Amber Pasiak

Data Analysis

Proposal

11/20/19

Working Title: FOIA Release Time 2008-2015: A Predictive Data Analysis

In 2017, Ted Bridis, for PBS, published an article stating that during Obama’s presidency, he set the record for lawsuits in defending the refusal of FOIA releases. In a First Amendment Watch article, a similar statement has been made about President Trump, claiming that he also has increased the challenges for the public to access records. For the final project I had wanted to see if there was a statistical significance between FOIA release time between President Obama and President Trump. I wanted to look at this due to numerous articles claiming that the release time has gotten worse. My aim was to assess whether there was any statistical significance of the data to verify, or disprove this. However, there is no public data for the years after 2018.

I was able to find open data from 2008-2015 on Kaggle and 2008-2018 on FOIA.gov. The FOIA.gov data is sectioned off by agency and then by the complication of a release. When uploaded into R, the data is all in character form and not numeric. Due to these inconstancies I would like to use only the Kaggle dataset. The Kaggle data set ends in 2015. With this limitation in mind, a predictive data analysis, or more specifically a quantitative forecast, could be performed. As I would be unable to directly compare the Kaggle data to the FOIA.gov data of 2015-2018 because the variables are so different and the project time constraint, this project will analysis President Obama’s FOIA release time from 2008-2015 and then through R make a prediction on the future progression of release time based on a continuation of the trends.

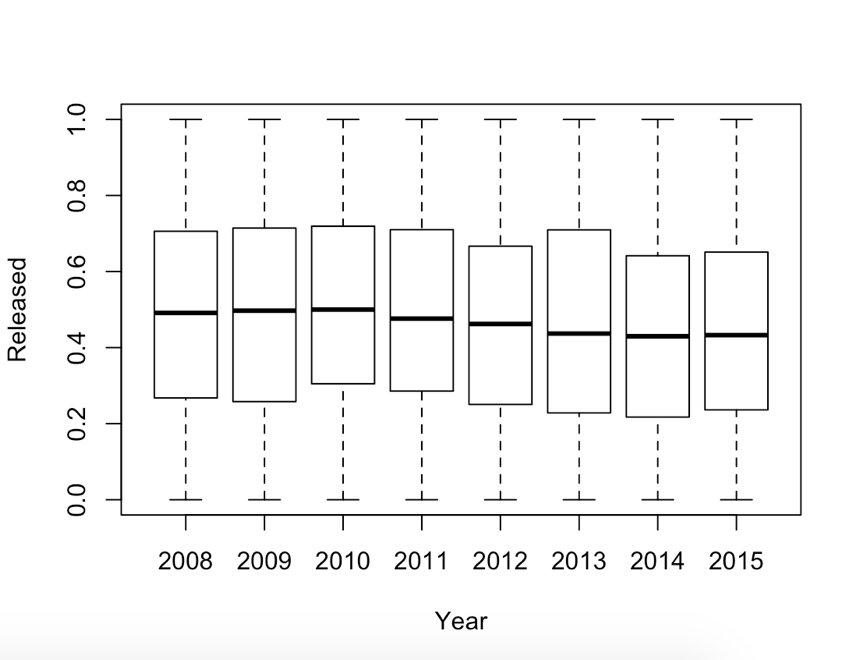
This ideally would be the first part of a larger project. The second half would take place a few years from now, when a comparison can be made between Obama and Trump and to see if the prediction was correct. The findings from this project and future projects could be important when assessing the implications made by journalists on the despairing state of access to public, government records.

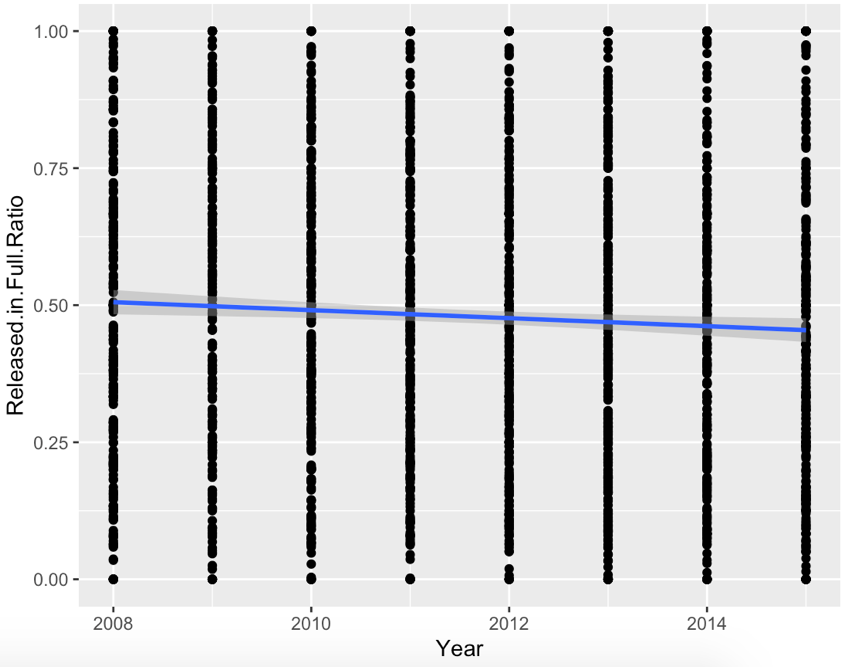
The dataset I will be using has 11 variables: Agency, Department, Year, Requests pending at the start of the year, Requests pending at the end of the year, Requests received, Requests processed, Released in full, Released in part, Denied in full. The released variables are in a ratio format. I was unable to find how this was computed in the metadata. I am also unsure how to make them all the same numeric values in order to make a direct comparison. This is another limitation of the dataset, yet with a little math I believe this can be accounted for. The variables that I will be taking a deeper look at include: Agency, Released in full, Denied in full, and Year.

