The Data Science Process

In the first few lessons of this course we discussed what data and data science are, and ways to get help. What we haven't yet covered is what an actual data science project looks like. To do so, we'll first step through an actual data science project, breaking down the parts of a typical project and then provide a number of links to other interesting data science projects. Our goal in this lesson is to expose you to the process one goes through as they carry out data science projects.

The Parts of a Data Science Project

Every Data Science Project starts with a question that is to be answered with data. That means that **forming the question** is an important first step in the process. The second step is **finding or generating the data** you're going to use to answer that question. With the question solidified and data in hand, the **data are then analyzed**, first by **exploring the data** and then often by **modeling the data**, which means using some statistical or machine learning techniques to analyze the data and answer your question. After drawing conclusions from this analysis, the project has to be **communicated to others**. Sometimes this is a report you send to your boss or team at work. Other times it's a blog post. Often it's a presentation to a group of colleagues. Regardless, a data science project almost always involves some form of communication of the projects' findings. We'll walk through these steps using a data science project example below.

A Data Science Project Example

For this example, we're going to use an example analysis from a data scientist named Hilary Parker. Her work can be found on her blog, and the specific project we'll be working through here is from 2013 and titled "Hilary: the most poisoned baby name in US history". To get the most out of this lesson, click on that link and read through Hilary's post. Once you're done, come on back to this lesson and read through the breakdown of this post.

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

Hilary's blog post

The Question

When setting out on a data science project, it's always great to have your question well-defined. Additional questions may pop up as you do the analysis, but knowing what you want to answer with your analysis is a really important first step. Hilary Parker's question is included in bold in her post. Highlighting this makes it clear that

she's interested in answer the following question:

> Is Hilary/Hillary really the most rapidly poisoned name in recorded American history?

The Data

To answer this question, Hilary collected data from the Social Security website. This dataset included the 1,000 most popular baby names from 1880 until 2011.

Data Analysis

As explained in the blog post, Hilary was interested in calculating the relative risk for each of the 4,110 different names in her dataset from one year to the next from 1880 to 2011. By hand, this would be a nightmare. Thankfully, by writing code in R, all of which is available on GitHub, Hilary was able to generate these values for all these names across all these years. It's not important at this point in time to fully understand what a relative risk calculation is (although Hilary does a *great* job breaking it down in her post!), but it is important to know that after getting the data together, the next step is figuring out what you need to do with that data in order to answer your question. For Hilary's question, calculating the relative risk for each name from one year to the next from 1880 to 2011 and looking at the percentage of babies named each name in a particular year would be what she needed to do to answer her question.

https://github.com/hilaryparker/names

Hilary's GitHub repo for this project

Exploratory Data Analysis

What you don't see in the blog post is all of the code Hilary wrote to get the data from the Social Security website, to get it in the format she needed to do the analysis, and to generate the figures. As mentioned above, she made all this code available on GitHub so that others could see what she did and repeat her steps if they wanted. In addition to this code, data science projects often involve writing a lot of code and generating a lot of figures that aren't included in your final results. This is part of the data science process too. Figuring out *how* to do what you want to do to answer your question of interest is part of the process, doesn't always show up in your final project, and can be very time-consuming.

Data Analysis Results

That said, given that Hilary now had the necessary values calculated, she began to analyze the data. The first thing she did was look at the names with the biggest drop in percentage from one year to the next. By this preliminary analysis, Hilary was sixth on the list, meaning there were five other names that had had a single year drop in popularity larger than the one the name "Hilary" experienced from 1992 to 1993.

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
Iesha	69	1992
Minna	68	1883
Ashanti	68	2003
Celestine	67	1881
Infant	67	1991

Biggest Drop Table

In looking at the results of this analysis, the first five years appeared peculiar to Hilary Parker. (It's always good to consider whether or not the results were what you were expecting, from any analysis!) None of them seemed to be names that were popular for long periods of time. To see if this hunch was true, Hilary plotted the percent of babies born each year with each of the names from this table. What she found was that, among these "poisoned" names (names that experienced a big drop from one year to the next in popularity), all of the names other than Hilary became popular all of a sudden and then dropped off in popularity. Hilary Parker was able to figure out why most of these other names became popular, so definitely read that section of her post! The name, Hilary, however, was different. It was popular for a while and then completely dropped off in popularity.

