

2018 DC R Conference

# Anomaly Detection with Time Series

Catherine Zhou

SENIOR DATA SCIENTIST + MANAGER



# Anomaly Detection with Time Series

... or how to know when something is terribly wrong



twitter @catherinezh

#rstats

#rstatsdc

#rladies

#codecademy

# About Me\_

- Proud New YorkeR
- Currently @ Codecademy
- Formerly @ JetBlue & NY/DC/Boston Sports Clubs

## ABSTRACT

With the rise of streaming data and cloud computing, data scientists are often asked to analyze terabytes of data. The sheer amount of data available leads to a lag time in identifying irregularities, resulting in lost time and revenue.

We can pinpoint these outliers through anomaly detection algorithms, which can be repurposed to monitor key metrics, website breakage, and fraudulent activity. I will demonstrate how we can build a system for anomaly detection to uncover blind spots in large datasets and reduce fire drills at work.

# Agenda

By the end of this talk you will be able to:

- **Analyze seasonal trends** in time series
- **Plot and visualize anomalies** in *Google Trends* data
- Use *anomalize* to **do this the tidy way**
- Explore different **anomaly detection algorithms**
- Explain **case studies** where outliers can be useful

PART ONE

# Time Series\_

Jared Lander



oooh, show how you do  
time series forecasting

Jared Lander

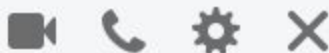


oooh, show how you do  
time series forecasting

**I have a complicated  
relationship with forecasting  
lol**



Jared Lander



oooh, show how you do  
time series forecasting

I have a complicated  
relationship with forecasting  
lol

**facebook**

**Basic Information**

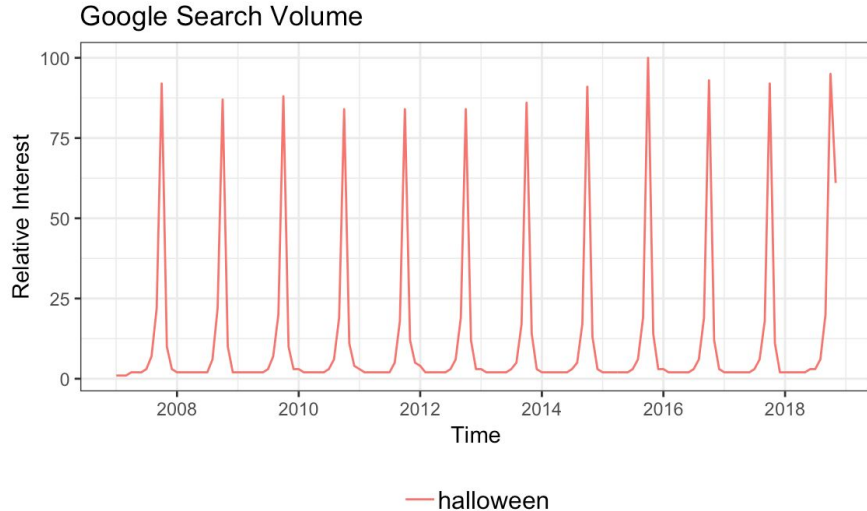
**Relationship  
Status**

It's complicated

# EXPECTATION

We want to work with data that is:

- Clean and well-organized
- Daily or weekly patterns
- Clear seasonal trends
- Key metrics to monitor
- Actionable insights



# EXPECTATION

We want to work with data that is:

- Clean and well-organized
- Daily or weekly patterns
- Clear seasonal trends
- Key metrics to monitor
- Actionable insights



Jared Lander



oooh, show how you do  
time series forecasting

I have a complicated  
relationship with forecasting  
lol

Ppl don't like hearing they  
don't have enough quality  
data to forecast well

## EXPECTATION

We want to work with data that is:

- Clean and well-organized
- Daily or weekly patterns
- Clear seasonal trends
- Key metrics to monitor
- Actionable insights

VS.

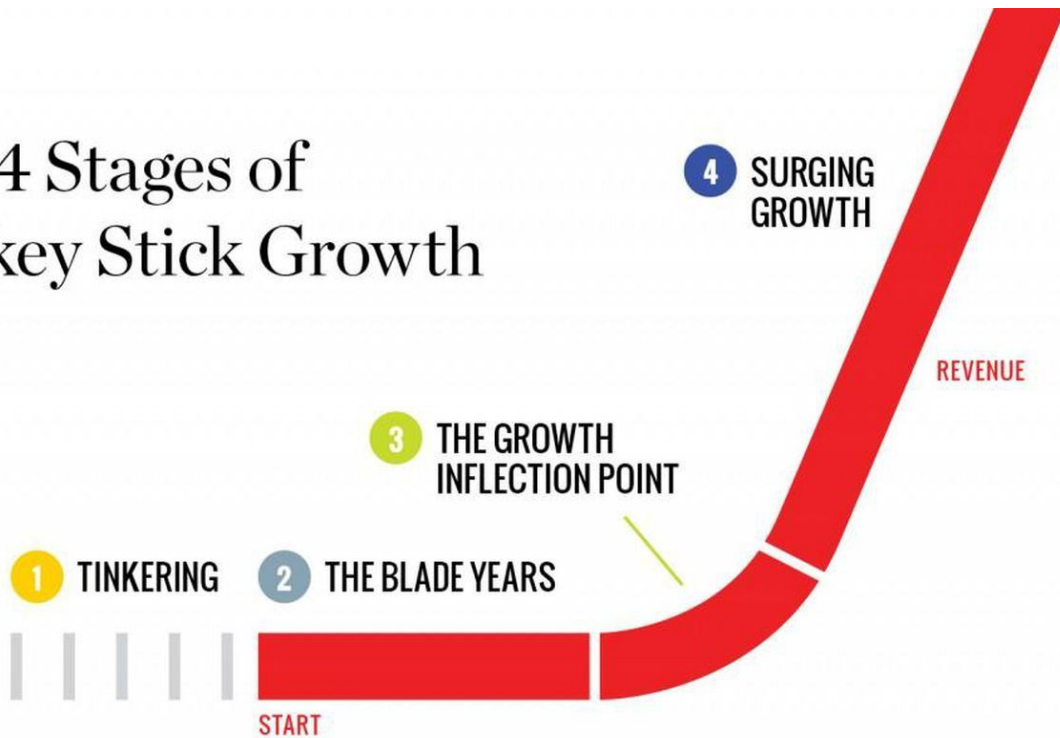
## REALITY

We often work with data that has:

