University of California Graduate Programs in Life Sciences, and the Cost of Rent Near Campus

[redacted]

Western Governors University

Table of Contents

A. Project Highlights

A1. Research Question:

This project was implemented to investigate whether or not doctoral graduate programs in life sciences are on a declining trend at the University of California, and to investigate if the cost of housing around a UC campus correlated to those application trends. Before any analysis was performed, I believed that at least some campuses would have declining trends in admission characteristics, and that they would be correlated with increased housing costs.

A2. Project Scope:

The scope of this project included:

- The collection and processing of data from the UC public database, and from Zillow's rental database.
- The analysis of admission characteristic data and rental data to produce relevant summary statistics and statistic visualizations
- Discussion of the delivered statistic analysis

Some considerations outside the scope of this project include non-rental housing costs such as home value or living expenses, as well as considerations of data outside the date range of the data (2015-2022). Also, the deliverables of this project show *correlations* between variables. It is important to remember that *causal* relationships are not being investigated by this analysis and while we can speculate on the cause of correlations, further investigation is outside the scope of this project.

A3. Solution Overview:

Tools:

Data was collected from public sources, such as the University of California Information Center, and from the Zillow Research web page. This data was downloaded as .CSV files and opened in Microsoft Excel for initial inspection and cleaning. This inistial cleaning removed any non-relevant data and identified where large portions of data was missing.

Next these .CSV files were loaded into data frames with Python through Jupyter Notebook for a more detailed incpection and cleaning. Here I used Python packages like Pandas to clean the data frames. I had to adjust the Zillow data so that it contained annual data values rather than monthly values, I removed Riverside form the costs data frames, and then used packages like matplot and scipy to analyze and visualize the data.

Methodologies:

To complete this project successfully, I followed an Agile methodology throughout. Data wrangling and data cleaning methods were used to collect and organize the data. Afterwards, exploratory and descriptive analytic methods were used to investigate the data and provide guidance in the design of the project. Finally, statistic analysis methods were used to test the data for significance and insight from which visualizations were created.

B. Project Execution

Project Plan:

The following plan submitted in Task 2 was implemented and completed successfully without any major deviations. The only changes were some minor additions to the visualizations created since they helped display a more robust understanding of the data as a whole. Specifically, I decied to create trend statistics and visualizations for the housing costs over time in each UC campus city – an addition to the original plan.

Task 2 plan for goals, objectives, and deliverables:

The goal of this project is to provide insight on the role of housing cost near UC campuses in the success of graduate programs and their students measured by their applications characteristics. I will specifically be investigating the life sciences doctoral graduate programs.

The primary objective is to investigate the existence of a correlation between life science graduate program admission characteristics (applications, admissions) and the cost of housing (specifically adjusted cost of rent) near UC campuses. To complete this objective, this project will first deliver a statistic analysis of trends in application characteristics for each campus, utilizing a series of Mann-Kendall tests to determine the significance of these trends. This analysis will include visualizations of each campus' trends as well as a table of Kendal Tau values and associated p-values.

The next objective is to investigate the correlation between the cost of rent near a campus, and that campus' previously determined program application trends. The deliverables required for this objective include tables containing the outcomes of a series of Pearson correlation calculations between rent costs and application numbers, as well as a visualization of the significant outcomes.

The project planning methodology:

An agile methodology was applied to the completion of this project. This project had defined goals to be completed by a single individual (myself), and was therefore best suited for a flexible goal-oriented methodology. The four foundations of an agile methodology include adaptivity, goal-oriented structure, an integrated approach, and learnable nature. Since this was a single-person project, adaptability was key to this project's success. As the data was investigated, different techniques and programmatic approaches had to be considered and implemented as their needs arose and became apparent. As stated previously, this project had clearly defined goals – the statistical analysis of defined data sets. The outcome of this project could be directly integrated into the decisions of relevant committees, such as financial committees on campus. This project was only the first step of a series of highly relevant questions concerning graduate programs and student finances. The outcomes of this project elucidated the path for learning more about these important interactions by revealing deeper, more nuanced questions concerning the role of student finances and academic success.

Project timeline and milestones:

Milestone or deliverable	Duration (hours or days)	Projected start date	Anticipated end date
Create Timeline and Project Plan	7 days	*10/11/2023*	*10/21/2023*
Gather and Clean Data	3 days	*10/21/2023*	*10/24/2023*
Foundational code and analysis of data	10 days	*10/24/2023*	*11/04/2023*
Visualization and Discussion of Statistical Analysis	2 days	*11/04/2023*	*11/06/2023*
Testing/Verification/Revision	1 day	*11/06/2023*	*11/07/2023*

This timeline was followed closely with no major deviations, aside from being ahead of schedule since building the foundational code was quicker than the allotted time for completion.

