Data Incubator Project Proposal

Suleyman Durgut, Ph.D Cand, Experiemental High Energy Physics, U. of Iowa

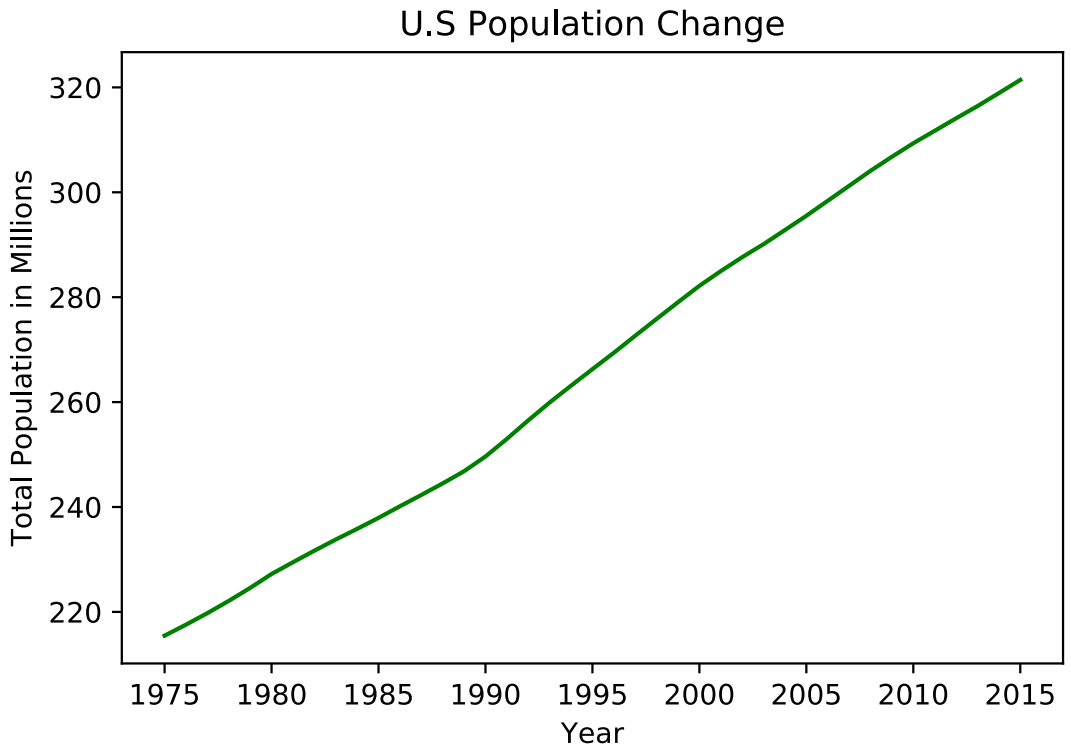
Motivation

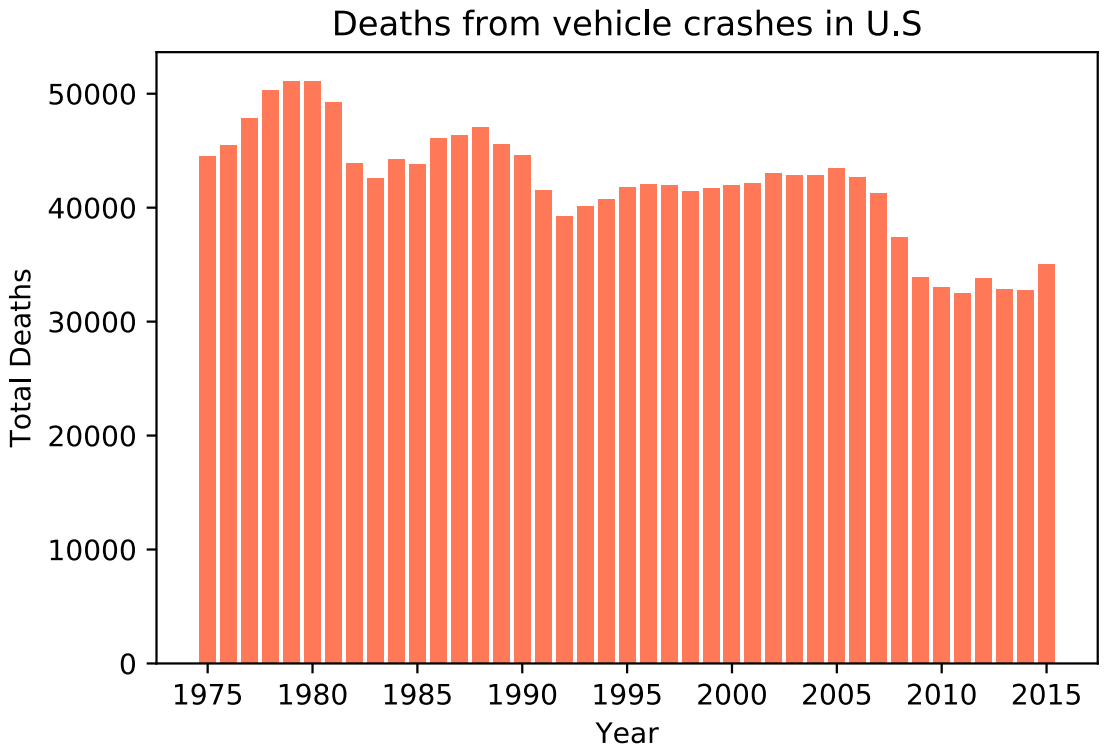
- In 2015, about 35k people were killed from motor vehicle crashes in US.
- In fact, some counties are more dangerous, and some are safer to drive.
- It would be nice to analyze fatality rates at county level.
- Future fatality rates can also be predicted at county level.
- Insurance companies might find this information useful.

Data sets & Tools

- Yearly vehicle crash data at county level from 1975 to 2015
  - Source: https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars
  - Size: 400 Mb
- Yearly population data at county level from 1969 to 2015
  - Source: https://seer.cancer.gov/popdata/download.html
  - Size: 370 Mb
- U.S FPIS codes for counties
  - Source: www.schooldata.com/pdfs/US_FIPS_Codes.xls
  - Size: 250 Kb
- Tools: Pandas, NumPy, Matplotlib

