Data Tidying and Analysis

Andrew Pichette

12/3/2019

```r
#load all packages
library(readxl)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ----------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(dplyr)
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```r
library(ggrepel)
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```r
getwd()
```

```
## [1] "C:/Users/Andrew/Desktop/R Final Project/Code"
```

```r
#read datasets into R
afghan_strike_data <- read.csv("../data/afghanistan_table.csv", stringsAsFactors = F)

global_terror_data <- read.csv("../data/globalterrorismdb_0919dist.csv", stringsAsFactors = F)
```

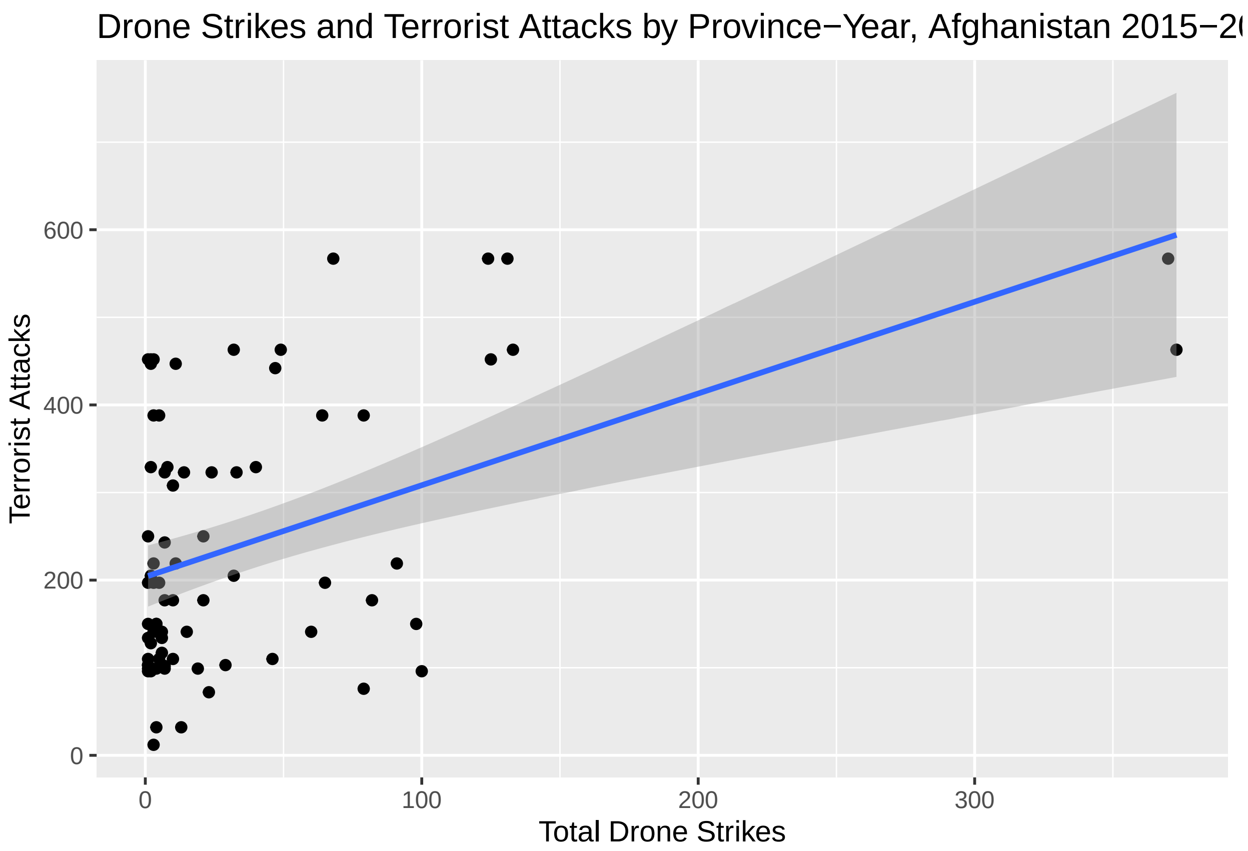
```
#create column "year" from "date"
afghan_table <- mutate(afghan_strike_data, year =
    format(as.Date(afghan_strike_data$Date, format = "%d/%m/%Y"), "%Y"))
```

```
#create "strikes_by_district" dataframe, observing target district, sum of max  number of strikes,  and

Max.Strikes <- select(afghan_table, Province, Minimum.strikes, Maximum.strikes, Minimum.total.people.kil
  group_by(Province, year) %>%
  summarize(Maximum.strikes = sum(Maximum.strikes))

Max.People.Killed <- select(afghan_table, Province, Minimum.strikes, Maximum.strikes, Minimum.total.peop
  group_by(Province, year) %>%
  summarize(Maximum.total.people.killed = sum(Maximum.total.people.killed))

strikes_by_district <- left_join(Max.Strikes, Max.People.Killed, by = c("Province", "year"))

strikes_by_district$year <- as.integer(strikes_by_district$year)
```

```
#shrink "global_terror_data" to observations of Afghanistan and select for desired columns to create af
afghan_terror_data <- global_terror_data %>%
