* + **Introduction**
    - There's a lot of publicly available data related to the pandemic
      * We wanted to use some of these data to find a story about the effectiveness of shutdown policies in the U.S.
    - All states have undergone a period of economic shutdown
      * Pandemic will continue
      * Motivation: States may consider shutting down again
    - Shutdown policies are sort of a proxy for social distancing
      * Many studies confirm that social distancing is effective at slowing spread of virus; It has huge benefits
      * Prevents huge loss of life and economic damage; Unfortunate to say both in the same breath
    - Difficulties in studying the effects of social distancing, especially in the US
      * The length of state shutdowns have ranged from two weeks to two months
      * Restrictions vary widely
    - Due to this variability:
      * It remains difficult to directly examine the effect of social distancing
      * Therefore, we focus on shutdown policies
  + **Objectives**
    - Explore the effects of economic shutdown policies on the transmission of SARS-CoV-2 and case-counts of COVID-19 within each state and nationwide.
      * Hypothesis 1: States that implemented relatively **l**onger shutdowns had fewer cases per capita and/or lesser R0 values at the time of reopening.
    - Identify states that are outliers or otherwise uniquely interesting for further consideration by Policy Team.
      * Hypothesis 2: States that implemented shutdowns before their R0 values exceeded 1 will have fewer cases per capita and/or lesser R0 values at the time of reopening.
      * Hypothesis 3: States that reopened while their R0 values were greater than 1 have more cases per capita and/or greater R0 values at present.
  + **Data**
    - **Shutdown policies**
      * Source: Official text of shutdown policies
    - **Mobility of individuals**
      * Source: Google Community Mobility Reports
        + The baseline is the median value, for the corresponding day of the week, during the 5-week period Jan 3–Feb 6, 2020.
        + Six place categories

Workplace, residential, and retail and rec were highly correlated

Modeling with workplace and residential did not significantly improve model fit over just using retail and rec

* + - * + Inflection points

Approximated the second derivative using the method of divided differences

Plots are for weekdays only

No obvious pattern there; Decided to use mobility-defined shutdown instead

* + - **Basic reproduction number (R0)**
      * Source: Rt.live
        + Estimates given from state's first case to present day
    - **Case and death counts**
      * Source: NY Times
  + **Exploratory analyses**
    - **Length of shutdown**
      * Here we test Hypothesis 1
      * Tested correlation with Kendall's Tau
        + Not Pearson CC because one of our dimensions in an ordinal, discrete variable, and therefore it is not normally distributed
        + Not Spearman's rank CC because we have many ties in the ranked data
    - **Clustering the states**
      * An attempt at aggregating the states, noting that they are not independent of one another
      * PCA
        + The PCs are basically linear combinations of the features
        + Data must be normalized before performing PCA
      * K-means clustering minimizes variance within clusters and maximizes distance between clusters
        + K-means bad because data are not exactly convex or arranged into spherical clusters
        + DBSCAN good to deal with outliers, but bad because the density of the data is not very uniform
        + Hierarchical good

Assumes very little about the structure of the data

Especially with connectivity restraints: only adjacent clusters can be merged

* + - * Unsupervised, so determined parameter k using: Silhouette analysis
        + A measure of how similar an item is to its own cluster and how different it is from other clusters
        + Red line is average silhouette score
  + **Modeling R0**
    - **Multi-level / Mixed-effects linear modeling**
      * In general
        + A type of generalized linear regression
        + Allows us to consider between and within subjects data;
        + Consider categorical variables that may contribute random variance to the observations
      * Well suited to data with nested/hierarchical subgroups
        + We have many random effects (i.e. 50 states)
        + Some variance between states might be explained by a higher level subgroup in which they're contained (e.g. regions)
        + Some characteristics at the national level might also be present at the state level, even though our data might be too limited to observe them
    - **Model diagnostics**
      * **Basic**
        + Random residuals: should have no correlation in residuals vs. fitted values for entire model and residuals vs. each predictor variable
        + Should not use predictors that are strongly correlated
        + R^2
        + F-test
        + ANOVA
      * **AIC (Akaike information criterion)**
        + AIC = 2k - 2 \* ln(L-hat)

k: number of parameters

L-hat: maximized likelihood function for the model

L-hat = P(c | theta-hat, M)

M: model

* + - * + Basic facts

Rewards goodness of fit and it penalizes complexity (added parameters, over-fitting)

Lower values are relatively better models

Describes relative quality of one model with respect to others; Does NOT describe the absolute quality

* + - * + Outperforms BIC even when "true" model is present in study
        + Asymptotically equivalent to leave-one-out cross-validation for ordinary/mixed-effect linear regression models
      * **BIC (Bayesian information criterion)**
        + BIC = k \* ln(n) - 2 \* ln(L-hat)

L-hat: maximized likelihood function for the model

L-hat = P(c | theta-hat, M)

M: model

n: number of observations

k: number of parameters

* + - * + Basic facts

Rewards goodness of fit and it penalizes complexity (added parameters, over-fitting)

Lower values are relatively better models

Describes relative quality of one model with respect to others; Does NOT describe the absolute quality

* + - * + Need to have very large n, even larger than k^2
      * **Chi-square goodness of fit tests**
        + Categorical data
        + Independence of rows/columns in contingency tables
        + Reject H\_0 when test statistic is large
      * **p**
        + The smallest size alpha at which H\_0 can be rejected, based on the observed value of the pre-determined test statistic
      * **SE**
      * **t**
        + Used to test hypotheses about an unknown population mean when the value of its SE is also unknown
  + **Conclusion**