

Introduction

In this project, exploratory data analysis (EDA) was conducted on large, structured and semi-structured datasets. With the help of EDA, important analysis about data and its relation with various parameters was studied. The different types of datasets analyses were twitter data set, New York Times ad click dataset, Real Direct dataset and live twitter data.

Problem 1: Data Acquisition

For this problem, the data source used was Twitter data using Twitter Search API, collected over a period of ten days. Twitter Streaming API is used in Problem 5. TwitteR and ROAuth were the packages that were used to carry out twitter data collection. TwitteR provided functions like `setup_twitter_oauth` which creates a connection to Twitter's Search API. The resulting list of retrieved tweets is stored in a data frame by using the function `twListToDF()`. Over a period of ten days, I created different data frames that stored that day's tweets. At the end, I combined all these data frames into one unit and stored this collective data as both cvs file and Rdata by using `write.csv` and `save` commands respectively. `tf` is the local data frame containing collected tweets. These are the column variables it has.

```
> colnames(tf)
[1] "text"           "favorited"      "favoriteCount"  "replyToSN"      "created"         "truncated"      "replyToSID"
[8] "id"            "replyToUID"     "statusSource"   "screenName"     "retweetCount"    "isRetweet"      "retweeted"
[15] "longitude"      "latitude"
> |
```

text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID	id	replyToUID	statusSource
1 RT @AndrewGuapo_: "Young Metro doesn't trust Don...	FALSE	0	NA	2016-02-25 23:59:59	FALSE	NA	703006559421513728	NA	<a href="
2 RT @steph93065: Trump doesn't need to change a thi...	FALSE	0	NA	2016-02-25 23:59:59	FALSE	NA	703006559329230849	NA	<a href="
3 @Thidran @Enjenovix @duckspeakeasy @IndJournal O...	FALSE	0	Thidran	2016-02-25 23:59:59	FALSE	703005778106568704	703006559169847296	176874550	<a href="
4 Debate another chance for Rubio, Cruz to try to slo...	FALSE	1	NA	2016-02-25 23:59:59	FALSE	NA	703006558309842944	NA	<a href="
5 RT @GregJaye: Hey all #SECprimary voters: Think abo...	FALSE	0	NA	2016-02-25 23:59:59	FALSE	NA	703006558083362816	NA	<a href="
6 RT @ReneSaldanajr: All of a sudden @realDonaldTrump...	FALSE	0	NA	2016-02-25 23:59:59	FALSE	NA	70300655752180737	NA	<a href="
7 Bernie's ego rivals Trump's. No doubt about it! #imWi...	FALSE	1	NA	2016-02-25 23:59:59	FALSE	NA	703006557458403328	NA	<a href="
8 Yes. When Trump mentions a candidate, the candidat...	FALSE	1	NA	2016-02-25 23:59:59	FALSE	NA	703006557328580608	NA	<a href="
9 Trump is the GOP's #1 Frankenstein monster. ...	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	70300655944255488	NA	<a href="
10 RT @JoeNBC: Trump Dominates South. Most distress...	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006555512430595	NA	<a href="
11 #What Trump, Clinton and the 2016 Election Mean for...	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006554849554434	NA	<a href="
12 RT @CommonWhiteGirl: Donald Trump's cousin drag...	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006554786766848	NA	<a href="
13 @andersoncooper just absolutely destroyed a Trump ...	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006553633239040	NA	<a href="
14 Former Mexican president says he will not pay for Do...	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006553138438144	NA	<a href="
15 "I'm votin' for Donald Trump" are fightin' words.ced	FALSE	1	NA	2016-02-25 23:59:58	FALSE	NA	703006552932745216	NA	<a href="

