1. Introduction

2020 US Election happened on November 3rd, which is a very popular topic and has numerous related discussions across various social applications and websites. Therefore, it is worth exploring that the tweets of the two popular candidates, Trump and Biden, near the election time.

There are three topics that are interesting to discover:

(1) Distribution of tweets of these two candidates in different aspects.
(2) Tweets discussions for them in each state vs Final votes for them in each state.
(3) Words and sentiments inside each tweet.

In this report, two datasets are used, and the prepossessing process of datasets, including wrangling, cleaning and checking, and exploration of the topics mentioned above are shown.

2. Data Wrangling

- **Datasets:**

a. US Election 2020 Tweets. Oct 15th 2020 - Nov 8th 2020, 1.72M Tweets (two dataset inside)

b. US Election 2020 Race to Presidential Election 2020 by County

2.1. Loading data into R and briefly look at the data

Dataset a:

```
> summary(data_trump)
  created_at          tweet_id            tweet               likes           retweet_count           source             user_id            user_name        
 Length:971088      Length:971088      Length:971088      Length:971088      Min.   :0.000e+00   Length:971088      Length:971088      Length:971088     
 Class :character   Class :character   Class :character   Class :character   1st Qu.:0.000e+00   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :0.000e+00   Mode  :character   Mode  :character   Mode  :character  
                                                                             Mean   :6.950e+12                                                            
                                                                             3rd Qu.:0.000e+00                                                            
                                                                             Max.   :1.323e+18                                                            
                                                                             NA's   :155                                                                  
 user_screen_name   user_description   user_join_date     user_followers_count user_location          lat                long               city          
 Length:971088      Length:971088      Length:971088      Length:971088        Length:971088      Length:971088      Length:971088      Length:971088     
 Class :character   Class :character   Class :character   Class :character     Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character     Mode  :character   Mode  :character   Mode  :character   Mode  :character  


   country           continent            state            state_code        collected_at      
 Length:971088      Length:971088      Length:971088      Length:971088      Length:971088     
 Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
> summary(data_biden)
  created_at          tweet_id            tweet               likes            retweet_count          source             user_id            user_name        
 Length:777073      Length:777073      Length:777073      Length:777073      Min.   :0.000e+00   Length:777073      Length:777073      Length:777073     
 Class :character   Class :character   Class :character   Class :character   1st Qu.:0.000e+00   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :0.000e+00   Mode  :character   Mode  :character   Mode  :character  
                                                                             Mean   :1.651e+12                                                            
                                                                             3rd Qu.:0.000e+00                                                            
                                                                             Max.   :1.283e+18                                                            
                                                                             NA's   :178                                                                  
 user_screen_name   user_description   user_join_date     user_followers_count user_location          lat                long               city          
 Length:777073      Length:777073      Length:777073      Length:777073        Length:777073      Length:777073      Length:777073      Length:777073     
 Class :character   Class :character   Class :character   Class :character     Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character     Mode  :character   Mode  :character   Mode  :character   Mode  :character  


   country           continent            state            state_code        collected_at      
 Length:777073      Length:777073      Length:777073      Length:777073      Length:777073     
 Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
```

Figure 1. summary about dataset a

Dataset b:

```
> summary(q3.vote)
    state              county            candidate            party            total_votes          won           
 Length:32177       Length:32177       Length:32177       Length:32177       Min.   :      0   Length:32177      
 Class :character   Class :character   Class :character   Class :character   1st Qu.:      3   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :     34   Mode  :character  
                                                                             Mean   :   4960                     
                                                                             3rd Qu.:    745                     
                                                                             Max.   :3028885                     
```

Figure 2. summary about dataset b

2.2. Dataset combination

Dataset a contains two csv file (one for Trump and one for Biden), it needs to be combined together first, the method for that is first adding a new column to record the tweet is about which candidate (using mutate() in R), and then using rbind() to put them together and deleting duplicate rows before transforming. The result can be seen below.

```
> glimpse(candidate_data)
Rows: 1,748,161
Columns: 22
$ created_at           <chr> "2020-10-15 00:00:01", "2020-10-15 00:00:01", "2020-10-15 00:00:02", "2020-10-15 00:00:02", "2020-10-15 00:00:08", "2020-10-15 0~
$ tweet_id             <chr> "1.316529221557252e+18", "1.3165292227484303e+18", "1.316529228091847e+18", "1.316529227471237e+18", "1.3165292523014513e+18", "~
$ tweet                <chr> "#Elecciones2020 | En #Florida: #JoeBiden dice que #DonaldTrump solo se preocupa por Ã©l mismo. El demÃ³crata fue anfitriÃ³n de ~
$ likes                <chr> "0.0", "26.0", "2.0", "0.0", "4.0", "2.0", "0.0", "0.0", "0.0", "0.0", "0.0", "0.0", "3.0", "2.0", "0.0", "3.0", "0.0", "0.0", "~
$ retweet_count        <dbl> 0, 9, 1, 0, 3, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 1, 2, 0, 1, 2, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 6, 6, 2, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,~
$ source               <chr> "TweetDeck", "Social Mediaset ", "Twitter Web App", "Trumpytweeter", "Twitter for iPhone", "Twitter for Android", "Twitter for i~
$ user_id              <chr> "360666534.0", "331617619.0", "8436472.0", "8.28355589206057e+17", "47413798.0", "1138416104.0", "7.674018410302095e+17", "9.007~
$ user_name            <chr> "El Sol Latino News", "Tgcom24", "snarke", "Trumpytweeter", "Rana Abtar - Ø±Ù†Ø§ Ø£Ø¨ØªØ±", "Farris Flagg", "Michael Wilson", "S~
$ user_screen_name     <chr> "elsollatinonews", "MediasetTgcom24", "snarke", "trumpytweeter", "Ranaabtar", "FarrisFlagg", "wilsonfire9", "sm_gulledge", "jami~
$ user_description     <chr> "ðŸŒ\220 Noticias de interÃ©s para latinos de la costa este de #EEUU\nâ \200â\217'ï¸\217 Facebook e Instagram\nâ \200ðŸ\217\231ï~
$ user_join_date       <chr> "2011-08-23 15:33:45", "2011-07-08 13:12:20", "2007-08-26 05:56:11", "2017-02-05 21:32:17", "2009-06-15 19:05:35", "2013-02-01 0~
$ user_followers_count <chr> "1860.0", "1067661.0", "1185.0", "32.0", "5393.0", "2363.0", "75.0", "766.0", "151.0", "8.0", "4622.0", "1396.0", "496.0", "2755~
$ user_location        <chr> "Philadelphia, PA / Miami, FL", "", "Portland", "", "Washington DC", "Perris,California", "Powell, TN", "Ohio, USA", "Pennsylvan~
$ lat                  <chr> "25.77427", "", "45.5202471", "", "38.8949924", "33.7825194", "", "40.225356899999994", "40.9699889", "", "", "41.87556160000000~
$ long                 <chr> "-80.19366", "", "-122.6741949", "", "-77.0365581", "-117.22864779999999", "", "-82.6881395", "-77.72788309999999", "", "", "-87~
$ city                 <chr> "", "", "Portland", "", "Washington", "", "", "", "", "", "", "", "Chicago", "San Diego", "City of Edinburgh", "", "", "City of Edin~
$ country              <chr> "United States of America", "", "United States of America", "", "United States of America", "United States of America", "", "Uni~
$ continent            <chr> "North America", "", "North America", "", "North America", "North America", "", "North America", "North America", "", "", "North~
$ state                <chr> "Florida", "", "Oregon", "", "District of Columbia", "California", "", "Ohio", "Pennsylvania", "", "", "Illinois", "California",~
$ state_code           <chr> "FL", "", "OR", "", "DC", "CA", "", "OH", "PA", "", "", "IL", "CA", "SCT", "", "", "SCT", "", "", "MI", "", "", "OR", "", "", "F~
$ collected_at         <chr> "2020-10-21 00:00:00", "2020-10-21 00:00:00.373216530", "2020-10-21 00:00:00.746433060", "2020-10-21 00:00:01.119649591", "2020-~
$ candidate            <chr> "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "T~
```

```
> dim(candidate_data)
[1] 1748161       22
```

```
> new_candidate_data <- candidate_data[!duplicated(candidate_data),]
> dim(new_candidate_data)
[1] 1748088       22
```

