

AI IS599 Course - Spring 2019



**TOYOTA**

# Research Design and Methodology

**Twitter Sentiment Analysis  
of Toyota brand by States in USA**

# Overview

- **Research object:** focus on sentiment and text measures and algorithms that are use to collect and analyze social media.
- **Scope:** Twitter data focused on Toyota auto brand in US states.
- **Design:** Using Twitter API to optimize @ and # toward particular Toyota searching.
- **Tool:** Twitter API, Python with PyCharm EDI (Tweepy, TextBlob, Word Count), Tableau.

# Outline

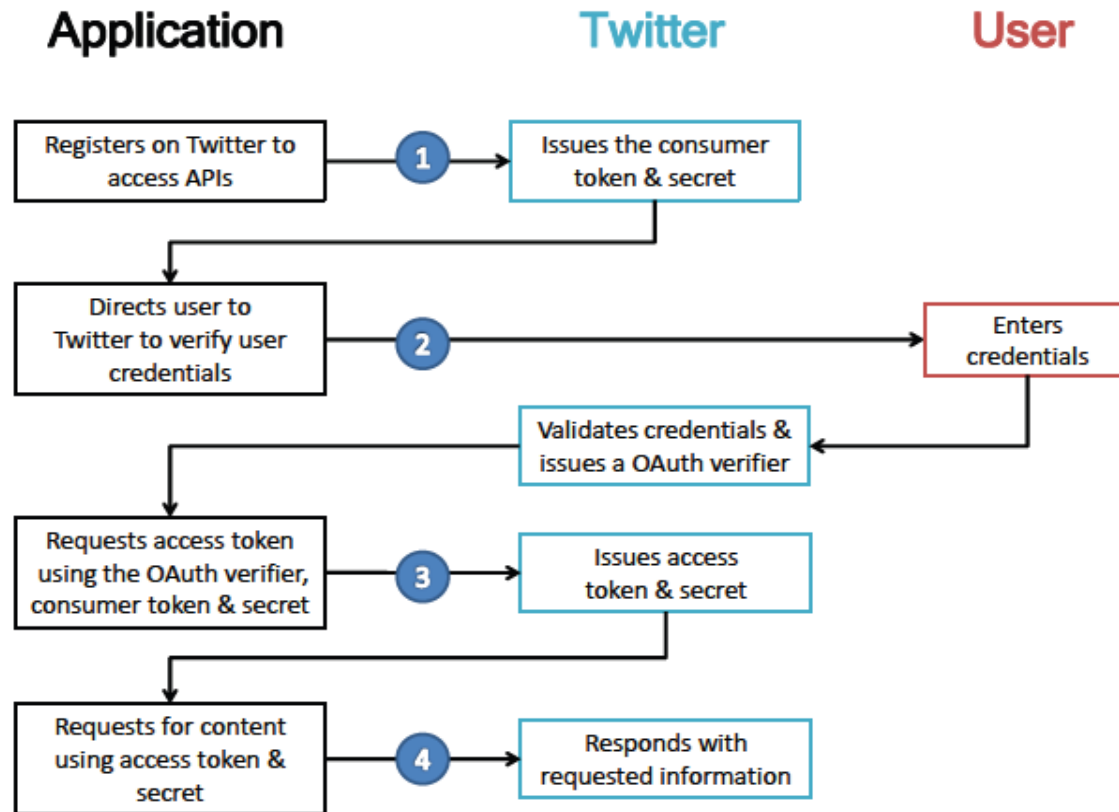
- Why Twitter?
- Crawling Twitter Data
- Collecting Data
- Storing Twitter Data
- Analyzing Twitter Data
  - Text Measure
  - Sentiment Analysis
- Visualizing Data

# Why Twitter?

- Twitter is a massive social networking site tuned towards fast communication.
- More than 140 million active user publish over 400 million everyday.
- Its speed and ease of publication have made it an important communication medium for people.
- Its hashtag as useful tool to collect related data.
- It has played a prominent role in socio-political events.

