**Comparing the Proportions of Cat-related Tweets in the United States**

Kaitlyn Lawlor

**Abstract**

Cats are very popular animals, especially when it comes to the internet. Internet users spend a tremendous amount of time engrossed in cat-related media when he or she could be using that time working or studying. Although cats appear to dominate the internet, there are very few studies that explain this phenomenon. The focus of this paper is to identify the region in the United States that is most interested in cats by analyzing Twitter posts that contain pre-defined terms relevant to felines. This study uses RStudio and the Twitter API to collect cat-related tweets from two separate days. The results indicate that the West is most interested in cats; the region had the highest proportion of cat-related tweets on both days. Furthermore, this work suggests that there is no association between geographic regions and the day the cat-related tweets were taken from.

1. **Background and Significance**

Twitter is an online microblogging service. There are approximately 313 active users and one billion unique visits monthly to sites with embedded tweets [1]. Unregistered users still have the capability to view posts through users’ Twitter feeds, thus, there is still communication between the person and the people that view the tweet. On the other hand, an important, distinct feature of Twitter that is essential to understand is the ability to “follow” another user. A Twitter user can follow another Twitter user and need not to follow back; following means the user is notified of all Tweets from the user that they followed. Although most people view Twitter as a form of entertainment or a source of news, others—such as researchers—are quite interested in analyzing trends via Twitter data, and furthermore, exploring sentiment analysis, microblogging and data mining.

Twitter can be described as a microblogging website where “users can describe their status in short posts distributed by instant messages, mobile phones, email or the Web” [2]. Some researchers are interested in discovering why so many people use Twitter as their form of communication, conducting studies to explore the topological characteristics of Twitter and its power as a new medium of information sharing” [3]. However, the service is not merely a place where people conglomerate to broadcast small, mundane occurrences in their lives, such as their feelings, thoughts, or current activity.

In fact, Twitter is a prime focus for analysis of data because some researchers believe that complex queries can be answered by streaming Data form Twitter. In the research article, “A Framework for Summarizing and Analyzing Twitter Feeds,” the authors stated their goal was to “build a summary of microblogging data, focusing on Twitter feeds, that can fit a limited memory budget, and can help answer complex queries” [4]. The authors, Xintian Yang, Amol Ghoting, and Yiye Ruan, develop a framework that is composed of multiple parts, such as SPUR (a batch summarization and compression algorithm) and TED-SPUR, a topic and event based analytics tool to support complex querying on dynamic data” [4]. The group was able to come up with an effective summarization framework that can be used to build summaries of Twitter messages and develop an algorithm to compress Twitter messages with low compression ratio [4].

Other important developments of analyzing Twitter data is not limited to finding algorithms—other researchers use visualization techniques to gather insight of events, such as TweetXplorer [5] or GeoTwitter [6]. Such events include but are not limited to—tracking levels of disease activity in the U.S., presidential elections, and measuring public anxiety related to stock market prices [7].

The focus of this research paper is identifying the region of the United States that is most interested in cats. Cats are said to dominate the internet. As of 2014, there were more than 2 million cat videos posted on Youtube.com with nearly 26 billion total views—an average of 12,000 views for each cat video; that is more views-per-video than any other category of YouTube content. [8]. There are many photos, memes (defined by Merriam-Webster as an idea, behavior or style that spreads from person to person within a culture), and internet-famous cats, such as Grumpy Cat—an internet celebrity so famous that the line to get a photo with her wrapped around a city block—that circulate the internet [9], [10].

Consumption of online cat-related media is heavily understudied, considering the large viewership of feline-related content. Internet users spend a tremendous amount of time engrossed in cat-related media when he or she could be using that time working or studying. On the contrary, it is estimated that up to 70 million cats in the U.S. are strays [11]. It is also estimated that 70-80 million dogs and 74-96 million cats are owned in the United States; approximately 37-47% of all households in the United States have a dog, and 30-37% have a cat [11]. This means that a household is more likely to have two cats, but there are less households that own at least one cat.

