

# The Data Science Process



The Data Scientist's Toolbox



# Not So Standard Deviations

A statistics (etc.) blog by Hilary Parker



 Search

About Me

Contact

Posted on January 30, 2013

[← Previous](#)

[Next →](#)

## Hilary: the most poisoned baby name in US history

I've always had a special fondness for my name, which — according to Ryan Gosling in "Lars and the Real Girl" — is a scientific fact for most people (Ryan Gosling constitutes scientific proof in my book). Plus, the root word for **Hilary** is the Latin word "hilarious" meaning cheerful and merry, which is the same root word for "hilarious" and "exhilarating." It's a great name.

Several years ago I came across [this blog post](#), which provides a cursory analysis for why "Hillary" is the most poisoned name of all time. The author is careful not to comment on the details of why "Hillary" may have been poisoned right around 1992, but I'll go ahead and make the bold causal conclusion that it's because that was the year that Bill Clinton was elected, and thus the year Hillary Clinton entered the public sphere and was generally reviled for [not wanting to bake cookies](#) or something like that. Note that this all happened when I was 7 years old, so I spent the formative years of 7-15 being called "Hillary Clinton" whenever I introduced myself. Luckily, I was a feisty feminist from a young age and rejoiced in the comparison (and [life is not about being popular](#)).

In the original post the author bemoans the lack of research assistants to perform his data extraction for a more complete analysis. Fortunately, in this era we have replaced human jobs with computers, and the data can be easily extracted using programming. This weekend I took the opportunity to learn how to scrape the social security data myself and do a more complete analysis of all of the names on record.

**Is Hilary/Hillary really the most rapidly poisoned name in recorded American history? An analysis.**

<https://hilaryparker.com/2013/01/30/hilary-the-most-poisoned-baby-name-in-us-history/>

# Not So Standard Deviations

A statistics (etc.) blog by Hilary Parker



Search

The Data Science  
Question being  
answered

About Me

Contact

Posted on January 30, 2013

[← Previous](#) [Next →](#)

## Hilary: the most poisoned baby name in US history

I've always had a special fondness for my name, which — according to Ryan Gosling in "Lars and the Real Girl" — is a scientific fact for most people (Ryan Gosling constitutes scientific proof in my book). Plus, the root word for **Hilary** is the Latin word "hilarious" meaning cheerful and merry, which is the same root word for "hilarious" and "exhilarating." It's a great name.

Several years ago I came across [this blog post](#), which provides a cursory analysis for why "Hillary" is the most poisoned name of all time. The author is careful not to comment on the details of why "Hillary" may have been poisoned right around 1992, but I'll go ahead and make the bold causal conclusion that it's because that was the year that Bill Clinton was elected, and thus the year Hillary Clinton entered the public sphere and was generally reviled for **not wanting to bake cookies** or something like that. Note that this all happened when I was 7 years old, so I spent the formative years of 7-15 being called "Hillary Clinton" whenever I introduced myself. Luckily, I was a feisty feminist from a young age and rejoiced in the comparison (and **life is not about being popular**).

In the original post the author bemoans the lack of research assistants to perform his data extraction for a more complete analysis. Fortunately, in this era we have replaced human jobs with computers, and the data can be easily extracted using programming. This weekend I took the opportunity to learn how to scrape the social security data myself and do a more complete analysis of all of the names on record.

**Is Hilary/Hillary really the most rapidly poisoned name in recorded American history? An analysis.**

## Is Hilary/Hillary really the most rapidly poisoned name in recorded American history? An analysis.

I will follow up this post with more details on how to perform web-scraping with R (for this I am infinitely indebted to my friend Mark — check out his [storyboard project](#) and be amazed!). For now, suffice it to say that I was able to collect from the [social security website](#) the data for every year between 1880 and 2011 for the 1000 most popular baby names. For each of the 1000 names in a given year, I collected the raw number of babies given that name, as well as the percentage of babies given that name, and the rank of that name. For girls, this resulted in 4110 total names.

The Data!



The code is available!