I am wondering if there is a way to create a timeline that shows important political moments that may change policy or have historically affected the release time. This could include amendments made, or events with a perceived threat. I want to be able to account for this uncertainty somehow, even if it is not used directly with the Kaggle dataset.

I have started to create a time series, however am unsure if I should chose the normal distribution, and the standard deviation, or I use the information I have from the data already. The mean for Released in full is .4794478 and the standard deviation for released in full is .278873. In “Data Analysis with R”, by Tony Fischetti, it shows this being chosen at random.

Through the analysis I have done thus far, I have been able to see a decreasing trend in FOIA release times from the years selected, as shown in the box plot and a scatterplot with a regression line below. The next step in this project is to create and analysis the time series, do an accuracy assessment and create a codebook.





**Data analysis plan:**

* Predictive Analysis: Quantitative Forecast
* Scatterplot with regression line
* Box Plot of mean
* Time series
* Autocorrelation?
* Accuracy assessment: histogram of residuals
* Codebook

**Code:** (Most of this so far is trial and error with help from google and the Data Analysis with R book, will be updated when project is finalized)

library(lubridate)

library(dplyr)

library(tidyverse)

library(gmodels)

library (broom)

library(GGally)

library (ggplot2)

install.packages("ggplot2")

install.packages("rcompanion")

install.packages("forecast")

library(forecast)

FOIA <- read.csv("../amberpasiak/Desktop/2019-Pasiak-DATALIB628-FOIA/CSV/requests.csv", header = TRUE)

summary(FOIA)

glimpse(FOIA)

head(FOIA)

tail(FOIA)

summary(FOIA$Released.in.Full.Ratio)

summary(FOIA$Denied.in.Full.Ratio)

summary(FOIA$Department)

ggplot(FOIA, aes (x= Year, y= Released.in.Full.Ratio)) +

geom\_point() + geom\_smooth(method = "lm")

plotmeans(Released.in.Full.Ratio ~ Year, data = FOIA, frame = FALSE)

#well that didnt work

hist(FOIA$Released.in.Full.Ratio,

col="gray",

main="",

xlab="Released Full")

x = FOIA$Released.in.Full.Ratio

plotDensityHistogram(x,

adjust = 1)

hist(FOIA$Year,

col="blue",

main="",

xlab="Year")

boxplot(Released.in.Full.Ratio ~ Year,

data=FOIA,

ylab="Released")

mean(FOIA$Released.in.Full.Ratio, na.rm = TRUE)

sd(FOIA$Released.in.Full.Ratio, na.rm = TRUE)

by(FOIA$Released.in.Full.Ratio, FOIA$Year, mean, na.rm= TRUE)

summary(FOIA$Year)

Requests\_Year <- select (FOIA, Year)

head(Requests\_Year)

FOIA %>% arrange (desc(Year))

Requests\_2008 <- FOIA %>%

filter (Year == "2008")

summary(Requests\_2008)

R2008 <- rnorm(329, mean= .4940, sd= .2788)

glimpse(R2008)

hist(R2008)

#okay something new, time series:

ts(FOIA$Year, start = 2008, end = 2015,frequency = 8 )

plot(FOIA)

FOIAYear <- ts(FOIA$Year, start = 2008, end = 2015,frequency = 8 )

plot(FOIAYear)

#already have a FOIAYear? Need a new name … nope, FOIA08.15, still years on X and Y

FOIA08.15 <- ts(FOIA$Released.in.Full.Ratio, start = 2008, end = 2015,frequency = 8 )

plot(FOIA08.15)

#Why is this so different than scatterplot? Am I doing this wrong or is this not possible?

