Statistics 141A Final Project
Fall 2016

Ruriko Imai
Sophie McGuinness
Shogun Sherchan

Introduction:
School construction has been a heavily discussed topic due to the high costs yet vague impact on student success. Many papers have studied the effect of spending on student's performance and teacher's morale around the country. However, none have looked at the type of construction project on which money is being spent. We pre-processed our data on construction projects in California Community Colleges in such a way that we could differentiate the types of construction projects. Did the school build a new library or sports facility? We believe this information tells an important story about school construction. This analysis will serve as a preliminary examination of spending trends across time, region, and spending category. The goal is to reveal never before seen information about the type of school construction spending in California community colleges. We will attempt to answer the following three questions: Which community colleges are spending the most and in which areas? How does this change from before and after the 2008 Financial Crisis? What are the most common construction projects by region?

Data and Methods:
We will use a previously unexplored categorical and continuous panel data set that includes data on every construction project in all community colleges in California since 1998 (n = 3,568). The primary variables we will focus on are the year the construction project was funded, the final cost of the construction project per student (in 2015 USD), the location of each college, and the type of construction. The categorical variables describing the type of construction are: educational buildings, career technical facilities, general facilities maintenance, administrative buildings, and community buildings.

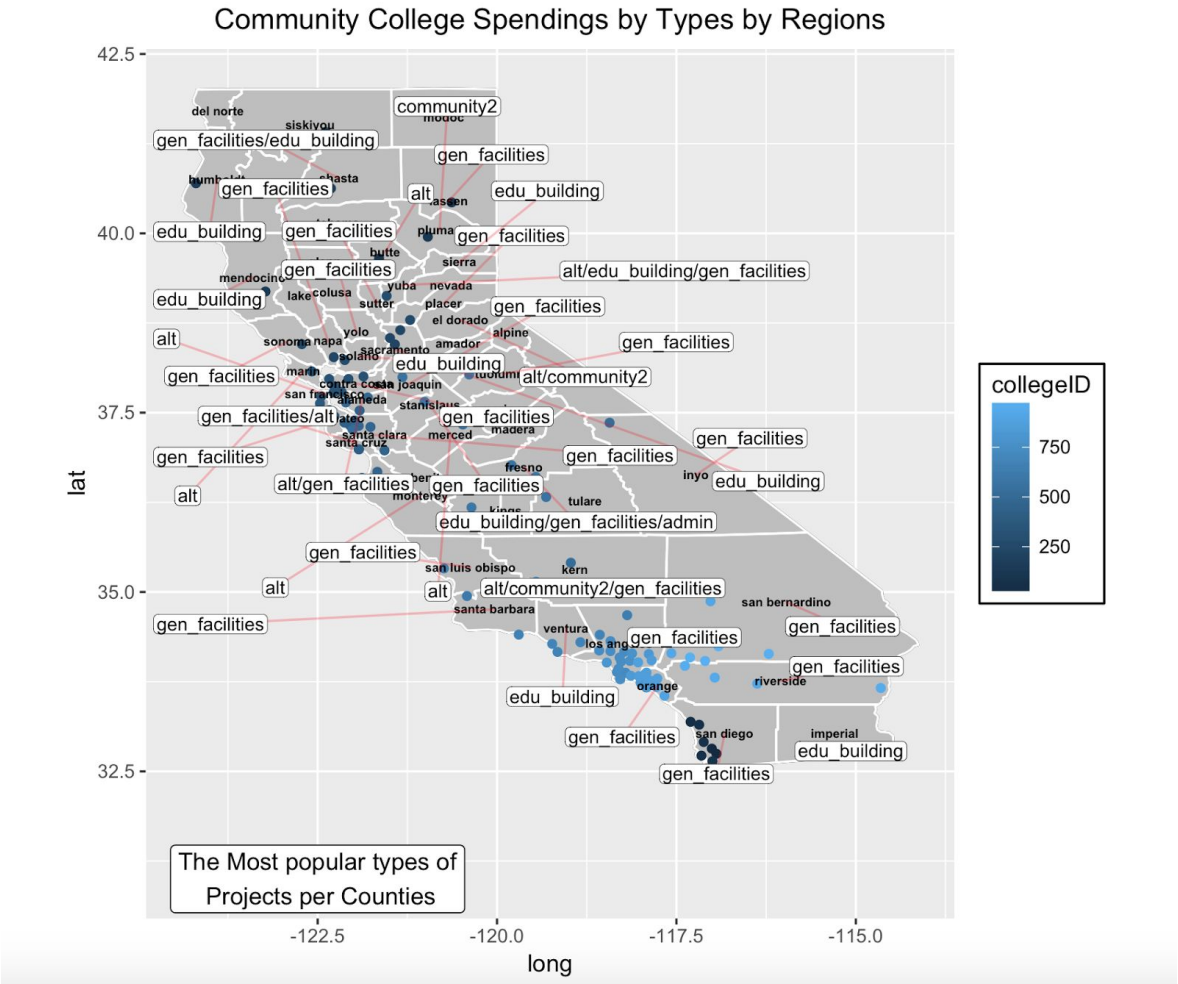
Pre - processing:
We used regex commands in Stata to identify keywords in the variable "scope" in order to code the six categorical variables. A complete list of the categorical assignments can be found in the pre-processing Stata Dofile. Entries with a 1 for "edu_buildings" include construction on classroom buildings and laboratories such as auto shops, welding studios, and science labs. "Career_tech" is a subcategory of "edu_buildings", and includes projects that are specific to vocational studies, such as machine shops, nursing labs, cosmetology, and any classrooms that are specifically geared toward vocational training. This is an interesting category because Career Technical Education (CTE) is becoming an increasingly popular alternative to four year degrees, and many economists believe it is essential to meeting the labor demands of our economy. Entries with a 1 in "gen_facilities" include maintenance projects like replacing toilets or building roads. "Admin" indicates projects related to upkeep or construction of administrative facilities. "Community1" includes facilities like libraries and student service centers that are focused directly on the student community. "Community2" projects are