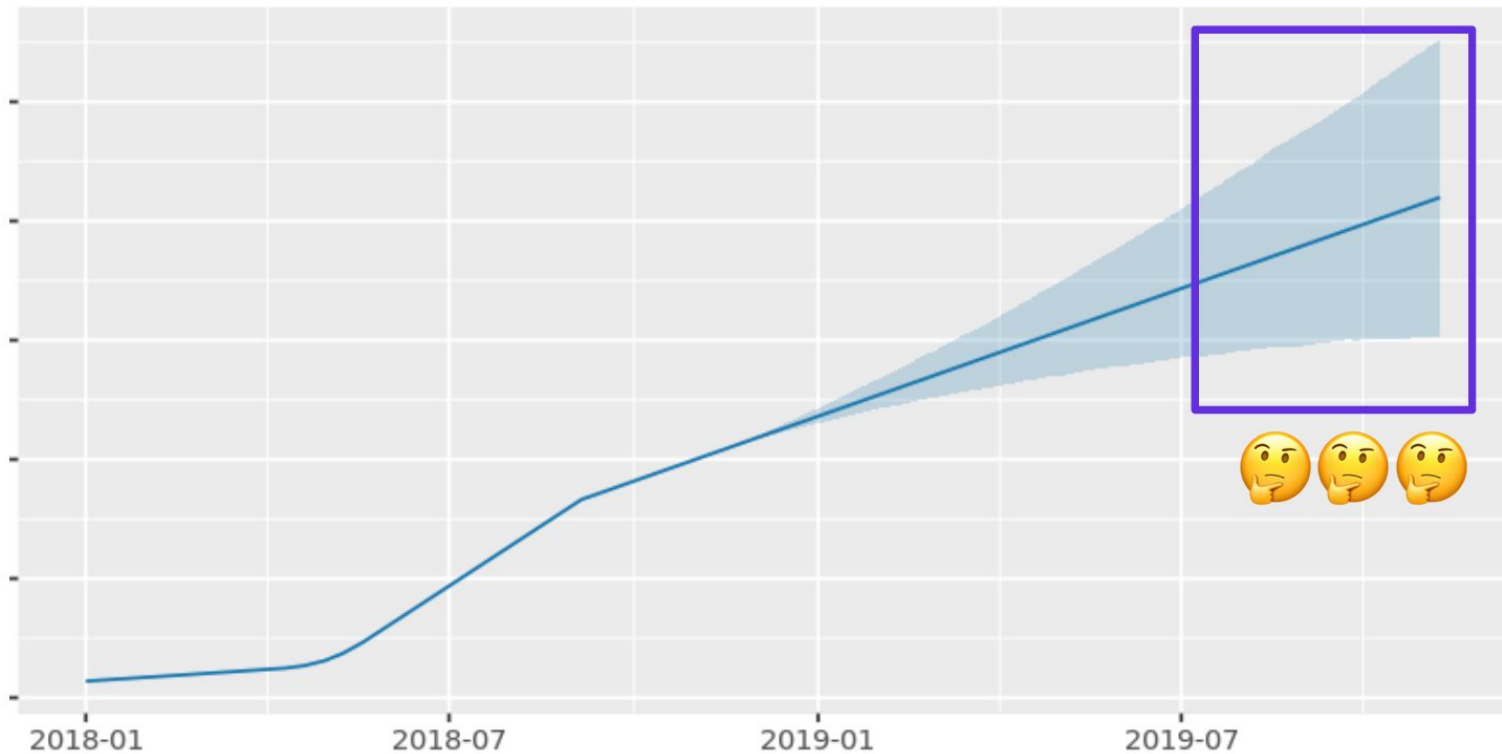
- Inconsistent trends and patterns
- Terabytes in size
- Multiple key metrics
  - Difficult to monitor
  - Difficult to interpret

For data scientists in tech, growth is a double-edged sword.

## The 4 Stages of Hockey Stick Growth



## Growth creates uncertainty in time series forecasting.



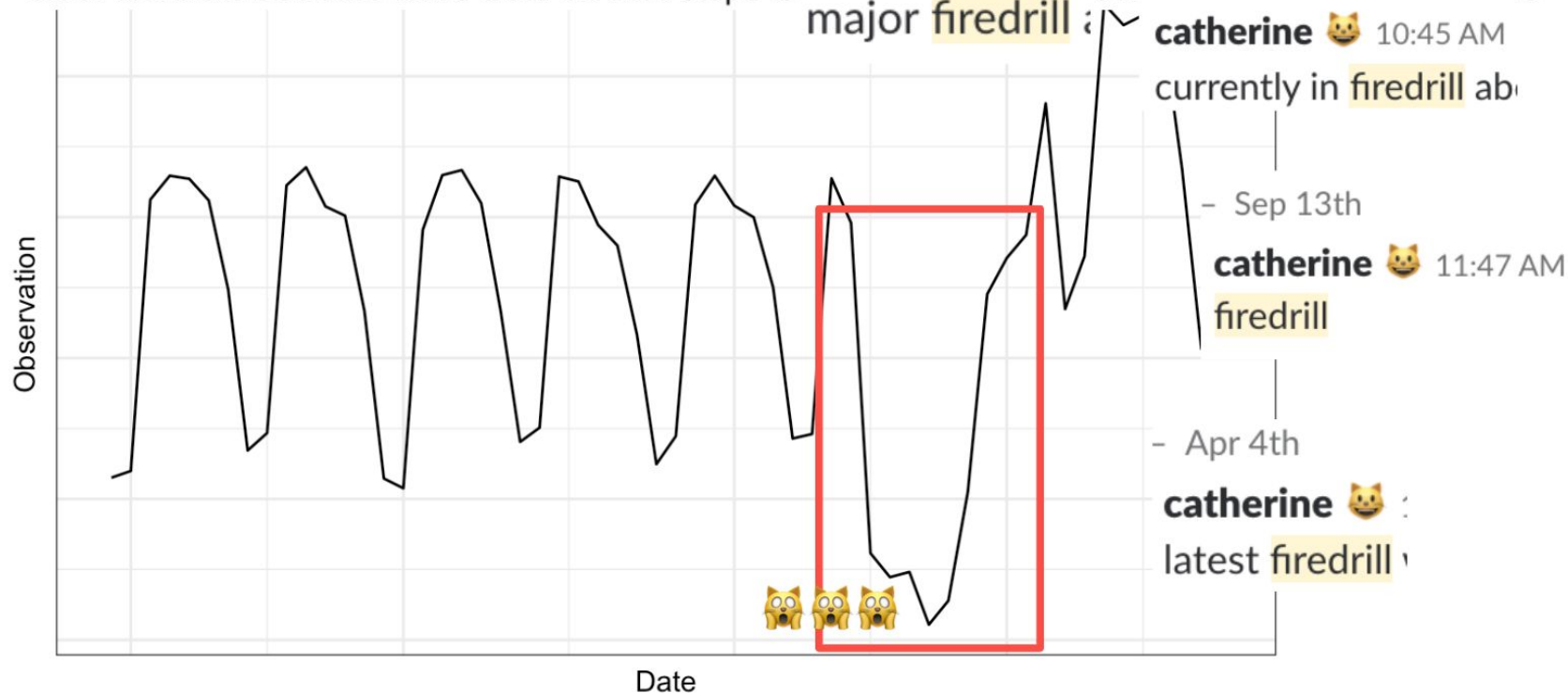
PART TWO

# Dealing with Fire Drills\_



something is **terribly wrong** 🔥🔥🔥

we're calling a **Code Red** investigation. **Code Red** means this is top priority and takes precedence over other tasks at hand until we're clear on next steps for



# DATA SCIENCE FIRE DRILLS

**catherine** 🐱 4:43 PM

my typical workflow:

- 1) start working on an analysis i'm excited about
- 2) fire drill, everything else is derailed
- 3) somehow still working on the fire drill and other related issues
- 4) think longingly about the analysis i was planning to work on

Jared Lander



What sort of anomaly detection?

Detecting anomalies in time series data (webpage visits, empty flights, etc)

Places I've worked have ended up getting more use out of that than brittle forecast models, to be honest

Reduced the number of firedrills

Anomaly detection on key metrics can lead to earlier detection of irregularities and reduce the number of fire drills.

We can be proactive instead of reactive.

# Applications of Anomaly Detection\_

- Fraud Detection
- KPI Monitoring
- Identify Breakage
- Workforce Planning
- Nature (e.g. weather)
- ... and more!

“monitor key metrics, website breakage, and fraudulent activity... we can build a system for anomaly detection to **uncover blind spots in large datasets** and reduce fire drills at work”