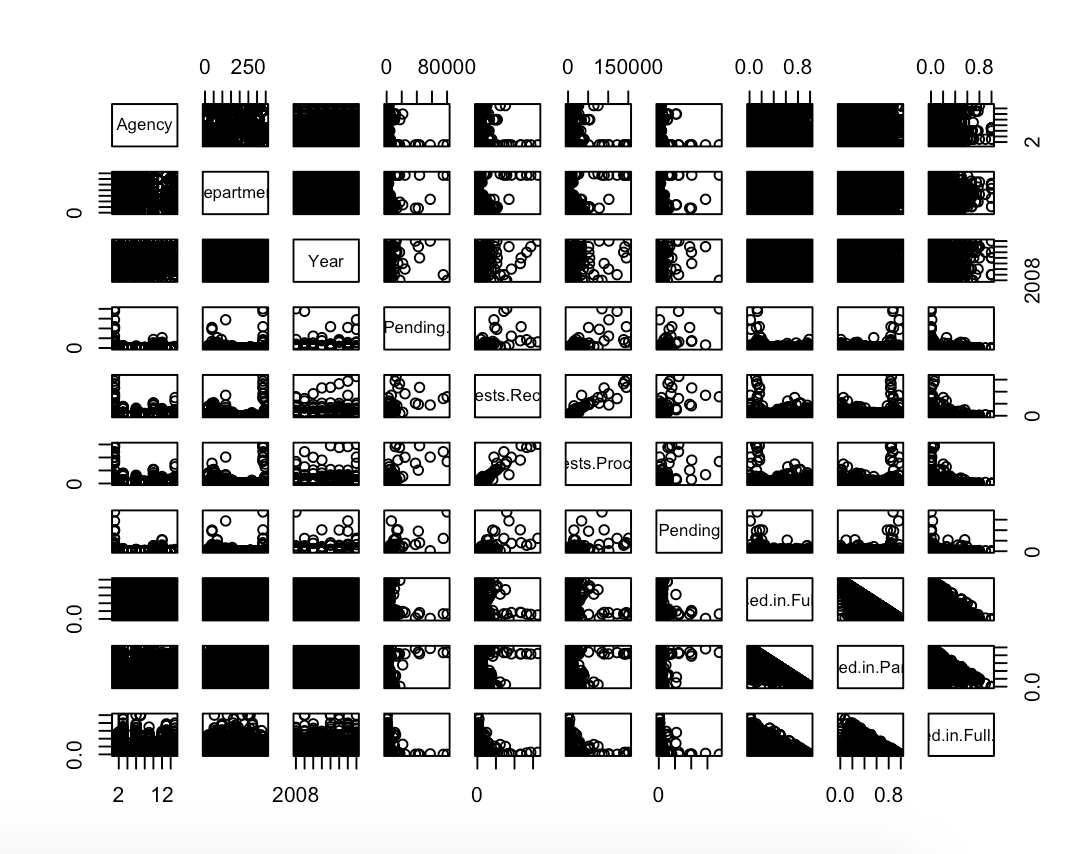
FOIA08.15P <- rnorm(329, mean= .4940, sd= .2788)

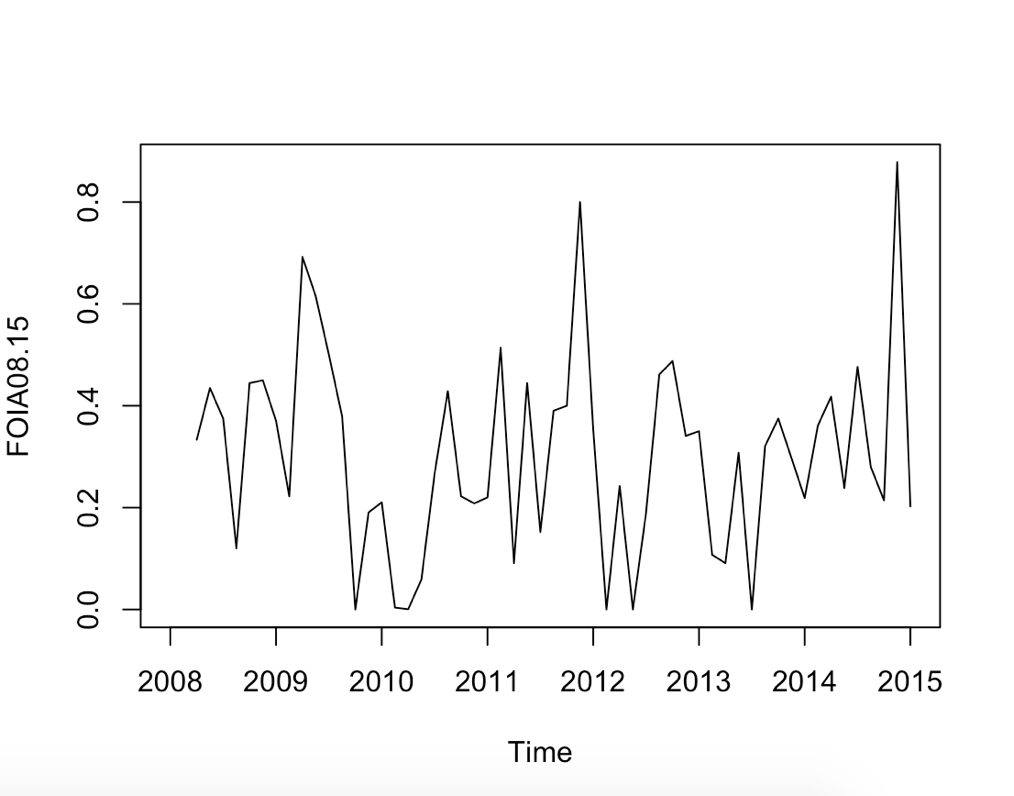
head(FOIA08.15P)

