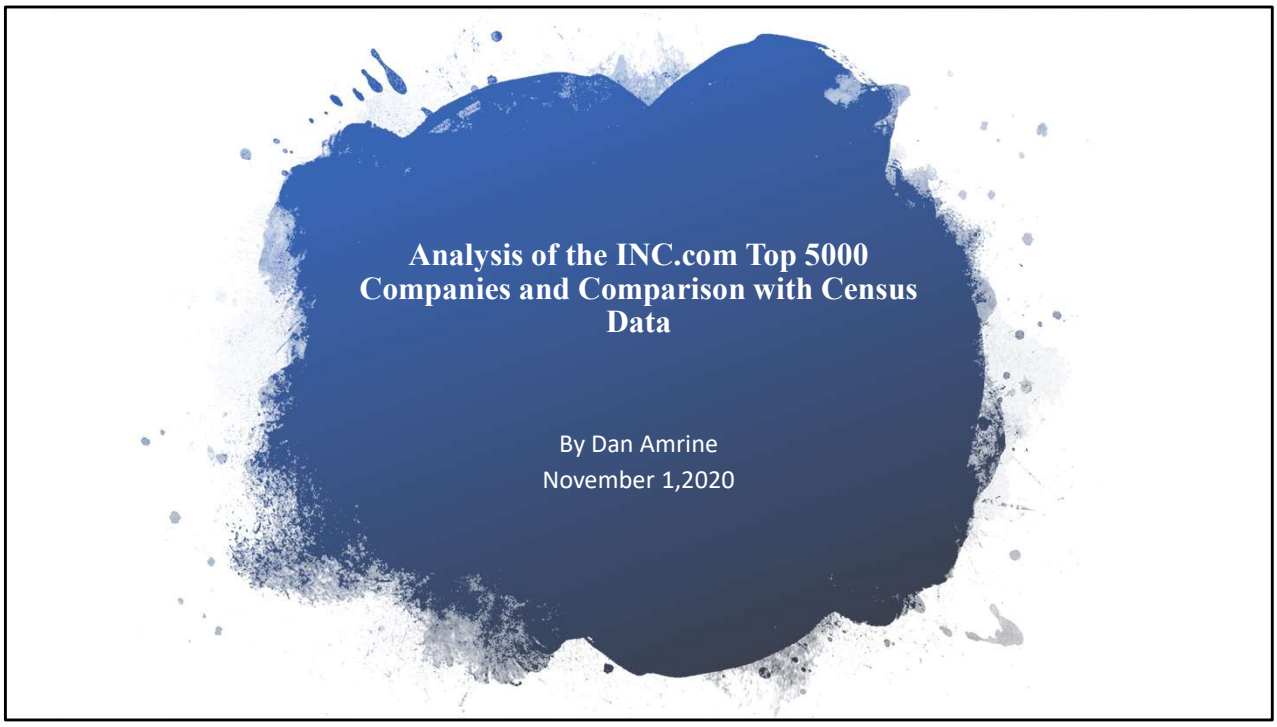
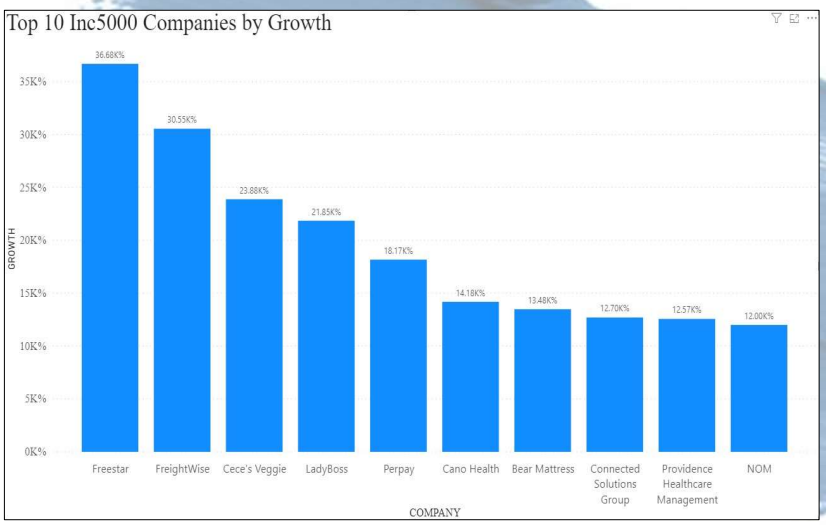

