**Analysis of the INC.com Top 5000 Companies and Comparison with Census Data**

Daniel Amrine

Colorado State University Global

MIS581-1: Capstone Business Intelligence and Data Analytics

Dr. Kimberly Ford

11/1/2020

**Analysis of the INC.com 2019 top 5000 Companies and Comparison with Census Data**

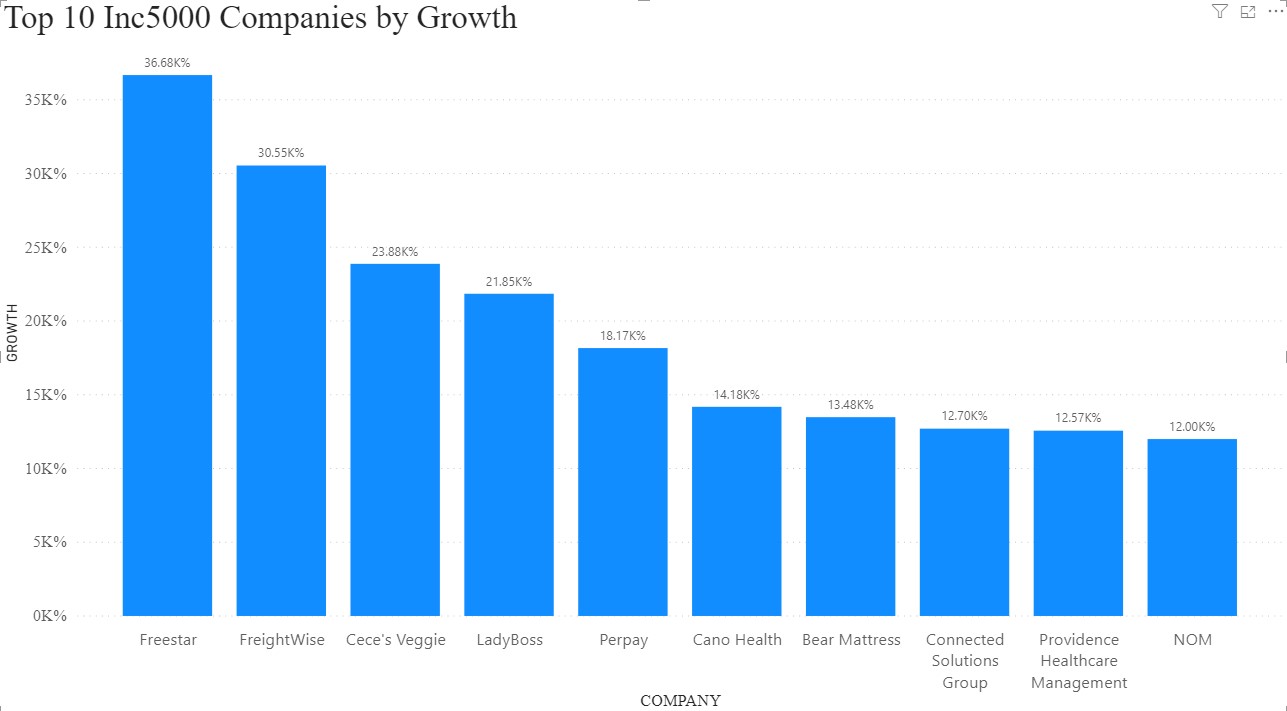
**Abstract**

Inc.com is a magazine devoted to private business and entrepreneurs. They publish a list of the 5000 fastest growing private business every year. The growth is calculated using the revenue from the previous year. These companies also must meet a strict criterion to be included in the list so this ensures a unique data set. The following analysis will first analyze this data set to identify key patterns in terms of geography and then compare the results with census data including earnings, health, education, population, and urban density. The comparison will determine if there are any strong or weak correlations and if it is possible to identify a combination of attributes that will either encourage or discourage the growth of these companies. The results show there are clear and strong relationships between the population of the state, the education, the poverty, and the growth of these companies. There are also negative relationships between the companies with the lowest growth and states with lowest population. However, this analysis also highlighted discrepancies in some of the population data and with the distribution of much of the data. The primary goal was to investigate the relationships between the growth of these specific companies using advanced data analysis. There is much more to be learned from this analysis and workflow.

The Inc.com website and magazine annually identifies the top 5000 growing businesses in the united states and publishes a list every year in a special issue and posts the list online with access to limited set of columns(Inc.5000, 2020). This dataset is unique because all the companies are private and Inc.com has verified all the revenue data and growth figures with the CEO of each company. For this analysis, the 2019 list will be used because in 2020 Inc.com took revenue off the public website and it now costs $4,995 to get the entire dataset. The companies ranked in the top 10 of these lists all have growth above 12,000% (Figure 1.)