This research paper attempts to identify the region of the United States that is most interested in cats by analyzing Twitter data. By discovering the answer, it is possible that certain methods, such as marketing, can be applied to users of that region. It would allow potential or future targeted marketing campaigns to know which region of the country to focus on. It would allow a company to have insight of where to set up their business based on the population of those who are interested in cats and how they should design their products, or what should be bred if there is a particular breed a region is enamored with. Currently, there is very little evidence pertaining to internet cat popularity and the region that is interested in cats.

1. **Methodology**

This research project utilizes RStudio, an open source interface for R, a programming language that is popular among statisticians [12]. Furthermore, specific R packages are required to carry out this project, “twitteR”, “ROAuth,” and “streamR.” All data is collected from the Twitter API. Thus, an API key will be necessary to collect the data from Twitter. These tweets must contain specific words, “Cats,” “cats,” “felines,” or “#cats.” The Twitter screen names are extracted to find the self-proclaimed location of the user. Regions of the US will be broken down into West, Southwest, Midwest, Southeast, and Northeast. Users that left the location blank will be disregarded. In addition, locations that are formatted as such—CITY,STATE—will be observed. Furthermore, some of the STATES are spelled out (Sacramento, California) and thus, will be shortened to their abbreviations. Words to the left of the comma will be ignored. The grepl function is used to check each state abbreviation to an element, returning TRUE or FALSE. The TRUEs are then counted out of the sample size. The United States Department of Labor is used to determine which states belong to the region. The sum function is used to count the number of TRUEs for each state respective to its region. Then, the count for each state of that region is added. Let *c* be represented as the number of tweets per individual state, *x* as the number of states in a region, *r* the total tweets per region, and *n* be the sample size across all regions.

**Figure 1.** Equation used to find the total tweets per region

The is the proportion of an individual state’s cat-related tweets over the total sample size of cat related tweets. To find a region’s proportion, we find the summation of for *x* amount of states that belong to *r* (region). The proportion represents the regional proportion in the United States.

The R package ggplot2 will be utilized for a clean visualization of the results after finding the proportions, specifically, to display bar charts of the proportions of cat-related tweets.

