Project outline

**Introduction**

* What we chose:
  + We chose to analyze several variables around covid-19 and see which variables were the most strongly correlated with covid-19 outcomes (deaths, cases, mortality rates)
* Why we chose covid 19:
  + During the brainstorming session – we evaluated a few different options but felt with the breadth of impact covid has had on everyone and the availability of data, we could explore the data in depth
* Variables we chose:
  + Weather
  + Age
  + Testing Density
  + Policy Decisions/Governmental factors
  + Population Density
  + Hospital Bed Density
* Why we chose these variables:
  + As covid has impacted virtually everyone on earth and has constant news coverage cycles, we often hear of what the clear risk factors are (i.e age, population density, preexisting conditions) and the potential risk factors (Hospital capacity, public policy decisions) but do not have a real understanding of the actual correlative values of these variables on these outcomes

**Summary of Data Wrangling and Hypotheses**

* We originally chose to analyze data at the county level but gathering and aggregating data at this level was proving to be a large hurdle to moving towards the analysis – we were able to find some data at the county level for our variables and not for others and chose to maintain a level of consistency across all the variables analyzed – thus, we chose to analyze the data at the state level
  + With this – we took any data we could find at the county level and grouped it to the state level to merge with any data we could find only at the state level
* Our data came from a variety of sources
  + John Hopkins Covid-19 Data
  + Kaggle NYT enriched dataset
  + Weather.gov
  + County/census Data
  + (Michael’s Datasets)
  + Erik’s Datasets
* Our hypotheses prior to investigation were based upon anecdotal evidence and what was being told to us via the news, data experts, and medical experts
  + Our hypotheses:
    - As Temperature rose, we would see a negative correlation with cases/deaths/mortality rate (defined as deaths per day/cases per day)
    - As Age rose, we would see a positive correlation with cases/deaths/mortality rate
    - As Testing Density rose, we would see a positive correlation with cases, but a negative correlation with deaths/mortality rate
    - As Population Density rose, we would see a positive correlation with cases, deaths, and mortality rate
    - As hospital bed density rose, mortality rate would decrease
    - Quickness of response to covid (closing down quickly), staying closed longer, and limiting reopening would all correlate with negative case rates and death rates

**Analysis**

* Talk about data analysis, results, compare to hypothesis for each of us

**Conclusions**

* Implications of the data/takeaways
* Limitations of the data
  + Not available at the granularity we would have liked OR difficult to wrangle a
* Obstacles
  + Time
  + Overly ambitious
  + Difficulty merging data
* Extras – interactive data from Michael