Conner Sexton
Wesleyan University and CT Data Collaborative
QAC Praxis Lab
25 May 2020

Economic Development and Migration in Connecticut

Introduction

The reliance on a particular industry for the sustained livelihood of a given midsized city's economy, becomes increasingly difficult with time. Today's economy favors the high-tech businesses, which, in turn, has led to the migration of big businesses to and from cities across the country. The migration of high-tech corporations into, usually larger, cities has resulted in an influx of young, educated graduates into these larger cities, which leads to more than just economic growth. This results in the increased cost of housing, increase of amenities within cities (e.g. increased need for transportation and housing infrastructure, and the gentrification areas that are often home to older, less educated populations). On the other hand, midsize cities, are struggling to keep up with this change, often experiencing the migration of large companies out of their economies, seeking the larger cities that have already attracted the young, educated workforces that they desire. Hartford and New Haven, Connecticut both serve as examples of this condition, both situated in close proximity to elite universities (Yale, Trinity, Wesleyan, etc.), but are experiencing the migration of companies such as General Electric and Alexion Pharmaceuticals from their economies. Connecticut is situated in commuting distance from both Boston and New York, which are large, well established high-tech cities that have the jobs, amenities, and atmosphere that university grads are so attracted to. This begs the question of how midsize cities can adapt to this changing economic landscape. Especially when they are situated in close proximity to elite universities that are churning out yearly cohorts of a young, educated workforce. A workforce that is attractive to these high-tech companies.

Connecticut has experienced a slow resurgence since the 2008 recession, with its 2017 economy coming in lower than its economy back in 2004, whereas over the past ten years, the "nationwide unemployment rate at a mere 3.8 percent, and [an] economy growing by 2.7 percent in 2018" (Yale Daily News, 2019). Additionally, the nondurable goods manufacturing industry and the finance and insurance market, two of Connecticut's major industries, have faced huge contractions between 2008 and 2018 (75% and 30%, respectively). Taking this into consideration, how has Connecticut's economy changed over the past decade in terms of employment by sector (are certain sectors contracting/expanding)? And how has the overall economy changed in relation to this? Additionally, in the articles cited in the introduction, above, how have communities changed in relation to the changing economy, over the past years? In terms of demographics, housing costs/availability, income distributions, infrastructure, etc.

Methodology

Due to the scale of this project, many of the forms of analysis used are exploratory. Additionally, our primary model implementation is the use of the random forest machine learning algorithm. We opted with random forests for multiple reasons. Firstly, its use of the bagging resampling with replacement method proved ideal considering the small amount of observations we were working with. Additionally, random forests return a variable importance metric, which measures the importance of each of our predictors in reducing the error term when

predicting our response. This allowed us to narrow down our scope, from the original 132 predictors.

Next, we used K-means cluster analysis. This method produces representations of our MSA's in a vector space, based on an inputted number of predictors (top N from random forest variable importance), and an inputted number of clusters. The algorithm then measures the distance of our MSA's from one another based on scaled values of these predictor variables. In performing this calculation, we are able to get a better sense of which MSAs are most similar/dissimilar to one another.

Finally, we perform an array of post-hoc analyses implementing basic linear regression models and pearson correlations to get a better sense of which attributes correlate highly with real GDP per capita. In so doing, we can characterize which attributes make for a 'more economically successful' MSA, using GDP as a gauge for this.

Data

The data used in this study was compiled from multiple sources[1]. Firstly, for geographic and time series purposes, we used a general time period of 2010 to 2018 (this can vary depending on the availability of data) and core base statistical areas (CBSA). For CBSAs, which include metro and micropolitan areas, we included all Connecticut CBSAs with the addition of the New York-Newark-Jersey City and Boston-Cambridge-Newton metropolitan statistical areas (MSA). The latter two MSAs were included to represent the New York and Boston metro areas, respectively.

For data on GDP, we used the Bureau of Economic Analysis's developer API tool to query MSA-level data from 2010 to 2018. We used the U.S. Census Bureau's new data tools to collect demographic data on age, educational attainment, employment statistics, household income, and population dynamics. Finally, for establishments by MSA, we used the U.S. Census Bureau's Statistics of U.S. Businesses repository for establishment numbers across MSAs as well as metrics on establishment sizes. The Census Bureau's SUSB repository only contained data up to 2016, therefore analysis that required the use of these metrics needed to be restricted to the time frame of 2010 to 2016.

Real GDP per Capita Overview

This study uses rGDP per capita as a primary response variable to gauge the economic development/success of a given MSA. We used data from the Bureau of Economic Analysis and the U.S. Census Bureau to calculate rGDP per capita values (GDP and population). As reflected in Figure 1, Connecticut's Bridgeport-Stamford-Norwalk MSA has the highest rGDP per capita of our MSA's of interest until around 2017/2018. Additionally, Boston-Cambridge-Newton and New York-Newark-New Jersey have similarly consistent growth in rGDP per capita from 2010 to 2018, with Boston's metro area eclipsing Bridgeport-Stamford-Norwalk in 2018. For clarity, we have delineated Connecticut MSA's with dashed lines to differentiate these from New York and Boston metro areas. The high rGDP per capita of Bridgeport-Stamford-Norwalk must be investigated further, but this is most likely due to its commuter economy. This MSA's proximity to the New York metro area, and access to the Metro North railway, allows its workforce to commute to the city for often higher paying jobs in New York and lower housing costs of suburban Connecticut.

[1] A more thorough description of our different data sources, codebooks, etc. can be found in our data descriptions document.

