UNIVERSITY OF CALIFORNIA
RIVERSIDE

Analysis on Universities' Impact on Home Price

Master's of Science

in

Computer Science

by

Yi Wu

March 2017

Committee:

Dr. Vagelis Hristidis, Chairperson
Dr. Vassilis Tsotras

ABSTRACT OF THE THESIS

Analysis on Universities' Impact on Home Price

by

Yi Wu

Masters of Science, Graduate Program in Computer Science
University of California, Riverside, March, 2017
Dr. Vagelis Hristidis, Chairperson

House has been an important element in American people's life. Mortgage is a big financial burden to many people. It is one of the most common and important investment tools. It also led to the biggest economical recession in 2008. People proposed many complicated models in order to understand how the home price change and what are the factors impact the home price. Schools is long known as one of the key factors impacting the home price. All parents want to have a mailing address in a high-ranking school district so the children can go to the good schools in that district. We also observe that the renting market near large universities is always hot. All those behaviors heat up the home price in certain areas. In this thesis, we analysis the universities' influence on home price quantitatively. We calculated the correlation between home price and university features in different types of homes based on home features like number of bedrooms and bathrooms. We also performed analysis on time series home price data and showed that the larger the university is nearby, the faster the home price increases.

Chapter 1

Data Source and Preprocess

The home price data is collected every week by previous and other current group members. It's crawled from major home price websites, including www.trulia.com, www.zillow.com and www.redfin.com, etc. The data is stored as a MySQL table on the group server. Other than the fields containing the data source, the table contains other fields describing the important aspects about the homes. Among the aspects, there are some important features we care about most, including "latitude", "longitude", "state", "zipcode", "numbed", "num_bath_full", "yearbuilt" and, of course, "saleprice".

The university data is collected from www.wikipedia.org. The data is also stored as a MySQL table on the group server. The table contains fields including "state", "postcode", "enroll", "latitude", "longitude", "acadStaff", "students", "underGrad" and "postGrad". The table contains 1200 universities.

The home price data is across many years and parsed differently by different group members. So we decide to select data only from June, 2016. This results in 183460 home records, which gives more than 100:1 home to university ratio. This ratio serves the purpose of the thesis very well. We output the table to a csv file and save it to the local machine.

We read the csv file with Pandas in Python and result in a 183460 * 39 table, called dataframe in Pandas. The table below (Table 1) shows the statistics of a few numerical columns in the dataframe.

	latitude	longitude	numbed	num_bath_full	num_bath_part	rentalprice_min	rentalprice_max	saleprice
count	183460.000000	183460.000000	183460.000000	183460.000000	183460.000000	183460.000000	183460.0	1.834600e+05
mean	33.545487	-93.689711	1.966156	1.393595	-0.697100	984.602317	-1.0	1.338273e+05
std	5.212622	14.886998	1.897425	1.669656	0.716976	3306.473516	0.0	6.412166e+05
min	19.090885	-161.786590	-1.000000	-1.000000	-1.000000	-1.000000	-1.0	-1.000000e+00
25%	30.361464	-94.455809	1.000000	1.000000	-1.000000	-1.000000	-1.0	-1.000000e+00
50%	33.323529	-87.718835	3.000000	2.000000	-1.000000	-1.000000	-1.0	0.000000e+00
75%	36.088063	-85.468457	3.000000	2.000000	-1.000000	1395.000000	-1.0	1.350000e+05
max	64.978010	-70.718870	12.000000	47.000000	1.000000	200000.000000	-1.0	1.500000e+08

Table 1. Sample statistics of numerical columns in the outputted csv file.

To find the relationship between universities and home price, we first pair each home to the nearest university. The number of home record is large, it's going to be a huge computation if we calculate the distance between each home and university. To save computation, we can subgroup the home and universities by city or zipcode. However, there are only 1200 universities. Not all city or zipcode have an university. On the other hand, a home can be on the or zipcode boundary and has a smaller distance to the university in the neighbor zipcode. To balance between speed and accuracy, we subgroup home and university by state. We compute each pair of home and university in each state and append the nearest university and the distance to each home. In cases a home has multiple nearest universities, we randomly choose one university to append. The distance is calculated using the home longitude and latitude (x1, y1) and university longitude and latitude (x2, y2).

$$distance = \frac{1}{69}\sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

We divide the square root by 69 to convert it to miles. Even though we should have used the orthodromic distance instead of the L2 distance, it's a good enough estimate when the distance between the two points are small compared to the sphere radius, which is the case in the project since we only consider the home if it's close to an university.

Chapter 2

Analysis on University Features

After the preprocess mentioned in last chapter, we have a big dataframe in each row containing a home record, the nearest university and the distance between them. As we know, the home price is very different in different locations. It's not fair to compare home price from different areas. One advantage the dataframe gives us is that we now can group the homes by their nearest university, so the comparison is performed locally.

In order to investigate the influence the universities have on home price, we first, for each university group, calculate the correlation (called "distCorr") between the home price and the distance to the university. The correlation is found by a Python package scipy.stat.pearsonr. We assume the university influence is very weak when distance is larger than 10 miles. So we confine the calculation in the homes within 10 miles to the university.

We first look at the statistics of the 1200 distCorr. We generate a boxplot for distCorr with a Python library, seaborn.