Total Population vs. Total Deaths

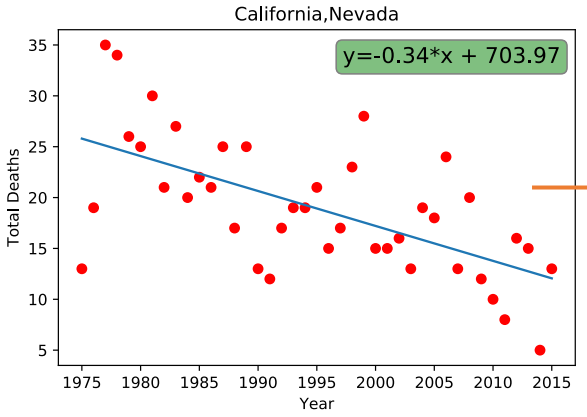


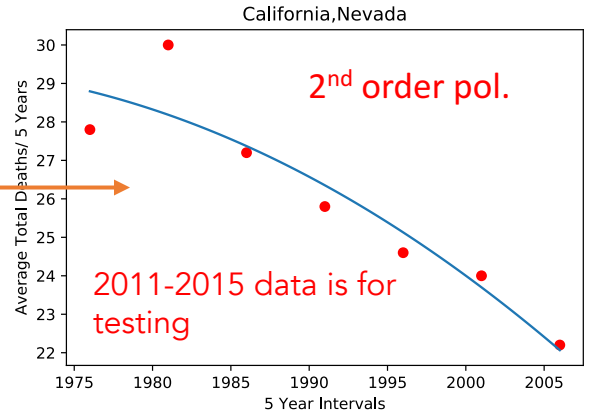


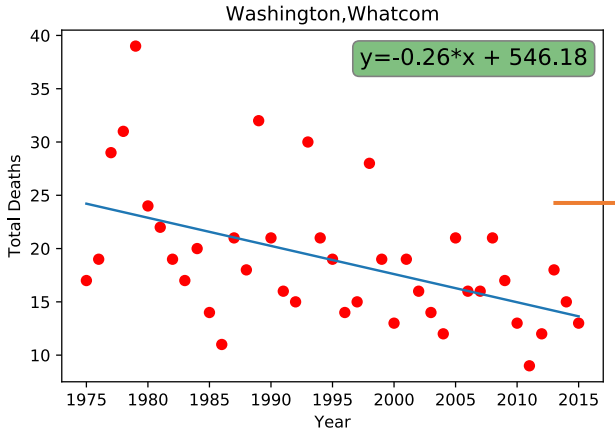
How are the population and fatality numbers changing at county level?

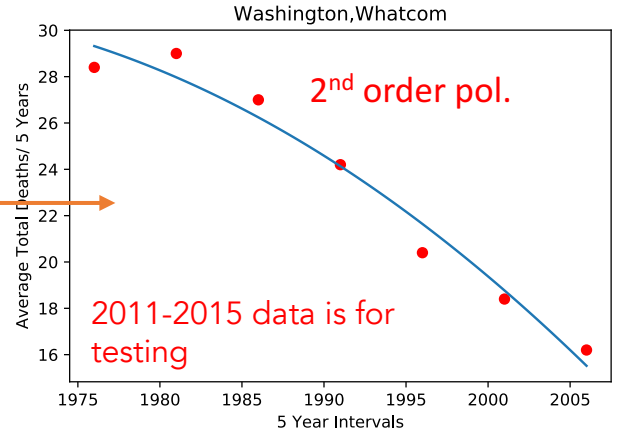
Methodology

- Purpose: Predict number of deaths per 10,000 people for each county.
- Divide the data sets as training and test.
- Perform polynomial/linear regression for both population and fatality data for each county in training data sets. More than 6000 distributions for the analysis!
- Using the fit parameters, get prediction results and compare it with the test data sets.
- First attempt was plotting each data yearly, however statistics is low for number of fatalities in low pop counties.
- Second attempt : Rebin the data sets into 5 year intervals and use the average of population and fatality information for each bin.
- Calculate (number of deaths/population)*10,000 people for a given time period and compare it with the test data for each county.

