Benjamin R. Everett
Capstone Proposal 1

Google Street View(GSV) Analysis:

Impetus:

I listened to a podcast about a year and a half ago regarding two data scientists that trained an algorithm to identify year, make and model of vehicles on the street via GSV. They were able to use this information to very accurately predict election results.

In my previous career I was a real estate broker. I've always noticed that real estate price prediction algorithms don't seem to adequately account for the things that seemingly most affected my clients' purchasing behavior. My clients would spend much more time discussing the number of trees on a block, the trash on the street, the number of abandoned houses on a block, and the construction on a block than anything else. In a gentrifying neighborhood, young buyers especially, are looking to purchase potential.

Real estate is fundamentally extremely local. In Philadelphia, even more so than other cities. I bought a house 4 blocks away from most of my other friends. It cost 5 times less. It is difficult for algorithms (I'll specifically pick on zillow's Zestimate) to take this into account. I wonder if analyzing the built and natural environment via GSV can shed some light on this.

1. What question are you trying to answer?

I would like to see if I can correlate the built and physical element of a city block to real estate prices.

2. What are the data and do you have them, and have you looked at them?

I can pull 25,000 images/day from the GSV API. I've tested this and the API will allow me to pull about 1 image/30sec. This is slow but I should be able to let it run nights.

Real Estate data is in the public domain. I've downloaded the public records for Philadelphia PA. I have 800,000+ rows and 70+ columns of data. But most importantly, I have address, sale price and sale date contained within the data.

3. What is the MVP (minimum viable product)?

Train a neural net to properly classify trees on a block. Then correlate the number of trees on a block to real estate prices.

4. MVP +, MVP ++, MVP +++, etc

Train the neural net to identify more physical characteristics pertaining to the block. E.g:

Number of broken windows, trash, condition of road/sidewalk, number of empty lots, condition of brick exterior are a few that come to find.

Benjamin R. Everett
Capstone Proposal 2

Baby Names!!:

Impetus:

We're having a baby in November. A few weeks ago my wife and I were trying to settle on names. We wanted to choose some names that are unique, but not too unique, but also not wildly popular. And most of all, we can't choose a name that is trending too steeply upwards.

Out of this grew a lot of fascinating questions.

1. What question are you trying to answer?

How much does popular culture, literature, previous popularity of a name affect the naming of babies? Are different states affected differently? Can I overlay socioeconomic income for a state or county, or education level with name and see what goes on?

2. What are the data and do you have them, and have you looked at them?

The social security administration a record of all names for babies going back to 1880. I've already downloaded, pickled, parsed, and done much EDA on these data sets. Based upon my own curiosity.

I have explored IMDB website to scrape movie and TV information. It does not seem they have a readily available API, so I'd have to scrape their website.

NYT does have an API, I can pull NYT best sellers from them.

3. What is the MVP (minimum viable product)?

Baysian analysis regarding the probability of a name taking on a certain rank before a popular cultural event.

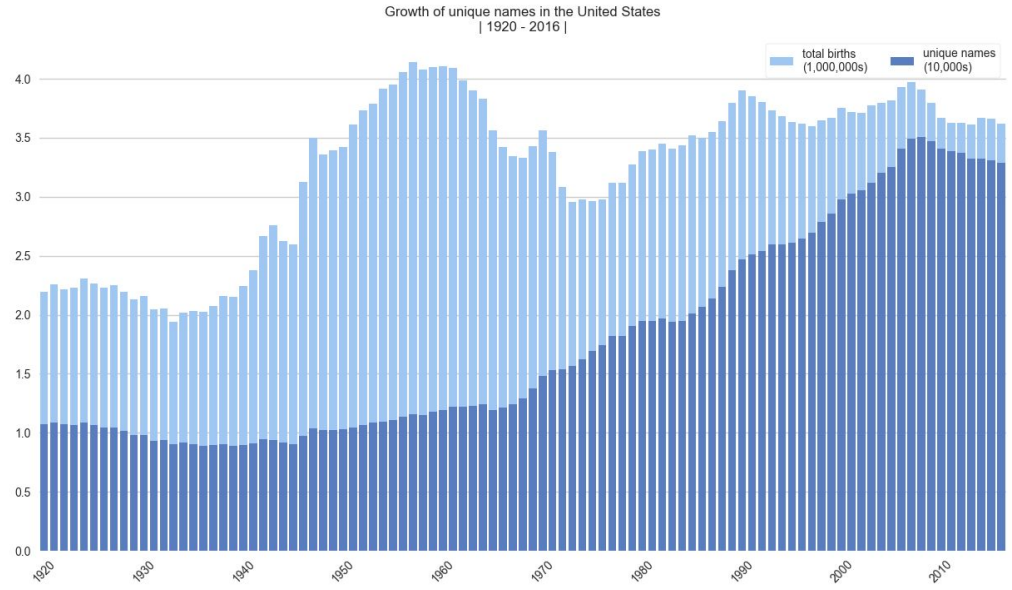
4. MVP +, MVP ++, MVP +++, etc:

Train an algorithm to predict names moving forward. E.g. what sort of names are predicted if I train the algorithm on names from 1920-1950 as opposed to 1990-2010?

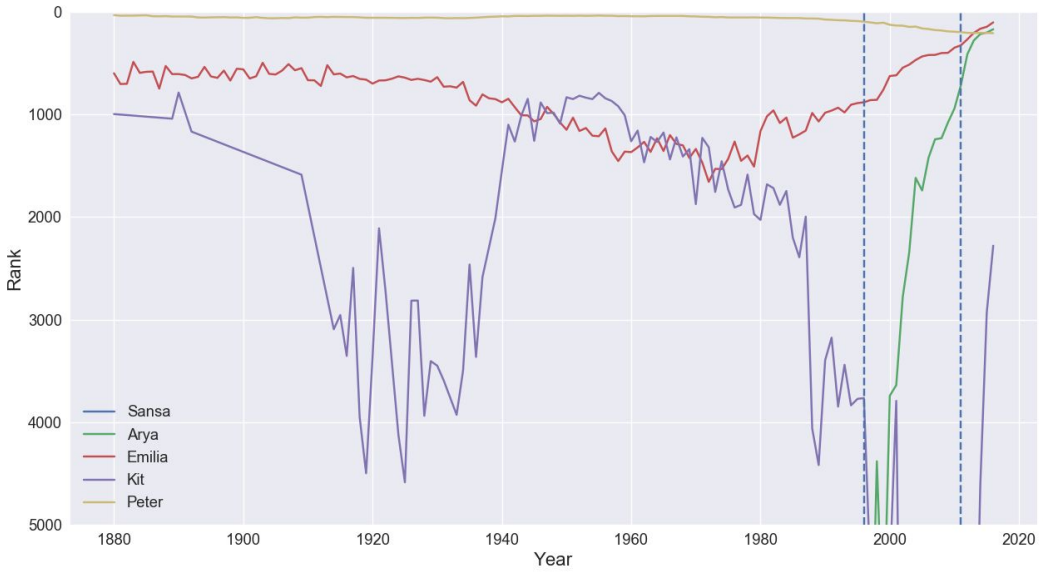
Create a prediction algorithm and website whereby people can enter their name, see the stats, but more importantly enter some preferences and have a unique name generated just for them and their upcoming baby(Thanks Frank!).