Figure 1

*Top 10 Inc5000 companies in 2019.*



The goal of this analysis is to investigate what drove the growth of these companies. There are many variables and attributes that go into business growth, but for this analysis it will focus on geography, the education of the people living near these companies, the poverty, and earnings. It is essential to understand how these businesses have grown so fast and this analysis is the initial step in this direction.

**Objectives and Overview**

The objectives of this analysis can be broken down into three parts. The first focuses on the 2019 list of companies. The objective is to find any unique or unknown relationships or patterns in the dataset. The next part is comparing a summarized dataset to geography in terms of what states the companies are in, the population of the states, and the amount of urbanization in each state. The third part is to then compare this dataset to the summarized poverty, earnings, and education of each state. The final objective is to show how these comparisons can be used to predict which business type will grow the fastest based on this list of 5000 companies across the country.

To begin this analysis a summary will be made of initial data set in terms of counts, medians, maximums, and minimums. The result will be a list of the states with a count of total companies in the list, the industry with the highest growth in each state, median growth, maximum growth, and minimum growth. The key to this analysis is grouping the dataset by state. In the next stage the target geographic data will be merged with this new dataset and then a correlation can be made between the three company growth attributes and the target attributes. Because this analysis will only focus on growth the revenue will be ignored for now.

The next stage will require a summary of the census data by state and then a merge with the census data and the original summarized dataset from the list of 5000 companies. Again, the analysis will use correlation to assess if there are any strong correlations between growth and the us census data. In the worst-case scenario, all the correlations are non-existent or so weak that they cannot be used for any future analysis. If this does happen then a new approach or set of variables will have to be found. In the case of finding strong correlations then the analysis will move into the final stage and use the relationships above to define a set of cases that either encourage or discourage business growth.

**Research Hypothesis**

There are multiple hypotheses required for this analysis. The following lists them in order of analysis.

1. Is each variable normally distributed? In this case the null hypothesis (Ho) is the data does not have a normal distribution and the alternate (Ha) is the data has a normal distribution.
2. There is a strong correlation between the state population, growth, and urban density and business growth. Ho: there is no correlation or a very weak correlation, Ha: The correlation is strong
3. There is a clear relationship between the population’s education, earnings, and poverty with the growth of these companies. Ho: there is no correlation or a very weak correlation, Ha: The correlation is strong.
4. If these relationships exist, it is then possible to identify the combination of these attributes that would either enhance the growth or discourage the growth of these companies. Ho: It is not possible because there are no correlations, Ha: It is possible to define a positive or negative impact on business growth based on the strong correlations.

**Literature Review**

It is difficult to find specific references that deals with this analysis. For this reason, the analysis will only focus on business growth and the following references helped build a framework on constitutes business growth in general. (Bonsu & Kuofie, 2019) mention how small businesses are the key to the GDP of the United States. For this reason, understanding business growth would help small businesses be more successful. This analysis can provide the foundation to build on by comparing these relationships with these larger companies identified by INC.com. The second hypothesis deals with population and geography. (Rosenblum et al., 2014) uses geography to correlate the location of Puerto Rican neighborhoods with the introduction of a cheaper type of heroin. A similar analysis can be quickly done by showing a map of where the 5000 companies are in relationship to city density.

This idea is what led to this analysis in the first place. To further cement how important small businesses are (Farlie et al., 2019) looks at the universe of start-ups in the US and combines it with a panel data set. It also confirms the importance of small business by highlighting how there are nearly 5 million new businesses created annual which leads to over 3 million jobs created annually. In this analysis the criterion Inc.com uses to select the companies are as follows (Inc 5000, 2020):

1. Be Privately-owned, based in the United States, and independent
2. Have started earning revenue by March 31, 2016
3. Had revenue no less than $100,000 in 2016
4. Had revenue no less than $2,000,000 in 2019
5. Revenue in 2019 exceeds revenue in 2016

There is no mention of number of employees or a specific limit to size, so this data set includes companies of all sizes. However, learning from these companies will be beneficial for small business.

