Mapping the Dynamics of the Davis Rental Market: A Web Scraping and Sentiment Analysis Approach

Allen Li, Shang Chen

October 4, 2025

Abstract

This project harnesses web scraping and natural language processing techniques to explore the dynamics of the rental market in Davis and its neighboring regions. We combine listings (Craigslist, Apartments.com, Zillow) with contextual data (NeighborhoodScout, BestPlaces) and Google reviews to analyze rents, safety, and livability. Our findings reveal that while Davis commands the highest rental prices driven by strong market demand, the correlation between rental price and reviews is weak. This indicates service quality is only weakly priced in this market. Additionally, challenges related to duplicated entries and availability in data quality across platforms are discussed, with recommendations for refining future analyses. Overall, this study offers valuable insights for students seeking affordable housing, property owners aiming to make improvements and investors looking for opportunities in the local real-estate market.

Keywords: Web Scraping, Natural Language Processing, Rental Market Analysis, Sentiment Analysis, Data Integration, Online Reviews, Housing Prices, Davis Rental Market.

1 Introduction

In Fall 2024 UC Davis enrolled 9,443 new undergraduates, in a city of about 65,000 residents, intensifying rental demand locally and in nearby towns and creating business opportunities for property owners and investors.

This report uses web-scraped data and visualization to characterize the local rental market. Our analysis is structured around three interconnected sub-themes:

1. **Rental Pricing Dynamics:** How do regional differences and distance to campus or downtown areas drive rental prices?
2. **Safety and Livability:** In what ways do region safety levels and living costs affect rental decisions?
3. **Online Reviews:** Which concerns in reviews are associated with willingness to pay above-average rents?

By integrating data from multiple sources, this report not only demonstrates effective web scraping and data cleaning skills but also offers a comprehensive look at the factors shaping local rental markets.

2 Data Source

2.1 Housing Information: Craigslist, Apartments, Zillow

We scraped listings from Craigslist, Apartments.com, and Zillow. Their posts often include address, price, number of rooms, photos and amenities.

To focus on our areas of interest, we limited our scraping to the Davis, Dixon, Woodland and West Sacramento regions (Figure 1). These are four common rental areas for UC Davis students, all within a 15-minute drive to campus under normal traffic conditions.

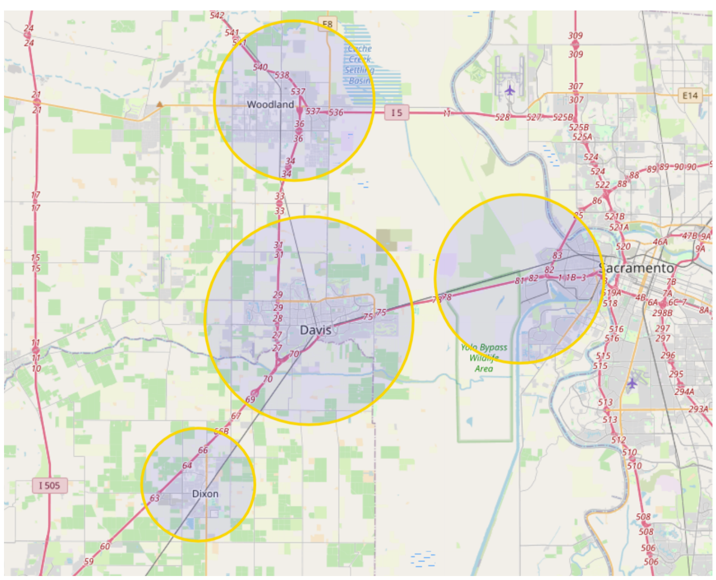

Figure 1: Four interested regions

Craigslist: Using `GET` requests with headers, we scraped rental listings from Craigslist. The extracted data was then parsed, extracted via `Xpath`. The captured key information such as apartment name (if available), address, price, number of bedrooms and bathrooms, area, and location were stored in

`DataFrames`.