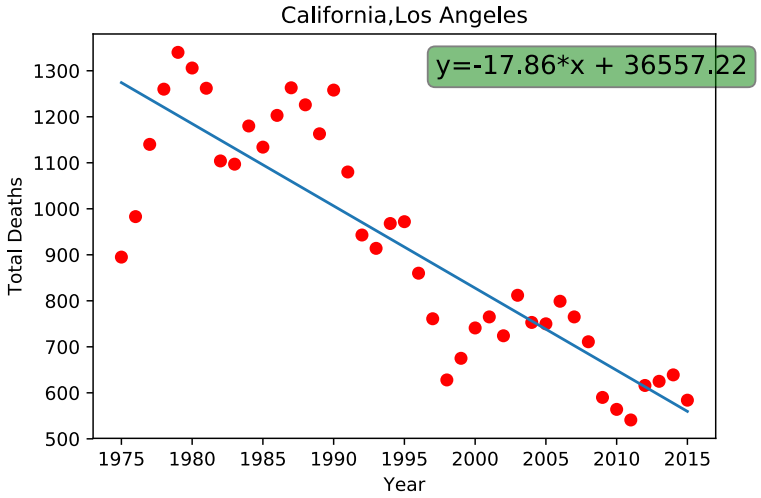


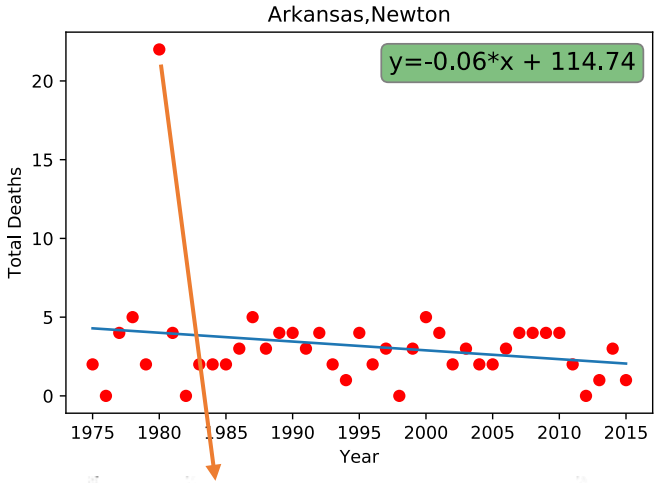




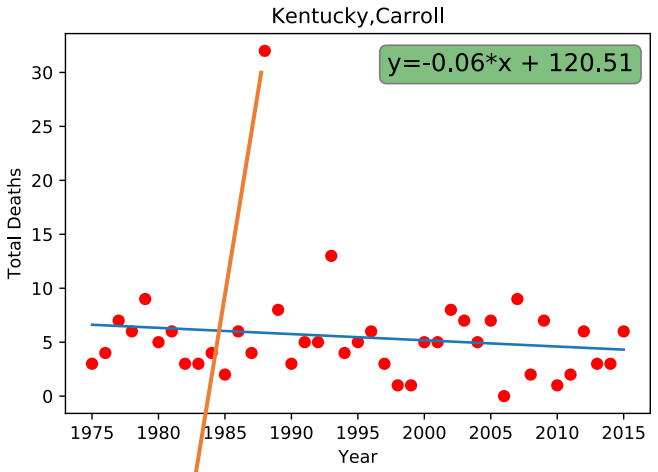


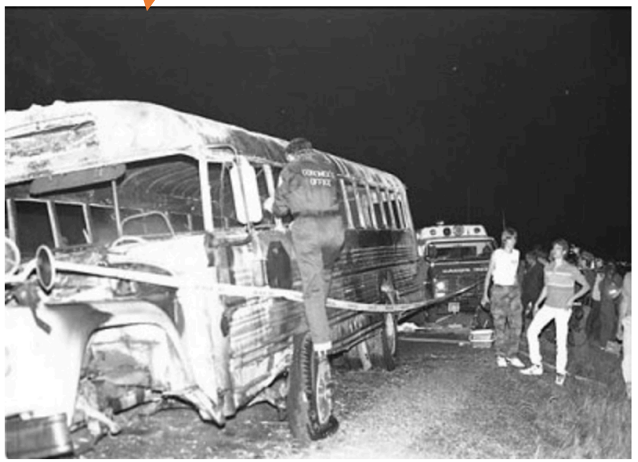
The Good & Bad



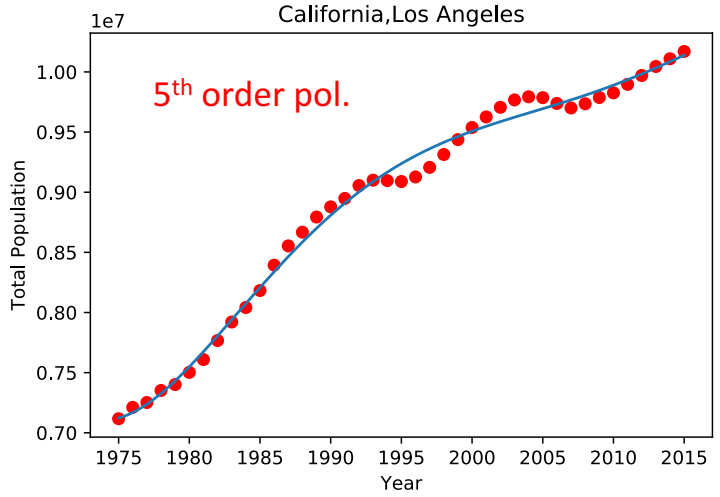


Chicago Tribune, June 6, 1980



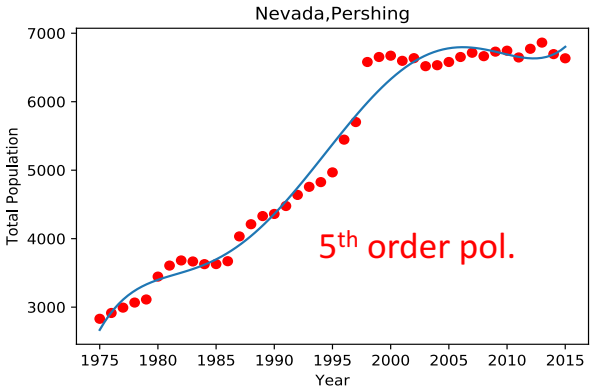


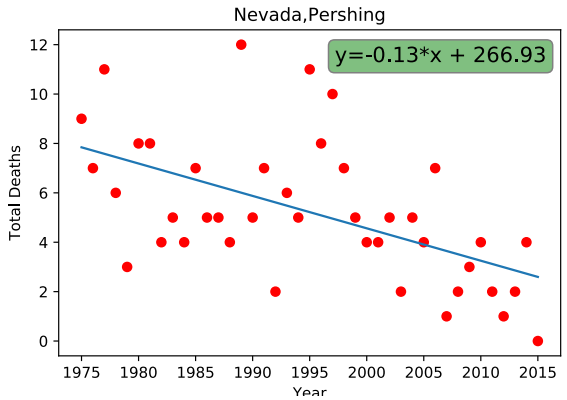
ANNIVERSARY OF TRAGEDY: Carroll County, Ky., Coroner Jim Dunn looks in the driver's window of the First