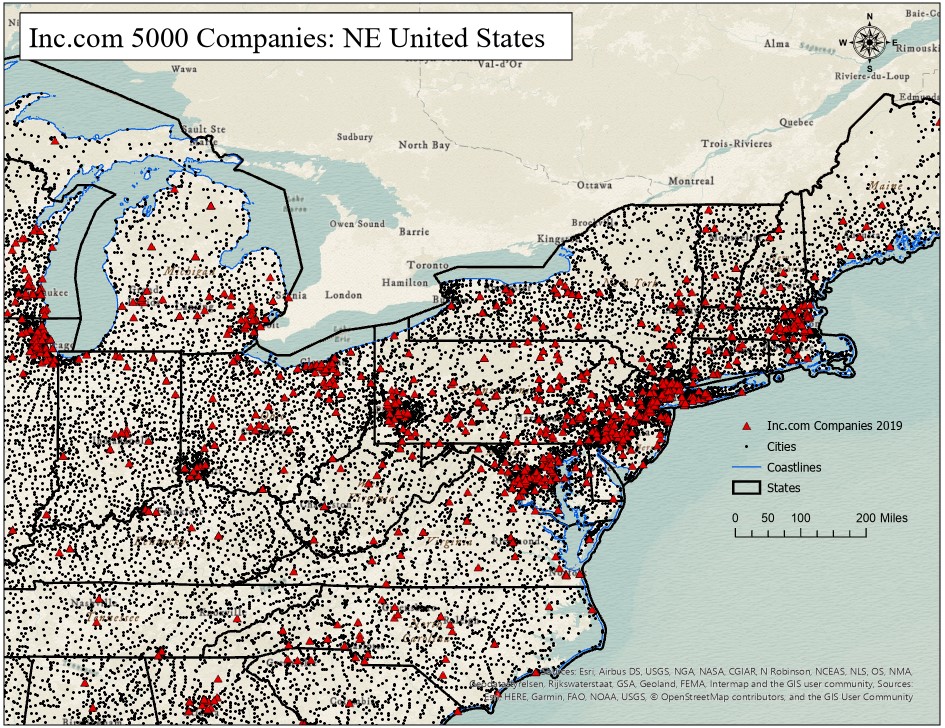
Finally, (Rupasingha & Wang, 2017) compare the access of capital top the growth of small business. This is the one factor that is not considered in this analysis because inc.com either does not track this or does not provide it with their public release of data. The final aspect of this analysis deals with trying to predict what groups of variables will either encourage or discourage business growth. (Wach 2020) provides a detailed template of models that could be used to predict business growth.

**Methodology and Research Design**

The initial method involves manually scraping the data from Inc.com. In 2019 they provided the entire table on a public website and by viewing 500 records at a time they could be copied and pasted into an excel sheet. When this process is complete the next step was to clean the location data so it could be linked to state and county. Several of the records has misspelled city names or used the name of a geographic region rather than a city. The following Figure 2 shows the location of all 5000 companies in 2019. There is a clear clustering around the point layer that represents the cities of the United States. Most of the companies are centered around New York, Baltimore, Washington DC, and Los Angeles. This analysis does not dig deeper into these relationships using maps, but it uses a dataset from 538 that creates an urban index for each state (fivethirtyeight, n.d.).

Figure 2

*A sample of a map of the 5000 companies with cities*



*Note: made with ArcGIS PRO by ESRI*

There are no companies in Alaska with a few in Hawaii. This analysis will focus on the companies in the lower 48 states.

The first hypothesis focuses on the normal distribution of each target attribute. To test this the data will be imported into a Python Notebook that will provide the code to load the data into data frames and then run the Shapiro-Wilk Test to see if each variable has a normal or gaussian distribution. The reason python was selected is because it can be easily made into a notebook which can be shared and then to test other variables all someone needs to do is alter a few lines of code and run the notebook.

The second and third hypothesis will be tested using SAS. Based on the findings of the first hypothesis a correlation will be selected and ran because in SAS this can be done quickly using a tool called data exploration. It quickly compares variables and can quickly generated the appropriate figures to help decide which variables have the strongest correlations. In the case no correlations are found or if Ho is not rejected than the analysis will end, and a different set of variables will be identified. The final hypothesis will be tested using Microsoft PowerBI. This last hypothesis will require the use of box plots as well as live reports that let you quickly filter and slice the variables into a visualization that will answer these questions.

The main limitation is missing data. Unfortunately, there are many aspects of the companies we are missing. This analysis was based on a limited set of data from public sources. To make this analysis complete there is a need to obtain proprietary data about the companies by wither purchasing the data from Inc.com or gathering from another company that tracks data like total employees, total costs, how much capital did they start with, and other key metrics. This also opens the door on what data sources are ethical and if the companies on this list want their data used for this kind of research.

The final limitation is the in one case, for Puerto Rico, there is a problem with the population data and the poverty data. There is a criterion used by the census to define poverty and this data set is the count of those groups that meet these criteria that live above or below the 2019 threshold of $31,275. According to this data set the number of households living below this threshold in Puerto Rico was 1,577,075 but the population of Puerto Rico in 2018 was 1,344,083. This difference led to an error in the results that had to be mitigated. Fortunately, this was the only case.