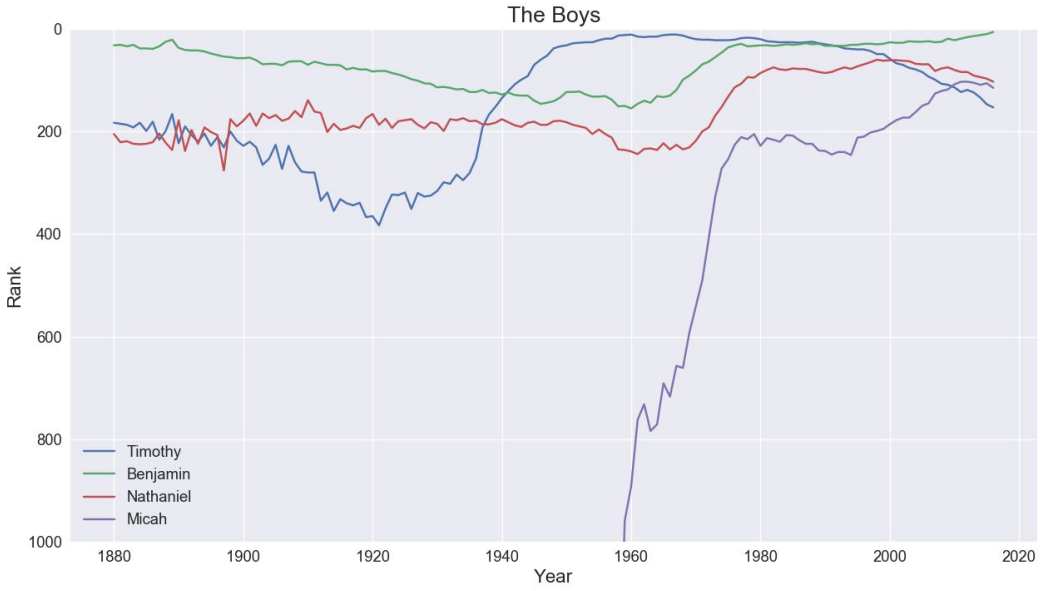
Does the uniqueness of a name affect earning potential of individuals? Don't know if I can track this down, but interesting side question.

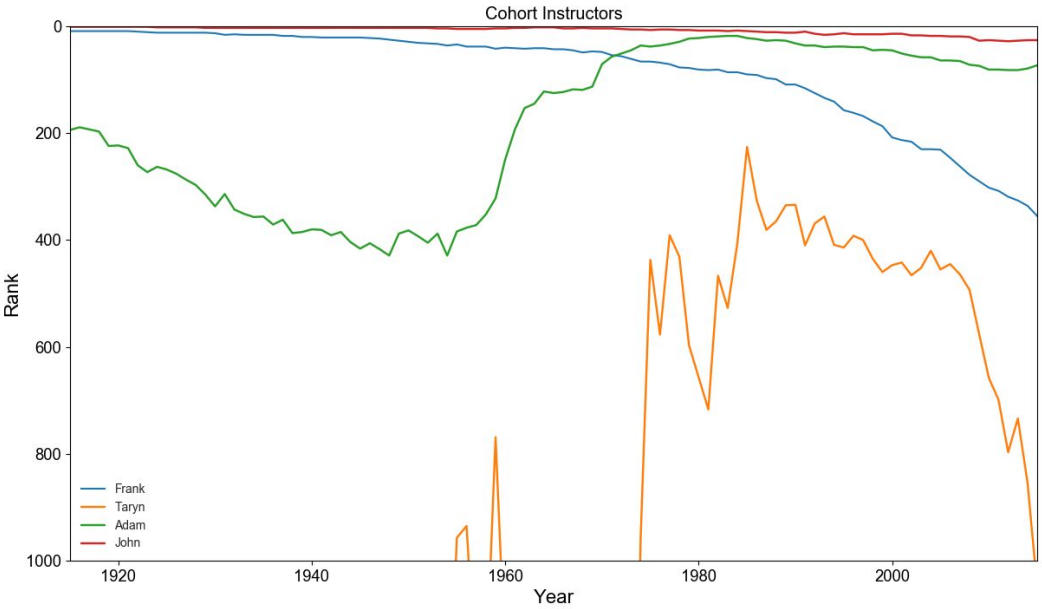
A few graphs for your viewing pleasure attached below.



A sampling of names from Game of Thrones and their ranks.First dashed line is the release of the books, the second dashed line is the start of the HBO series




The Boys
Rank
Year
0
200
400
600
800
1000
1880
1900
1920
1940
1960
1980
2000
2020
Timothy
Benjamin
Nathaniel
Micah


Cohort Instructors
Rank
Year
0
200
400
600
800
1000
1920
1940
1960
1980
2000
Frank
Taryn
Adam
John