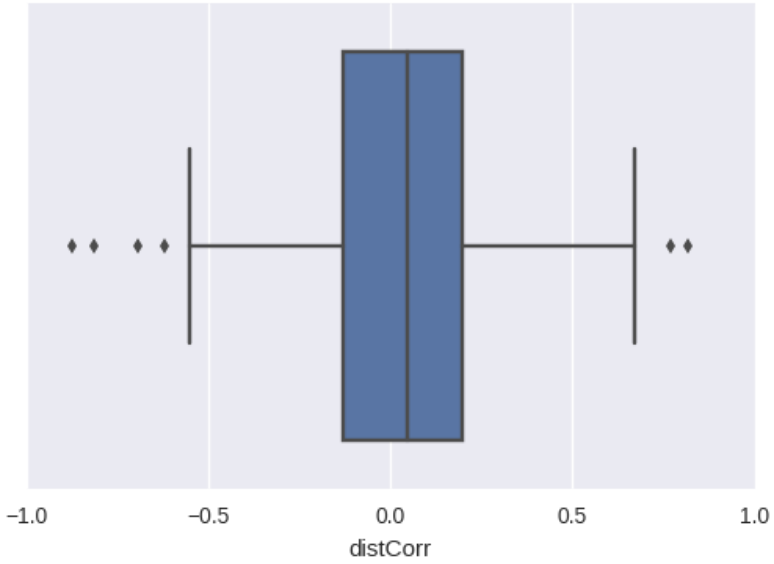


Figure 1. Boxplot of distCorr

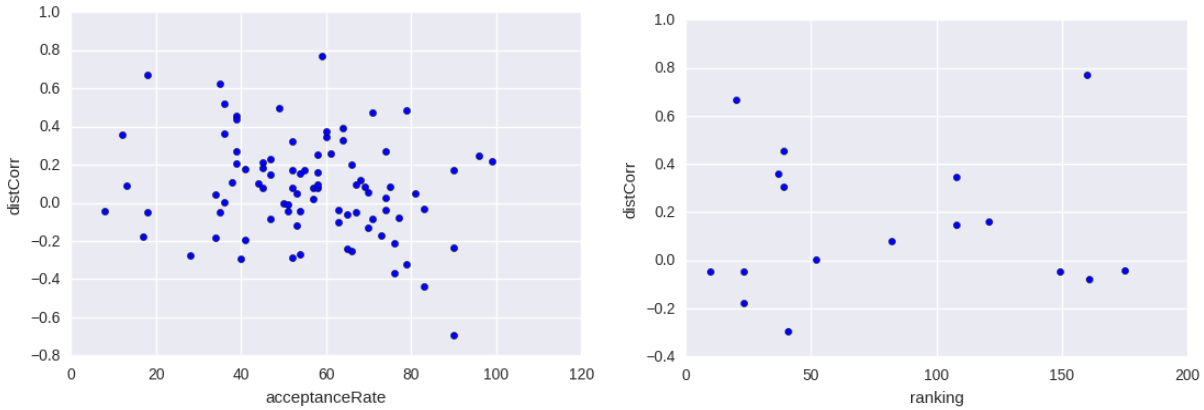
50% percent of distCorr is in zero's close vicinity. It's not a good idea to use correlation between distance and all home price within 10 miles of the university. More careful subgrouping is needed to find the distance's impact on home price.

We also compute the correlation between distCorr and numerical features to see if the home price is depend on any university feature. We summarize the correlation between distCorr and features in Table 2. The library scipy.stat.pearsonr calculates two values: correlation and two-tailed p value. Large correlation indicates strong dependence between the two variables. With small p value we have little confidence accepting the null hypothesis (no correlation) and thus large confidence assume they are correlated.

The plots below (Figure 2) are distCorr vs different features in university table. Each university with a non-null in feature value is represented by a dot. The distCorr is calculated on all homes within 10 miles of the university. The plots in Figure 2 don't show a clear correlation between distCorr and any of the university features. Therefore, we need to consider a certain type of the homes to better explore the university's influence on home price.

distCorr vs.	correlation	Two tail p value
Acceptance Rate	-0.213	0.0476
ranking	0.0399	0.8792
enroll	0.0585	0.5775
acadStaff	-0.2228	0.0531
num of students	0.0585	0.5691
num of underGrad	0.1610	0.1830
num of postGrad	0.0086	0.9461

Table 2. Correlation between distCorr and university features.



