Capstone Progress Report

Concept Slides

Data Science Capstone & Ethics (ENGI 4800)

Alexandra Sudomoeva (as5402)
DongGu Kim (dk2983)
Haotian Zeng (hz2494)
Chengzhang Xu (cx2188)
Yang Gao (yg2499)

Why SMEs?

SMEs - small and medium enterprises with total employee count below 500 people.

Small business growth and performance fluctuations represent an exciting and relevant research subject for a number of reasons:

1. SME performance can be used to define the national financial health
    - They drive **48%** of the job market in the US
    - They constitute approximately **52%** of net job growth
2. SME deaths can have serious negative consequences on the economy
    - Only **50%** of all new small businesses survive after the first 4 years
    - SMEs generate a significant amount of innovation that further fosters the economy

Problem Statement

Project Focus: For the purpose of this study, we wanted to focus on discovering and better understanding the drivers behind small and medium business formation, growth, and dissolution.

Goal: Forecast and measure SME growth distribution based on the quantitative dependency derived by test a number of factors suggested by previous research

Final Outcome:

- High accuracy forecasting model around SMEs' future growth by state
- Detailed report on significant predictive variables found during regression analysis

Related Literature

Most research tends to overlap on four main areas as drivers of SMEs' performance:

1. Significant economic fluctuations like trade wars, crises, economic downturns *(Kirchhoff; Brown and Batista; Broke and Evans)*
2. The financial sources, lending barriers, and overall health *(Brown and Batista)*
3. Regional conditions around taxation, population density, and general state/county economy *(Nicholls and Foreman-Peck; Dachin and Rusei)*
4. Third-party variables such as union strength, big corporation takeovers, and individual industry indicators *(Nicholls and Foreman-Peck; Kirchhoff)*

Data Collection and Description

The data collection process was strongly guided by the things learned from related research. We have aggregated our pull on an annual basis spanning for 1992-2016.

Target (Bureau of Labor Statistics, United States Courts)

- Count of firms in each growth category (birth, death) by company size

Regional (National Science Foundation; United Census Bureau; Bureau of Labor Statistics)

- Education level, union representation, region area and other descriptives

Economic (Bureau of Economic Analysis; United Census Bureau)

- Regional unemployment, GDP, personal income, regional price deflators

Financial (R - Cran)

- Indexes: NASDAQ, Dow Jones, Russell 1000
- Individual Companies: Walmart, GE, Coca-Cola, Caterpillar

Initial Data Exploration: Overview

We chose to prioritize the arguments about the importance of geography into our data structure due to a prevailing number of literature on the subject as well as the rich variety of features available.

The questions we attempted to answer were the following:

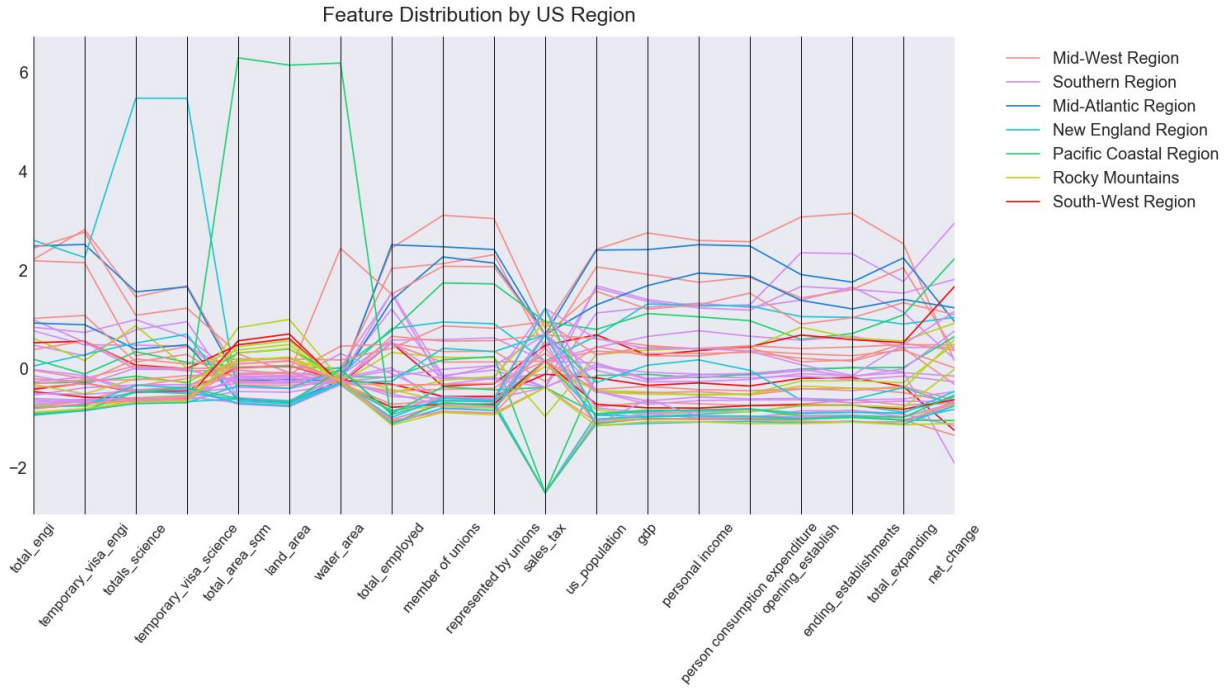
- ➔ What are some of the general characteristics of chosen features for SME performance?
- ➔ How is SMEs' growth distributed across various states? What are some of the factors that might skew the distribution (if any)?
- ➔ Which feature variables show the highest possible correlation with the growth target for counties and states? And what is the direction of the relationship (positive vs negative)?
- ➔ Is there any multicollinearity between tested SME performance drivers we should consider in further developing the regression model for states and counties?
- ➔ Does big corporation growth has an effect on SME deaths in the corresponding region (corporation's headquarters)?

Initial Data Exploration

What are some of the general characteristics of SMEs' performance and tested features?

A geographic pattern can be observed:

The Rocky Mountain region has low value of GDP and personal income, while the Mid-Atlantic region has higher value of total number of employees.



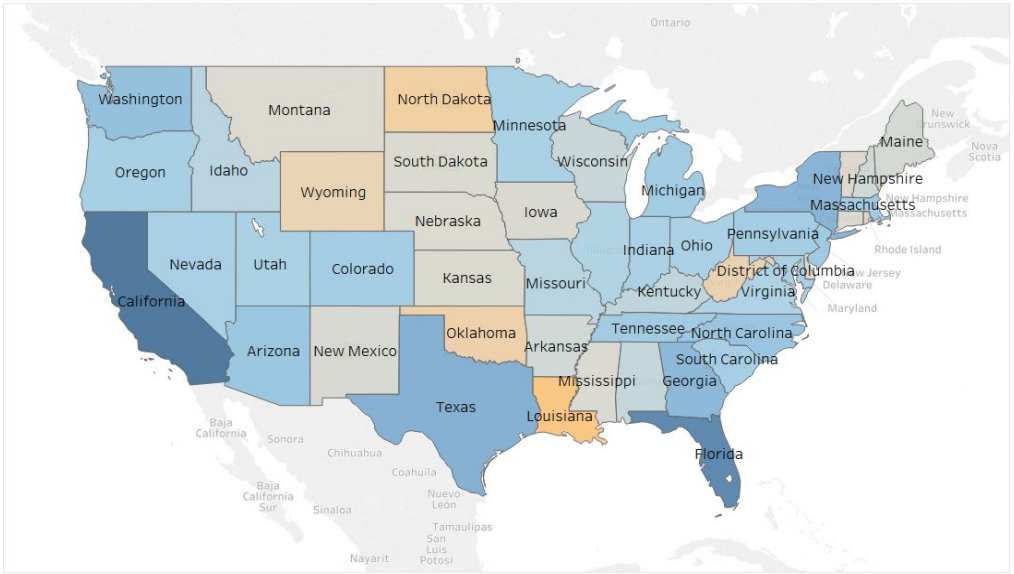
Initial Data Exploration

What are some of the general characteristics of SMEs' performance and tested features?