# Crawling Twitter Data: OAuth

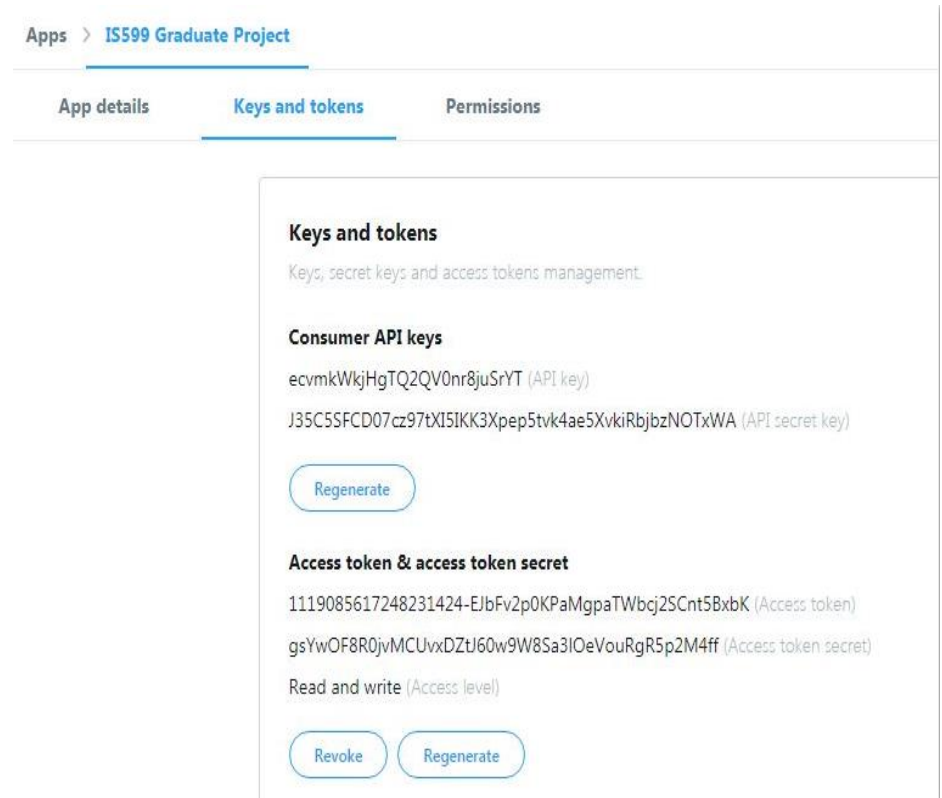
- Open Authentication (OAuth) process chart



# Crawling Twitter Data: Keys and tokens

In Key and Tokens on app, there are 4 values to access Twitter from Python code:

- API key
- API secret key
- Access token
- Access token secret



The screenshot shows the 'Keys and tokens' management page for the 'IS599 Graduate Project' app. The page has three tabs: 'App details', 'Keys and tokens' (which is selected), and 'Permissions'. The 'Keys and tokens' section includes a description: 'Keys, secret keys and access tokens management.' It displays the 'Consumer API keys' with the 'API key' (ecvmkWkJHgTQ2QV0nr8juSrYT) and the 'API secret key' (J35C5SFC07cz97tXI5IKK3Xpep5tvk4ae5XvkiRbjzNOTxWA). There is a 'Regenerate' button for the consumer keys. Below this, the 'Access token & access token secret' section shows the 'Access token' (1119085617248231424-EJbFv2p0KPaMgpaTWbcj2SCnt5BxbK) and the 'Access token secret' (gsYwOF8R0jvMCUvxDZtU60w9W8Sa3IOeVouRgR5p2M4ff). The 'Access level' is listed as 'Read and write'. There are 'Revoke' and 'Regenerate' buttons at the bottom of the access token section.

