Feng Chia University Outstanding Academic Paper by Students

Title: What Affects Property Rent in San Jose?

Author(s): Yang Jiang, Ke-Ying Mao, Kuan-Hao Su, Hsuan-Cheng Lee

Class: 2nd year of SJSU-FCU 2+2 Bachelors Program in Business Analytics

Student IDs: D0571973, D0572056, D0565664, D0565677

Course: Introduction to Data Analytics

Instructor: Dr. Cathy W. S. Chen

Department: SJSU-FCU 2+2 Bachelors Program in Business Analytics

Academic Year: January 12, 2017-2018

Abstract

This paper examines the various factors influencing property rent in San Jose, by using multiple regression to investigate ten variables. We then utilize forward selection, backward elimination, stepwise regression, Mallow's Cp, and Adjust-R^2 to select the best model. Our findings suggest that an increase in the internal conditions of property, such as number of bedrooms and baths, has positive effects on the rent. Geographical factors also have various influences, as evidence reveals that rent declines when the property is closer to San Jose State University, whereas the overall rent for San Jose decreases from South to North. Furthermore, rent for houses located in a community is higher than that for non-community houses.

Keywords: multicollinearity; model selection; outlier detection; regression analysis; residual analysis.

Table of Contents

1. Introduction

Housing is one of the most important issues for students travelling abroad for studies. Although with the progress in information technology, students are able to easily obtain data from the Internet, it is still a time-consuming process for them to completely understand a foreign rent market and estimate expenditure costs. Furthermore, students in the Department of Business Data Analytics are planning to attend San Jose State University after completing their two-year study at Feng Chia University in Taiwan, and thus it would be beneficial to know where the best place is to stay under various characteristics. Only 14% of students live on-campus at San Jose State University (U.S. News & World Report, n.d.), and on-campus rooms and living areas are small in comparison to the living room, resulting in it being uneconomical for students. As a result, students may prefer to live off-campus, but may also be subject to difficulties in choosing suitable residential areas.

We acquire and collect data for this study from three different sources using web crawler (A web robot that can get data from websites). The three sources used are (i) Trulia, an online residential real estate website; (ii) Google Maps Distance Matrix Application Programming Interface (API) – an API that provides routes and distances between housing and San Jose State University; and (iii) CrimReport - a website providing the latest cases reported by the police department. Details on the collection of data are available on the GitHub (Basically a website that store code). This study thus looks to understand the factors influencing rent values of property in San Jose and highlight the most suitable off-campus areas available to rent for students attending San Jose State University.

2. Method

We first employ summary statistics and scatter plots to provide a basic understanding of the dataset. Second, we use multiple linear regression to build the model. Third, we utilize studentized residual to detect outliers and Cook's distance to identify influential points. Lastly, we apply five different model selection methods to pick out the best model.

2.1 Data Description

The paper examines property rent in San Jose through 11 variables. Data come from Trulia,

Variable	Definition	Source
Response variable		
price	Price of housing in dollars	Trulia (December 2017)
Structural		
sqft	Size of housing in square feet	Trulia (December 2017)
beds	Number of bedroom in the housing	Trulia (December 2017)
baths	Number of bathrooms in the housing	Trulia (December 2017)
Geographical		
lat	Latitude of the housing	Trulia (December 2017)
lng	Longitude of the housing	Trulia (December 2017)
dis	Walking distance between the housing and San Jose State University	Google Maps Distance Matrix API (December 2017)
Attitude		
photoCount	Number of photos uploaded by the owner in the description page	Trulia (December 2017)
pet	Pet dummy: = 1 if renter can have pets in the house; = 0 otherwise	Trulia (December 2017)
Safety		
isComm	Community dummy: = 1 if housing in a community; = 0 if otherwise	Trulia (December 2017)
crim	Number of crime cases around the housing within 1.45 km²	CrimReport (November 2017 - December 2017)
discrim	dis divided by crim	—

Table 1. Definition of variables

CrimeReports, and Google Maps Distance Matrix API, comprising 100 observations selected from full dataset (contains 846 observations) using simple random sampling. Table 1 lists the details of the variables.

We then classify predictors into five factors. The structural factor focuses on internal conditions of the property; geographical factors reflect the housing location; attitude factors deal with permission granted by the owner to have pets and efforts made to rent out the house (efforts are observed through the variable “photoCount”; the more photos posted by the owners onto the webpage, the greater the effort is to rent out the housing); and the safety factor accounts for the level of safety of the property.