Analysis of the INC.com Top 5000 Companies and Comparison with Census Data
By Dan Amrine
November 1,2020

Agenda:

- Introduction
- Objectives
- Research Hypothesis
- Methodology
- Findings
- Conclusion and Recommendations

Introduction:



INC.com list of top 5000 growing companies from 2018-2019

What drives this growth?

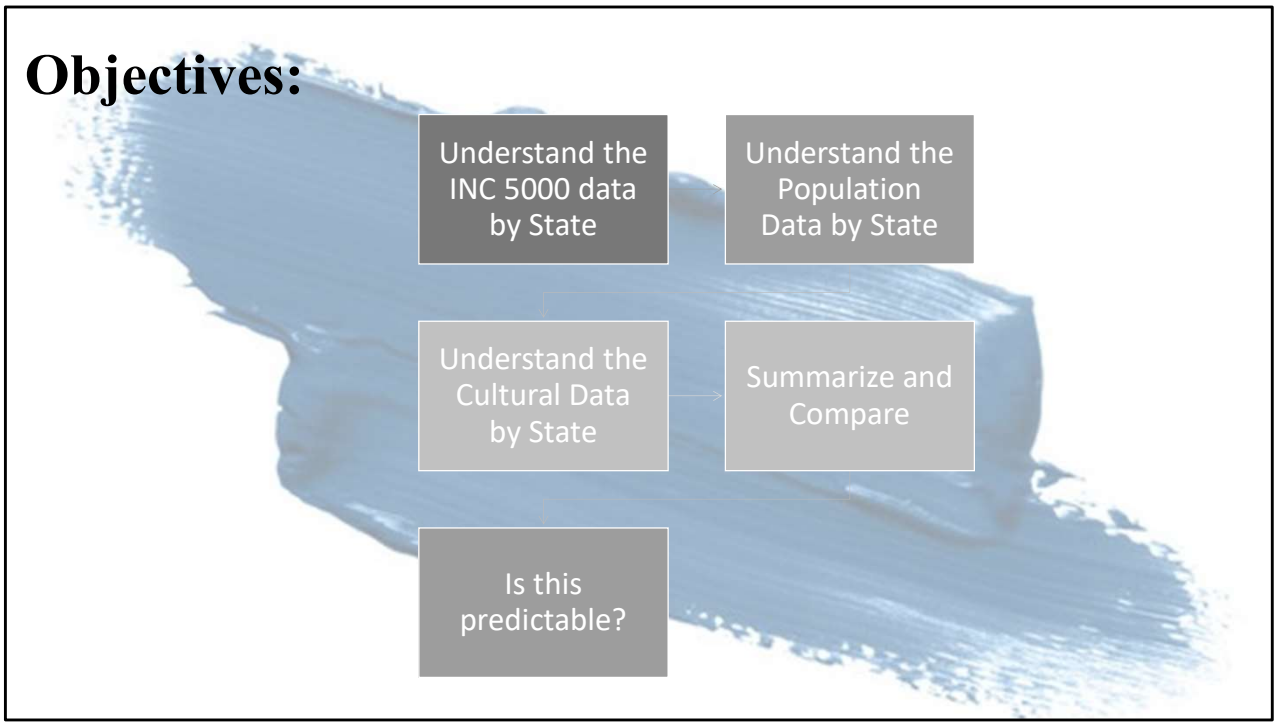
This project compares the growth of these companies with Census data collected by the United States.

The main goal is to see if there is a relationship with the growth and these attributes:

- Population
- Poverty
- Education
- Income
- Urban Density

Main Point: "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. " Be sure to impress this is an investigation and the initial step in the workflow.

Objectives:



Main points:

-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.

By prediction is it possible to use the results from summarize and compare to model the outcome of business growth?