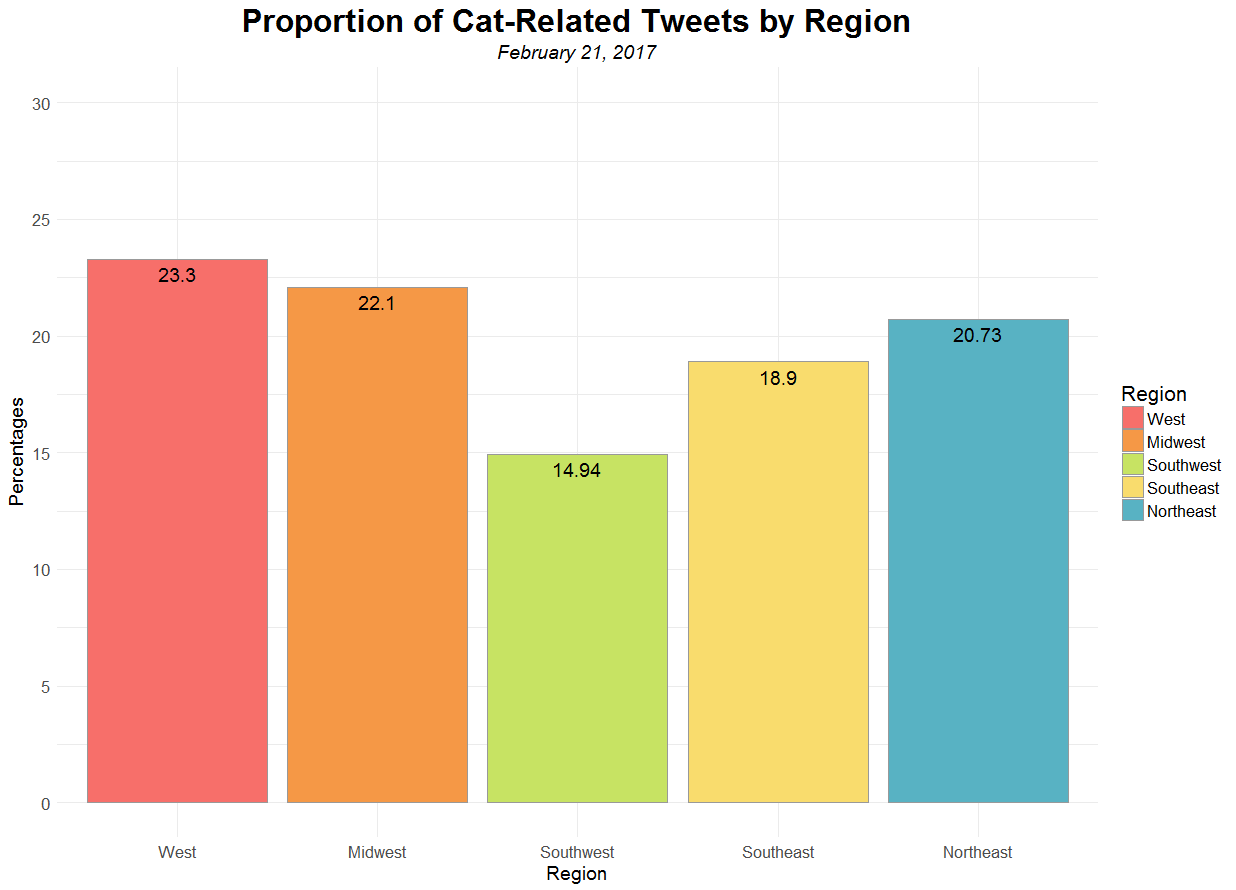
The ggmap R library will be used to graph each individual tweet to its respective state. The geocode function is used to find the longitude and latitude coordinates for 48 states—Hawaii and Alaska are excluded from this study—and assigned to the state in a CSV file. In addition, for each state, the count of tweets per state for February 21, 2017 and February 23, 2017 are also included as separate columns. The CSV file is later imported into RStudio and used to map tweets to the US map as another method for visualization.

In addition, Pearson’s chi-squared test is used to analyze if the distribution of tweets across geographic regions is associated with the day.