Percent of baby girls given a name over time for the 14 most poisoned names
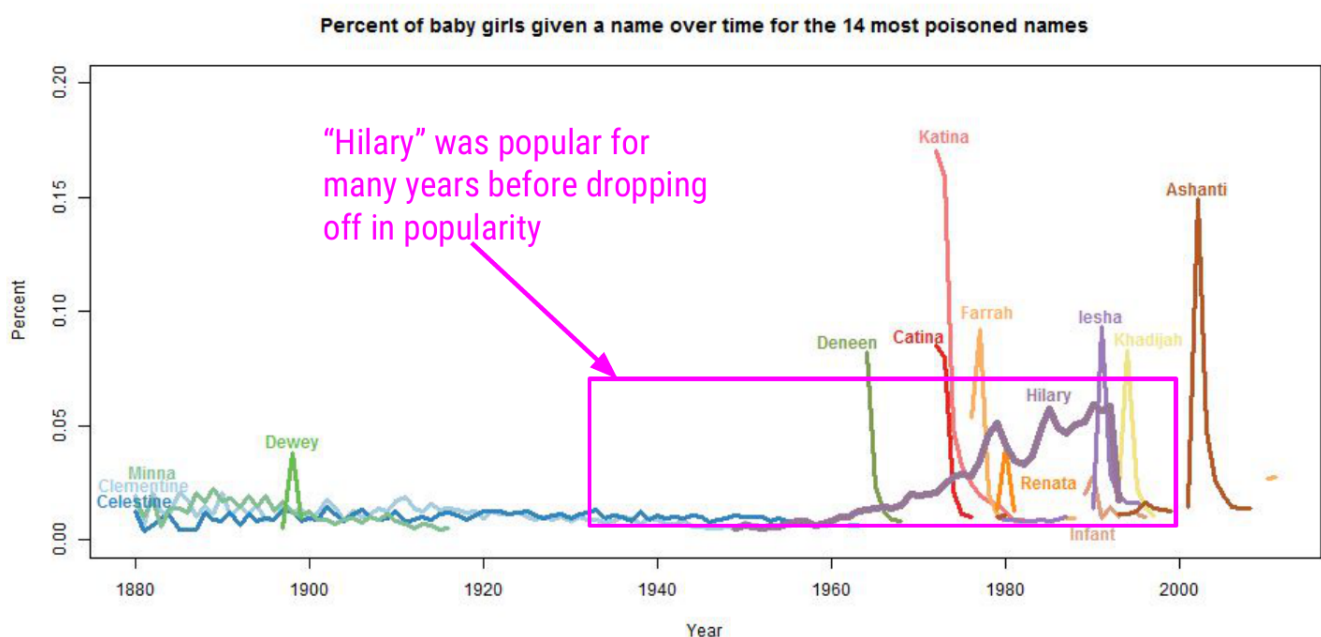
14 most poisoned names over time
To figure out what was specifically going on with the name Hilary, she removed names that became popular for short periods of time before dropping off, and only looked at names that were in the top 1,000 for more than 20 years. The results from this analysis definitively show that Hilary had the quickest fall from popularity in 1992 of any female baby name between 1880 and 2011. ("Marian"'s decline was gradual over many years.)

