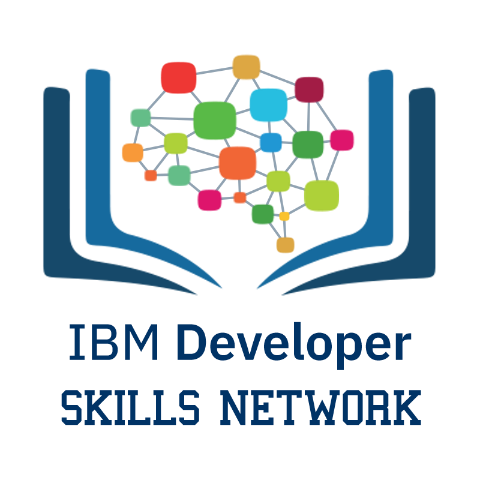
**Capstone Project Report**

**Timothy Scott Marinaro**

**Applied Data Science Capstone, by IBM**

**October 17, 2020**





# **Business Problem & Audience**

# **Background and Problem Statement**

# Graduates of America’s top-ranked public universities are extremely boastful about their college experiences, certain that they enjoyed a collegiate experience superior to that of all others. Long after graduation, they brag not only about the academic reputations of their chosen universities (which are highly quantified, and supported by a cottage industry of data providers) but also about their experiences in these iconic American college towns, each purporting to have chosen the best and most quintessential.

# This calls into question how important setting actually ends up being in the collegiate experience – essentially a question of heterogeneity, though one which has not been subject to the same rigor as has the academic quality of the schools in question.

# The underlying question we’ll try to address with this project is: are all college towns the same?

# **Objective**

# Using various data about the large, highly ranked US public universities and the cities in which they reside, we’ll try to answer the following questions:

# Do any of these college towns really stand out from the crowd in offering a unique experience via bars, restaurants, parks, and more?

# Do they cluster into a handful of types?

# Or are all of these well-known “college towns” more or less inseparable despite the insistence of those who lived out their college years there?

# **Audience**

# Primarily, this would speak to those students considering attending schools of this variety, helping them determine how much emphasis they should potentially place on basing their selections on other than academic ranking. And, in fact, some of the ways these cities and towns are quantified in the process could well become the basis for predictive models – helping students determine which locations and schools might best fit their preferences in terms of amenities.

# **Data Selection and Utilization**

While there is a good deal of available data which can inform our project, it does not reside in any one singular data set. As such, we relied on collecting, joining, and normalizing data from various sources. This includes:

* + Scraping certain data from HTML sites directly
  + Import .csv files that have been created or found from other sources
  + Making API calls

Once these data have been selected and culled, they were brought into an environment (a Juypter Notebook) in which data science tools could be deployed – in this case, Python and a host of associated libraries which can be seen in the submitted notebook – in order to clean, structure, and manipulate them toward reaching conclusions.

While we considered a great number of potential sources, those ultimately selected included:

* + *U.S. News and World Report* 2021 rankings of US Public Universities, particularly those ranked among the top 40 schools
  + A suite of geolocation and city population data from https://simplemaps.com/data/us-zips
  + Foursquare location data via their developer API

The main challenge was to make uniform the values in the various data sources (cities names that were structured in a common way, e.g.) and to stitch together a variety of useful dataframes from smaller tables by marrying a common element. It was these dataframes which allowed us to actually start to apply data science methods.

# **Methods**

While we did employ certain basic statistical techniques such as correlation, our approach relied most predominantly on visualization, which we found to be extremely enlightening in understand our data. We made heavy use of a variety visualization techniques afforded by our imported libraries. This included, in particular, developing:

* + Scatter Plots
  + Box Plots
  + Geospatial Mapping
  + Histograms

The resultant visualizations can be seen in the notebook and PowerPoint presentation. But, suffice to say, each of these were instrumental in drawing initial and subsequent conclusions – both finding apparent patterns worth validating, but also demonstrating outliers, other lines of questioning we might pursue, etc.

In order to help us find patterns that wouldn’t otherwise be very obvious to us, we turned to the popular unsupervised machine learning clustering algorithm of K-means. To leverage the available Foursquare location data, we stood on the shoulders of our week 3 assignment and turned to K-means clustering, a partitioning technique which relies on the similarity and dissimilarity of various attributes. In this case, the main criteria used was the preponderance (or absence) of certain types of venues in the cities in question, as derived from the Foursquare location data. The schools in question were grouped into 5 clusters based on their similarities in prevalence of each type of venue in the overall Foursquare data set.

# **Results & Discussion**

The existence of outliers in various measures (student enrollments, city population, etc.) in and of itself suggests some degree of heterogeneity.

Once our list of comparable schools came into focus, we could see at first-glance some of the theme we would have expected in terms of the frequency of various types; namely, casual food and drink options with a social bent. However, closer comparisons called out some more nuanced findings that suggested individuality. College towns are actually extremely distinct in terms of the frequency of occurrence of certain venue types:

* Of the 60 distinct venue types appearing in the table created – a top ten most-appearing venue types by school – almost half (29) appear only once in the table
* Another 13 show up on twice.
* 6 of the 16 college towns had 3 unique venue types among their top 10
* And only 1 school (The University of Georgia) has no unique venue types in its 10 most-occurring

Also of note, once the clustering was complete, was the extent to which geography didn’t seem to be a major determinant in the ways college towns tended to cluster. Among our 5 resultant clusters:

* Two schools in Florida matched up closely with one each in Texas and Indiana
* Northern CA schools did seem to cluster, but The University of Virginia also joined their cluster
* One cluster included schools in CA, IL, MD and NC;
* We also saw UCSD be assigned to its own cluster, likely because it was the lone to record beach venues, and moreover as its most frequent venue type
* Our other cluster features two schools in the upper Midwest and one in GA

# One of our first visualizations suggested another line of questioning worth pursuing. The schools we initially plotted as a scatter on the basis of student enrollments and population suggested that, a few outliers aside, populations of these college towns hovered pretty closely around the mean of 180K. This made us wonder if those students interested in attending the highly ranked large public universities in the US were more or less required to live in cities of a certain size. To test this, we increased our sample size to include all of the top 40 schools and then plotted the occurrence by city population using a histogram (leaving out Los Angeles, as it would likely mean significant skew). The resultant histogram suggested that students do not need settle for living in cities of certain sizes – that there is a lot of diversity in the sub-500k population. But because this diversity only became apparent once we included the additional schools ranked 21-40, it begged the question of how correlated ranking might be to population. Bearing in-mind a small data set (which can make correlation a bit more fraught) we found that:

# There were virtually no correlations of note in the table

# The highest correlation was between latitude by ranking, but only at .264.

# The comparison we looked at between ranking and city population is so miniscule .04, as to suggest that the factors are almost completely independent given our data set after all

# **Conclusion**

Even despite what seemed to be some apparent patterns of homogeneity among our examined college towns at first glance, this examination basically proved that there’s great diversity in the schools. Among the key takeaways, it could be said that highly-ranked public universities in the United States:

* Come in a variety of sizes
* Are in cities of widely divergent geographies and population sizes
* Offer wide arrays of amenities even when geographies are alike; the converse that geography doesn’t necessarily dictate amenities also being true

The near-complete lack of correlation across the factors we examined -- geospatial data, city population, and school ranking, e.g. – suggests that these schools, and the cities in which they reside, are fiercely unique and that the choice of setting for one’s collegiate experience is significant and should factor into the decision-making process. Also, in good news for would-be students of these schools: there’s a wealth of choice at their disposal – they can select from a variety of high-quality public universities across a wealth of truly unique settings.