

**University clustering based on academic profile and surrounding venues**

**Shaleen Parikh**



June 22, 2020

Table of Contents

[Introduction 2](#_Toc43744938)

[1.1: Problem: 2](#_Toc43744939)

[1.2: Data: 2](#_Toc43744940)

# Introduction

## 1.1: Problem:

In the United States, millions of high school seniors are faced with the task of applying to universities and may experience difficulty as they attempt to figure out which universities are a good match for them. With thousands of universities to choose from, it is difficult to find a match. Students want to know about the available facilities near the universities and the academic profile of the university in order to narrow down their choices. Using the Foursquare API and other data to cluster universities, we can help high school seniors discover universities that have similarities based on their respective surrounding venues as well as their academic attributes. Together, the clusters will allow high schoolers and other university applicants to find personally suitable colleges based on both location and academic status.

## 1.2: Data:

An external dataset provided by IPEDS Data Center, which is in a CSV format, has information about the universities, including their name, latitude, longitude, address, city, SAT scores, ACT scores, price, acceptance rate, and student-to-faculty ratio. This data will be stored in a Python pandas dataframe so that meaningful analysis can be done on the data. First, the academic attributes (SAT/ACT scores, student-to-faculty ratio, acceptance rate) of the universities will be used to cluster them into categories.

Another data source that will be used is the Foursquare API. The latitudes and longitudes of the universities will be passed to the API in order to receive insights on the surroundings of the universities, including the nearby venues. This will require the use of the explore keyword in order to get a prespecified number of venues within a prespecified radius of the universities. Since the venue information is returned as JSON, it will be necessary to filter the unnecessary information and store the necessary information in a Python pandas dataframe. The categories of the venues will be sorted based on their frequency. This information will then be used to cluster the universities based on the nearby venues.

Together, these data sources will allow for two types of clusters to be formed to characterize each university: one with the academic characteristics, which come from the IPEDS Data Center dataset, and one with the characteristics of surrounding venues, which come from the Foursquare API.