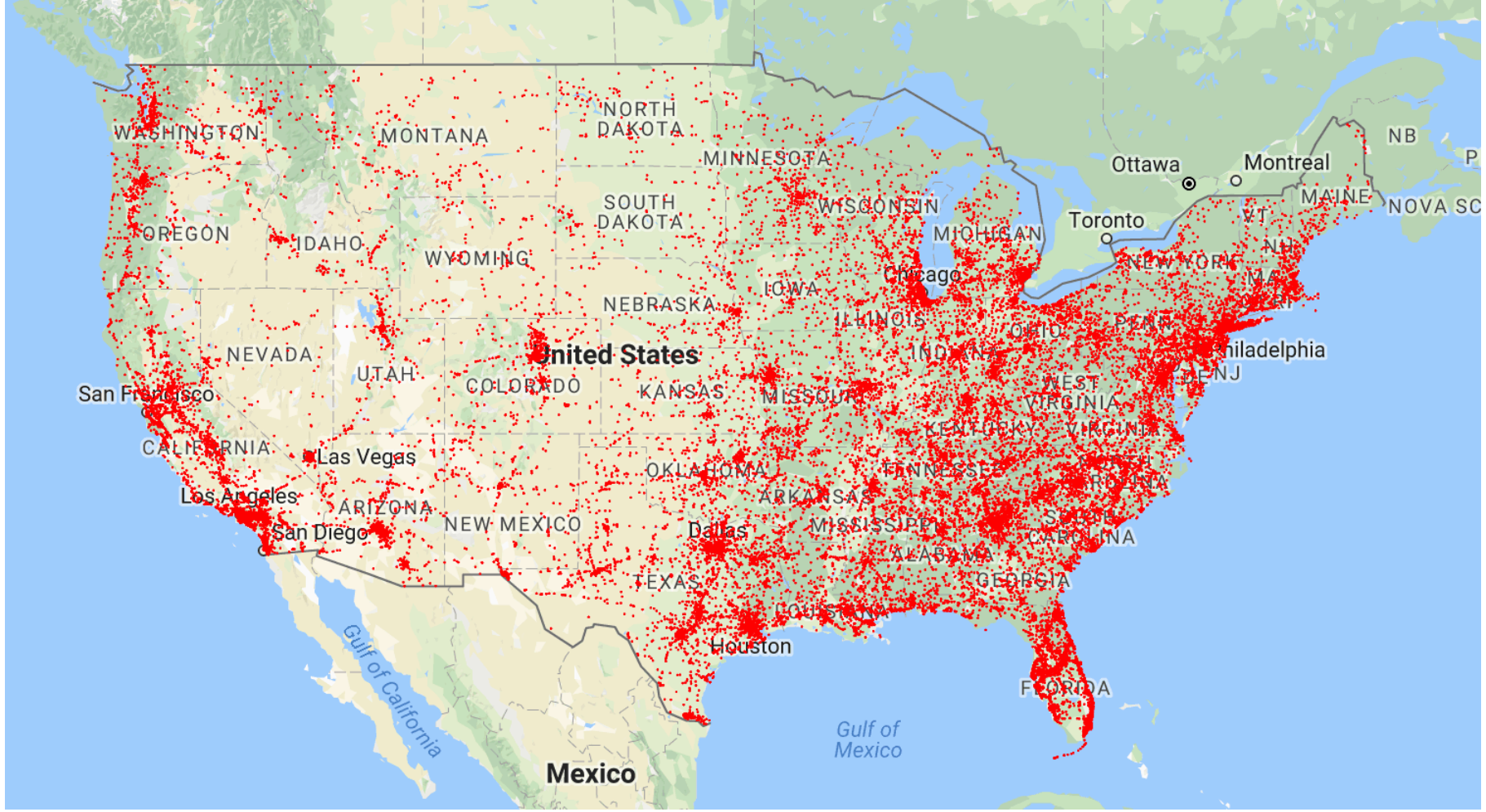
Forecasting fatality rates from motor vehicle crashes in US

Motivation

- In 2015, about 35k people were killed from motor vehicle crashes in US.
- Some counties are more dangerous, compared to others.
- Fatality rates can be analyzed for each county, fatality rates can also be predicted by analyzing the historical data.