Figure 3. results for combining and duplicate deleting

2.3. Checking for NA values

Dataset a (vis_dat() function in R):

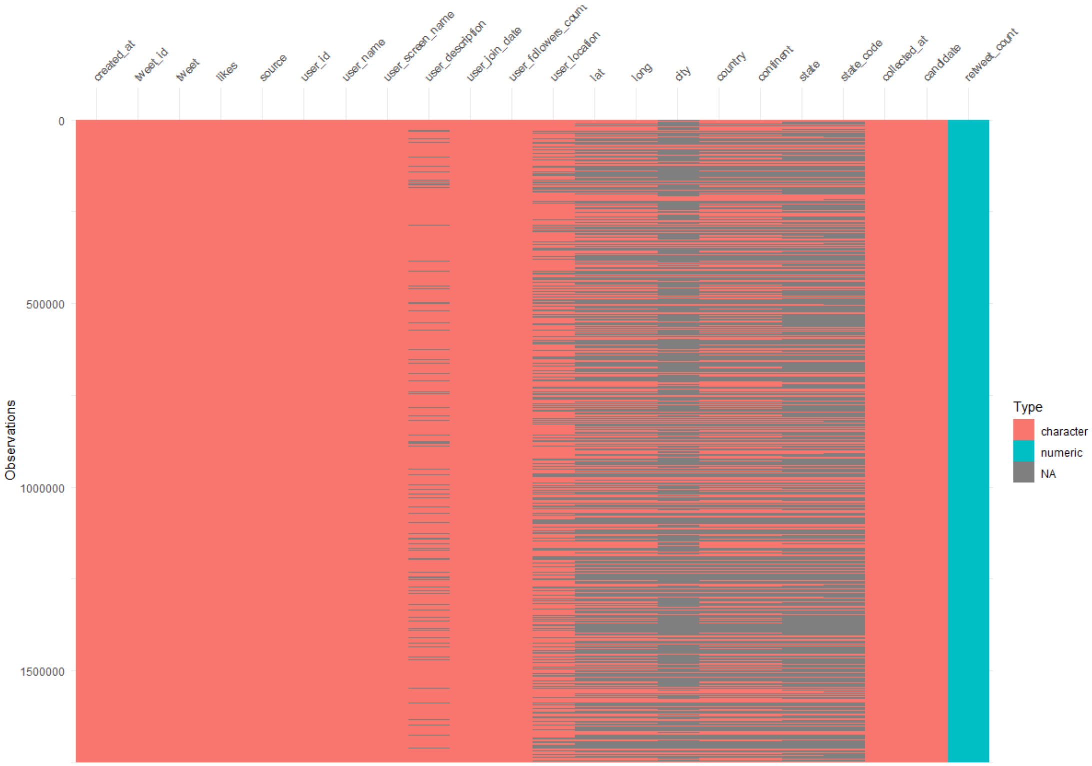

Figure 4. NAs in dataset a

Since some columns are useless, such as user_description, and some columns have around 50% NAs, for example lat and long. It is more suitable to clean NAs in different sub-set for different questions to keep as much as possible data that can be used. In addition, it can be observed that dataset a has only one numeric values, but some columns such as likes, retweet_count should change to number, so dataset a need to do some data type change before transforming to sub-set and cleaning.

Dataset b(using sum() to sum na for each columns):

```
> sapply(q3.vote, function(x) sum(is.na(x)))
      state      county   candidate       party total_votes         won 
          0           0           0           0           0           0 
```

