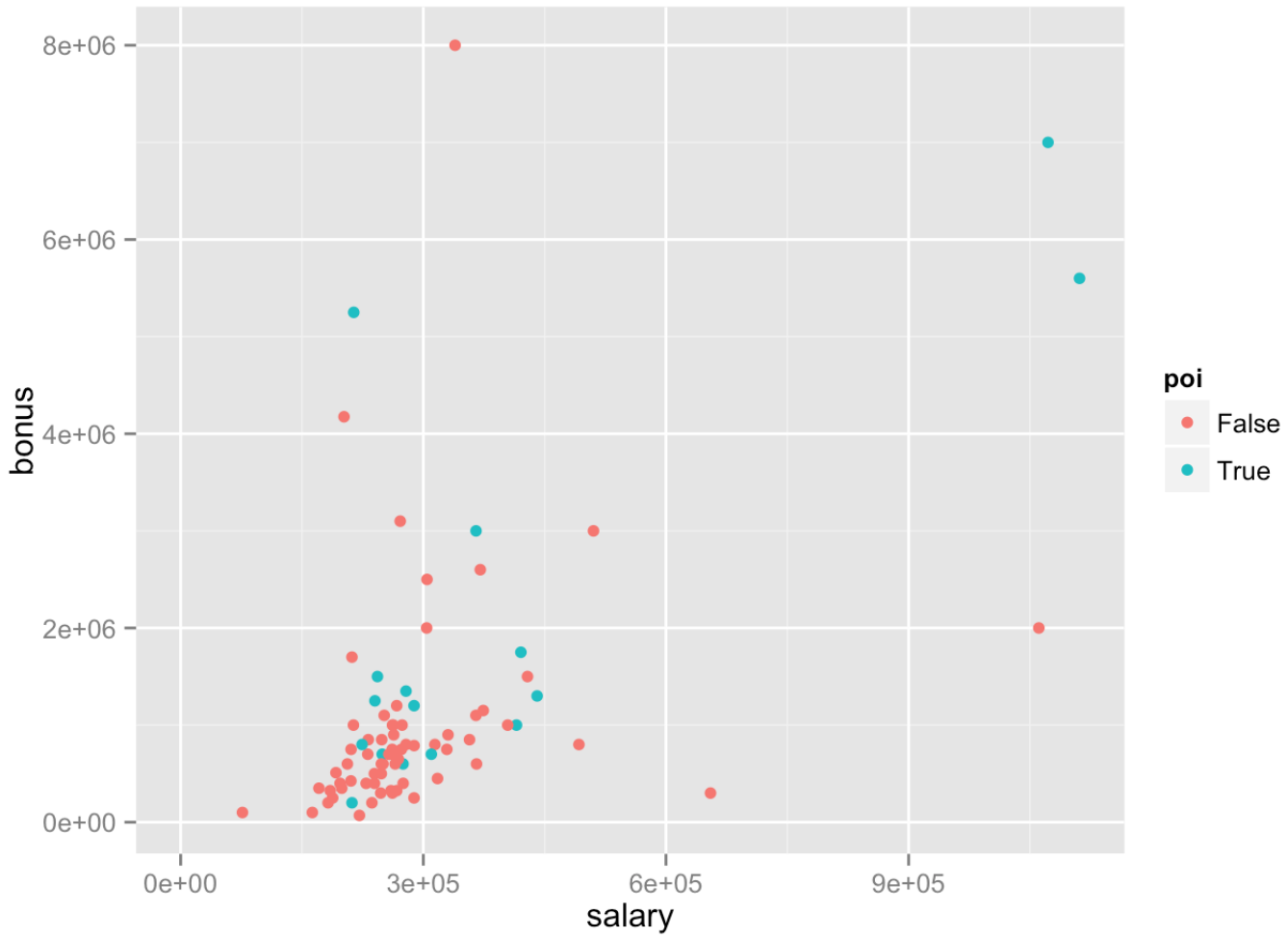
```
library(ggplot2); library(caret)
```

```
## Loading required package: lattice
```

```
enron = read.csv('dataset.csv')
```

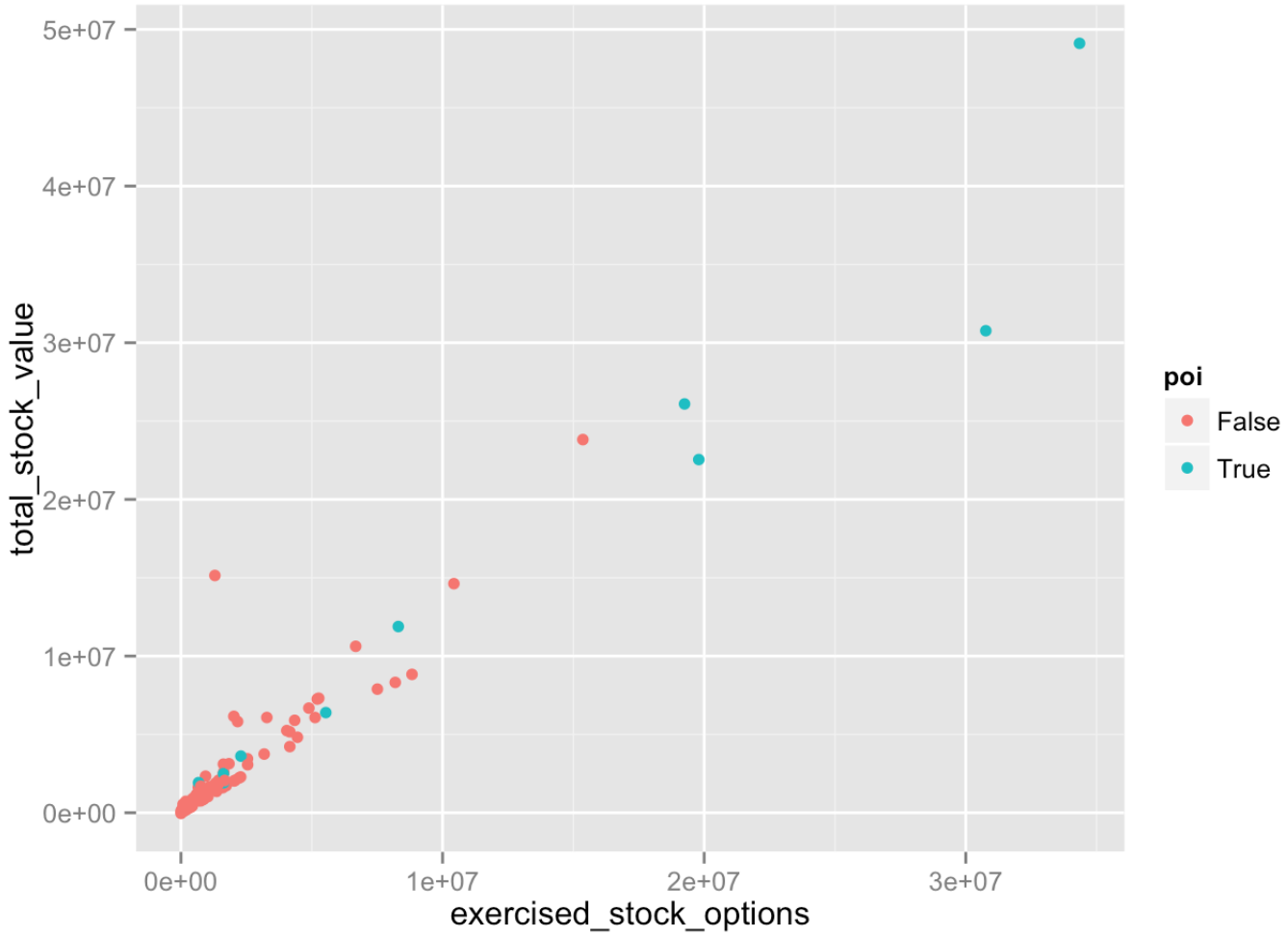
```
ggplot(enron, aes(x = salary, y = bonus, color = poi)) + geom_point()
```

```
## Warning: Removed 64 rows containing missing values (geom_point).
```



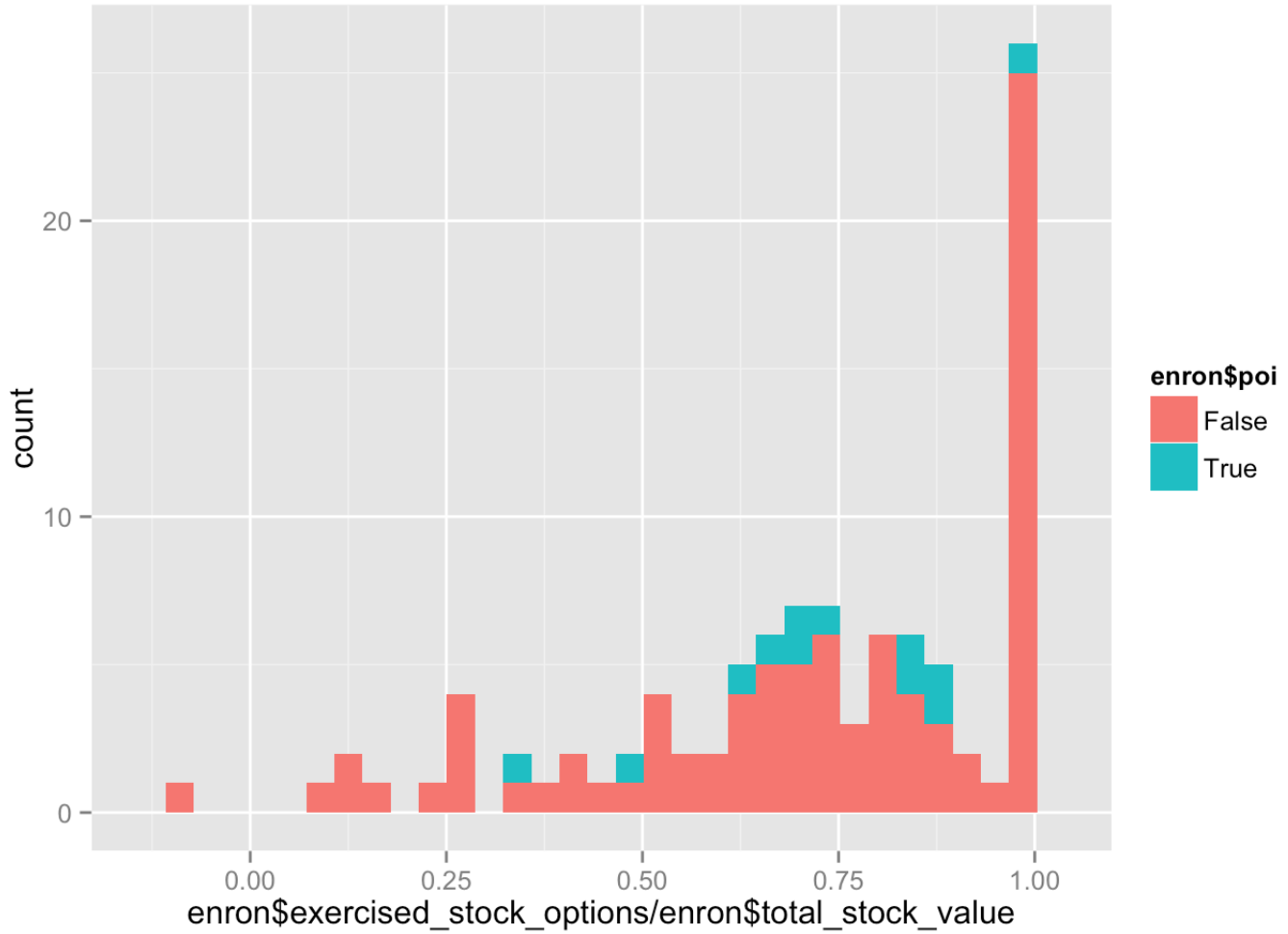
```
ggplot(enron, aes(x = exercised_stock_options, y = total_stock_value,
                  color = poi)) + geom_point()
```

```
## Warning: Removed 45 rows containing missing values (geom_point).
```



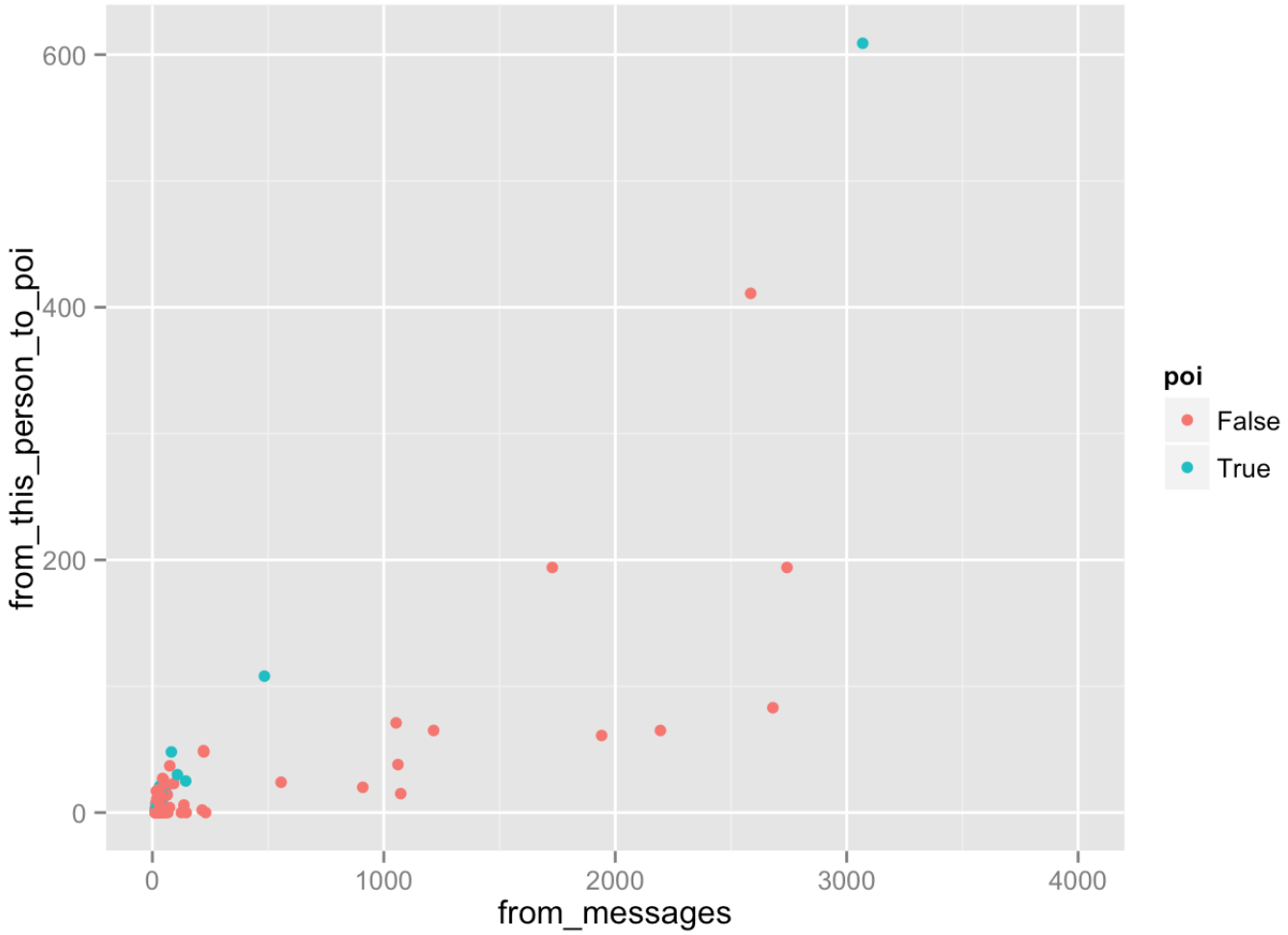

```
qplot(enron$exercised_stock_options/enron$total_stock_value, fill = enron$poi)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



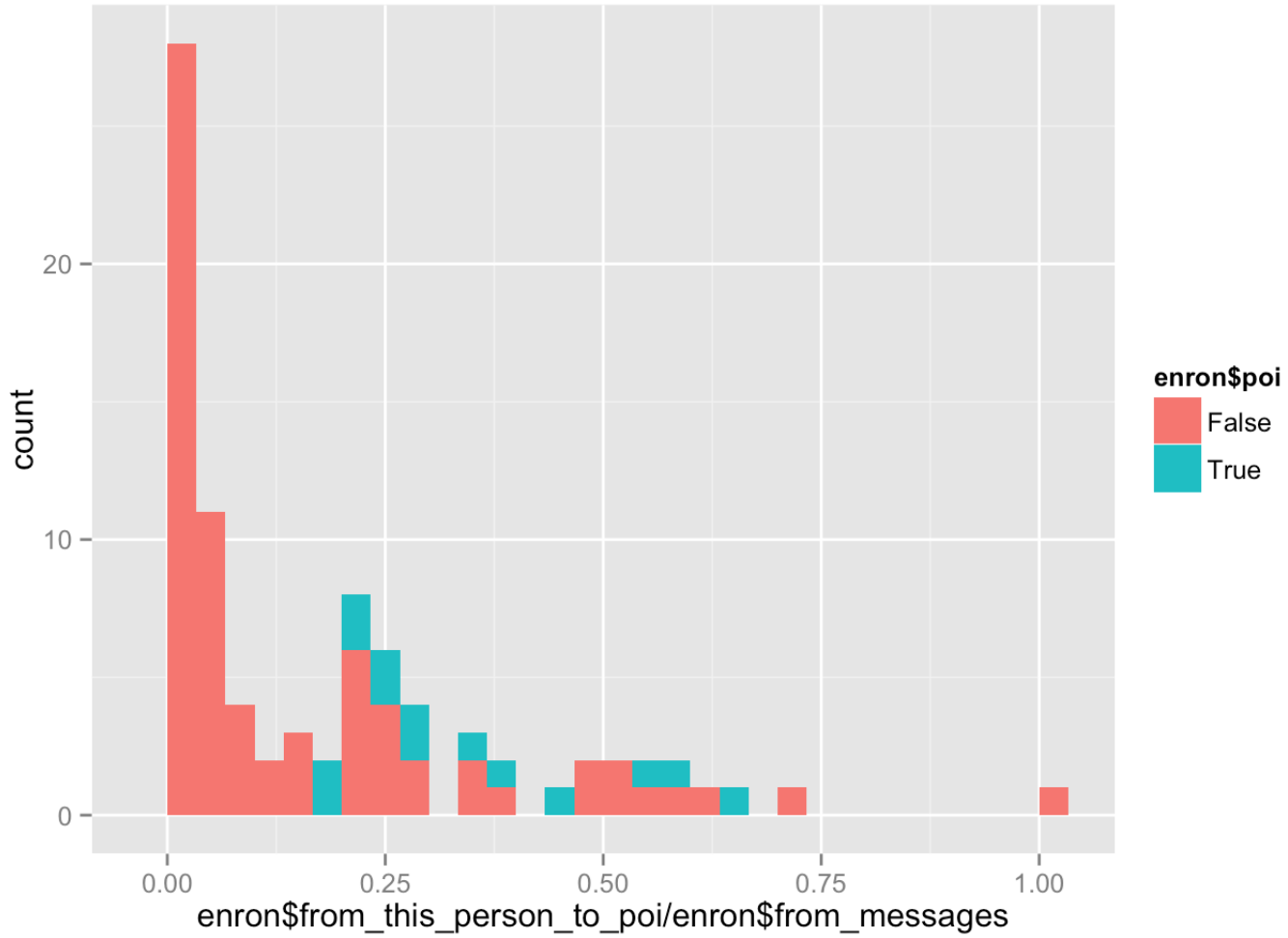