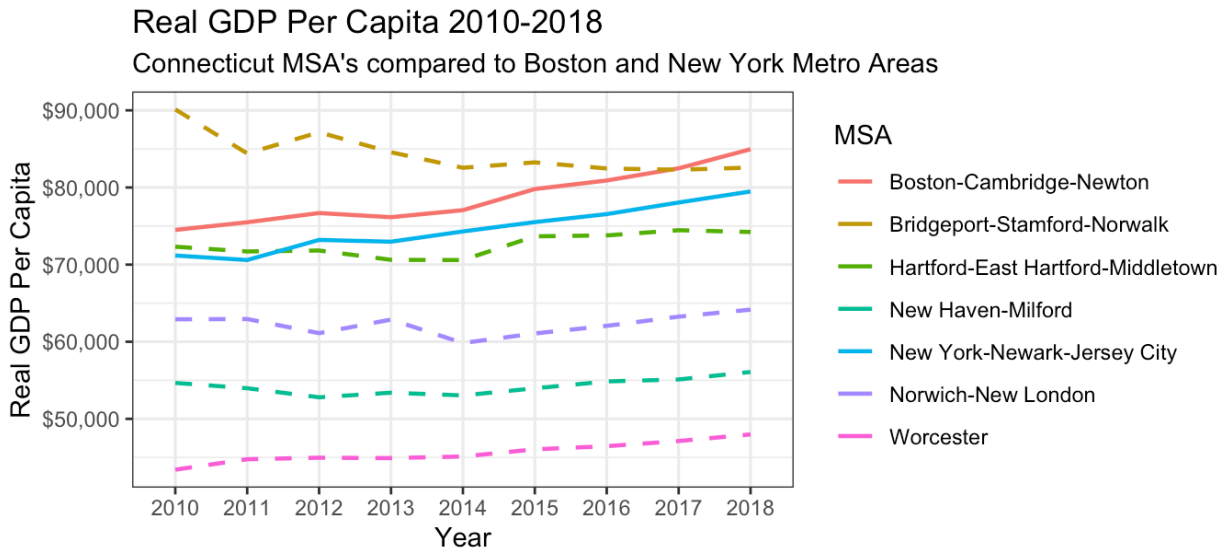

Figure 1

When observing the percent change in rGDP per capita, the Boston and New York metro areas have the highest percentage change, consistently, from 2011 to 2018. As reflected in Figure 2, all of our observed MSA's took a hit in 2013, with the percent change in rGDP per capita dipping below 0%, which can most likely be attributed the lasting effects of the Great Recession. Additionally, as Boston and New York metro areas recover in 2014, with a positive percent change, Connecticut dips farther into the negatives, with a 1.49% decrease in rGDP per capita across its 5 metropolitan statistical areas.

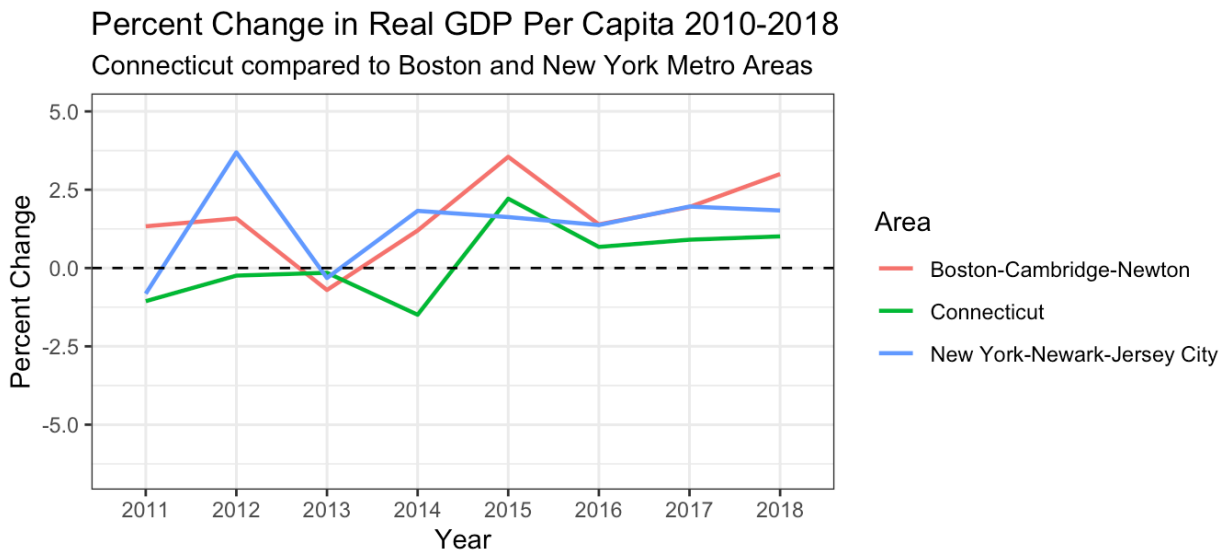

Figure 2

Machine Learning Approach

We decided to use the random forest machine learning approach to get a better sense of which attributes are most predictive of a given MSA's real GDP per capita. Random forests are beneficial in this sense as they output a variable importance metric, which gives us an idea of which predictor attributes are most influential in predicting our response.

Random forests are an ensemble machine learning method, meaning they combine regression trees with the bagging resampling method. Multiple "weak trees" or weak models are run on samples of our training set (sampling with replacement) and the predictions from these models are aggregated by taking the arithmetic mean. While this sacrifices some interpretability, it offers far more accurate predictions than traditional linear regression methods and allows us to understand how important our predictor variables are to the overall model.

To train our data, we split our sample with a 70/30 split, where 70% of the observations were randomly selected for our training set and 30% for the testing set. Our final model used 100 distinct, parallelized regression trees, with a root-mean-square-error term of 4,929.68 and an R-squared value of 0.933. Therefore, our model was able to predict within $4,929.68 of the observed rGDP per capita and our predictors explained 93.33% of the variance in our response variable. Table 1 and Figure 3 display the results of our model and top 20 most important variables respectively:

Model Results:	RMSE	R-Squared	MSE
	4,929.68	0.933	3,350.24

Table 1 Random forest model results.

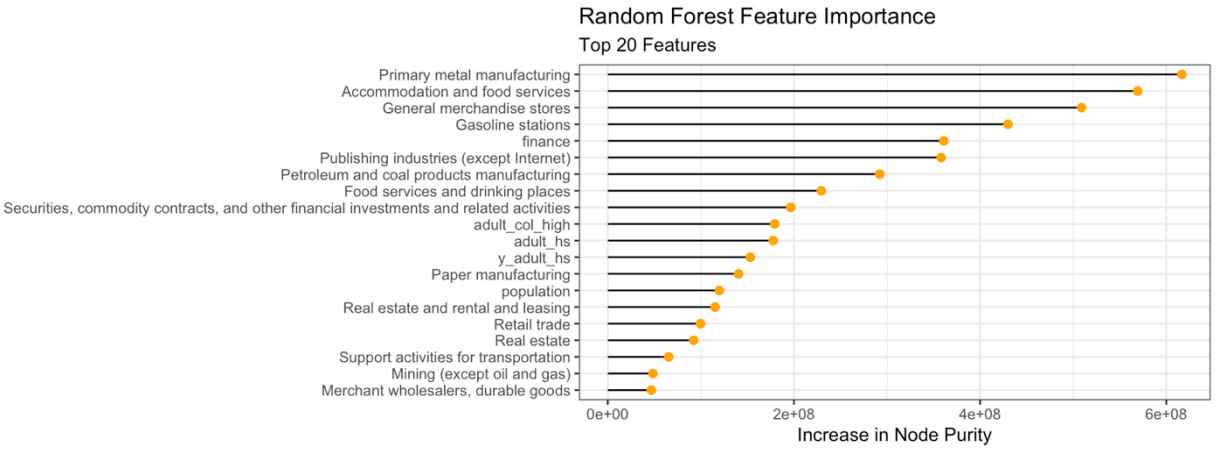

Figure 3

Considering our model originally had 132 distinct predictor variables, reducing these to 20 important features is helpful to get a better sense of which predictors were most influential in predicting rGDP per capita. Notably, many of the features reflect NAICS industries in the amenities and housing areas. For example: accommodation and food services (2), general merchandise stores (3), gasoline stations (3), food services and drinking places (8), real estate and rental and leasing (14), real estate (17), support activities for transportation (18). These NAICS features reflect the percentage of all establishments in a given MSA that the respective NAICS industry accounts for. Additionally, some of these features reflect the educational attainment percentage of a given MSA's population (e.g. % of adults who have completed high school, have completed college or more, and young adults who have completed high school).