A geographic pattern can be observed:

Louisiana, North Dakota, Wyoming, Oklahoma show a high density of company death, but most states remained on the positive side of the scale.

Net Change by state



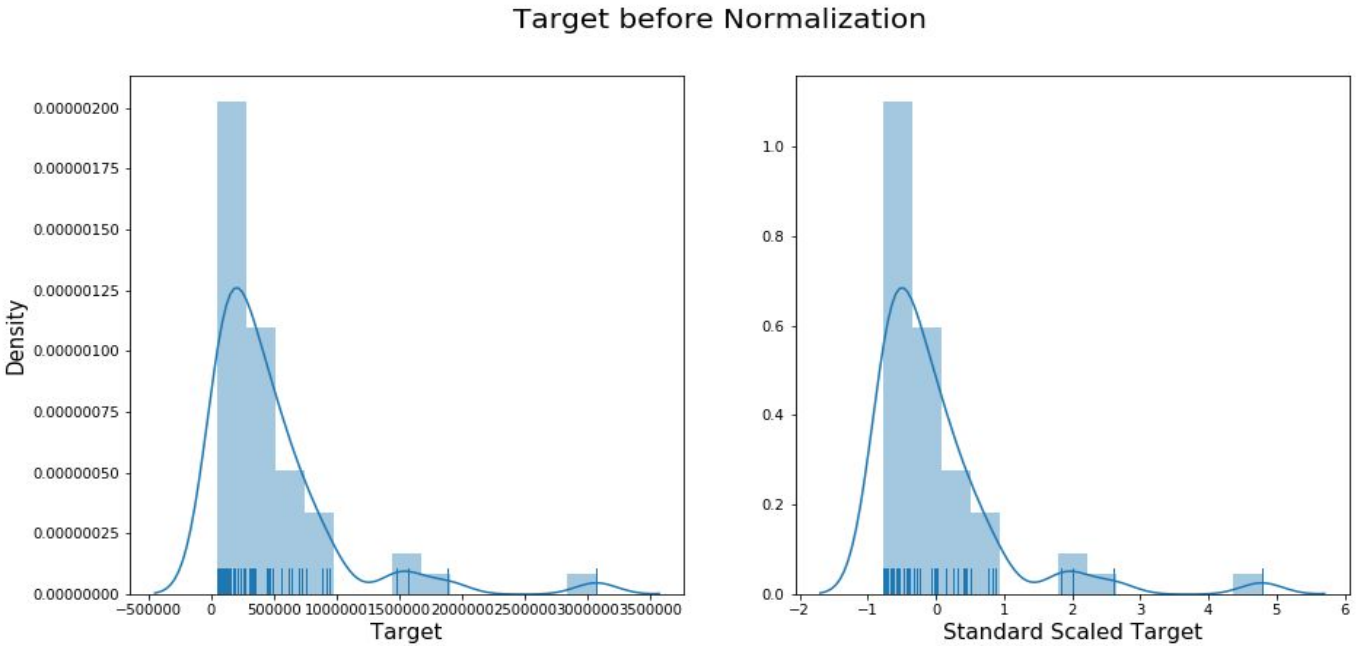
Map based on Longitude (generated) and Latitude (generated). Color shows sum of Net Change. The marks are labeled by State. Details are shown for State. The view is filtered on State, which excludes Alaska and Hawaii.