C. Data Collection Process

The data selection and collection did not differ significantly from the original plan. The UC data was filtered prior to download from the UC database.

The Zillow data was collected directly from the Zillow Research website as planned.

There were no changes to data governance or security, since I only used the public, anonymous data that was included in the original plan.

C.1 Advantages and Limitations of Data Set

One advantage of this dataset is that the UC data is wholly complete, since it is directly from the UC itself. I was able to find the dgraduate admissions data on the UC Information Center website, and filter it to show only doctorate programs, and only life sciences programs before downloading it. The Zillow dataset was also advantageous in the fact that it had data already separated by City. This was perfect for investigating rent costs around a UC campus, since each campus is separated by city also.

One disadvantage of this data was the fact that the Zillow data was supplied as a list of monthly values – while the UC data was supplied as annual values. This meant that I had to translate the Zillow data to annual values before I could make comparisons. Also this dataset went only back to 2015, so it (and therefore my analysis) was limited in that factor.

D. Data Extraction and Preparation

The data extraction for the UC data was simple, and the only unexpected obstacles in the extraction of the data was initial issues with encoding. I used Notepad to open the .CSV files and check their encoding, and found that some files were not UTF-8 as was expected. I then used Python to open the .CSV files with the proper encoding – this was more appropriate than trying to change the encoding since this was a more direct process.

The Zillow data contained data for every city in the US – while I only needed data for cities with UC campuses. As a preparation step in Excel, I removed all non-relevant cities, and also identified the fact that the city of Riverside had nearly zero data collected and would have to be dropped from the financial analysis. This seemed appropriate to do since any cities other than UC campus cities were irrelevant. Using Excel filters to show only cities of interest was an appropriate preparation method for selecting only relevant data for extraction. Also, I did not realize until investigating the extracted data, that the Zillow data contained monthly values – unlike the annual values provided by the UC. I had to group the Zillow data by year, and sum the values together to create annual values. I did this preparation step using Pandas in Python.

E. Data Analysis Process

E.1 Data Analysis Methods

Mann-Kendall Tests were used to analyze the trends in the application rates by year. Mann-Kendall Tests are perfectly appropriate for this since they are used to analyze trends in time series data. The Mann-Kendall Tests provided Kendall Tau values that like between -1 and 1 to represent the direction of a trend.

Pearson Correlation Tests were used to determine correlation between the application rates and rent costs near each campus. This was an appropriate method since Pearson Correlation tests are used to determine the correlation between two data series. The Pearson Correlation Tests provided correlation coefficients to represent the correlation between the rent costs and the application rates.

P-values were utilized to determine the significance of the trends and correlations investigated. Kendall Tau values and correlation coefficients were assigned p-values to determine their significance. P-values <= 0.05 were determined to be significant. This is an appropriate method for determining significance because this is a well established, broadly accepted, and simply implemented method.

E.2 Advantages and Limitations of Tools and Techniques

The tools used for this project included Microsoft Excel, Python, and Jupyter Notebooks. This was more than sufficient for the scope of this project since Excel has the advantage of allowing easy visual access to data for inspection and preparation, Python has a plethora of packages designed for data analysis, and Jupyter Notebooks allows for easy organization of code and code output through its use of visual blocks and markdown language. One limitation to this toolset is the limited ability to visualize data with Python alone. While Python packages like matplot

made visualizations that were sufficient for this project – deeper, more complex visualizations would be more readily constructed with something like R or Tableau.

This project included primarily descriptive analysis techniques which have the advantage of being very straightforward with easily translatable outcomes since it directly describes the data being investigated; however it is limited by the dataset itself and cannot make complex predictions.

E.3 Application of Analytical Methods

The following steps were used to apply the analytical methods listed in part E.1:

1. Data Collection:

Data was downloaded from the UC Information Center and from the Zillow Research web page.

2. Data Preparation:

CSV files were opened in Excel for initial visual inspection and removal of irrelevant data.

CSV files were read into a dataframe and initial inspection and exploration was performed using Python/Pandas exploration tools. Dataframes were inspected for null/missing values and any obvious errors.

The data frames were prepared for analysis by setting indexes and properly naming columns.

The Zillow data had to be prepared by grouping the columns by the year value in their name and then summing each yeat together to get annual values of rent costs.