Finally, the percentage of an MSA's population employed in the finance sector (feature: finance (5)) is highly influential in predicting rGDP per capita.

Due to the random forest algorithm's lack of the coefficients we would traditionally interpret from a linear regression model, we must conduct some post analyses on the aforementioned features to get a better sense of how the affect an MSA's rGDP per capita. Therefore, we ran a regression tree model to see which features it would select and how these features contribute both to the prediction or rGDP per capita and how they influence different predictions. Table 2 and Figure 4 display the results and visual representation of this regression tree model, respectively:

Model Results:	RMSE	R-Squared	MAE
	5,295.85	0.866	4,153.5

Table 2 Regression tree model results.

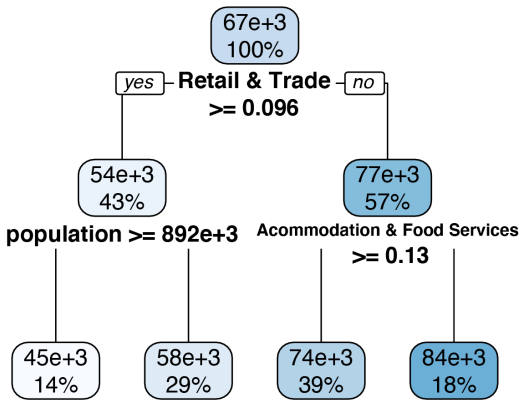

Figure 4 Regression tree final model - visual representation.

This regression tree model used 100 trees, just as our random forest model had. While this model does not predict as well as our random forest model, it allows us to interpret how our most important features influence the prediction of rGDP per capita. As displayed in Figure 4, the three features the model chose were retail and trade, population, and accommodation and food services. Additionally, this shows that the higher an MSA's percentage of both retail and trade establishments and accommodation and food services establishments, the higher our predicted rGDP per capita is. Our initial hypothesis states that a greater focus on amenities will attract a young, educated workforce and hypothetically reflect an increased rGDP per capita. While these results show that amenity establishments correlate with a higher rGDP per capita, they do not necessarily show the relationship between this young, educated workforce and amenity establishments, therefore we will need to perform some post-hoc analyses to get a better sense of this relationship.

The following post-hoc analyses will be largely informed by the variable importance methods we employed through our random forest modeling. As our random forests model helps us narrow down which features in our sample are most predictive of rGDP per capita, this drastically simplifies which features should be focused on, out of the 132-feature sample. More

specifically, we will focus on population, the education/age of MSA populations, employment of MSA populations, and finally, the NAICS establishment characteristics across MSA's.

K-Means Cluster Analysis

To get a better sense of which MSA's are most similar to one another, based on the aforementioned 20 attributes, we performed a K-means cluster analysis. For ease of interpretability, we chose 3 clusters, which produced a within sum of squares (WSS) value of 73.6%. This measure evaluates the variability within each of the three clusters, which in our case is fairly high. Despite a high WSS, the K-means clusters still offer a rough template in comparing the MSA's of interest. Figure 5 depicts the representation of these MSA's in vector space and the relative distance of the MSA's from one another. These distances are used to approximate which cluster each MSA "belongs" to.

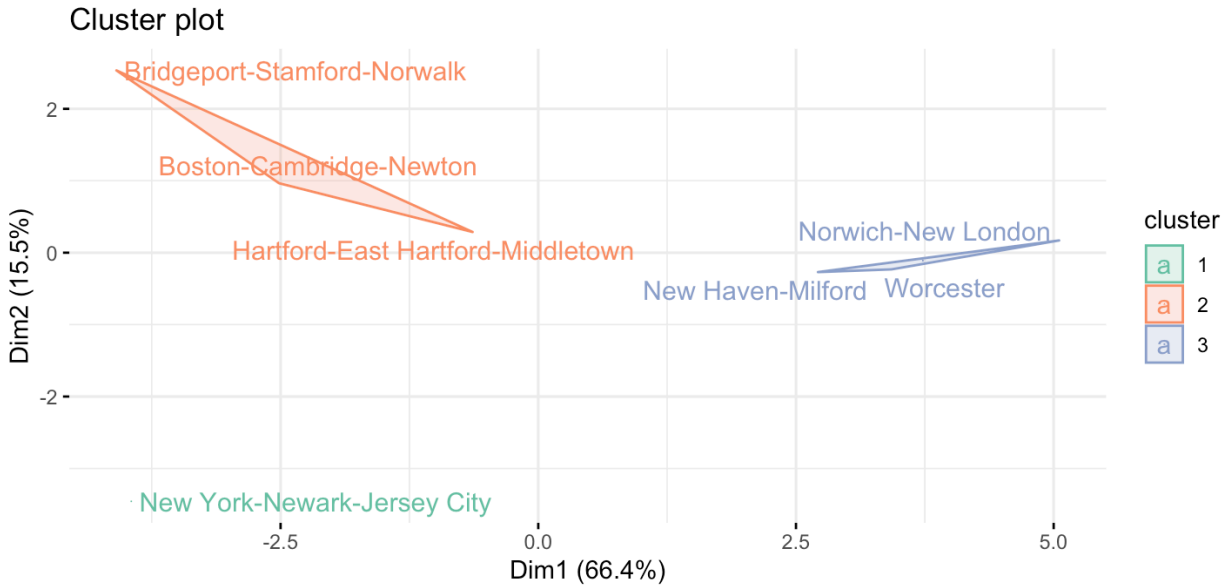

Figure 5

As depicted in Figure 5, the New York metro area MSA is given its own cluster, as it's measures across our 20 attributes are most unique compared to the remaining 6 MSA's. Cluster 2 is also notable in that it shows that Boston's metro area shares similar characteristics, across the 20 attributes, to Bridgeport-Stamford-Norwalk and Hartford-East Hartford-Middletown. Figure 6 shows a breakdown of the scaled attributes across these 3 clusters, allowing us to compare the characteristics of these MSA clusters.