  filter(country_txt == "Afghanistan", iyear > 2014) %>%
  select(iyear, provstate) %>%
  add_count(provstate) %>%
  distinct()


colnames(afghan_terror_data) <- c("year", "provstate", "total_incidents")
```

```
#merge "afghan_terror_data" onto "strikes_by_district" to create "strikes_and_terror_df". This join pro
strikes_and_terror_df <-left_join(strikes_by_district, afghan_terror_data, by = c("year", "Province" =
  na.omit() %>%
  unite(strikes_and_terror_df, 1:2, sep = "-")

colnames(strikes_and_terror_df) <- c("district_year", "max_strikes", "max_killed", "terrorist_attacks")


#manually removing rows for "Unknown" provinces
final_data <- strikes_and_terror_df[c(1:67, 71:79), ]

colnames(final_data) <- c("province_year", "max_strikes", "max_killed", "terrorist_attacks")
```

```
# test plot to visualize district-year strikes vs terrorist attacks
ggplot(data = final_data, aes(x = max_strikes, y = terrorist_attacks)) +
  geom_point()+
  geom_smooth(method = "lm", size = 1) +
  xlab("Total Drone Strikes") +
  ylab("Terrorist Attacks") +
  ggtitle("Drone Strikes and Terrorist Attacks by Province-Year, Afghanistan 2015-2018")
```

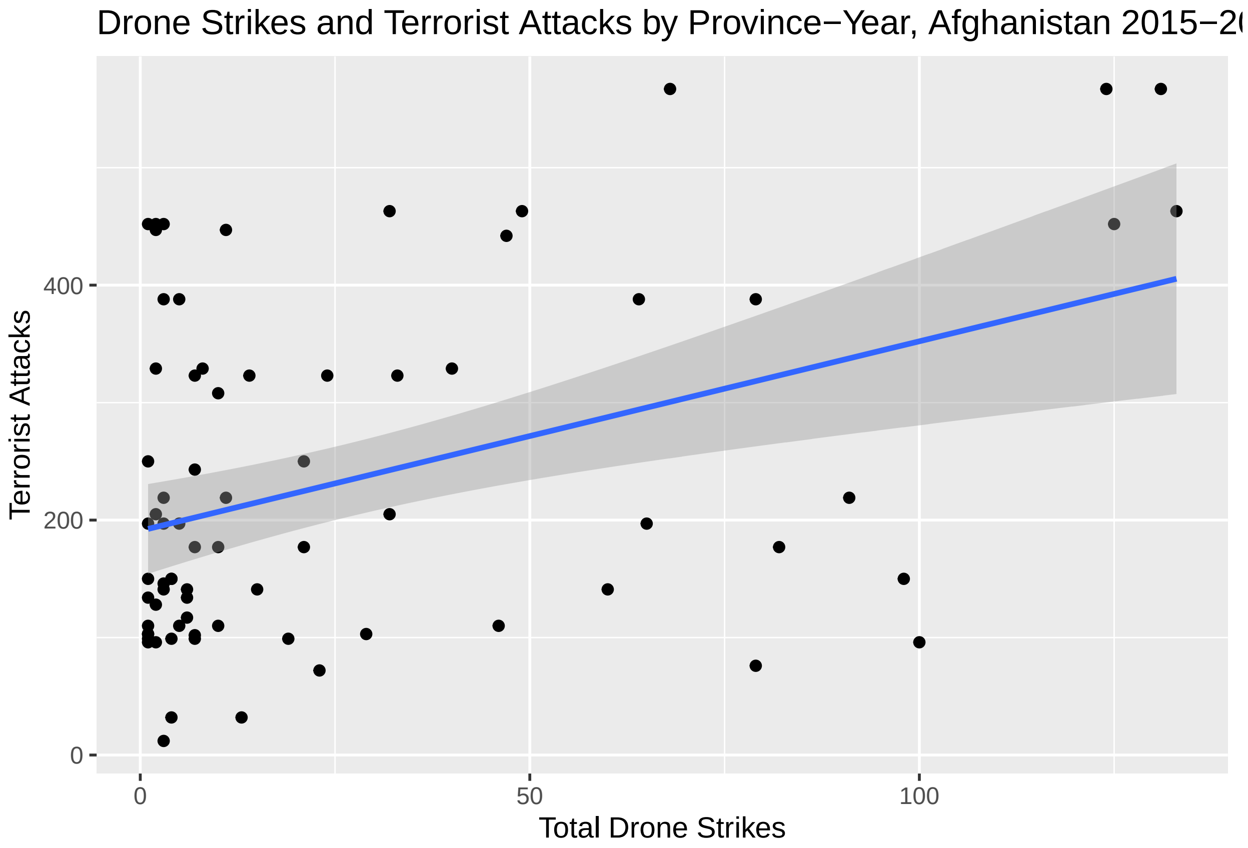


```r
#create a regression table
mod.1 <- lm(formula = terrorist_attacks ~ max_strikes, data = final_data)
mod.2 <- lm(formula = terrorist_attacks ~ max_killed, data = final_data)
mod.3 <- lm(formula = terrorist_attacks ~ max_strikes + max_killed, data = final_data)


stargazer(mod.1, mod.2, mod.3, title = "Regression Results", type = "text",
covariate.labels = c("Drone Strikes", "Casualties From Strikes"),
omit = "Constant", dep.var.labels = "DV: Terrorist Attacks",