Assembly of God Church bus from Radcliff, Ky. after the bus was hit by a truck driven by a drunken driver on May 14, 1988. Twenty-seven people died in the accident. (Staff file photos by Mark Campbell)

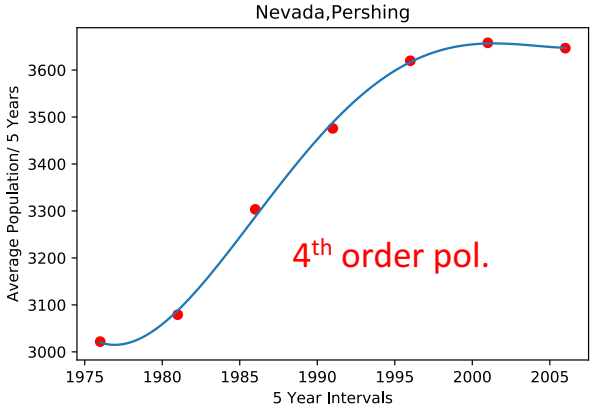


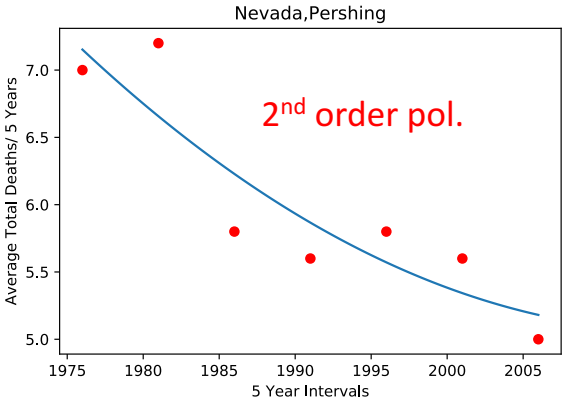
Results

- Training data: 1976-2010 (7 bins)
- Test data: 2011-2015 (1 bin)
- Use 4th order pol. for population, 2nd order pol. for fatality data after rebinning.