3. Mann-Kendall Tests:

First I visualized the application rates for each campus over the years to get a basic look at the shape of the trends. Then I used the KendalTau function from the scipy package to generate kendall-tau values for the trends in application rates for each campus, along with p-values for

each kendall-Tau value. I loaded all of these results into a dataframe and created a column to track “trending” or “no trend” depending on the p-value. Rows with p-values <= 0.05 were considered to be trending. The result was a dataframe where each row had a UC campus, a kendall-Tau value, a p-value, and a “trending” value. This entire process was repeated for admission rates for each campus as well.

4. Pearson Correlation Tests:

For the Pearson correlations, I first visualized the mean annual cost of rent in each UC campus city, and visualized the trends in these costs over the years. Next, I used the pearsonr function from Python’s scipy package to perform the Pearson Correlation Tests on the cost data and the application data. I loaded the results into a dataframe and created a list of significant correlations by pulling out only correlations whose p-value was <= 0.05. I then visualized these significant correlations in a graph.

F. Data Analysis Results

F.1 Statistical Significance

Mann-Kendall Tests:

The null hypothesis for the mann-kendall tests was that there is no significant trend in the increase or decrease of application or admission rates between the years of 2013 and 2022 (the range of the dataset). These tests generated Kendall-Tau values between -1 and 1 that represent the direction of a trend, as well as p-values to determine the significance of those trends. We were able to reject the null hypothesis for 7/10 UC campuses. UCLA, UCSD, UCM, UCB, UCI, UCSB, and UCSF all had significant (and positive) kendall-tau values. This would suggest that these campuses all have positive trends in their annual application rates.

Pearson Correlation Tests:

The null hypothesis for the Pearson Correlation Tests is that there is no significant correlation between the cost of rent in a UC city, and the application rates for that city's campus. These tests generated correlation coefficients for each campus, along with an associated p-value to determine the significance of the correlation coefficient. It was determined that there is significant evidence to reject the null hypothesis in the case of UCB, UCLA, UCM, and UCSD. These campuses all had positive correlation coefficients and p-values <= 0.05, suggesting that these campus' application rates and their city's rent costs are significantly (and positively) correlated.

F.2 Practical Significance

The practical significance of the analysis performed is a more robust understanding of the role that housing costs play in the success of graduate programs in the UC system, measured by the application characteristics of the life science doctoral graduate programs across campuses. Interestingly, this information was originally intended to inform the decision-making at the administrative level of the UC, however; the positive correlations between housing costs and application rates may actually suggest that the popularity of the UCs are effecting the living cost of the city – information that would be useful to the city administration. It is important to note that the direction of cause and effect are *not* tested by these methods and the direction of the causal relationship is still unverified. That being said, this information could help city planners project living costs as the UC campuses grow, or inform the more popular campuses on the expected living costs in their home cities in relation to their growing popularity.

F.3 Overall Success

This project can be considered a complete success, since the following were all completed successfully:

- Visulizations were created to show the trends of application rates for each campus
- Visualizations were created to show the average annual cost of rent in each campus city and to show the trends of these costs over the years
- Statistic analysis was performed to determine the significance of trends in application rates over the years for each campus
- Statistic analysis was performed to determine the nature and significance of correlations between UC campus application rates and the cost of rent in their home cities

G. Conclusion

G.1 Summary of Conclusions

This project intended to investigate the nature of changes in application rates for doctoral graduate programs in life sciences at the University of California, and their correlations with the trends in rental costs near each campus. For this investigation, application data was collected directly from the UC information center, and cost of rent data was downloaded from the Zillow Research website.

First this project showed that the Los Angeles, San Diego, Merced, Berkeley, Irvine, Santa Barbara, and San Fransisco campuses all have positive trends in their application rates since 2013. These trends are statistically significant, with Los Angeles and San Diego having exceptionally strong trends towards increasing application rates. For a more robust understanding of the application trends, this project also looked at the same trends in admission rates. This suggested that only Los Angeles, Irvine, Riverside, and Santa Barbara had positive trends in the admission rates. The most interesting trend data is can be summarized in the following table:

Campus	**Application Trend**	**Admission Trend**
Berkeley	+	None
Merced	+	None
San Diego	+	None
San Fransisco	+	None
Riverside	None	+

This table suggests that Berkeley, Merced, San Diego, and San Fransisco are receiving more applications year-by-year, but are not accepting more students. Speculatively, this could be because of increased popularity, without the capacity to take more students. Surprisingly, Riverside is not receiving more applications, but is admitting more students. Speculating once more, this could be because the application rates have been well above a threshold of capacity, and recent additions to the Riverside campus have increased the capacity for new students – allowing the campus to admit more students since 2013.