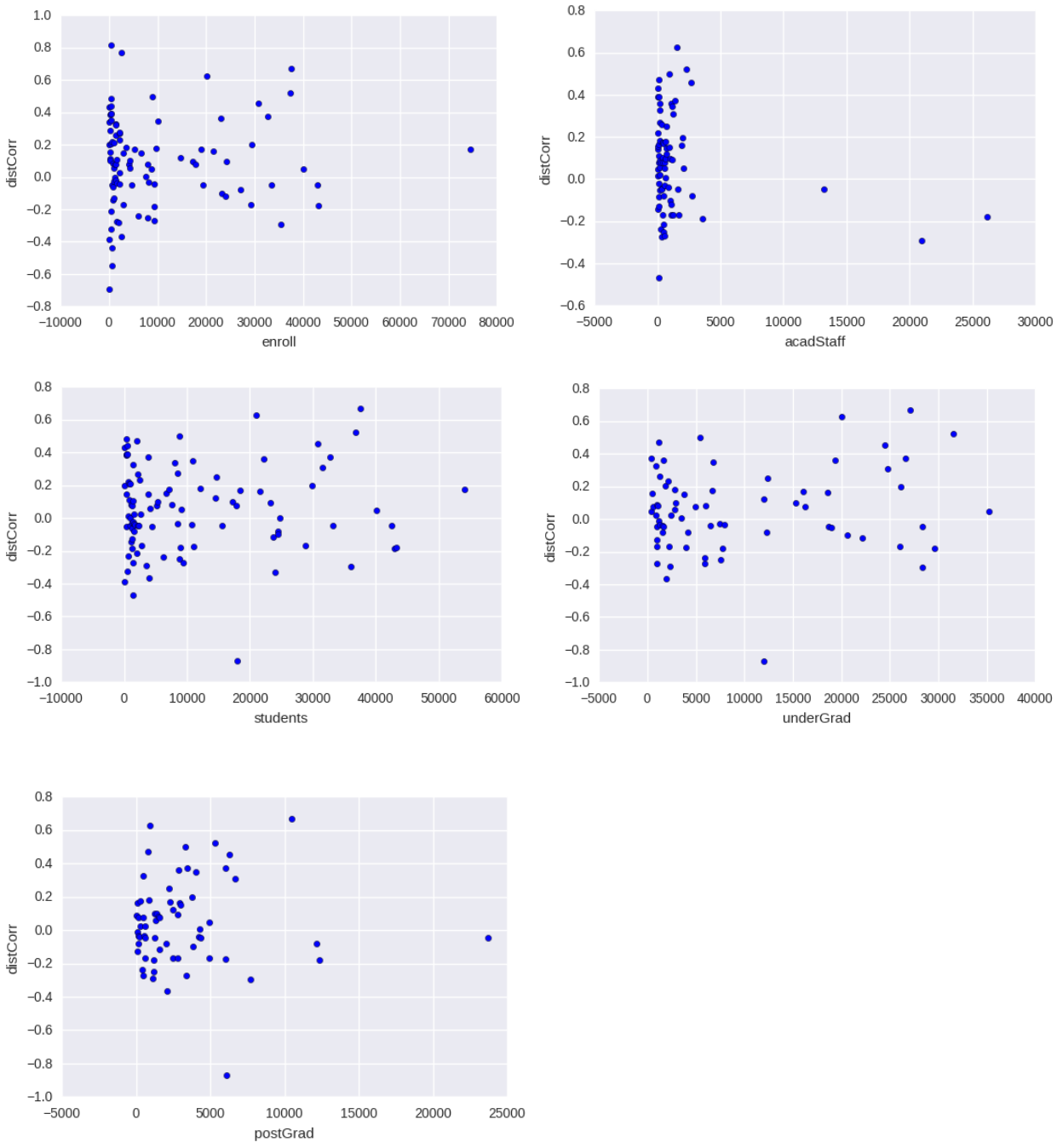


Figure 2. Plot distCorr versus numerical university features. Each dot represents an university. The plots don't show a clear correlation between distCorr and any numerical university features. The distCorr is computed with all homes within 10 miles of an university.

Chapter 3

Analysis on Home Features

We want to confine the exploration in one state to eliminate the extra variance brought by different state. Since the number of homes and universities are both sufficiently large in California, we choose California to be our target state. We count the number of homes paired to each university and find the top four universities with most homes. The universities and the number of homes within their 10 miles are listed in Table 3.

University	Num of homes within 10 miles
California State University - Sacramento (CSUS)	188
William Jessup University (WJU)	157
Soka University of America (SUA)	157
California State University - Bakersfield (CSUB)	157

Table 3. The name of the California universities that have most homes within 10 miles and the number of homes.

We continue to divide the homes within 10 miles based on certain features and explore if the universities have influence on those certain group of homes. One important home features is “Record_type”, which has value of “rent” and “sale”. However, “rent” doesn’t exist in the dataset we choose. So we ignore the record_type and look at other home features.

3.1 Number of Bedrooms

We first consider the number of bedrooms. Within each universities’ 10 miles vicinity, we divide the homes to subgroups based on the number of bedrooms the home has, and calculate the correlation between the home price and the distance to the university. As mentioned in chapter 2, we are particularly interested in large correlation (close to 1 or -1) and small p value (we usually consider 0.05 as small). We summarize the correlation and p value in Table 4. We fill the table field with “correlation/p-value (number of homes)”. We ignore the group if the number of homes is smaller than 5 to make the correlation statistically significant.

We can also plot the incidence of the home records in Figure 3, and the incidences in the highlighted subgroups (highlighted in Table 4) in Figure 4.

In most cases, the price doesn't show a dependence on the distance, while in the cases where the price does show a dependence on the distance, it shows a negative one. We expect the university students to rent the houses and so that pushes the house price higher. However, not all house types are the primary target types for students. Students typically don't target on the living environment, but rather care more about living expense. So they often seek for more bedroom houses so the rent and utilities expense can be shared. For example, one bedroom homes and studio are typically rent by family or working singles. This group of people needs more space and have the financial ability to pay for the expense. On the other hand, college students are more likely to rent a three or four bedroom houses. So we expect the price in those houses depend more on the distance to the university. Our assumption is supported by the data. We see a stronger correlation in the 4 bedroom homes.