Hypothesis:

- **Does each variable have normal distribution**?
  - Ho: It does not have normal distribution
  - Ha: It does have normal distribution

- **Is there a strong correlation between the state population, growth, urban density, and Business Growth**?
  - Ho: There is no correlation
  - Ha: There are strong or weak positive or negative correlations.

- **Is there a strong correlation between the population's education, earnings, and poverty with the growth of these companies**?
  - Ho: There is no correlation
  - Ha: There are strong or weak positive or negative correlations.

- **Is it possible to identify a combination of these attributes that would either enhance or discourage company growth**?
  - Ho: It is not possible or there is no definitive combination
  - Ha: It is possible to find a specific combination

- This is the key slide for the analysis, this is what we are setting out to answer.

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.

The test of normality is mainly to ensure the data should be used for correlation however based on the results we know it is not. This may be a difficult point to sell.
Try and stress that correlation is used on data without a normal distribution daily.

Literature and Limitations:

Literature	Limitations
➢ Difficult to find examples of this analysis ➢ However, it is clear small business growth is an essential part of economy ➢ Maps have been used to find correlation ➢ The following is the criterion to included in this list 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	➢ Missing Data such as number of employees, initial capital, total costs, and other proprietary attributes. ➢ There is a problem with the population in Puerto Rico this is not a serious limitation as we are focusing on the lower 48 states. ➢ This led to a very thorough review of all the attributes used to find and mitigate any other issues.

Please stress these statistics:

"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.

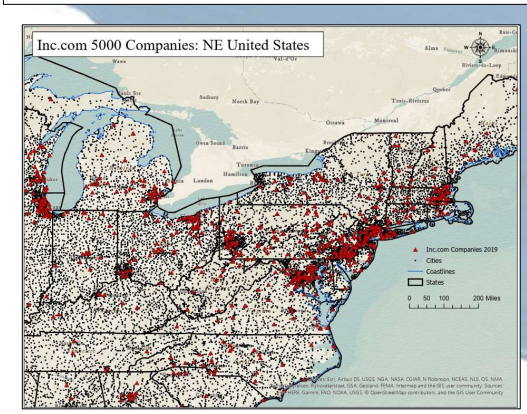
This idea is what led to this analysis in the first place. To further cement how important small businesses are (Farlie et al., 2019) looks at the universe of start-ups in the US and combines it with a panel data set. It also confirms the importance of small business by highlighting how there are nearly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

“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.”

“

Methodology:

Collecting and Mapping	Discussion:
	➢ This process involved scraping the 5000 records from the public Inc.com website into an excel sheet. ➢ To make the map the names of the cities had to be corrected and when a company used a regional or common name it had to be linked to the closest census designation. ➢ This data is at the city level with corresponding state and county.

This was the initial process that led to the entire project. After spending several weeks scraping the records and then performing an initial analysis it was clear this unique data set and warranted
Further investigation.

The decision was made to compare it to census data because this data is publicly available and also has multiple scales from city to state.

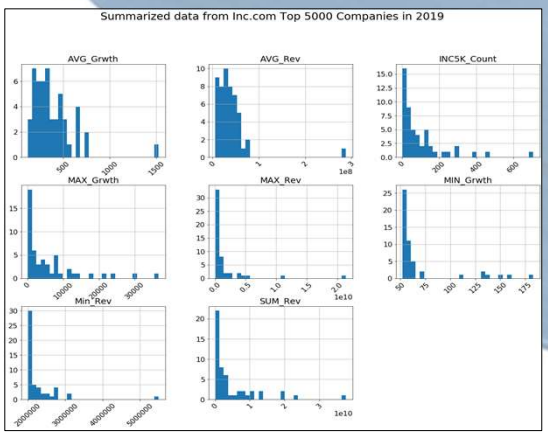
Methodology (cont.):

Software and Languages	Tools and Analysis
➢ Python: Used to test the data distribution ➢ SAS Studio used to explore the relationships between the data ➢ Microsoft PowerBI for advanced visualization and live functionality ➢ Microsoft Excel for the initial analysis	➢ Shapiro – Wilks Test of Gaussian (Normal) Distribution ➢ Pearson Correlation Coefficient ➢ Boxplot visualization which allows the comparison between three variables