The second part of this project intended to investigate correlations between the application rates, and the cost of rent in UC campus cities. The campuses with statistically significant correlations between their applications rates and the cost of rent in the campus' home cities were Merced, Los Angeles, San Diego, and Berkeley. Maybe the most surprising outcome of this investigation is the fact that these correlations are strongly positive correlations. This suggests that as rent increases, so too does the application rates. This may be less surprising when stated in the reverse; that is, as a UC campus begins to receive more applications, the cost of rent also increases in that campus' city. It is important to remember that the correlations do not explain the causal relationship between data, and it is likely many other factors that drive the popularity of a campus and the cost of rent together.

G.2 Effective Storytelling

For all visualizations, Python was used in Jupyter notebooks. Specifically, I used matplot to create plots from the analyzed data. These tools are sufficient for the visualizations required in this project, and are readily deployable within Jupyter notebooks, within the same file. While other tools may be capable of more dynamic or complex visualizations, the availability of these tools and their sufficient nature made them most appropriate for all of the following visualizations. Each visualization has been included and are described in detail below.

Figure 1:

The first visualization is a series of trend lines for the application rates of campuses between 2013 and 2022. This helps us understand the popularity of a campus measured by the number of applications in a given year, and how that popularity has changed over time.

Figure 2:

The second visualization is a line graph of admission rates for each campus. This was not split up into subplots since the scale of the admission rates is small enough that the lines for each campus do not get too crowded. This visualizations helps us understand the capacity for admission each campus has, and suggests that while different campuses have different capacities for admission – the annual trends in admission are much more static.

Figure 3:

This visualization is a simple bar chart showing the average annual cost of rent in each UC campus' home city. This chart exists entirely to help give us foundational context for which campuses are in the most expensive cities.

Figure 4 and 5:

The fourth and fifth visualizations are line graphs that show the annual cost of rent in each city over time. First as a stacked line graph, then as a series of subplots so we can inspect each city individually. These visualizations help us understand how the cost of rent has changed over the years around each campus.

Figure 6:

The last visualization is a bar chart showing the correlation coefficients for campuses with statistically significant correlations between the cost of rent and their application rates. Had there been any negative correlations, this would have been a waterfall graph, but since all correlations were positive, a bar chart seemed appropriate. This helps us understand the strength of the correlation for each campus, as well as provides us with a list of which campuses have significant correlations.