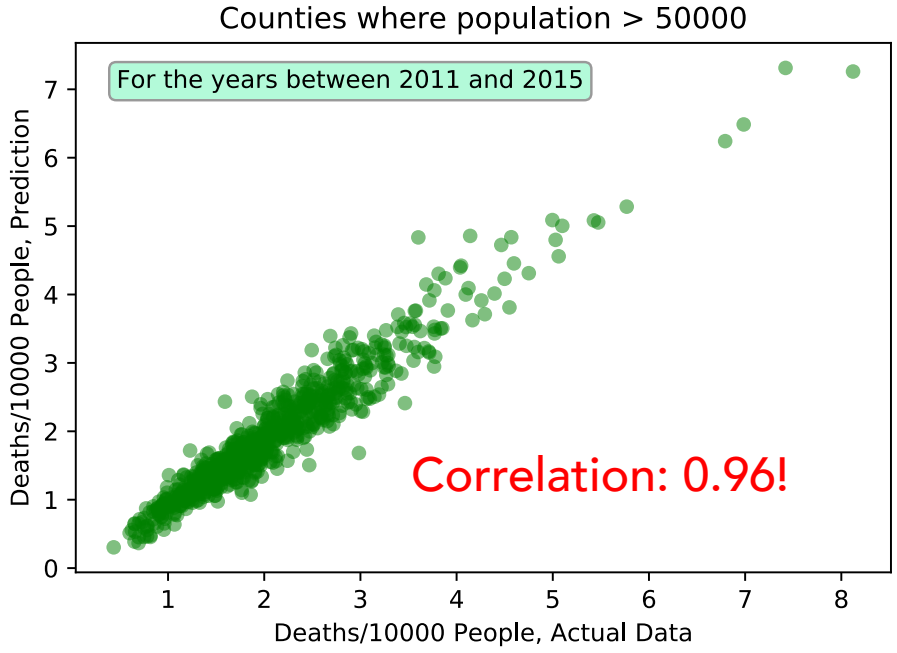








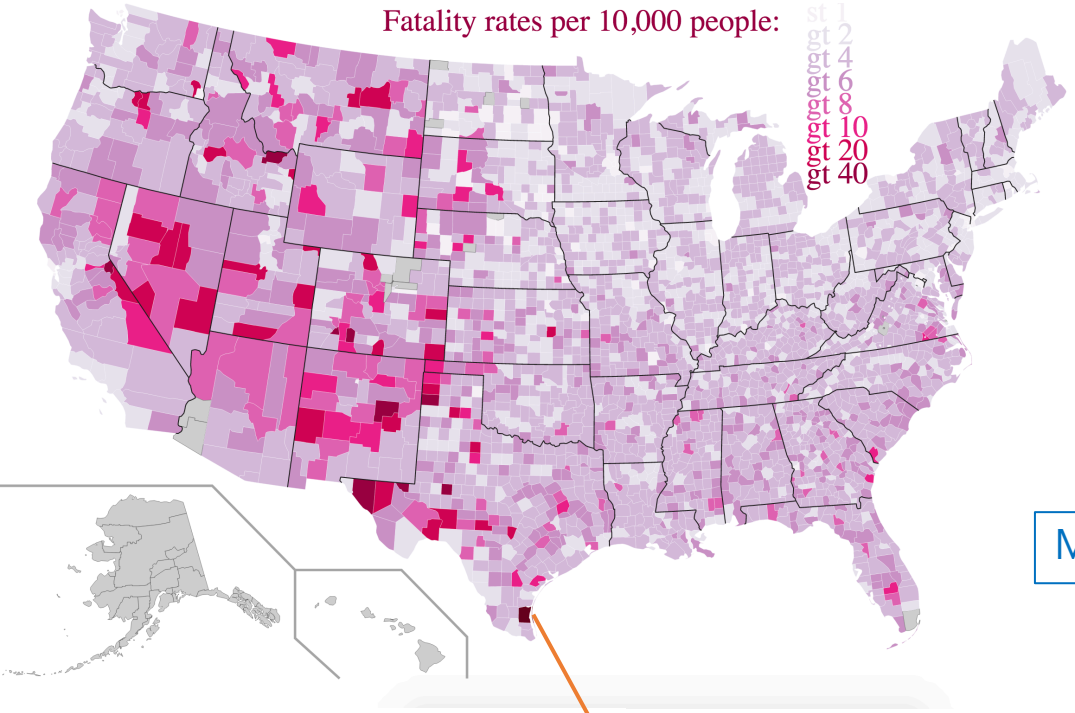
- Pearson correlation coefficient: 0.88 for the whole data.
- Some counties have very low population <10k, low statistics.
- Check the correlation if county population > 50k, correlation: 0.96.