Net Change

-59,908 555K

Initial Data Exploration

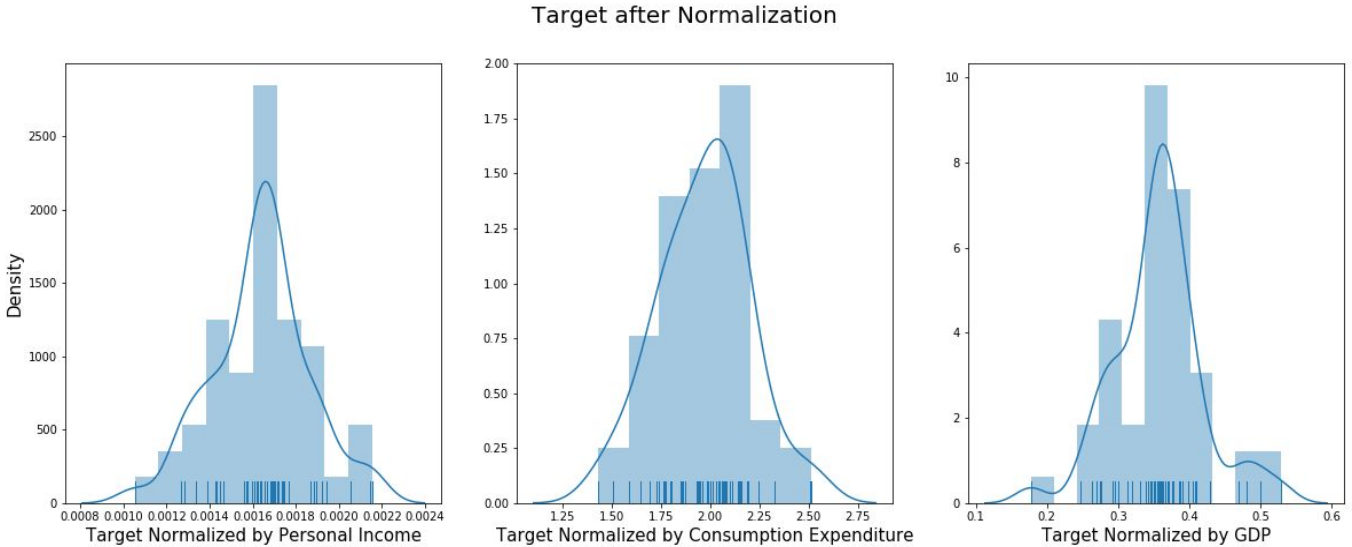
How is SMEs' growth distributed across various states?



Distribution of raw/standard-scaled target-value appears to be extremely skewed to left. It needs to be normalized for better quality of analysis and training ML models based on probabilistic assumption.

Initial Data Exploration

How is SMEs' growth distributed across various states?