```
ggplot(enron, aes(x = from_messages, y = from_this_person_to_poi, color = poi)) +
     geom_point() + scale_x_continuous(limits = c(0,4000))
```

```
## Warning: Removed 62 rows containing missing values (geom_point).
```



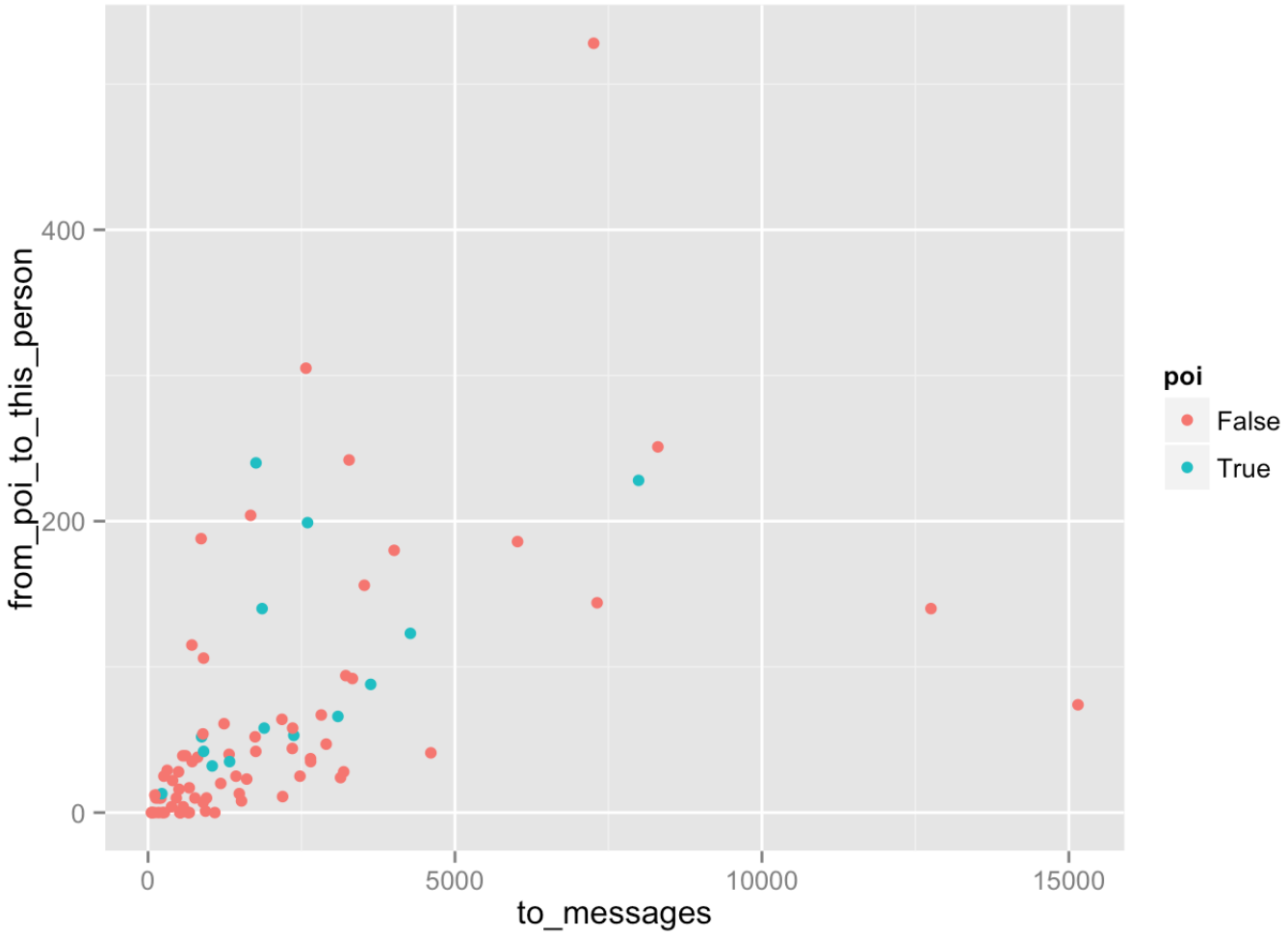

```
qplot(enron$from_this_person_to_poi/enron$from_messages, fill = enron$poi)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



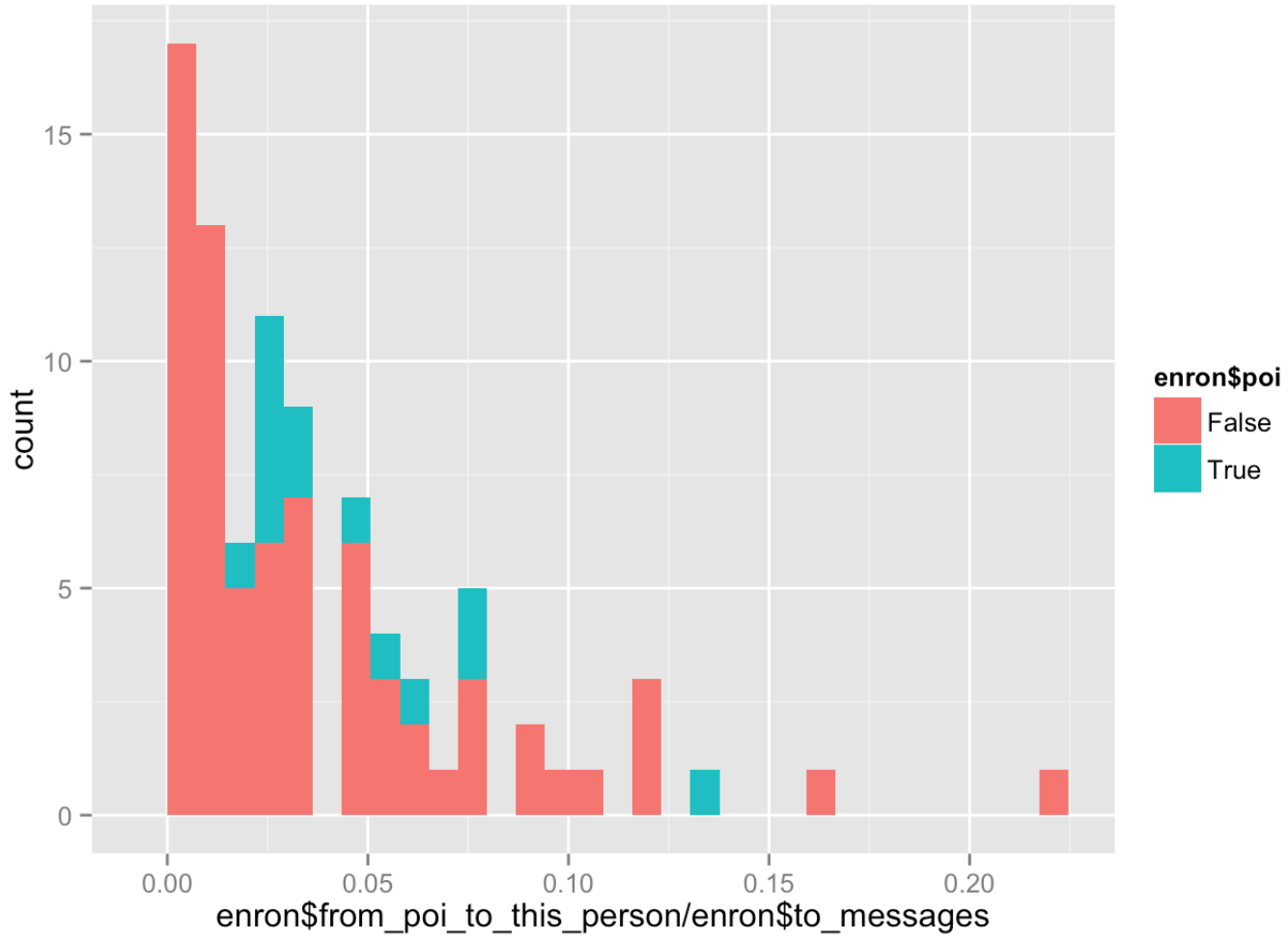

```
ggplot(enron, aes(x = to_messages, y = from_poi_to_this_person, color = poi)) +
     geom_point()
```

```
## Warning: Removed 59 rows containing missing values (geom_point).
```



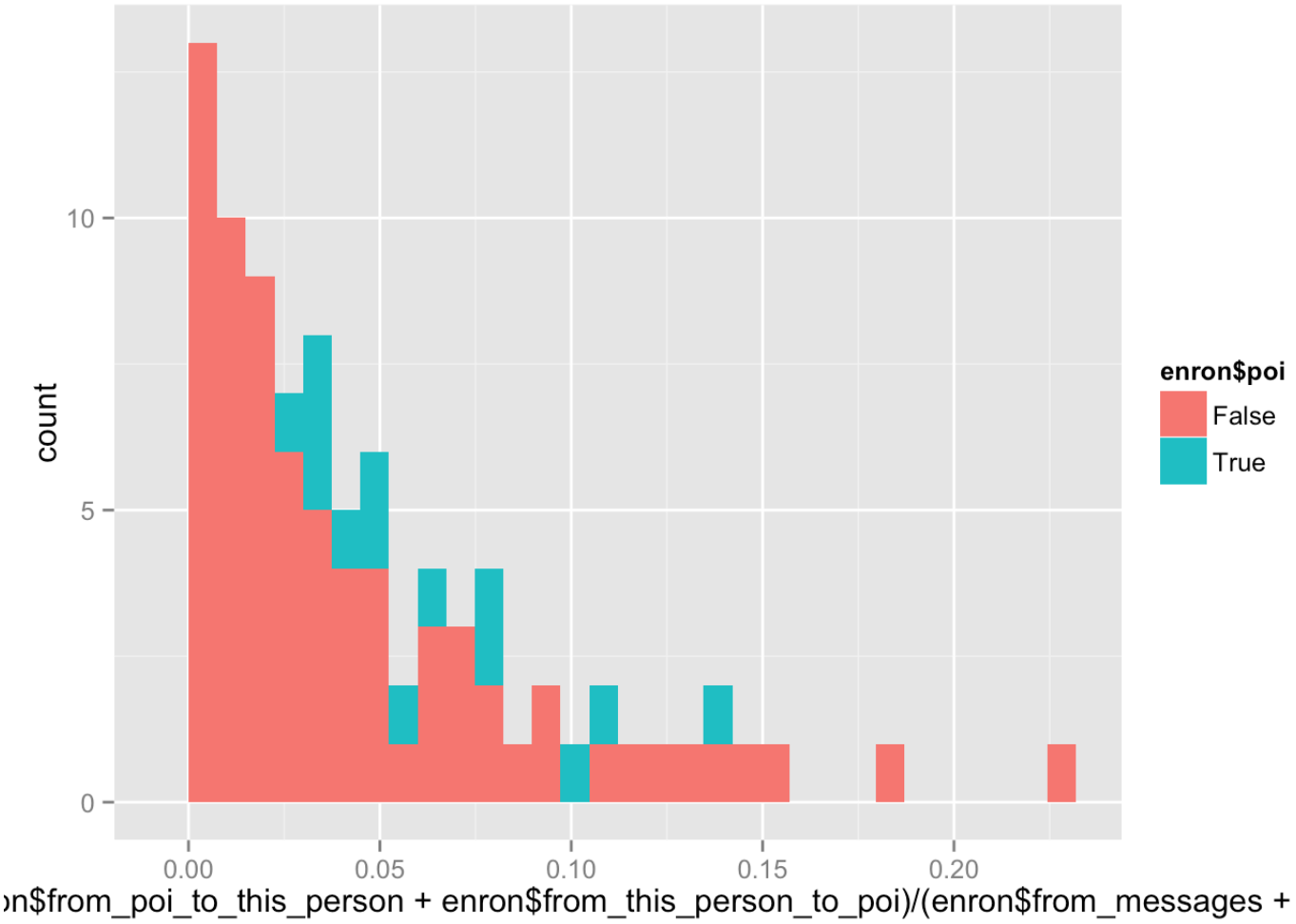

```
qplot(enron$from_poi_to_this_person/enron$to_messages, fill = enron$poi)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
qplot((enron$from_poi_to_this_person + enron$from_this_person_to_poi)/
          (enron$from_messages + enron$to_messages), fill = enron$poi)
```