2.2 Basic Statistics

In order to provide a basic idea of variables influencing rent in San Jose, Table 2 provides their summary statistics. The statistics include μ (mean), σ (standard deviation), minimum, 1Q (lower quantile), median, 3Q (upper quantile), and maximum. The statistics in the table show that rent per month per person in San Jose is approximately US$1,207. Furthermore, on average, 77 crimes take place around housing on a monthly basis, which is a rather high number. We calculate the square root of crim, nature log of price, disCrim, and sqft due to their large range.

Variable	μ	σ	Minimum	1Q	Median	3Q	Maximum
price	3140.250	1184.49	975.000	2440.000	3125.000	3647.500	9200.000
sqft	1373.960	932.465	175.000	906.000	1214.500	1531.500	9000.000
beds	2.610	1.207	1.000	2.000	3.000	3.000	9.000
baths	1.915	0.817	1.000	1.000	2.000	2.000	5.000
photoCount	13.150	9.441	0.000	7.000	12.000	20.000	58.000
lat	37.308	0.044	37.209	37.288	37.314	37.336	37.414
lng	-121.886	0.061	-122.031	-121.922	-121.890	-121.844	-121.740
dis	6.265	3.179	0.300	3.850	6.850	8.750	11.900
isComm	0.080	0.273	0.000	0.000	0.000	0.000	1.000
pet	0.230	0.423	0.000	0.000	0.000	0.000	1.000
disCrim	43.126	126.925	0.093	2.916	10.252	27.462	1007.500

Table 2. Summary statistics of the variables.

2.3 Plots

Figure 1 demonstrates a positive correlation between the size of housing and price, thereby indicating that an increase in this leads to higher rent. In Figure 2, we directly remove an outlier (marked in black circle), and Figure 3 is the scatter plot after removal of this observation. The results indicate that property rent rises as the number of bedrooms increases.

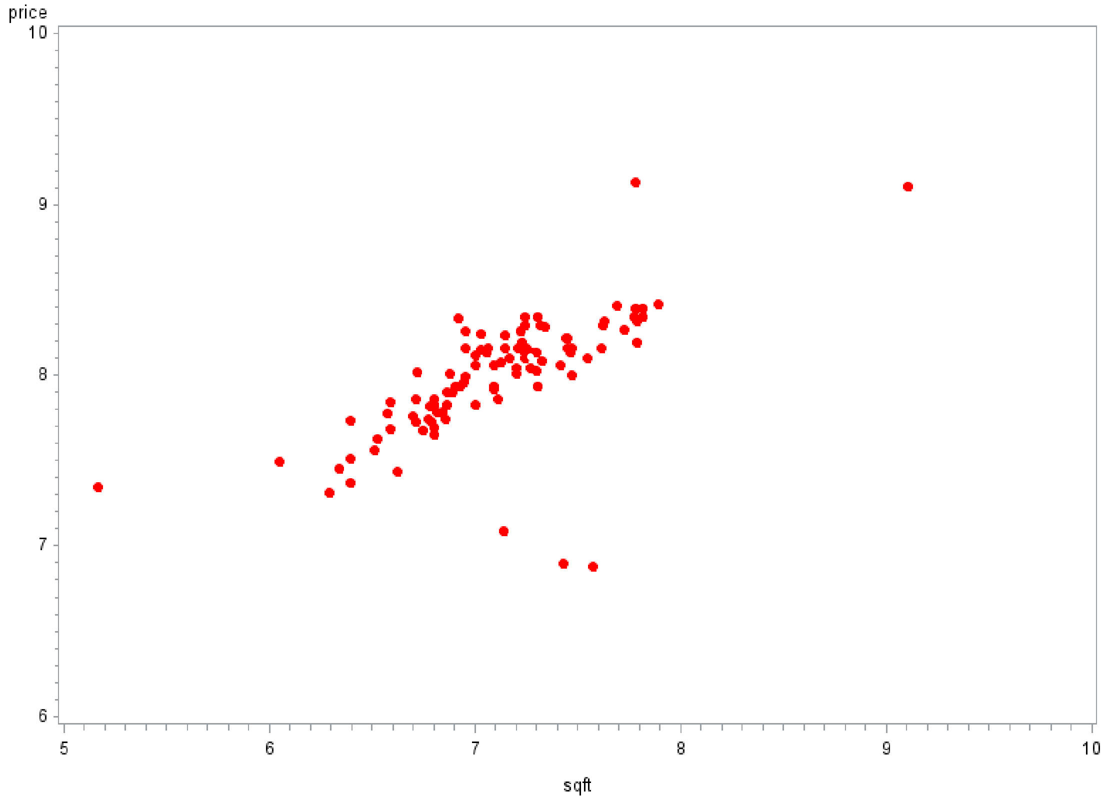


Figure 1. Scatter plot - sqft versus price.