Apartments: `GET` requests were also sent to retrieve data. Instead of `Xpath`, we applied BeautifulSoup here to parse and extract data. A `for-loop` was implemented to scrap through all available pages and collect data efficiently. Notably, since apartments will automatically find nearby cities, we filtered the data by retaining only apartments with region in the address. Similar information was stored.

Zillow: We used `PUT` requests with different payloads, gathering four `json` files. We noticed that the data in these files fell into two groups: One containing a field called `['hdpData']` , which held key information, one without the field. After investigation, we found that if a listing included multiple room types, then the `['hdpData']` field would present in its structure. To handle this, we applied an `if` condition to separate the two groups. For listings with multiple room types, we were limited to scrap only the price and minimum number of rooms. The data was properly converted and stored into `DataFrames`, extracting the relevant information.

2.1.1 Data Quality Assessment

We collected 353 listings from Craigslist. However, 220 of them are duplicated, identified by having the same address, price, number of rooms, and area. Listings from Zillow and Apartments tend to have a better quality, with 169 listings and 99 listings, respectively. We chose to use Zillow data to conduct our analysis.

Among the 169 listings, 107 of them are from Davis, 35 are from West Sacramento, 22 are in Woodland and only 5 of them are in Dixon. This reflects the higher concentration of rental listings in Davis, which is in high-demanding, while other towns that are further away from school have fewer listings.

2.2 Living Costs and Safety Data: Neighborhoodscout and Bestplaces

While looking for rentals, people commonly care about the living expense and safety levels. To understand them better, we collected data from Neighborhoodscout, a data platform specializing in providing detailed U.S. neighborhood analytics, and Bestplaces, a data platform focusing on providing analytics in overall livability, for Davis, West Sacramento, Woodland, and Dixon. BeautifulSoup was used to parse and extract key information.

2.3 Google Reviews: Data Collections

Google is currently the most widely used platform for all types of searches, including reviews of housing. Consumers rely on Google reviews to make their final decisions. According to SOCi research (2023), businesses that rank highly in Google local searches gain a significant competitive edge.

We used Google Maps Platform - Place API to retrieve reviews for apartments listed in our Zillow `DataFrames`. Google Maps uses `place_id` to differentiate buildings and some Zillow listings (e.g., house) may not have corresponding reviews, we first applied an `if` condition to fetch the `place_id` for apartments only:

- If the Zillow listing includes the apartment name, the request sent to GoogleAPIs would use `query: name` along with its `latitude-longitude` pair.
- If the Zillow listing does not include the apartment name, we would make a request to fetch the nearest apartment based on its `latitude-longitude`. We then compared the distance between the listed building and the nearest identified apartment. If the distance falls within 0.1 mile, we assume they referred to the same apartment.

Due to Google's regulations and the presence of invalid reviews (e.g., overly short or blank), we limited our search to the 5 most relevant reviews. Requests were made based on the valid `place_id`, and we collected both overall ratings and the top 5 relevant reviews.

Across regions, we collected 320 reviews from 77 apartments. Out of the 77 apartments, 50 received the maximum of 5 reviews, while 7 had only 1 review, such as 301 D Apartments only got one review in the past years. This highlights the inconsistency in review availabilities.

3 Data Analysis and Findings

3.1 Rental Price Analysis

To analyze rental prices, we calculated the unit price for each listing by dividing the total rent by the number of bedrooms. For studios (where bedrooms are listed as 0), the unit price is considered equal to the total rent. The standardized data allows comparisons across different apartments. Additionally, we excluded listings with a unit price below $600, as such prices are highly unrealistic for the local market and likely represent errors or misleading posts.

Examining the filtered data, Davis and Woodland stand out with the highest average unit price at $1,475 and $1,471, reflecting strong demand from students. West Sacramento follows with an average of $1,415. Dixon remains the most budget-friendly choice at $1,190.