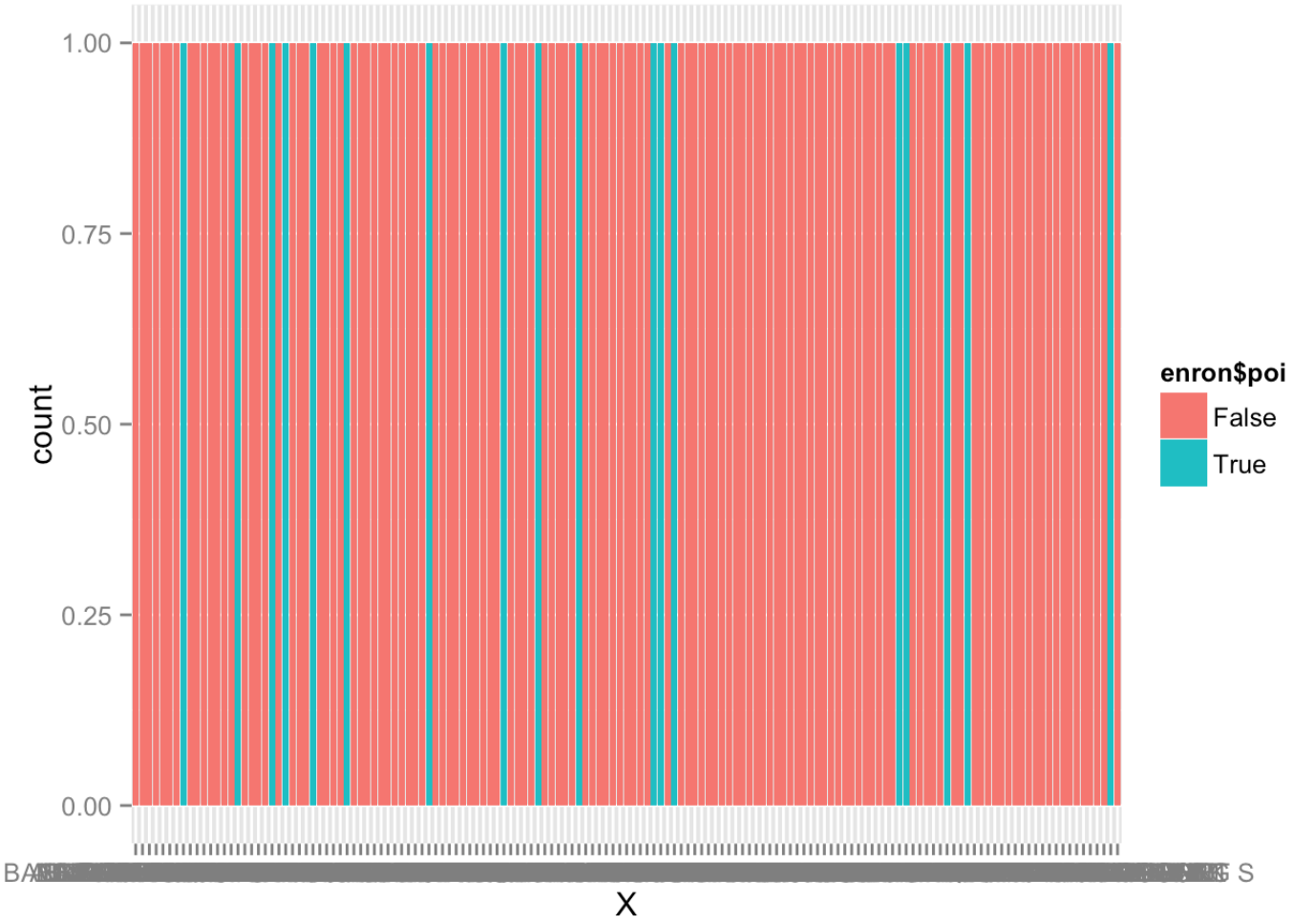
```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



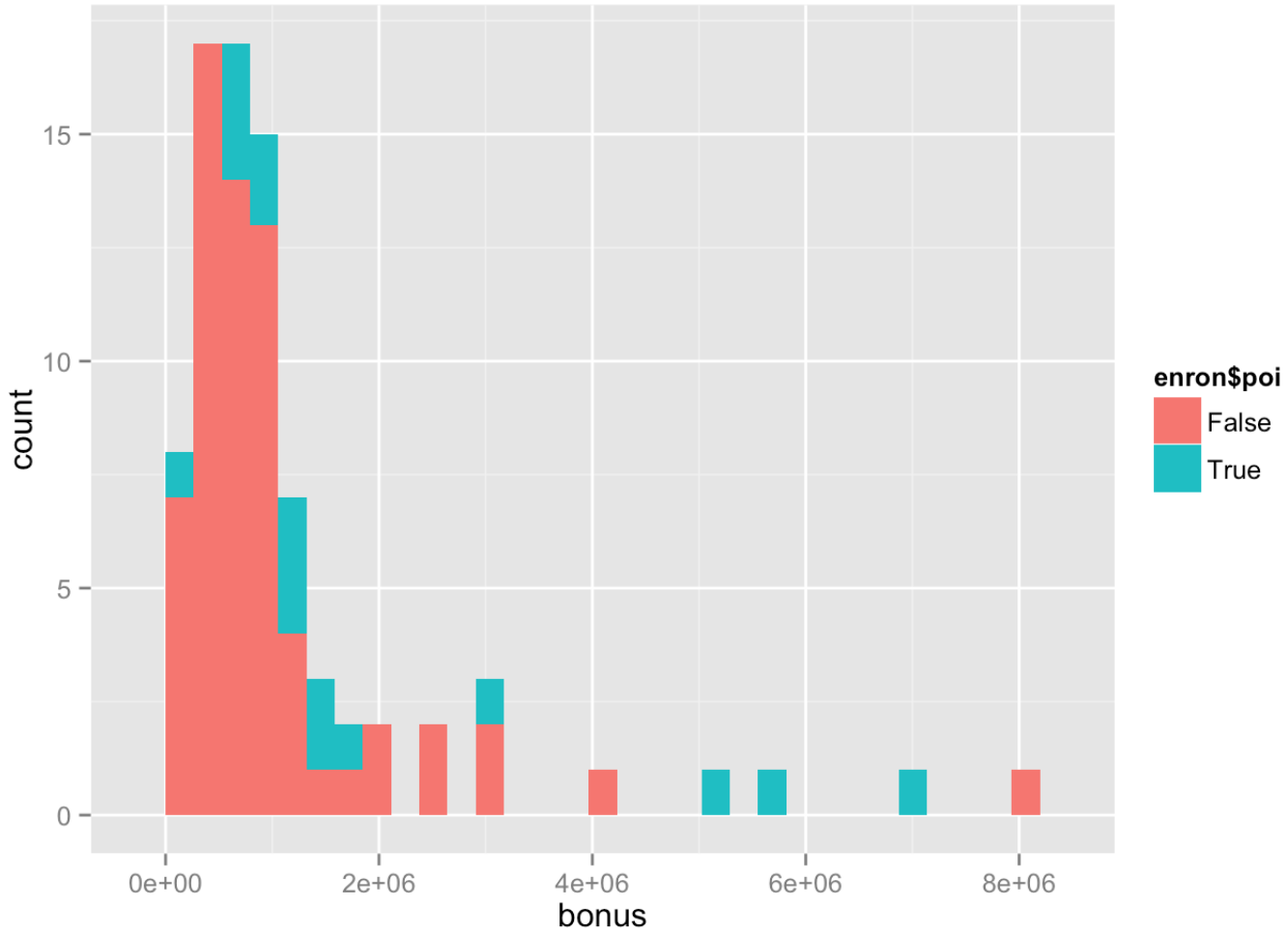

```
for (i in seq(2:length(names(enron)))) {
  print(qplot(enron[,i], fill = enron$poi) + xlab(names(enron)[i]))

}
```

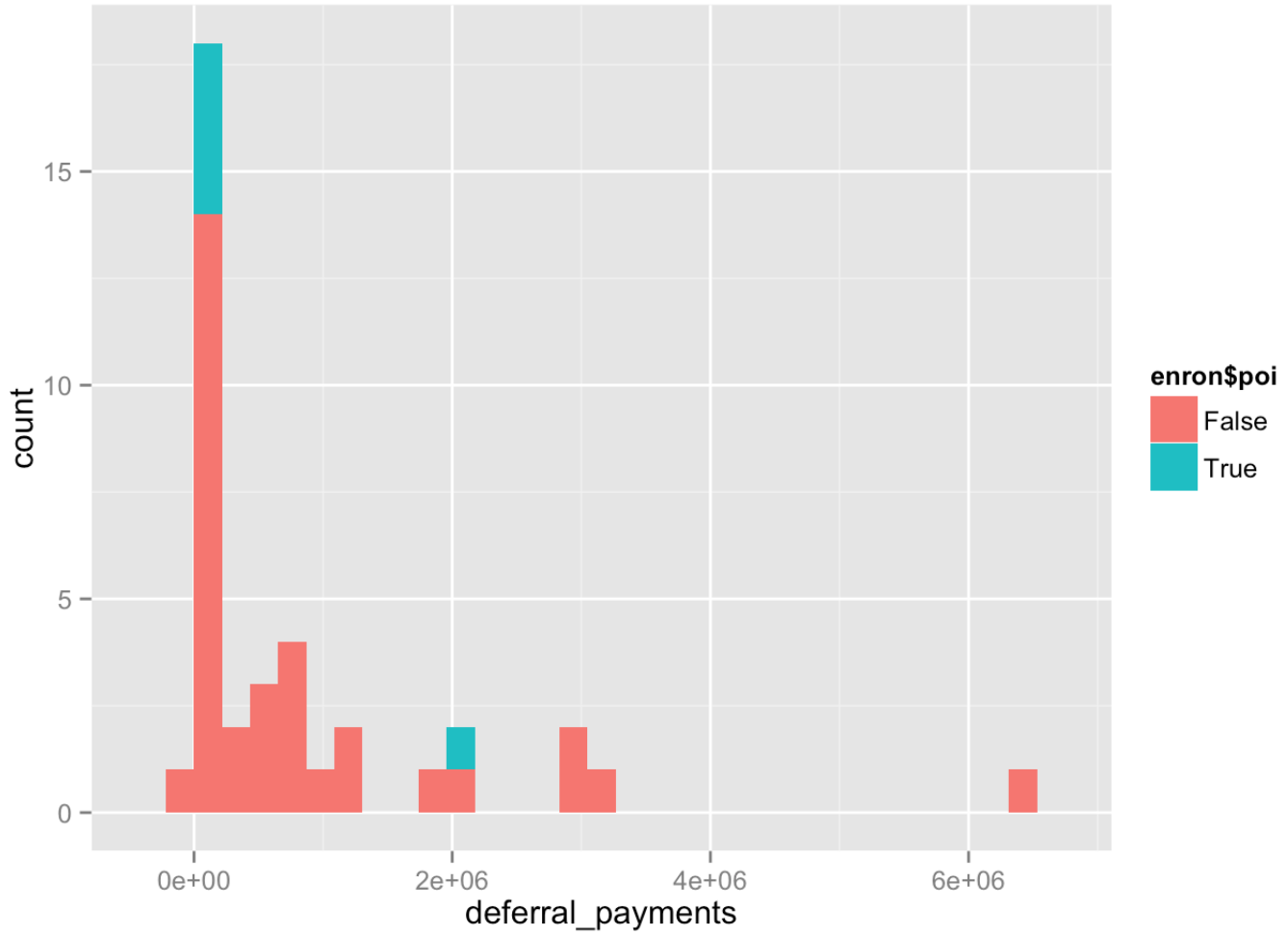
```
## Warning: position_stack requires constant width: output may be incorrect
```



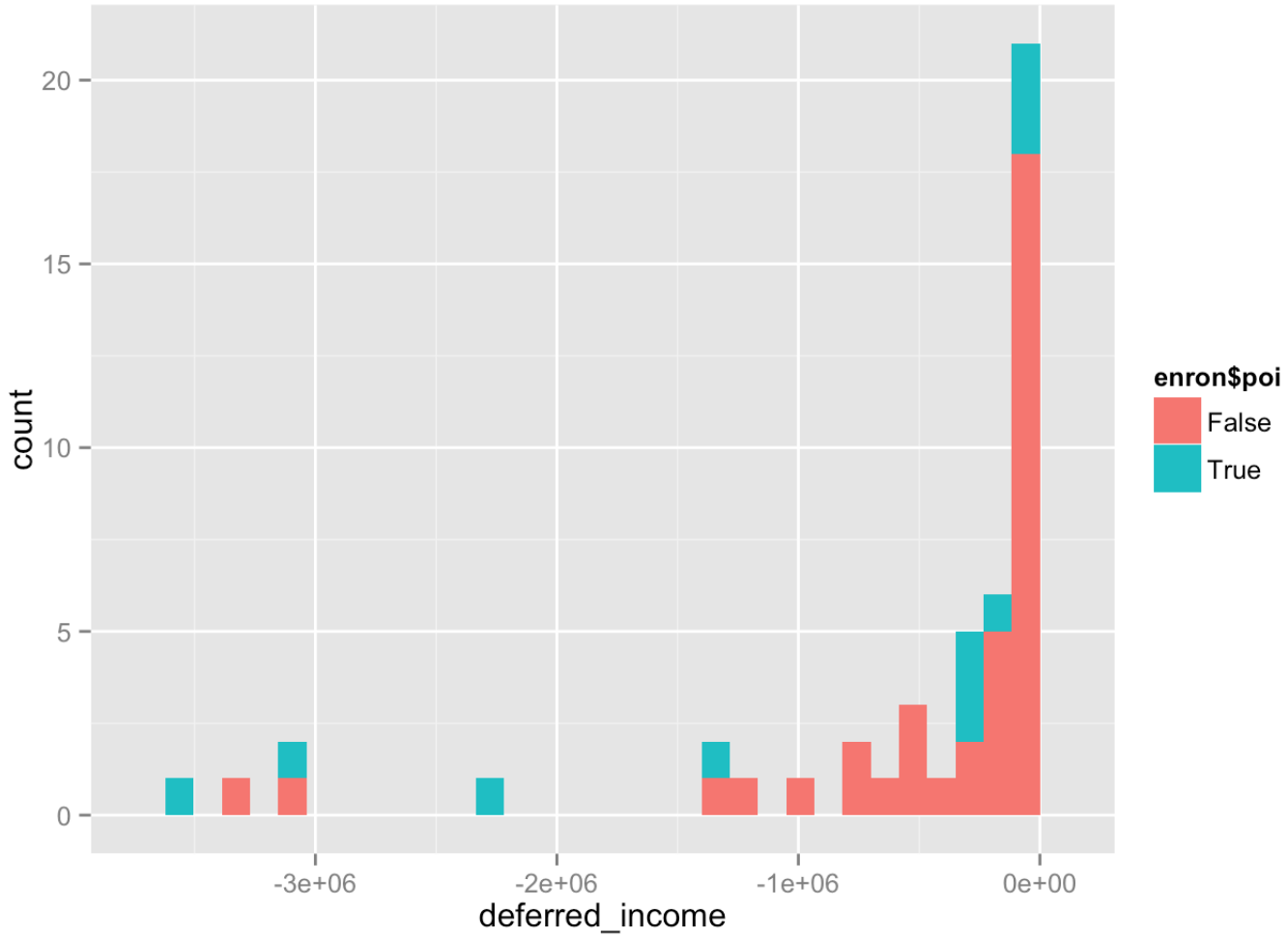

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



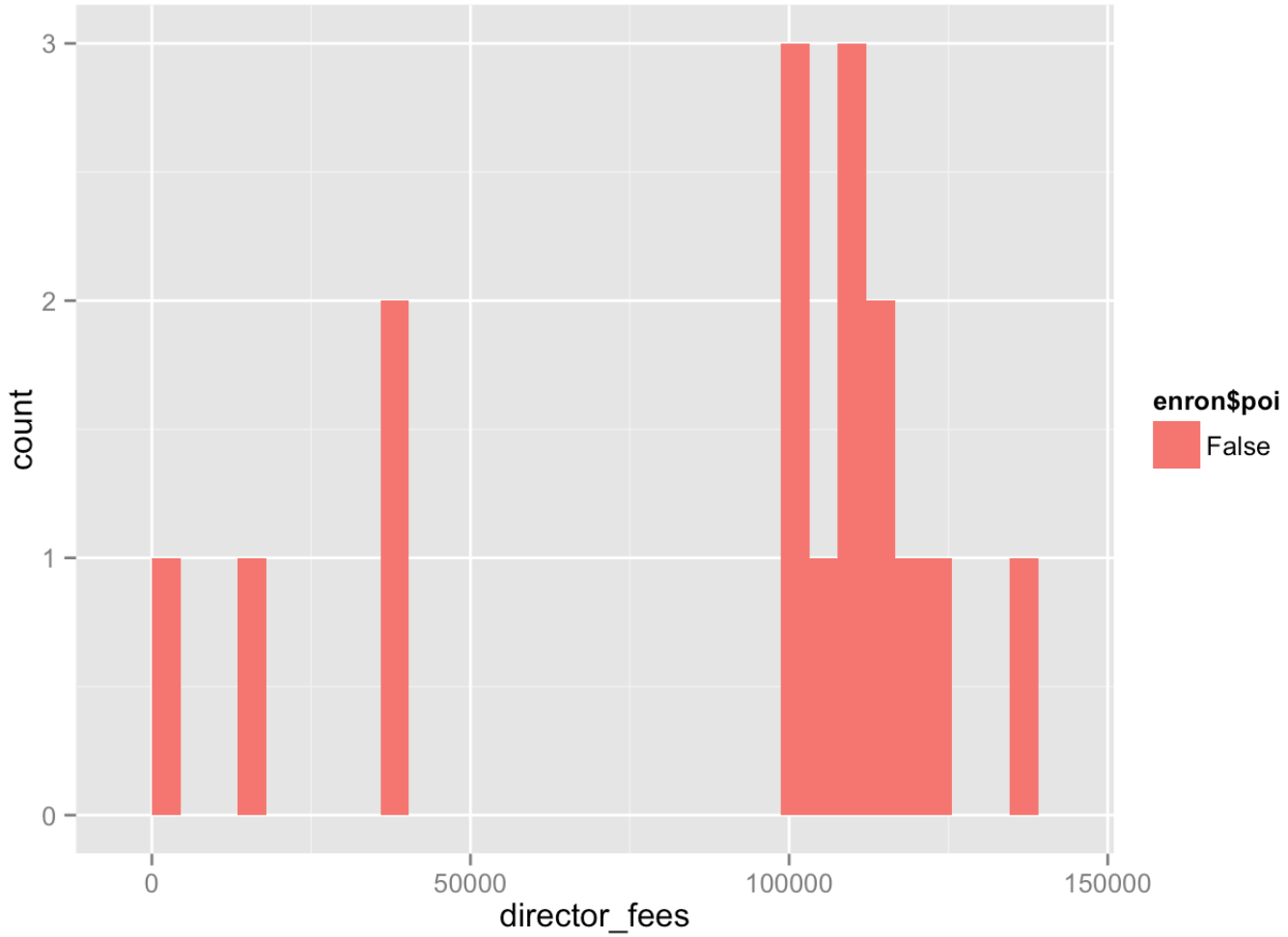

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

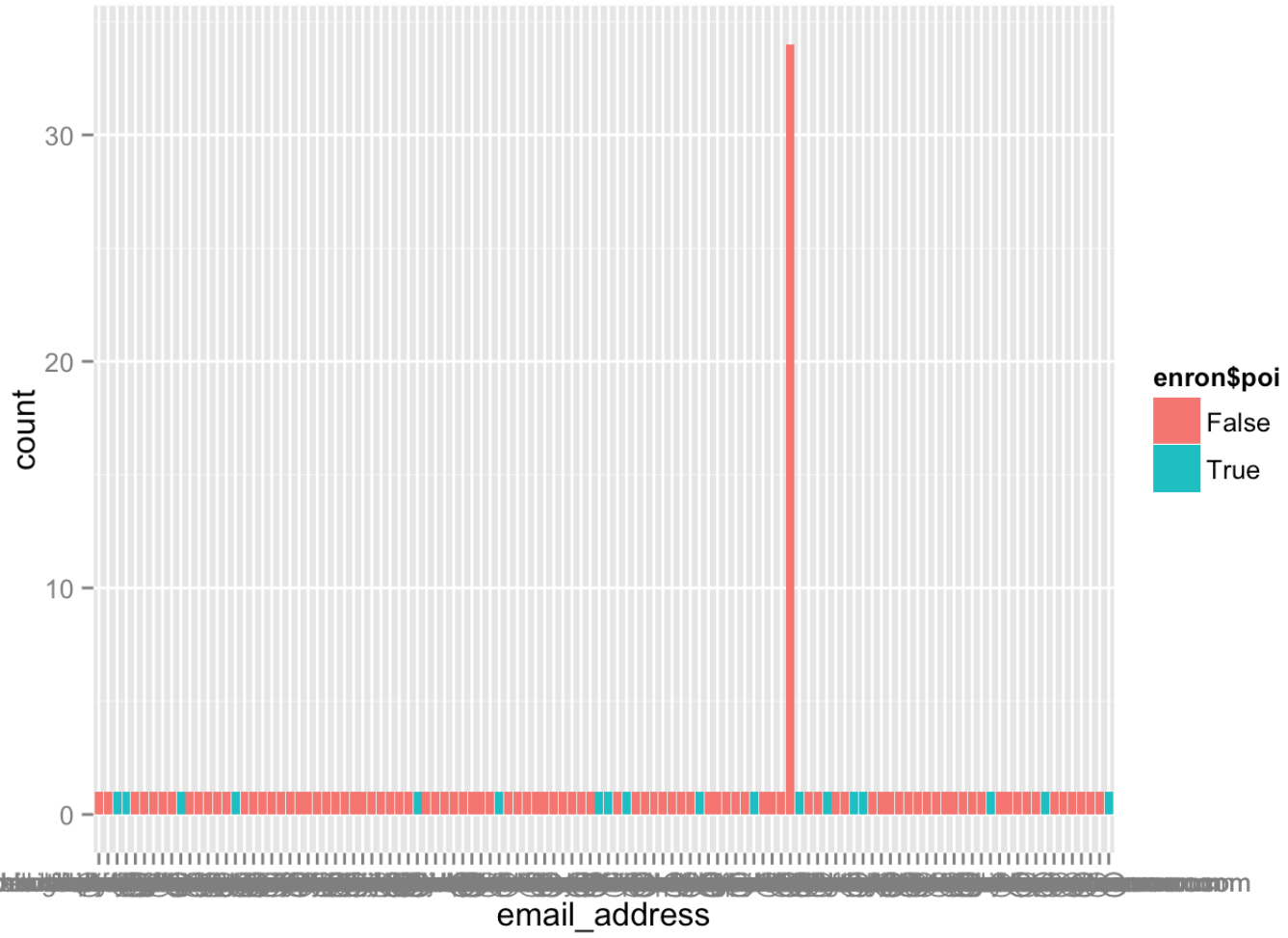