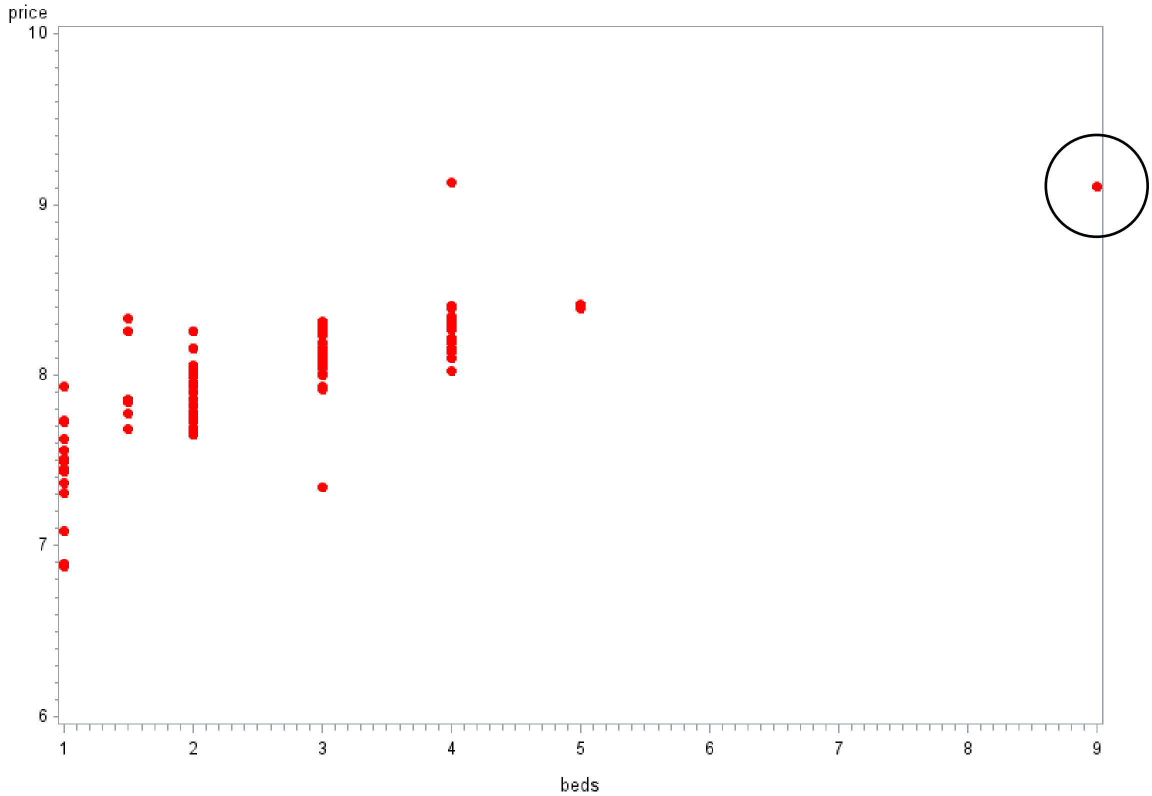


Figure 2. Scatter plot – No. of bedrooms versus price

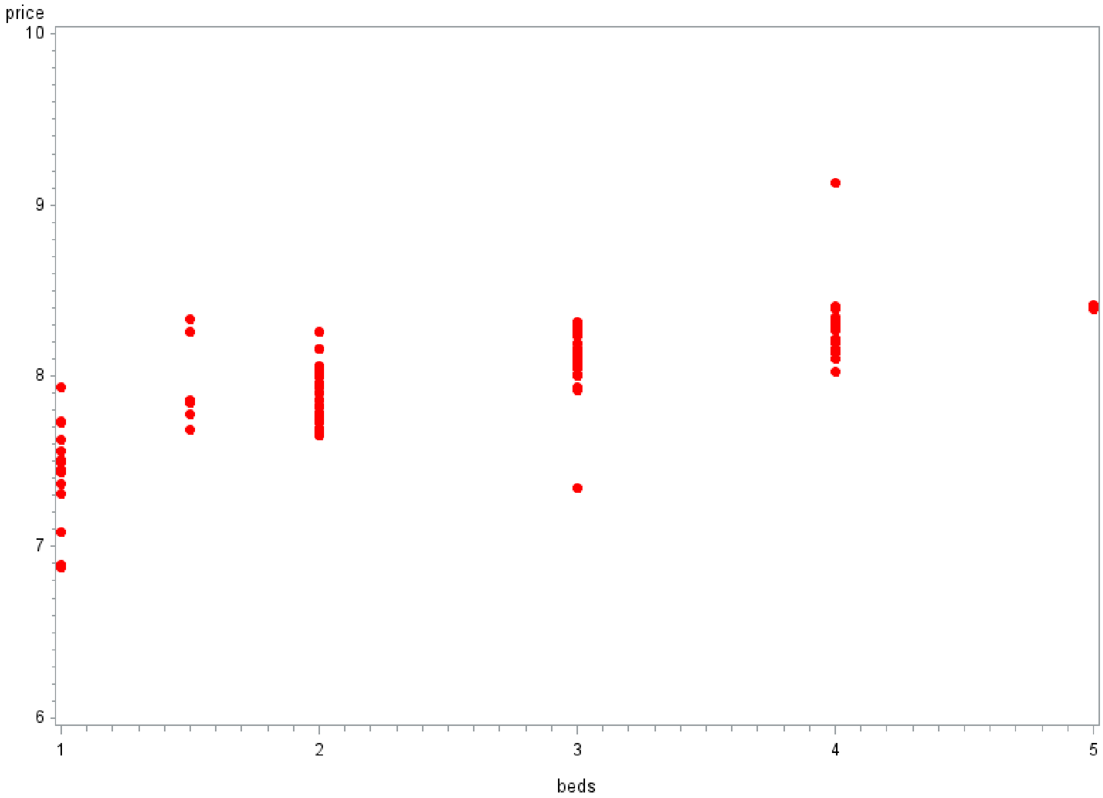


Figure 3. Scatter plot – No. of bedrooms versus price after removing the outlier.