**Findings**

The results of this analysis rejected the null on some of the hypotheses but not all of them. From the distribution of the attributes (Figure 3) and the following Shapiro-Wilks Tests (Figure 4) this data does not have a gaussian distribution. Ideally, this data should have a normal distribution before attempting correlation but because of the nature of this data finding attributes that have this distribution is very difficult. The null of the first hypothesis was not rejected.

The decision was made to continue with the analysis rather than hunting for data with a normal distribution or applying a transformation that forces the data into a normal distribution. The main purpose of this analysis is to answer all the proposed hypotheses.

Figure 3

*Distribution of columns in INC5000 dataframe.*

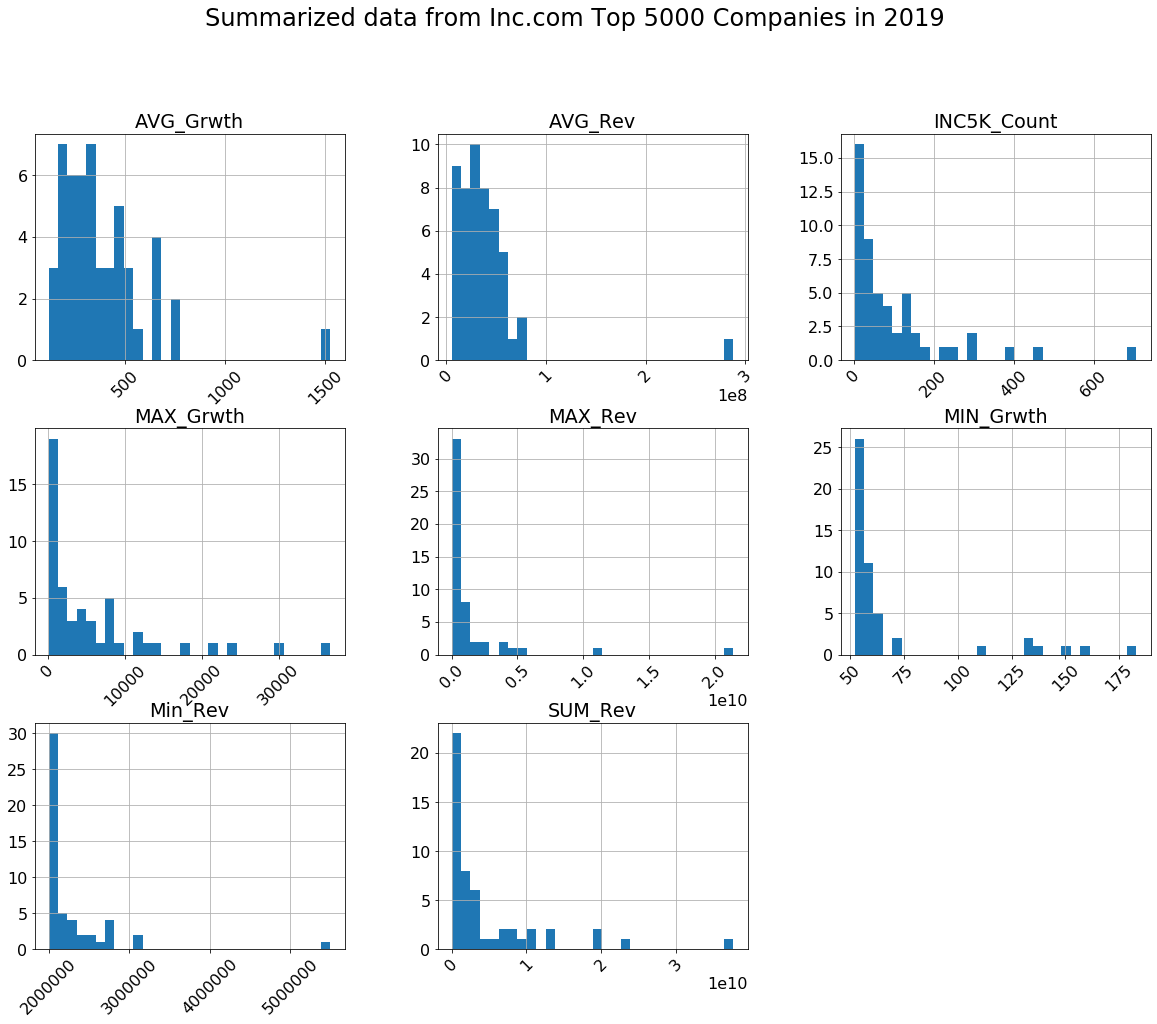


Figure 4

*The Shapiro-Wilks Test results in python for an attribute.*



A second attempt was made to find normal distribution by converting the poverty and education data from counts to a rate of value per 100,000. However, this still produced the same results.