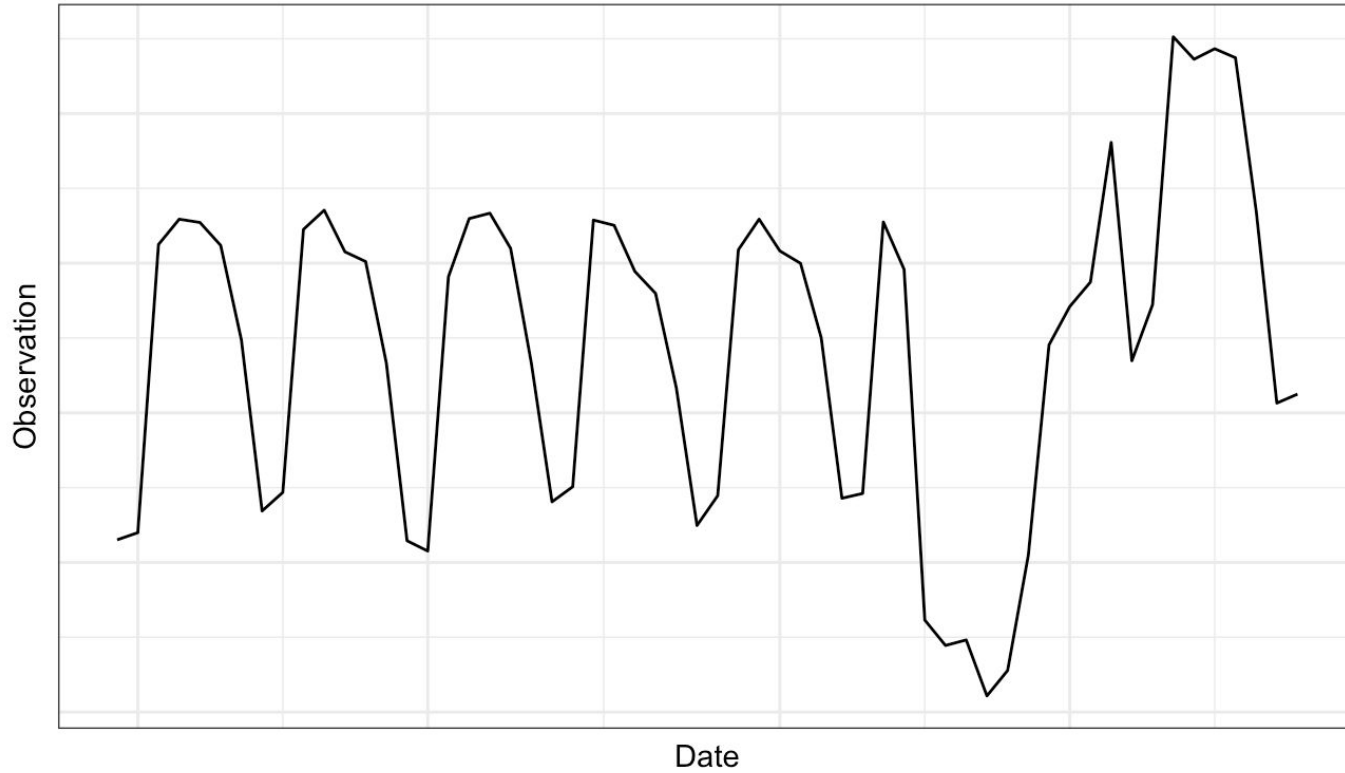
- my DC R promise, stated at the beginning of this talk

PART THREE

# Anomaly Detection\_

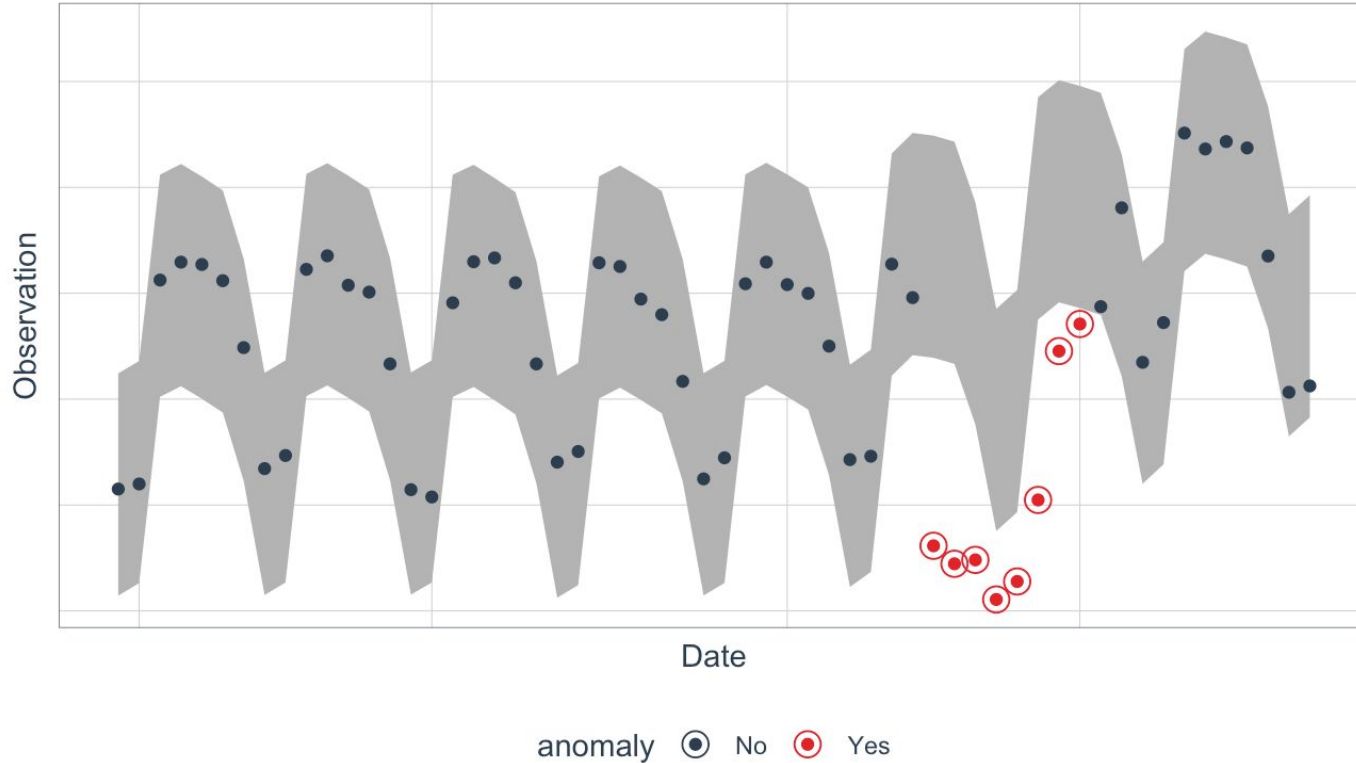
# Before...

Time Series Data



# AFTER!

Time Series With Anomalies Detected



LIVE CODE SESSION

# Let's get started!

Follow along:

twitter @catherinezh

github @cattystats

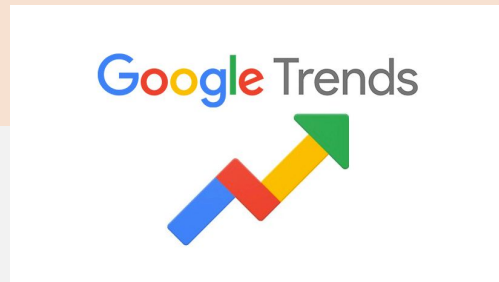
<https://github.com/cattystats/>

Anomaly\_Detection



## 1. CREATE A DATA FRAME

```
#install.packages("gtrendsR")
library(gtrendsR)
google_trends_df = gtrends(
  c("Vote"), #keywords -- start with one
  gprop = "web", #choose: web, news, images, froogle, youtube
  geo = c("US"), #only pull results for US
  time = "2004-01-01 2018-11-08")[[1]] #timeframe
```



```
> as.tibble(google_trends_df)
```

```
# A tibble: 179 x 6
```

	date	hits	keyword	geo	gprop	category
	<dtm>	<int>	<chr>	<chr>	<chr>	<int>
1	2004-01-01 00:00:00	5	Vote	US	web	0
2	2004-02-01 00:00:00	7	Vote	US	web	0
3	2004-03-01 00:00:00	7	Vote	US	web	0
4	2004-04-01 00:00:00	5	Vote	US	web	0
5	2004-05-01 00:00:00	5	Vote	US	web	0
6	2004-06-01 00:00:00	5	Vote	US	web	0
7	2004-07-01 00:00:00	10	Vote	US	web	0
8	2004-08-01 00:00:00	14	Vote	US	web	0
9	2004-09-01 00:00:00	21	Vote	US	web	0
10	2004-10-01 00:00:00	46	Vote	US	web	0

```
# ... with 169 more rows
```

```
>
```

install + load **gtrendsR**:  
choose a keyword that  
interests you

## 2. PREPARE DATA

### install + load tidyverse and anomalize

```
#install.packages("anomalize")
library(tidyverse)
library(anomalize)

google_trends_df_tbl = google_trends_df %>%
  mutate(date=lubridate::ymd(date)) %>%
  tbl_df()
```