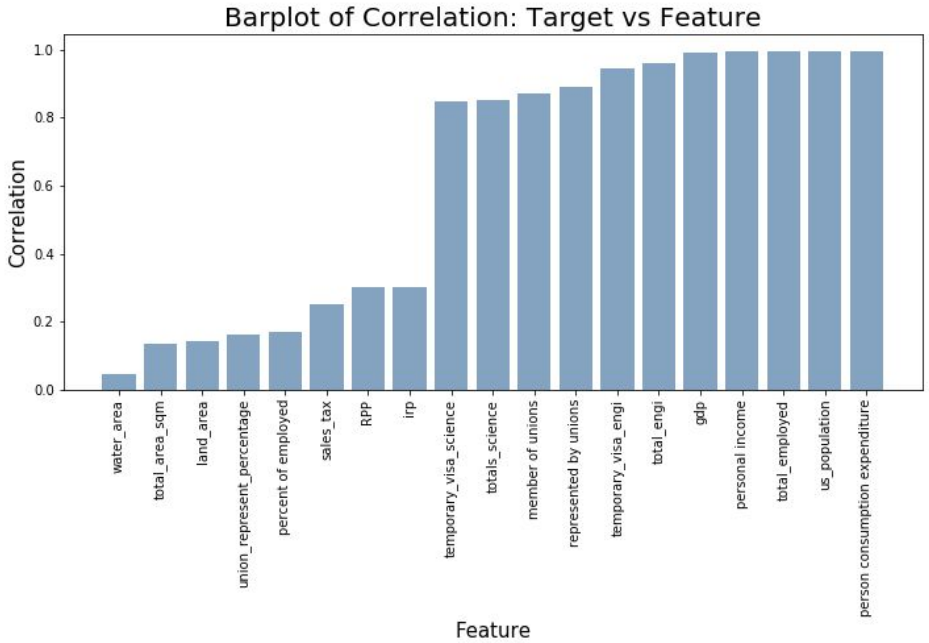


We chose to test some of our data features as normalization factors to examine how well each factor performs. The distribution of target normalized by personal income appears to fit better into shape of normal distribution. Hence, personal income has the strongest impact on SMEs growth.

Initial Data Exploration

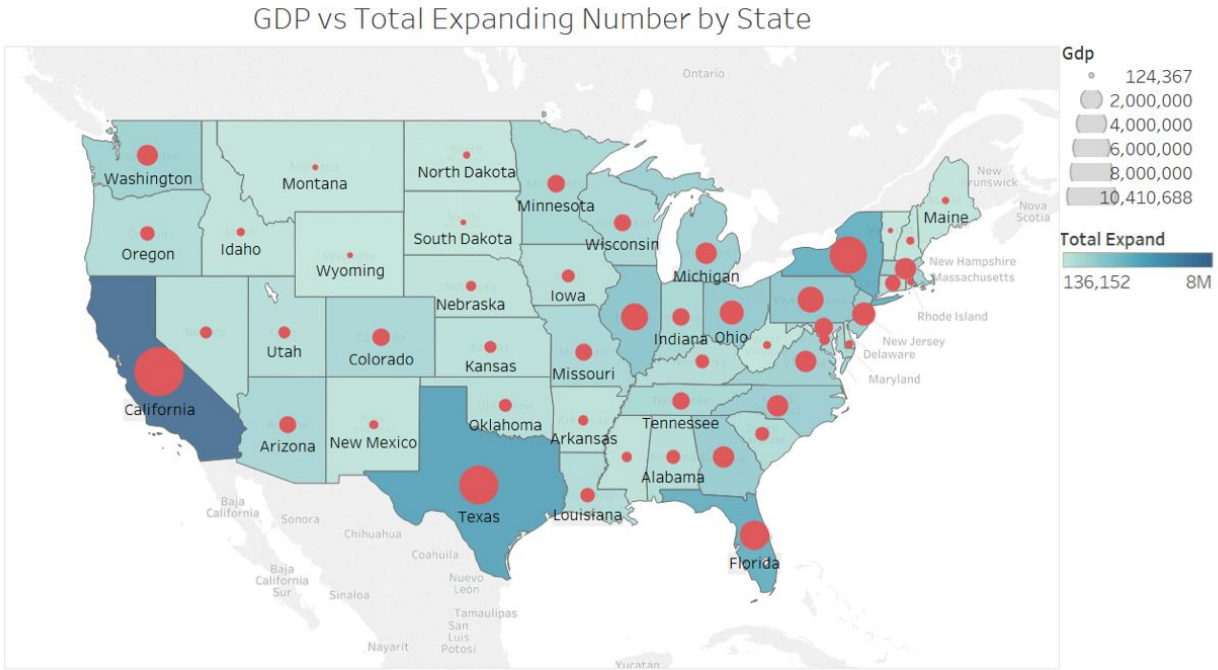
Which feature variables show the highest correlation with the growth target?