To test the second and third hypothesis the dataset was combined and imported into SAS Studio. Figures 5 and 6 show the results of the data exploration process. This process compares the correlation of all the selected variables. In Figure 5 there is a strong correlation with count of companies and total state population in 2018.

Figure 5

*Data exploration of hypothesis question number two*

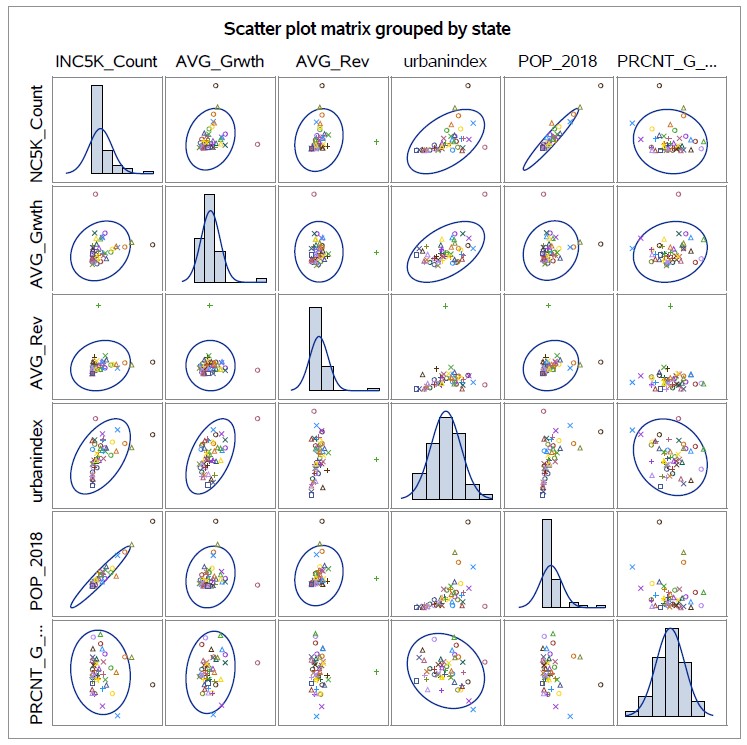


Figure 6

*Data exploration for hypothesis question number three*

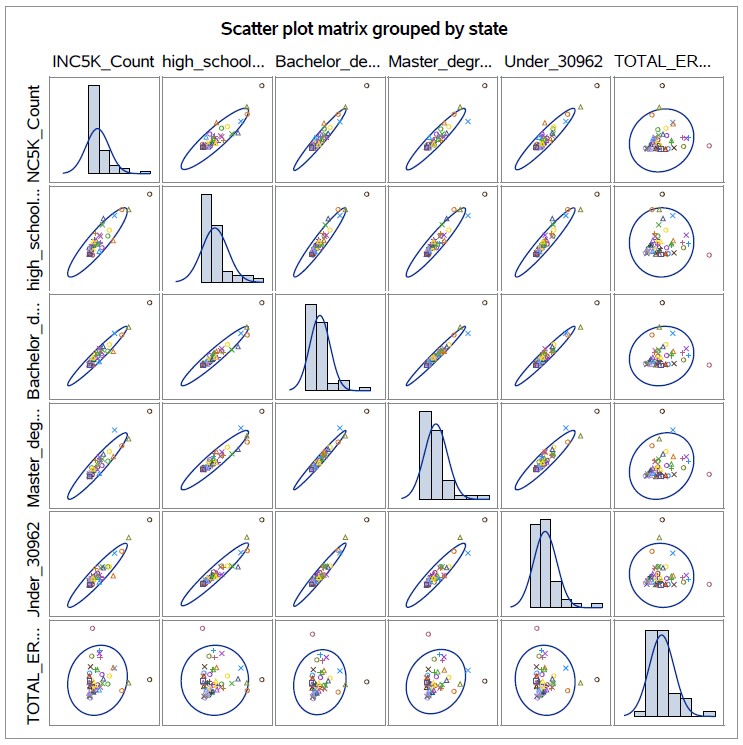
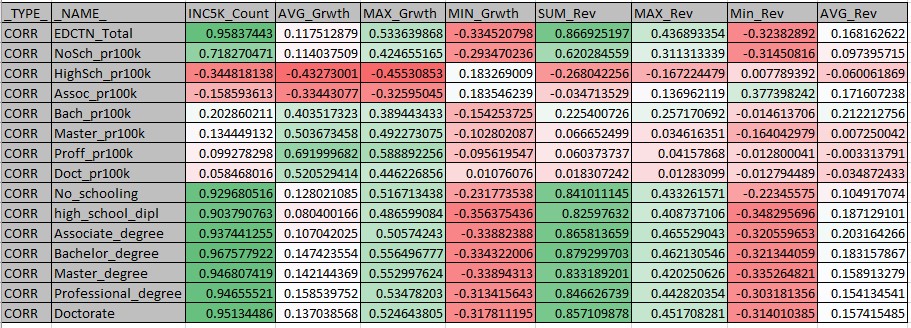