Data sets & Tools

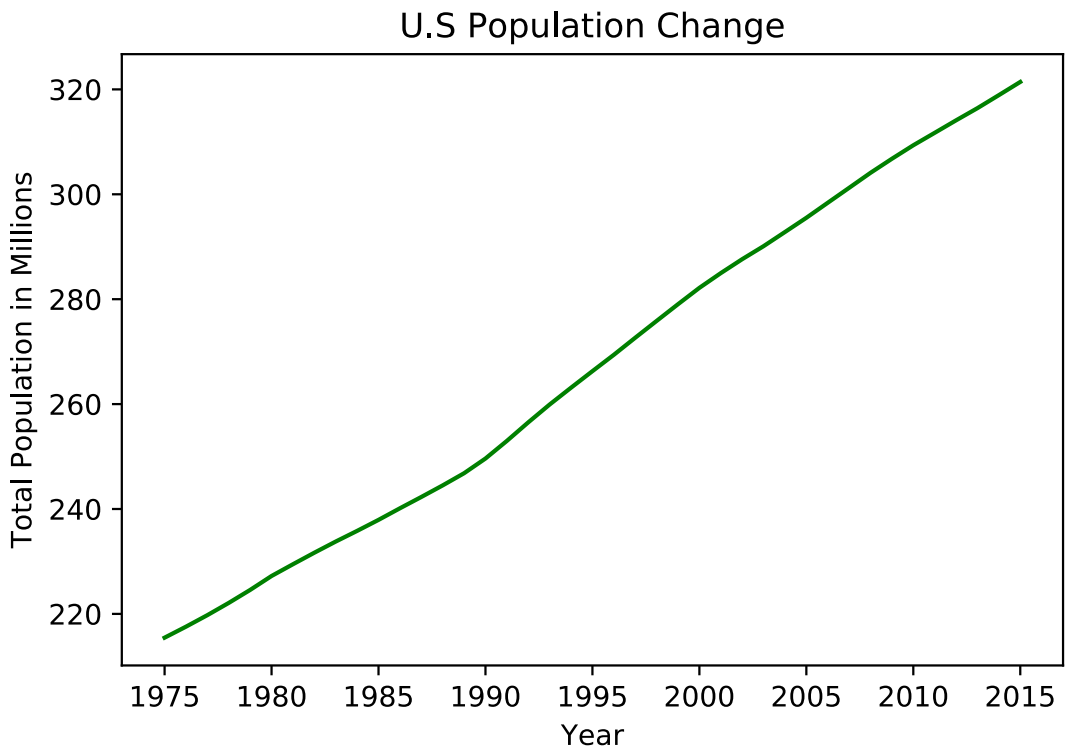
- Historical **Accident** Data
  - https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars
  - Size: 300 Mb
- Historical **Population** Data
  - Source: https://seer.cancer.gov/popdata/download.html
  - Size: 1.5 GB
- County **Geo** Data
  - Source: https://en.wikipedia.org/wiki/User:Michael_J/County_table
  - Size: 1MB