```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

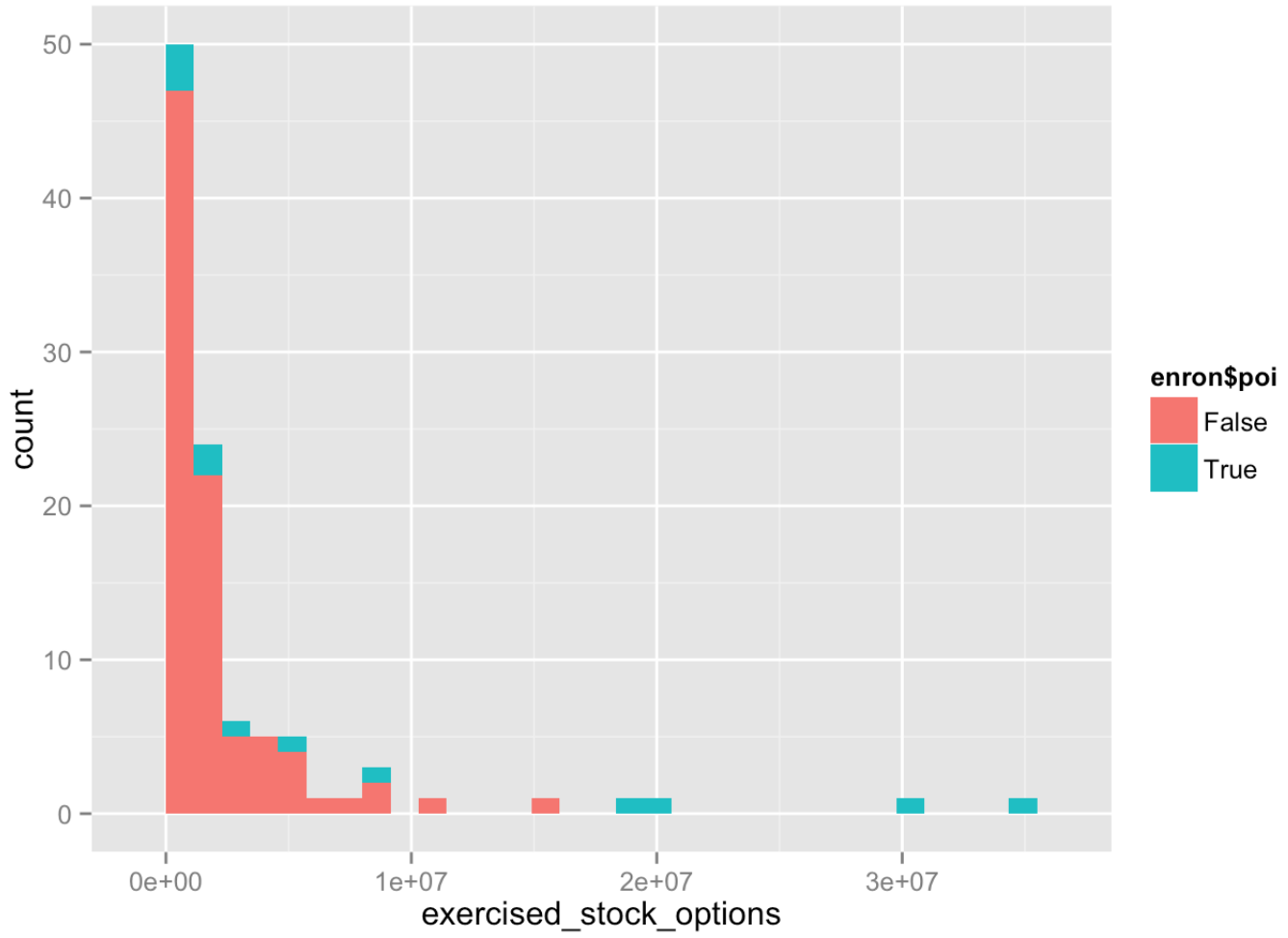



```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

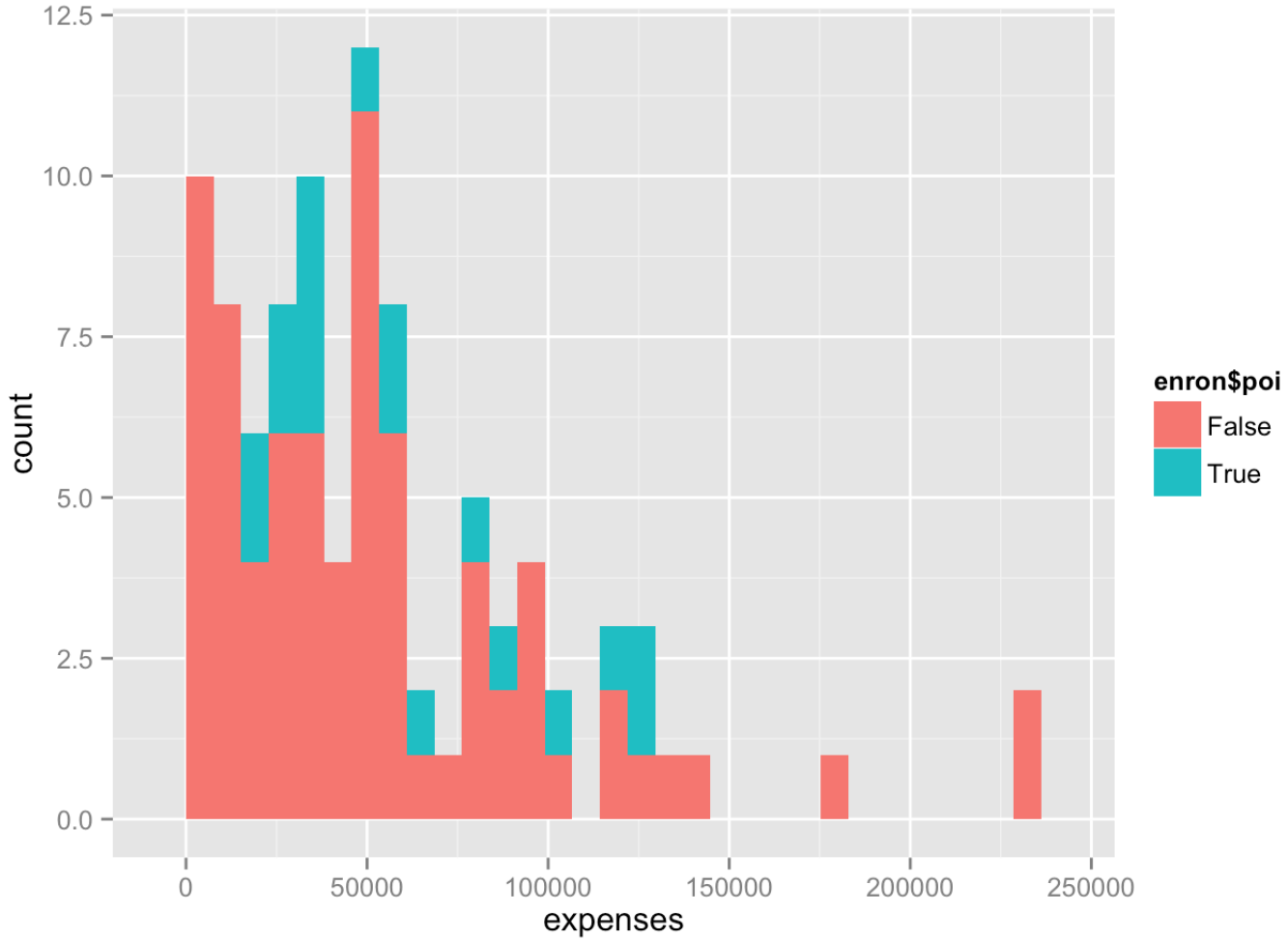

3
2
1
0
count
0
50000
100000
150000
director_fees
enron$poi
False


30
20
10
0
count
email_address
enron$poi
False
True

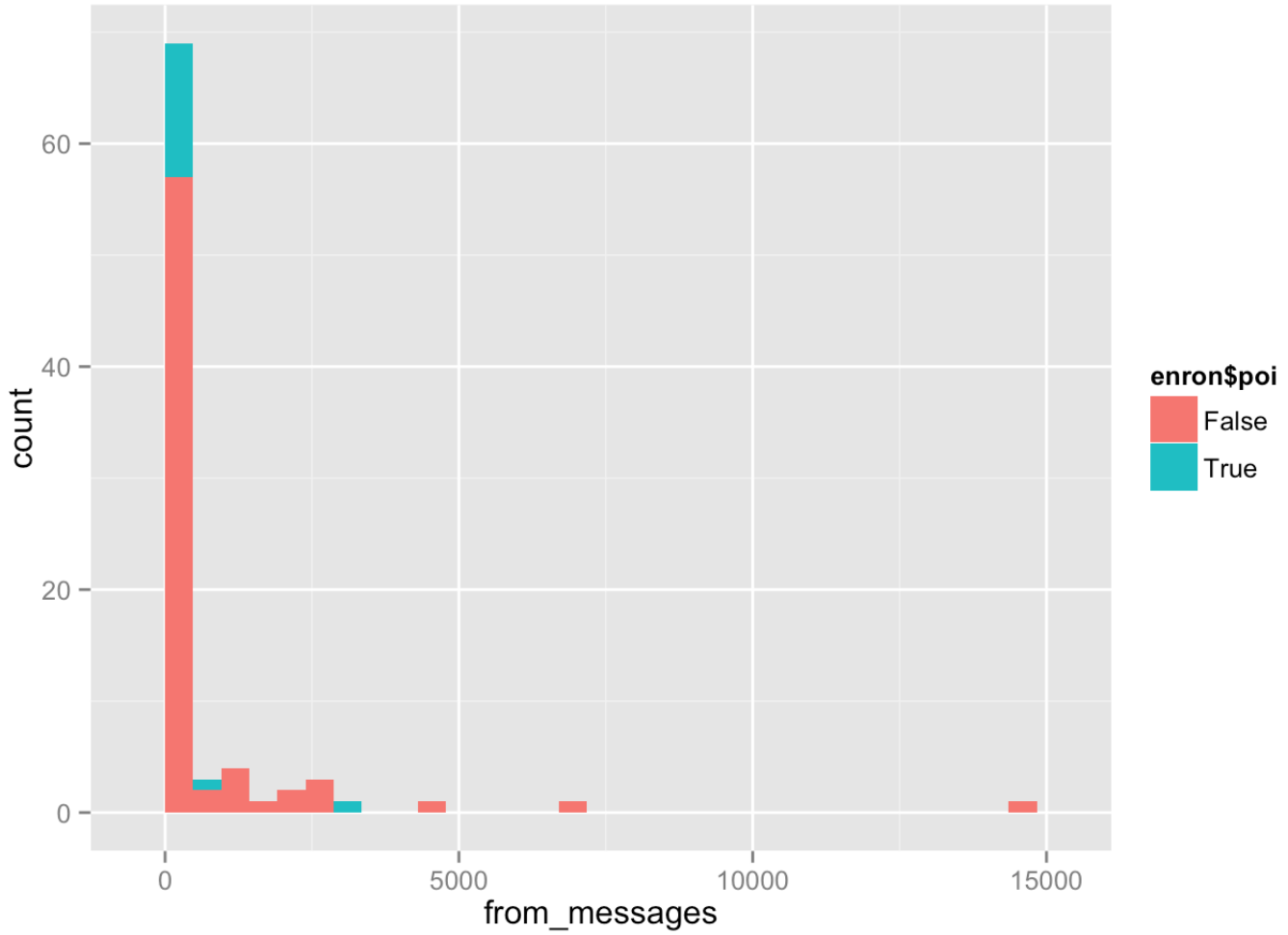
```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