construction projects that can be used by the greater community, such as theaters, sports facilities, and gym equipment.
In addition to these categorical variables, we also used regex to assign ID numbers to the community colleges so that they could be identified consistently. We merged in CPI information to convert all dollar amounts to 2015 USD. Finally, we used the college IDs to merge in data from the State of California on enrollment numbers. This allowed us to create the variable "spend_student", which is the amount of the construction project in per-student dollars.

Method:
The method that will be used to analyze this set of data will be mostly comprise of statistical analysis of graphs and patterns that is uncovered through the data mining process. However, we will also be using computational methods as a process of constructing graphs and further delving into the exploration of the given data set.

Section I: Visualization of Spendings of Community Colleges throughout California

To attain a better understanding of the data, we constructed a clear visual image of the community colleges located in California. The map plots the locations of all the schools, unique by their college ID. The visual image helps us locate several clusters of where the mass of the schools are located, such as the County of Los Angeles, Santa Clara area, Orange county, and so on. The white label indicates the most common type of project/projects that has been executed by the schools in the individual counties. Clearly, the majority of the schools spread across the state has similar execution on types of projects, that is, on general facilities. Although the overall execution of projects are on general facilities throughout the 15 years (1998 - 2016), the rises and falls of the types of projects may vary every year. These curious results obtained by the geographical map may raise some questions as to how much money was spent overall, and the amount used by each community college on the type of constructions. In the following section, we will explore more closely by analyzing spendings by years and type of constructions.



Section II: Spending per student and trends in Spending from 1998 to 2016

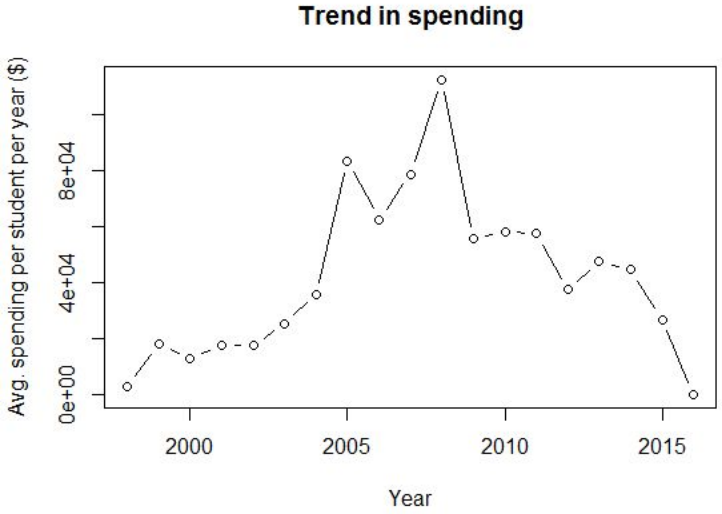
In order to fully understand the nature of the spending in community colleges across California between 1998 to 2016, it is important to summarize the data which provides a snapshot of the entire data. In this section, we also explore how spending patterns change across the 2008 Financial Crisis, which had a huge impact on real estate prices and construction projects all throughout the economy.