- Most features tested have a relatively high correlation with target
- Correlation of percent of employed is low while that of total employed is high
- Possible explanation is that distribution of target value before normalization is highly skewed by absolute economic size of state rather than relative records



Initial Data Exploration

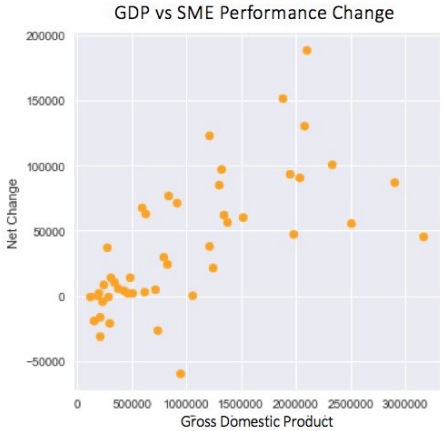
Which feature variables show the highest correlation with the growth target?

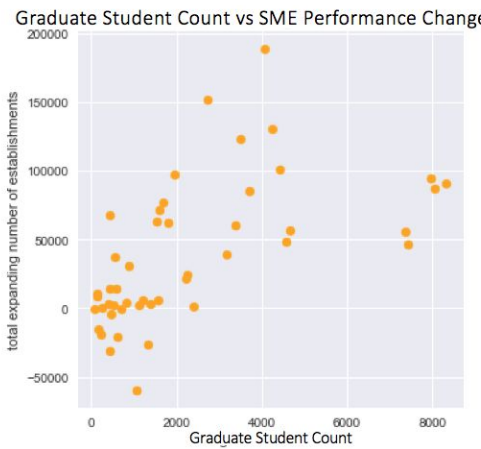


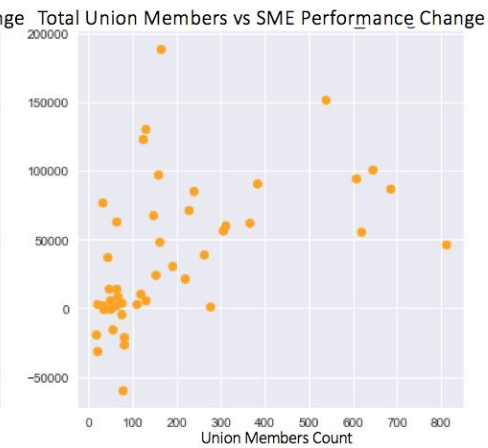
The map plot below reveals that states with higher GDP tend to have a higher SME growth.

Initial Data Exploration

Which feature variables show the highest correlation with the growth target?