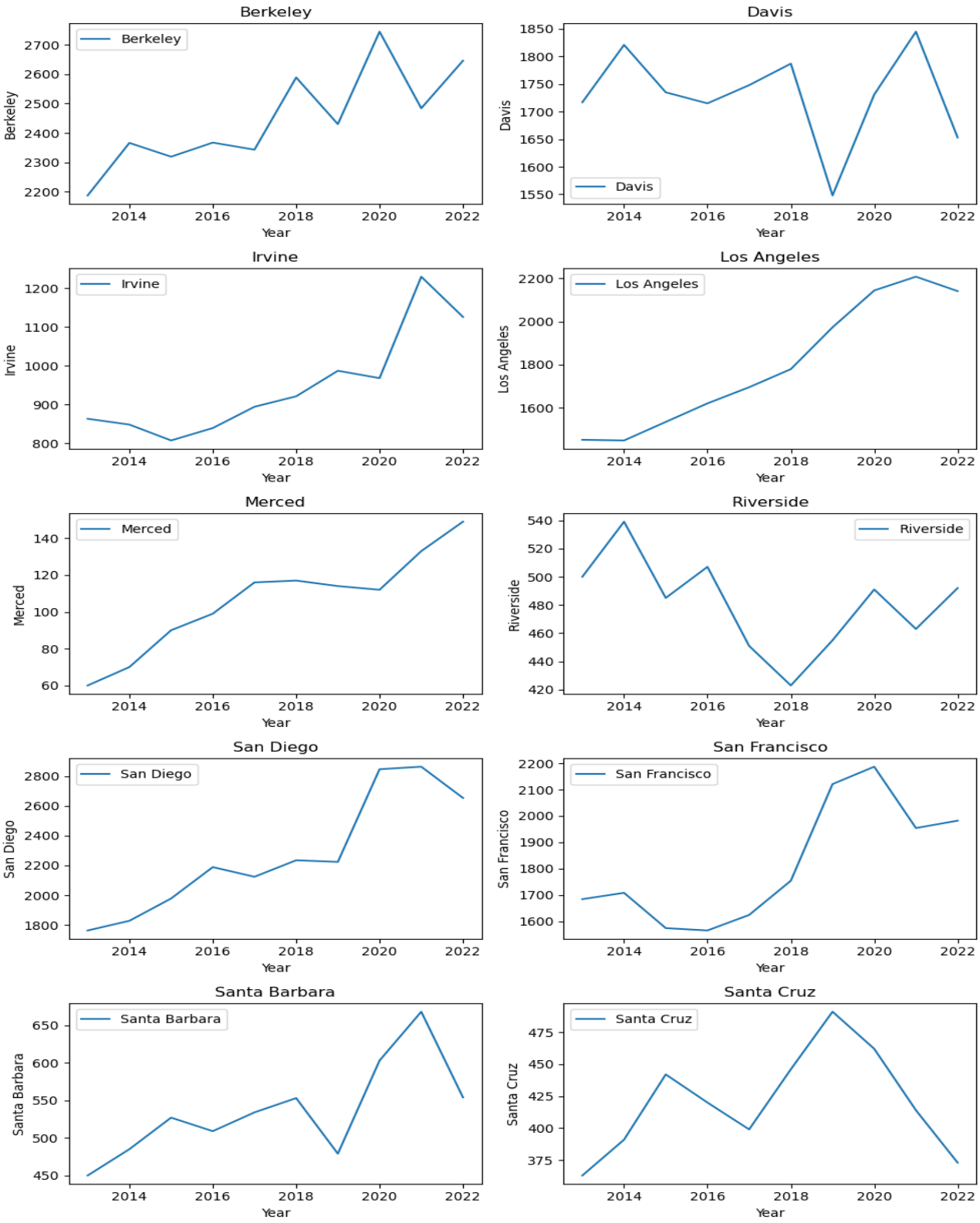

Figure 1.
Trend lines for application rates at each UC campus

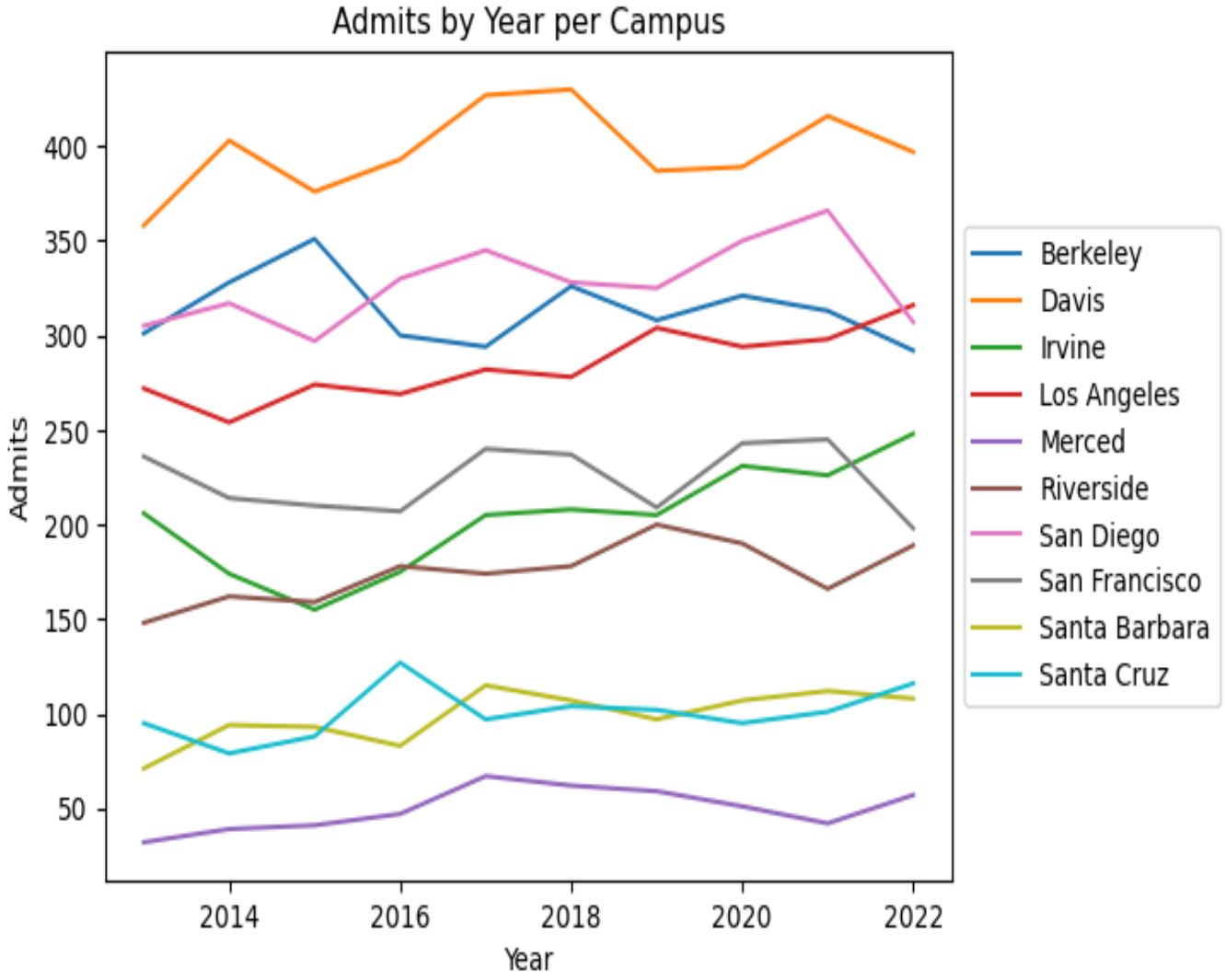

Figure 2.