Figure 5. NAs in dataset b

2.4. Changing datatypes for dataset a (lieks, retweet_count, lat, long -> type number)

```
> glimpse(new_candidate_data)
Rows: 1,748,088
Columns: 22
$ created_at            <chr> "2020-10-15 00:00:01", "2020-10-15 00:00:01", "2020-10-15 00:00:02", "2020-10-15 00:00:02", "2020-10-15 00:00:08", "2020-10-15~
$ tweet_id              <chr> "1.316529221557252e+18", "1.3165292227484303e+18", "1.316529228091847e+18", "1.316529227471237e+18", "1.3165292523014513e+18",~
$ tweet                 <chr> "#Elecciones2020 | En #Florida: #JoeBiden dice que #DonaldTrump solo se preocupa por Ã©l mismo. El demÃ³crata fue anfitriÃ³n d~
$ likes                 <dbl> 0, 26, 2, 0, 4, 2, 0, 0, 0, 0, 0, 0, 3, 2, 0, 3, 0, 0, 1, 3, 0, 0, 1, 1, 1, 0, 1, 0, 2, 2, 8, 14, 5, 6, 1, 0, 1, 0, 0, 0, 0, 1~
$ retweet_count         <dbl> 0, 9, 1, 0, 3, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 1, 2, 0, 1, 2, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 6, 6, 2, 0, 1, 0, 0, 0, 0, 0, 0, 0, ~
$ source                <chr> "TweetDeck", "Social Mediaset ", "Twitter Web App", "Trumpytweeter", "Twitter for iPhone", "Twitter for Android", "Twitter for~
$ user_id               <chr> "360666534.0", "331617619.0", "8436472.0", "8.28355589206057e+17", "47413798.0", "1138416104.0", "7.674018410302095e+17", "9.0~
$ user_name             <chr> "El Sol Latino News", "Tgcom24", "snarke", "Trumpytweeter", "Rana Abtar - Ø±Ù†Ø§ Ø£Ø¨ØªØ±", "Farris Flagg", "Michael Wilson", ~
$ user_screen_name      <chr> "elsollatinonews", "MediasetTgcom24", "snarke", "trumpytweeter", "Ranaabtar", "FarrisFlagg", "wilsonfire9", "sm_gulledge", "ja~
$ user_description      <chr> "ðŸŒ\220 Noticias de interÃ©s para latinos de la costa este de #EEUU\nâ \200â\217'ï¸\217 Facebook e Instagram\nâ \200ðŸ\217\23~
$ user_join_date        <chr> "2011-08-23 15:33:45", "2011-07-08 13:12:20", "2007-08-26 05:56:11", "2017-02-05 21:32:17", "2009-06-15 19:05:35", "2013-02-01~
$ user_followers_count  <dbl> 1860, 1067661, 1185, 32, 5393, 2363, 75, 766, 151, 8, 4622, 1396, 496, 2755, 6402, 828, 2755, 967, 49, 101, 275, 5974, 1185, 3~
$ user_location         <chr> "Philadelphia, PA / Miami, FL", NA, "Portland", NA, "Washington DC", "Perris,California", "Powell, TN", "Ohio, USA", "Pennsylv~
$ lat                   <dbl> 25.77427, NA, 45.52025, NA, 38.89499, 33.78252, NA, 40.22536, 40.96999, NA, NA, 41.87556, 32.71742, 55.95335, NA, 51.08342, 55~
$ long                  <dbl> -80.1936600, NA, -122.6741949, NA, -77.0365581, -117.2286478, NA, -82.6881395, -77.7278831, NA, NA, -87.6244212, -117.1627714,~
$ city                  <chr> NA, NA, "Portland", NA, "Washington", NA, NA, NA, NA, NA, NA, "Chicago", "San Diego", "City of Edinburgh", NA, NA, "City of Ed~
$ country               <chr> "United States of America", NA, "United States of America", NA, "United States of America", "United States of America", NA, "U~
$ continent             <chr> "North America", NA, "North America", NA, "North America", "North America", NA, "North America", "North America", NA, NA, "Nor~
$ state                 <chr> "Florida", NA, "Oregon", NA, "District of Columbia", "California", NA, "Ohio", "Pennsylvania", NA, NA, "Illinois", "California~
$ state_code            <chr> "FL", NA, "OR", NA, "DC", "CA", NA, "OH", "PA", NA, NA, "IL", "CA", "SCT", NA, NA, "SCT", NA, NA, "MI", NA, NA, "OR", NA, NA, ~
$ collected_at          <chr> "2020-10-21 00:00:00", "2020-10-21 00:00:00.373216530", "2020-10-21 00:00:00.746433060", "2020-10-21 00:00:01.119649591", "202~
$ candidate             <chr> "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", ~
```

Figure 6. Changing results for dataset a

2.5. Sub-set for each question and clean

Q1 (dataset a): three sub-sets which used for analysing the tweets number distribution in general and in terms of time, source and location.

(1) Used for analysing in general. The columns likes, retweet_count, user_followers_count and candidate is used. The screenshot below also shows that it has NA values in some column, since the number of NAs is very low in this dataset (1748088 rows with around 200-300 NAs), so the NAs are simply deleted in this sub-set.

```
> dim(q1.general)
[1] 1748088       4
> sapply(q1.general, function(x) sum(is.na(x)))
               likes        retweet_count user_followers_count            candidate 
                 293                  260                  539                    0 
```

```
> q1.general <- na.omit(q1.general)
> dim(q1.general)
[1] 1747542       4
```

Figure 7. transforming and cleaning for sub-set(1)

(2) Days distribution analysis (created_at, candidate column). When loading it into tableau and generating chars, there are 2 null values, so simply filter them.

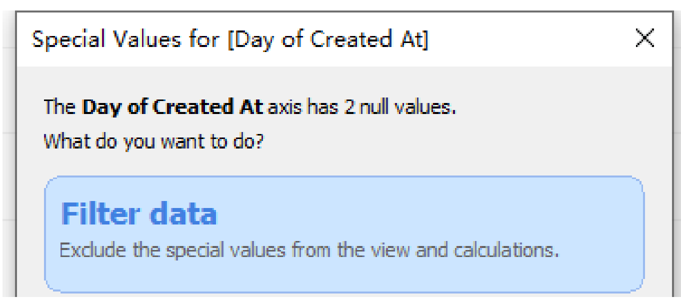

Figure 8. transforming and cleaning for sub-set(2)

(3) Source and location analysis (source, lat, long ,country ,state, candidate column), deleting all NAs for this dataset.

```
> dim(q1.location_source)
[1] 1748088       6
> sapply(q1.location_source, function(x) sum(is.na(x)))
   source       lat      long   country     state candidate 
     1849    947106    947106    951586   1167283         0 
> q1.location_source <- q1.location_source[!(is.na(q1.location_source$source)
 | is.na(new_candidate_data$lat) | is.na(new_candidate_data$long)),]
> dim(q1.location_source)
[1] 800748      6
```

Figure 9. transforming and cleaning for sub-set(3)

Q2 (dataset a): likes, retwet_count, tweet, candidate columns are chosen, deleting NAs. Also, an additional column is needed named id to record the original rows before unnest tweets.

```
> sapply(q2, function(x) sum(is.na(x)))
        likes retweet_count         tweet     candidate            id
          293           260            20             0             0
> glimpse(q2)
Rows: 1,748,088
Columns: 5
$ likes         <dbl> 0, 26, 2, 0, 4, 2, 0, 0, 0, 0, 0, 0, 3, 2, 0, 3, 0,~
$ retweet_count <dbl> 0, 9, 1, 0, 3, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 1, 2, ~
$ tweet         <chr> "#Elecciones2020 | En #Florida: #JoeBiden dice que ~
$ candidate     <chr> "Trump", "Trump", "Trump", "Trump", "Trump", "Trump~
$ id            <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ~
> q2 <- na.omit(q2)
```

Figure 10. transforming and cleaning for Q2

Q3 (dataset a and dataset b):