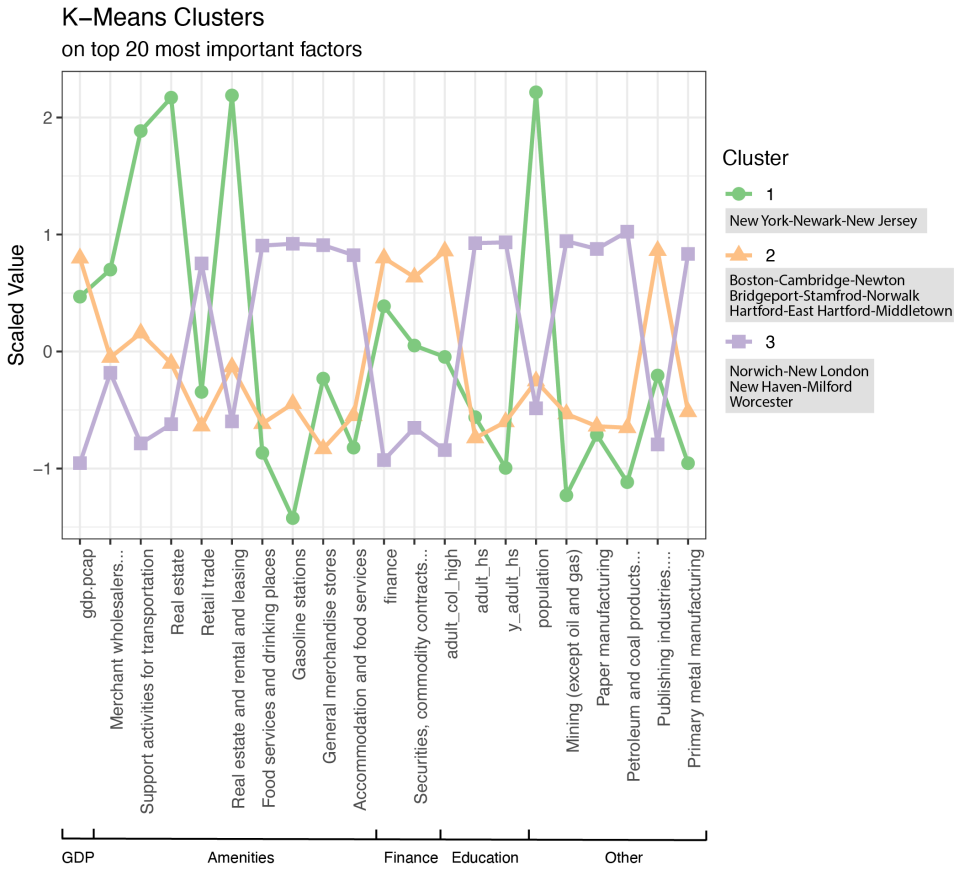

Figure 6

Firstly, cluster 2 has the highest average rGDP per capita across the 3 clusters, which translates to a higher percentage of adults with a college education or higher, individuals employed in the finance industry, NAICS establishments in the securities, commodity contracts area. The New York-Newark-New Jersey cluster is also interesting as it's percentage of NAICS establishments in the general amenities industry is consistently higher than the other two clusters.

Migratory Patterns

An important factor when observing the economic development of a given MSA is the net migration experienced. Net migration is the number of individuals entering a given MSA divided by the number of individuals leaving. We used reported net migration estimates from the U.S. Census Bureau from the years 2010 to 2018. Additionally, to control for varying population

sizes across MSAs, we divided this value by the yearly population estimate of each MSA. Figure 7 reflects these values:

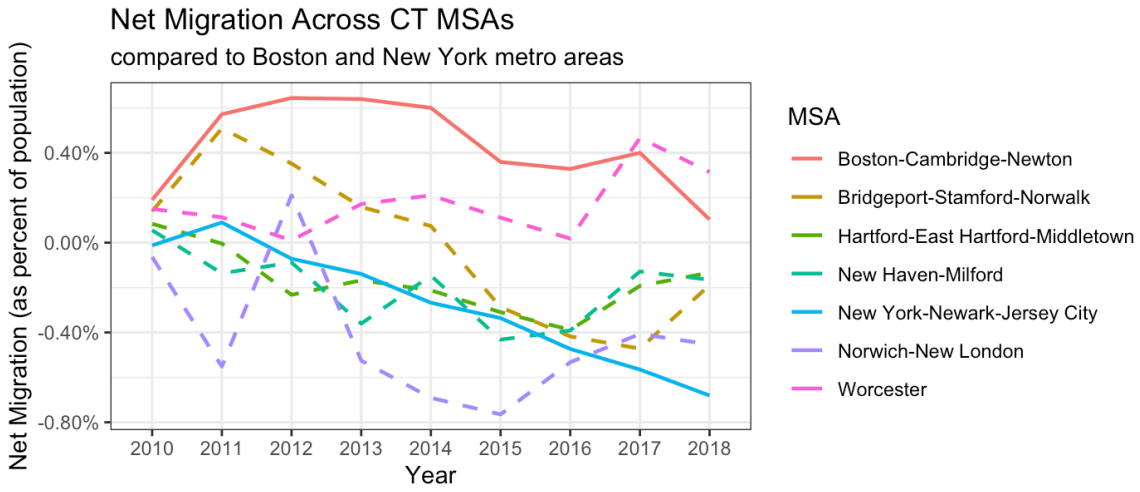

Figure 7

With Connecticut MSA's delineated with dashed lines, we can observe how net migration has changed between 2010 and 2018 for these 7 MSAs. It's clear that the Boston metro area has experienced a steady positive net migratory pattern, which may signal its growth in popularity over the years. New York's metro area has seen a consistent decline in net migration, dipping below 0% in 2012. This signals that there are more individuals leaving this MSA than entering, from 2011 to 2018. Additionally, the Bridgeport-Stamford-Norwalk MSA has an observed decrease in net migration from 2011 to 2017, dipping below 0% between 2014 and 2015. The further decline in in-migration for this MSA around 2016/2017 may be attributed to General Electric's announcement that it would move headquarters to Boston in January of 2016.

Age and Education

Age:

Taking into consideration the hypothesis that MSA's with more successful economies must attract younger, educated workforce populations, Figure 8 and Table 3 reflect the age distributions across our MSA's of interest.