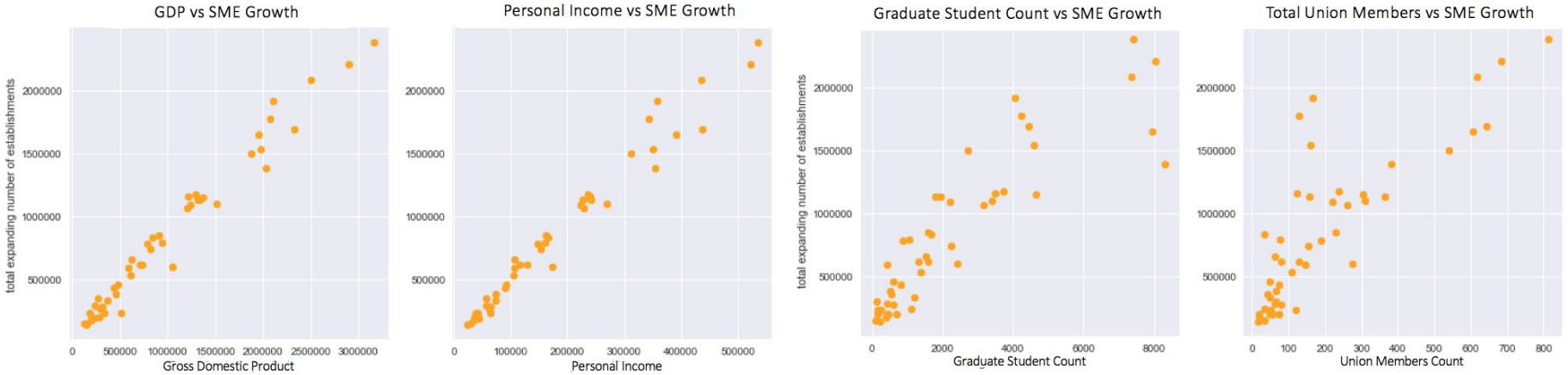




The scatter plots show that there are some positive relationships between *GDP* and *personal income* with the *net change number of establishments*.

Initial Data Exploration

Which feature variables show the highest correlation with the growth target?

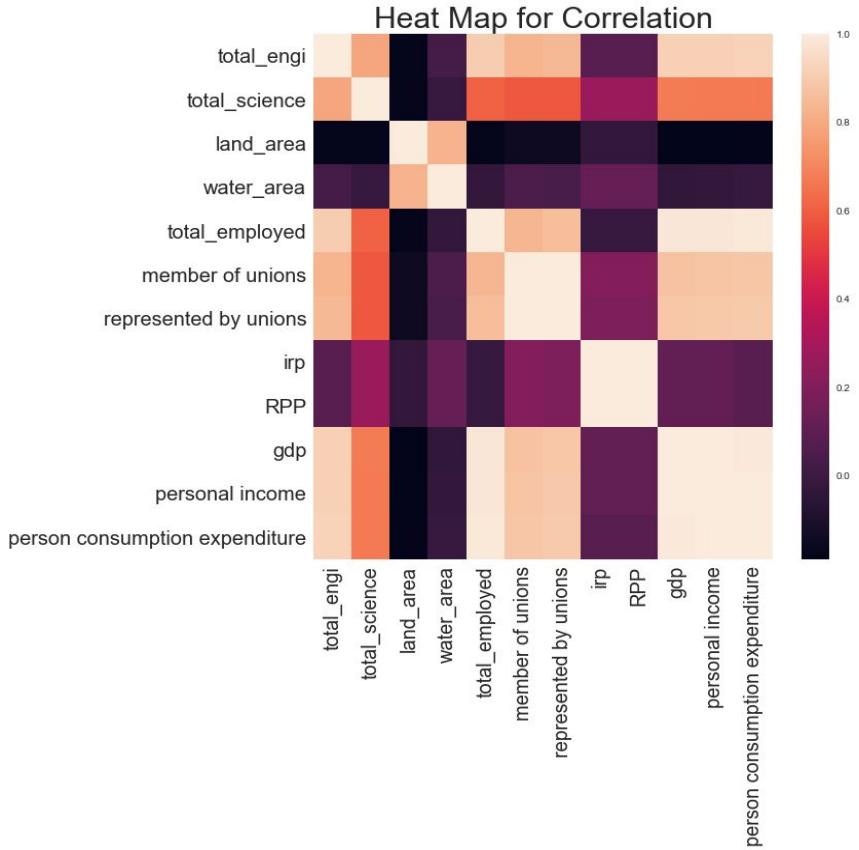


There are strong positive relationships between *total numbers of expanding and opening establishments* with *economic indicators*.

Initial Data Exploration

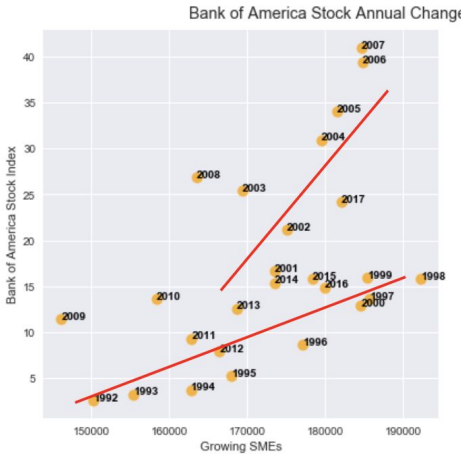
Is there any Multicollinearity?