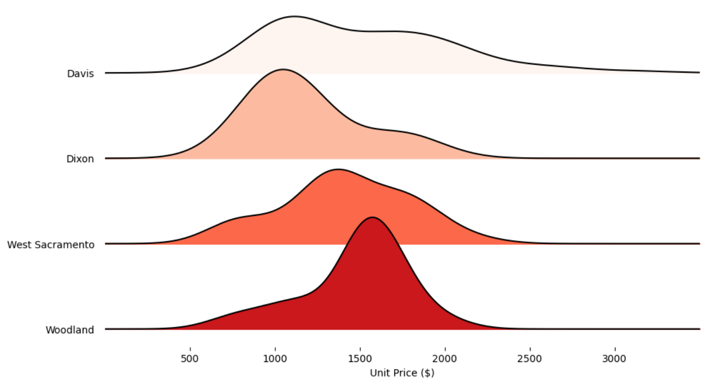

Figure 2: Distribution of Unit Prices Across Regions

The ridgeline plot (Figure 2) shows the distribution of unit prices across four regions: Davis, Woodland, West Sacramento, and Dixon. Davis and West Sacramento exhibit the highest rental price variation, with long tails indicating a broad range of prices. Woodland has relatively stable prices while it does include a few extreme values. Dixon, despite having the lowest median rent, also shows the least variability. However, the limited sample size in Dixon (only 7 listings) may influence this observation.

Building on this, the rental price distribution in Davis is further explored through a map visualization via Folium (Figure 3). This map demonstrates rental unit prices per bedroom across different locations in Davis. Each red circle represents a rental unit, with its color intensity indicating the price—darker red for higher rental prices. The clustering of dark red circles in central Davis suggests that rental prices are generally higher in that area, particularly near the university and downtown. Meanwhile, more dispersed lighter markers on the periphery indicate that units more than 15 minutes by bike are more affordable.

With the same visualization method, a similar map (Figure 4) of Woodland reveals rental price patterns in that area. Rental units are concentrated around Main Street and major roads. Notably,

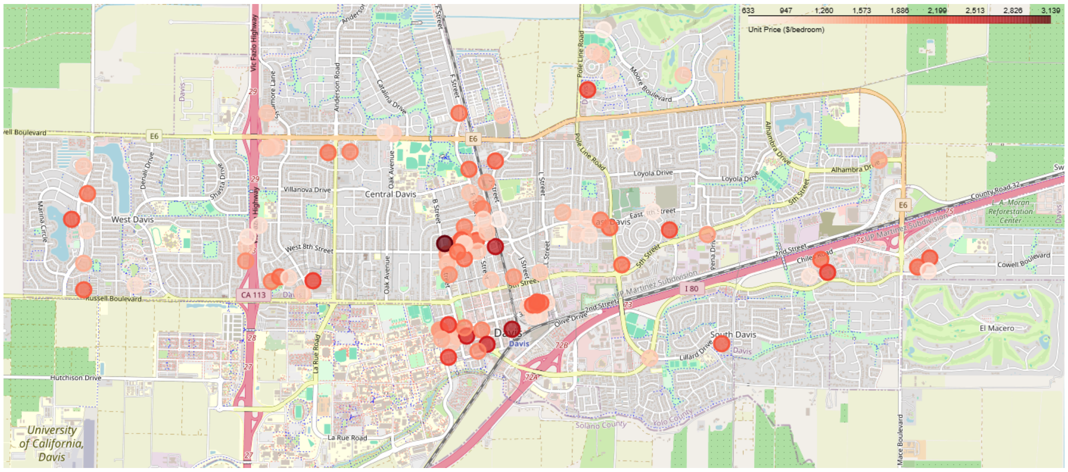

Figure 3: Davis rentals

YoloBus 42A & 42B run along East Street and Main Street, providing a direct link to West & Old Sacramento and Davis. For students with vehicles, commuting via CA 113 is another option, with Exit 36 and Exit 37. The multiple transportations make Woodland a practical and cost-effective alternative for student housing.

