"tried to scare my cat with a banana but i didn't expect this to happen!"

"My cats have trained me"

"THERE'S SO MANY CATS ON TWITTER!"

Comparing Proportions of Cat Related Tweets in the United States

Kaitlyn Lawlor
April 28, 2017

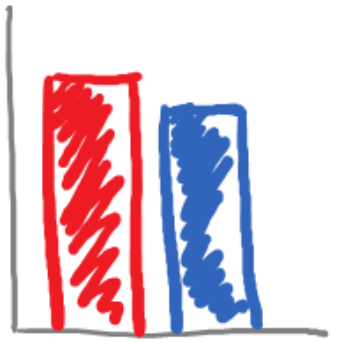
Background & Significance

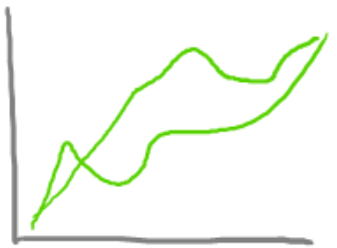
Youtube

popularity

marketing

Background & Significance (cont.)

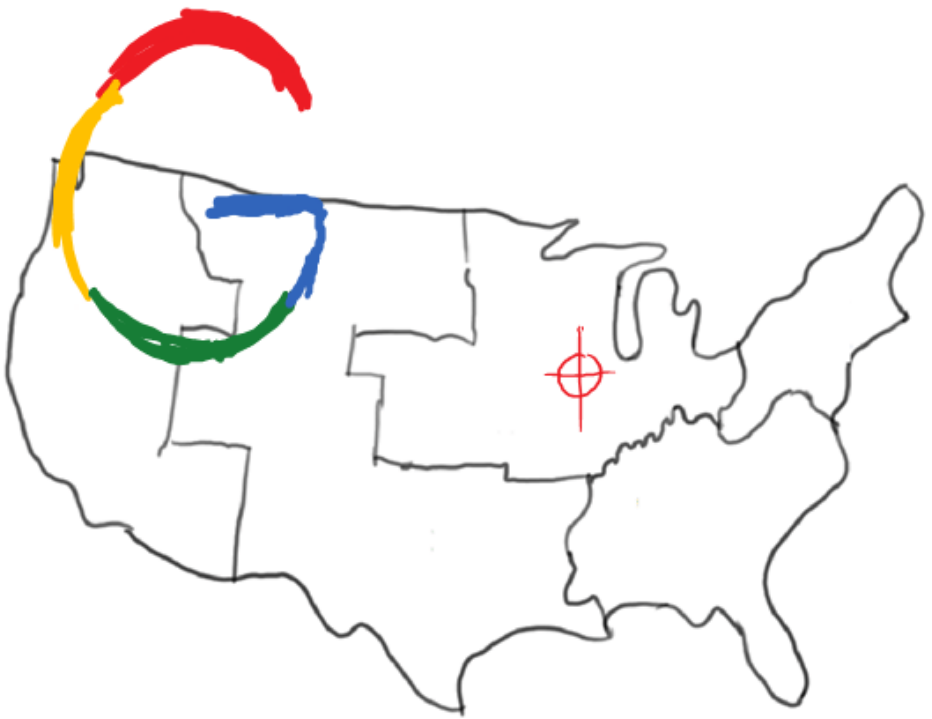




Studies

Objective: Report the region of the U.S. that is most interested in cats.

Methodology



- catTweets.df 5000 obs. of 16 variables
- catTweets2.df 5000 obs of 16 variables

"Cats" OR "Cats" OR "felines" OR "#cats"