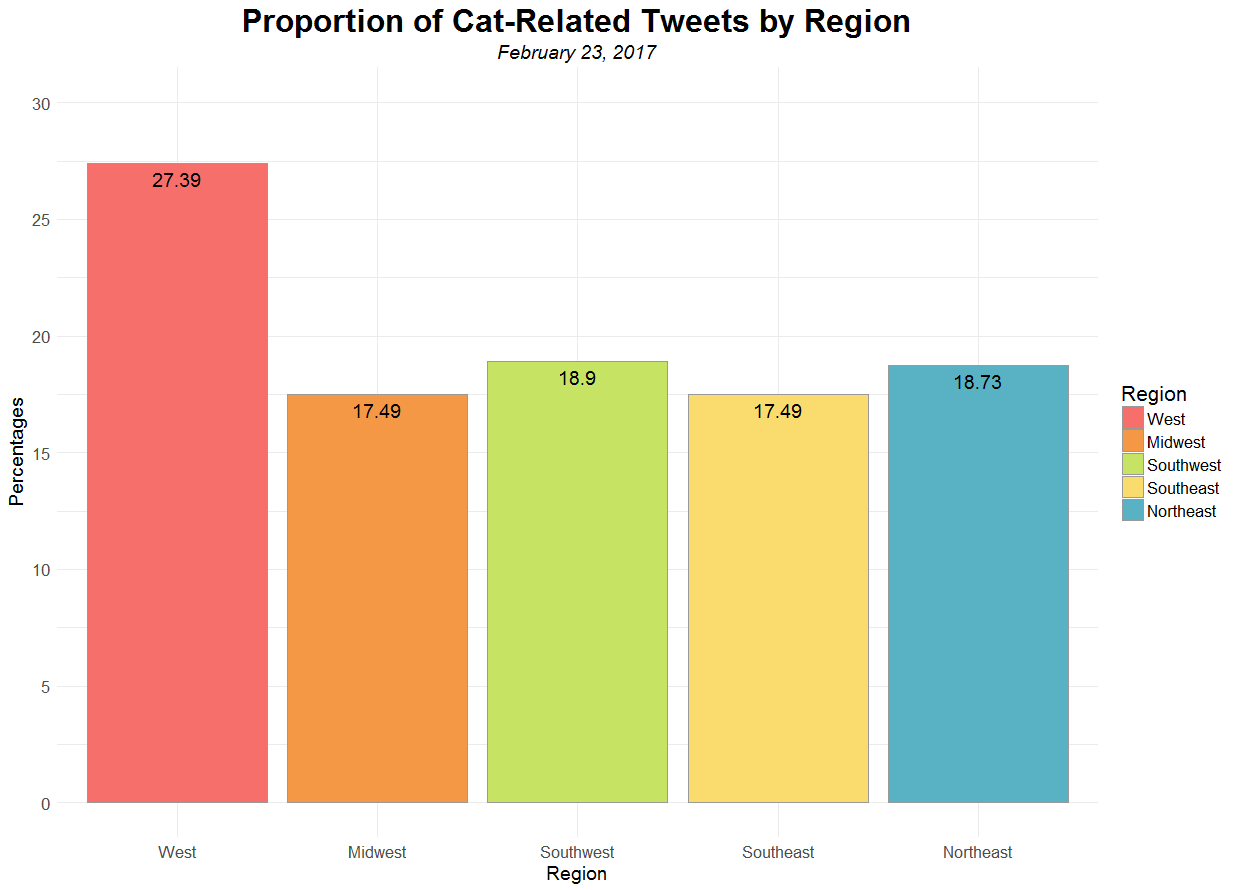
The files and code developed for this project are uploaded to Github using Git Bash. They may be found at: <https://github.com/ktlawlor/Senior-Research>.