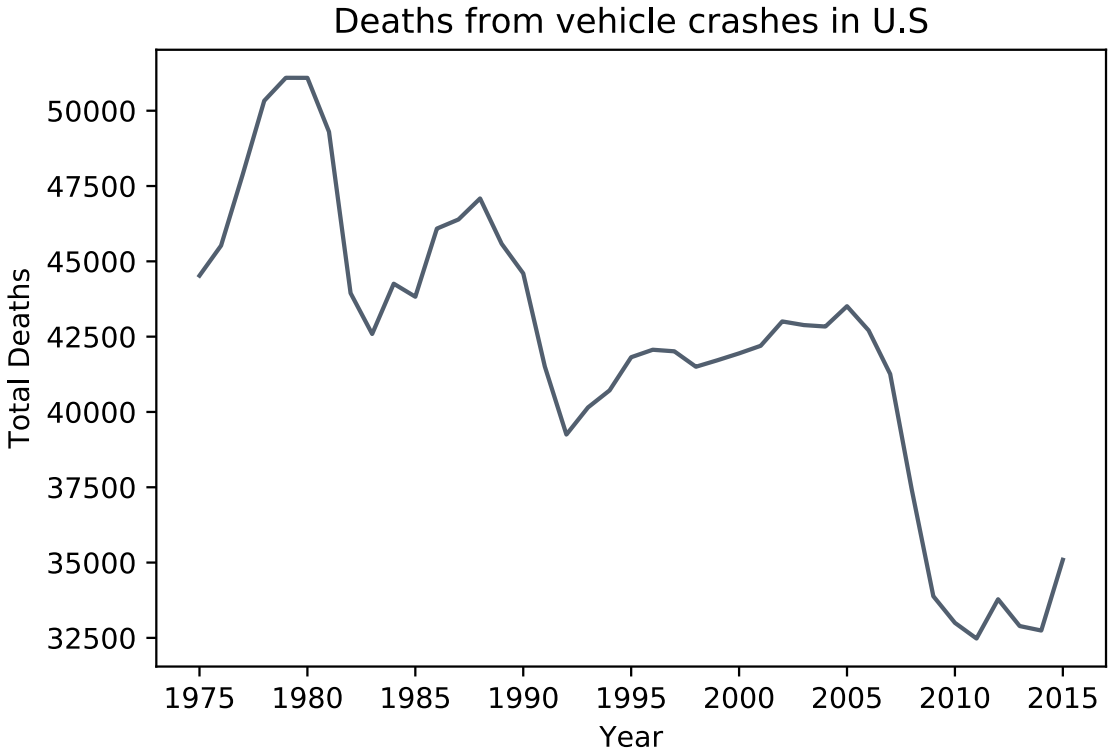


- All the accidents from 2011 to 2015 where there is at least one person lost his life.
- Yearly average car insurance rate in Iowa ~1060$, in New York City ~2900$, and in Kenedy, TX ~650$[1].
- How about the fatal accident rates per county?