## Analysis of most poisoned names in US

6 commits

1 branch

0 releases

1 contributor

Branch: master

New pull request

Create new file

Upload files

Find file

Clone or download



hilaryparker added 2014 analysis for R Meetup

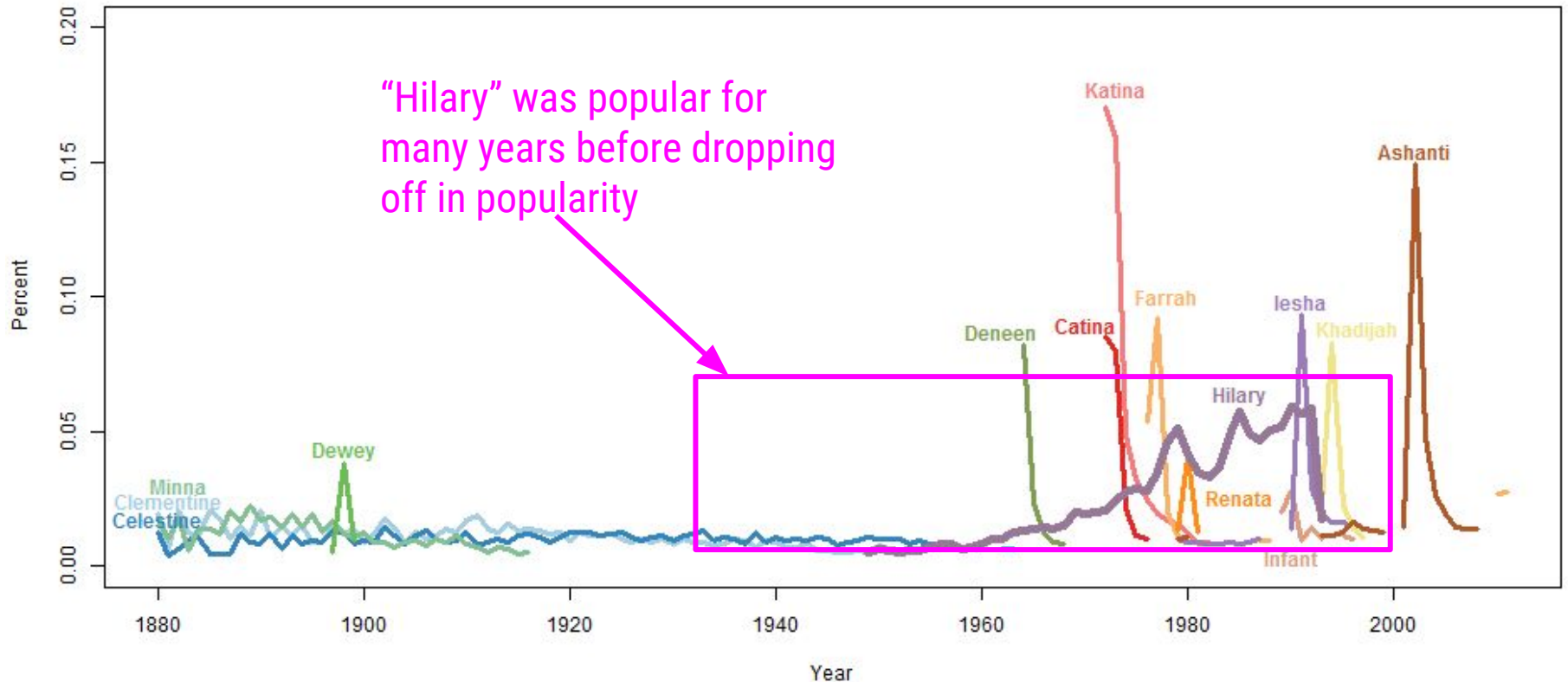
Latest commit bb1ab7e on Nov 7, 2016

NYCR_hillary_2014	added 2014 analysis for R Meetup	2 years ago
NYCR_hillary_2015	added 2014 analysis for R Meetup	2 years ago
cache	initial commit	5 years ago
config	changes for Strata ignite talk	4 years ago
graphs	changes for Strata ignite talk	4 years ago
lib	initial commit	5 years ago
munge	initial commit	5 years ago
reports	initial commit	5 years ago
src	changes for Strata ignite talk	4 years ago
.gitattributes	initial commit	5 years ago
citizens	initial commit	5 years ago