1. **Results**

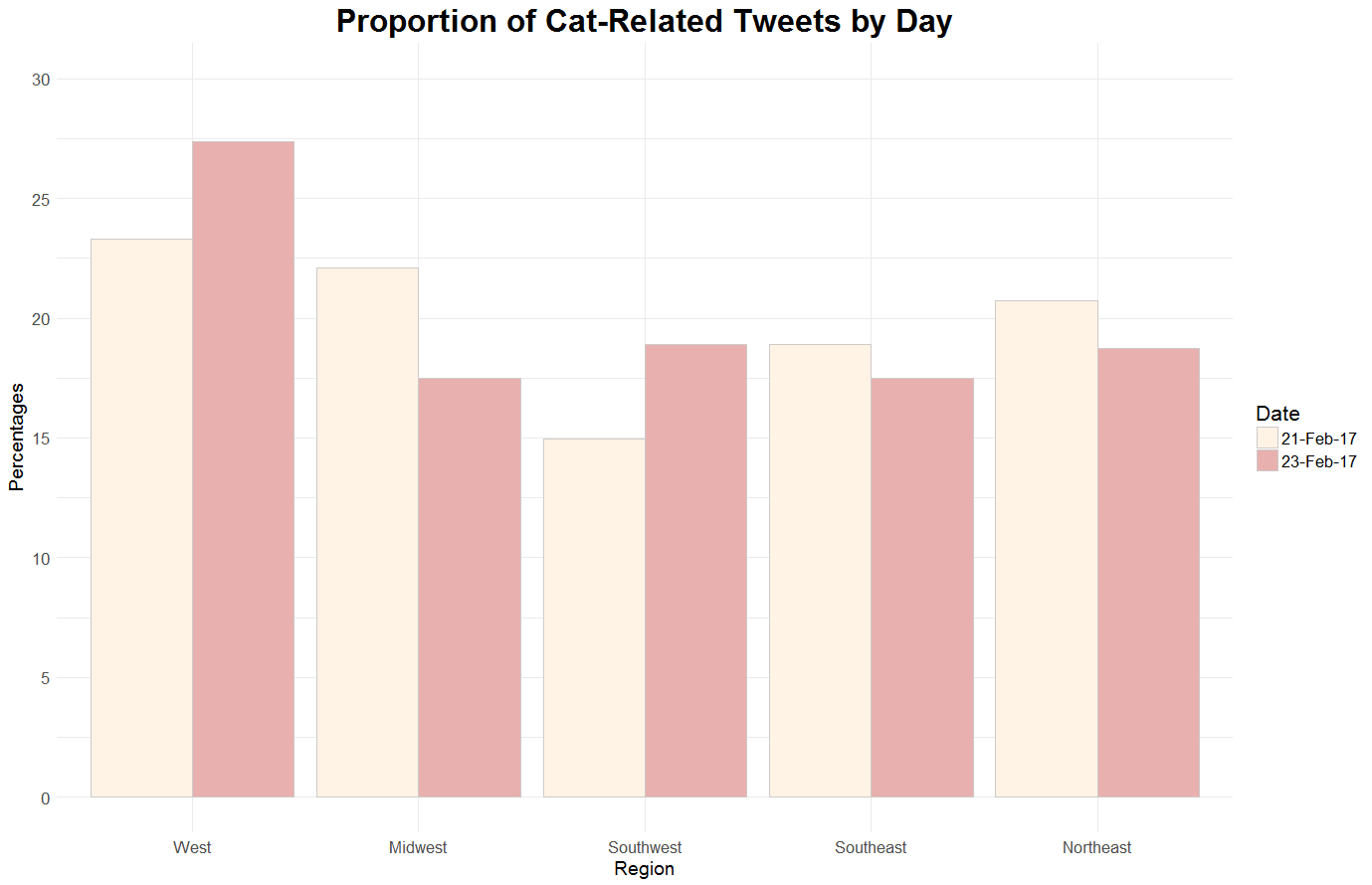
I collected five thousand tweets from February 21, 2017 and five thousand tweets from February 23, 2017. These two sets contained at least one of the following terms: “Cats,” “cats,” “felines,” or “#cats.” After narrowing down the number of tweets by location—taking the second element of the vector, in other words, the term after the comma—I found for the first set only 1,298 tweets were relevant to the study. For day two, 1,056 tweets out of 5,000 could be considered. After analyzing which tweets were from the United States, I found that 0.13% tweets were relevant to the study for the first day and 0.11% tweets for the second day, thus, a total of 656 tweets for February 21, 2017 and 567 for February 23, 2017.