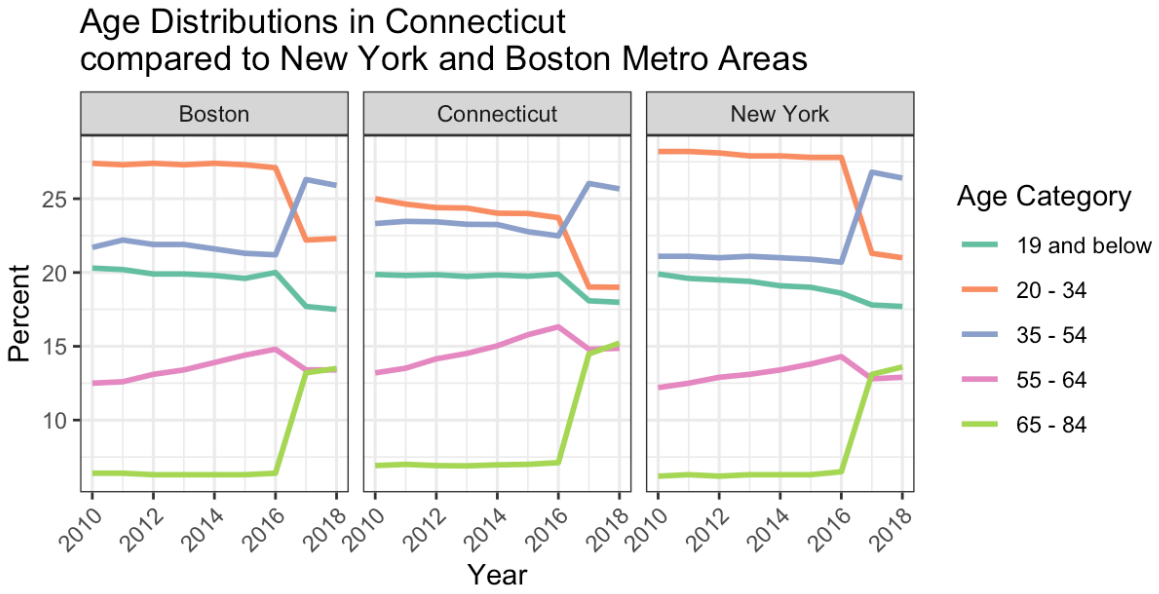

Figure 8

	Age Category:				
	19 and Below	20 to 34	35 to 54	55 to 64	65 to 84
Boston	19.4%	26.2%	22.7%	13.5%	7.9%
New York	19.0%	26.5%	22.2%	13.1%	7.9%
Bridgeport-Stamford-Norwalk	19.9%	23.4%	23.7%	13.4%	9.1%
Hartford-East Hartford-Middletown	19.9%	23.5%	23.2%	14.3%	8.8%
New Haven-Milford	19.9%	24.2%	22.5%	14.3%	8.8%
Norwich-New London	19.4%	23.4%	23.2%	15.1%	8.8%
Worcester	20.2%	23.9%	23.8%	13.6%	7.8%

Table 3 Percentage of population for age categories (averaged between 2010 and 2018).

Based on Figure 8, the Boston and New York metro areas have a greater percentage of individuals aged 20 to 34 years old, when compared to Connecticut as a whole. This aligns with the more granular numbers reflected in Table 3 as well. When analyzed at the MSA level for Connecticut, the New Haven-Milford MSA has a marginally higher percentage of individuals in this age category, which is surprising as Bridgeport-Stamford-Norwalk has a consistently higher rGDP per capita than the other MSA's in our study. While New York and Boston metro areas have consistently higher percentages of young adult populations, this discrepancy within CT MSA's leads us to reconsider the hypothesis of rGDP per capita being correlated with a greater percentage of young adults.

Education:

Educational attainment is an important attribute to analyze as there is a large body of existing literature that has pointed towards a correlation between economic success and high educational attainment within an area's population. The primary category of interest in our study is that of young adults with college degrees or higher (25 years or older, bachelors and higher). Additionally, we analyze the category of adults with college degrees and higher, as this category

was an important variable in our rGDP per capita prediction model. The following Figure 9 and Table 4 display these metrics:

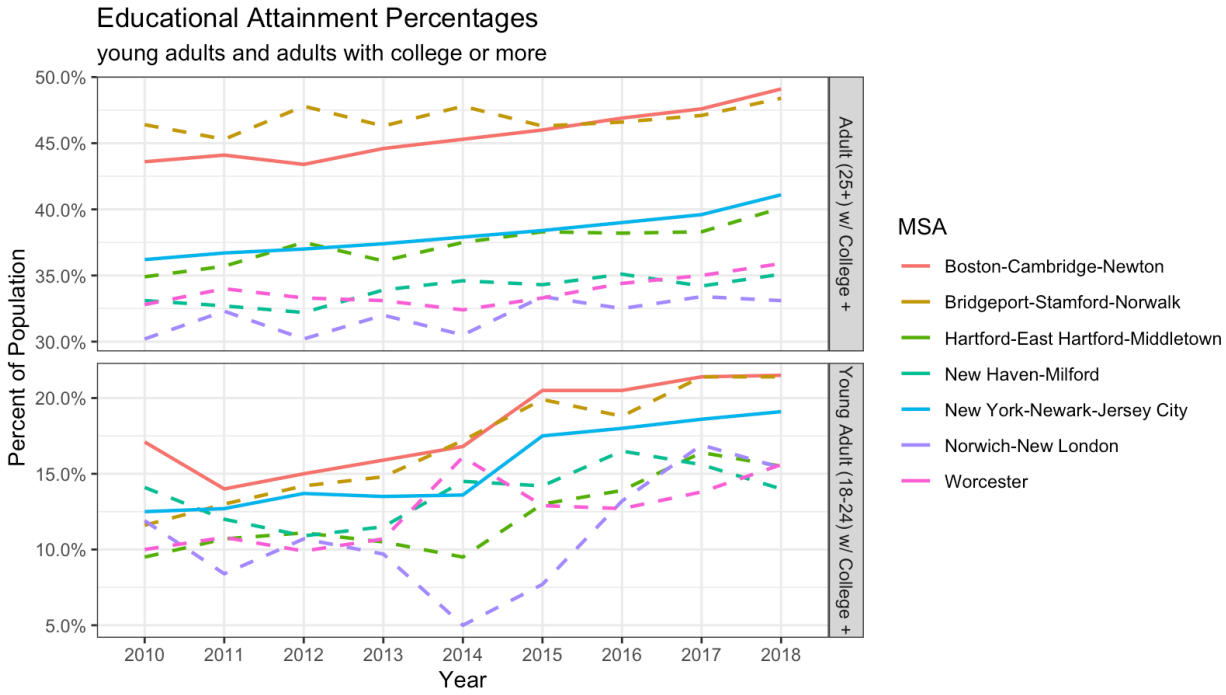

Figure 9

	Age/Educational Attainment:	
	Adult (25+) w/ College+	Young Adult (18-24) w/ College+
Boston	45.6%	18.1%
New York	38.1%	15.5%
Bridgeport-Stamford-Norwalk	46.9%	16.9%
Hartford-East Hartford-Middletown	37.4%	12.2%
New Haven-Milford	33.9%	13.7%
Norwich-New London	32.0%	11.0%
Worcester	33.8%	12.5%

Table 4 Percentage of population for educational attainment/age categories (averaged between 2010 and 2018).