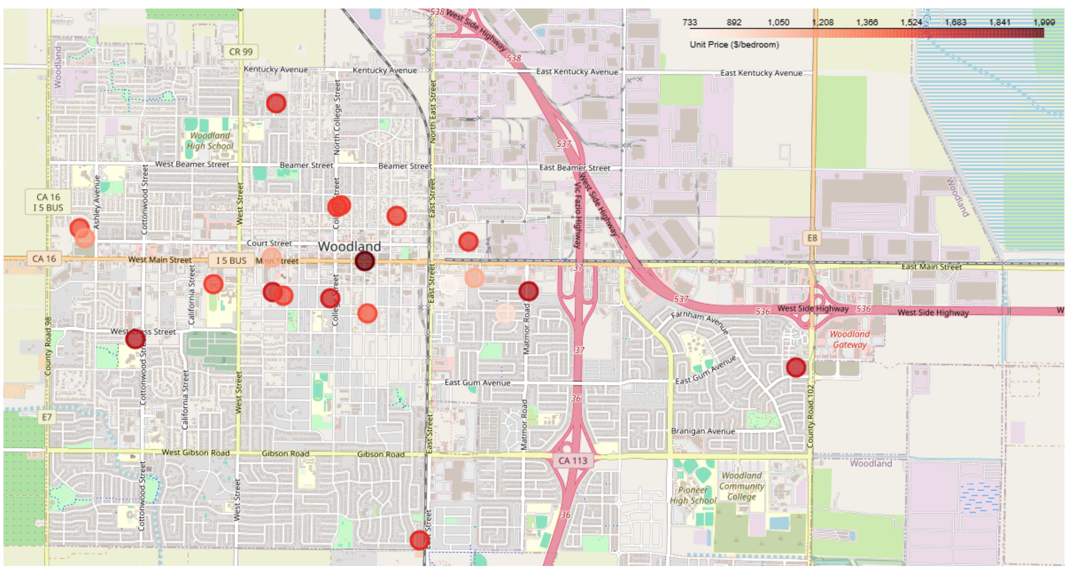

Figure 4: Woodland rentals

3.2 Safety and Livability

3.2.1 Crime Data Overview

We investigated the crime data in these regions, with the common sense that people would take neighborhood safety level into account while relocating. As shown in the table, all four regions are relatively safer than a major city like San Francisco. Among them, Woodland has the highest safety rating, making it the safest option. Davis also ranks well, while West Sacramento and Dixon have slightly higher crime rates.

The *Total Crime Index* represents how much safer the region is compared to the national average, with 100 being the safest score, covering both violent crimes and property crimes.

Table 1: Crime Statistics by Region

Region	Total Crime Index	Violent Victim Probability
Woodland	14	1 in 255
Davis	4	1 in 427
West Sacramento	31	1 in 465
Dixon	24	1 in 386
San Francisco	1	1 in 142

The *Victim Probability* represents the chance of a resident becoming a crime victim in violent crimes.

3.2.2 Living Cost Overview

We also studied the data from Bestplaces. The table shows that Woodland, Davis and Dixon have the highest utility index (108.6), suggesting that energy costs for heating/cooling may be higher there. It is probably due to their less energy-efficient buildings. In terms of grocery costs, Davis (110.6) is the most expensive among the four regions, which aligns with its nature as a college town. In contrast, West Sacramento (105.3) has the lowest grocery index, making it more affordable for daily essentials.

Table 2: Living Costs by Region

Region	Grocery Index	Utility Index
Woodland	105.6	106.6
Davis	110.6	108.6
West Sacramento	105.3	101.4
Dixon	108.1	108.6
San Francisco	116.6	97.5

The *Grocery Index* represents the cost of groceries compared to other regions in U.S. Higher than 100 is more expensive than the national average.
The *Utility Index* represents the average cost of heating or cooling a typical residence for the area, including electricity, natural gas and other fuels.

3.3 Analysis on Online Reviews

3.3.1 Word Cloud

Understanding what renters prioritize is the key to identifying trends in the Davis rental market. To gain some insights into the most frequently discussed issues in apartment reviews, we further conducted text analysis by word cloud.

To refine the analysis, we filtered out two groups of stopwords:

- **Sentiment stopwords** (e.g., good, bad, excellent, poor) to focus on the main themes of the text.
- **Custom stopwords** (e.g., apartment, unit, rent, lease) to remove common housing-related terms that are not meaningful to our analysis.