Apps > IS599 Graduate Project

App details Keys and tokens Permissions

**Keys and tokens**  
Keys, secret keys and access tokens management.

**Consumer API keys**  
ecvmkWkJHgTQ2QV0nr8juSrYT (API key)  
J35C5SFC07cz97tXI5IKK3Xpep5tvk4ae5XvkiRbjzNOTxWA (API secret key)  
Regenerate

**Access token & access token secret**  
1119085617248231424-EJbFv2p0KPaMgpaTWbcj2SCnt5BxbK (Access token)  
gsYwOF8R0jvMCUvxDZtU60w9W8Sa3IOeVouRgR5p2M4ff (Access token secret)  
Read and write (Access level)  
Revoke Regenerate

# Crawling Twitter Data

- Install Python
  - Integrated Development Environment (IDE): PyCharm
- Install Tweepy library and create the Twitter Developer Application to get authentication from Twitter
- Tweepy is tool to collect data from Twitter, it enables Python to communicate with the Twitter platform

```
6 consumer_key = 'ecvmkWkjHgTQ2QV0nr8juSrYT'
7 consumer_secret = 'J35C5SFCD07cz97tXI5IKK3Xpep5tvk4ae5XvkiRbjbzNOTxWA'
8
9 access_token = '1119085617248231424-EJbFv2p0KPaMgpaTWbcj2SCnt5BxbK'
10 access_token_secret = 'gsYwOF8R0jvMCUvxDZtJ60w9W8Sa3lOeVouRgR5p2M4ff'
11
12 auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
13 auth.set_access_token(access_token, access_token_secret)
```

# Collecting Data: User Information

- Unique tweet ID, Date, User Name, Location, number of retweet, number of favorite, Tweet.
- Key parameter: Location, Tweet.

```
30 with open(file_name, mode='w', newline='', encoding='UTF-8') as my_file:
31     my_writer = csv.writer(my_file, delimiter=',')
32     my_writer.writerow(['Id', 'User_Name', 'Date', 'Location', 'Retweet', 'Likes', 'Text'])
33     # 'User_ID', 'Date', 'Tweet_Id'
34     for i in fetch_tweets:
35         current_ID = str(i.id)
36         current_Txt = str(i.text)
37         current_Date = str(i.created_at)
38         current_ScreenName = str(i.user.screen_name)
39         current_Location = str(i.user.location)
40         current_Retweet = str(i.retweet_count)
41         current_Likes = str(i.favorite_count)
```



# Collecting Data: Data Format

```

34     for i in fetch_tweets:
35         current_ID = str(i.id)
36         current_Txt = str(i.text)
37         current_Date = str(i.created_at)
38         current_ScreenName = str(i.user.screen_name)
39         current_Location = str(i.user.location)
40         current_Retweet = str(i.retweet_count)
41         current_Likes = str(i.favorite_count)
42         my_writer.writerow([current_ID, current_ScreenName, current_Date,
43                             current_Location, current_Retweet, current_Likes, current_Txt])

```

	A	B	C	D	E	F	
1	Id	User_Name	Date	Location	Retweet	Likes	Text
20	1.13E+18	aimeectait	5/30/2019 22:48	Nashville, TN	1	0	RT @DiscoveryEd: Parents
21	1.13E+18	Rachel_at_TTO	5/30/2019 22:47	Tempe, AZ	0	0	Heading to California for the
22	1.13E+18	QuiggleMasen	5/30/2019 22:42		0	0	@Toyota why does my 200
23	1.13E+18	maxinasti	5/30/2019 22:40	New York, NY	37	0	RT @Toyota: Have fun shif
24	1.13E+18	warhawk86	5/30/2019 22:39	Dallas Texas	88	0	RT @Toyota: This is the wa
25	1.13E+18	amciglobal	5/30/2019 22:30	Global	0	0	Rumor has it the next @Toy
26	1.13E+18	186thSt	5/30/2019 22:28		1	11	The Toyota engineers and t
27	1.13E+18	Toyotaoftacoma	5/30/2019 22:24	Tacoma WA	0	2	#TBT 1990 @Toyota Crow
28	1.13E+18	MawahidAbbas	5/30/2019 22:24		0	4	@Toyota it would be even t
29	1.13E+18	DiscoveryEd	5/30/2019 22:21	Global	1	2	Parents, don't be afraid to
30	1.13E+18	twitsandbox1	5/30/2019 22:19		3	0	RT @VisitFortWorth: Goat
31	1.13E+18	TOMeHEGER	5/30/2019 22:10	your "impossibil	1	1	@ktumulty I would say this
32	1.13E+18	kannaka4748817	5/30/2019 22:10	Memphis, TN	88	0	RT @Toyota: This is the wa