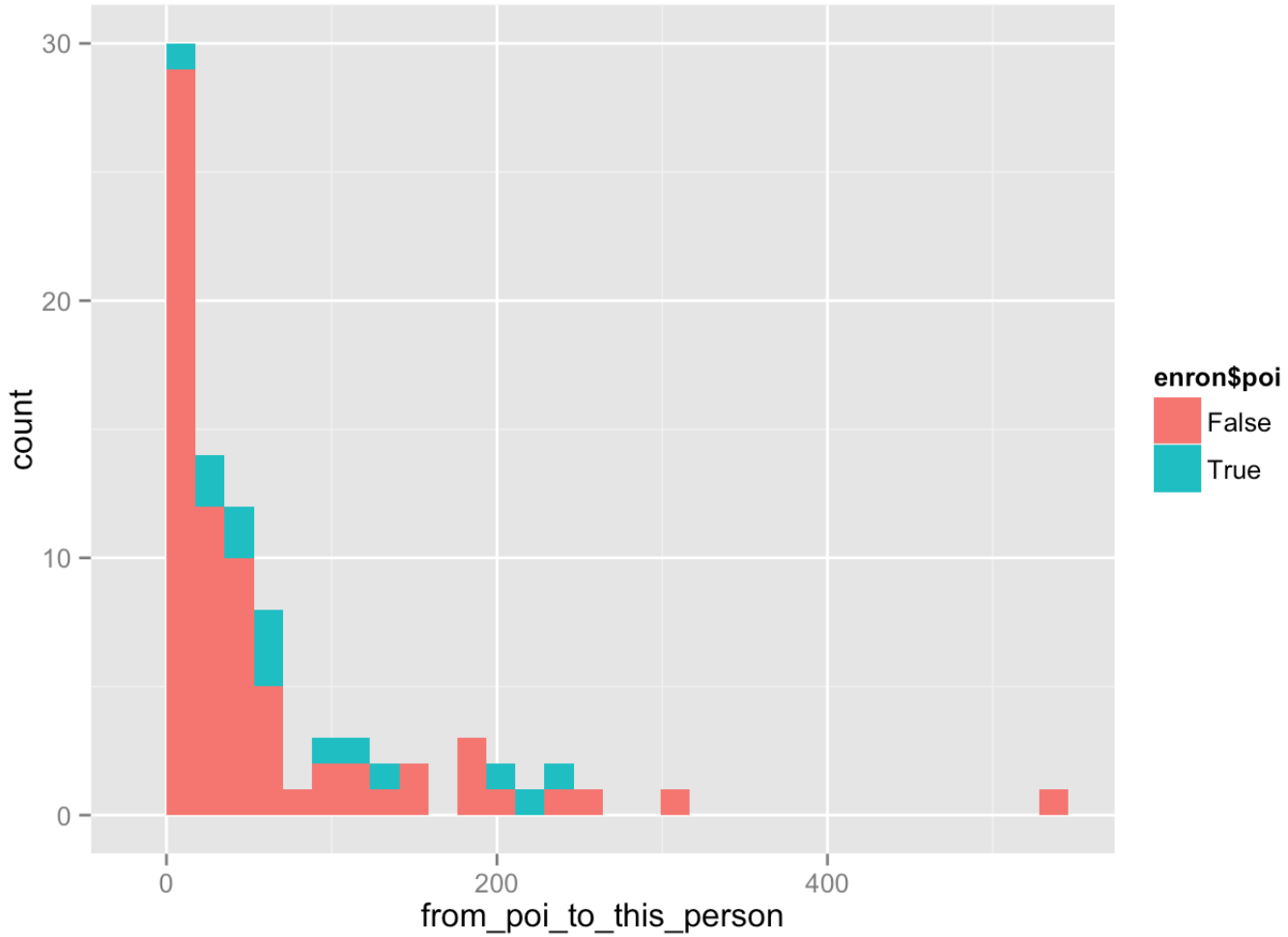

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



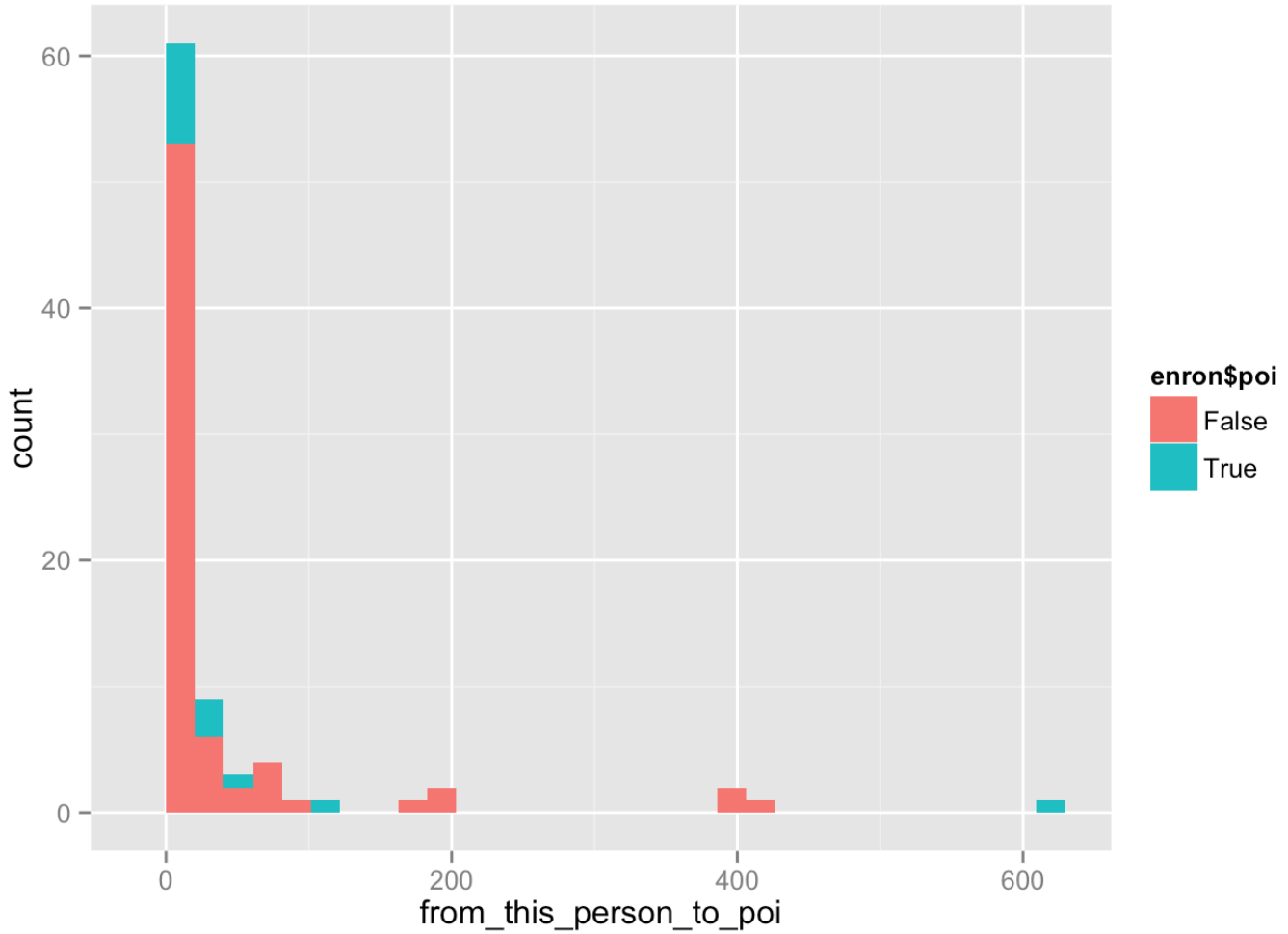

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



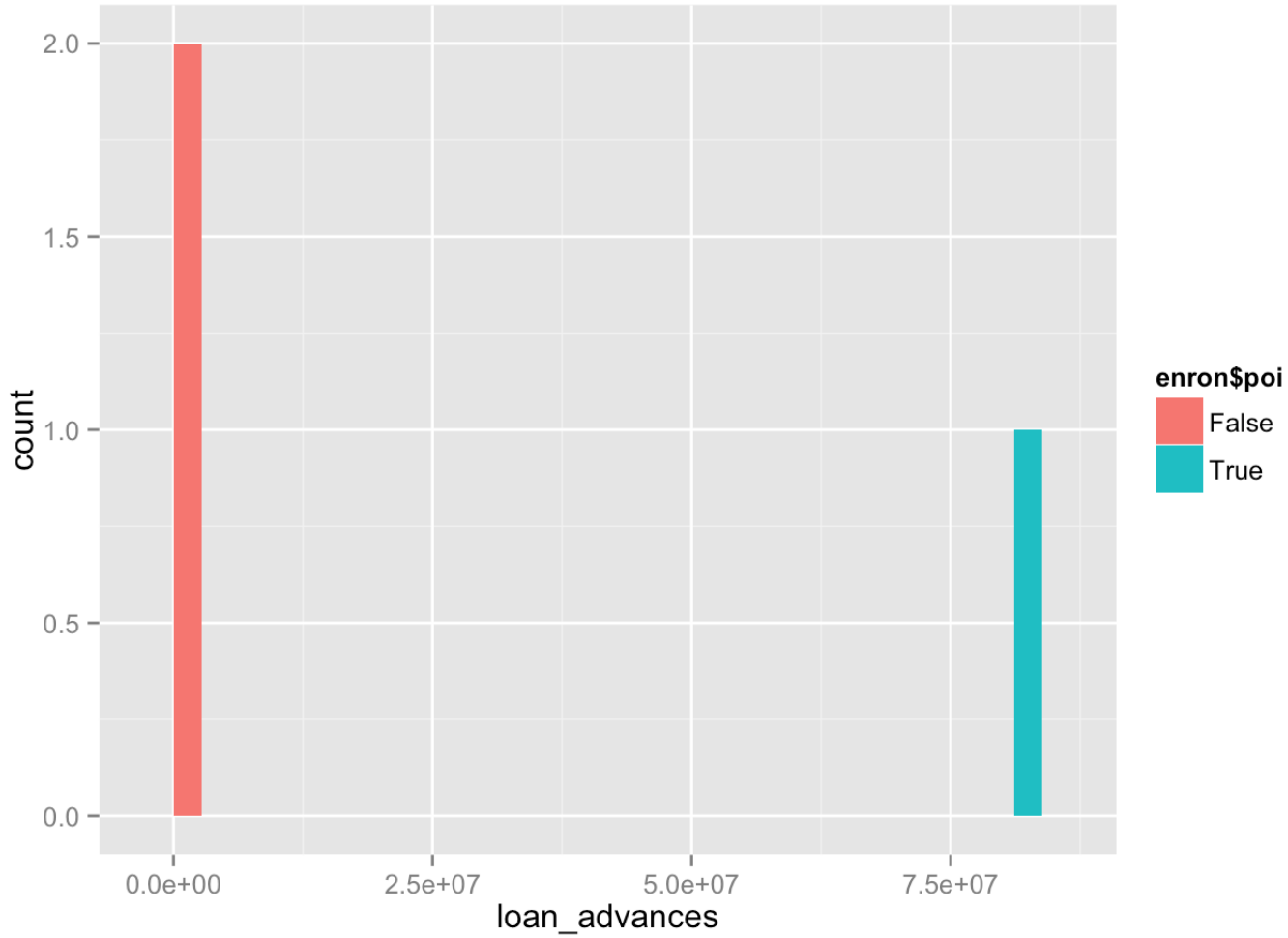

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



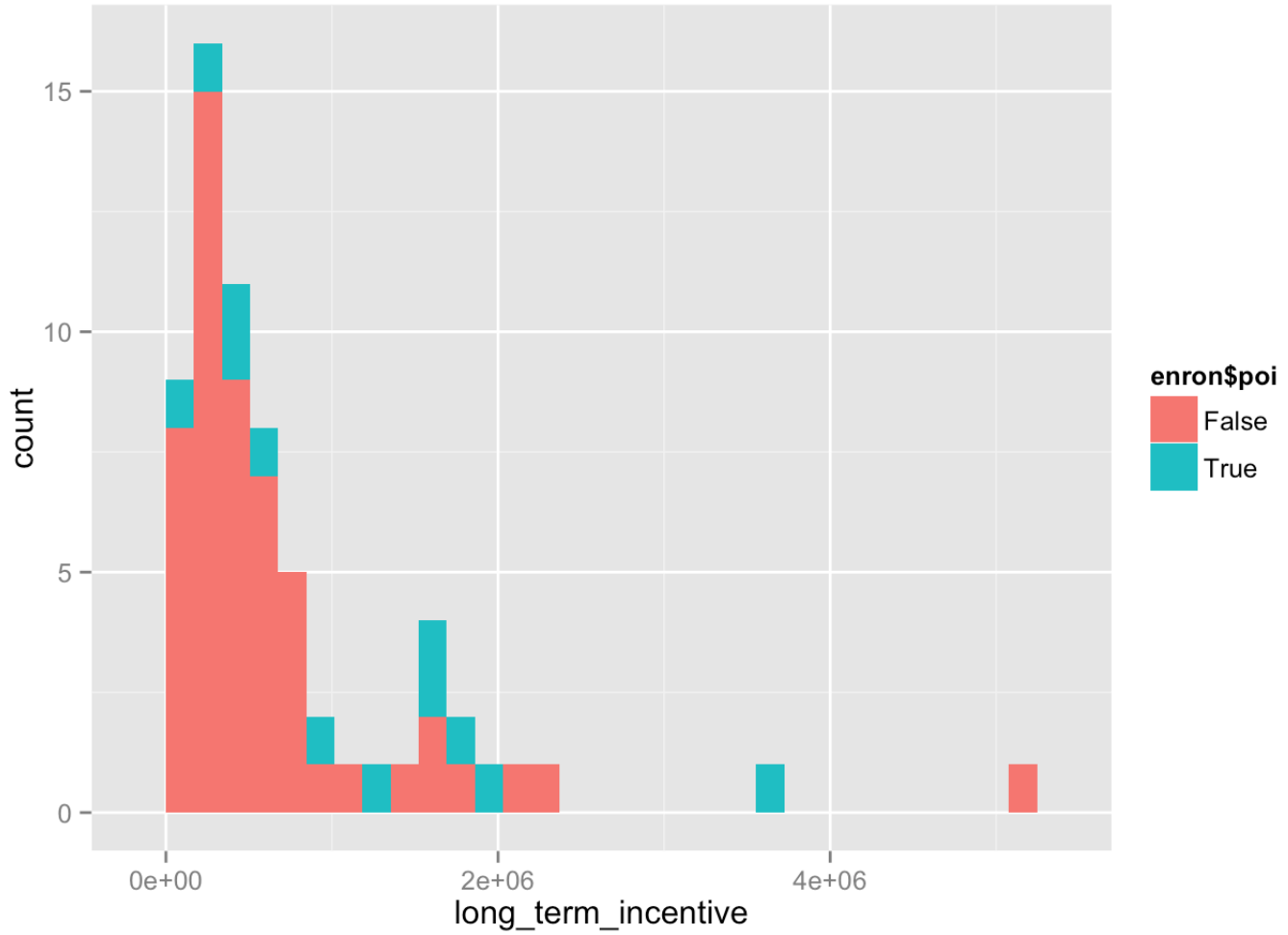

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



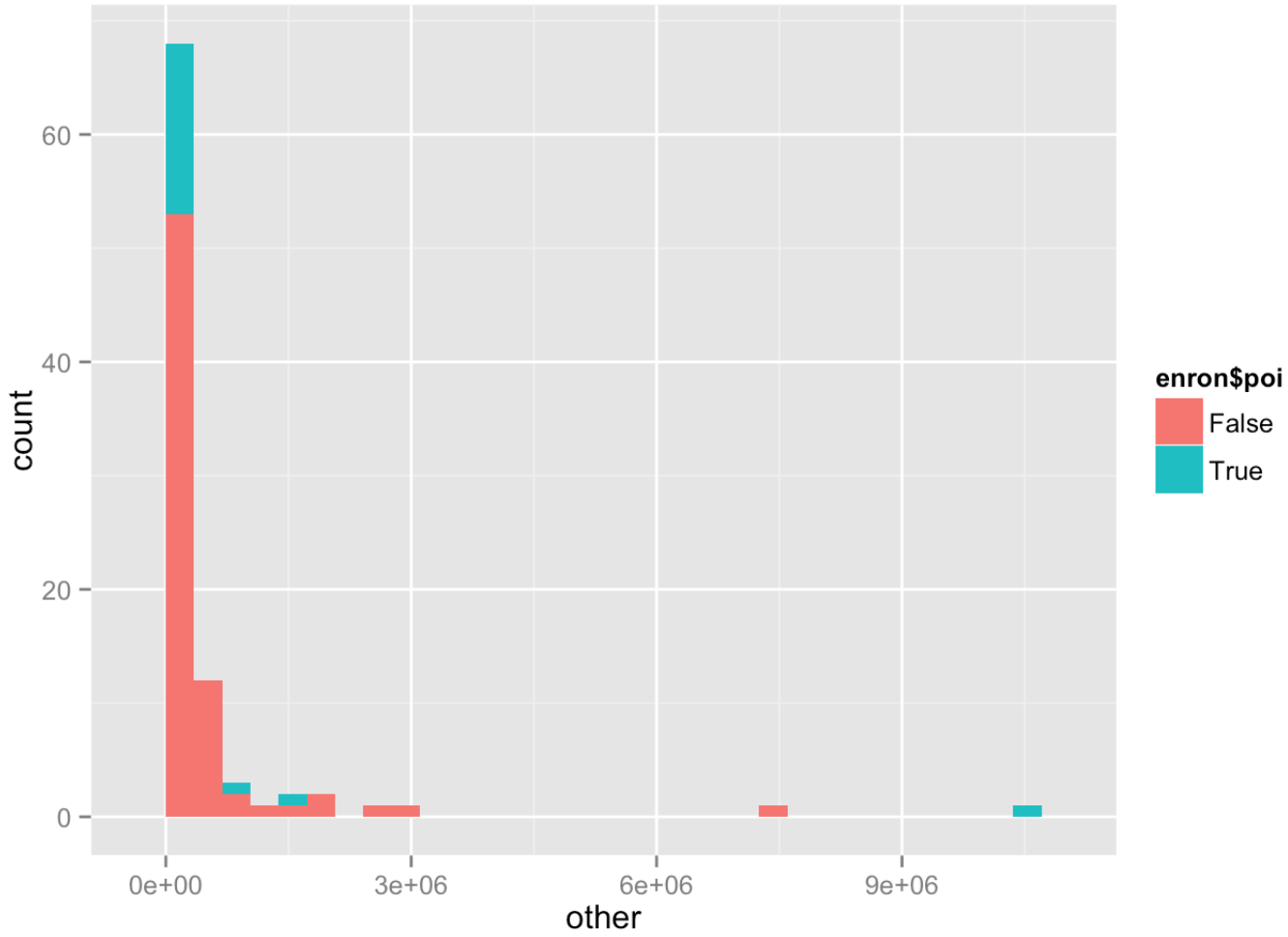

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

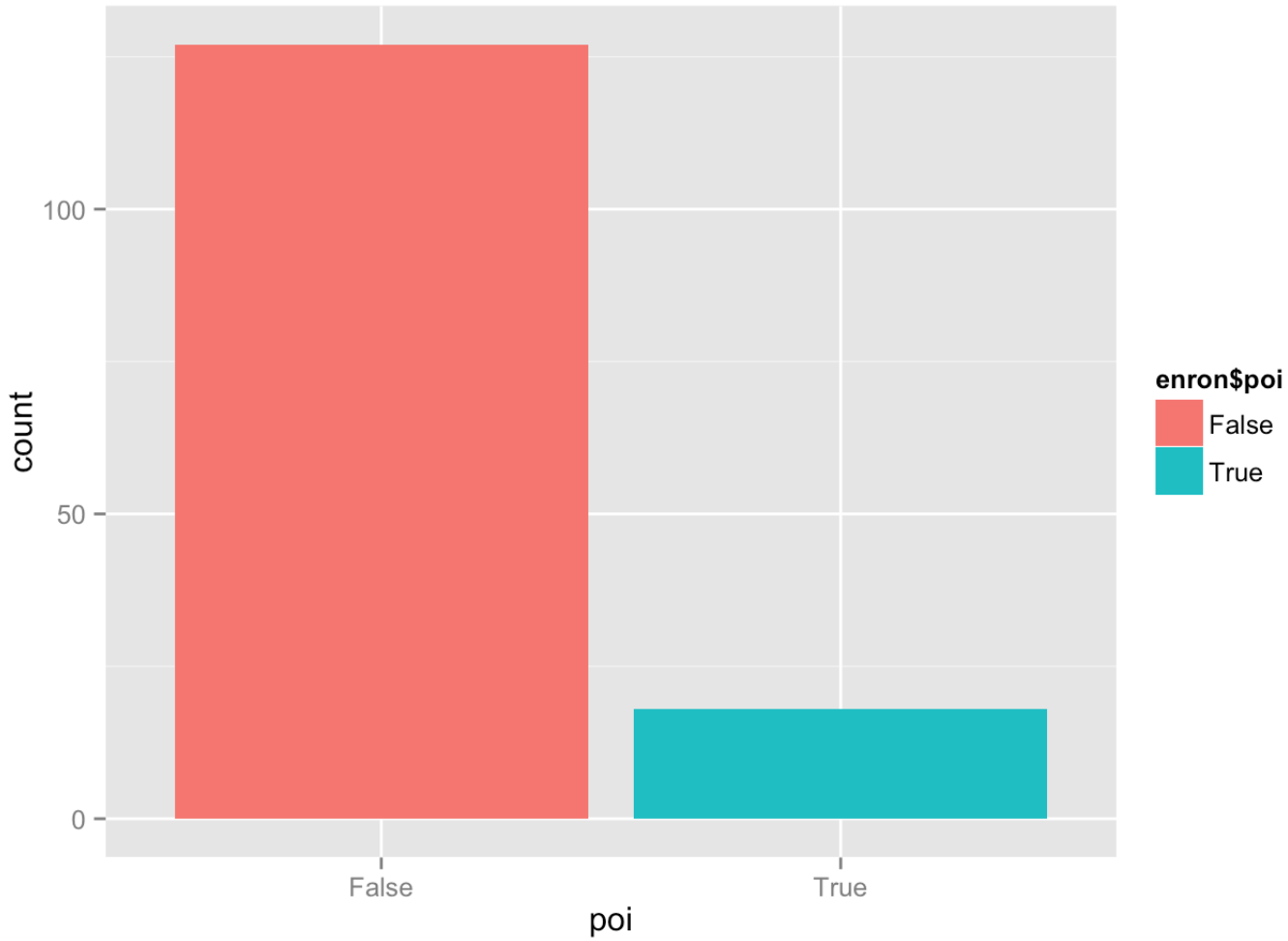



```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

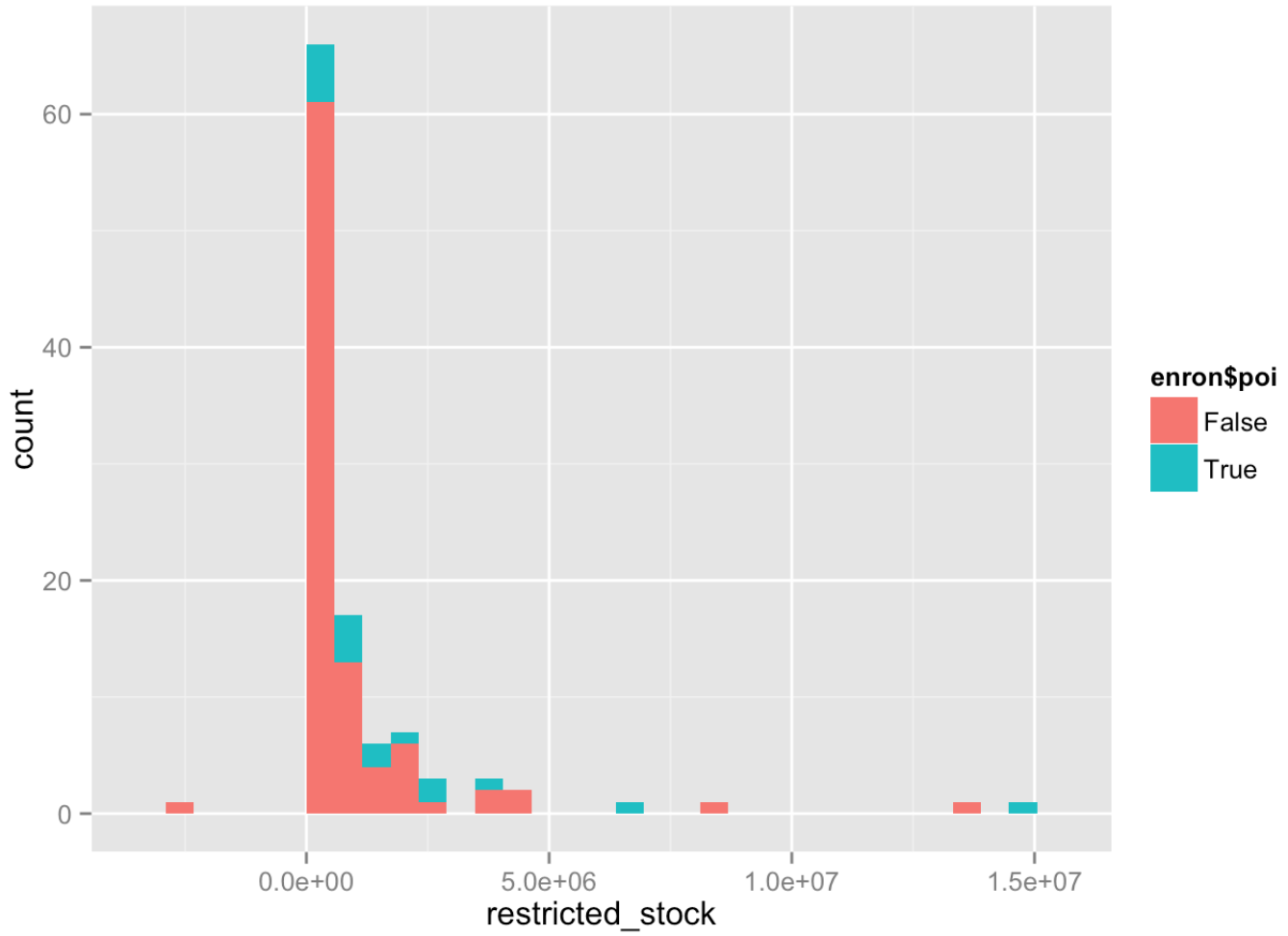