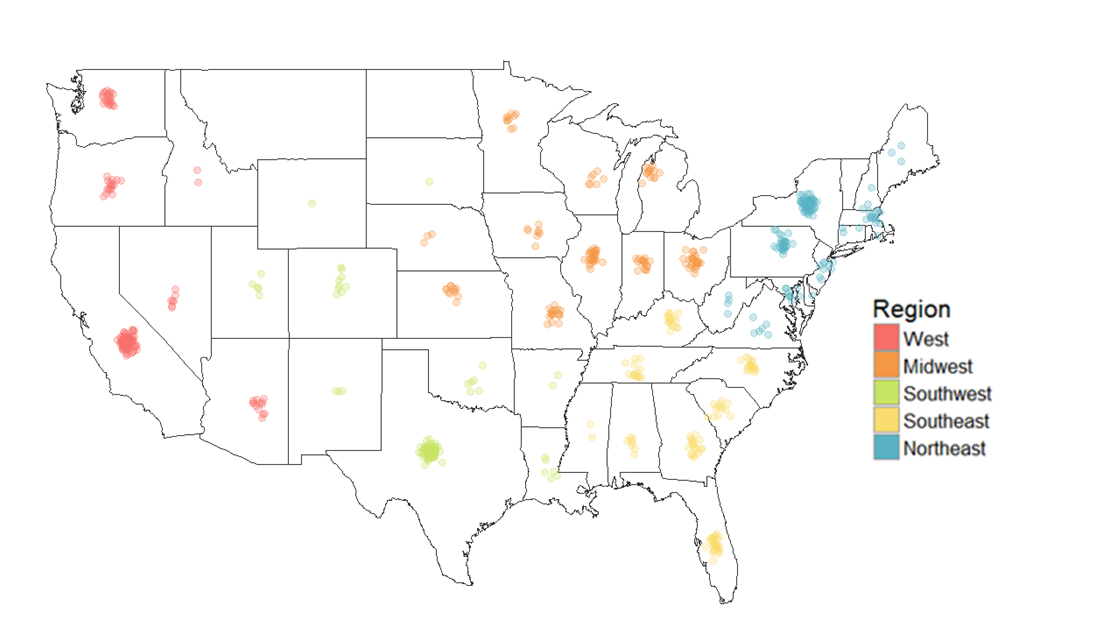
**Figure 2**: *Bar chart of the proportions of cat-related tweets by region on 2-21-2017. Total of 656 tweets.*



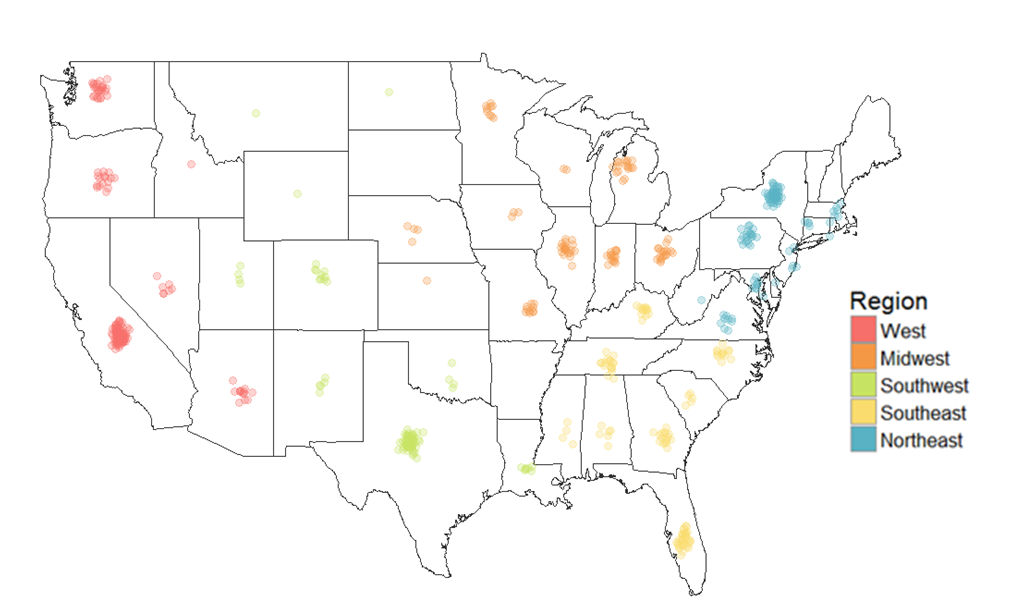
**Figure 3**: *Bar chart of the proportions of cat-related tweets by region on 2-23-2017. Total of 567 tweets.*