Since the goal of this project is to study the spending behavior, table of some important descriptive statistic are provided below.

	Spending Per Student	College ID	College Name
Total spending (98 - 16)	$ 795,476.10	NA	All colleges
Average spending per year (98 - 16)	$ 41,867.16	NA	All colleges
Standard Deviation (98-16)	$ 29,487.49	NA	NA
Total spending (09 - 16)	$ 328,023.50	NA	All colleges
Average spending (09 - 16)	$ 41,002.94	NA	All colleges
Standard Deviation (09 - 16)	$ 19,756.89	NA	NA
Total spending (98 -08)	$ 467,452.60	NA	All colleges
Average spending (98 - 08)	$ 42,495.69	NA	All colleges
Standard Deviation (98-08)	$ 35,928.48	NA	NA
Biggest Spending on a single project	$ 7,253.26	971	Cooper Mountain Community College
Biggest Spending overall (98-16)	$ 23,815.15	742	Los Angeles Harbor College

One surprising piece of information that can be taken from the table above is that average spending before and after 2008 hasn't changed drastically. However, there is a large difference in standard deviation of average spending before and after 2008. Further investigation of this difference in standard deviation could unlock the spending behavior of community college across California since 1998.

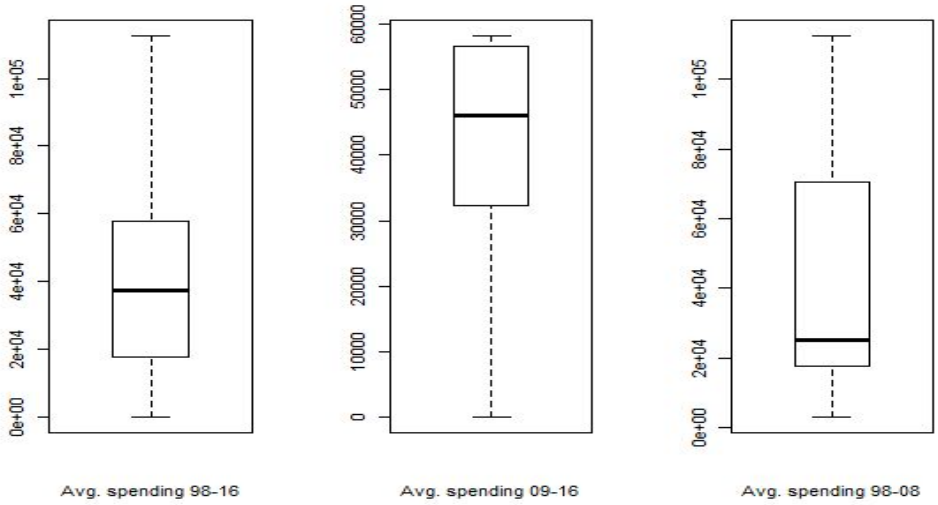
One aspect of descriptive statistics that could help gather important information about the data are graphs.



There seems to be an obvious trend in the data. The overall trend is somewhat parabolic. However, more interesting feature of the graph above is the deviation away from the initial trend. The critical point seems to be the year 2008 in which the trend started to deviate in a different direction. One can argue that both trends before and after 2008 is somewhat linear in nature. The graph also explains the large difference in standard deviance of average spending before and after 2008 that is reported in the table above. The large difference in standard deviation is due to series of large jumps in spending before 2008, especially after 2004, compared to mostly gradual decrease in spending after 2008. There large boost in

spending after 2004 drastically affected the spread of average spending before 2008 since the jumps resulted in larger range in average spending. This phenomenon is also responsible for the similarity in average spending before and after 2008. The graph depicts that there were decrease in spending in 2008 and onwards, with one sharp drop after 2008, however, the decreases were gradual enough to keep the average spending at the 1998 - 2008 level.

Another way of visualizing the difference in standard deviation and similarity in average spending is through box-plot. Below is the boxplot of average spending of 98-16, 98-08, and 09-16.

