#Introduction:

Hurricane Harvey is an Atlantic hurricane that has attack America for several times in the past year. The most recent attack happened on August 13 when the National Hurricane center detect the hurricane on the western east of Africa. It hit United States on Texas and Louisiana on August 23rd.

In this project, I am going to use the information from twitter to identify possible the most serious hurt area and its possible moving route.

#Method:

1. Basic assumption:
   1.1 The severity of the hurricane affects the times people tweets.
   1.2 The location information in the location column the tweet original display is accurate and represent the people's real location.
   1.3 The missing and useless data is random.

2. Data extraction:
   The data comes from twitter from Sep 4th to Sep 8th. For each day, I pooled the 10,000 tweets data from the API(except for Sep 4th) and use the location information as the basic data source to identify the path of hurricane and use the number of the same location occur as the severity of the hit.
   Also, I poll the data again using the same method for Sep 29th. Still with the same key words. This dataset is used as reference for correction.

3. Data cleaning:
   3.1 Delete the data without "location" information(NAs). and create new dataset called cHHs.
   3.2 Further cleaning the data: including separate the location information by city and state, showing all the characters in upper form and cleaning the spaces. Then, separate the dataset into two part, the one with useful state information and the one without useful state information. Abandon those without useful state information since most of it are typed and have nonsense information compared to those with state.
   3.3 Since the hurricane hurt TX and LA, extract the time each city shows in TX and LA, both in the data that containing a state information and without state information And separate the dataset based on different dates. Also, extract MD, too, as reference.

4. Data showing:

   The graphing process containing several different steps.

   4.1 basic row data exploration. Graphing the first 20 cities that have the most frequency of showing for different dates using ggplot (geom_bar) and for the reference dates, too. Also, the multiplot function from the internet helped to merge all the graph together.

   4.2 Exclude other confounders using the data 9/29 as reference.

   In order to exclude other confoundings that may have effect on twitter number, I extract the twitter date at Sep 29th. Since September 29th is the date almost a month away from the hit of hurricane Harvey, we can suppose that the twitter number in that date can represent usual twitter number.

   I get the frequency of each city that shows up in the twitter for the top 40 cities each day from Sep 4th to Sep 8th and Sep 29th. And then, using the frequency in the hurricane days to subtract the frequency of none hurricane days and get the absolute differences.

   The reason that I didn't use relative risk as indicator is that some cities are in hurricane hit dates are not present in non-hurricane hit dates. These cities have the frequency of 0 and we cannot use them as dominator.

   After we get the absolute difference, that is the one indicator I would use for the following step, including showing them on the map.

   4.3 using ggmap package and ggplot to graph the area on the map as to visualize the data.

5. Statically comparing and assumption checking

   5.1 comparing the frequency of Maryland cities showed up and TX/LA cities showed up. Test the hypothesis of whether most hit area got the most tweets number.

   5.2 Comparing the frequency of hurricane dates and non-hurricane dates. Both 5.1 and 5.2 are done by the adjusting. If after adjusting, there are more TX/LA cities shows up than before adjusting, it would support the assumption

   5.3 Using the number difference between reference date and hurricane date as the new number to identify the most several hurt areas.

#result:

1. Raw data:

   Before the adjusting, we can see that there are several non-hurricane hit state are in the top 20 cities that have the most tweets, but most of them are from hurricane hit area. The cities that are in the top 20 among the 5 days are:

2. Basic map after adjusting comparing different dates:

From the graph below, Huston always the biggest proportion in the map. This is reasonable since Boston is the capital of the TX. It is difficult to tell whether Huston is the most hit city. However, as the time passed on, the trend are moving to the northwest of the state, this may show the pattern as the hurricane moves into the mainland.

3. Assumption checking:

3.1 I am going to use bar plot to show each date, the top 20 cities before adjusting and after adjusting to test the hypothesis 1 and 2.

From the bar plot, we could see clear difference between adjustment and non-adjustment data. Taking September 4th as an example, before adjustment, there are 3 cities in non-hurricane area showing up in the first 20 cities, but after adjustment, there is only 1 city. It shows that, there is a difference between the tweet pattern before and after the hurricane hurt.

3.2 Identify the most serious hurt area:

From the result below, the city that have the most tweets after adjustment is Huston, Austin, Corpus Christi, Katy, other cities changed a lot base on different dates.

#Discussion.

In this project, I predict the most hurt area by counting the number of tweets sending from each city. From the result, Huston and Austin get the most frequency of showing up and this is the same as what I found from the actual damage information from the internet. However, there are few limitations about this analysis.

As the result, although that the results are most similar to the actual serious hurt area, there is an exception. Dallas, which is also serious hurt, not shown itself in the map. In order to make sure why this happen, I check again the original dataset before the adjustment, and there it is, always in the best 10 of the city list. So, what happened is, it shows up in high frequency on date 9/29, too. And after the subtraction, it disappear. This might happen by on the specific day of 9/29, people in Dallas accidently tweet a lot by some reason.

However, by using the twitter, it can give a basic map of what is happening across the country, but not that precisely.

Citations:

1. Multigraph function: http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/

2. https://en.wikipedia.org/wiki/Hurricane_Harvey
3. Twitter R package and tutorial: http://geoffjentry.hexdump.org/twitteR.pdf