**Figure 4**: *Percentages of cat-related tweets across region by day.*

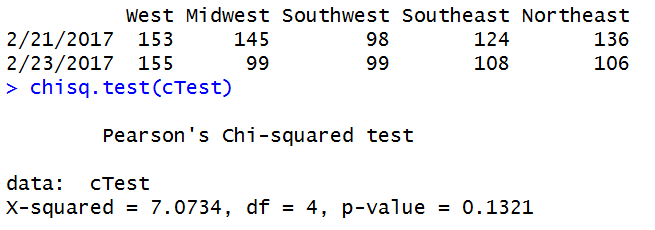


**Figure 5**: *Tweets per state on February 21, 2017.* *Each point represents one tweet.*



**Figure 6**: *Tweets per state on February 23, 2017. Each point represents one tweet.*

As seen in **Figure 2**, the region with the highest proportion of cat-related tweets is the West, with a percentage of 23.3%. In **Figure 3**, the West is still the region with the highest proportion of cat-related tweets—27.39%. There is a 4.09% difference in the percentage of cat-related tweets for the West between February 21, 2017 and February 23, 2017. This variance can be observed in **Figure 4**. In **Figure 5** and **Figure 6**, I analyzed cat-related tweets separately by state. It is evident that the states with the highest count of cat tweets are California, Texas, and New York.



**Figure 7**: *Chi-squared test to test for independence between region and day*.

From **Figure 4**, it is evident that the proportions for each region are similar between days. In order to test for independence, a chi-squared test—specifically Pearson’s chi-squared test—was performed to demonstrate that the probability distribution of region is not affected by day. The alpha value used to run the test was 0.05. The hypotheses formulated are as follows:

H0: The distribution of tweets across geographic regions is not associated with the day.

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

**Figure 7** displays a p-value of 0.131, which is greater than 0.05. Thus, I fail to reject the null hypothesis. Therefore, there is not enough statistically significant evidence to suggest that the distribution of cat-related tweets across geographic regions is associated with the day.

1. **Discussion**

I expected to find that the northeast will have the most cat-related tweets. It is possible that the region that has cats as the more common household pet is more likely to mention cats in their tweets.

The American Veterinary Medical Association (AVMA) conducts a breakdown of pet ownership by state every five years. In 2011, the two top states with the most cat-owning households were Vermont and Maine, 49.5 percent and 46.4 percent of households possessing cats respectively [13].

According to the results of this study, Maine had very few cat-related tweets. This result contradicts the fact that Maine is one of the top states in the northeast that is most interested in cats. In fact, New York had the highest count of cat-related tweets on both days. New York has a higher population density than Maine [14] which could have an effect on the number of cat-related tweets per state and furthermore, region.

However, I also anticipated the region to have the most cat-related tweets to be the west. Washington is also a state that is highly populated with cat owners—62.7 percent [13]. Internet famous cats like Colonel Meow or Cooper the “Cat-tographer” originate from Washington. However, the results indicate that California has the highest count and proportion of cat-related tweets in the western region. Again, the population is much higher in California than Washington—37, 253, 956 to 6,724,540. It is difficult to predict the outcome of the results—even though cats are the more favored household pet in a state doesn’t signify it will be the most popular in a region—one or two states are not representative of the Twitter demographic, let alone the total population of a region of the U.S. As stated previously, the population of a state or region could sway the proportion of cat-related tweets in its favor. It is possible that the region that has a higher population would appear to be the most interested in cats, simply because would be more people to tweet about cats.

There is not much knowledge or studies done on the consumption on cat-related media. This research project collects cat-related Tweets from two specific dates from the Twitter API. Other uses of Twitter include tracking levels of disease activity in the U.S., presidential elections, and measuring public anxiety related to stock market prices [7]. Further research in cat-related studies include sentiment analysis for cat-related tweets—can they be classified as positive, negative, and neutral posts? If so, which is more common?