(1) Dataset a: Using filter() to choosing only US data, choosing state and candidate as the subset and filter the NAs, after that group them by state using group_by() and using summarise() to count the total tweets for each state. Then divided into two sets by different candidates.

```
> head(q3.state.count)
# A tibble: 6 x 3
# Groups:   state [3]
  state   candidate   sum
  <chr>   <chr>     <int>
1 Alabama Biden       864
2 Alabama Trump       849
3 Alaska  Biden       429
4 Alaska  Trump       311
5 Arizona Biden      3248
6 Arizona Trump      2865
```

```
# Groups:   state [54]
   state                sum_biden
   <chr>                    <int>
 1 Alabama                    864
 2 Alaska                     429
 3 Arizona                   3248
 4 Arkansas                   469
 5 California               25814
 6 Colorado                  2687
 7 Connecticut                878
 8 Delaware                   331
 9 District of Columbia      7055
10 Florida                  13278
# ... with 44 more rows
```

```
> q3.state.trump
# A tibble: 53 x 2
# Groups:   state [53]
   state                sum_trump
   <chr>                    <int>
 1 Alabama                    849
 2 Alaska                     311
 3 Arizona                   2865
 4 Arkansas                   613
 5 California               31140
 6 Colorado                  3618
 7 Connecticut               1141
 8 Delaware                   245
 9 District of Columbia      9683
10 Florida                  16554
# ... with 43 more rows
```

Figure 11. transforming and cleaning for Q3(1)

(2) Dataset b: Using filter() to filter the candiate Joe Biden and Donald Trump for two sets separately and then selecting state and total_votes columns and group by state to sum the total_votes in each state (This dataset does not have NAs).

```
> q3.vote.biden
# A tibble: 51 x 2
   state                biden_total
   <chr>                      <int>
 1 Alabama                   849648
 2 Alaska                    153405
 3 Arizona                  1672143
 4 Arkansas                  423932
 5 California              11109764
 6 Colorado                 1804352
 7 Connecticut              1080680
 8 Delaware                  296268
 9 District of Columbia      317323
10 Florida                  5297045
# ... with 41 more rows
```

```
> q3.vote.trump
# A tibble: 51 x 2
   state                trump_total
   <chr>                      <int>
 1 Alabama                  1441168
 2 Alaska                    189892
 3 Arizona                  1661686
 4 Arkansas                  760647
 5 California               6005961
 6 Colorado                 1364607
 7 Connecticut               715291
 8 Delaware                  200603
 9 District of Columbia       18586
10 Florida                  5668731
# ... with 41 more rows
```

Figure 12. transforming and cleaning for Q3(2)

3. Data Checking

3.1 Q1 subset:

There is data inconsistency in country column, change 'United States of America' to 'United States'. In addition, in this column, there are three data that cannot been recognised by tableau, and the problem is that the data is the state name instead of country name, so change them to the right name.

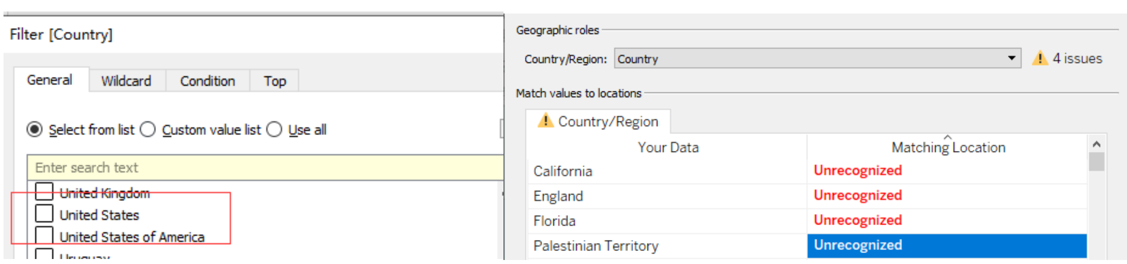

Figure 13. date checking in Q1

3.2 Q2 subset:

The tweets column contains many words with number, special character, web address and some common stop words, so they should be cleaned before unnest. However, after initial cleaning and unnesting, the wordcloud below shows that some frequently used words are meaningless in this particular situation, such as trump, biden and election, so they should be regarded as stop words and deleted.

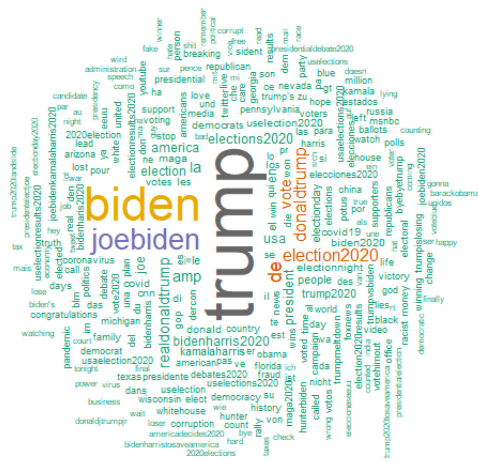

Figure 14. date checking in Q2

3.3 Q3 subset:

When loading data into tableau and check, there is no error in the subset.

4. Data Exploration

- **Non-text data exploration (Q1 & Q3)**

Q1: What is the difference of the number of related discussions, like or retweet of tweets between Trump and Biden in general or in terms of time, location, and source?