# Collecting Data: User's Tweets

- The corpus contains one car brand, Toyota : '#Toyota', '@Toyota'
- Tweets are filtered in English
- Region: US states
- Date: May 10 – Jun 3

```
15 file_name = 'Toyota.csv'
16 search_words = "@Toyota"
17 region = '39.8,-95.583068847656,2500km'
18 number records = 2
19 until_date = '2019-06-02' #YYYY-MM-DD
20 since_date = '2019-05-31' #YYYY-MM-DD
21 # 48 states of USA '39.8,-95.583068847656,2500km'
22
23 api = tweepy.API(auth)
24 fetch_tweets = tweepy.Cursor(api.search, q=search_words, lang='en',
25                               geo = region, until = until_date, since = since_date).items()
```

# Storing Twitter Data

- Saving to CSV file
- use the “with open()” with “mode=‘w’” to make a new file.
- use the “writerow()” function to write each row.

```
30 with open(file_name, mode='w', newline='', encoding='UTF-8') as my_file:
31     my_writer = csv.writer(my_file, delimiter=',')
32     my_writer.writerow(['Id', 'User_Name', 'Date', 'Location', 'Retweet', 'Likes', 'Text'])
33     # 'User_ID', 'Date', 'Tweet_Id'
34     for i in fetch_tweets:
35         current_ID = str(i.id)
36         current_Txt = str(i.text)
37         current_Date = str(i.created_at)
38         current_ScreenName = str(i.user.screen_name)
39         current_Location = str(i.user.location)
40         current_Retweet = str(i.retweet_count)
41         current_Likes = str(i.favorite_count)
```

```
54 with open(file_name, mode='r', encoding='UTF-8') as my_file:
55     csv_reader = csv.reader(my_file, delimiter=',')
56     line_count = 0
57     for row in csv_reader:
58         if line_count == 0:
59             line_count = line_count + 1
60         else:
61             txt = row[1]
62             print(txt)
63             print("-----")
```

# Sampling

- Collect data from Twitter using the twitter API.
- Search keywords: “@Toyota” “#Toyota”
- 52 states of US
- Date: May 10 – Jun 3

	Number of records
Data collected from API	16,016
Data selected by USA:	6,193
Data by detailed state	5,736

# Analyzing Twitter Data: Text Measures

- Text Measure:
  - Finding topics and trends in Tweet by counting words.
  - Cleansing data from unnecessary characters such as retweets, hashtags, punctuations, html links and special symbols.

```
20 final_txt = txt.lower()
21 stop_words = set(stopwords.words('english'))
22 tokenizer = RegexpTokenizer(r'\w+|\$[\d\.]+\S+')
23 word_tokens = tokenizer.tokenize(final_txt)
24 filtered_sentence = [w for w in word_tokens if not w in stop_words]
25 for w in filtered_sentence:
26     if w not in final_dictionary:
27         final_dictionary[w] = 1
28     else:
29         final_dictionary[w] = final_dictionary[w] + 1
30
31 line_count = line_count + 1
32
33 sorted_d = sorted(final_dictionary.items(), key=operator.itemgetter(1), reverse=True)
```

```
1 import csv
2 import nltk
3 from nltk.tokenize import word_tokenize
4 from nltk.tokenize import RegexpTokenizer
5 from nltk.corpus import stopwords
6 from collections import OrderedDict
7 import operator
```

# Analyzing Twitter Data: Text Measures

Example of data after counting word

	A	B	C
1	Word <input type="text"/>	Count <input type="text"/>	Word <input type="text"/>
2	new	554	new
3	@hyundai	254	@hyundai
4	2020	252	2020
5	rav4	235	rav4
6	@honda	227	@honda
7	supra	219	supra
8	great	184	great
9	love	180	love
10	@mercedesbenz	175	@mercedesbenz
11	alabama	174	alabama
12	corolla	172	corolla
13	sales	127	sales
14	features	125	features
15	camry	116	camry

# Analyzing Twitter Data: Sentiment Analysis

## What is Sentiment Analysis (SA)?

- Sentiments are feelings, opinions, emotions, likes/dislikes, good/bad.
- SA, know as Opinion Mining is a study of human behavior in which we extract user opinion and emotion from plain text.
- It identifies the opinion or attitude that a person has towards a topic or an object
- The opinion expressed in a text is positive, neutral or negative.