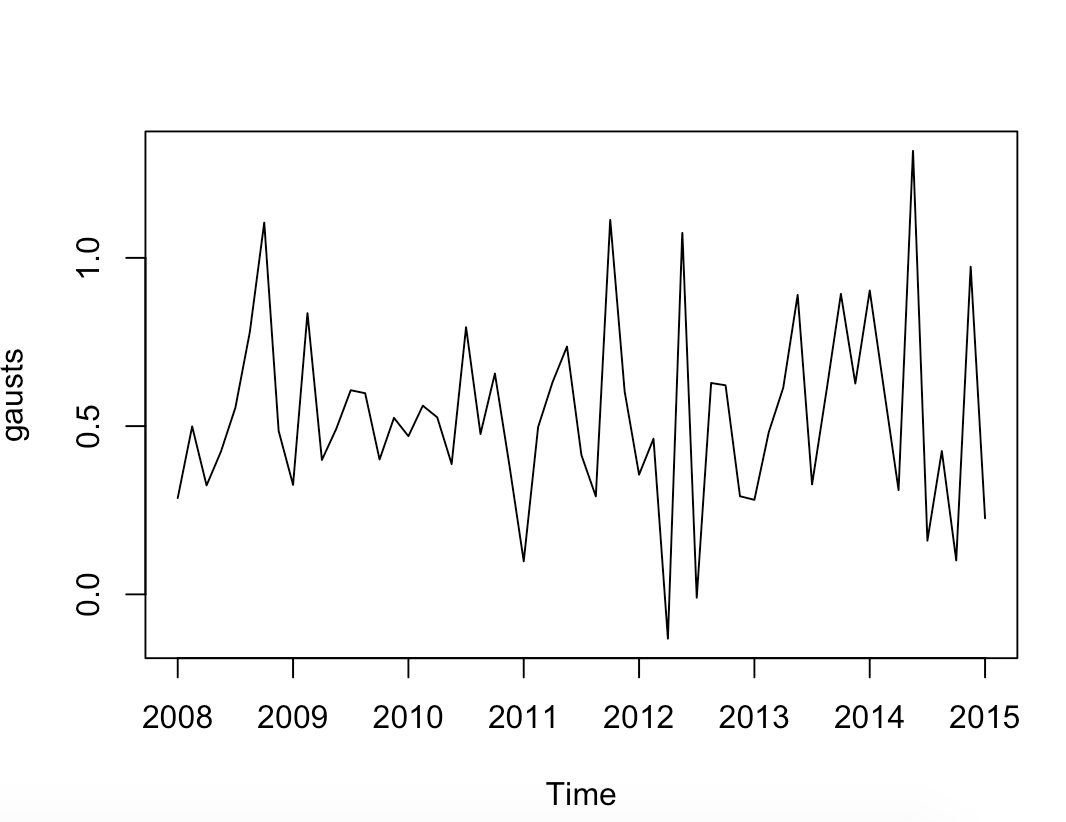
gausts <- ts(FOIA08.15P, start= 2008, end = 2015, frequency = 8)

plot(gausts)

**Other Images:**



****

****

**Working Resources:**

First Amendment Watch at New York University. The Limits of Transparency and FOIA Under Trump. Retrieved from <https://firstamendmentwatch.org/deep-dive/the-limits-of-transparency-and-foia-under-trump/>

Bridis, T. (2017, March 14). In Obama’s final year, U.S. spent $36 million in FOIA lawsuits. *PBS.* Retreieved from <https://www.pbs.org/newshour/nation/obamas-final-year-u-s-spent-36-million-foia-lawsuits>

The United States Department of Justice. Retrieved from <https://www.justice.gov/oip/foia-library>

Fischetti, T. Data Analysis with R. (2018).