Based on Figure 9, it's clear that the Boston metro area and Bridgeport-Stamford-Norwalk MSA have a far higher percentage of adults with a college degree or higher amongst their populations. This is interesting in that we had assumed that the New York metro area populations would have higher percentages of adults with college degrees or higher. When observing the lower graph in Figure 9, Boston, New York, and Bridgeport-Stamford-Norwalk all follow a fairly steady upward trend in percentage of young adults with a college degree or higher, from 2010 to 2018. The remaining 5 MSA's are more sporadic throughout this time frame. This can most likely be attributed to the higher propensity of this age group to migrate from place to place. As young adults pursue higher education institutions and first/entry-level

jobs within these MSA's, their propensity to 'move around' is fairly high. Additionally, the dip experienced in the Norwich-New London MSA around 2014/2015 reflects an identical dip along the same time frame for our net migration metrics. We hypothesize that Norwich-New London's 'negative peak' in net migration along this time frame can largely be attributed to a population of educated young adults.

Table 4 contains average percentages along this 2010 to 2018 time frame for our 7 MSA's of interest. Again, Boston's metro area shares similar values with Connecticut's Bridgeport-Stamford-Norwalk MSA. This is a similarity that should be noted as these two MSA's also appeared within the same cluster during our K-means cluster analysis.

To get a better sense of how different percentages of these age groups affects the rGDP per capita of a given MSA, we ran a basic linear regression. Our initial model used rGDP per capita as its response and the percentage of young adults with a college education or higher as our single predictor. As reflected in Table 5, this model did not predict rGDP per capita very well, with a R-squared value of .18 (our single predictor explaining 18% of the variance in our response). Although, our predictor did correlate with our response at a 99% confidence level. The coefficient for this percentage shows that for every 1% increase in an MSA's population of young adults with a college education or higher, the rGDP per capita increases by $1,751.2 on average. Our next model controlled for variables such as percentage of adults with a college degree of higher as well as three of our age groups: 19 and below, 20 to 34, and 35 to 54. The results were far stronger, with our 5 predictors explaining 81.9% of the variance in rGDP per capita (R-squared: 0.819). Although, the coefficient for our young adult educational attainment predictor became negative, meaning a higher percentage in the population of young adults with a college degree or higher results in a decreased rGDP per capita. Additionally, the coefficient for adults with a college degree or higher shows that for every 1% increase in the population of these adults, the rGDP per capita is expected to increase by $2,194.4. The full results are displayed in Table 5, although they diverge from our hypothesis.

Model 1:		Estimate	Std. Error	T value	P value
	(Intercept)	43,603.6	7,057.6	6.178	<.001 ***
	y_adult_col_high	1,751.2	514.4	3.405	.0014 **
		F-Statistic:	11.59	Adj. R-squared	0.18
Model 2:					
	(Intercept)	322,621.7	65,778.6	4.905	<.001 ***
	y_adult_col_high	-1,055.6	373.2	-2.829	.007 **
	adult_col_high	2,194.4	214.6	10.226	<.001 ***
	age_one	-10,331.4	2,025.2	-5.101	<.001 ***
	age_two	-2,189.6	907.5	-2.413	.02
	age_three	-2,736.8	1,618.3	-1.691	.098
		F-Statistic:	44.3	Adj. R-squared	0.82

Table 5 Age/Educational Attainment Regression Models

Ultimately, while the results of our models do not align with our hypothesis, they do show an interesting relationship between the percentage of adults, aged 25 and older who have attained a Bachelor's degree or higher, and rGDP per capita is strong and positive. Additionally, this attribute is an important variable in our random forest models, as depicted in Figure 3.

Furthermore, we can confidently conclude that this attribute is an important factor when considering the economic success/development of a given MSA.

NAICS Establishments

In analyzing the relationship between NAICS establishment percentages and distributions across our 7 MSA's, we decided to choose those that aligned most closely with our initial hypotheses. We selected 4 from our top 20 most important variables, as outlined in Figure 3, which included: Accommodation and Food Services, Real Estate and Rental and Leasing, Support Activities for Transportation, and Securities, commodity contracts, and other financial activities. We believe these 4 NAICS industries do a good job of encompassing a few attributes within our MSA's: amenities, travel/tourism, commuter economies, and the financial sector. Figure 10 outlines the overall correlations observed between rGDP per capita and percentage of overall establishments attributed to each NAICS industry of interest, by MSA. Additionally, Table 6 reflects the pearson correlation coefficients for these comparisons.

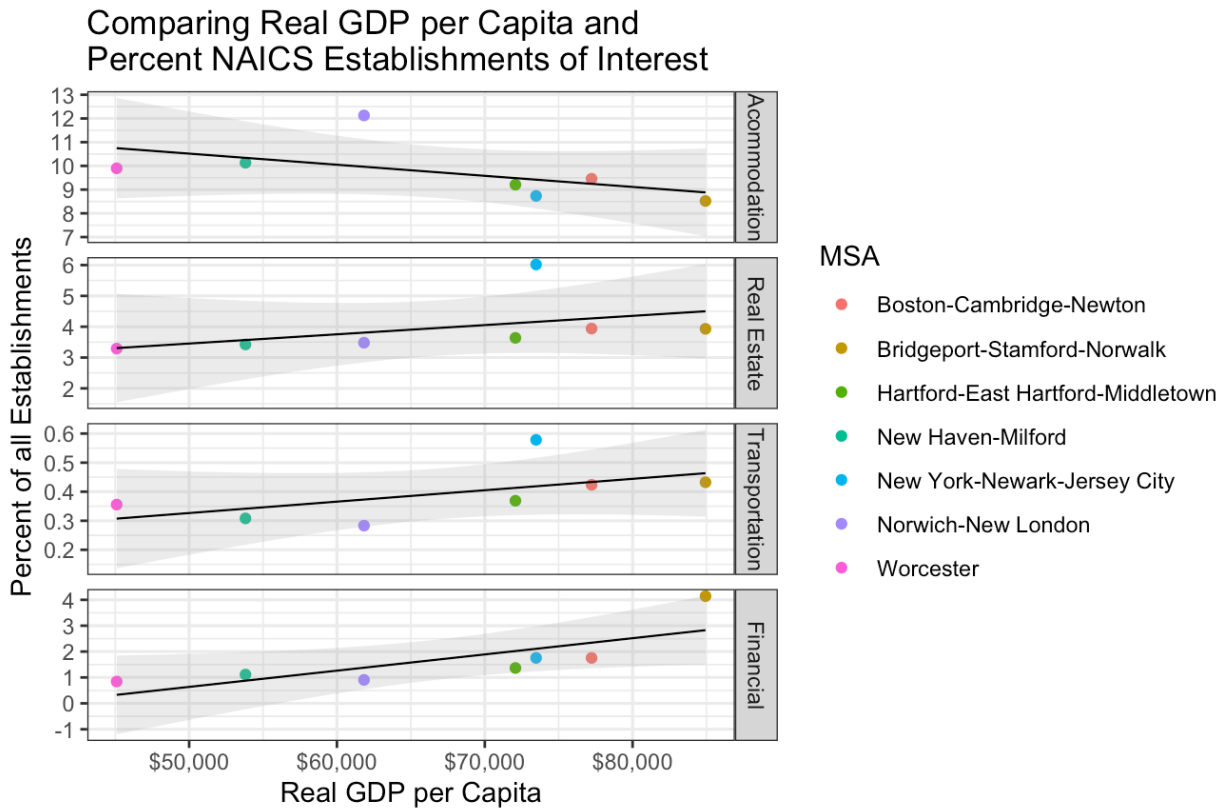

Figure 10 Percent of establishments attributed to each NAICS industry of interest vs. average real GDP per capita (2010 to 2016).