```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

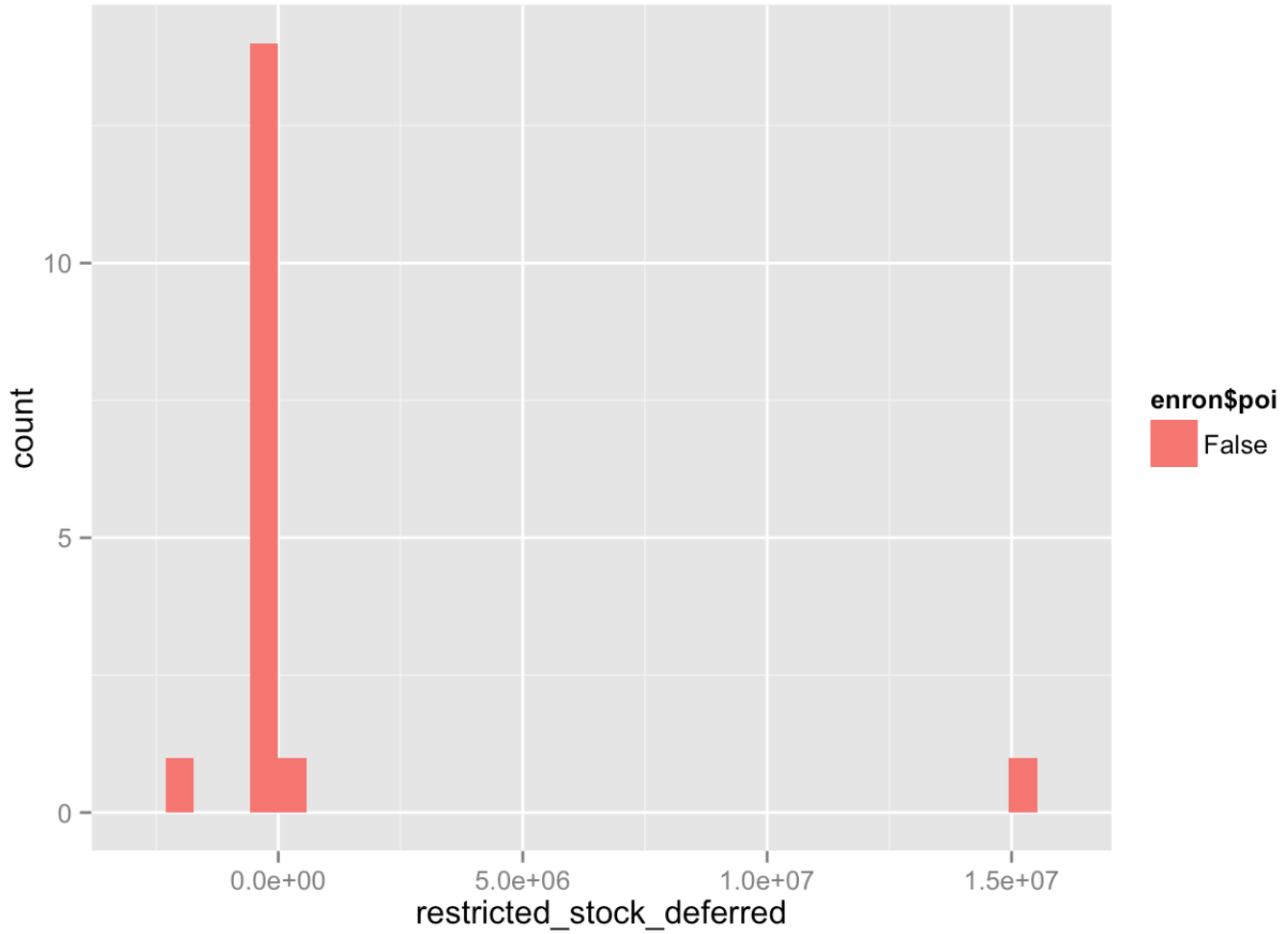

60
40
20
0
count
0e+00
3e+06
6e+06
9e+06
other
enron$poi
False
True


100
50
0
count
False
True
poi
enron$poi
False
True

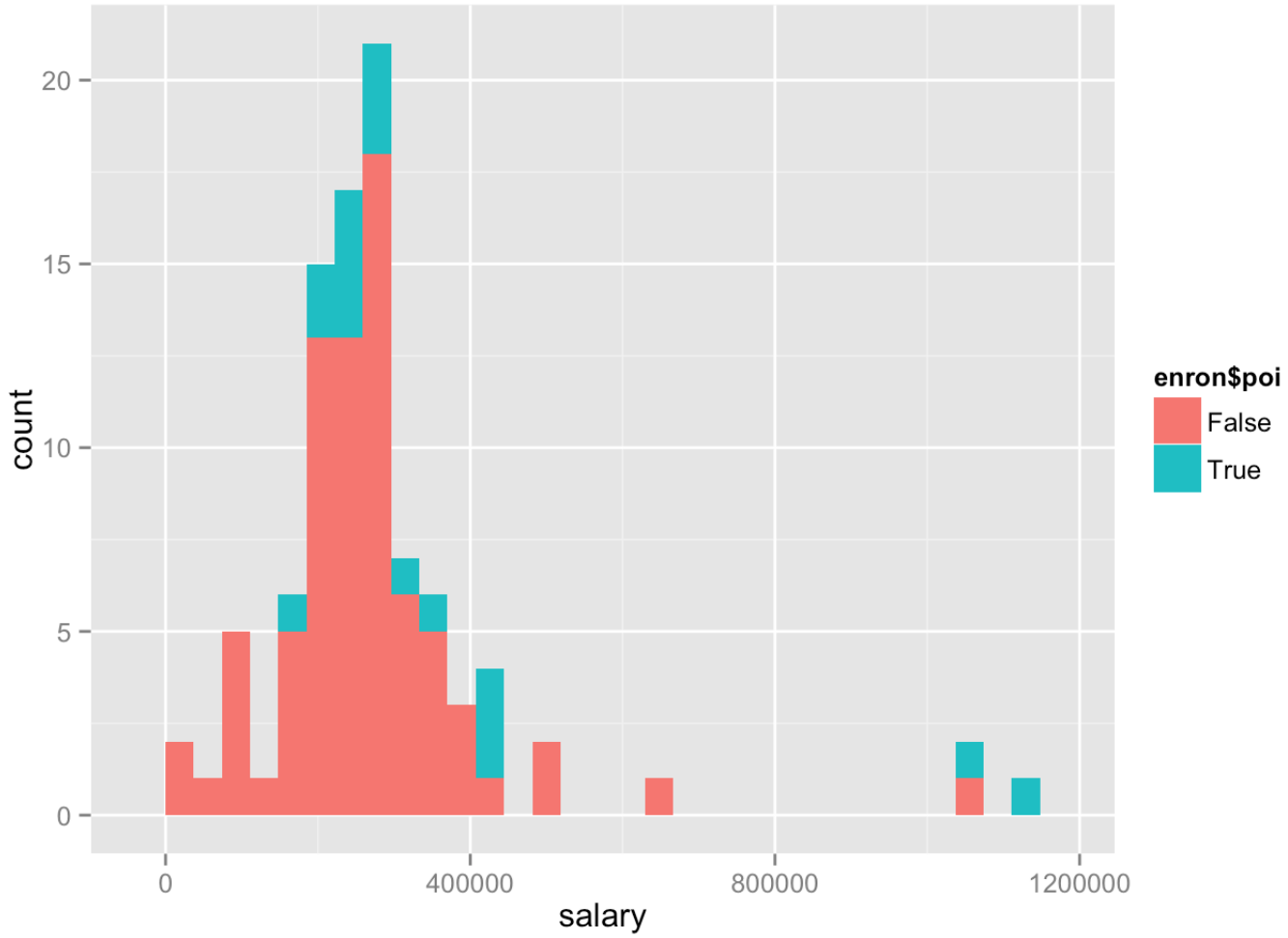
```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



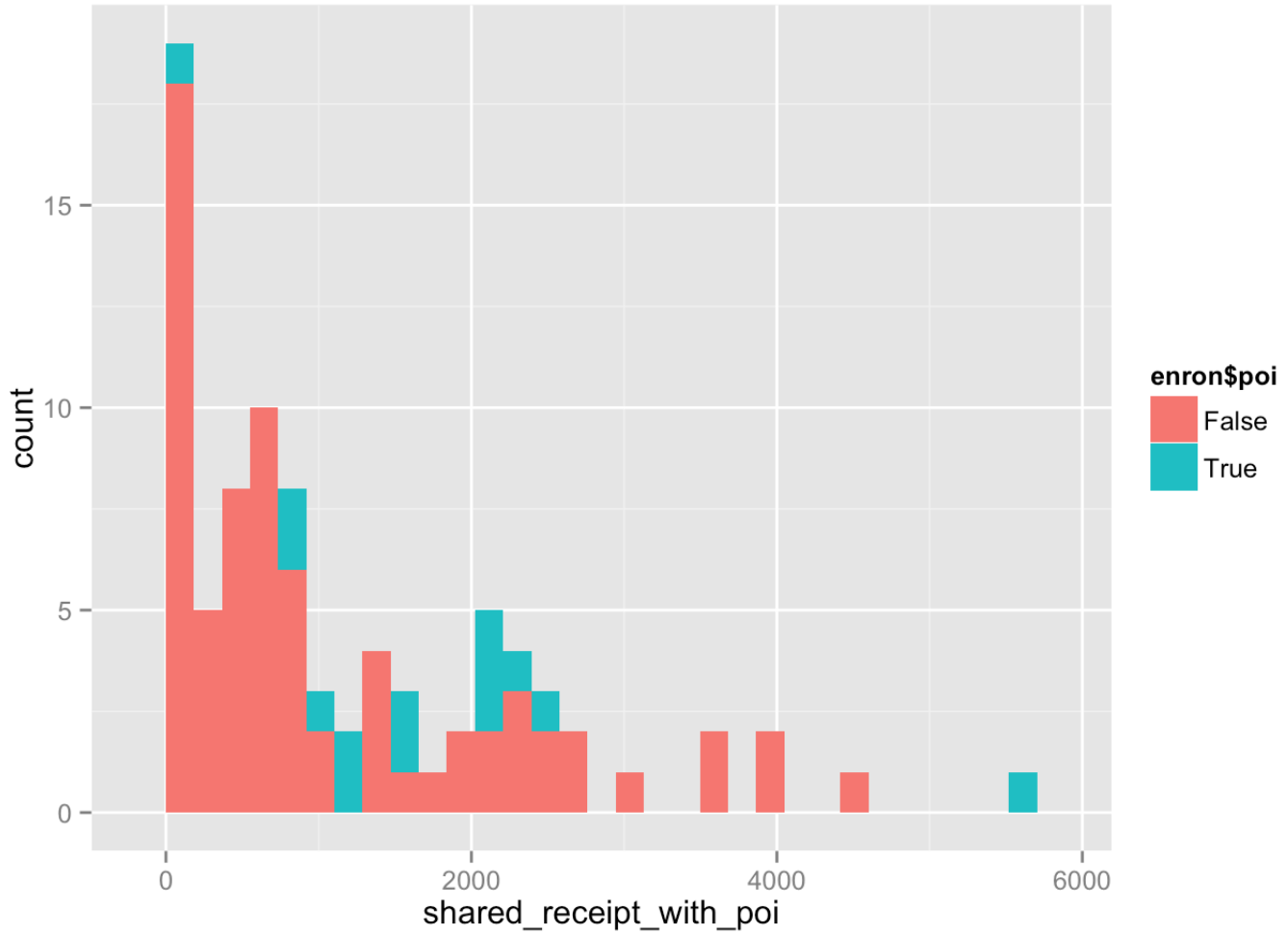