keep.stat="n", style = "ajps",
out = "regression-table.txt")
```

```
##
## Regression Results
## --------------------------------------------------
##                           DV: Terrorist Attacks
##                         Model 1  Model 2  Model 3
## --------------------------------------------------
## Drone Strikes           1.047***          0.687***
##                         (0.237)           (0.229)
## Casualties From Strikes          0.548*** 0.436***
##                                  (0.099)  (0.102)
## N                          76       76       76
## --------------------------------------------------
## ***p < .01; **p < .05; *p < .1
```

```r
#Given the presence of clear outliers in the data, I've chosen to subset two observations out to see th

final_data_no_outliers <- final_data[-c(20, 52), ]
```

```r
#replotting with outliers removed as test
ggplot(data = final_data_no_outliers, aes(x = max_strikes, y = terrorist_attacks)) +
  geom_point()+
  geom_smooth(method = "lm", size = 1) +
  xlab("Total Drone Strikes") +
  ylab("Terrorist Attacks") +
  ggtitle("Drone Strikes and Terrorist Attacks by Province-Year, Afghanistan 2015-2018")
```



```r
#create another test regression table with outliers removed
mod.1 <- lm(formula = terrorist_attacks ~ max_strikes, data = final_data_no_outliers)
mod.2 <- lm(formula = terrorist_attacks ~ max_killed, data = final_data_no_outliers)
mod.3 <- lm(formula = terrorist_attacks ~ max_strikes + max_killed, data = final_data_no_outliers)


stargazer(mod.1, mod.2, mod.3, title = "Regression Results", type = "text",
covariate.labels = c("Drone Strikes", "Casualties From Strikes"),