Trend lines for admission rates at each UC campus

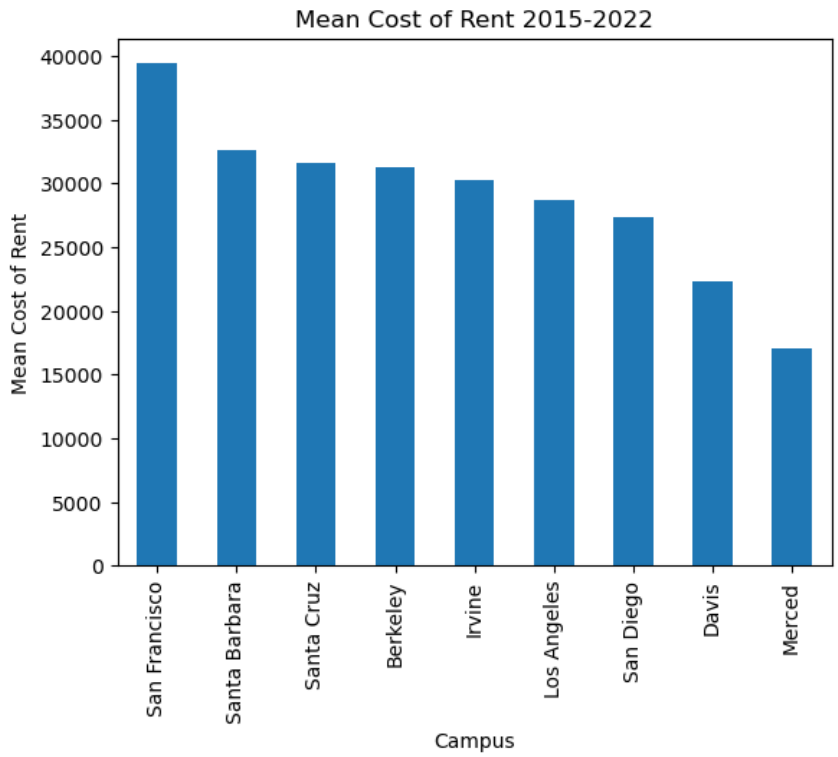

Figure 3.
Average annual cost of rent in UC campus cities

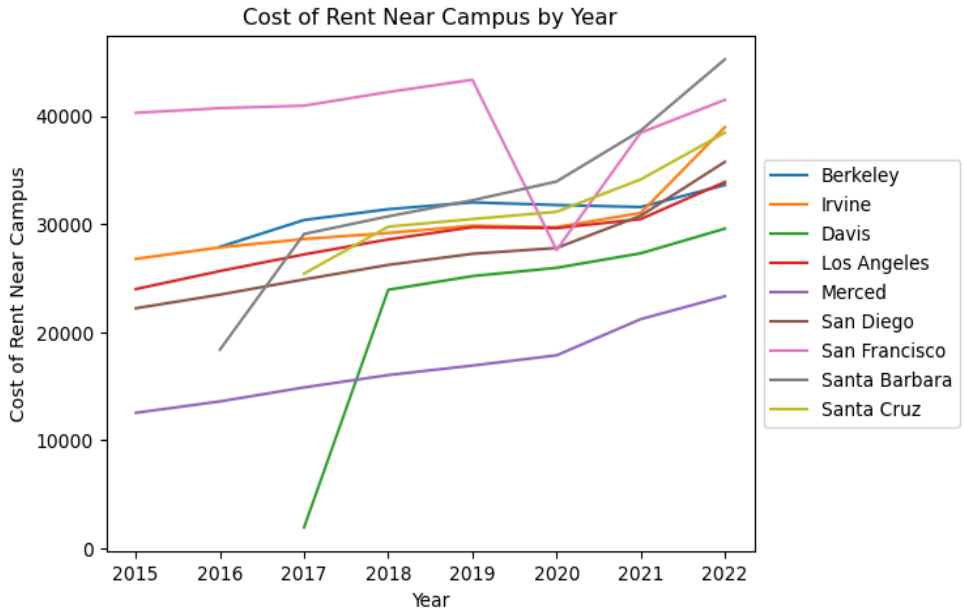