rGDP per capita vs.	cor	95% Confidence:		p-value
Accommodation	-0.53	-0.71	-0.29	<.001
Real Estate	0.44	0.19	0.64	.0014
Transportation	0.54	0.30	0.71	<.001

Financial	0.76	0.61	0.86	<.001

Table 6 Pearson correlation test results

As reflected in Table 6, we have consistently significant p-values from our pearson correlation tests. We used data from all MSA's across all observed years (2010 to 2016) to increase our degrees of freedom. Our correlation coefficient for establishments in the Accommodation and Food Services NAICS sector was moderately strong and negative (-0.53). This was surprising as we had hypothesized that NAICS sectors related to amenities would be positively correlated with rGDP per capita. Additionally, the correlation coefficient for establishments in the Securities, commodity contracts, and other financial activities NAICS sector was strong and positive (0.76). This was our strongest correlation, of those we tested, therefore this leads us to presume that a higher percentage of establishments in this NAICS sector will lead to a higher rGDP per capita for a given MSA. Below, we have broken down the changes in these establishment attributes across our time frame, by MSA.

Accommodation and Food Services:

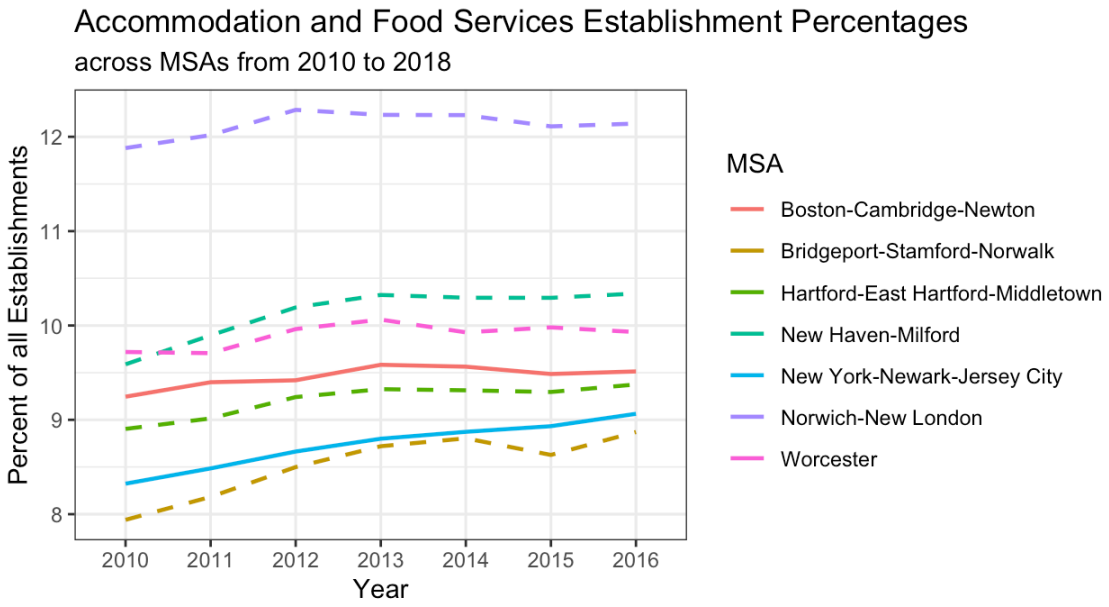

Figure 11

As displayed in Figure 11, Norwich-New London has the highest percentage of establishments in the Accommodation and Food Services NAICS industry from 2010 to 2016. Considering the Norwich-New London MSA has a median rGDP per capita (within our sample; see Figure 1), it is interesting to see how much higher its percentage is compared to the remaining MSAs. This could also explain why this sector's aforementioned pearson correlation coefficient was negative, as these values most likely influenced this relationship. Additionally, it seems like Connecticut generally has a higher percentage of establishments in this NAICS sector, on average, compared to the New York and Boston metro areas.

Real Estate and Rental and Leasing:

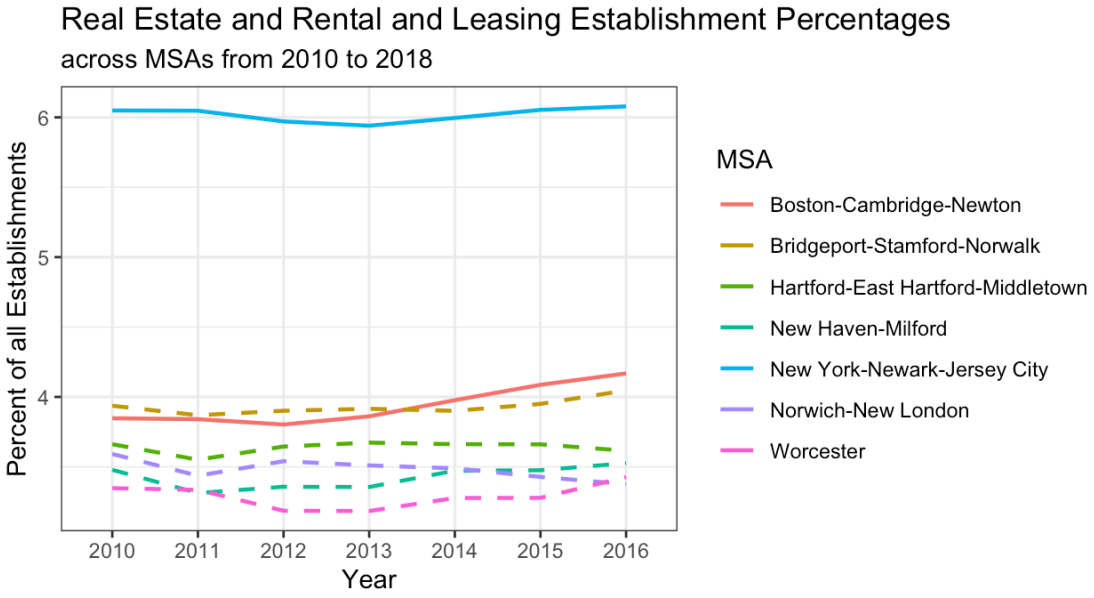

Figure 12

Figure 12 depicts the year-to-year percentages for establishments in the Real Estate and Rental and Leasing NAICS sector. The New York metro area has a much higher percentage of establishments in this NAICS sector compared to the remaining 6 MSAs. As we mentioned previously, we are treating this NAICS sector as a rough proxy for the characterization of general tourism/travel, as it accounts for rentals and leases. This aligns with more subjective characteristics of the New York metro area, as it's notorious for its high housing costs – as it would make more sense to temporarily live in this MSA, rather than live permanently here.