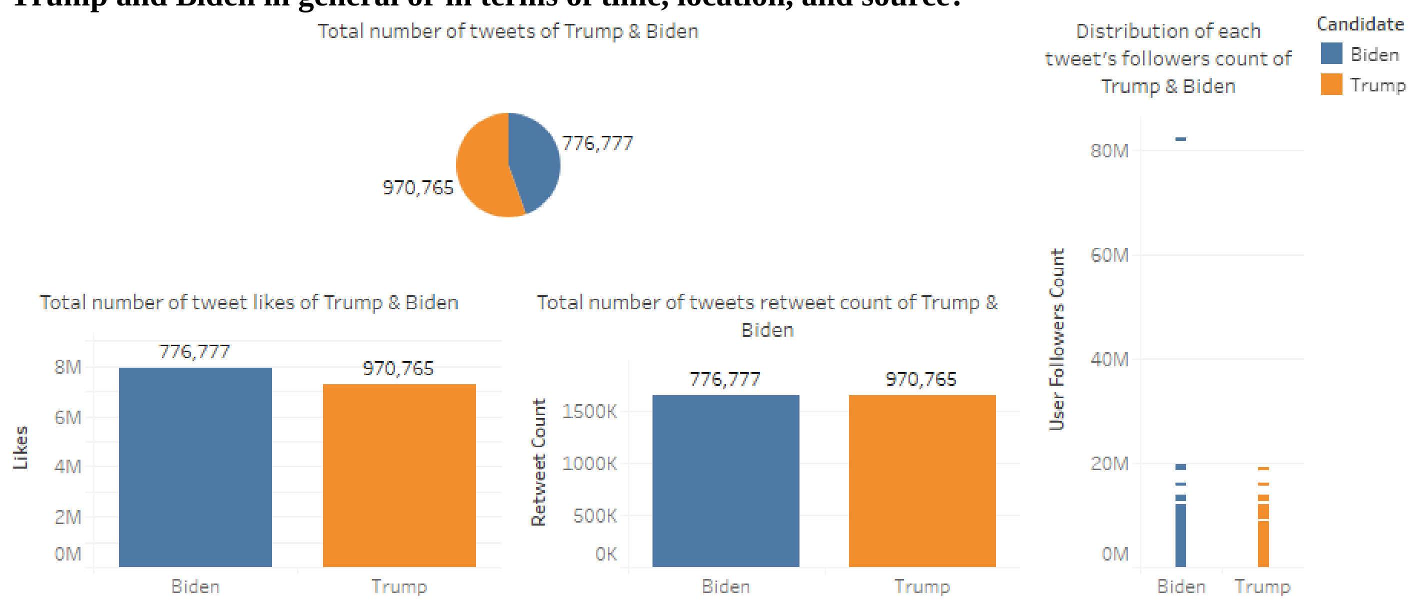

Figure 15. tweets general distribution

The figure above illustrates the distribution of number of tweets in general. First, total number of tweets and the sum of retweet count about Trump is slightly more than that of Biden. However, the sum of likes of the tweets about Biden is more than Trump's. In terms of the peoples followers count distribution, the peoples that tweets about these two people are similar, while for Biden, it has slightly more 'famous followers', which has a large amount of followers.

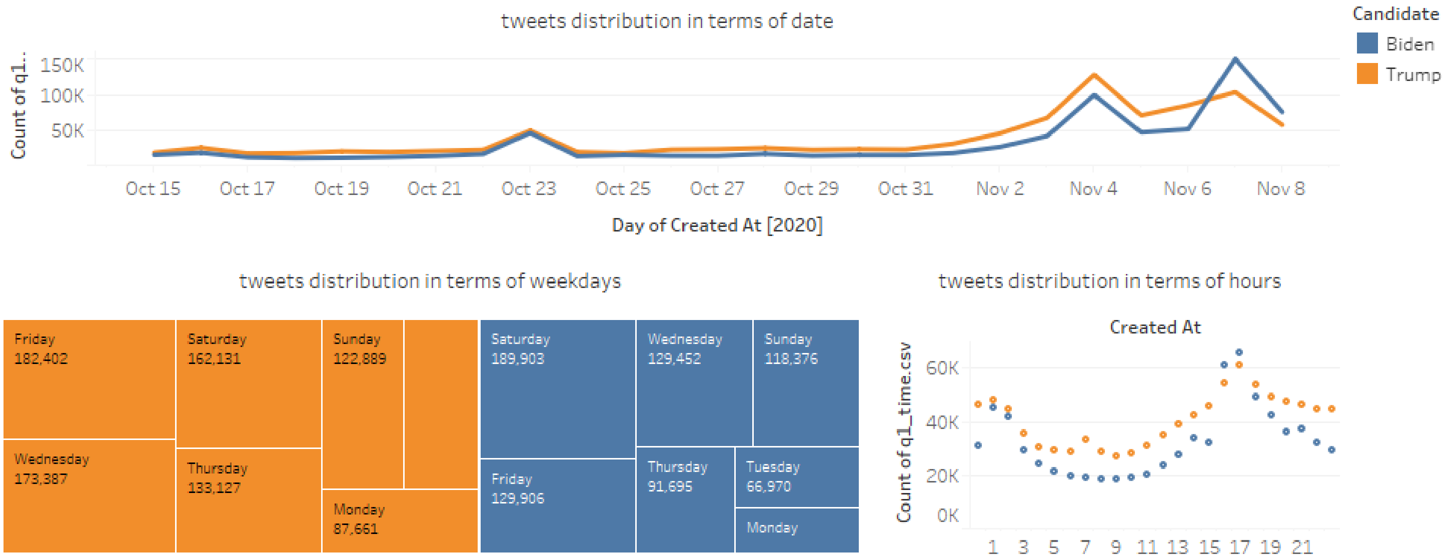

Figure 15. tweets distribution in terms of time

The above figure demonstrates the tweets distribution in terms of time. First, the tweets about these two people increased significantly and fluctuated between Nov 2and Nov 8, and number of tweets about Biden is lower than Trump's before Nov 7th. In terms of weekdays, tweets about Trump are the most, while for Biden, Saturday has the most tweets amount. With regards to different, peoples that tweets about Biden are more active from 15:00 to 17:00.

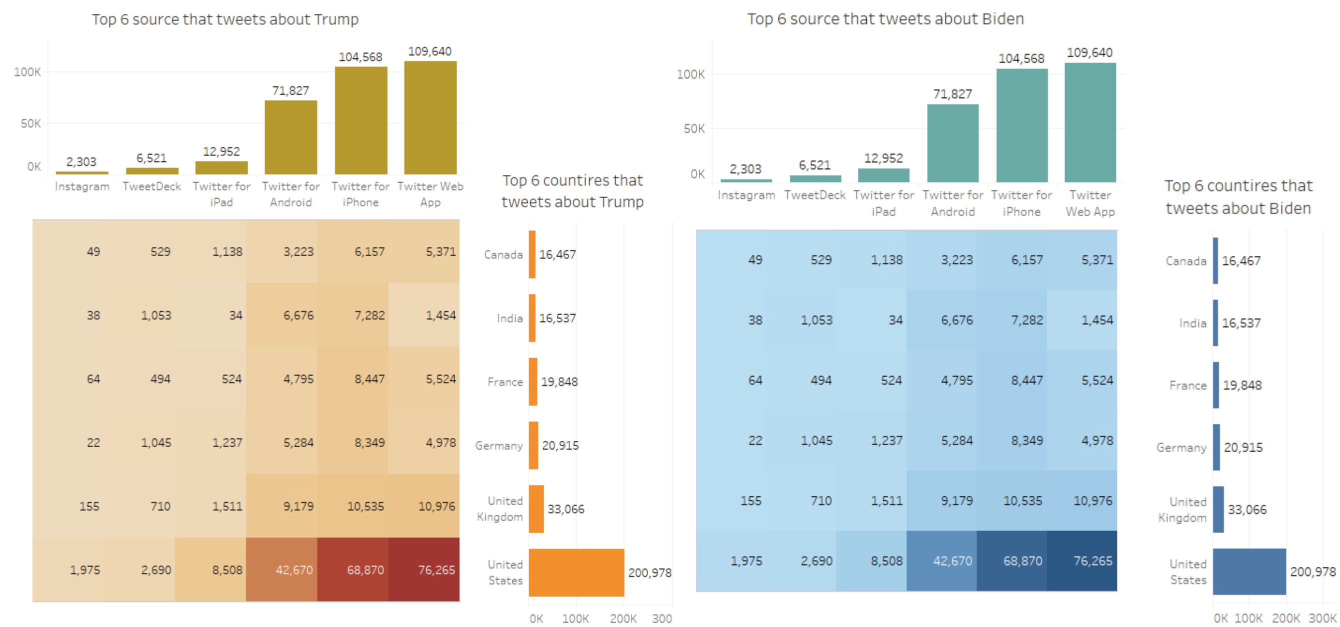

Figure 16. tweets distribution in terms of source and location

The heatmap above shows the tweets distribution in the top 6 sources and top 6 countries. The number of tweets about them in US dominates in the whole records. The distributions for these two candidates are similar as shown above, but it is interesting to note that the US people that tweets about Biden use iPhone to tweets more, which tweets about Trump are more frequently created from Web App.