# Sentiment Analysis

- **Need of SA:**

- Rapid growth of available subjective text on the internet
- To make decisions

- **Applications:**

- Organizations: brand analysis, new product perception, etc.
- Individuals: finding behavior towards products or topics
- Social Media: finding general opinion about recent trending topics in particular area.
- Ads Placements: ads in the user-generated content



# Sentiment Analysis

- Install TextBlob which is a Python library for processing textual data. It provides a simple API for diving into common Natural Language processing tasks such as noun phrase extraction, sentiment analysis, classification and more.
- Cleansing data: ignore retweets, remove non-ascii characters, normalize case, remove URLs.

```
75     # Ignore retweets
76     if re.match(r'^RT.*', tweet['orig']):
77         continue
78
79     tweet['clean'] = tweet['orig']
80
81     # Remove all non-ascii characters
82     tweet['clean'] = strip_non_ascii(tweet['clean'])
83
84     # Normalize case
85     tweet['clean'] = tweet['clean'].lower()
86
87     # Remove URLs. (I stole this regex from the internet.)
88     tweet['clean'] = re.sub(r'http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|'
89                             r'|!*\(\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+', '', tweet['clean'])
90
```

# Sentiment Analysis

- Fix classic tweet lingo

```
90 # Fix classic tweet lingo
91 tweet['clean'] = re.sub(r'\bthats\b', 'that is', tweet['clean'])
92 tweet['clean'] = re.sub(r'\bive\b', 'i have', tweet['clean'])
93 tweet['clean'] = re.sub(r'\bim\b', 'i am', tweet['clean'])
94 tweet['clean'] = re.sub(r'\bya\b', 'yeah', tweet['clean'])
95 tweet['clean'] = re.sub(r'\bcant\b', 'can not', tweet['clean'])
96 tweet['clean'] = re.sub(r'\bwont\b', 'will not', tweet['clean'])
97 tweet['clean'] = re.sub(r'\bid\b', 'i would', tweet['clean'])
98 tweet['clean'] = re.sub(r'wtf', 'what the fuck', tweet['clean'])
99 tweet['clean'] = re.sub(r'\bwth\b', 'what the hell', tweet['clean'])
100 tweet['clean'] = re.sub(r'\br\b', 'are', tweet['clean'])
101 tweet['clean'] = re.sub(r'\bu\b', 'you', tweet['clean'])
102 tweet['clean'] = re.sub(r'\bk\b', 'OK', tweet['clean'])
103 tweet['clean'] = re.sub(r'\bsux\b', 'sucks', tweet['clean'])
104 tweet['clean'] = re.sub(r'\bno+\b', 'no', tweet['clean'])
105 tweet['clean'] = re.sub(r'\bcoo+\b', 'cool', tweet['clean'])
106 #tweet['clean'] = re.sub('@toyota', tweet['clean'])
```

# Sentiment Analysis

- If input a string to TextBlob, the polarity will range from -1.0 to 1.0
  - $\geq 0.1$  means the text has a positive sentiment
  - $\leq -0.1$  means the text has a negative sentiment

```
132 for tweet in tweets:
133     tweet['polarity'] = float(tweet['TextBlob'].sentiment.polarity)
134     tweet['subjectivity'] = float(tweet['TextBlob'].sentiment.subjectivity)
135
136     if tweet['polarity'] >= 0.1:
137         tweet['sentiment'] = 'positive'
138     elif tweet['polarity'] <= -0.1:
139         tweet['sentiment'] = 'negative'
140     else:
141         tweet['sentiment'] = 'neutral'
142
143 tweets_sorted = sorted(tweets, key=lambda k: k['polarity'])
```