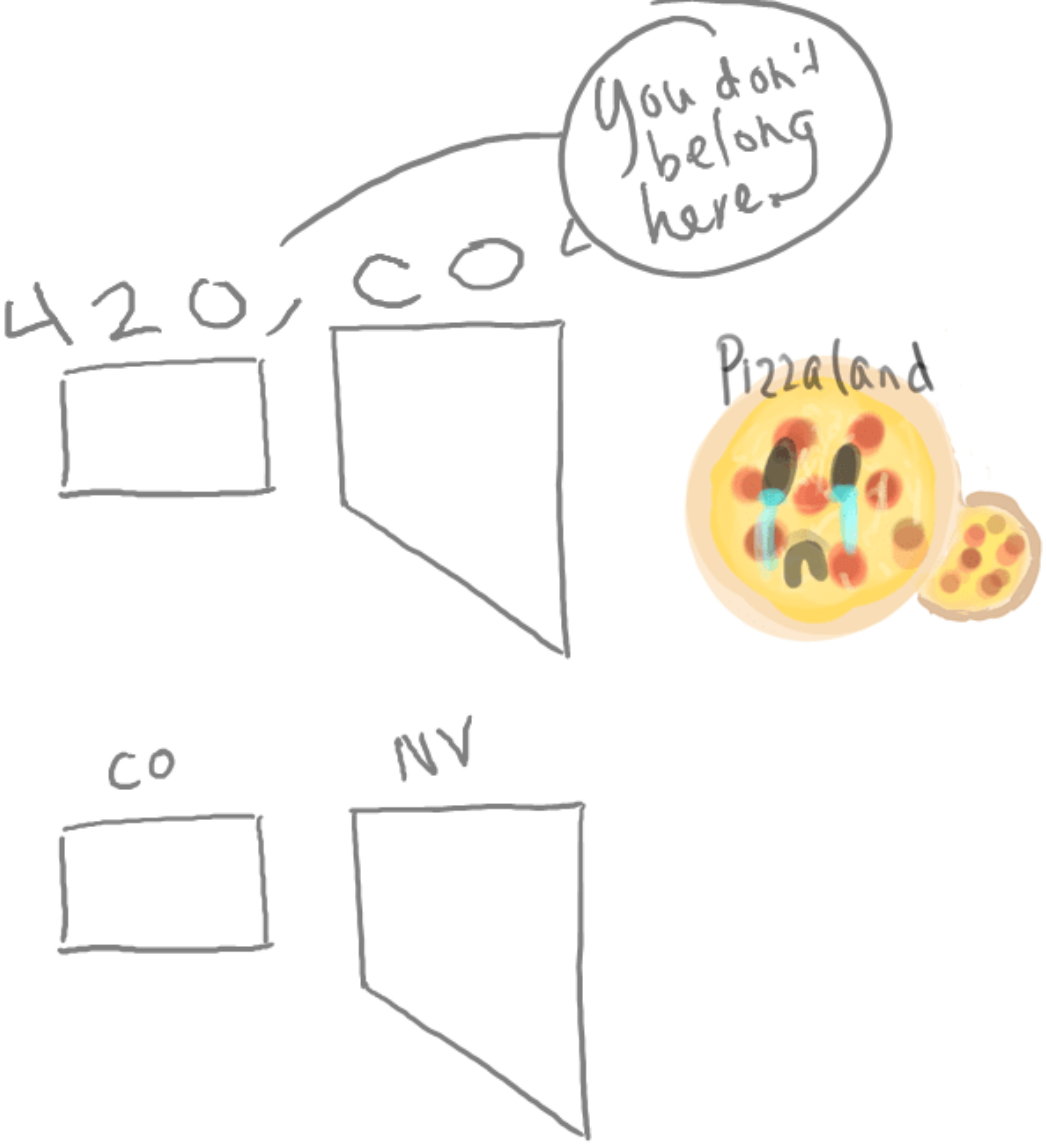
- twitteR
- ggmaps