Figure 4.
Annual cost of rent in UC campus cities by year

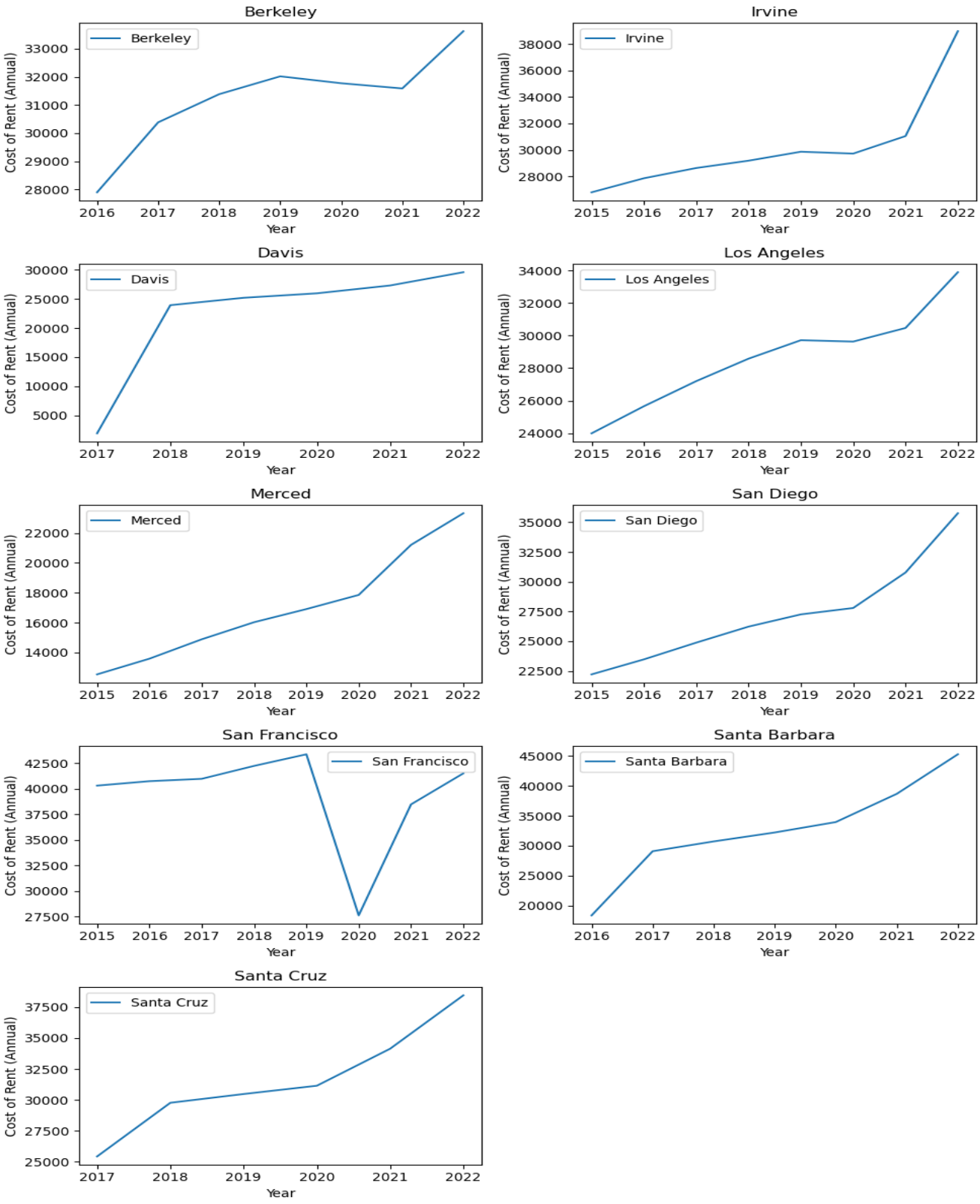

Figure 5.
Annual cost of rent in UC cities by year (subplots)