1. https://www.valuepenguin.com/average-cost-of-insurance

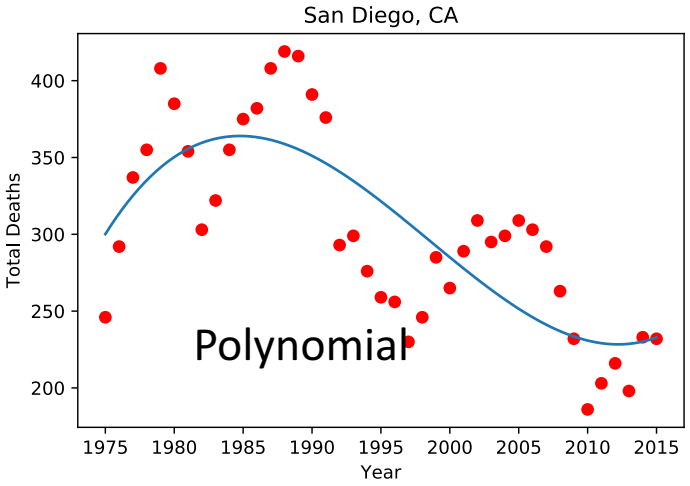
Total Population vs. Total Deaths

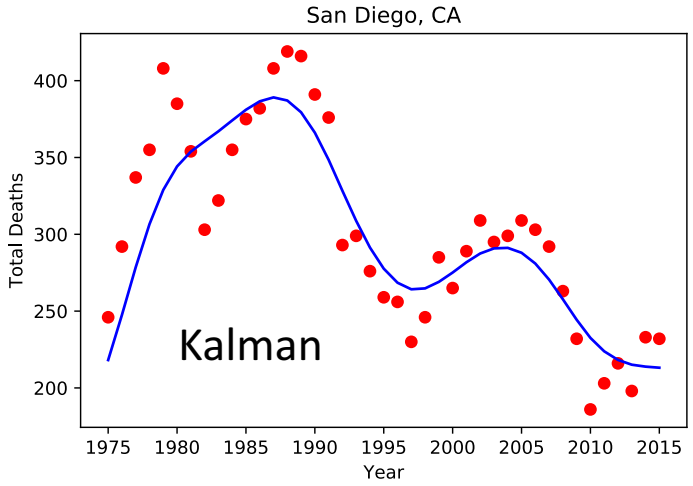


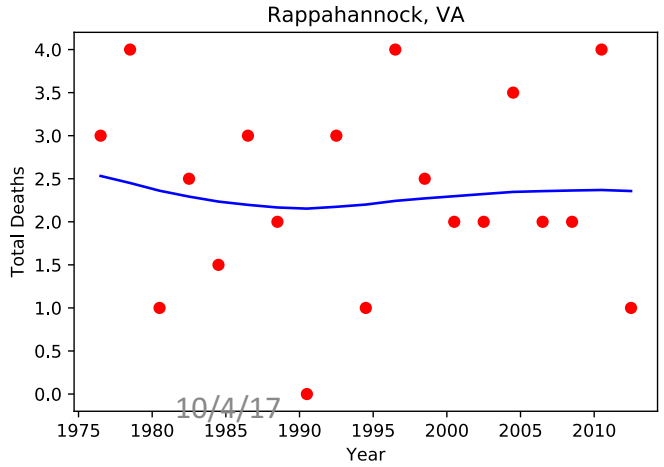


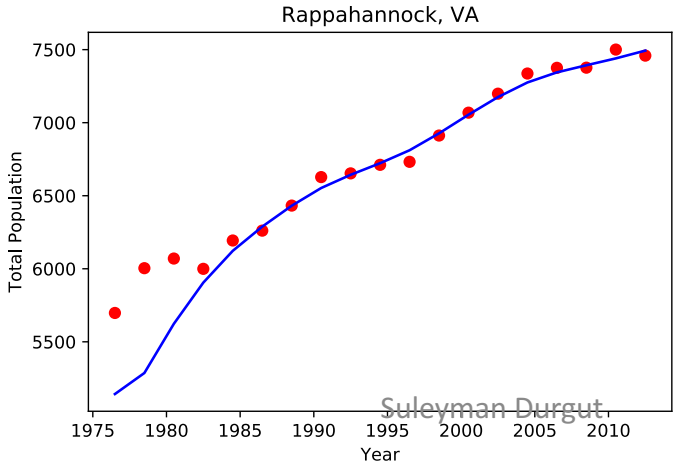
How are the population and fatality numbers changing at county level?