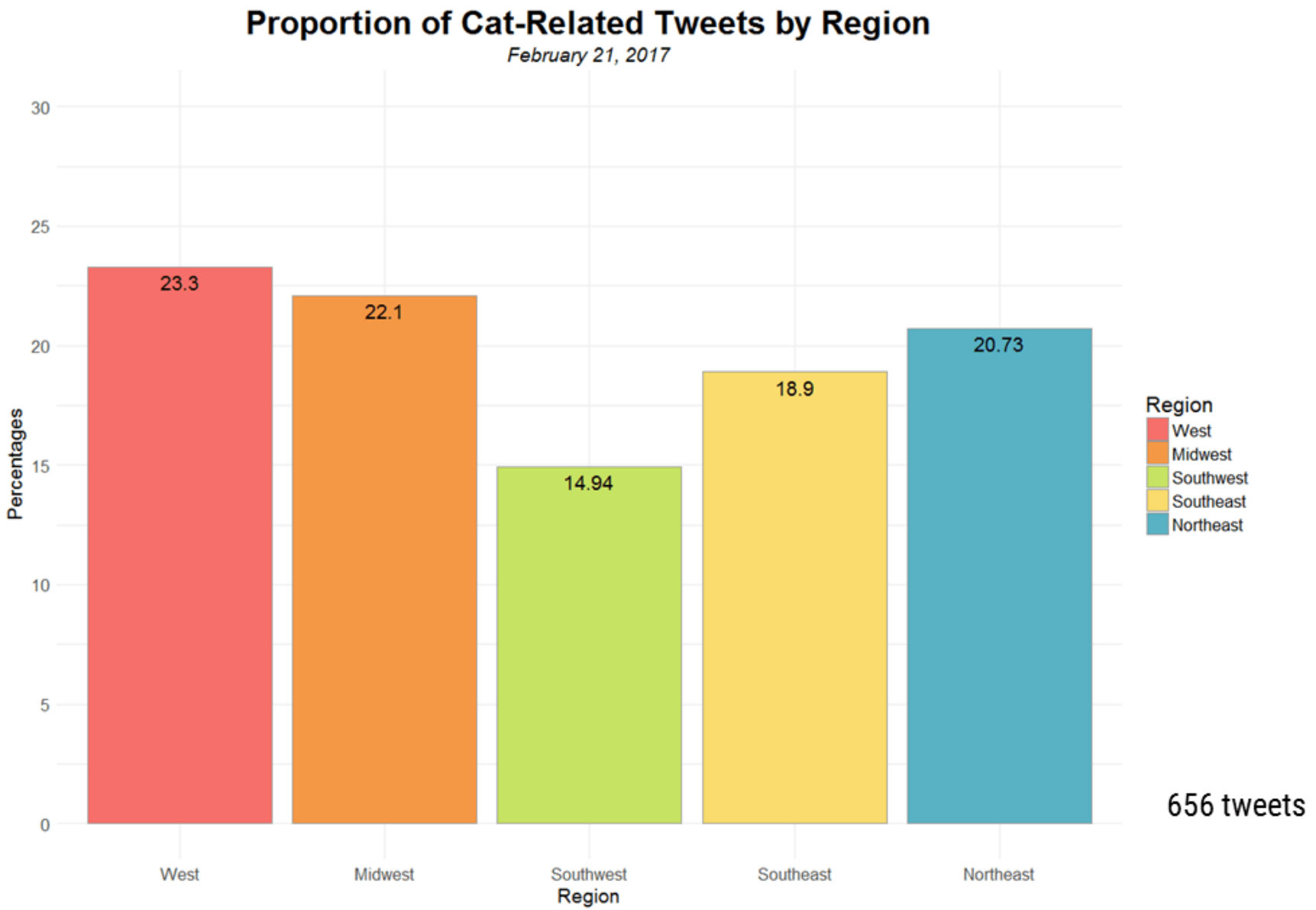
Methodology (cont.)