Num of bedrooms	CSUS	WJU	SUA	CSUB
1	-0.02 / 0.97 (5)		(3)	
2	-0.12 / 0.51 (34)	-0.73 / 0.0006 (18)	0.09 / 0.63 (32)	-0.37 / 0.19 (14)
3	-0.15 / 0.19 (80)	-0.25 / 0.094 (47)	-0.098 / 0.55 (39)	-0.36 / 0.0065 (55)
4	-0.43 / 0.0052 (41)	0.08 / 0.58 (49)	-0.36 / 0.015 (45)	-0.39 / 0.0016 (63)
5	-0.62 / 0.19 (6)	0.095 / 0.62 (30)	0.18 / 0.37 (27)	-0.15 / 0.69 (9)
6	(2)	(3)	(2)	

Table 4. Summarization of correlation between home price and distance to the paired university in each university's 10 miles vicinity subgrouped by number of bedrooms. The subgroup with less than 5 homes are ignored in the table. The four universities are chosen because they have most homes within 10 miles. We are interested in subgroups with a large correlation (abs(correlation) close to 1) and small p-value (smaller than 0.05). The subgroups satisfying this condition are highlighted.

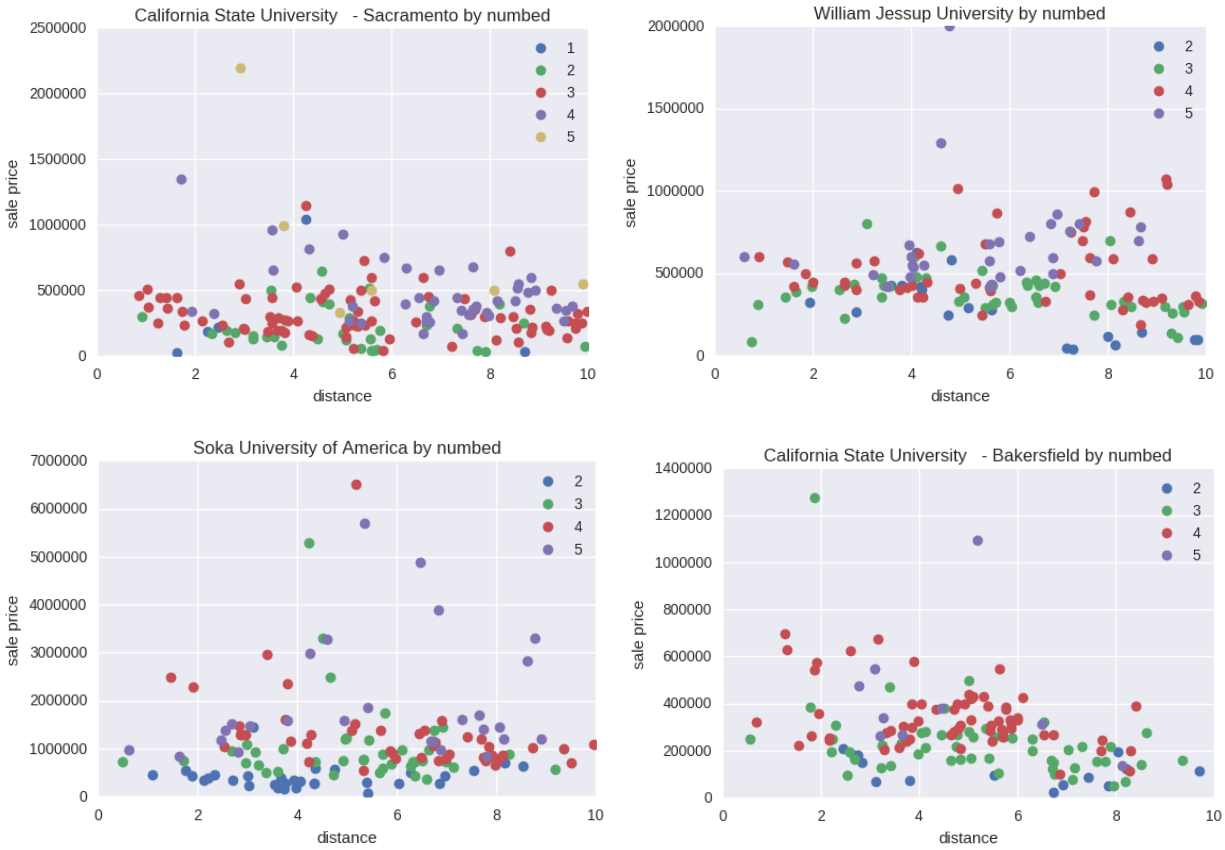


Figure 3. Plot sale price versus distance in different subgroups within each of the four universities' 10 miles vicinity. Each subgroup is plotted with different colors. In most subgroups, the trend is almost flat. The sale price's dependence on distance is weak. However, in some subgroups, the sale price shows a negative dependence on the distance.