Table 1									
text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID	id	replyToUID	statusSource
RT @AndrewGuapo_: "Young Metro doesn't trust Donald Trump" @Metroboon #Houston = Sdk, https://t.co/NAqg76HfLuz	FALSE	0	NA	2016-02-25 23:59:59	FALSE	NA	703006559421513728	NA	Twitter for Android
RT @steph93065: Trump doesn't need to change a thing. @Renace The party is the people. NOT party leaders &amp; donors. @jaketapper	FALSE	0	NA	2016-02-25 23:59:59	FALSE	NA	703006559329230849	NA	Twitter for Android
@Thidran @Enjenovix @duckspeakeasy @IndJournal Only a few, but I stand against the Trump Con.	FALSE	0	Thidran	2016-02-25 23:59:59	FALSE	703005778106568704	703006559169847296	176874550	TweetDeck
Debate another chance for Rubio, Cruz to try to slow Trump https://t.co/WE1dU8f #xosai https://t.co/TYvY2wP	FALSE	1	NA	2016-02-25 23:59:59	FALSE	NA	703006558309842944	NA	TweetDeck
RT @GregJaye: Hey all #SECprimary voters: Think about this: Trump is the Forerun	FALSE	0	NA	2016-02-25 23:59:59	FALSE	NA	703006558083362816	NA	Twitter Web Client
RT @ReneSaldanajr: All of a sudden @realDonaldTrump's "Build that Wall" takes on a whole new meaning: https://t.co/hpYghnWZ	FALSE	0	NA	2016-02-25 23:59:59	FALSE	NA	70300655752180737	NA	Twitter Web Client
Bernie's ego rivals Trump's. No doubt about it! #imWinning @ProgressiveParty https://t.co/helULMoc	FALSE	1	NA	2016-02-25 23:59:59	FALSE	NA	703006557328580608	NA	Twitter Web Client
Yes. When Trump mentions a candidate, the candidate's response should be, "What the heck does that mean, exactly?" https://t.co/RNPYOCaw	FALSE	1	NA	2016-02-25 23:59:59	FALSE	NA	703006554849554434	NA	Twitter Web Client
Trump is the GOP's #1 Frankenstein monster. Now he's #1 strong enough to destroy the party. https://t.co/leic8t1Q	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	70300655944255488	NA	Twitter Web Client
RT @JoeNBC: Trump Dominates South. Most distressing for conservatives is fact voters know he's most liberal &amp; support him anyway https://t.co/yf	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006555512430595	NA	Twitter for iPhone
#What Trump, Clinton and the 2016 Election Mean for! Stocks Hit a Record High https://t.co/gwHJH6SE	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006554786766848	NA	WordPress
RT @CommonWhiteGirl: Donald Trump's cousin dragged him through the mud in his obituary https://t.co/0fryhMMW	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006553633239040	NA	Twitter for iPhone
@andersoncooper just absolutely destroyed a Trump Supporter/Republican strategist in Sdkx Sdkx Sdkx Sdkx Sdkx Sdkx have it all Anderson, have it all	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006553138438144	NA	Twitter for iPhone
Former Mexican president says he will not pay for Donald Trump's "telly" war!	FALSE	0	NA	2016-02-25 23:59:58	FALSE	NA	703006553138438144	NA	Twitter for iPhone
"I'm votin' for Donald Trump" are fightin' words! Sdkx Sdkx	FALSE	1	NA	2016-02-25 23:59:58	FALSE	NA	703006552932745216	NA	Twitter for iPhone
RT @PisaPierlin: Bernie supporters Vs. Trump supporters https://t.co/GVND8qMcZ	FALSE	0	NA	2016-02-25 23:59:57	FALSE	NA	70300655234710400	NA	Twitter for Android
RT @Zacharycohen: Hillary can't beat trump. Bernie can. Fast.	FALSE	0	NA	2016-02-25 23:59:57	FALSE	NA	703006552035164180	NA	Twitter for iPhone
@mowes So, remember that convo we had a while back about Trump/Cruz, Cruz/Trump getting the nomination? What was that?	FALSE	0	increws	2016-02-25 23:59:57	FALSE	7030039583820440448	703006550825444160	14489996	Twitter Web Client
@knucklets @McGrahamClegg saw this thought of you... https://t.co/E2chd27y	FALSE	2	knucklets	2016-02-25 23:59:57	FALSE	70284384913637120	70300655036057904	2894973208	Echofon
RT @liandrumham: It would be SUPER cool to have a president who rates women's looks on a scale of 1-10, right guys? https://t.co/0Ks4gKJ8	FALSE	0	NA	2016-02-25 23:59:57	FALSE	NA	70300654992550400	NA	Twitter for iPad
RT @abrynthome: 21 Questions for Donald #Trump https://t.co/eyv97t6a	FALSE	0	NA	2016-02-25 23:59:57	FALSE	NA	703006549745274880	NA	Twitter for iPhone
RT @DefendingHUSA: TODAY'S POLLS VA - TRUMP +14% GA - TRUMP +28%	FALSE	0	NA	2016-02-25 23:59:57	FALSE	NA	70300654942137344	NA	Mobile Web (MSI)

This is the cvs file with tweets, stored on disk.