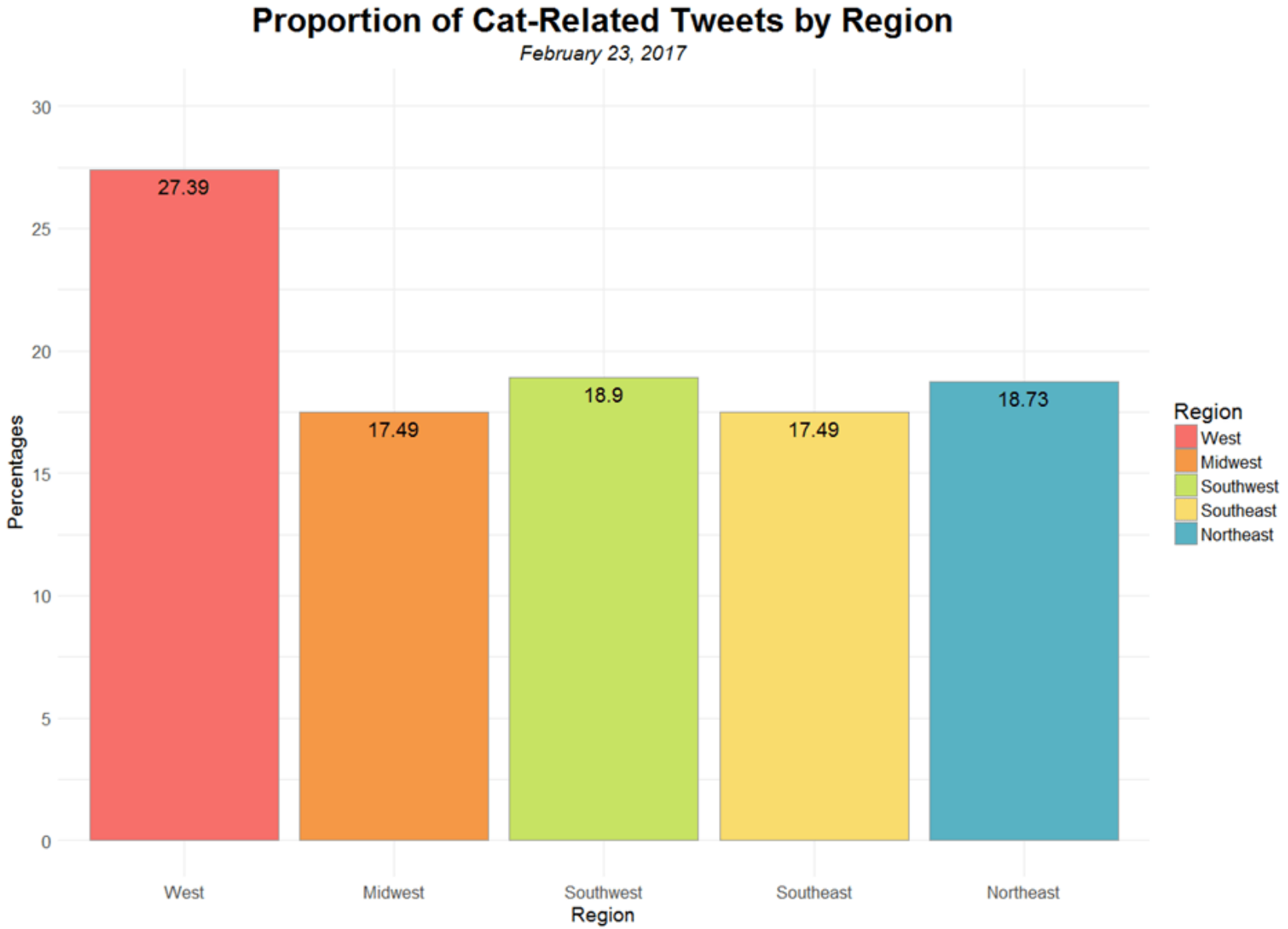
```
52  s = strsplit(catTweetUserLocation, ",")
53  # returns the 2nd element of a vector
54  get2nd <-function(x) {
55    x[2]
56  }
57
58  # extract the states
59  states = sapply(s, get2nd)
60
61  # remove any blank spaces
62  states = gsub(" ", "", states)
63  states = gsub("Alabama", "AL", states)
64  states = gsub("Alaska", "AK", states)
65  states = gsub("Arizona", "AZ", states)
```