Q3: Compare the vote by state with the tweets by state and see the difference.

The circular bar chars below shows the top 30 tweets states for Trump and Biden respectively, California contains the most number of tweets for each candidate, while New York has the second number of tweets about them, and Biden is slightly more popular in Texas, while for Trump, he is more popular in Florida.

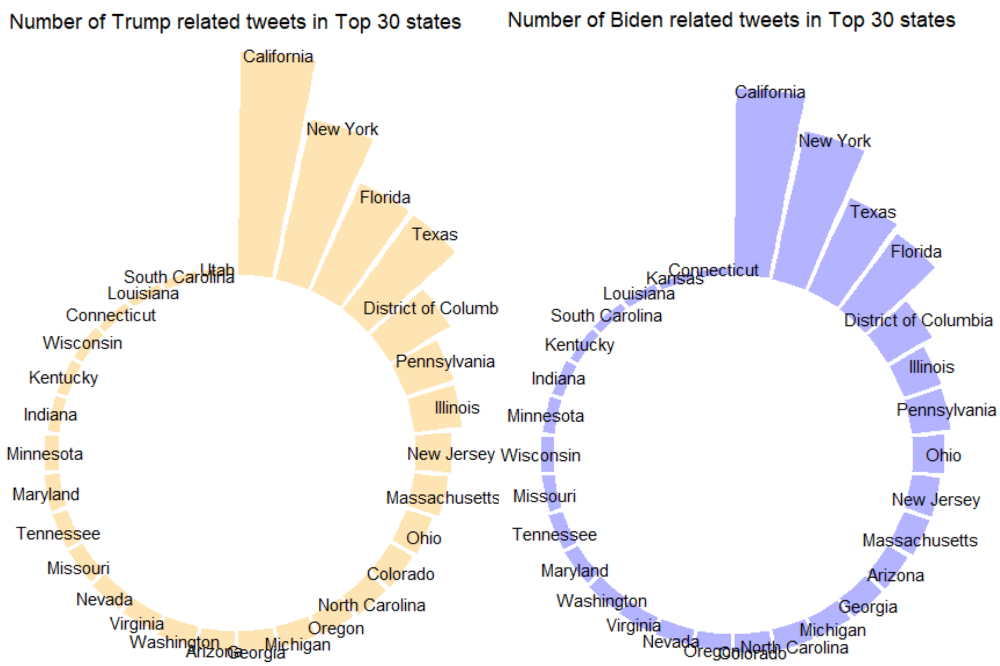

Figure 17. tweets distribution in terms of state in US

Before comparing the votes by state and tweets by state, the data needs some prepocessing to help visualization: adding a new column name Trump Minus Biden, which is the value of Trump's tweets minus Biden's for tweets dataset, and Trump's votes minus Biden's for vote dataset.

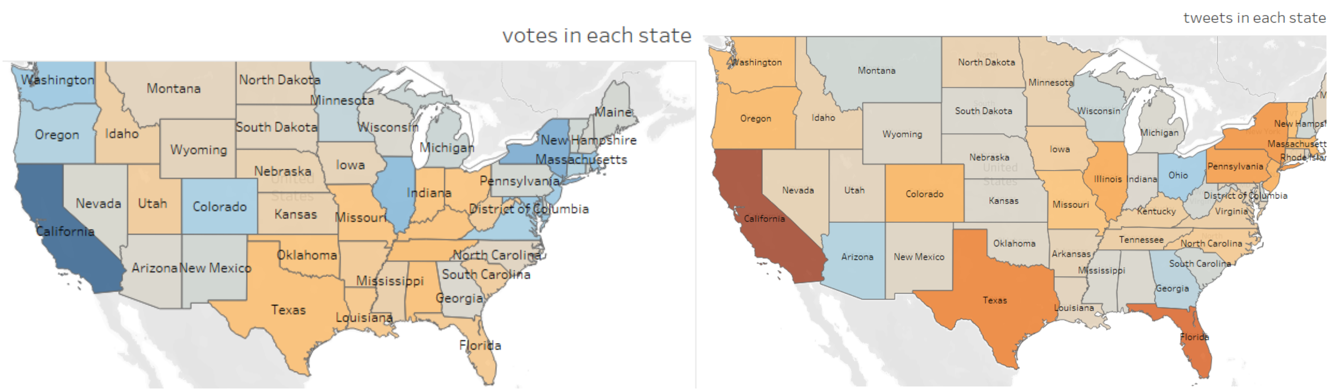

Figure 18. tweets distribution VS votes distribution in each state

The colorperth map above shows the 'winner' in each state (who has more votes or related tweets), the orange color means Trump is the 'winner' in this state and blue means Biden has more votes or tweets in this state. In general, Trump seems to have more dicussions across all America. It is interesting to note that although tweets in some regions, such as Calfornia, are more about Trump, but in fact, Biden has more votes in these states. Therefore, it is important to analysis the content in tweets to discover the reasons behind in Question 2.

To discover more about the correlation between number of tweets and the real total votes in each state, a standardlization process has performed before plotting the scatter plot, in which the votes or number of tweets for each state and each person are divided by the sum of total votes or number of tweets respectively. The standradlized dataset is shown below.

```
> q3.trump.standard
# A tibble: 51 x 5
# Groups:   state [51]
   state                sum_trump trump_total    vote    tweet
   <chr>                    <int>       <int>   <dbl>    <dbl>
 1 Alabama                    849     1441168 0.00476 0.0193
 2 Alaska                     311      189892 0.00174 0.00255
 3 Arizona                   2865     1661686 0.0161  0.0223
 4 Arkansas                   613      760647 0.00344 0.0102
 5 California               31140     6005961 0.175   0.0805
 6 Colorado                  3618     1364607 0.0203  0.0183
 7 Connecticut               1141      715291 0.00640 0.00959
 8 Delaware                   245      200603 0.00137 0.00269
 9 District of Columbia      9683       18586 0.0543  0.000249
10 Florida                  16554     5668731 0.0928  0.0760
```

```
> q3.biden.standard
# A tibble: 51 x 5
# Groups:   state [51]
   state                sum_biden biden_total    vote   tweet
   <chr>                    <int>       <int>   <dbl>   <dbl>
 1 Alabama                    864      849648 0.00564 0.0104
 2 Alaska                     429      153405 0.00280 0.00187
 3 Arizona                   3248     1672143 0.0212  0.0204
 4 Arkansas                   469      423932 0.00306 0.00517
 5 California               25814    11109764 0.169   0.135
 6 Colorado                  2687     1804352 0.0176  0.0220
 7 Connecticut                878     1080680 0.00574 0.0132
 8 Delaware                   331      296268 0.00216 0.00361
 9 District of Columbia      7055      317323 0.0461  0.00387
10 Florida                  13278     5297045 0.0867  0.0646
# ... with 41 more rows
```