All text was converted to lowercase and punctuation was removed to standardize. We then used `NLTK`'s word tokenizer to split the text. Since only nouns and adjectives were interested in, Part-Of-Speech(POS) tagging was applied to classify and select the `NN` (noun) words or `JJ` (adjective) words. These selected words were aggregated into a single text string and processed by WordCloud library. The parameter `collocations` is set to be `False` to ensure that only single word is displayed.

To compare review trends across different regions, we generated word clouds separately for Davis and West Sacramento:

Figure 5: Davis Word Cloud

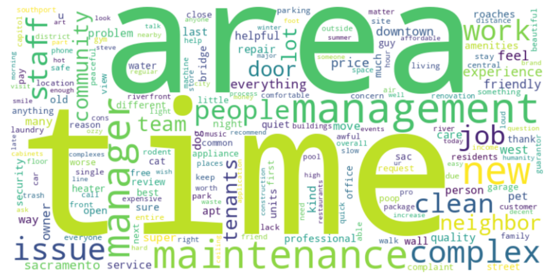

Figure 6: West Sacramento Word Cloud

In both locations, words *management*, *maintenance*, *staff* and *time* appear frequently in reviews. For Davis (Figure 5), words such as *complex*, *time* and *problem* stand out, implying people have concerns about maintenance efficiency and manager's response. In contrast, West Sacramento's word cloud (Figure 6) unique highlighting words are *area*, *new* and *community*, which may indicate that residents value location, neighborhood and new-built housing more.

3.3.2 Sentiment Analysis

In order to delve into the reviews, we planned to conduct a sentiment analysis from which we can understand the renters' real attitude compared to the rating they made on Google Maps.

We pipelined the process using a pre-trained roBERTa model from Hugging Face which is originally trained on Twitter data. However, due to the limitation of 512 tokens of the model, we could only deal with reviews shorter than the threshold. To fully utilize the data we scraped, we assigned a summarization task with the help of Facebook-based BART model and condense the long review to 400-500 tokens.

The result contains two attributes - `label` (positive, negative or neutral) and its corresponding `score` which represents the highest possibility among the three labels. With the aim of rescaling the score to the normal 0 - 5 rating format, we applied the following formula:

$$Score_{transformed} = \begin{cases} Score, \ label = positive \\ -Score, \ label = negative \\ 0, \ label = neutral \end{cases}$$

$$Score_{rescaled} = \frac{5}{2}(Score_{transformed} + 1)$$

From the review data and the sentiment analysis above, we already obtained the `rating` and `sentiment score` of each review and the `overall_rating` of one apartment. Then we could draw the differences between the three metrics via Folium. For instance, the left bar plot (Figure 7) showed that though Alvarado Sunset Apartments has a moderate overall rating, the most relevant reviews reflect a lower rating and the sentiment score is even lower, approaching 0, which indicates the apartment left a terrible impression on renters. The right bar plot demonstrated an opposite situation. Greenbriar Apartments shows an increasing trend of ratings and the highest score of sentiment analysis means renters' real attitudes are more positive than the ratings alone reflect.

In summary, incorporating sentiment analysis alongside traditional rating metrics provides a more accurate and comprehensive understanding of renters' true feelings upon an apartment. While overall ratings offer a general snapshot, sentiment scores can uncover the underlying emotions expressed in the reviews, revealing nuances that the ratings might overlook. This integrated approach ultimately delivers a more reliable evaluation of renters' satisfaction and the apartment's performance.

