*Covid-19 Exploratory Analysis*

TTU MSDS Business Intelligence - Group 10

*Coy, Juan  
De Los Santos, Jonathan*

# Overview

This analysis aims to examine confirmed cases of the novel Coronavirus (Covid-19) and associated deaths in the state of Texas with large, publicly available datasets that cover a diverse set of health and socioeconomic dimensions.

The datasets are as follows:

## [New York Times US Covid-19 Data](https://github.com/nytimes/covid-19-data)

The foundation of our Covid-19 analysis is this cumulative count of confirmed cases and deaths related to the virus. This is offered as both a historical daily tally of new cases at the selected geographic grain and a daily refreshed total count for those same geographies. We opted for the latter as it is smaller and represents the cumulative cases that we need for analysis.

## [County Health Rankings](https://www.countyhealthrankings.org/explore-health-rankings/measures-data-sources/county-health-rankings-model)

This program is cosponsored by the University of Wisconsin Public Health Institute and publishes county-level data related to length and quality of life. This includes health behaviors, clinical care, social and economic factors, and physical environment. Each of these are attributes that will be useful to explore relative to Covid-19 transmission.

## [CDC Social Vulnerability Index (SVI)](https://svi.cdc.gov/Documents/Data/2016_SVI_Data/SVI2016Documentation.pdf)

The Center for Disease Control (CDC) publishes this index as a measure of a community’s resilience to any type of disaster, including disease outbreak. This contains some especially interesting attributes related to human and household density that are directly related to what is understood about the transmission of Covid-19. These factors might be particularly useful for business leaders who are evaluating their own Covid-19 mitigation strategies around employee and operational policies.

# Data Cleaning

## Data Quality

Our data consists of 3 individual datasets: 1) Covid-19 Data (New York Times GitHub) 2) County Health Rankings (2020) 3) CDC Social Vulnerability. The quality of data was sound, considering only minor data munging was required in order to join the 3 different datasets.

## Missing data

* Covid-19 Missing Data
  + Probable Cases
  + Probable Deaths
* Health Rankings Missing Data
  + Poor or fair health ratio
  + Poor physical health days ratio
  + Poor mental health days ratio
  + Adult smoking ratio
  + Excessive drinking ratio
  + Air pollution ratio
  + Drinking water violations ratio
* Missing Values & Related Variables
  + Numeric values associated with “Confirmed Deaths” contained missing values.
  + Numeric values associated with “Primary Care Physicians” contained missing values.
  + Numeric values associated with “Mental Health Providers” contained missing values.

## Missing Data Cleanup

* We were able to leverage several of the techniques we learned in ISQS-6339. One of the techniques we used to reduce bias was to calculate the average values of a column and update the missing values with the column average
* In other instances, we dropped columns not necessary for analysis or leveraged the fillna() to populate missing values with the 0.

# Data Merging

## Common Elements

* 5 Digit FIPS Code
* State
* County

## Multi-level Measurements in final dataset

* The CDC-SVI dataset is census-based, so the original granularity is by tract. The desired granularity was county, which is made up of several tracts.
* This required a grouping in the get\_cdc\_svi\_data(): ETL function to roll up tracts into their respective counties.
* The decision was made to group by the sums of the raw values measures like population, and the means of the percentile attributes.

## Variables More Valuable when Combined

* The County Health Rankings dataset was divided into “raw” fields that represented different ratios from row to row, “numerator” fields that represented the actual raw counts, and a “denominator” that was also not consistent.
* After consulting the documentation, we opted to create ratios from the numerator and the relevant population which was also provided.

## Value of Merged Datasets

# The large collection of established population health and economic indicators was specifically chosen to help us uncover possible correlations between any of those variables and Covid-19 metrics at a glance.

# This virus and the problems it presents to public health are new, so the value we are trying to glean is determining which factors are worth exploring and further refining in future research.

# Visualizations and Analysis

## Characteristics of good visualization

* We chose to move forward with visualizations that would easily communicate our message to the end user.
* We were able to adhere to the key characteristics of association by ensuring similar objects were grouped together and differentiation choosing not to group dissimilar objects.
* We also made it a point to introduce the map of the great state of Texas, which is familiar to all users and thus taps into the characteristic of prior knowledge.
* This is a key element that helps achieve our goal of helping the end user consume the image in an effortless way, so they may focus on the analysis.