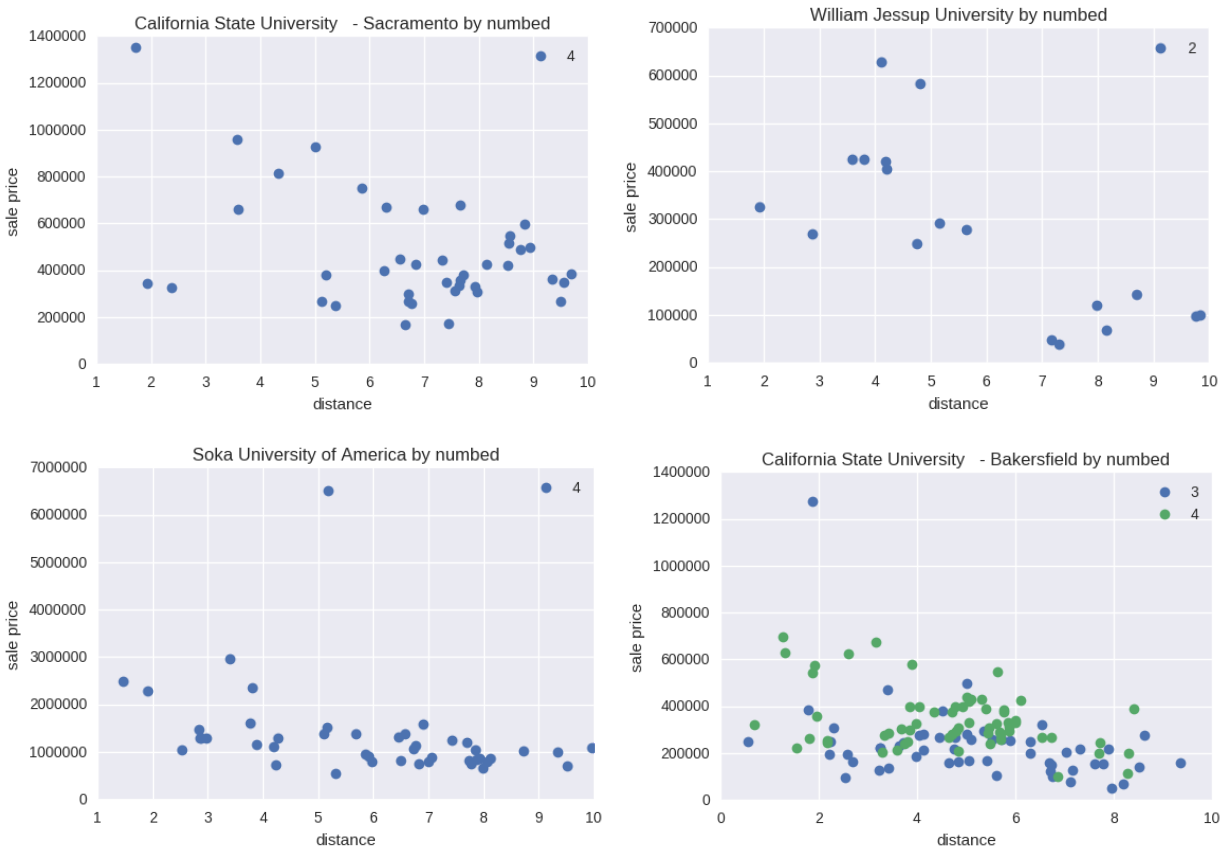


Figure 4. Plot sale price versus distance in highlighted subgroups within each of the four universities' 10 miles vicinity. Each subgroup is plotted with different colors. The sale price shows a negative dependence on the distance.

3.2 Number of Bathrooms

We can perform a similar analysis on the number of bathrooms. As before, we subgroup the homes by number of bathrooms and find the correlation between price and distance to the university. We summarize the correlation in Table 5. The field is filled as "correlation / p-value (number of homes)". The subgroups with large correlation and small p-value are highlighted. We also plot the incidences in Figure 5 and the highlighted incidences in Figure 6.

The result is consistent with the analysis on the number of bedrooms. Three and four bedroom houses usually come with two or three bathrooms. So we expect to see a better dependence on homes with two or three bathrooms.

Among the four universities, only WJU and CSUB have such small p values, and overall four groups. WJU with 4 full bathrooms exhibit a strong linear dependence between sale price and distance, which may be an interesting group to look at. Among the other three groups, CSUB with 3 full bathrooms has strongest correlation and smallest p value.

Num of full bathrooms	CSUS	WJU	SUA	CSUB
1	-0.27 / 0.08 (43)	-0.83 / 0.08 (5)	0.58 / 0.06 (11)	-0.09 / 0.68 (22)
2	-0.197 / 0.09 (77)	-0.35 / 0.003 (68)	0.21 / 0.14 (54)	-0.33 / 0.002 (87)
3	-0.26 / 0.13 (36)	0.10 / 0.50 (45)	-0.23 / 0.14 (42)	-0.64 / 0.0005 (26)
4	0.67 / 0.14 (6)	0.47 / 0.03 (21)	-0.41 / 0.05 (23)	(2)
5	(3)	0.71 / 0.12 (6)	0.10 / 0.75 (12)	
6	(1)	(1)	(3)	

Table 5. Summarization of correlation between home price and distance to the paired university in each university's 10 miles vicinity subgrouped by number of bedrooms. The subgroup with less than 5 homes are ignored in the table. The four universities are chosen because they have most homes within 10 miles. We are interested in subgroups with a large correlation (abs(correlation) close to 1) and small p-value (smaller than 0.05). The subgroups satisfying this condition are highlighted.