Figure 19. standradlized result

After that, the scatter plot with a curve fitting for each person is shown in below figures. In general, the correlation is similar, and when the amount of tweets increases, Biden seems to have more votes than Trump according to the fitting cruve.

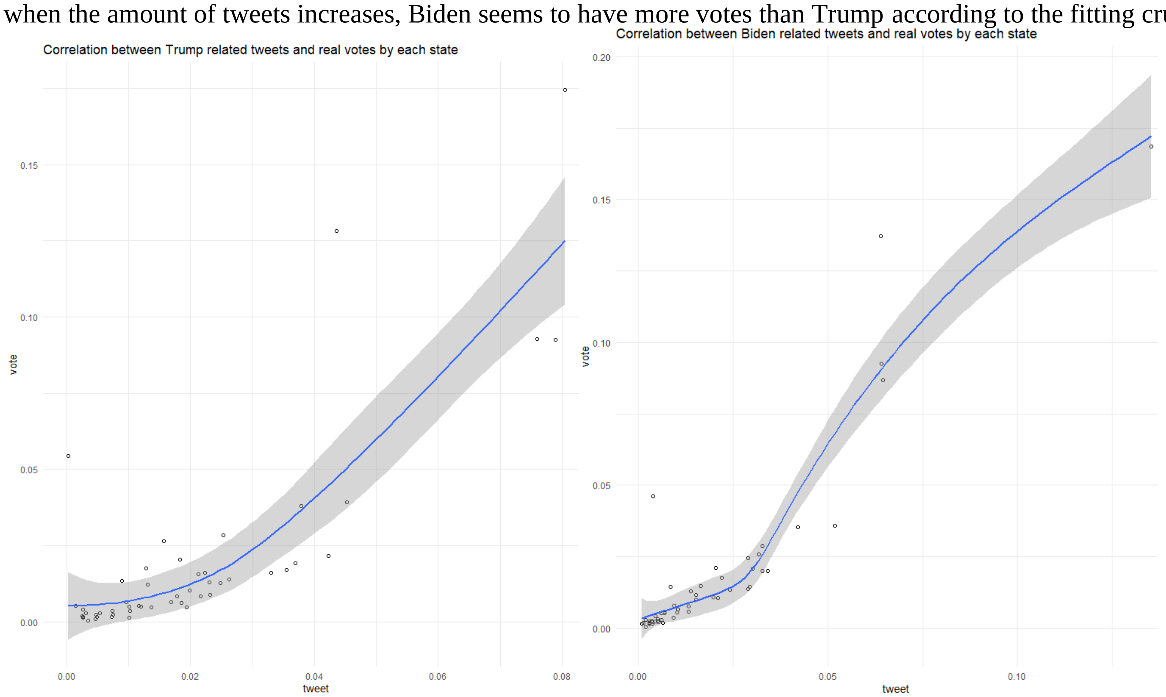

Figure 20. correlation between tweets and votes by state

- **Text data analysis (Q2)**

Q2 What are the frequent words in tweets content? Can we analyze and compare the sentiment inside the content?

To analysing text data (tweets column), the clean tweets should first been divided into words using unnest_tokens() function and count each word's frequency. The two word clouds below shows the words that used in Trump's related tweets and Biden's related tweets respectively. Vote is the common frequent word, and realdonaldtrump and bidenharries are both frequent in both word cloud, which indicates that the two people are often talked together by people. Some words, such as china, people, win and covid are both popular in two tweets set, and in Trump's word cloud, covid, republicans, foxnews, country are more frequently mentioned in tweets. In Biden's word cloud, Obama, congratulation, country are more actively shown in tweets.

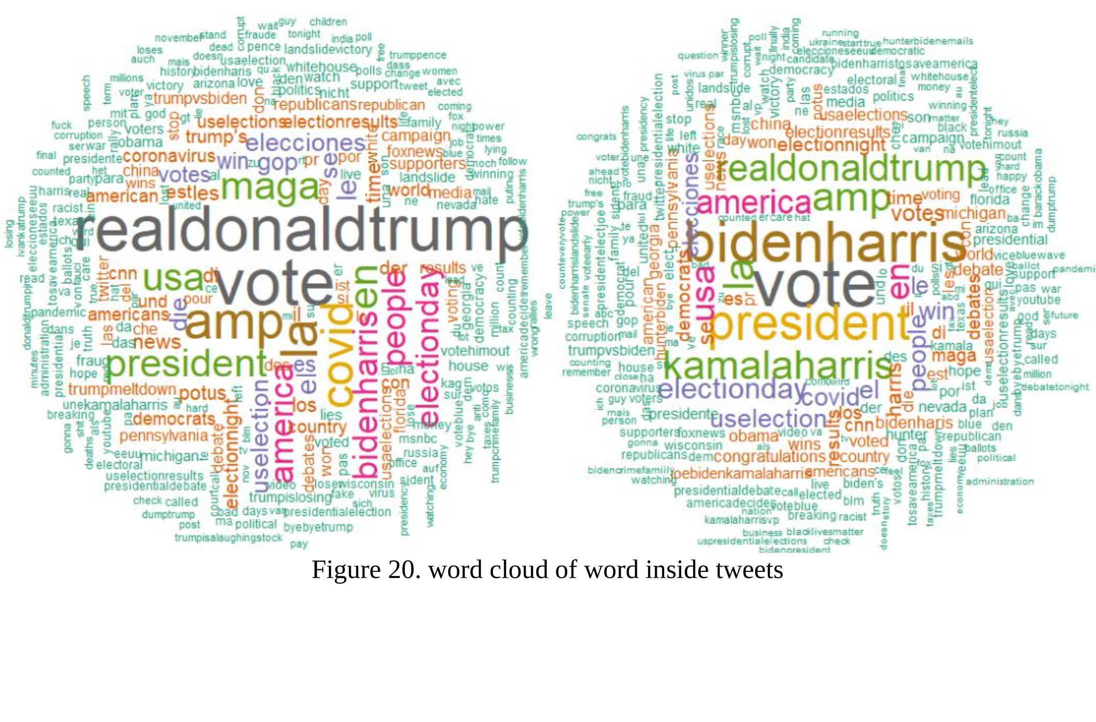

Figure 20. word cloud of word inside tweets

The radar char below is the sentiment analysis of the tweets about the two candidates. The inner_join() function is used for joining the word sub-set with the sentiment data NRC Word-Emotion Association Lexicon (Mohammad & Turney, 2013). In this figure, Biden's related tweets have more positive, anticipation and trust words, for Trump, the positive and negative words count are similar, the fear words count is also very large, so some tweets about Trump may contains negative and fear emotion.