## Natural Processing

* We feel our graphs adequately observe the guidelines for natively processing images primarily because we kept out graphs simple, without the bells and whistles that could introduce confusion.
* By keeping our graphs informative, but minimalistic, we maximize our odds of the end user understanding our data. One characteristic that might have helped end users more easily consume our graphs is rendering in 3D, considering we are born into this world viewing 3D objects.
* Our Raw\_Covid\_CorrMatrix worksheet in Tableau was necessary to provide the end user an overview of the various variables. While we saw this graph as necessary, it is likely the more difficult to consume from an end user perspective.

## Visualizations

While exploring the final merged dataset, a few decisions were made at this stage to support our exploration of possible relationships:

* Calculated fields for Covid cases and deaths *per-capita*
  + The data is presented at a per-county level and we are usually examining percentages, so raw counts weren’t providing useful analyses.
* Moving Averages
  + Covid-19 reporting can be inconsistent, and the erratic spikes can obscure obvious trends. This is a compromise with our data that could probably be addressed with more advanced statistical modeling.
* Polynomial regression
  + Nearly all attributes were better represented by a polynomial curve to the second power, whether the correlation was positive or negative.

## Correlation Matrix

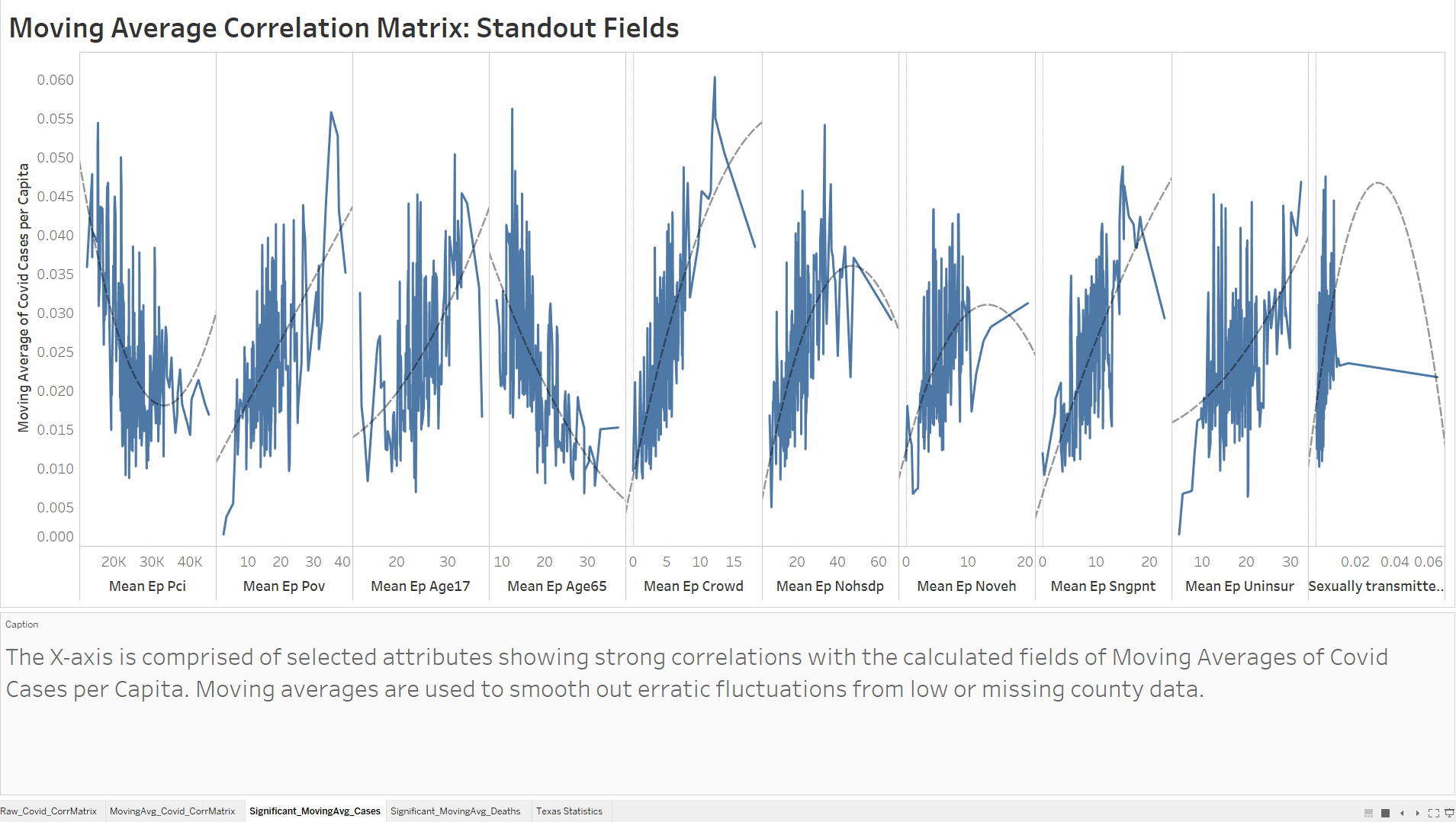
A picture containing timeline

Description automatically generated

* As discussed in the overview, our primary objective was to dig through a large number of factors to find possible correlations across different lenses related to Covid-19.
* We began with a correlation matrix to see which of these factors appeared to fit our lines, the highly correlated fields are pulled out and discussed below.