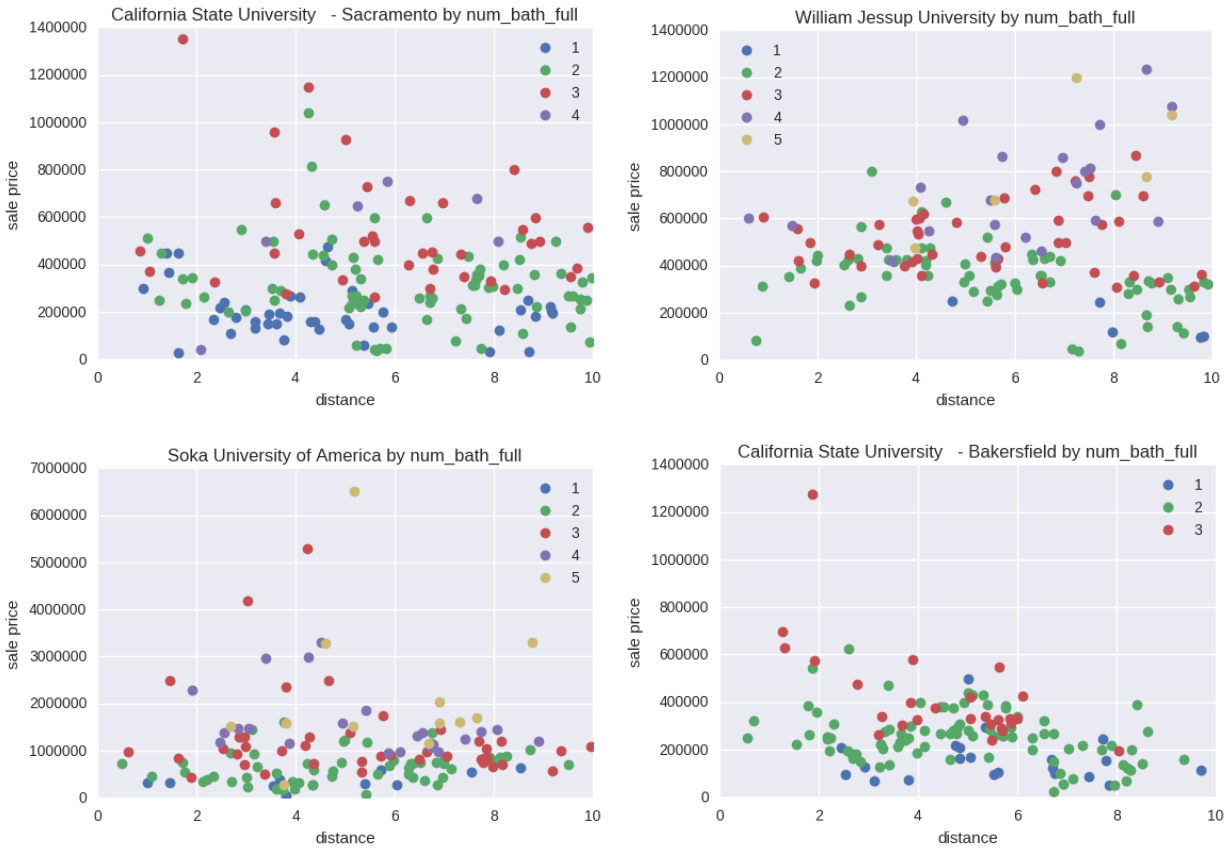


Figure 5. Plot sale price versus distance in different subgroups within each of the four universities' 10 miles vicinity. Each subgroup is plotted with different colors. In most subgroups, the trend is almost flat. The sale price's dependence on distance is weak. However, in some subgroups, the sale price shows a negative dependence on the distance.

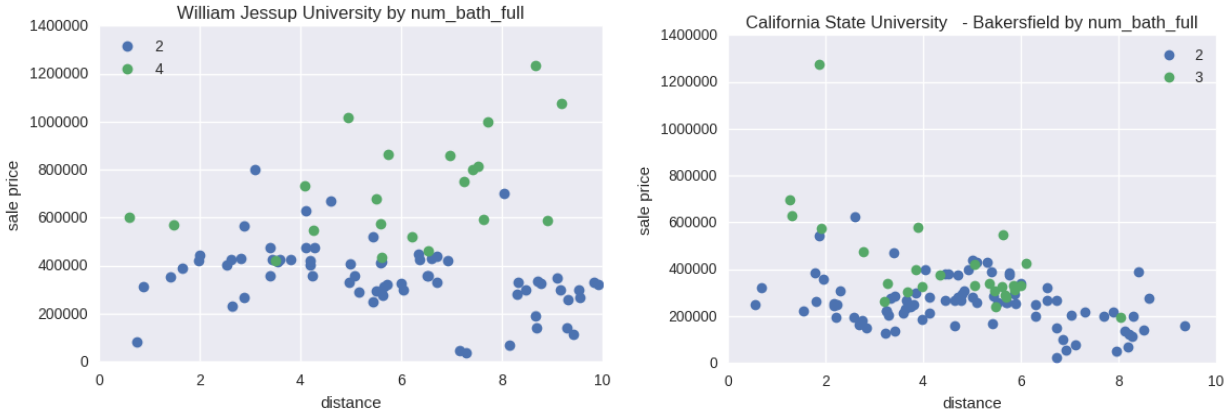


Figure 6. Plot sale price versus distance in highlighted subgroups within each of the four universities' 10 miles vicinity. Each subgroup is plotted with different colors.