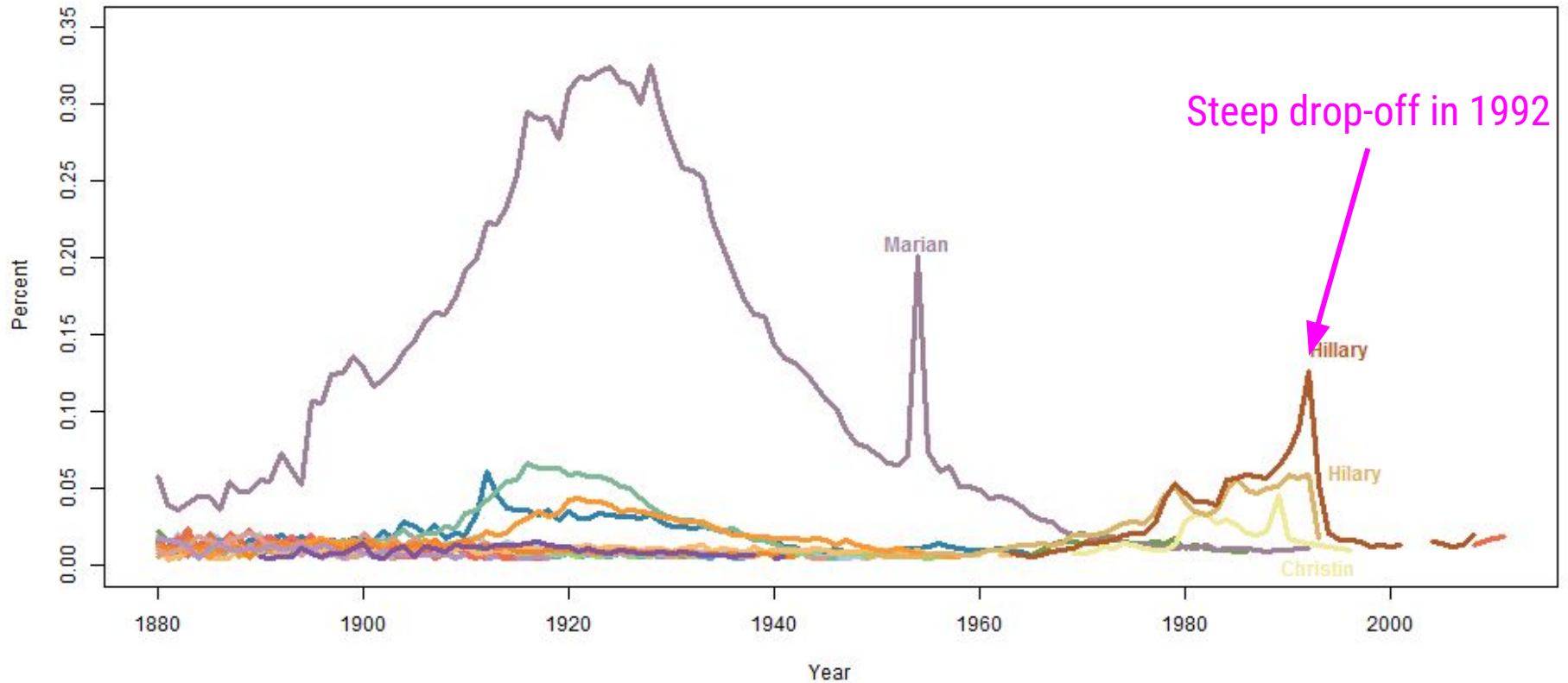
Name	Loss (%)	Year
Farrah	78	1978
Dewey	74	1899
Catina	74	1974
Deneen	72	1965
Khadijah	72	1995
Hilary	70	1993
Clementine	69	1881
Katina	69	1974
Renata	69	1981
Ilesha	69	1992
Minna	68	1883
Ashanti	68	2003
Celestine	67	1881
Infant	67	1991