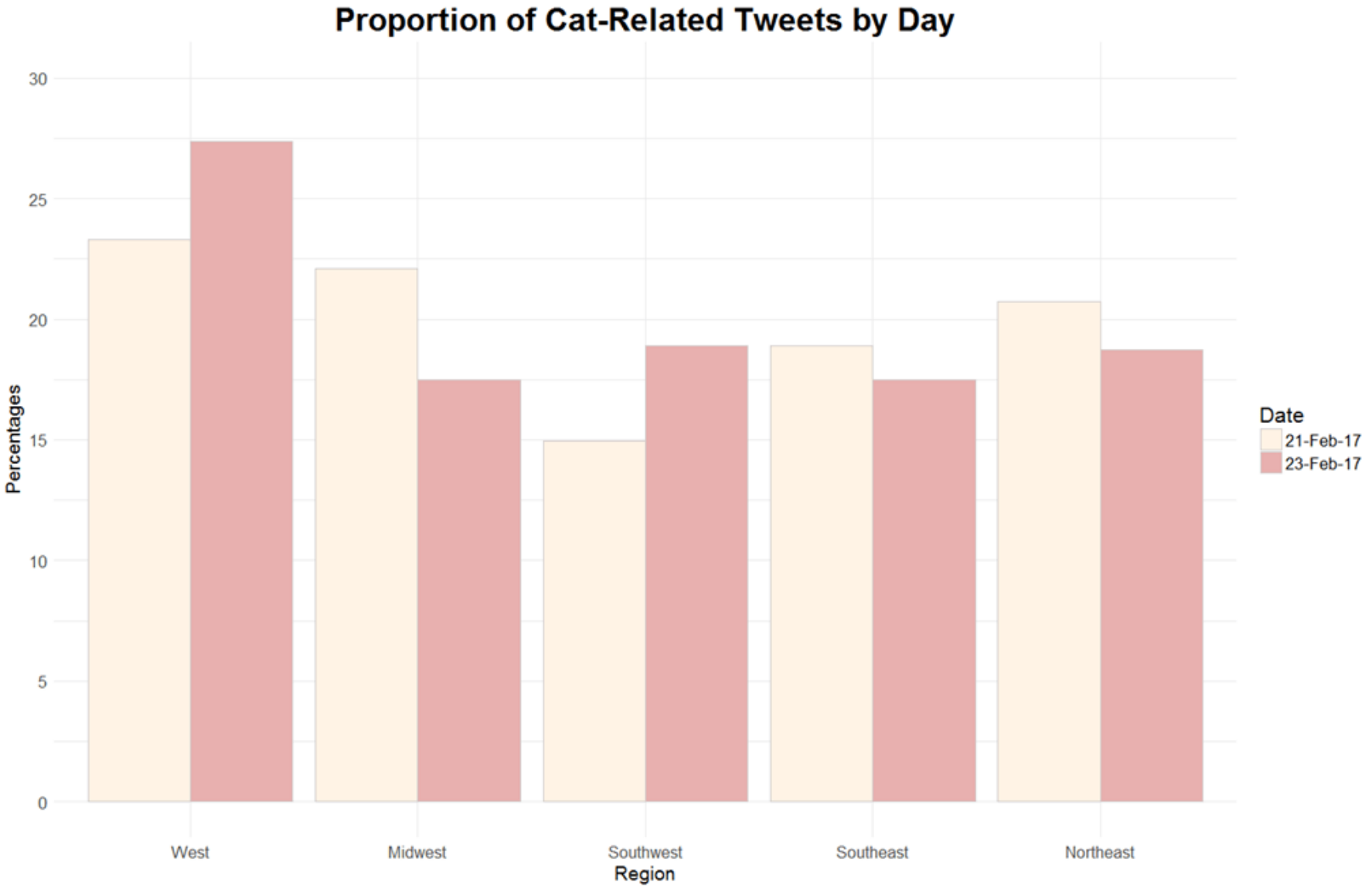
```
#WEST = 153
#Arizona, California, Neva
AZ <- grepl("AZ", states)
CA <- grepl("CA", states)
ID <- grepl("ID", states)
NV <- grepl("NV", states)
WA <- grepl("WA", states)
OR <- grepl("OR", states)
sum(AZ,CA,ID,NV,WA,OR)
```