3.3 All Homes With a Certain Number of Bedrooms/ Bathrooms

The previous analysis is performed in the local area of a certain university. In this section, we perform the analysis across all universities and try to pick out which university features affect the home price most.

Each time, I filtered out all homes with 3 bedrooms, 4 bedrooms or 3 bathrooms, group them by the nearest universities. I calculated the correlation, named distCorr, between the sale price and distance for all homes within 10 miles to each university. I continued to calculate the correlation between distCorr and some university features. I summarized the correlations and p values in Table 6, formatted as (correlation / two tailed p-value)

Corr / p value	acadStaff	students	underGrad	postGrad
3 bedrooms	-0.247 / 0.119	-0.078 / 0.592	-0.157 / 0.352	-0.121 / 0.496
4 bedrooms	-0.240 / 0.171	-0.109 / 0.528	0.120 / 0.527	0.121 / 0.547
3 full bathrooms	-0.241 / 0.177	-0.139 / 0.405	-0.010 / 0.960	-0.018 / 0.932

Table 6. The correlation between distCorr and some university features

The distCorr vs different university features were plotted below in Figure 7. Each dot represents an university.

When calculating the correlation on all universities, the universities with nan in university field have to be removed. So the correlation and plot could not represent the whole picture. I found 10 universities with the largest distCorr (close to 1) and smallest distCorr (close to -1) in each group. Table 7 is a summary of the mean value of the distCorr, reported as mean(feature value of the largest 10 distCorr universities) / mean(feature value of the smallest 10 distCorr universities)

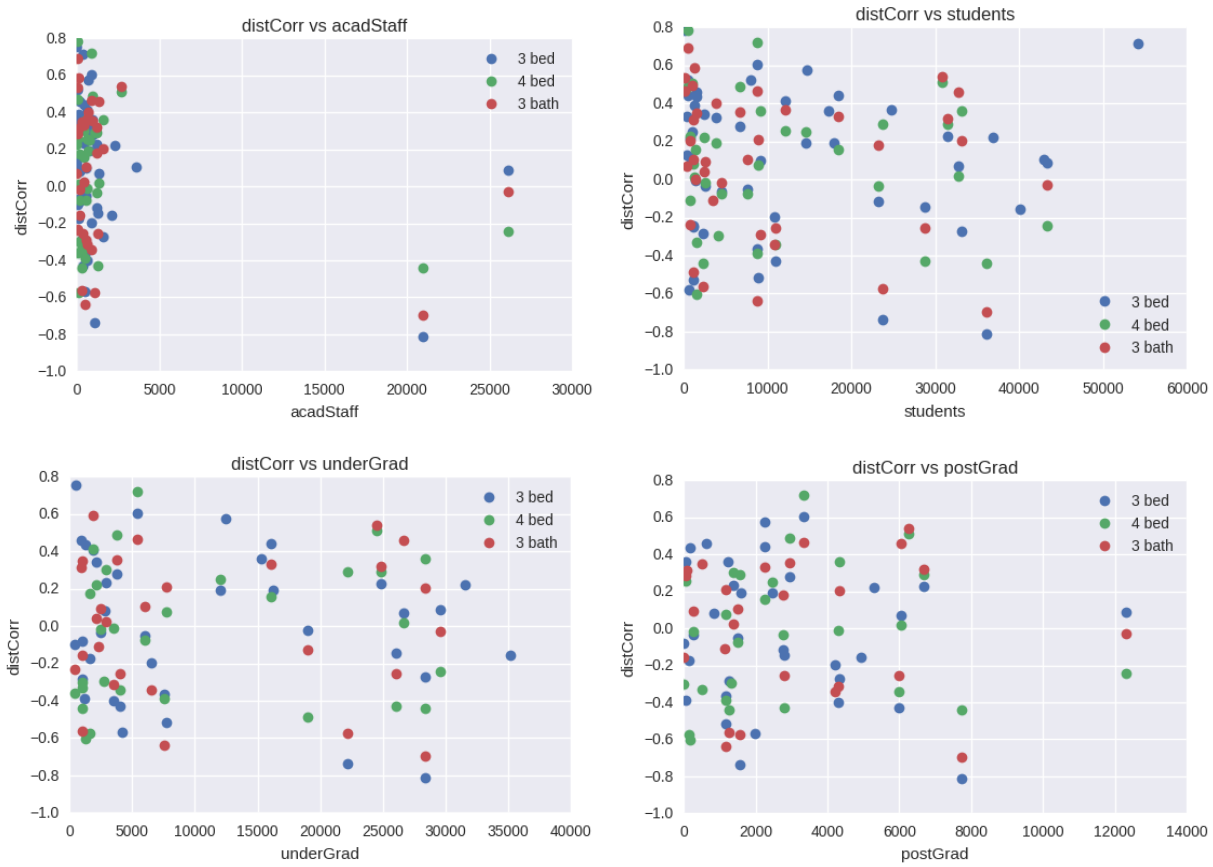


Figure 7. Plots of distCorr versus some of the university features. Each university is represented by a dot. Each bedroom/ bathroom subgroup is represented by a different color.