### 3. ANOMALIZE!

## anomalize!

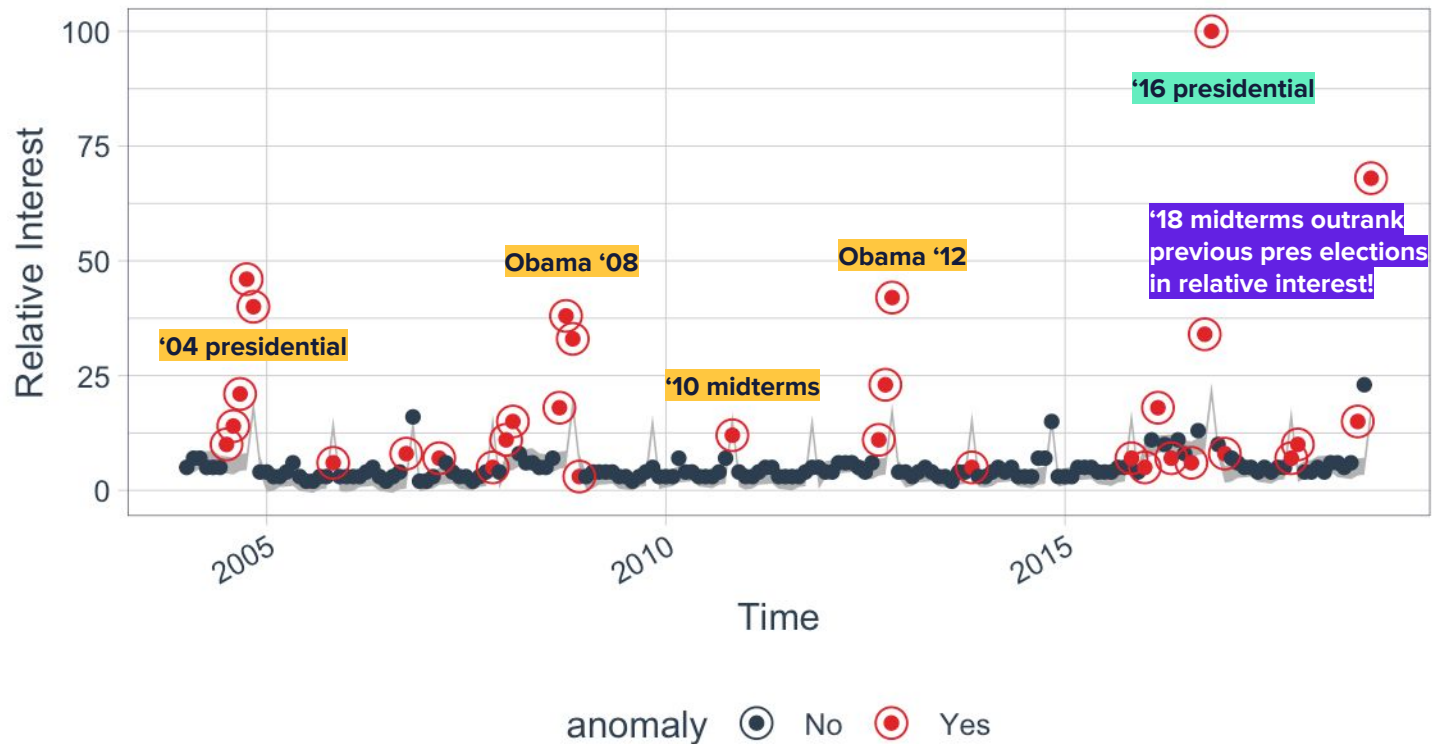
```
google_trends_df_tbl %>% # Twitter and GESD
  time_decompose(hits, method = "twitter", trend = "1 year") %>%
  anomalize(remainder, method = "gesd") %>%
  time_recompose() %>%
  # Anomaly Visualization
  plot_anomalies(time_recomposed = TRUE) +
  labs(title = "Google Trends Data - Twitter + GESD
Method", x="Time", y="Relative Interest", subtitle = "United States search volume
for 'Vote' between Jan'04-Nov'18"
  )
```

### 3. ANOMALIZE... TADA!

**KEYWORD: VOTE**

#### Google Trends Data - Twitter + GESD Method

United States search volume for 'Vote' between Jan'04-Nov'18



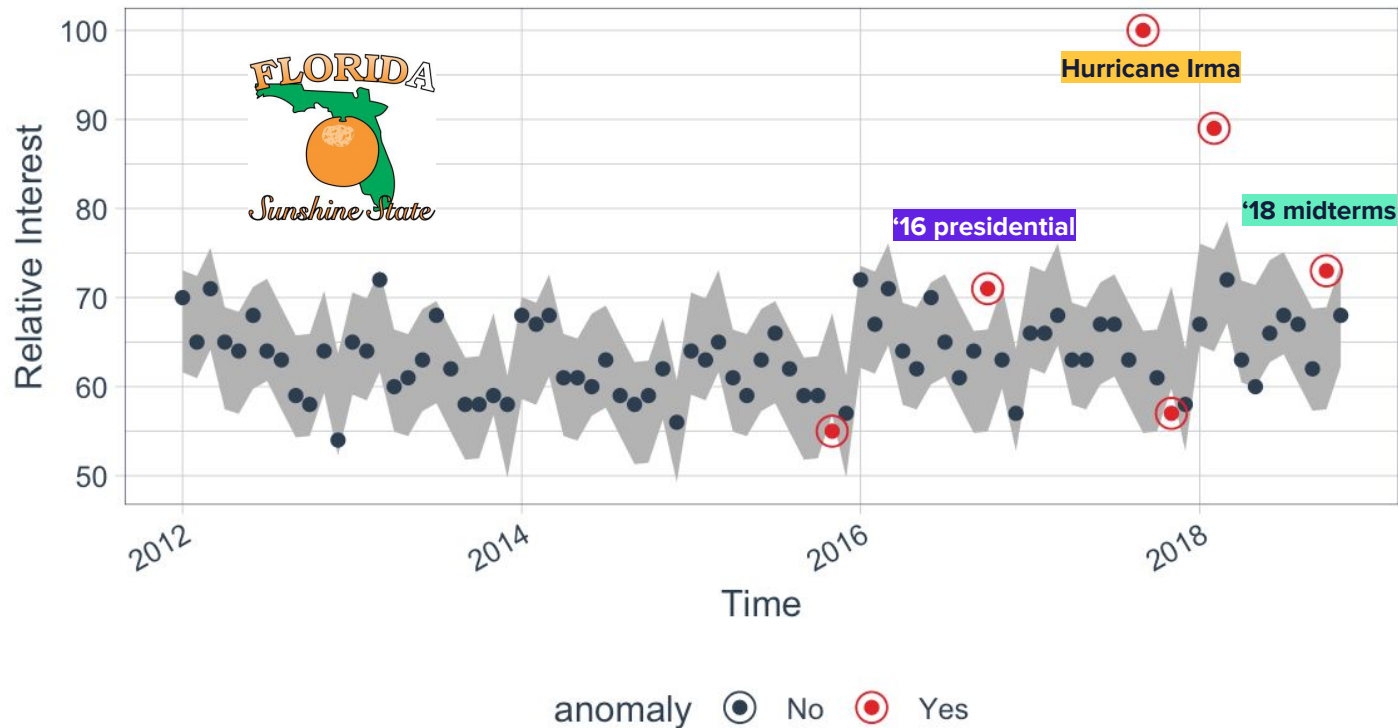
LET'S TRY THIS WITH...