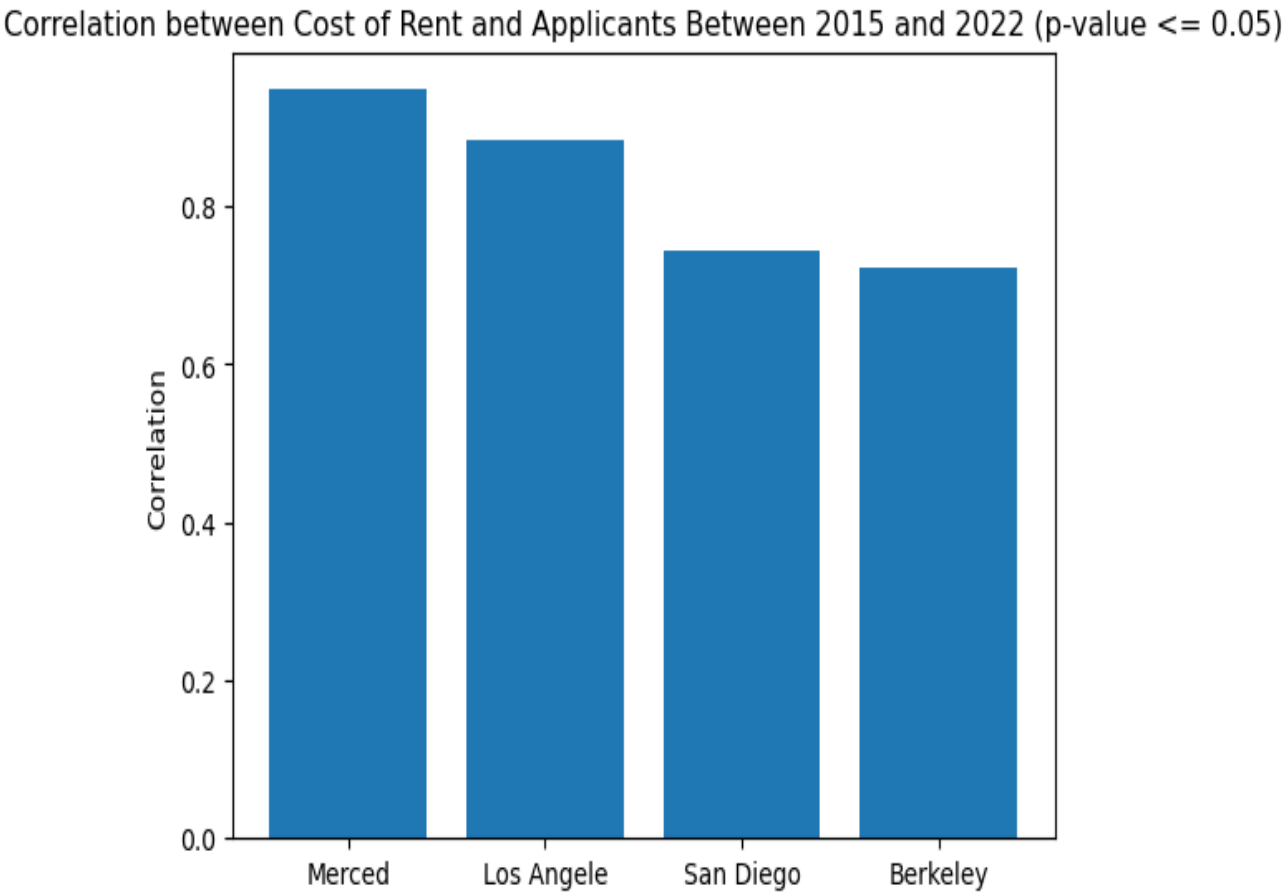

Figure 6.
Statistically significant correlations between cost of rent and application rates (2015-2022)

G.3 Recommended Courses of Action

Recommendation 1:

The first research question asks if there are significant trends in the application characteristics of each campus. From this project's findings, we can see that some campuses have positive trends in their applications, and no trend in their admissions. This suggests that these campuses are "at capacity" for their incoming student admissions. I would recommend expanding the life sciences graduate programs for these campuses, since they are deeply over-saturated with applications.

Recommendation 2:

The second research question asked if there are correltations between the application rates and the cost of rent in a campus' city. This project's finding show that several campuses do have statistically significant correlations between their application rates and the cost of rent – and interestingly, all of these correlations are positive. This could be evidence that as a campus becomes more popular, so to does the city in which it resides – or visa versa. I would suggest that campuses be aware of this correlation and take this information into considerations when projecting housing costs for students in coming years. If a campus is expecting to become a more desirable academic option or if a campus expands their campus and expects to become more popular for applicants, they should account for changes in housing costs for students when budgeting student funds.

H Panopto Presentation

Project Summary Video:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id[redacted]

Code Sumary Video:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=[redacted]

References

No sources were cited.