Week 1 Assignment

In this assignment, you will explore data on holidays and events in Russia and Russian trolls tweets. It will be up to you to clean and combine the raw data using the best practices outlined in the readings for this week. Please be sure to set your working directory to your folder on Dropbox.

Data

Four datasets are provided:

- `islamist_groups.csv`: A list of Islamist terrorism groups.
- `GTD.csv`: Dataset of all incidents of terrorism between 1970 to 2018, collected and maintained by the Global Terrorism Database.
  - `eventid`: An unique incident identifier
  - `iyear`: Year of event
  - `imonth`: Month of event
  - `iday`: Day of event
  - `country_txt`: Country where the event occurred
  - `provstate`: Province or state within the country where the event occurred
  - `city`: City where the event occurred
  - `gname`: Name of the perpetrator group
  - `summary`: Description of the event
- `Russian_Holidays.csv`: List of annual Russian holidays
  - `Month`: Month of holiday
  - `Day`: Day of holiday
  - `Holiday`$_{N}ame : Name of holiday$`Notes` : $Description of holiday$
  - `Religious`: Indicator for whether holiday is a religious holiday
  - `Public`: Indicator for whether holiday is a public holiday
  - `Political`: Indicator for whether holiday a political holiday

- `IRA_tweets.csv`: Panel data of IRA-produced tweets from 2015 to 2018.
  - `Date`: Date
  - `day`: Day of the week
  - `tweet_count_islam`: Number of IRA tweets about Islam
  - `tweet_count_blm`: Number of IRA tweets about BLM movement
  - `tweet_count_all`: Total number of IRA tweets

Guidelines

1 Constructing Panel Data

Construct a panel of IRA tweet count, terrorist events, and Russian holidays at the day level from January 1 2015–June 30 2018. As you clean and prep the datasets prior to constructing the panel, check for common problems we encounter when working with data (eg., missing observations, duplicates, etc.) When constructing the panel, check whether the data is balanced or unbalanced. Note down whether any of them were problems you encountered and how you fixed them. The final dataset should contain the following:

- A date variable
- The number of total tweets, tweets about BLM, and tweets about Islam per day
- Indicator columns for whether: (a) a terrorist event occurred in Russia; (b) an islamist terrorist event occurred in Russia; (c) it is a holiday; (d–f) whether it is a public/religious/political holiday specific

(a) How many total observations are there in the final panel?

(b) On how many days were there a terrorist or islamist terrorist event in Russia?

(c) How many holiday days were there in total from January 1 2015 to June 30 2018? How many days with public holidays, religious holidays, and political holidays?

2 Descriptive Statistics & Data Visualizations

(a) Write a function that calculates descriptive statistics (N, mean, min, median, max, and S.D.). Then, loop over each of the three tweet count variables in the data to

generate those descriptives and output a 3 x 7 dataframe of these with the variable name.

(b) Try different plots to visualize the distribution of each tweet count variable. Which type of graph is most effective in your opinion, and why?

(c) What do you notice about the distribution of these variables? Can you think of any problems that this may cause for analysis? What could you do to remedy these problems? Try them and report on whether they worked.

(d) Create a time-series plot of these tweet count variables in a way that you think is meaningful. Explain what considerations you thought through and what the figure shows.