KEYWORD: FLORIDA

## Google Trends Data - Twitter + GESD Method

United States search volume for 'Florida' between Jan'12-Nov'18

Florida: The swingiest swing state

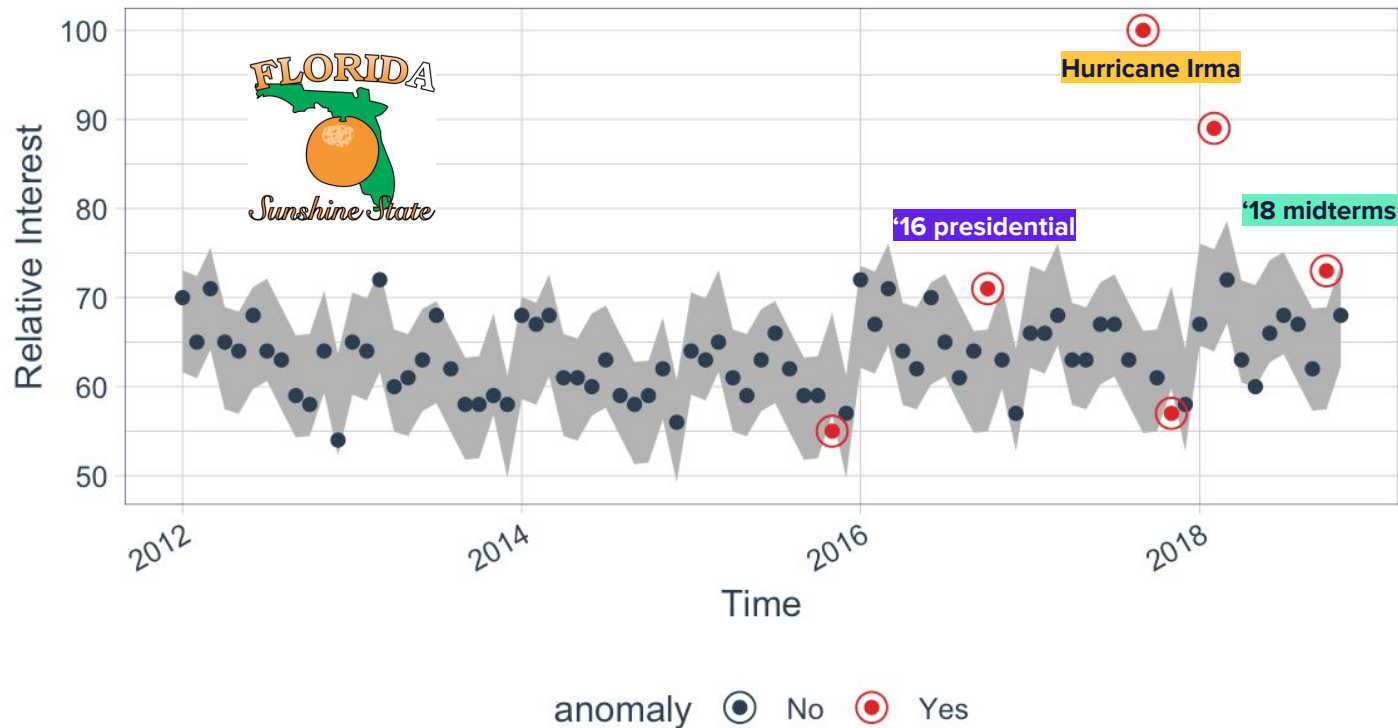


LET'S TRY THIS WITH...

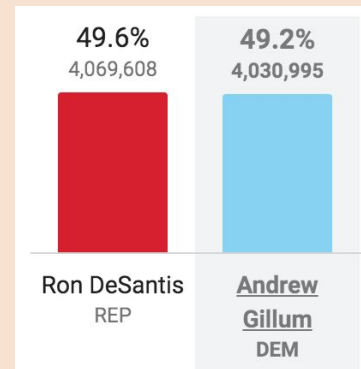
KEYWORD: FLORIDA

## Google Trends Data - Twitter + GESD Method

United States search volume for 'Florida' between Jan'12-Nov'18



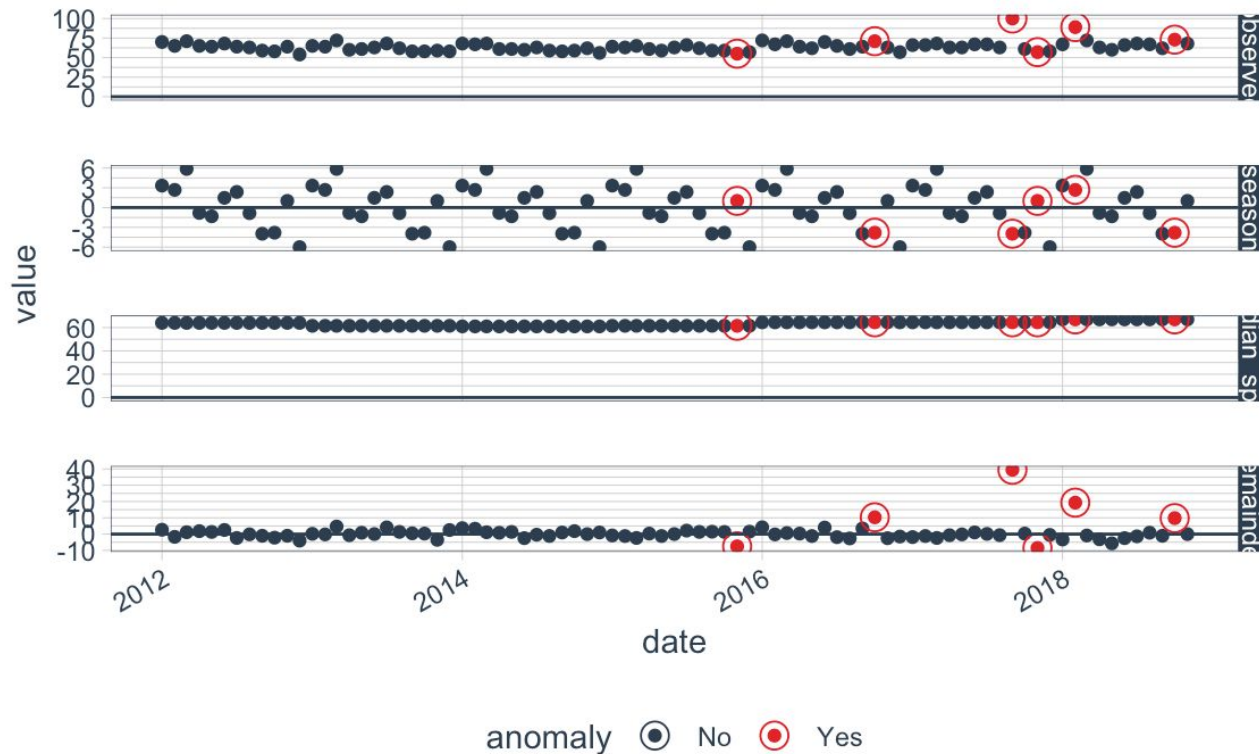
Florida: The swingiest swing state





## KEYWORD: FLORIDA

```
google_trends_df_tbl %>%  
  time_decompose(hits, method = "twitter",  
                 frequency = "1 year", trend = "1 year") %>%  
  anomalise(remainder, method = "gesd", alpha = 0.05, max_anoms = 0.2) %>%  
  plot_anomaly_decomposition()
```



**plot\_anomaly\_decomposition()**

visualize inner workings  
of how algorithm detects  
anomalies in the  
“remainder”