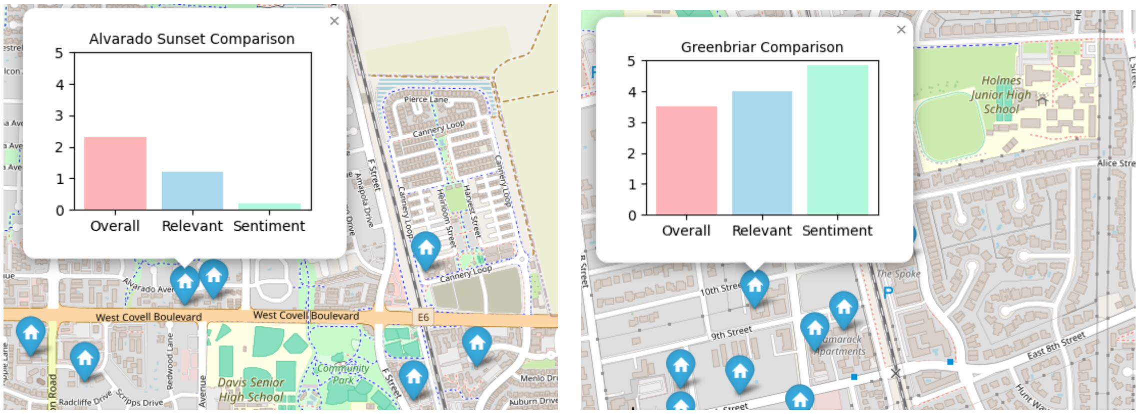

Figure 7: Review Score Comparisons

3.4 Rental price vs Sentiment scores

With the rental prices and sentiment scores, we further studied the relationship between them. Figure 8 and 9 demonstrate bubble plots with rental price as the x-axis and average sentiment scores as the y-axis, in West Sacramento and Davis, respectively. As the bubble grows larger and lighter in color, the average sentiment score increases, suggesting that residents had a good living experience. Additionally, from left to right, the unit rental price goes up.

Figure 8: Sentiment score vs. Rental Price (West Sacramento)

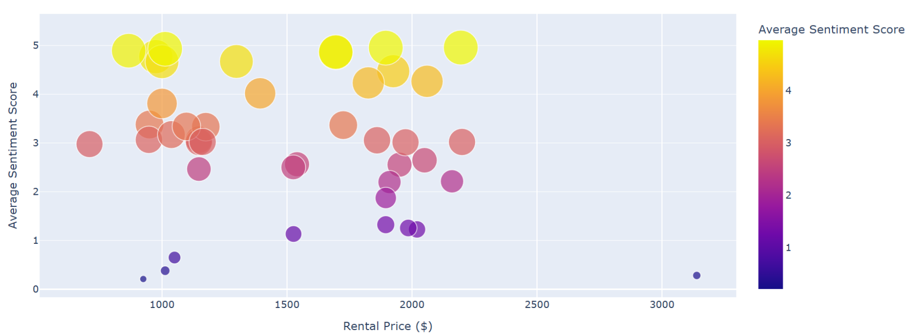

Figure 9: Sentiment score vs. Rental Price (Davis)

A common hypothesis is that higher-priced apartments yield higher satisfaction. Figure 8 shows a mild clustering consistent with this pattern. The bubble plot for West Sacramento. There is a clustering in the top right quadrant, where apartments with higher rental prices generally have higher sentiment scores. Many of them are more expensive apartments, above $1700, compared to West Sacramento's average rent, while they tend to have sentiment scores above 4.

However, this is not the case in Davis (Figure 9): there is no pattern between the size/color of bubbles and their distribution. Apartments with sentiment scores in the mid-range make up as majority, no matter what prices they are offering. We observed plenty of small bubbles, sentiment scores range from 1 to 3, falling on the right side of the plot, indicating that sentiment scores do not affect rental prices in Davis. This is consistent with capacity-constrained, student-driven demand in Davis.

In short, sentiment scores do not influence the rental market as much as one might expect. This is problematic because it implies that apartments would have little incentive to improve their services based on reviews: owners and agents can continue charging high rents regardless of the living experience they provide. Meanwhile, renters are forced to accept unsatisfactory living conditions at normal prices.

4 Caveats Discussion

In this report, as plenty of web scraping and natural language processing skills have been applied, we retain some unsolved challenges and concerns:

4.1 Data Quality and Methodological Limitations

The quality of online data varied across different websites. Because Craigslist is user-generated and open-posting, duplicates are common. Many owners or agents intentionally make multiple posts to maintain the visibility. We spot significant amount of duplicates listing in Craigslist data set, taking up almost 70% of the listings, which damage the data quality of Craigslist.