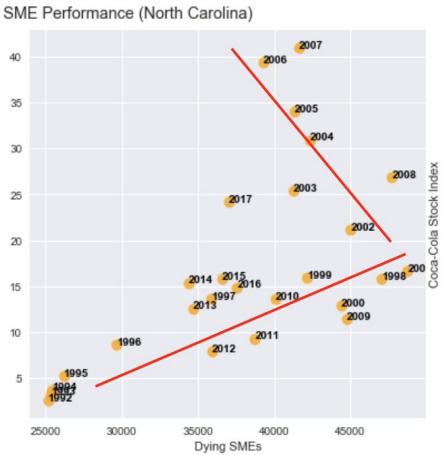
- Interdependency between tested variables can create a strong bias in our model
- Most variables show relatively low levels of correlation
- But some variables should not be included in the same model:
  total number of graduate students has high correlation with employment, unions, and economic variables

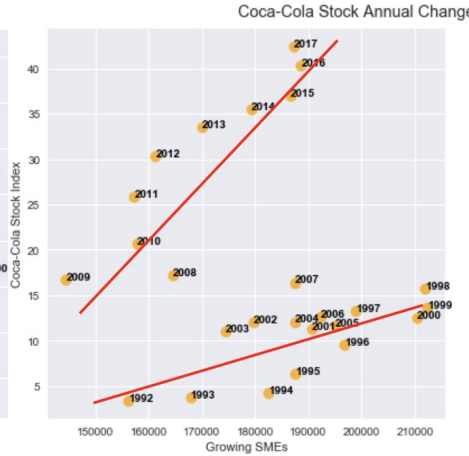


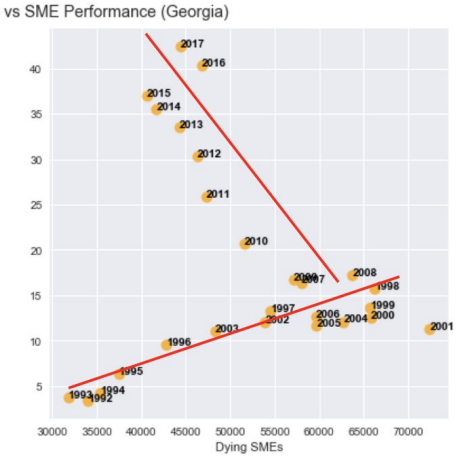
Initial Data Exploration

Will big corporation growth affects SME deaths?









- In North Carolina, *Bank of America* led some negative relationship associated with SME deaths from 2000 to 2007 and from 2010 to 2017.
- In Georgia, *Coca-Cola* has a varying strong relationship associated with SME deaths since financial crisis occurring in 1998 and 2008.

Next Steps

First Cycle

- Explore how the combination of variables affect our target value
- Train Machine Learning models and search for the optimal parameter setting
- Evaluate critical drivers that affect success and failure of small businesses

Beyond that

- Search for additional features to improve performance of model
- Process text data and examine how sentimental score is related to small business economy

Our Goal

- Identify critical economic indicators affecting small business pattern
- Reach an accurate model forecasting trends in small business or target value

Team Member Contribution

In putting the first progress report together, each team member has contributed to a variety of analysis:

- *Alexandra Sudomoeva* has worked on the introduction, literature review, data collection and description as well as the concept slides
- *DongGu Kim* has also contributed to the literature review and most of the visualization analysis for second and third questions of interest
- *Haotian Zeng* focused his forces on data cleaning and analysis also contributing to the concept slides
- *Chengzhang Xu* was the driving force behind most of data collection and visualization that focused on answering third and fourth questions of interest
- *Yang Gao* was responsible for collecting the finance data as well as significantly contributing to the analysis of the first and fifth questions in the EDA segment