Figure 6 brings out some shocking results. It seems strange that the count of companies per state would have such strong positive correlations with all levels of education obtained by those over 25 and households living under an annual salary of $30,962. The correlation tool in SAS Studio was used to dig deeper. Figure 8 shows a more detailed analysis of the correlations in excel using conditional formatting. The Dark Green is the greatest positive correlation, and the dark red is the most negative correlation.

Figure 8

*Detailed correlation table imported into Excel from SAS*

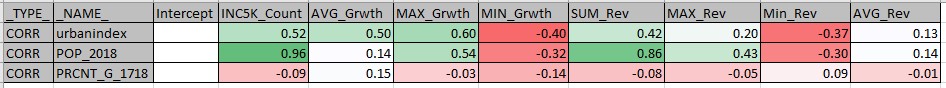


Based in this portion of the analysis both hypothesis two and three rejected the null and pushed the alternate as there are indeed some strong correlations between the summarized growth, revenue, and counts of the companies. However, more investigation is needed into why the count of education obtained seems to have such strong correlation with the count of companies per state and not with the growth.

To test the fourth and final hypothesis the same dataset was loaded into Microsoft PowerBI. Based on the correlation analysis the population of the state has a strong positive correlation with the total number of companies by state. However, the population has no correlation with average growth, but does have a weak negative correlation with minimum growth. Figure 9 shows this correlation table. To confirm this the population had to be grouped into categories using a custom equation in PowerBI. The result were three categories representing population size. *Class\_500K* are states with less than 500,000 people based on population in 2018. *Class\_5mil* are states with a population between five hundred thousand and 5 million and *Class\_GT5mil* are states with more than 5 million. A *boxplot* is then used to compare the counts of these classes with company count per state and summarized company growth per state (Figures 10 and 11).

Figure 9

*Population correlation table*



Based on this Figures 10 and 11 show the relationship with the new population class, the count of companies by state and the minimum growth by state, respectively. This analysis has shown that states with a population more than 5 million contain most of the fastest growing companies, but the four states with population lower than 500,000 have the best performing companies out the companies with the slowest growth. There are more cases to build and this analysis has mode this possible.

The Green points represent outliers and the 2 state with highest counts that have a population greater than 5 million are California and Texas. California has more than 700 of the companies on the list and they are all clustered around Los Angeles. Colorado and Utah have the most companies on the list for states with population between five hundred thousand and 5 million. In terms of minimum growth, the outliers for the state in *Class\_5mil* were West Virginia, Rhode Island, and New Mexico. The companies in the list in West Virginia averaged a growth of 149%. The outliers reveal how further investigation is necessary as well as a possible change in scale to get more granular in terms of counties or cities.