Support Activities for Transportation:

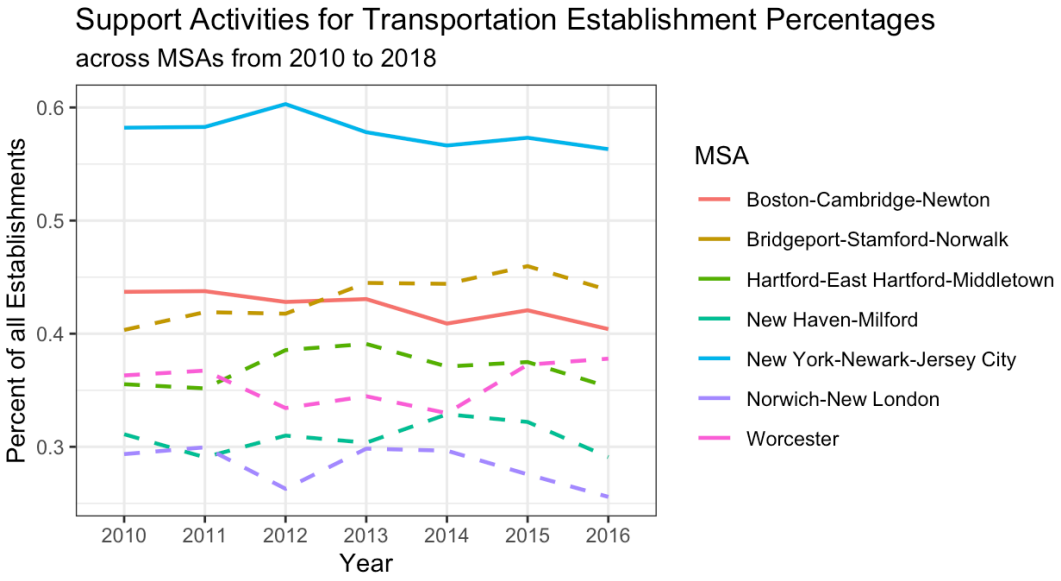

Figure 13

Figure 13 depicts the year-to-year percentages for establishments in the Support Activities for Transportation NAICS sector. Again, the New York metro area has a consistently higher percentage of establishments in this NAICS sector. Additionally, Bridgeport-Stamford-Norwalk and Boston-Cambridge-Newton have marginally higher percentages than the other MSA's. This would make sense considering these MSA's are known to have large commuter workforces.

Securities, commodity contracts, and other financial activities:

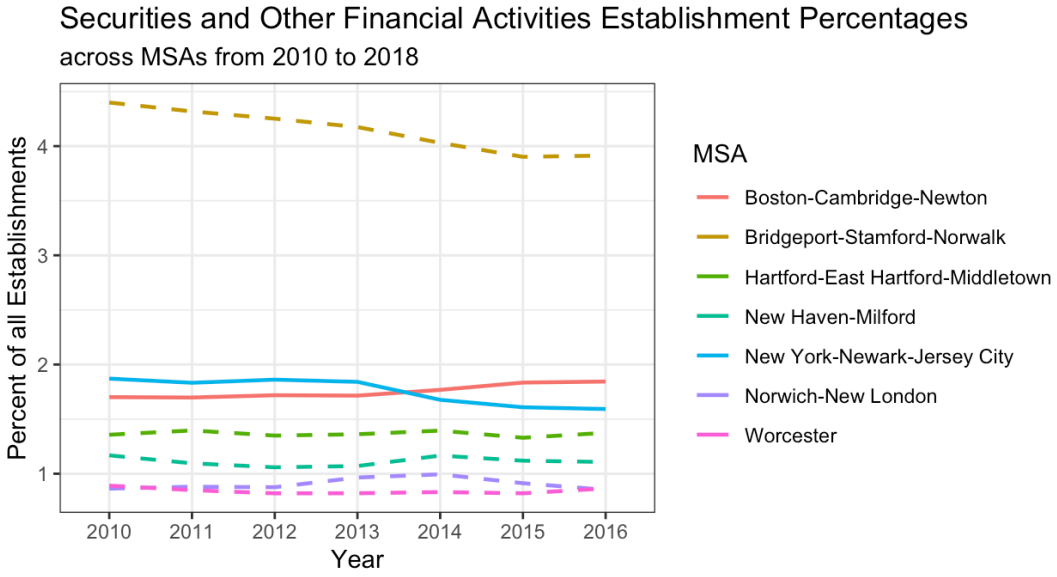

Figure 14

Figure 14 depicts the year-to-year percentages for establishments in the Securities, commodity contracts, and other financial services NAICS sector. Bridgeport-Stamford-Norwalk has by far the highest percentage of establishments in this sector, with the New York and Boston metro areas marginally higher than the remaining MSAs. While Bridgeport-Stamford-Norwalk's proximity to the New York metro area, makes it a hotspot for commuters, Figure 14 shows how this MSA also has many establishments in the financial sector, of its own. It is important to note, although, that the percentage of establishments in this sector, for Bridgeport-Stamford-Norwalk, has steadily declined from 2010 to 2016.

Discussion/Implications

We found that the Bridgeport-Stamford-Norwalk MSA holds the most similarities to the New York and Boston metro areas. It's high GDP, percentage of adults and young adults with a college education or higher, and high percentage of establishments in the financial NAICS sector all contribute to this. Additionally, we find that an MSA's percentage of adults, aged 25 and higher, with a college degree or higher, is a good indicator of economic success in a given MSA, based on real GDP per capita. And lastly, Hartford-East Hartford-Middletown and Worcester are two Connecticut MSA's that we see as particularly interesting. Firstly, Hartford-East Hartford-Middletown has a high real GDP per capita and shares similar characteristics to Boston's metro area, as it shared a cluster in our K-means cluster analysis. Secondly, Worcester has held a steady increase in positive net population migration. Although its percentage of adults and young adults with a college degree or higher is relatively low compared to the remaining MSAs, we

believe Worcester may be a favorable destination for those seeking retirement, with a relatively higher percentage of establishments in the Accommodation and Food NAICS sector.

While our study is generally broad and requires additional, more specific analyses, it has laid some groundwork in terms of offering a working methodology for exploratory data analysis techniques and data acquisition techniques. A large portion of this project was involved with the collection of data from multiple sources, learning new API tools, and aggregating this data into clean data sets. We hope that this project will be extended in the future to include additional years and analytic techniques to get a better sense of where Connecticut is headed economically, especially considering the current and future effects of the COVID-19 pandemic.