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



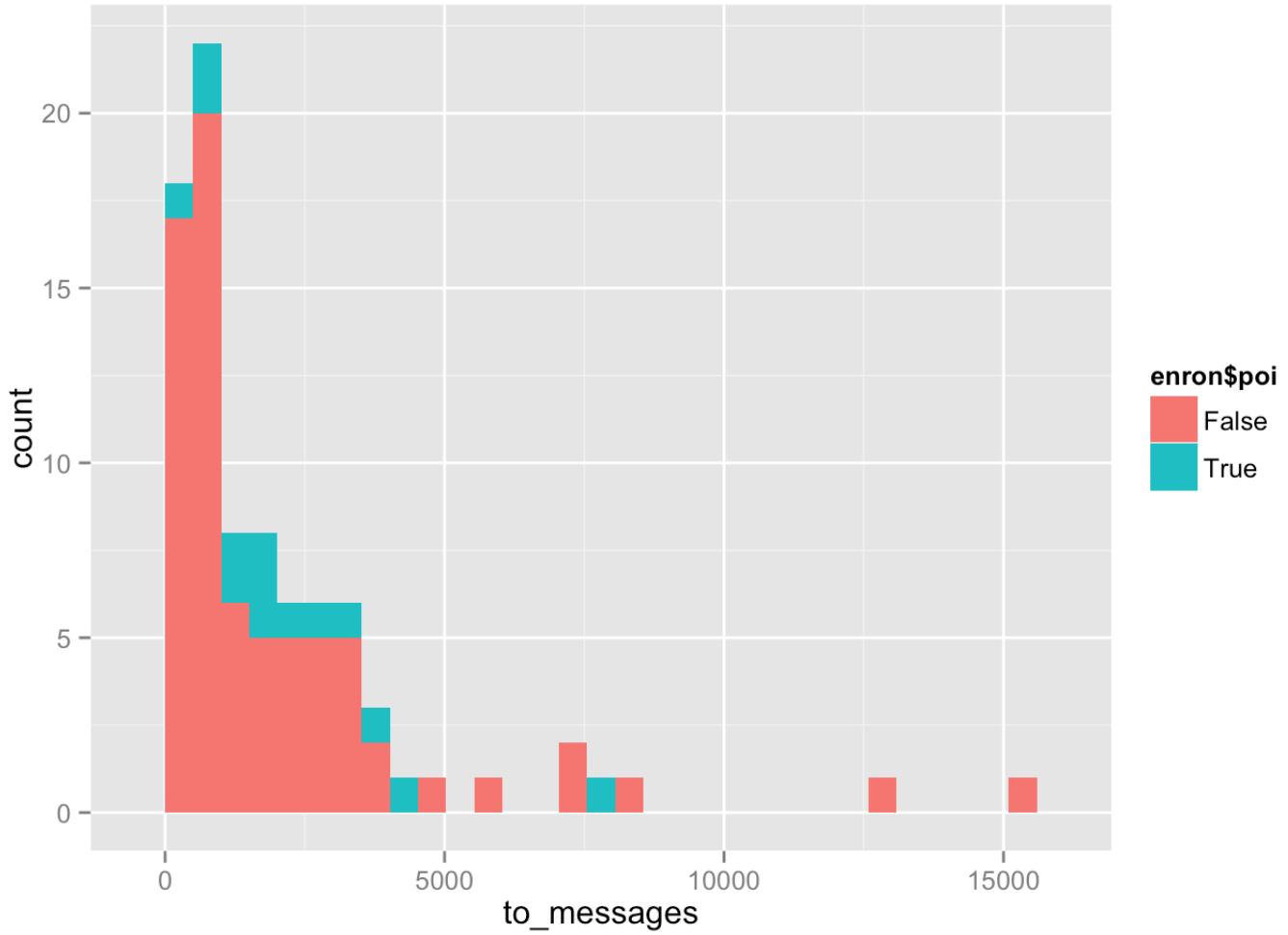

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



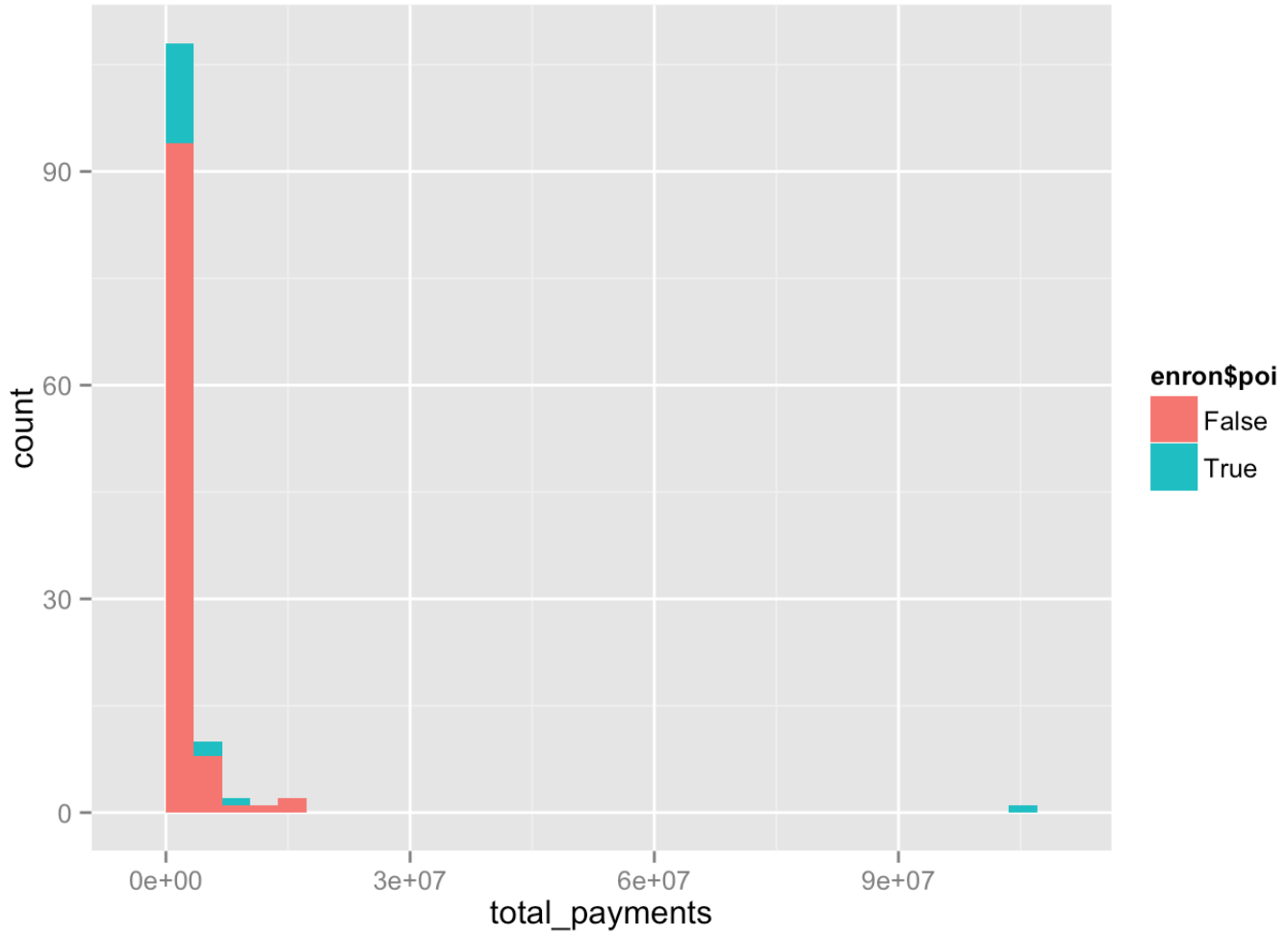

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



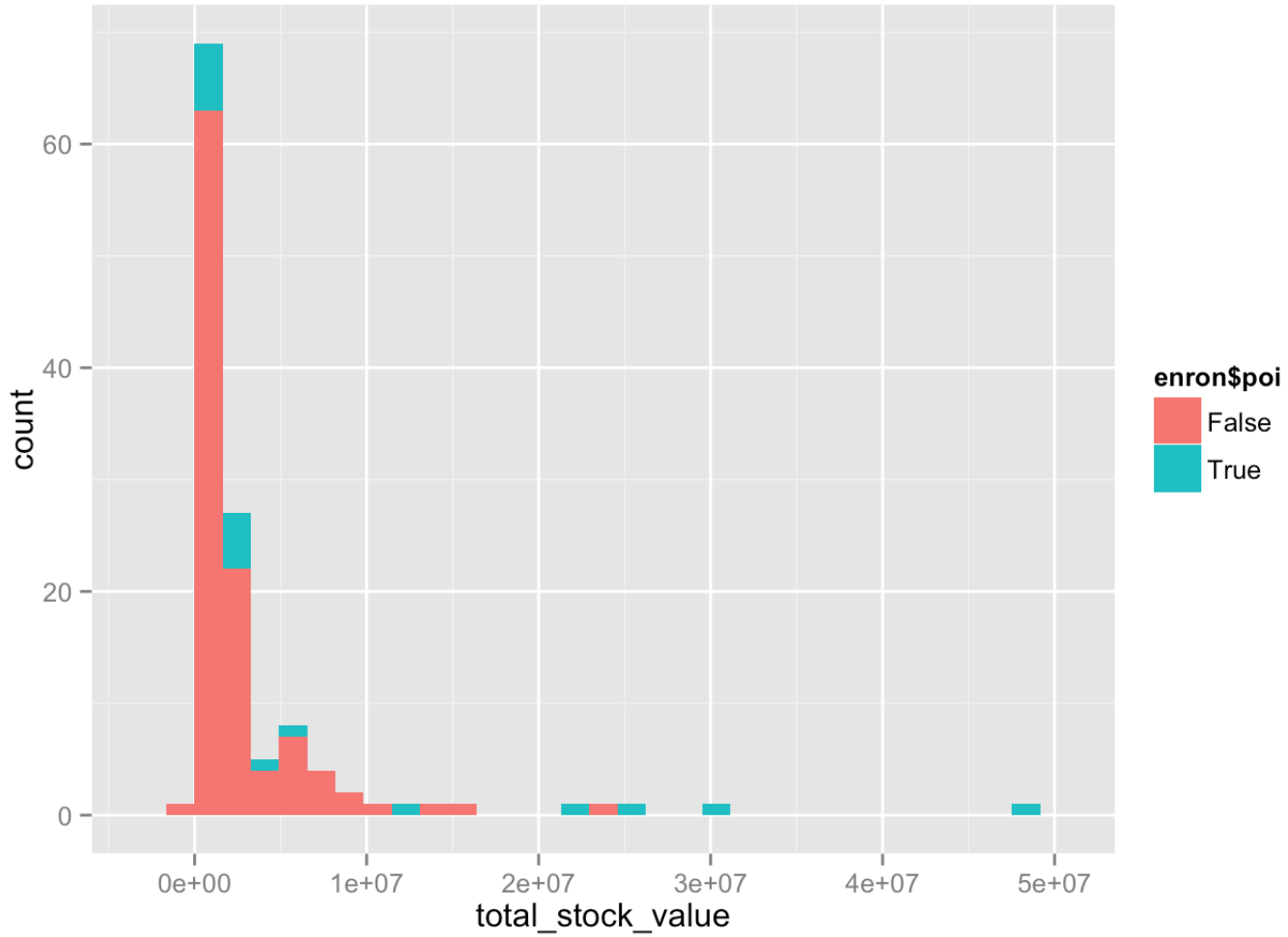

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

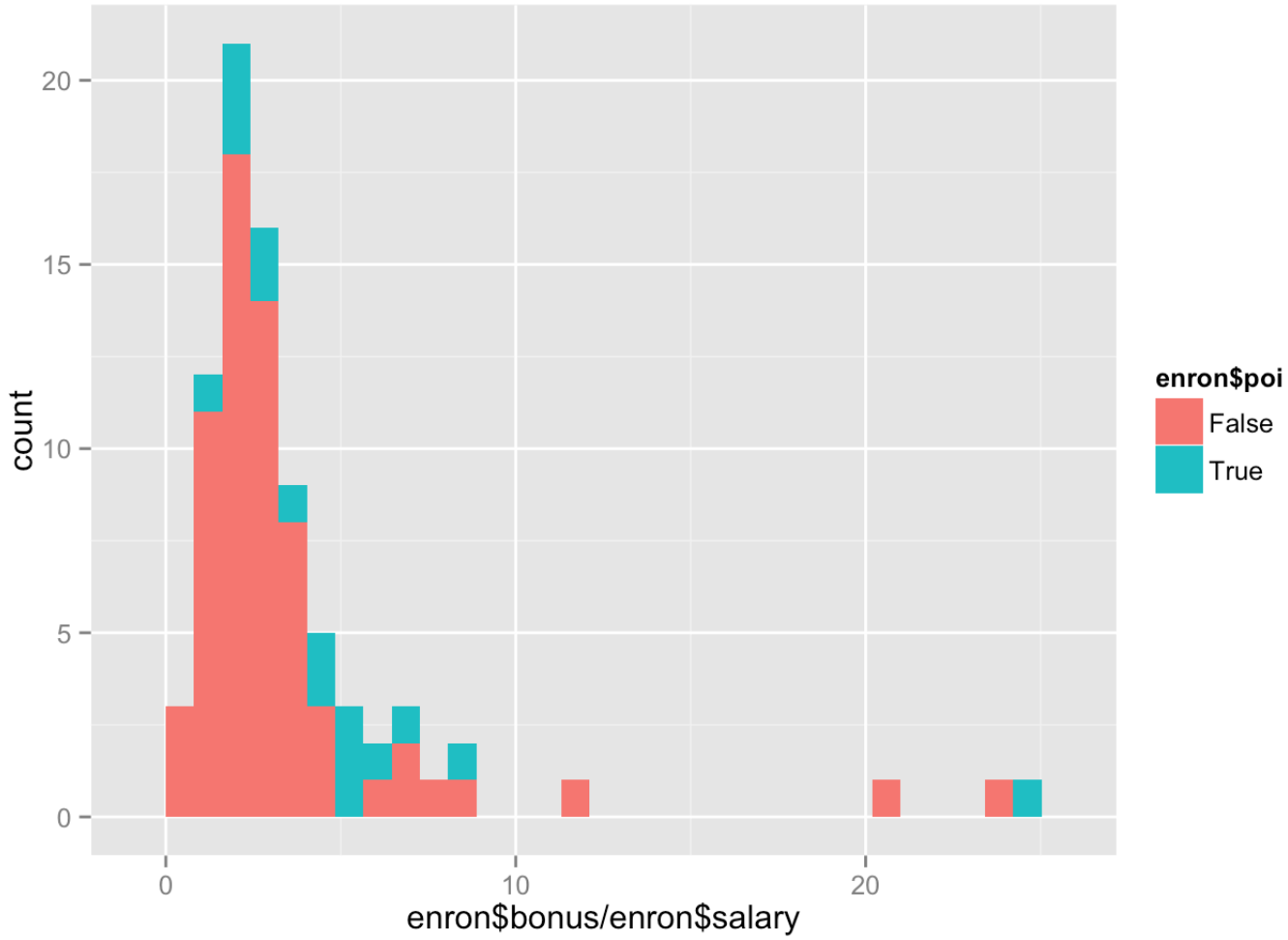