Zillow's outer display pages are not standardized; without API access we scraped only what is visible on the display page. However, not all apartments fill out the name of the apartments(See Figure 10).

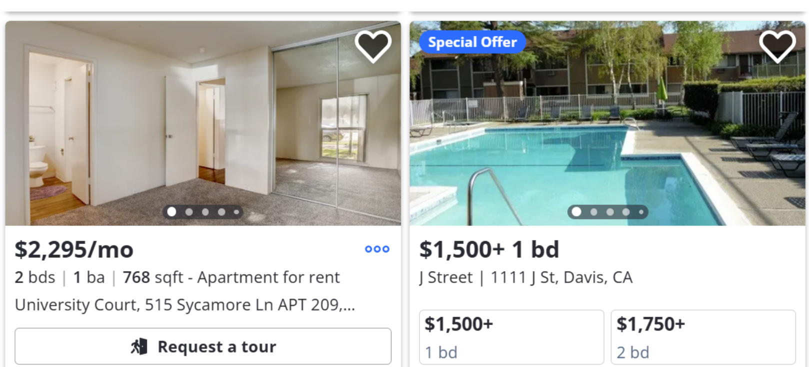

Figure 10: Screenshot from Zillow.com (taken on 2025-03-13 12:12pm PST)

Moreover, during our analysis of Google Reviews, we encountered some issues: One is caused by user-claimed spots on Google Maps. For example, a spot simply marked as *Apartments* appeared with one review containing blank text and three photos. The another one is caused by unknown reason: With the name of the apartment *Wake Forest*, Google Reviews matched Wake Forest University, a private university in NC, for us. Based on our current methodology, these ambiguous spots cannot be filtered out effectively.

4.2 Future Analysis Directions

For future analysis, more advanced skills can be used to catch duplicate and unclear listings. Integrating additional data sources or documented API access would help standardize the data, while crosscheck with metadata can improve listing verification as well.

5 Conclusion

This project explored key factors influencing rental prices and resident's choices around Davis. Our findings highlight that Davis consistently maintains the highest rental prices due to strong demand, while Woodland, West Sacramento, and Dixon offer more affordable alternatives. Crime data suggests that Woodland is the safest option, while West Sacramento and Dixon have slightly higher crime rates, which may contribute to their lower rental prices. However, living costs (e.g., groceries, utilities) appear to have minimal influence on rental decisions.

A further study into word clouds and sentiment analysis reveals that service quality is a significant concern for residents. Contrary to expectations, sentiment scores and rents are weakly related—higher rents do not consistently coincide with better reviews. This suggests that in a high-demand market like Davis, where students have limited housing choices, rental prices are driven more by location and availability rather than service.
These findings directly address the initial research questions and share some practical suggestions:

- **For students:** The report provides them with adequate information to help to find a balance between rents and other factors.
- **For owners or agents:** They should recognize that management responsiveness and maintenance are what people care the most. Alternatively, they can explore offering optional premium services, such as in-time maintenance, quicker response, as an additional paid option for people who prioritize service.
- **For investors:** They can better plan their investments based on market demands, identifying both high-rent regions with consistent demand and lower-cost regions with different risk-reward dynamics.

By integrating rental prices, living conditions and online reviews, this report offers a comprehensive perspective on the rental market around Davis. Ultimately, it highlights that while location is the main driver of rental prices, while location primarily sets rents, service quality remains a central concern that is not consistently priced.

References

[1] UC Davis Enrollment Management. (2023). UC Davis Record Enrollment Brings Increased Diversity, HSI Eligibility. Retrieved from https://enrollmentmanagement.ucdavis.edu/news/uc-davis-record-enrollment-brings-increased-diversity-hsi-eligibility

[2] Wikipedia. (2023). Davis, California. Retrieved from https://en.wikipedia.org/wiki/Davis,_California

[3] Soci.ai. (2023). State of Google Reviews. Retrieved from https://www.soci.ai/insights/state-of-google-reviews/

[4] United States Census Bureau. (2023). QuickFacts: Davis City, California. Retrieved from https://www.census.gov/quickfacts/fact/table/daviscitycalifornia#