Figure 10

*Boxplot of total count by state vs population class*

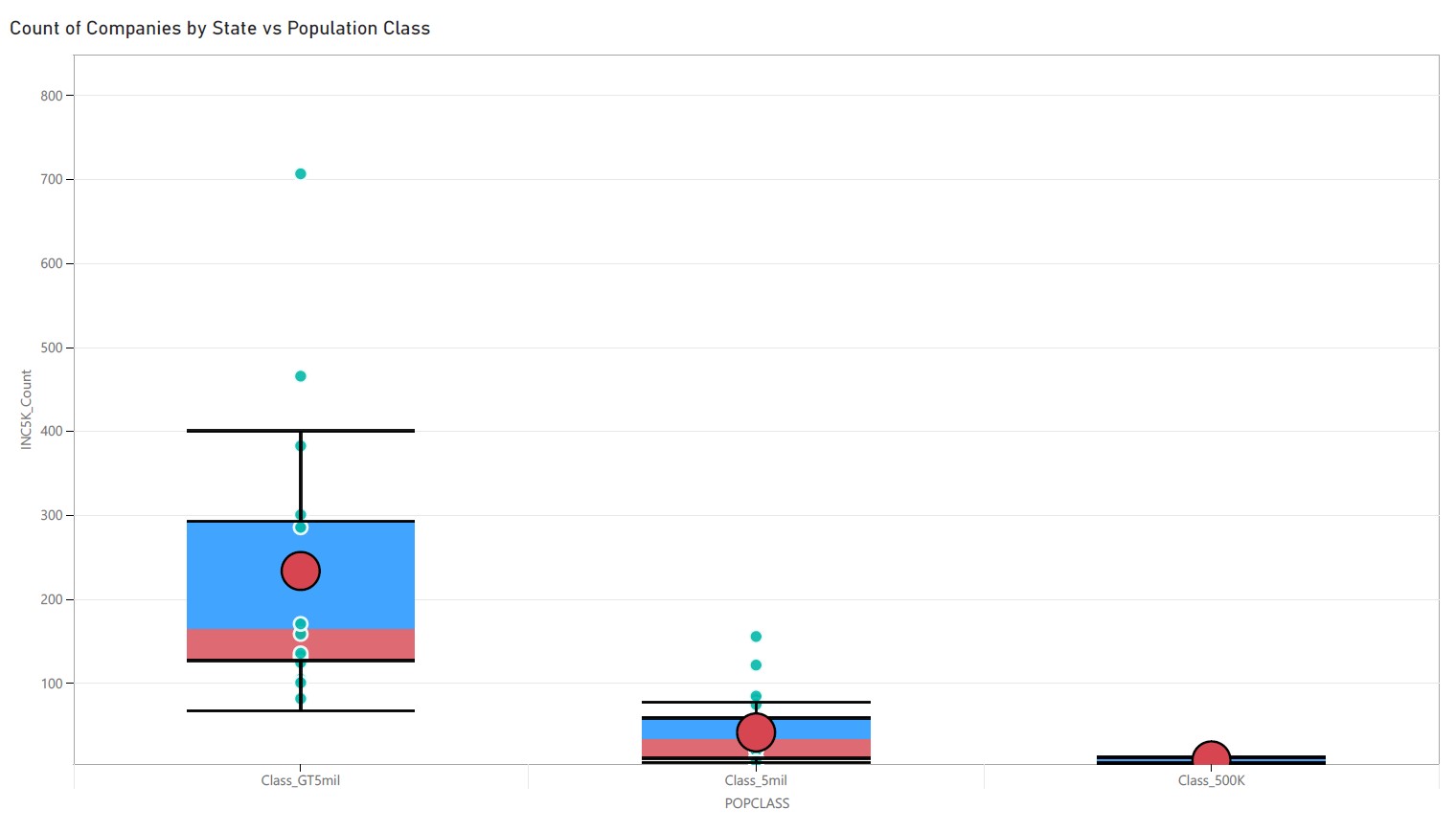
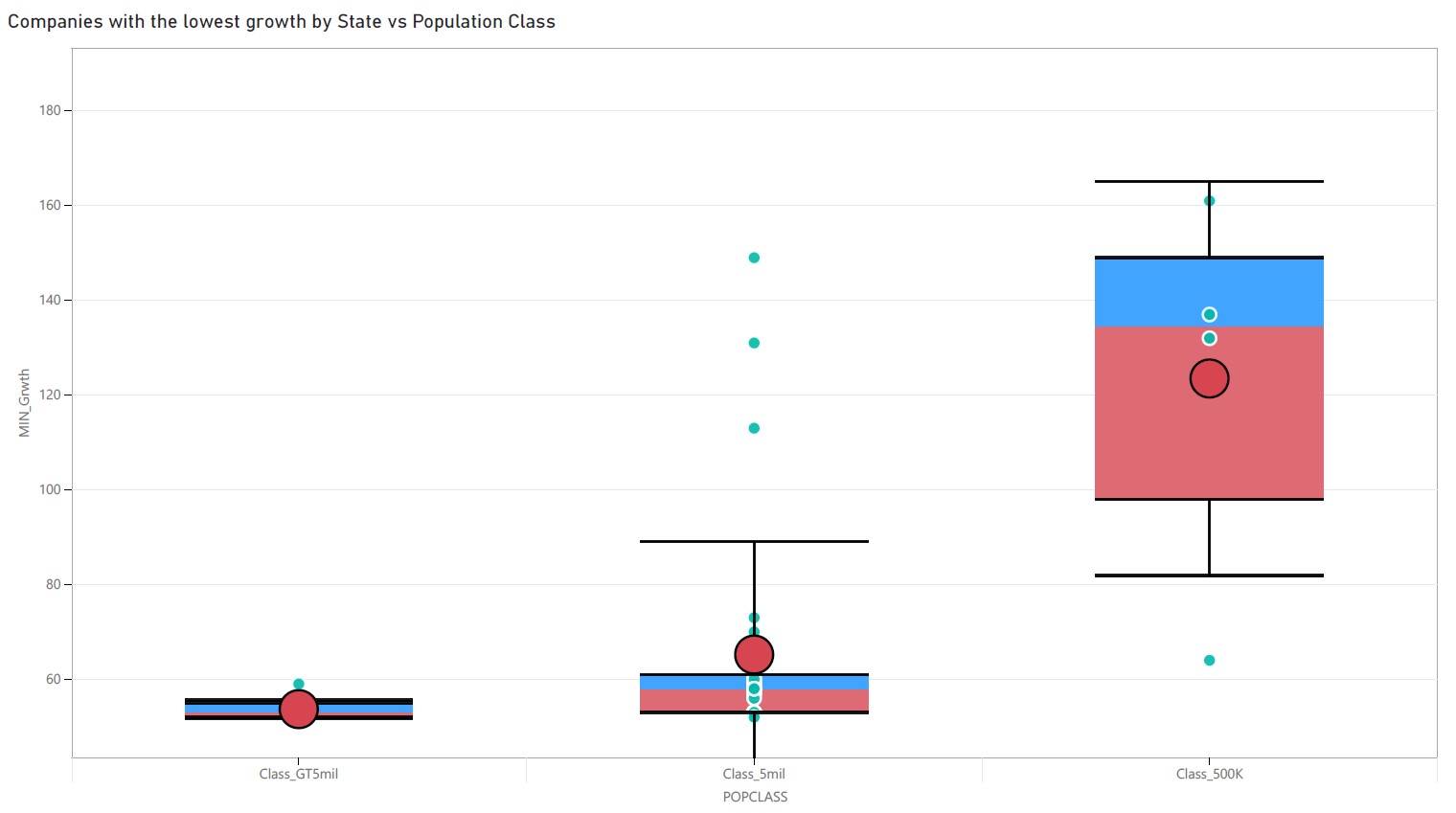