#### 4. EXPLORE METHODS BASED ON TIME SERIES ATTRIBUTES

**anomalize** cheat sheet:

**Twitter + GESD** better for highly seasonal data

**STL + IQR** if seasonality is not a major factor

adjust **trend period** using domain knowledge

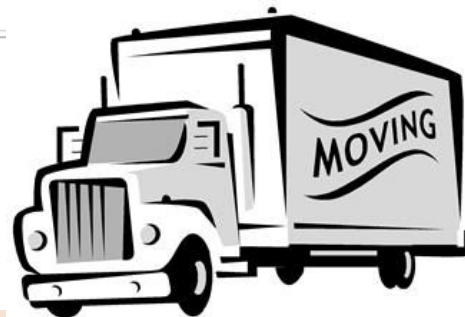
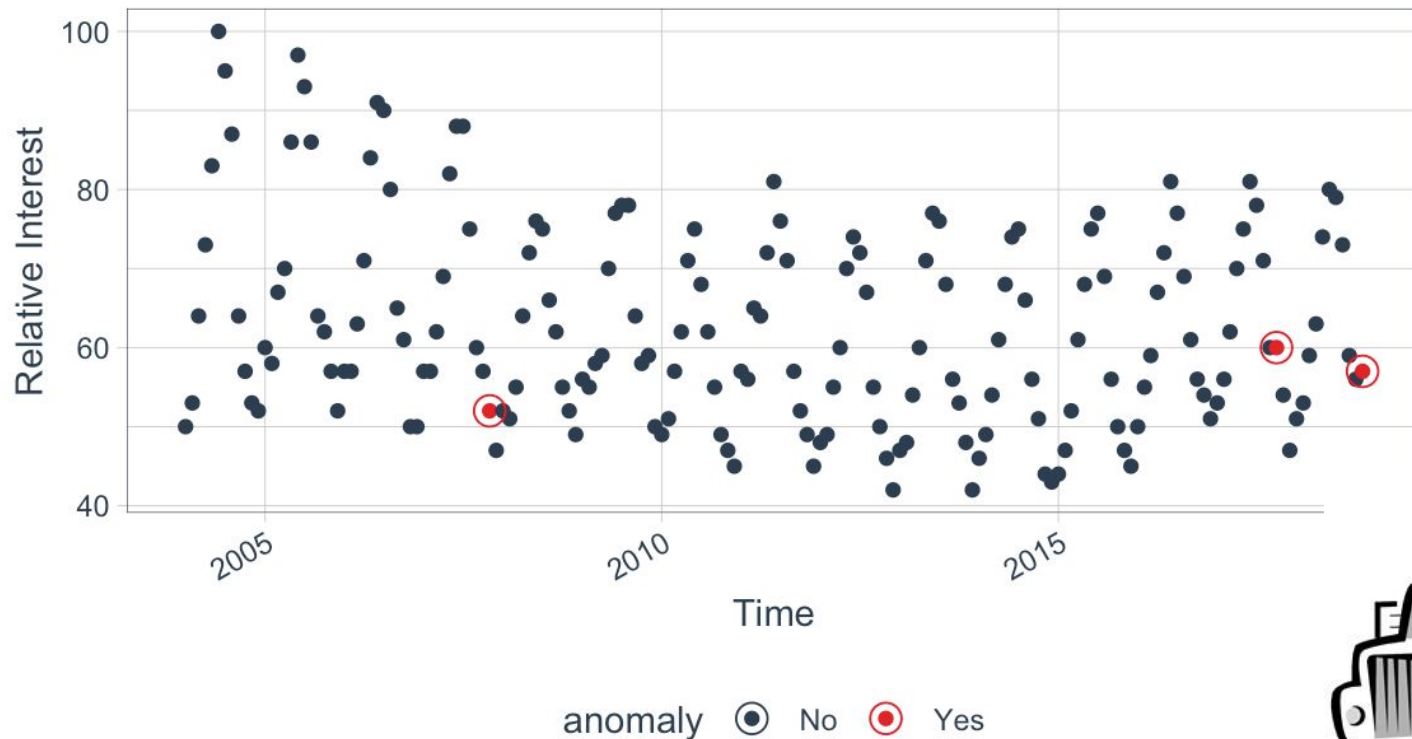


STL + IQR

KEYWORD: MOVERS

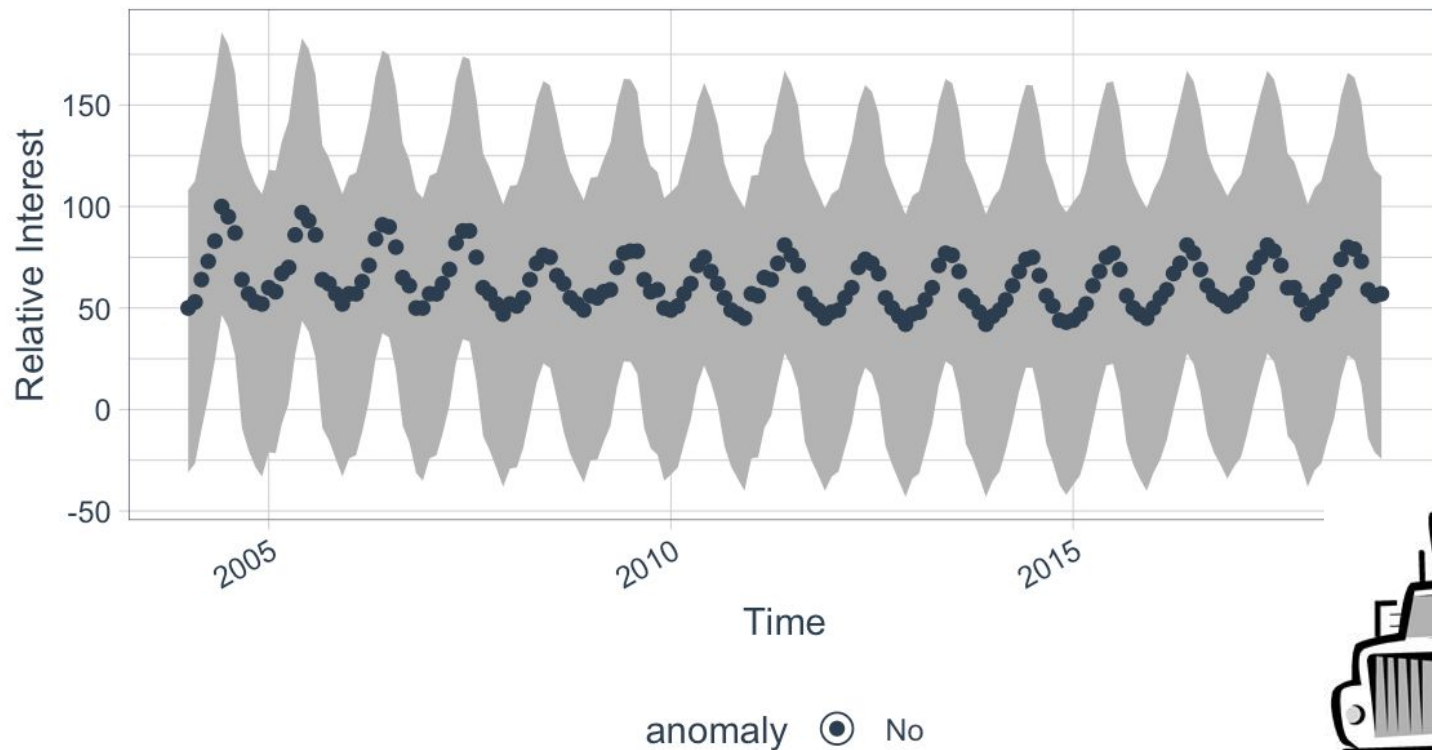
## Google Trends Data - STL + IQR Method