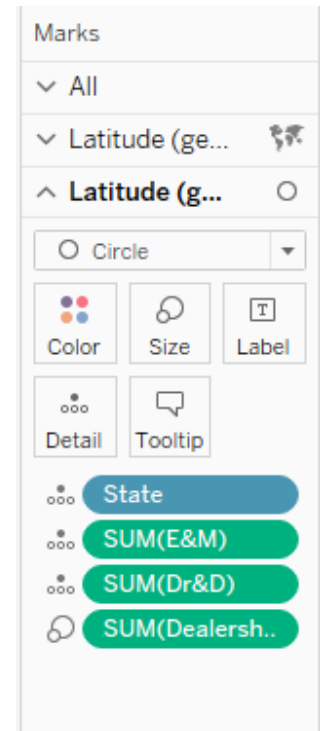
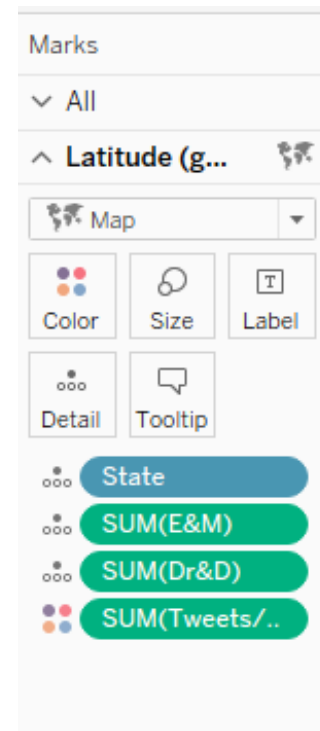
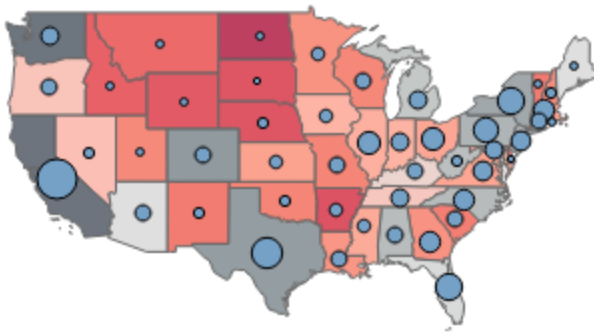
# Sentiment Analysis

Example of data after analyzing sentiment: polarity column, sentiment column and cleansing text

	B	G	H	I
1	Id	Popularity	Sentiment	Cleansing
20	1.12E+18	0.14	Positive	the celebration of our fans continues at @toyota bassmaster texas fest benefitting @tpwdpa
21	1.12E+18	0.00	Neutral	bassmaster elite series pro @brian_snowden joins tommy sanders to preview the 2019 @toy
22	1.13E+18	0.27	Positive	enter daily through september 29th for a chance to win a 2019 @toyota tacoma trd off-road
23	1.13E+18	0.50	Positive	fun night cheering on @nationals thanks to @thechildrensinn and @toyota for great seats!
24	1.13E+18	-0.05	Neutral	@toyota if i purchase a pre-owned 2018 camry (private party)
25	1.13E+18	-0.30	Negative	i think its crazy that @cityofbhamal (and jefferson county
26	1.13E+18	0.39	Positive	@mazdausa @mazdausa @toyota are supporting alabamas abortion laws by building a faci
33	1.13E+18	0.00	Neutral	tomorrow on espn2! @toyota bassmaster texas fest at lake fork benefitting @tpwdfish airs at
34	1.13E+18	0.12	Positive	bass fishing tip: gerald swindle says you need only these three crankbait colorsto catch more
35	1.13E+18	0.27	Positive	enter daily through september 29th for a chance to win a 2019 @toyota tacoma trd off-road
36	1.13E+18	0.00	Neutral	@martintruex_jr @toyota @bassproshops i wish you could spare a ticket and garage pass fc
37	1.13E+18	0.00	Neutral	@eileenmdhrita @al_labor @alsenaterepubs @adolsecretary @govemorkayivey @alwork
38	1.12E+18	-0.50	Negative	@toyota ill already be paying for a 14th tire by the time that i get offer. what a joke.
39	1.12E+18	0.00	Neutral	@toyota why should i buy a tacoma over a ridgeline? the ridgeline won truck of the year?
40	1.12E+18	0.50	Positive	@toyota @mrpeanut @nutmobile_tour @fredricaasbo @toyota: let me copy your homework
41	1.12E+18	0.70	Positive	@toyota the charity is a good vehicle? why by a prius when that is an option?
42	1.12E+18	0.17	Positive	@toyota that is cool
43	1.12E+18	0.50	Positive	@toyota when i think highlander i see more ground clearance and a more rugged look this v
44	1.12E+18	0.06	Neutral	@good4beans @gymrat10 @toyota most people i know paid less taxes this year at an aver
45	1.12E+18	0.00	Neutral	@good4beans @gymrat10 @toyota definitely not a trump fan and why is it that you automa