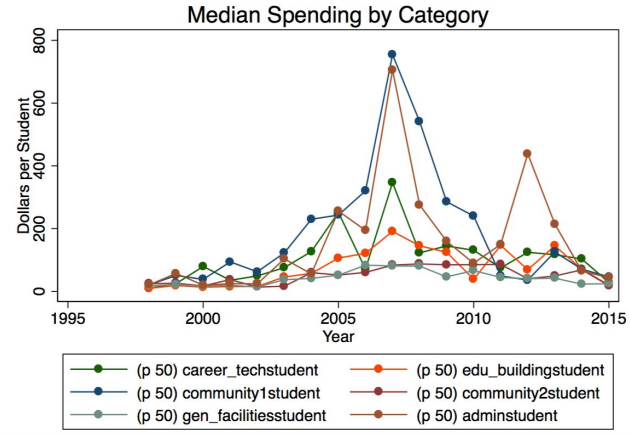


After analyzing the above box plot, it is clear that the average spending from 98-08 has the largest standard deviation as the data spreads from low 100's to over 100,000. However, it is rather difficult to pinpoint the location of the average spending due to its asymmetrical structure. With some experience in reading boxplot, it is possible to approximate the average with certain degree of accuracy. For instance, the plot of 98 – 08 is right skewed, therefore, one would expect the average to be somewhere to the right of the median since the points in the upper quantile has more weight and pulls the average towards it. $40,000 is a good approximation of average spending which is around the spending average reported in the table above.

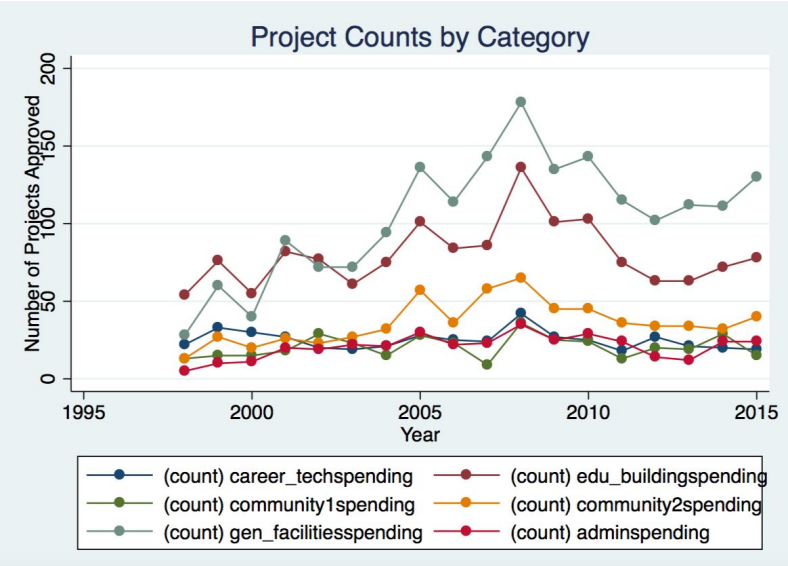
We also analyzed how spending changed across categories before and after 2008. The graph below shows the median spending per student in each of the six spending categories. Once again we see an overall trend of large increases of spending leading up to 2007 and 2008, followed by a sharp collapse and downward trend. It is interesting to note that the largest increases and collapses were in community1- student community, and administrative spending. It seems that these categories received a relatively large amount of financial attention, because they were in the middle of the pack from 1998 to about 2003, and since 2010 have come back down to the levels of the other categories.

Another interesting finding here is that the only spending category that increases during the period of 2008 to 2010 is career technical spending (shown in green).To ensure that these

trends were not driven by outliers we graphed the median spendings and performed an outlier analysis (below).

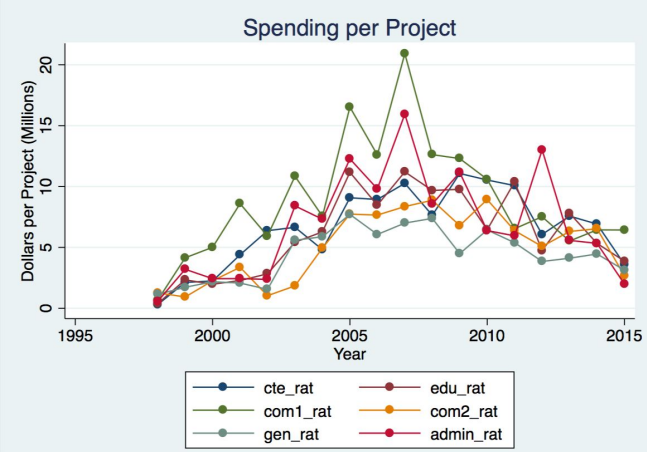


Another interesting metric we examined is the count of projects per category over the years. We can see that general facilities projects are the most numerous even though the previous plot showed that their median amount spent is among the lowest. This plot also indicates a rise and fall with the Financial Crisis. However, administrative spending and community1 spending, which the previous plot showed to have the sharpest booms and busts, are now revealed to have remained at a constant number of projects over time. This means that the monetary size of administrative and community1 projects must have driven the spike in their respective spendings. We also see that general facilities and educational building projects experienced the largest spike in number of projects, whereas they seemed to be relatively constant in the previous graph of median spending.