United States search volume for 'Movers' between Jan'05-Nov'18



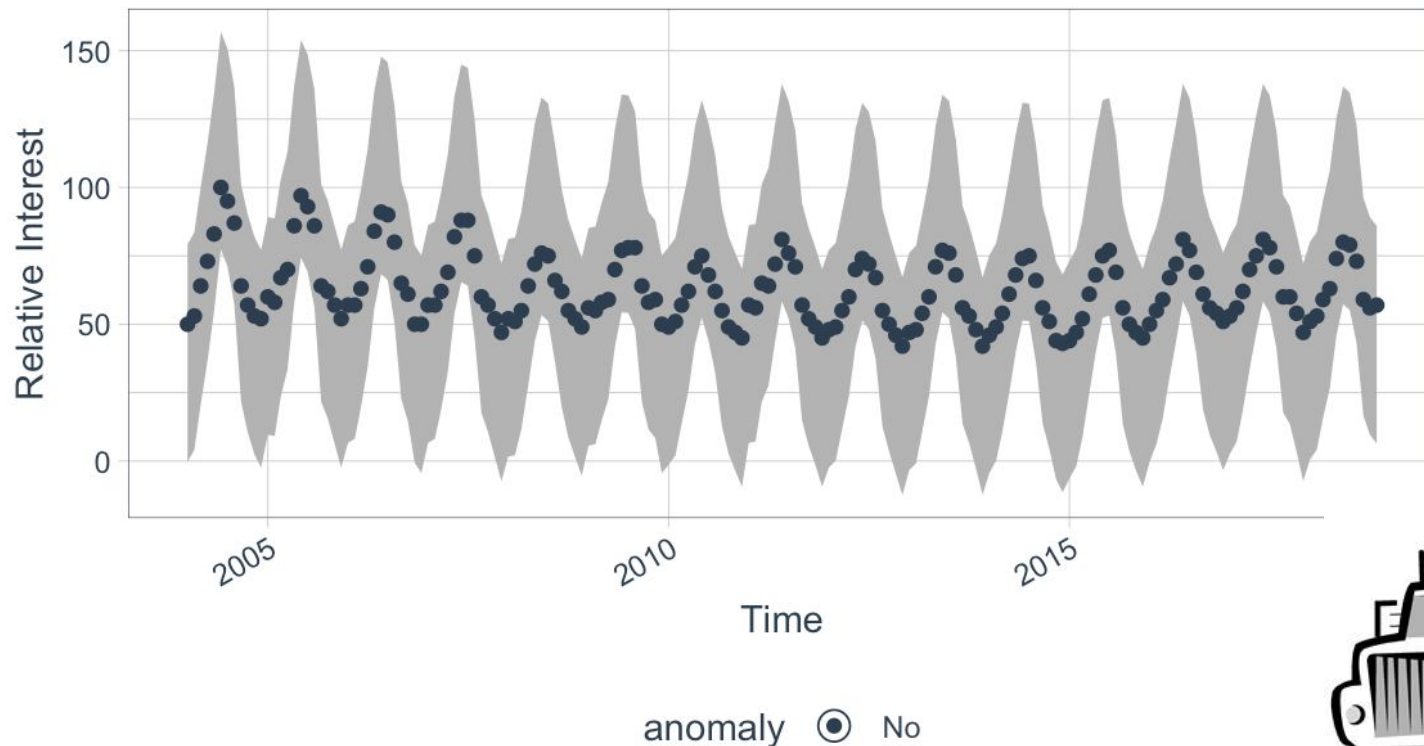
## Google Trends Data - Twitter + IQR Method

United States search volume for 'Movers' between Jan'05-Nov'18



## Google Trends Data - Twitter + GESD Method

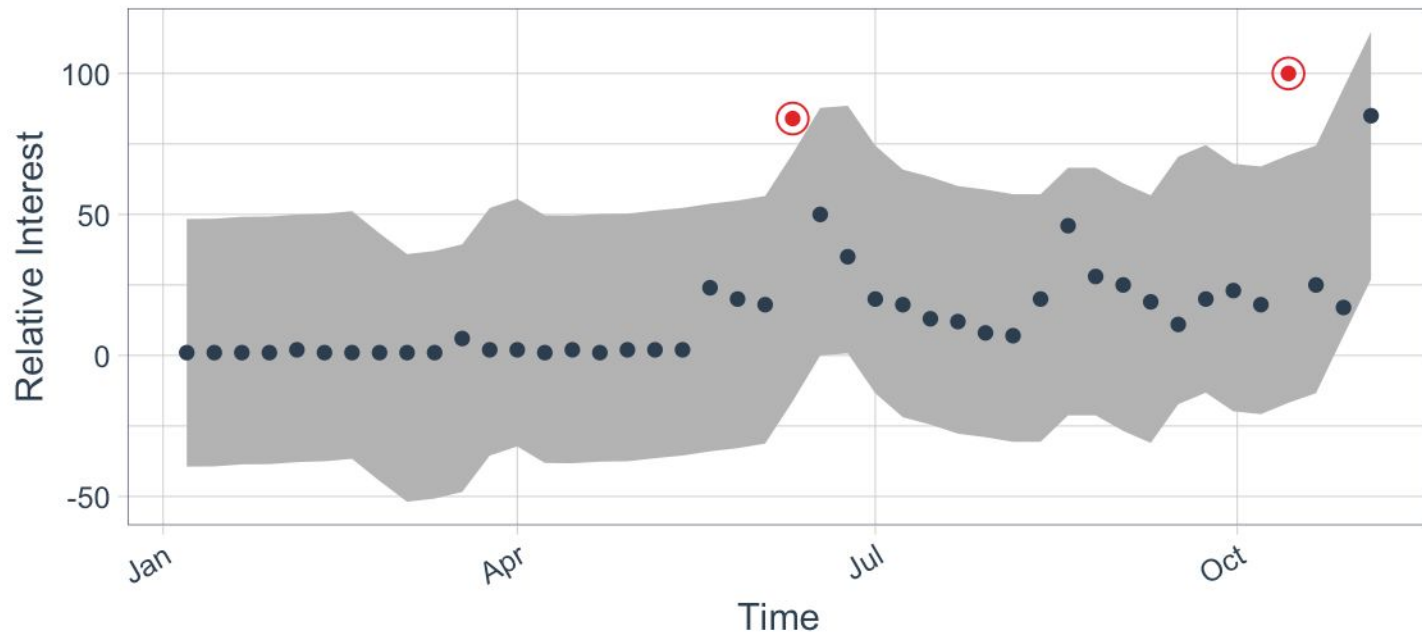
United States search volume for 'Movers' between Jan'05-Nov'18



**TRY THIS ON DIFFERENT KEYWORDS**

## Google Trends Data - STL + IQR Method

United States search volume for 'Pete Davidson' between Jan-Nov'18

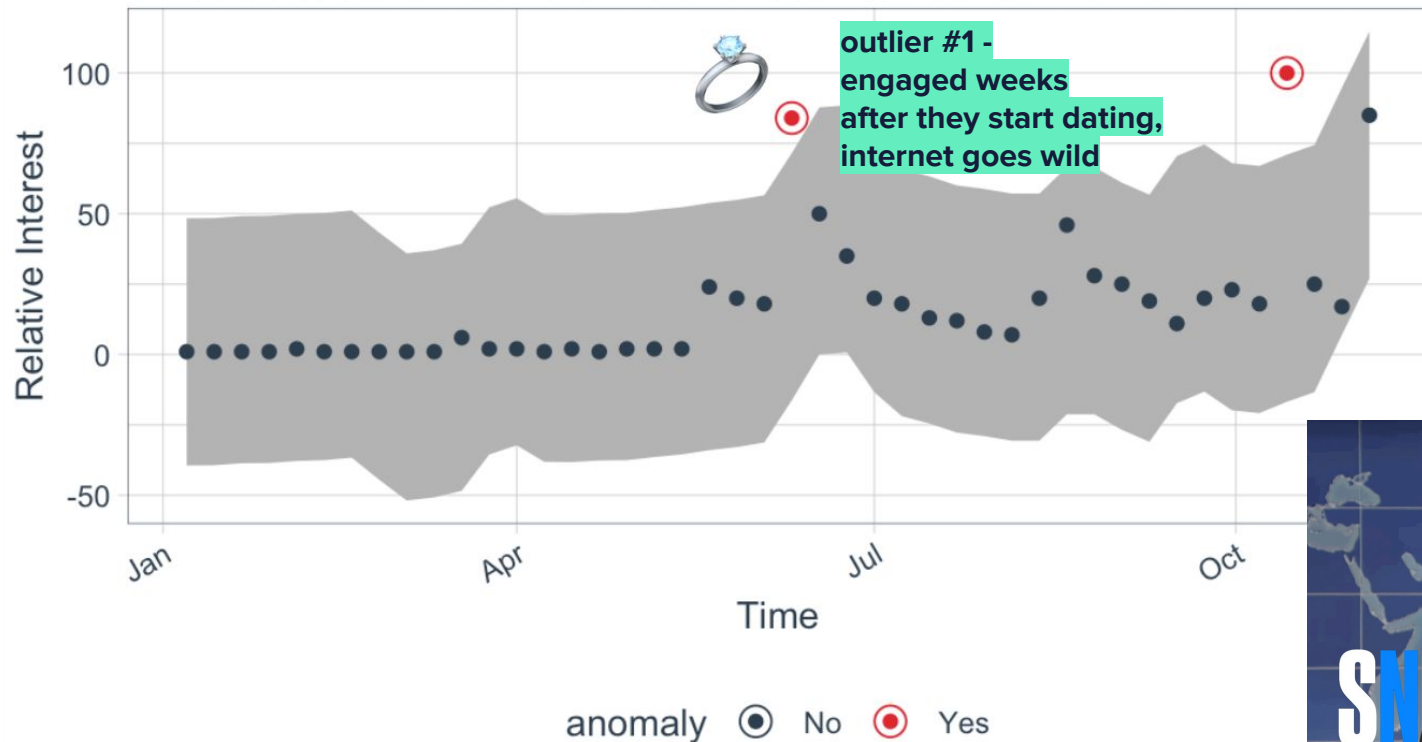


recent news, and  
2018 data only --  
seasonality is not  
really a factor, so  
we go back to  
using STL + IQR