pattern matching

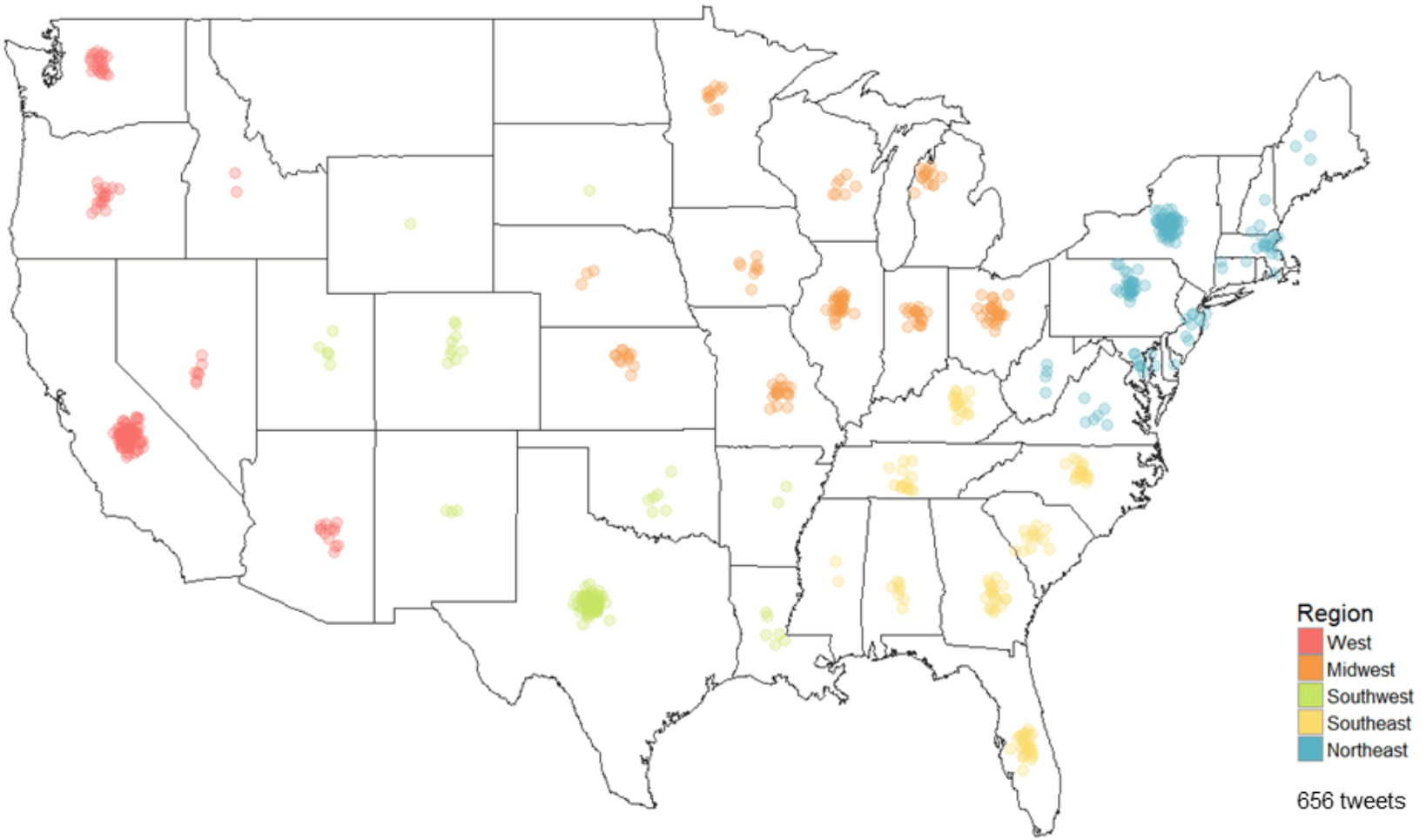



Proportion of Cat-Related Tweets by Region
February 21, 2017
Percentages
30
25
20
15
10
5
0
23.3
22.1
14.94
18.9
20.73
West
Midwest
Southwest
Southeast
Northeast
Region
Region
West
Midwest
Southwest
Southeast
Northeast
656 tweets


Proportion of Cat-Related Tweets by Region
February 23, 2017
Percentages
30
25
20
15
10
5
0
27.39
17.49
18.9
17.49
18.73
West
Midwest
Southwest
Southeast
Northeast
Region
Region
West
Midwest
Southwest
Southeast
Northeast


Proportion of Cat-Related Tweets by Day
Percentages
30
25
20
15
10
5
0
West
Midwest
Southwest
Southeast
Northeast
Region
Date
21-Feb-17
23-Feb-17

Map of Twitter Posts from Feb. 21, 2017



Chi-Squared Test

H_0: The distribution of tweets across geographic regions is not associated with the day.

H_A: The distribution of tweets across geographic regions is associated with the day.

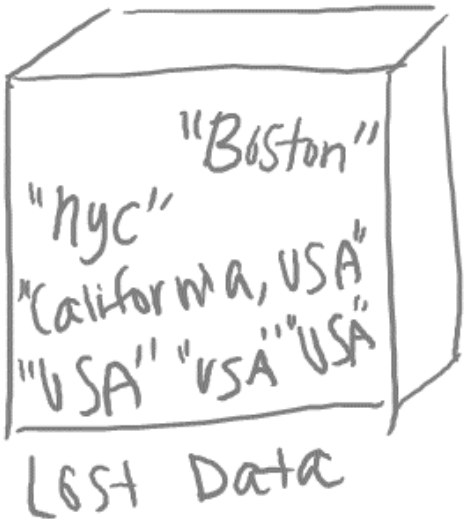
```
           West Midwest Southwest Southeast Northeast
2/21/2017  153     145        98       124       136
2/23/2017  155      99        99       108       106
> chisq.test((cTest))

        Pearson's Chi-squared test

data:  (cTest)
X-squared = 7.0734, df = 4, p-value = 0.1321
```

Significance Level: 0.05

Fail to reject null hypothesis

Limitations

Lots of data is lost
Only text-based tweets are considered
Not representative of total population



Future Work

Use a larger sample size
Longer time periods
By state instead of region

Acknowledgements

- Dr. Garrett Dancik
  - Professor at Eastern Connecticut State University
- Nathan Lawlor

Thank you 

Summary

- Objective: Find the region of the U.S. that is most interested in cats
- For both days, the West region had the highest proportion of cat tweets, 23.3% (153/656) and 27.3% (155/567)
- The null hypothesis is not rejected $0.1 \not< 0.05$
- Software: TwitteR package