y: state-level infection rate

X: Google mobility data and other factors

* Intrinsic characteristics:
  + COVID-19 vulnerability index: combination of socioeconomic, epidemiological, healthcare etc. factors. <https://www.precisionforcoviddata.org/>
* Meteorological: weather, humidity, season?
  + Winter months: stay indoors more
  + Temperature/humidity
* Special factors: policies, holidays.
  + Emergency declaration, school closure, stay-at-home order, reopening
  + Reopening: mask mandates, vaccine mandates
* Infection trajectory in the past month: Markov chain, combination of mobility factors is ONE state? -> output increase, decrease
* Google search volume for specific phrases
* INCLUDE PRIOR INFECTION RATE -> ideal range (1, 30, 60)

Initial Data Exploration: One for each category of mobility?

* Pearson coefficient per month (x = daily mobility data, y = daily cases)
* Tease out relationship between density and infection rate. How does infection rate vary across different mobility changes for different population densities?
  + X = change in mobility, Y = infection rate
  + Histograms = low, medium, high density regions averages

Different models to try

* Simple Regression
* Neural Network

DL with a variety of features: <https://ieeexplore.ieee.org/abstract/document/9179729>

Mobility data and neural networks: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7755147/#R11>

Feature Engineering

* Normalize
  + - Defining model

Feature importance

* Modify functions to include days\_prior variable (14, 30, 60 etc)
* Try top 30 days: if most important is still 30th, use anti-correlation feature importance
* Chapter 8.1, 8.2, 9.2, 9.6

Try Ensemble Learning

Try Deep Neural Network

* LSTM (next step)

Throw away feature (not needed):

* Try Lasso?
* Production: want low latency, throw away feature
* Ablation analysis
* After model done: xml explainable machine learning. -> Explains deep neural network

Dalle2