	3 bedrooms	4 bedrooms	3 bathrooms
acadStaff	373 / 7520	927 / 3870	751 / 2950
students	12531 / 14107	6885 / 15478	6076 / 14631
underGrad	4816 / 15628	11227 / 10648	10605 / 12388
postGrad	2069 / 3112	4188 / 2209	4802 / 3627
mean(top10 distCorr) / mean(bottom 10 distCorr)	0.624 / -0.623	0.570 / -0.515	0.589 / -0.441

Table 7. Statistics of university feature values and distCorr. In the first 4 rows, the field value is filled as mean(feature value of the largest 10 distCorr universities) / mean(feature value of the smallest 10 distCorr universities).

The larger the university size is, the more influence it has on the price of homes close to it. As a result, when the acadStaff and students numbers are large, the home price has a stronger dependence on the distance to the university, thus the correlation is more negative. However, if sorting by feature value, acadStaff for example, the dependence is not strong in the universities with most staff number.

But when the university is large, is the distance to the university the dominant factor determining the home price? We pick out the 9 universities with most academic staff and their distCorr in Table 8.

acadStaff	26139	20974	3600	2283	2096	1591	1358	1270	1215
distCorr	0.09	-0.81	0.10	0.22	-0.15	-0.27	0.07	-0.15	0.23

Table 8. The 9 universities with most academic staff and their distCorr. The distCorr value is not always negative. We conclude that the university is one of the important factor but not the most important one in terms of determining the home price.

Table 8 shows even though university size has an impact on the home price, it's not the most dominant one. When other factors present, the correlation between price and distance to university is not clear. As the data suggest, large university doesn't guarantee a strong correlation, however, strong correlation usually implies a large university.

Chapter 4
Time Series Analysis on Home price

In section 2.1 we find the 4 bedroom homes is the most relevant home type to universities. In this chapter, we focus on the historical data in this home type. The data can be downloaded from http://www.zillow.com/research/data/.

The data are collected every month since April 1996 and report the average home price in each zip code. For each zip code, I calculate the percent change on every April price, and find the mean and standard deviation of those numbers. Let's call them zip_mean and zip_std. Then, as before, I find the mean and standard deviation of zip_mean in large university zips and the same in no university zips.

large university zips:
mean(zip_mean) = 0.038
std(zip_mean) = 0.022

no university zips:
mean(zip_mean) = 0.023
std(zip_mean) = 0.038

Compared to the values in zips with no universities, the mean of percent change in zips with large universities is larger while the standard deviation of the percent change is smaller. We argue in the areas with large universities, the house price increases more rapidly and stably. While in the no university areas, the house price is subject to other factors and can vary more.

Appendix A

Saving Data Locally

To avoid server congestion, and most importantly, enable usage of programming libraries of self choice, we decide to save the data locally.

The data is originally saved as a MySQL table on the lab server. Naively, we use the command below to port out the data.

```
select *
from MySQLtable
where year(crawl_time) = 2016
into outfile '/tmp/file.csv'
fields terminated by ','
enclosed by '"'
lines terminated by '\n';
```

However, one of the table fields is "description". This field contains very complicated long strings, and some of the strings contains quotation marks, which cause conflict with the "enclosed" command. As a result, some lines in the csv file have more fields. We proposed four possible solutions.

1. Since we read the csv file with Pandas in Python, we use pd.read_csv(file, chunksize=1) to read one line a time and discard the lines with more fields. This is generally not a good approach because Pandas process data in a batch and construct a structured data structure. Reading line individually definitely kills this bonus Pandas provides.

2. Use single quote to enclose each field. The output file is enclosed by single quotation as expected, but each field is also converted into string.

3. Use Python file IO to filter the unwanted lines. We open the outputted file and read lines individually. We split the lines by comma and count the number of fields, and write the lines with a certain number of fields to a new file. However, the description usually contains "," and line.split(",") will split at those commas.Surprisingly, when there is no quote in the description, the Pandas parser doesn't make a mistake. I see no easy fix to make my parser as good as the Pandas one. It also didn't help to count the number of quote in a line,

because when a field was outputted as \N when it was original NULL. Note the \N was not surrounded by quote and there was no control of how many the fields would be.

4. Discard the record when outputting the file from MySQL. I hesitated to try this approach at the beginning because telling if description contains “\”” is expensive, and my dataset is large. But now it seems to be the only easy way to get the job done.

   ```
   select *
   from MySQLtable
   where year(crawl_time) > ‘2016-06-15’ and description not like ‘%”%’
   into outfile ‘/tmp/file.csv’
   fields terminated by ‘,’
   enclosed by ‘”‘
   lines terminated by ‘\n’;
   ```

This query gives me 183460 records and I can load them to Pandas with low_memory=False.