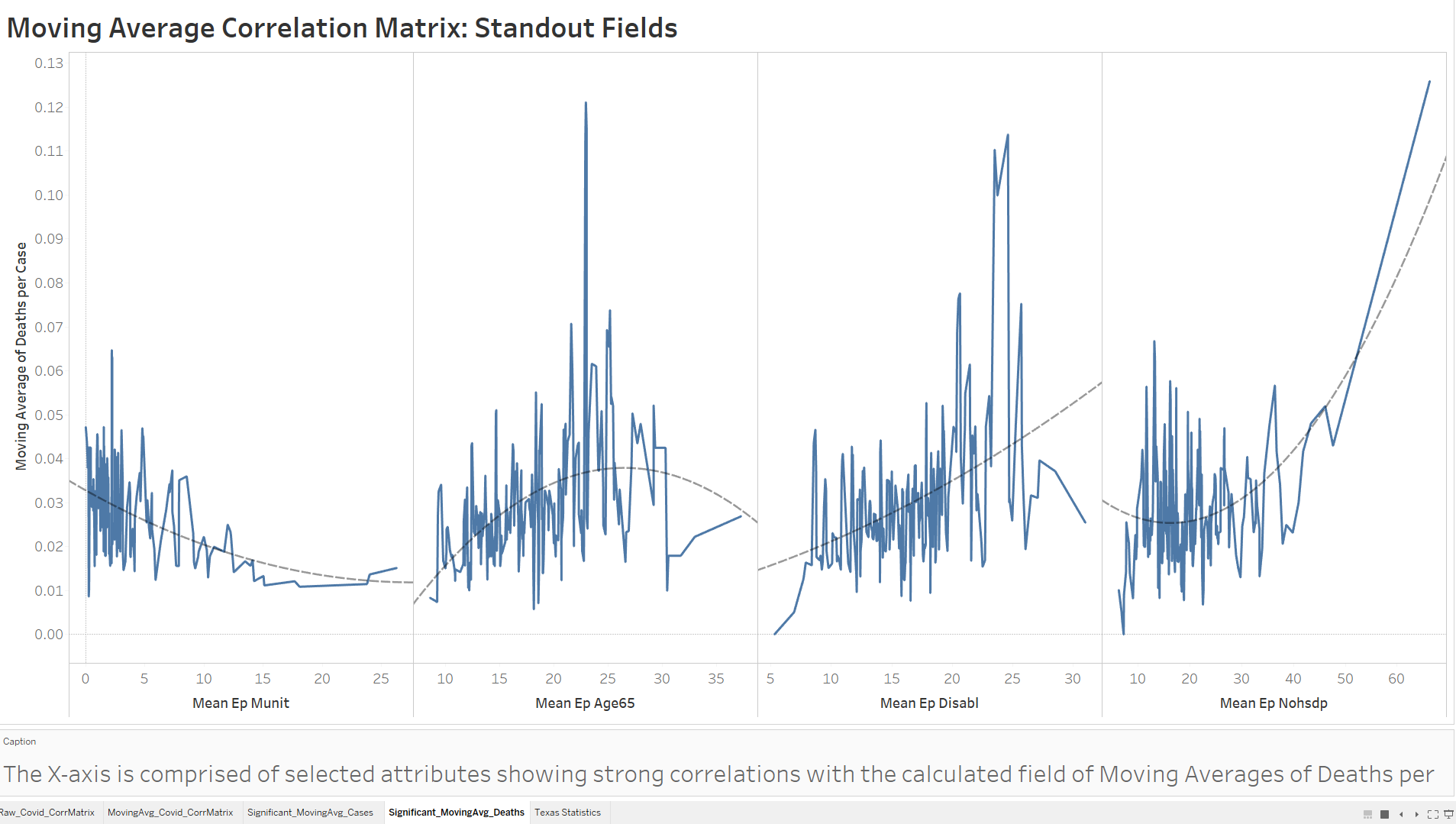
## Visualizations

### Factors Correlated with Covid-19 Cases



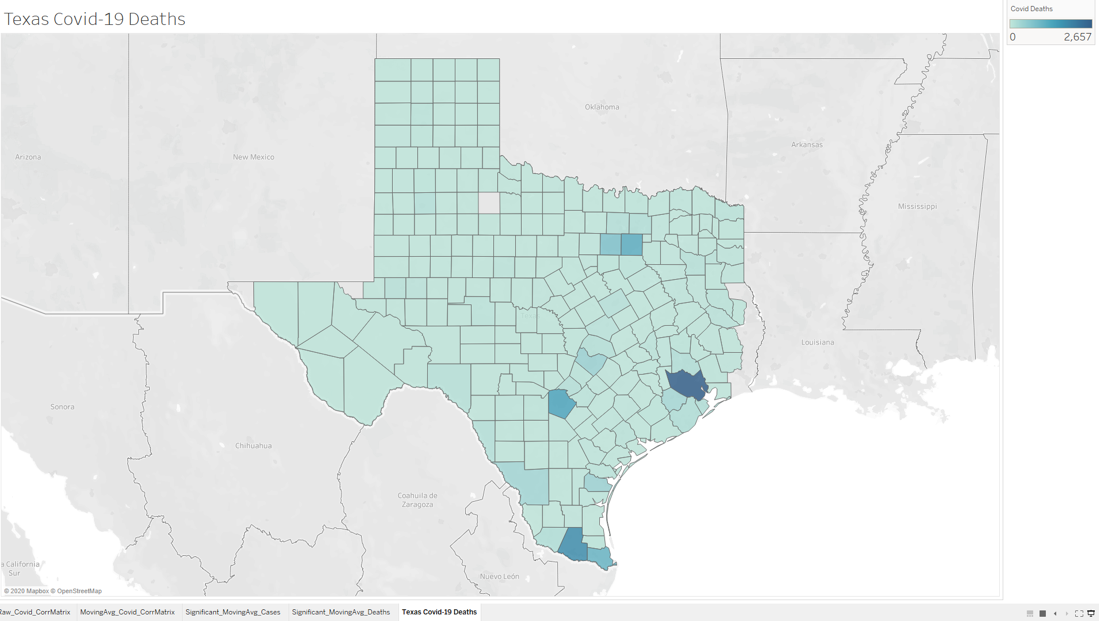
Calculated fields of Covid Cases per Capita, Covid Deaths per capita, and Deaths per Covid Case are show with polynomial (second power) regressions.

Factors Correlated with Covid-19 Deaths



Calculated fields of Covid Cases per Capita, Covid Deaths per capita, and Deaths per Covid Case are show with polynomial (second power) regressions.

### Geographic Distribution of Covid-19 Cases



# ETL Flow

Diagram

Description automatically generated

# Code Instructions

## Function Overview

* Support function
  + divide\_this(): divide, otherwise return 0
    - used for constructing the ratios in the health rankings data
* Get functions
  + get\_health\_ranking\_data()
    - ETL function for health rankings
  + get\_texas\_covid19\_data()
    - ETL function for Texas covid-19 data
  + get\_cdc\_svi\_data()
    - ETL function for CDC Social Vulnerability Index (SVI)
* Main function main()
  + App entry point handling the merge and creation of final csv
  + Final CSV is named final\_merged\_datasets.csv and is saved to the same directory as the file

## Requirements

* Can be found in the Files and Paths section after library import
* CDC SVI Data
  + The CDC SVI data is a local CSV in the same directory as this python file
  + ==If the CSV name is changed from SVI2018.csv, update the argument to os.path.join==
* Python file name
  + The name of the python file is an input to finding the local csv and performing the final export
  + ==If file name is changed from "Covid\_Final\_ETL.py, update the argument to os.path.abspath==

## Misc

* us-counties and analytical\_data2020\_0.csv are provided as reference, but they are not required to run the ETL as these are pulled in with GET requests
* This file uses #region and #endregion markers used by VS Code if they are supported by your IDE