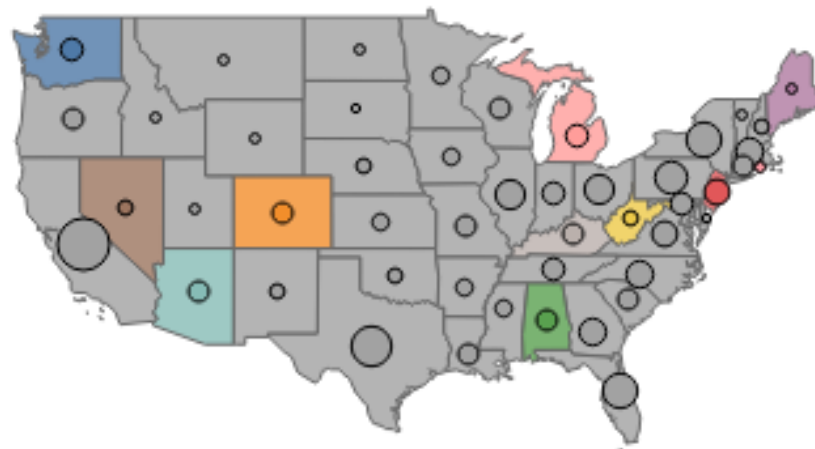
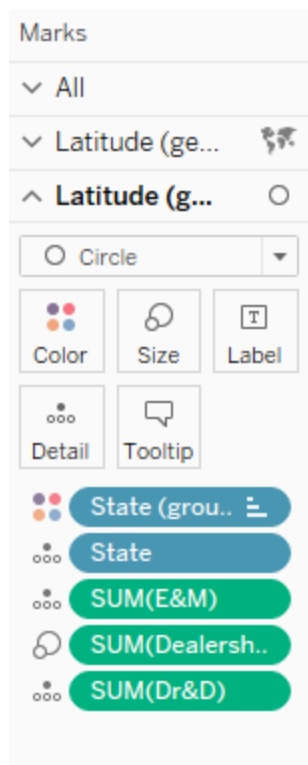
# Visualizing Data: Geo-Spatial Information

- Use Tableau to create Geo-Spatial Heatmaps
- On Tableau platform, import Excel file in Data Source
- Apply 2 layers: one with US states map, and one with circle
- Add more information in detail box



# Visualizing Data: Geo-Spatial Information







- To visualize specific regions, group function is very useful to hide unnecessary areas and highlight important ones with automatic colors



# Visualizing Data: Textual Information

- Wordart.com is useful and friendly-use webpage to create word cloud maps
- Import collect csv file
- Optional shapes, fonts, layout and styles to customize the map.



WORDS				
<div>  Import            Add            Remove            Up            Down            Options         </div>				
Filter	Size	Color	Angle	Font
toyota	158	Default	Default	Default ▼
car	19	Default	Default	Default ▼
new	15	Default	Default	Default ▼

**THANK YOU!**