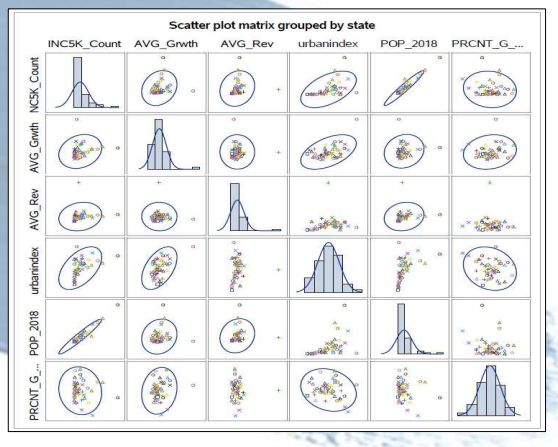
A quick list of the software and analysis

Results:

Distribution: Not Normal



Data Exploration: There are Correlations



H1:

"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."

Not normal distribution

H2: and H3:
“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 6 (image on right) 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. “

Results: SAS, Excel, and Correlation

6 With Variables:	Under_30962 Prcnt_UND_30962 POV30962_pr_100K TOTAL_ERNGS MALE_ERNGS FEMALE_ERNGS
8 Variables:	INC5K_Count AVG_Grwth MAX_Grwth MIN_Grwth SUM_Rev MAX_Rev Min_Rev AVG_Rev

Pearson Correlation Coefficients, N = 49

	INC5K_Count	AVG_Grwth	MAX_Grwth	MIN_Grwth	SUM_Rev	MAX_Rev	Min_Rev	AVG_Rev
Under_30962	0.94020	0.09627	0.51815	-0.30828	0.83969	0.39520	-0.29659	0.12884
Prcnt_UND_30962	0.02805	-0.04189	0.00304	-0.04205	-0.03781	-0.16544	-0.07795	-0.17295
POV30962_pr_100K	0.02805	-0.04189	0.00304	-0.04205	-0.03781	-0.16544	-0.07795	-0.17295
TOTAL_ERNGS	0.11679	0.49925	0.47403	-0.11217	0.09390	0.12133	-0.01604	0.09695
MALE_ERNGS	0.06331	0.41901	0.40276	-0.03933	0.05438	0.10197	0.05450	0.09461
FEMALE_ERNGS	0.14327	0.52038	0.48738	-0.10911	0.11198	0.12260	-0.03499	0.07689

← SAS Results

Total Correlation Results

TYPE	_NAME_	INC5K_Count	AVG_Grwth	MAX_Grwth	MIN_Grwth	SUM_Rev	MAX_Rev	Min_Rev	AVG_Rev
CORR	EDCTN_Total	0.95837443	0.117512879	0.533639868	-0.334520798	0.866925197	0.436893354	-0.32382892	0.168162622
CORR	NoSch_pr100k	0.718270471	0.114037509	0.424655165	-0.293470236	0.620284559	0.311313339	-0.31450816	0.097395715
CORR	HighSch_pr100k	-0.344818138	-0.43273001	-0.45530853	0.183269009	-0.268042256	-0.167224479	0.007789392	-0.060061869
CORR	Assoc_pr100k	-0.158593613	-0.33443077	-0.32595045	0.183546239	-0.034713529	0.136962119	0.377398242	0.171607238
CORR	Bach_pr100k	0.20286021	0.403517323	0.389443433	-0.154253725	0.225400726	0.257170692	-0.014613706	0.212212756
CORR	Master_pr100k	0.134449132	0.503673458	0.492273075	-0.102802087	0.066652499	0.034616351	-0.164042979	0.007250042
CORR	Proff_pr100k	0.099278298	0.691999682	0.588892256	-0.095619547	0.060373737	0.04157868	-0.012800041	-0.003313791
CORR	Doct_pr100k	0.058468016	0.520529414	0.446226856	0.01076076	0.018307242	0.01283099	-0.012794489	-0.034872433
CORR	No_schooling	0.929680516	0.128021085	0.516713438	-0.231773538	0.841011145	0.433261571	-0.22345575	0.104917074
CORR	high_school_dipl	0.903790763	0.080400166	0.486599084	-0.356375436	0.82597632	0.408737106	-0.348295696	0.187129101
CORR	Associate_degree	0.937441255	0.107042025	0.50574243	-0.33882388	0.865813659	0.465529043	-0.320559653	0.203164266
CORR	Bachelor_degree	0.967577922	0.147423554	0.556496777	-0.334322006	0.879299703	0.462130546	-0.321344059	0.183157867
CORR	Master_degree	0.946807419	0.142144369	0.552997624	-0.33894313	0.833189201	0.420250626	-0.335264821	0.158913279
CORR	Professional_degree	0.94655521	0.158539752	0.53478203	-0.313415643	0.846626739	0.442820354	-0.303181356	0.154134541
CORR	Doctorate	0.95134486	0.137038568	0.524643805	-0.317811195	0.857109878	0.451708281	-0.314010385	0.157415485