## PETE DAVIDSON

### Google Trends Data - STL + IQR Method

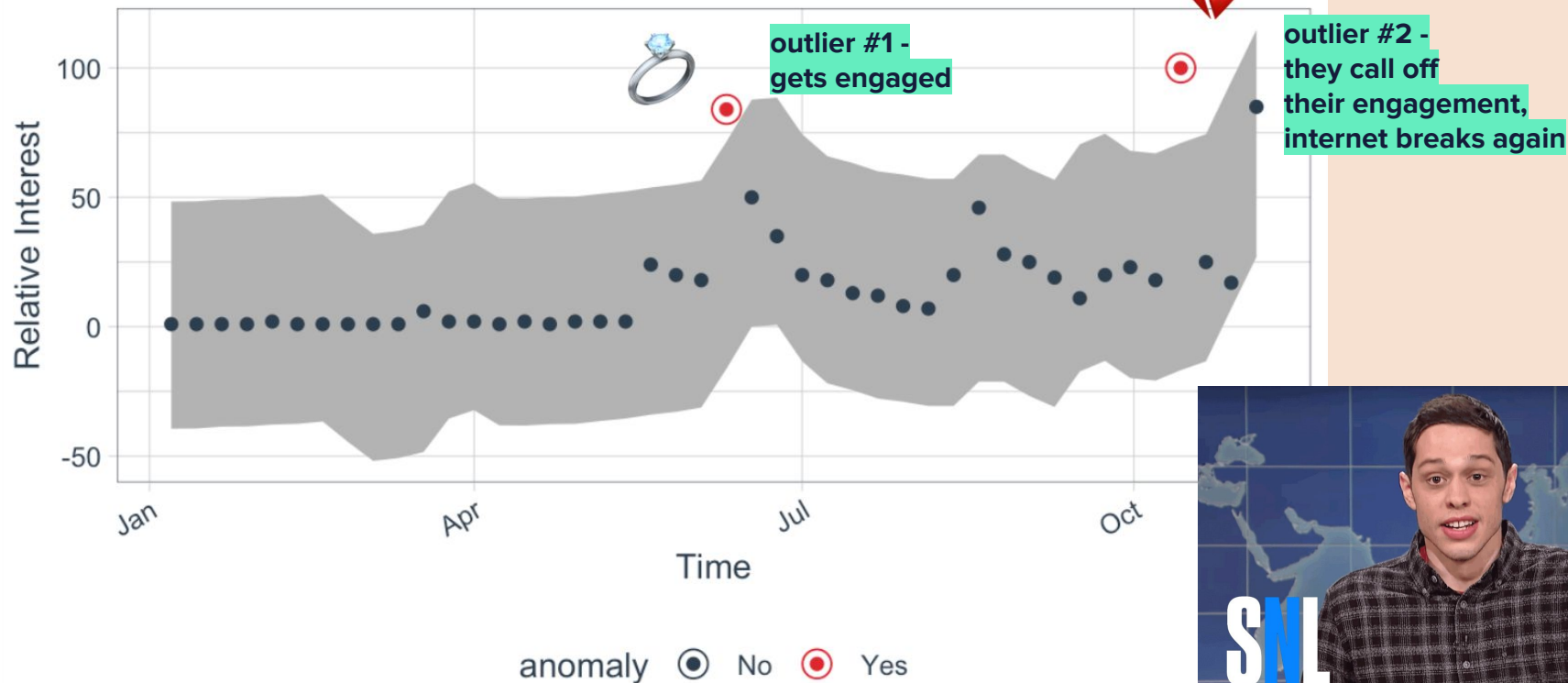
United States search volume for 'Pete Davidson' between Jan-Nov'18



## PETE DAVIDSON

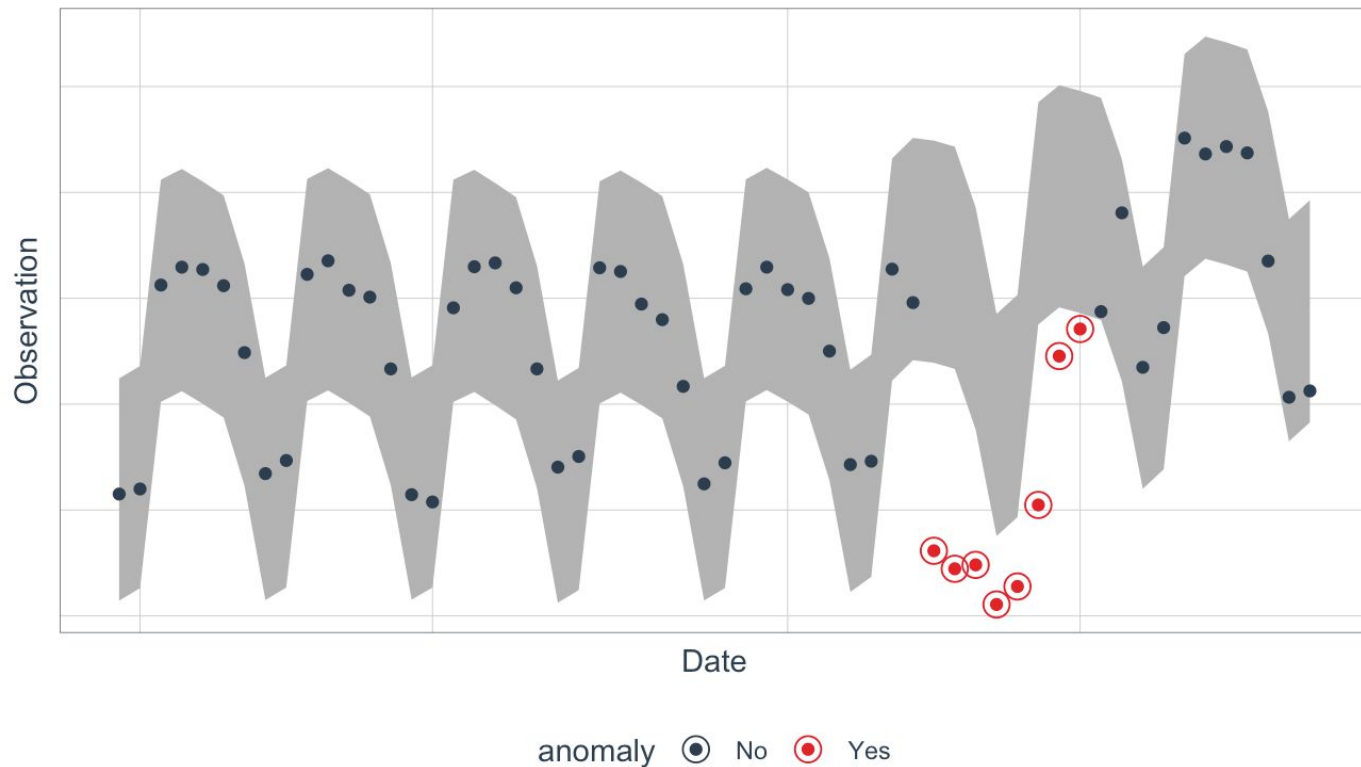
### Google Trends Data - STL + IQR Method

United States search volume for 'Pete Davidson' between Jan-Nov'18



## TRY THIS AT HOME!

Time Series With Anomalies Detected



## Keywords To Try:

- NBA
- Weather
- Politicians
- Elon Musk
- Holidays
- Memes
- Events

... and anything else you might think of!



# Additional Resources

- [R Code + Notebook](#)
- [Introducing Anomalize](#)
- [Github: Anomalize](#)
- [Codecademy](#)

**GOOD LUCK AND HAVE FUN!**

twitter @catherinezh

github @cattystats

#rstats

#rstatsdc

#rladies