Percent of baby girls given a name over time for the 14 most poisoned names





Percent of baby girls given a name over time for the 39 most poisoned names, controlling for fads



# Not So Standard Deviations

A statistics (etc.) blog by Hilary Parker



 Search

About Me

Contact

Posted on January 30, 2013

[← Previous](#)

[Next →](#)

## Hilary: the most poisoned baby name in US history

I've always had a special fondness for my name, which — according to Ryan Gosling in "Lars and the Real Girl" — is a scientific fact for most people (Ryan Gosling constitutes scientific proof in my book). Plus, the root word for **Hilary** is the Latin word "hilarious" meaning cheerful and merry, which is the same root word for "hilarious" and "exhilarating." It's a great name.

Several years ago I came across [this blog post](#), which provides a cursory analysis for why "Hillary" is the most poisoned name of all time. The author is careful not to comment on the details of why "Hillary" may have been poisoned right around 1992, but I'll go ahead and make the bold causal conclusion that it's because that was the year that Bill Clinton was elected, and thus the year Hillary Clinton entered the public sphere and was generally reviled for [not wanting to bake cookies](#) or something like that. Note that this all happened when I was 7 years old, so I spent the formative years of 7-15 being called "Hillary Clinton" whenever I introduced myself. Luckily, I was a feisty feminist from a young age and rejoiced in the comparison (and [life is not about being popular](#)).

In the original post the author bemoans the lack of research assistants to perform his data extraction for a more complete analysis. Fortunately, in this era we have replaced human jobs with computers, and the data can be easily extracted using programming. This weekend I took the opportunity to learn how to scrape the social security data myself and do a more complete analysis of all of the names on record.

**Is Hilary/Hillary really the most rapidly poisoned name in recorded American history? An analysis.**

<https://hilaryparker.com/2013/01/30/hilary-the-most-poisoned-baby-name-in-us-history/>

## How Hilary Parker gave credit to others' ideas and work:

- Linking to a [blog post](#) a similar question had been asked previously
- Linking to the [Social Security website](#) where she got the data
- Linking to a link about where Hilary [learned about web scraping](#)



## 1. Introduction

### 2. Exploratory Analysis

### 3. Model Building

### 4. Data Source Appendix

### 5. Feature Appendix

### 6. Data Wrangling Appendix

### 7. Data Visualization Appendix

### 8. Modeling Appendix

# Predicting Spatial Risk of Opioid Overdoses in Providence, RI

*Jordan Butz and Annie Streetman*

*May 3, 2018*

## 1. Introduction

### 1.1 How to Use This Document

This project was produced as part of the University of Pennsylvania Master of Urban Spatial Analytics Spring 2018 Practicum (MUSA 801), instructed by Ken Steif, Michael Fichman, and Karl Dailey. This document begins with a case study of predicting spatial risk of opioid overdoses in Providence, Rhode Island and is followed by a series of appendices that discuss [data wrangling](#), [data visualization](#), [data sources](#), [feature engineering](#), and [model results](#). Navigate through the document either by using the panel at the left, or by clicking the hyperlinks throughout the document.

### 1.2 Abstract

This project seeks to build a spatial risk model of opioid overdose events for the City of Providence, Rhode Island by examining current overdose locations, community protective resources, risk factors, and neighborhood characteristics. Assigning a level of risk to each area of the city can assist Providence and local stakeholders in strategically allocating resources in a way that will achieve the greatest impact. As of January 2018, Providence is implementing a Safe Stations program, where people struggling with substance abuse can come to any of the City's 12 fire stations to be connected with supportive services. The spatial risk model will help Providence's Department of Healthy Communities determine other areas at high risk of overdose events where the City could site additional interventions or supplement their communications efforts.



# Maëlle

Home

About

Bio

R packages

Publications

-  Twitter
-  GitHub
-  ResearchGate
-  Email

© 2018. All rights reserved.