Polynomial Regression vs. Kalman Filtering

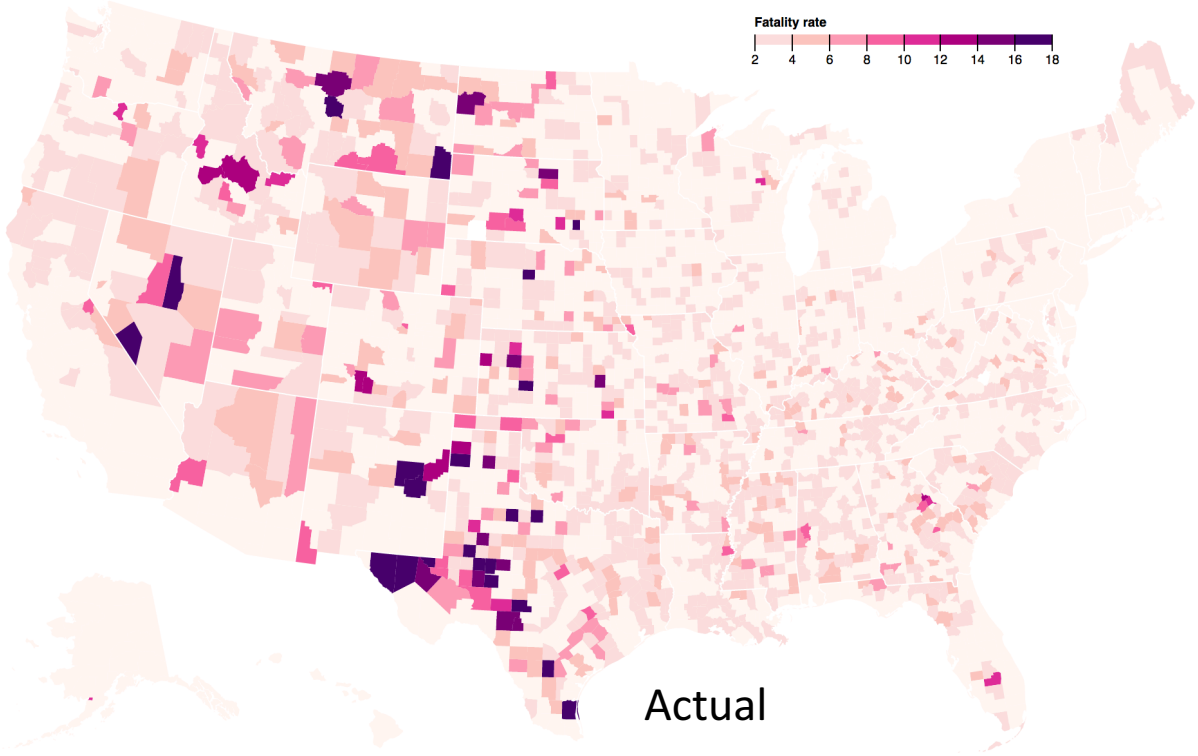


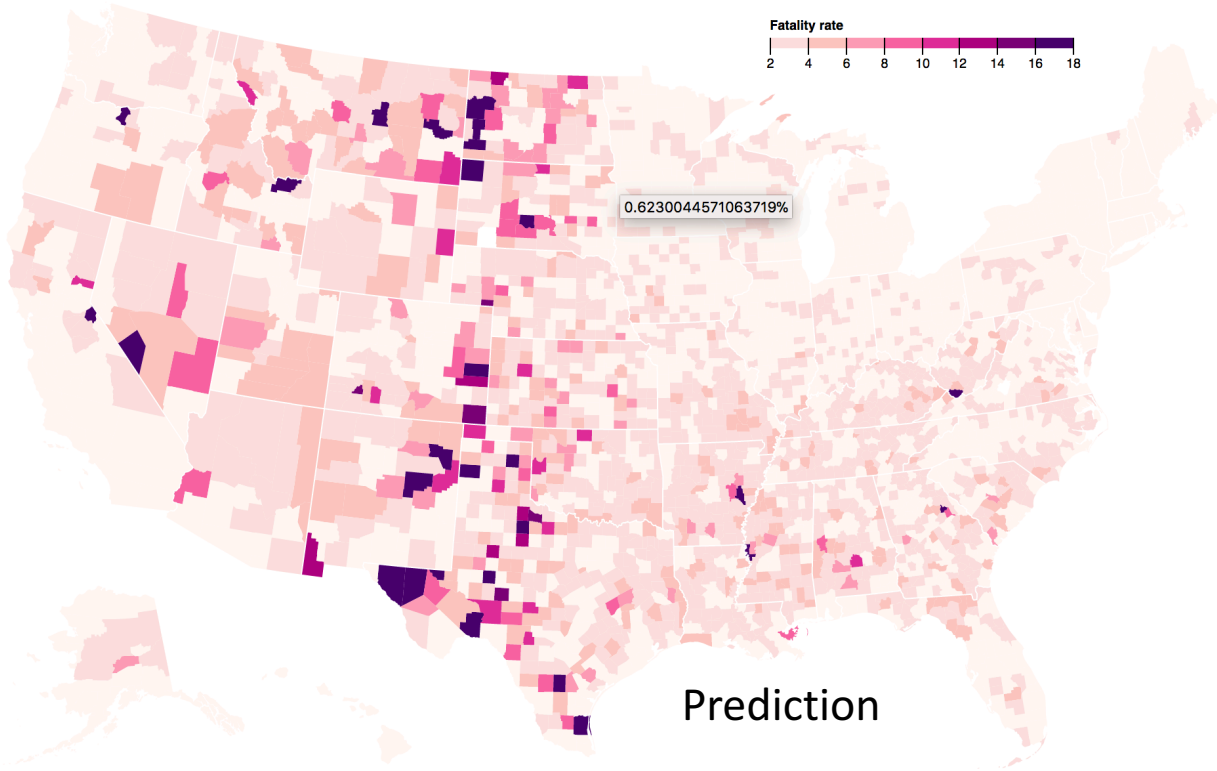






- Fitting 3k county data automatically is problematic, the right order of the polynomial needs to be determined for each case.
- Kalman Filtering corrects itself from previous measurements.
- Due to low statistics rebin the data with 2 year bin width. 1975-2013 is the train data, 2013-2015 is the test data.
- Both population and accident data need to be fitted separately due to missing data.
- Fatality estimation/ Population estimation gives the rate for each county.





- Rate means: # of fatalities/10k people in a given county.
- Correlation is 78% where county population > 50000 otherwise it is 51%
- Tried to use KNeighborRegressor but it decreased the correlation score(Problem stems from dividing the dataset as train and test.)
- It looks like New York City(0.27/10k ppl) is pretty safe compared to Kenedy, TX(~49/10k ppl).
- Dataset also includes number of deaths, weather information, drunk driving and arrival time after the accident. Using these information, I can try to predict whether the driver was drunk or the relation between arrival time, and number of deaths…