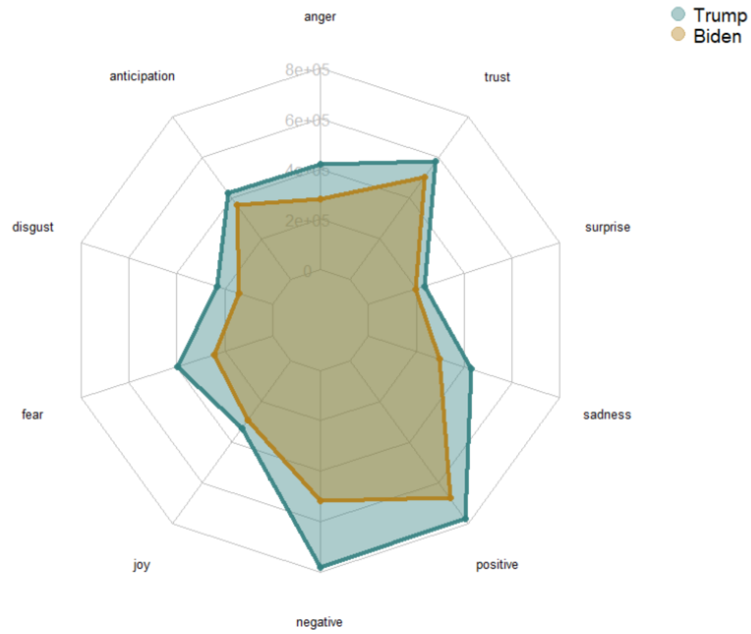

Figure 21. radar charts of sentiment words count inside tweets

In order to observe the positive and negative emotions in tweets instead of in words, another sentiment analysis dataset is used ("afinn"), and the words are first inner join with this dataset, then the data are grouped by the row id, which has been mentioned before and aims to recognize each original tweet, the sentiment score is calculated by the sum the value of each word's sentiment score in a tweet. The histograms below are the distribution of tweet sentiment score for each people's related tweets. It is obvious that Biden's related positive tweets are more than the negative tweets, but Trump's data are opposite, the negative tweets are more.

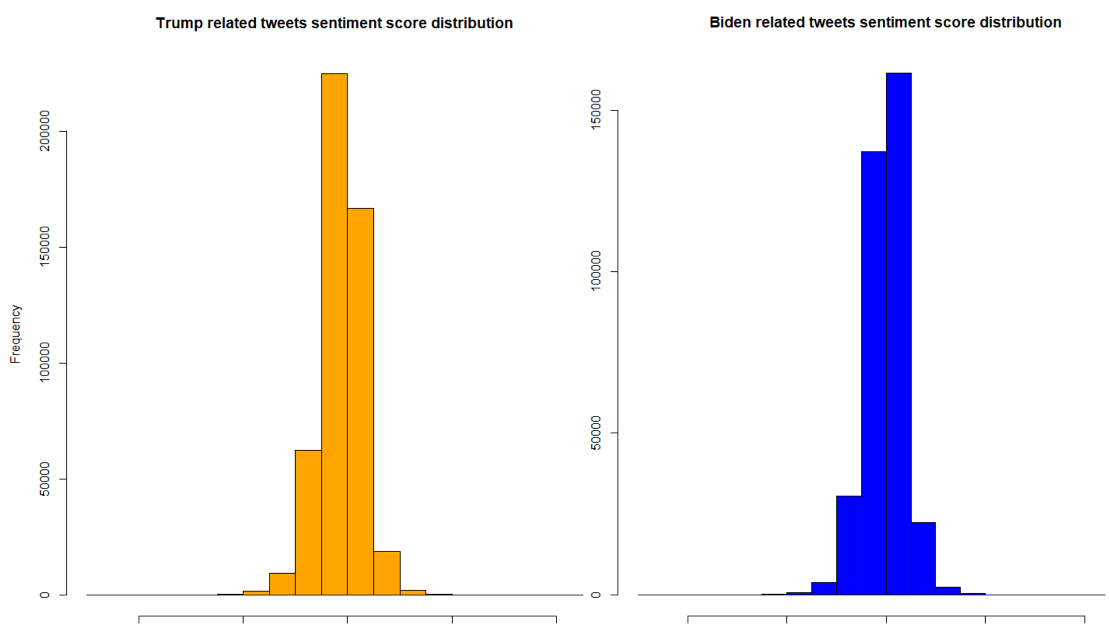

Figure 22. histograms of sentiment score of each tweet

5. Conclusion

In conclusion, in this report, the US 2020 Election tweets are analysed in different aspect, and three question are answered by the exploration and various plots. The three questions are:

1. What is the difference of the number of related discussions, like or retweet of tweets between Trump and Biden in general or in terms of time, location, and source?

2. What are the frequent words in tweets content? Can we analyse and compare the sentiment inside the content?

3. Compare the vote by state with the tweets by state and see the difference.

In terms of its distribution, Trump is more popular in most of time in general or in terms of time and location, but Biden related tweets has more likes. For the source, Trump is more popular on Web App, while a large number of peoples that tweets on Biden are iPhone users.

When it comes to the words and sentiment inside tweets, frequent words in the two candidates' related tweets are similar, but has some differences, which is interesting to discover, and Biden seems to have more positive related tweets than Trump.

By comparing the real votes and tweets, Trump often has more tweets discussion than Biden in different states, but in some states like California, although he is popular on Twitter, Biden has significant more votes than him. In addition, it is shown in this report that there is a correlation between tweets and votes.

In conclusion, the datasets that chosen are suitable to answer the questions.

6. Reflection

In this project, the use of various kinds of data representation ways and plots are learned, and the prepossessing, which includes cleaning, wrangling and checking, is really important and it could influence the data exploration.

However, there is a limitation in sentiment analysis, which is that the words have not been process by stemming and may contains the same word but with different forms, which needs to be improved. In addition, the sentiment tweet analysis can consider more information included such as retweet count and number of likes to generate various analysis.

7. Bibliography

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, *29*(3), 436-465.