## Where to live in the US

16 Nov 2017

I was fascinated by this [xkcd comic](#) about where to live based on your temperature preferences. I also thought it'd be fun to try to make a similar one from my R session! Since I'm no meteorologist and was a bit unsure of how to define winter and summer, and of their relevance in countries like, say, India which has monsoon, I decided to focus on a single country located in one hemisphere only and big enough to offer some variety... the USA! So, dear American readers, where should you live based on your temperature preferences?

## Defining data sources

### Weather data

The data for the original xkcd graph comes from [weatherbase](#). I changed sources because 1) I was not patient enough to wait for weatherbase to send me a custom dataset which I imagine is what xkcd author did and 2) I'm the creator of [a cool package accessing airport weather data](#) for the whole world including the US! My package is called "riem" like "R Iowa Environmental Mesonet" (the source of the data, a [fantastic website](#)) and "we laugh" in Catalan (at the time I wrote the package





## David Robinson

Chief Data Scientist at  
DataCamp, works in R and  
Python.

- Email
- Twitter
- Github
- Stack Overflow

### Subscribe

### Recommended Blogs

- DataCamp
- R Bloggers
- RStudio Blog

## Text analysis of Trump's tweets confirms he writes only the (angrier) Android half

I don't normally post about politics (I'm not particularly savvy about polling, which is where data science has had the largest impact on politics). But this weekend I saw a hypothesis about Donald Trump's twitter account that simply begged to be investigated with data:



Todd Vaziri

@tvaziri



Every non-hyperbolic tweet is from iPhone (his staff).

Every hyperbolic tweet is from Android (from him).

3:20 PM - Aug 6, 2016

14.2K 10.4K people are talking about this



## Tidying and mapping Toronto open data

Dec 27, 2017

According to the 2017 [Open Cities Index Results](#), the city of Toronto ranks second in Canada in terms of open data maturity. With 250+ data sets, this initiative makes it easy to access information on business, culture, environment, finance, health, parks + rec, public safety, transportation, and more.

I was curious to poke around open data and learn something new along the way, both in terms of Toronto and the R ecosystem. I'd never done any sort of mapping in R so chose something that could be visualized easily and was interesting to me, and chose to look into the data on Toronto Public Health's sexual health clinics, available [here](#).

According to the site, it includes: clinic name, location (i.e., intersection), address, contact number, drop-in and appointment hours, and services provided by each clinic. Since some of these attributes are plural (two kinds of hours and *services*, as in more than one service), I anticipated that the data wouldn't be in the tidy format (each observation is one row, and each column is a variable) that is preferable for easy manipulation and plotting. And I was right!

Clinic Name	Clinic Location	Address	Contact Number	Operational Hours		Services
				Drop-in	By Appointment	
Birth Control and Sexual Health Centre	Dufferin St./Lawrence Ave. W.	Suite 403, 960 Lawrence Ave. W., Toronto, On M6A 3B5	416-789-4541		Monday: 4 pm - 6 pm Tuesday: 4 pm - 7 pm Wednesday: 12 noon - 5 pm Thursday: 5 pm - 8 pm Friday: 12:30 pm - 4:30 pm	Birth control counselling Low cost or free birth control Free condoms Plan B (emergency contraceptive pill) STI testing and free treatment HIV testing Pregnancy testing, counselling and referral Sexuality and relationship counselling Anonymous HIV testing (including the rapid HIV test)
Names/addresses across multiple lines						
Black Creek Community Health Centre (Sheridan Mall Site)	Jane St./Wilson Ave.	North York Sheridan Mall, Unit 5 2202 Jane St., Toronto, On M2M 1A4	416-249-8000	Tuesday: 4 pm - 6 pm		Birth control counselling Low cost or free birth control Free condoms Plan B (emergency contraceptive pill) STI testing and free treatment HIV testing Pregnancy testing, counselling and referral Sexuality and relationship counselling
Black Creek Community Health Centre (Yorkgate mall Site)	Jane St./Finch Ave. W.	1 York Gate Blvd., Toronto, On M3N 3A1	416-246-2388	Wednesday: 1 pm - 4:30 pm		Birth control counselling Low cost or free birth control Free condoms Plan B (emergency contraceptive pill)

# Summarizing: The Data Science Process



The Data Scientist's Toolbox