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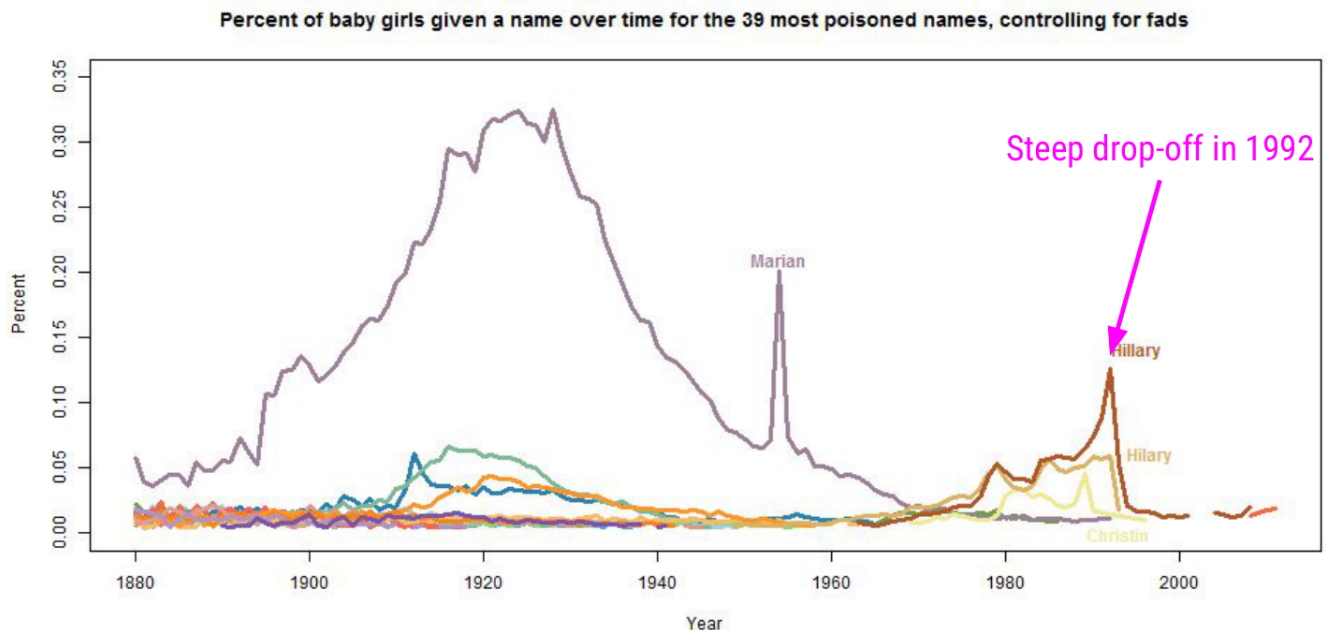
39 most poisoned names over time, controlling for fads

Communication

For the final step in this data analysis process, once Hilary Parker had answered her question, it was time to share it with the world. An important part of any data science project is effectively communicating the results of the project. Hilary did so by writing a wonderful blog post that communicated the results of her analysis, answered the question she set out to answer, and did so in an entertaining way.

Additionally, it's important to note that most projects build off someone else's work. It's *really* important to give those people credit. Hilary accomplishes this by:
- linking to a blog post where someone had asked a similar question previously
- linking to the Social Security website where she got the data
- linking to where she learned about web scraping

What you can build using R

Hilary's work was carried out using the R programming language. Throughout the courses in this series, you'll learn the basics of programming in R, exploring and analysing data, and how to build reports and web applications that allow you to effectively communicate your results. To give you an example of the types of things that can be built using the R programming and suite of available tools that use R, below are a few examples of the types of things that have been built using the data science process and the R programming language - the types of things that you'll be able to generate by the end of this series of courses.

Prediction Risk of Opioid Overdoses in Providence, RI

Masters students at the University of Pennsylvania set out to predict the risk of opioid overdoses in Providence, Rhode Island. They include details on the data they used, the steps they took to clean their data, their visualization process, and their final results. While the details aren't important now, seeing the process and what types of reports can be generated is important. Additionally, they've created a Shiny App, which is an interactive web application. This means that you can choose what neighborhood in Providence you want to focus on. All of this was built using R programming.

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.

⟶○⟵
---

https://pennmusa.github.io/MUSA_801.io/project_5/index.html

Prediction of Opioid Overdoses in Providence, RI

Other Cool Data Science Projects

The following are smaller projects than the example above, but data science projects nonetheless! In each project, the author had a question they wanted to answer and used data to answer that question. They explored, visualized, and analysed the data. Then, they wrote blog posts to communicate their findings. Take a look to learn more about the topics listed and to see how others work through the data science project process and communicate their results!

- Text analysis of Trump's tweets confirms he writes only the (angrier) Android half, by David Robinson
- Where to Live in the US, by Maelle Salmon
- Sexual Health Clinics in Toronto, by Sharla Gelfand

Summary

In this lesson, we hope we've conveyed that sometimes data science projects are tackling difficult questions ('Can we predict the risk of opioid overdose?') while other times the goal of the project is to answer a question you're interested in personally ('Is Hilary the most rapidly poisoned baby name in recorded American history?'). In either case, the process is similar. You have to form your question, get data, explore and analyse your data, and communicate your results. With the tools you'll learn in this series of courses, you will be able to set out and carry out your own data science projects, like the examples included in this lesson!