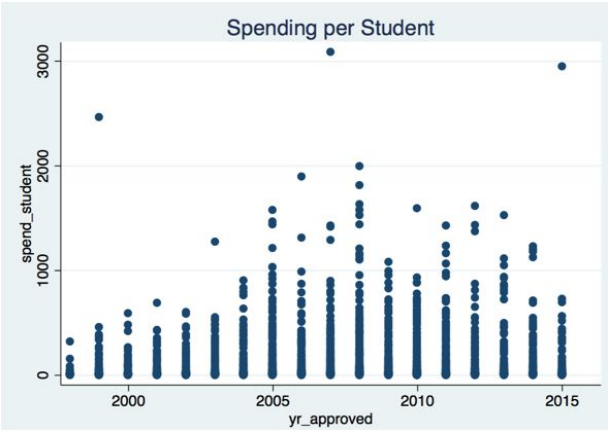
The next graph confirms our suspicions about the growth in monetary size of projects by graphing spending per project over time. The lines represent overall spending in each category

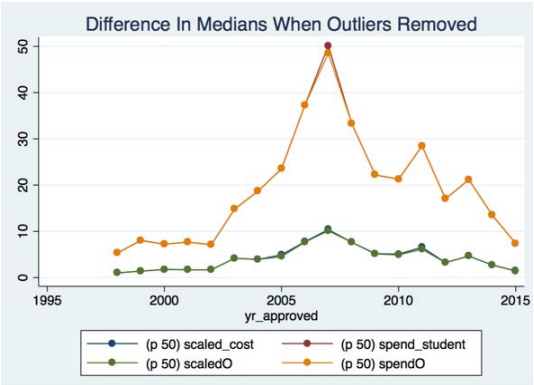
in each year divided by the number of projects approved in that year. It accounts for student population at the school by using per student dollars. While all categories grew in size leading up to 2008, we see that community1 and administrative projects grew the most. Administrative spending, which was among the smallest project categories in the late 90s and early 2000s, surges to become one of the largest, leading one to question why office and administrative buildings became as expensive as the libraries and student centers contained in community1. Once again, we see that career technical spending per project (cte_rat) is the only category to show a net increase from 2009 - 2011. If we had more time and resources, it would be interesting to track down the rise and perseverance of CTE spending.



Analysis of Outliers:

The variable spending per student exhibited only 3 major outliers. Removing these outliers had little effect on our analysis. The second plot below shows spending per student in red, then spending per student with outliers removed in orange. The plot of the median changes only very slightly when outliers were removed.





Conclusion:

We have shown that the overall types of constructions are quite similar throughout California. However, from the above analysis, the financial situations and needs of the individual schools each year allowed some interesting raises and falls of the construction types the community colleges decided to spend on.

We also see evidence of effects from the 2008 Financial Crisis on construction spending around the state. While all categories of student spending rose and fell with the real estate bubble, it is interesting that a small rise occurred in career technical spending in the period from 2009-2012. Analyzing the size of each project category reveals a large growth in the size of community1 and administrative projects in particular. This means that schools were mostly building bigger libraries, student centers, and administrative buildings during the real-estate bubble rather than educational facilities. It would be interesting to see how the effects of these types of projects endured after the collapse, and whether community colleges would have been better off if those projects had been devoted to classroom buildings rather than office buildings.

This analysis is an important first step in identifying trends in construction spending in the most populous and diverse state in the United States. Further research into the effects of these spending trends on student success and community property values could add an important piece to the puzzle of school construction in America. One issue with our analysis of the spending categories is that the categories are not mutually exclusive. It was necessary to allow for overlapping categories because the construction projects often have two or three types of projects lumped into one bundle of spending. In the future, we could subset the data in order to make the categories mutually exclusive, and therefore analyze the categories more closely. Given more time and better resources, we would be most interested in pursuing the causes and effects of the rise in career technical education spending after the Great Recession.

Individual Role of the Group Members:

Sophie - Pre-processing and analysis of change in spending categories.
Shogun - Analysis of graphs.
Ruriko - Pre-processing to access locations of the schools, coding, analysis of mapped results.