Figure 11

*Boxplot of minimum growth by state vs population class*



**Conclusion and Recommendations**

This analysis set out to test four specific hypotheses. The first dealt with the distribution and the null was not rejected. The second and third pushed the alternate hypothesis in terms of strong or weak correlations and the fourth pushed the alternate hypothesis because these correlations can be used to build cases about the company growth. However, there are many more hypotheses that can now be answered because of the process has been defined and the analysis can be applied to any other set of variables. This analysis has shown how there is a possible workflow that can be applied to any data set. If the analyst wants to continue with the process there are more correlations that need to be investigated and better datasets to use in the future.

Going forward, a different test about the distribution should be utilized and tested as the results of the education correlations may be a product of using a Pearson correlation on data that does not have normal distribution. The relationship between education per 100,000 and the growth should be investigated more as there is an expectation of higher education should warrant faster growth, but the findings of this analysis do not support that assumption. The goal of this analysis was not only to test the hypotheses but to also build a workflow than can now be applied to other variables and tests.

**References**

Bonsu, S., & Kuofie, M. (2019). Small business survival. *Journal of Marketing & Management*, *10*(1), 51–63.

Brownlee, J. (2018, August 14). 17 Statistical hypothesis tests in Python (cheat sheet). Machine learning mastery. https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/

US Census Bureau. (2020, January 22). Poverty thresholds: the united states census bureau. https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-thresholds.html

Fairlie, R. W., Miranda, J., & Zolas, N. (2019). Measuring job creation, growth, and survival among the universe of start-ups in the United States using a combined start-up panel data set. *ILR Review*, 72(5), 1262–1277. <https://doi.org/10.1177/0019793919862764>

Fivethirtyeight/data. (n.d.). GitHub. https://github.com/fivethirtyeight/data/tree/master/urbanization-index

Inc. 5000 2020: An exclusive guide to America’s most inspiring entrepreneurs. (n.d.). Inc.com. <https://www.inc.com/inc5000>

Rosenblum, D., Castrillo, F. M., Bourgois, P., Mars, S., Karandinos, G., Unick, G. J., & Ciccarone, D. (2014). Urban segregation and the US heroin market: A quantitative model of anthropological hypotheses from an inner-city drug market. *International Journal of Drug Policy*, 25(3), 543-555. <https://doi.org/10.1016/j.drugpo.2013.12.008>

Rupasingha, A., & Wang, K. (2017). Access to capital and small business growth: Evidence from CRA loans data. *Annals of Regional Science*, 59(1), 15–41. https://doi.org/10.1007/s00168-017-0814-9

Wach, K. (2020). A typology of small business growth modelling: A critical literature review. *Entrepreneurial Business & Economics Review*, *8*(1), 159–184. <https://doi.org/10.15678/EBER.2020.080109>