One of the limitations of this project is samples were taken from two days. It possible that an event could have happened those days that may have fluctuated the number of cat-related tweets or decreased it. In addition, the demographic of Twitter users do not represent the general population. Not everyone who uses Twitter is going to post about cats in their tweets or even use their Twitter account to post statuses at all. One of Twitter’s functions, “retweeting,” was not considered while collecting data. Those who like cats can simply “retweet” a cat-related post that they found on someone’s profile. Furthermore, when collecting data from the Twitter API, some of the tweets did not have any of the pre-defined terms within them. Instead, one of the terms was in the Twitter screen name. The tweets were getting included in the sample because at least in one of the fields, one of the pre-defined terms, i.e. “cat,” was mentioned. Thus, there is a confounding variable that could influence which region of the U.S. prefers cats. In addition, Internet users frequently post images of their own felines on social media platforms, further increasing the amount of online cat-related visual content available to Internet users [8]. Only text-based tweets were collected, not pictures, so this will also influence the results.

In addition, the original sample size of cat-related tweets was cut significantly because of improper formatting of locations or the locations themselves. Had Geocoding been utilized, it would have been easier for Google’s Geocoding API to turn the addresses from text to latitude and longitude pairs, therefore, plotted on a map [15].

**Bibliography**

[1] “Company | About,” *Twitter About*. [Online]. Available: https://about.twitter.com/company. [Accessed: 22-Feb-2017].

[2] A. Java, X. Song, T. Finin, and B. Tseng, “Why We Twitter: Understanding Microblogging Usage and Communities,” in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, New York, NY, USA, 2007, pp. 56–65.

[3] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a Social Network or a News Media?,” in *Proceedings of the 19th International Conference on World Wide Web*, New York, NY, USA, 2010, pp. 591–600.

[4] X. Yang, A. Ghoting, Y. Ruan, and S. Parthasarathy, “A Framework for Summarizing and Analyzing Twitter Feeds,” in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2012, pp. 370–378.

[5] F. Morstatter, S. Kumar, H. Liu, and R. Maciejewski, “Understanding Twitter Data with TweetXplorer,” in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2013, pp. 1482–1485.

[6] A. M. MacEachren *et al.*, “SensePlace2: GeoTwitter analytics support for situational awareness,” in *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2011, pp. 181–190.

[7] A. Signorini, A. M. Segre, and P. M. Polgreen, “The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic,” *PLOS ONE*, vol. 6, no. 5, p. e19467, May 2011.

[8] “Emotion regulation, procrastination, and watching cat videos online: Who watches Internet cats, why, and to what effect? (PDF Download Available),” *ResearchGate*. [Online]. Available: https://www.researchgate.net/publication/278161046\_Emotion\_regulation\_procrastination\_and\_watching\_cat\_videos\_online\_Who\_watches\_Internet\_cats\_why\_and\_to\_what\_effect. [Accessed: 02-Mar-2017].

[9] O. B. Waxman and N. Carbone, “Grumpy Cat Is Not Impressed by TIME’s Photo Shoot,” *Time*.

[10] D. Thompson, “The Wealth of Grumpy Cat,” *The Atlantic*, 08-Dec-2014.

[11] “Pet Statistics,” *ASPCA*. [Online]. Available: http://www.aspca.org/animal-homelessness/shelter-intake-and-surrender/pet-statistics. [Accessed: 02-Mar-2017].

[12] A. Vance, “R, the Software, Finds Fans in Data Analysts,” *The New York Times*, 06-Jan-2009.

[13] “AVMA releases new stats on pet ownership, ranking top/bottom 10 states.” [Online]. Available: https://www.avma.org/News/PressRoom/Pages/TopBotomTenStatesForPets.aspx. [Accessed: 28-Mar-2017].

[14] A. S. D. Website Services & Coordination Staff, “2010 Demographic Profile - U.S. Census Bureau.” [Online]. Available: https://www.census.gov/popfinder/. [Accessed: 07-May-2017].

[15] “Batch Geocoding with R and Google maps,” *R-bloggers*, 12-Oct-2013.