omit = "Constant", dep.var.labels = "DV: Terrorist Attacks",
keep.stat="n", style = "ajps",
out = "../Results/regression-table.txt")
```

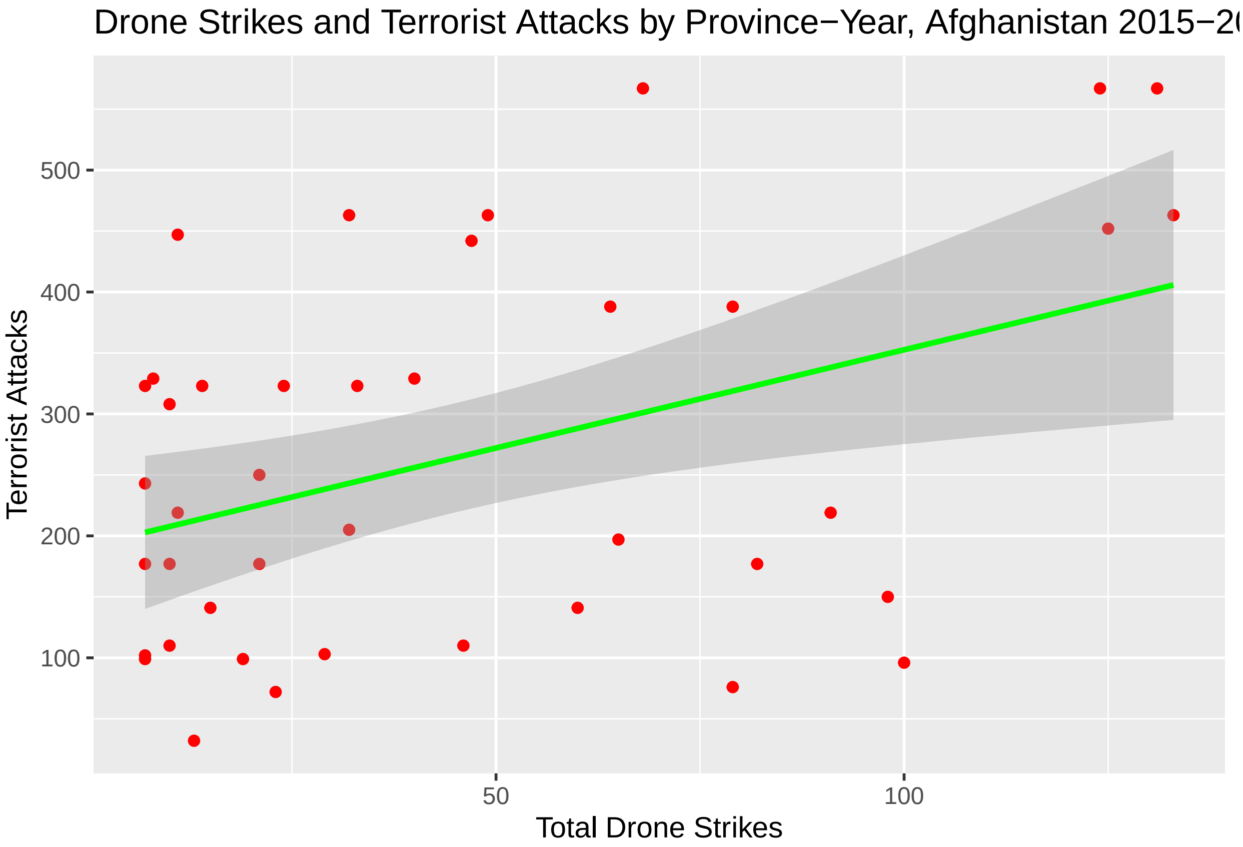
```
##
```

```
## Regression Results
## --------------------------------------------------
##                           DV: Terrorist Attacks
##                         Model 1  Model 2  Model 3
## --------------------------------------------------
## Drone Strikes           1.613***           0.373
##                         (0.437)           (0.503)
## Casualties From Strikes          0.540*** 0.485***
##                                  (0.095)  (0.121)
## N                          74       74       74
## --------------------------------------------------
## ***p < .01; **p < .05; *p < .1
```

```r
#Per Pete's recommendation, declutter the above visualizations by removing some of the observations in

final_data_decluttered <- final_data_no_outliers %>%
  arrange(max_strikes) %>%
  filter(max_strikes > 6)
```

```r
#repeat visualization with new dataframe "final_data_decluttered"
ggplot(data = final_data_decluttered, aes(x = max_strikes, y = terrorist_attacks)) +
  geom_point(color = "red")+
  geom_smooth(method = "lm", size = 1, color = "green") +
  xlab("Total Drone Strikes") +
  ylab("Terrorist Attacks") +
  ggtitle("Drone Strikes and Terrorist Attacks by Province-Year, Afghanistan 2015-2018") +
  ggsave("province_year_strikes_attacks.jpg", path = "../Results")
```

```
## Saving 6.5 x 4.5 in image
```


Drone Strikes and Terrorist Attacks by Province−Year, Afghanistan 2015−2
Terrorist Attacks
500
400
300
200
100
50
100
Total Drone Strikes