```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



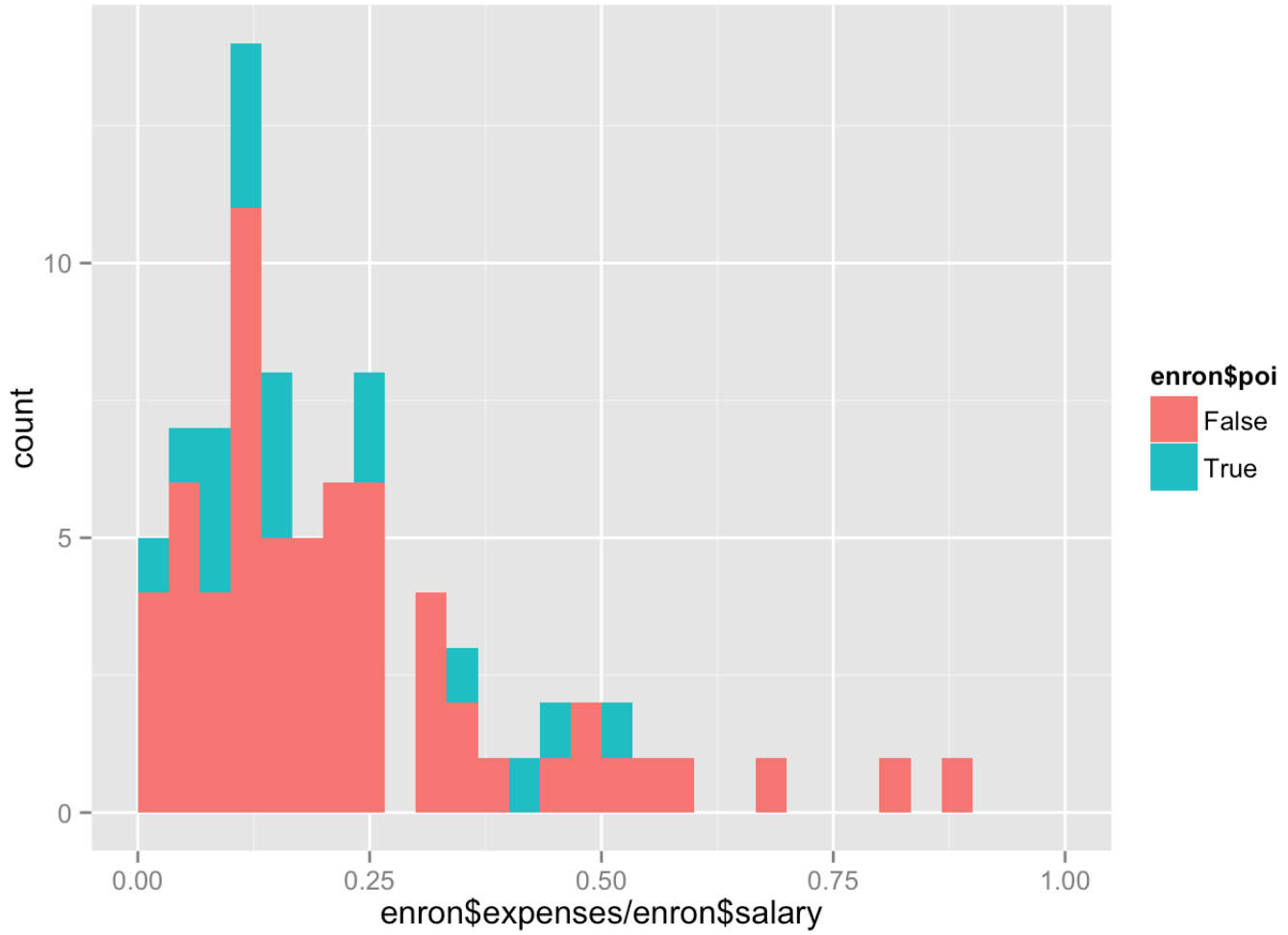

```
qplot(enron$bonus/enron$salary, fill = enron$poi)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



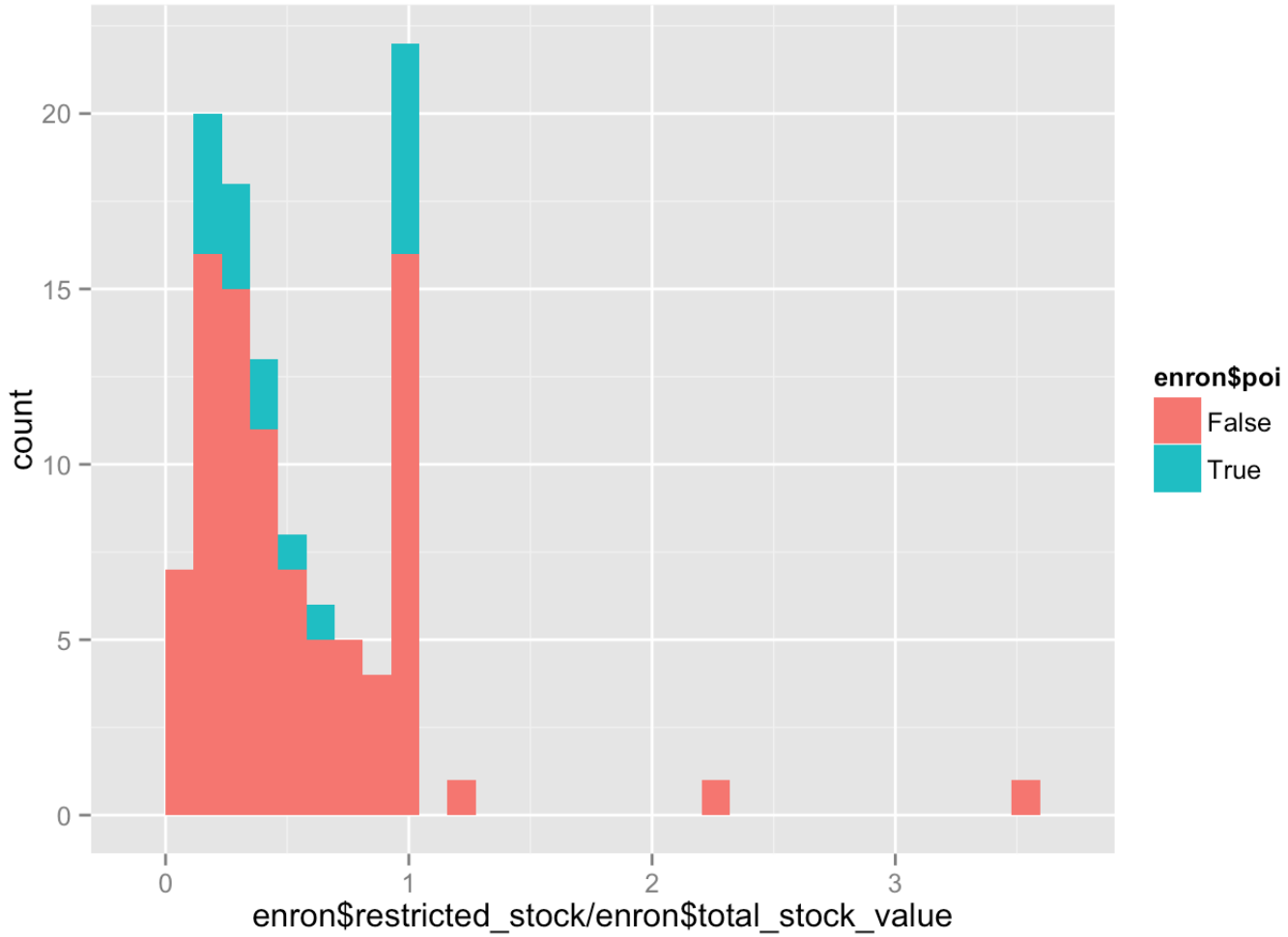

```
qplot(enron$expenses/enron$salary, fill = enron$poi) +
    scale_x_continuous(limits = c(0,1))
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



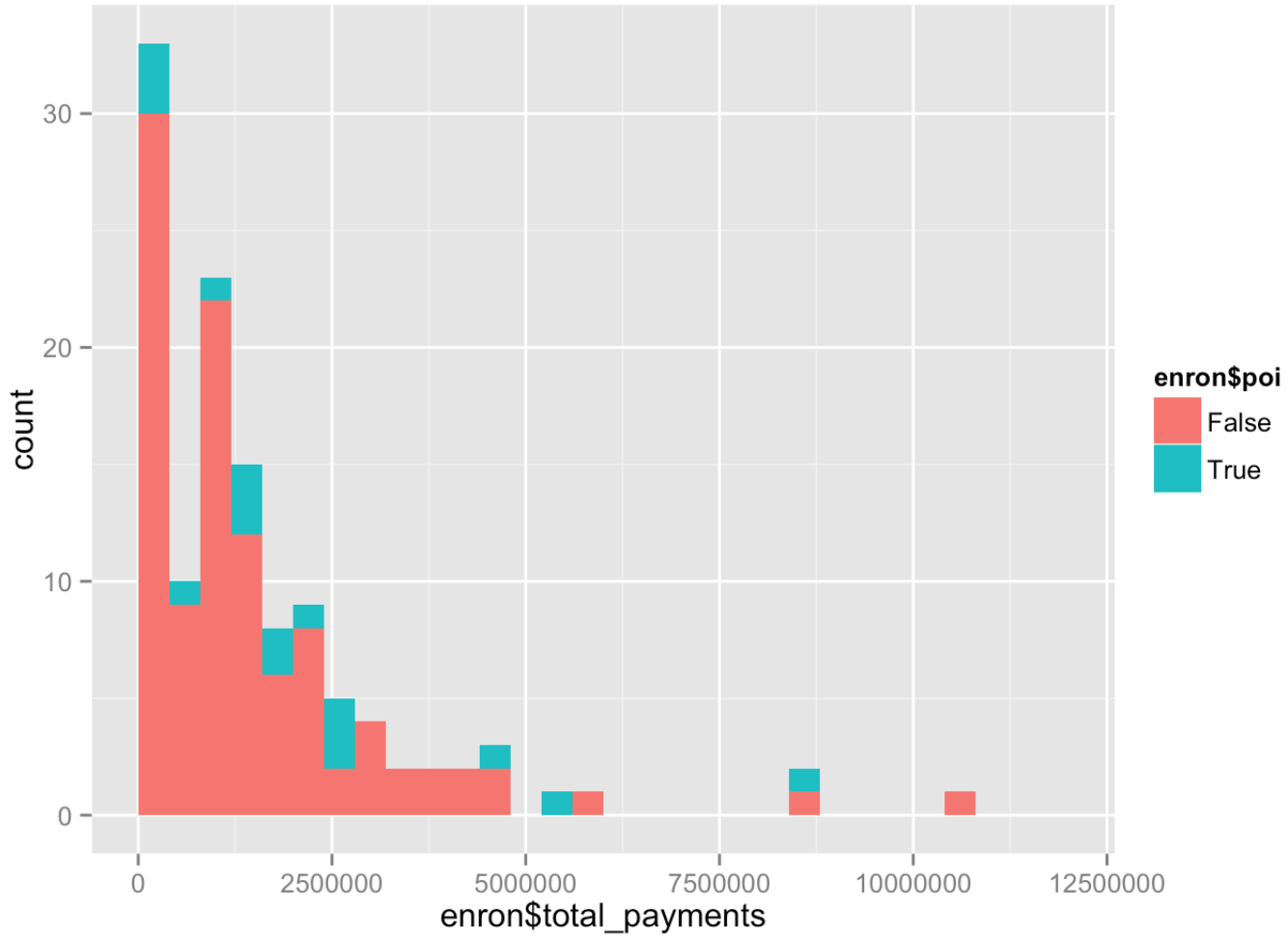

```
qplot(enron$restricted_stock/enron$total_stock_value, fill = enron$poi)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



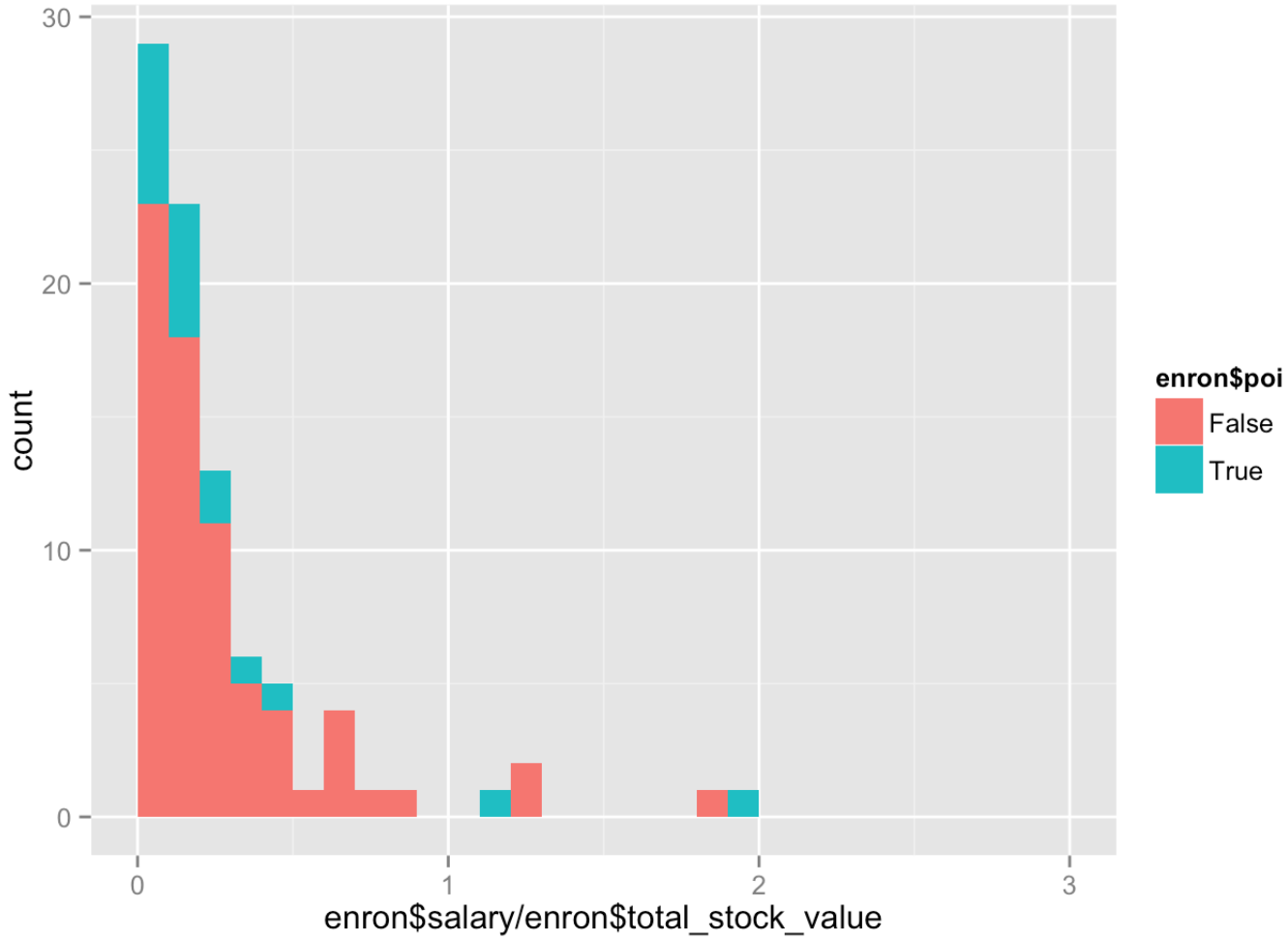

```
qplot(enron$total_payments, fill = enron$poi) + scale_x_continuous(
    limits = c(0,1.2e7))
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
qplot(enron$salary/enron$total_stock_value, fill = enron$poi) +
    scale_x_continuous(limits = c(0,3))
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```


30
20
10
0
count
0
1
2
3
enron$salary/enron$total_stock_value
enron$poi
False
True