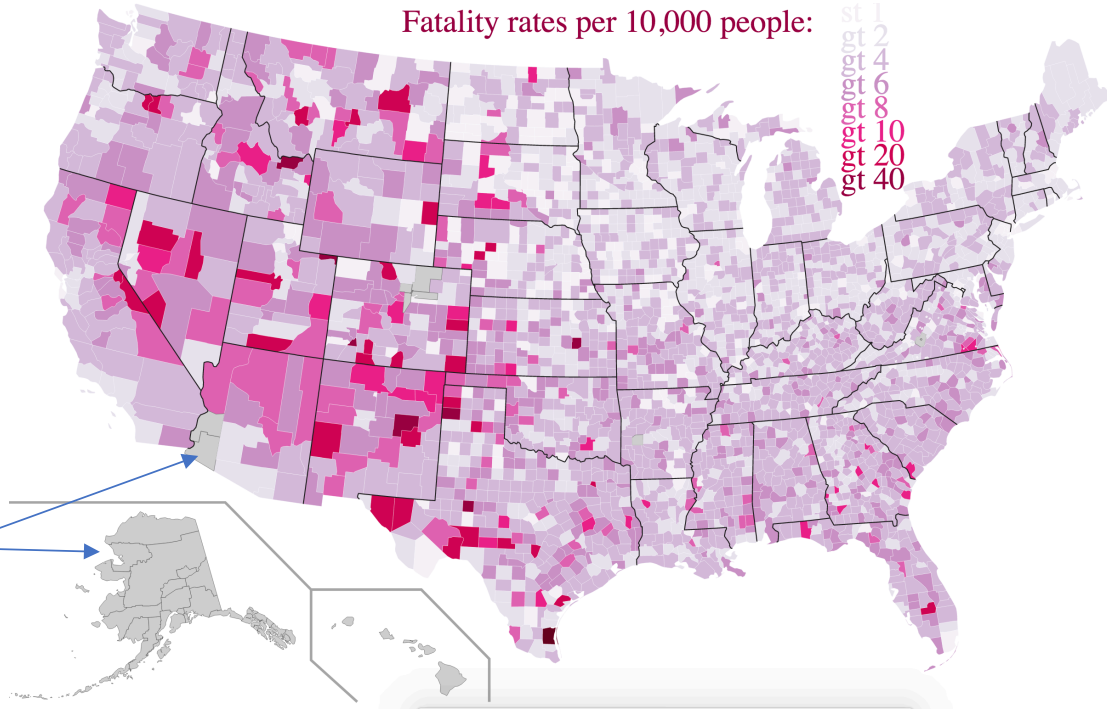
Durgut

Results

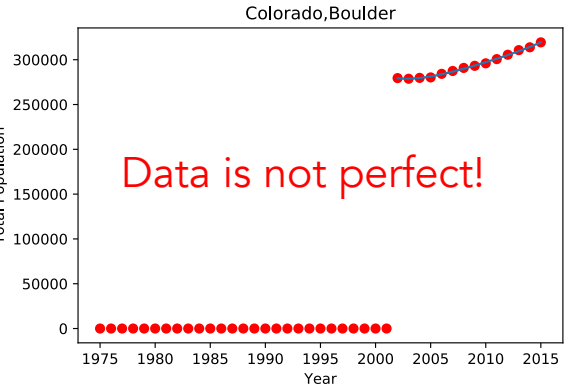
Actual Data



Prediction

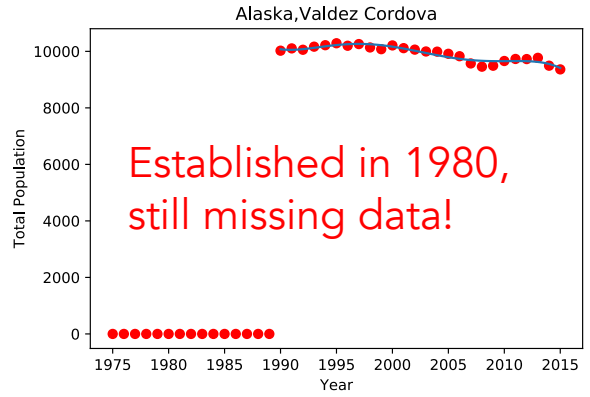


Missing data!



Data is not perfect!

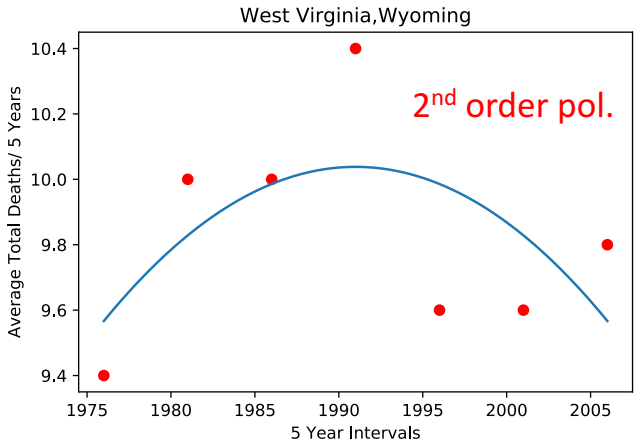
- Ignore missing data for now.
- Kenedy, TX is by far the most dangerous place!

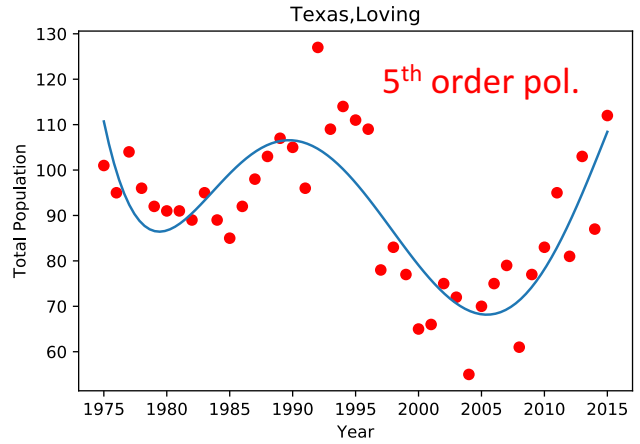


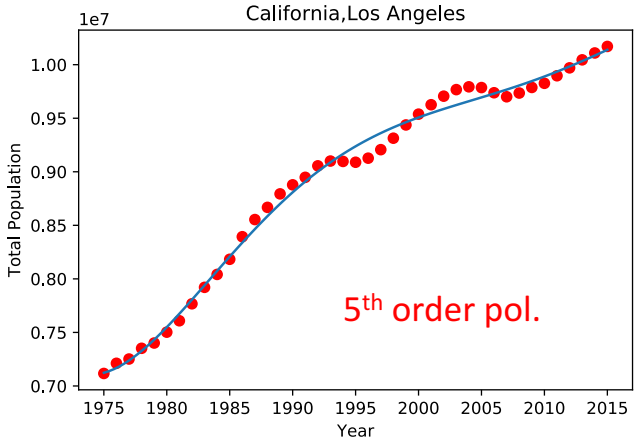
Established in 1980, still missing data!

Outlook

- Find the missing data from other sources.
- Not all regressions are perfect, order of the polynomial in the fit equation for each county can be improved by performing an *F-Test.*
- Correlation number considers, **Loving, TX** and **Los Angeles, CA** has the same weight. Data points need to be reweighted.







Thank you!