Scale	
1	Strong Positive
0.5	Weak Positive
0	None
-0.5	Weak Negative
-1	Strong Negative

TYPE	_NAME_	Intercept	INC5K_Count	AVG_Grwth	MAX_Grwth	MIN_Grwth	SUM_Rev	MAX_Rev	Min_Rev	AVG_Rev
CORR	urbanindex		0.52	0.50	0.60	-0.40	0.42	0.20	-0.37	0.13
CORR	POP_2018		0.96	0.14	0.54	-0.32	0.86	0.43	-0.30	0.14
CORR	PRCNT_G_1718		-0.09	0.15	-0.03	-0.14	-0.08	-0.05	0.09	-0.01

"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. "

"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

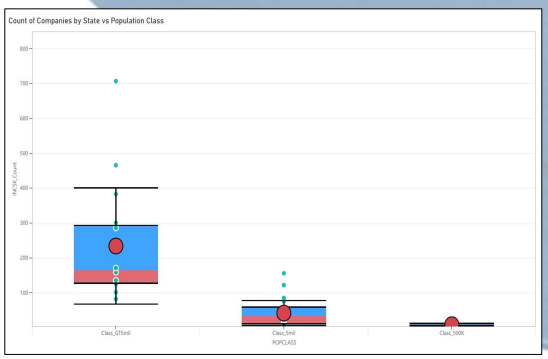
strong correlation with the count of companies per state and not with the growth. “

“However, the population has no correlation with average growth, but does have a weak negative correlation with minimum growth. Figure 9 (Bottom Table) shows this correlation table. “

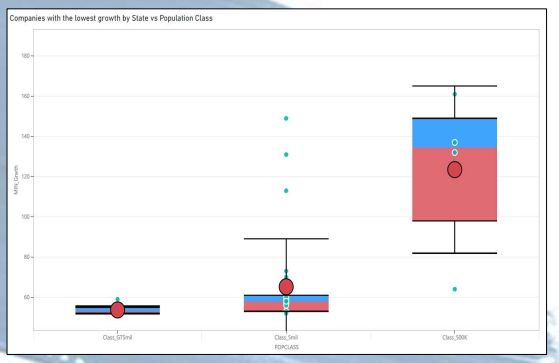
Make a note of how the education correlations seem strange and need further investigation.

Results: Box Plots

Box Plot: Population vs Total Count of Inc 5000 Companies



Box Plot: Minimum growth per state vs population



Please note the statistics and the outliers. It's important to stress this is one case and there are many more to be investigated.

"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. “

Conclusion:

- Of the original 4 hypotheses this analysis rejected the null on the last three but not the first. The data does not have normal distribution but there are strong correlations, and it is possible to combine them to build a case
- There is a negative correlation with low growth and population. The strongest correlation showed the more population the more companies on the list you have
- This analysis has shown the even though lower population decreases the count, the states with a population less than 500,000 improve the companies with a lower growth rate
- This is the initial pass on the data and because of this workflow many more cases will be discovered
- With the established workflow the data can be refined, and more relationships investigated

The main goal of this entire workflow is to sell it as something that can be applied to other hypothesis. There is more to be discovered based on these 5000 companies and we need to secure resources so we can continue this analysis.

"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. “

“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

Although these were not used in the slides, there are references in the notes.