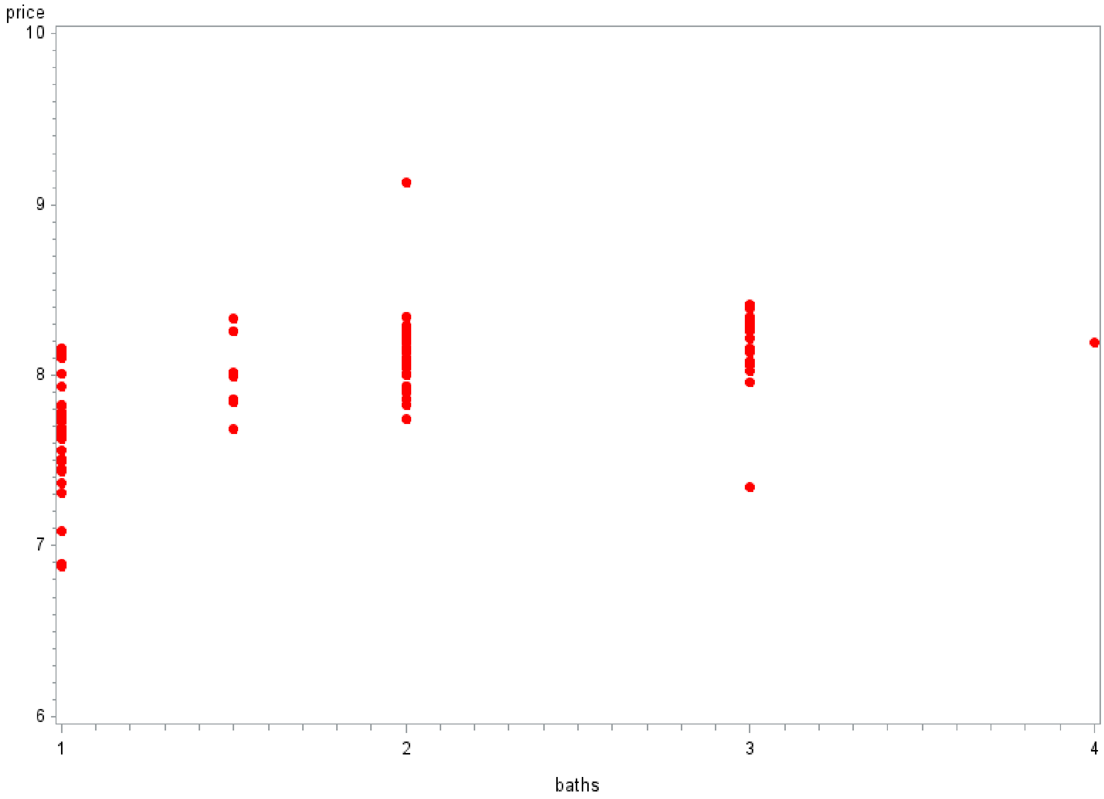


Figure 4. Scatter plot – No. of baths versus price.

Figure 4 exhibits a positive correlation between the number of bathrooms and price, which indicates that a greater number of baths lead to higher rent.

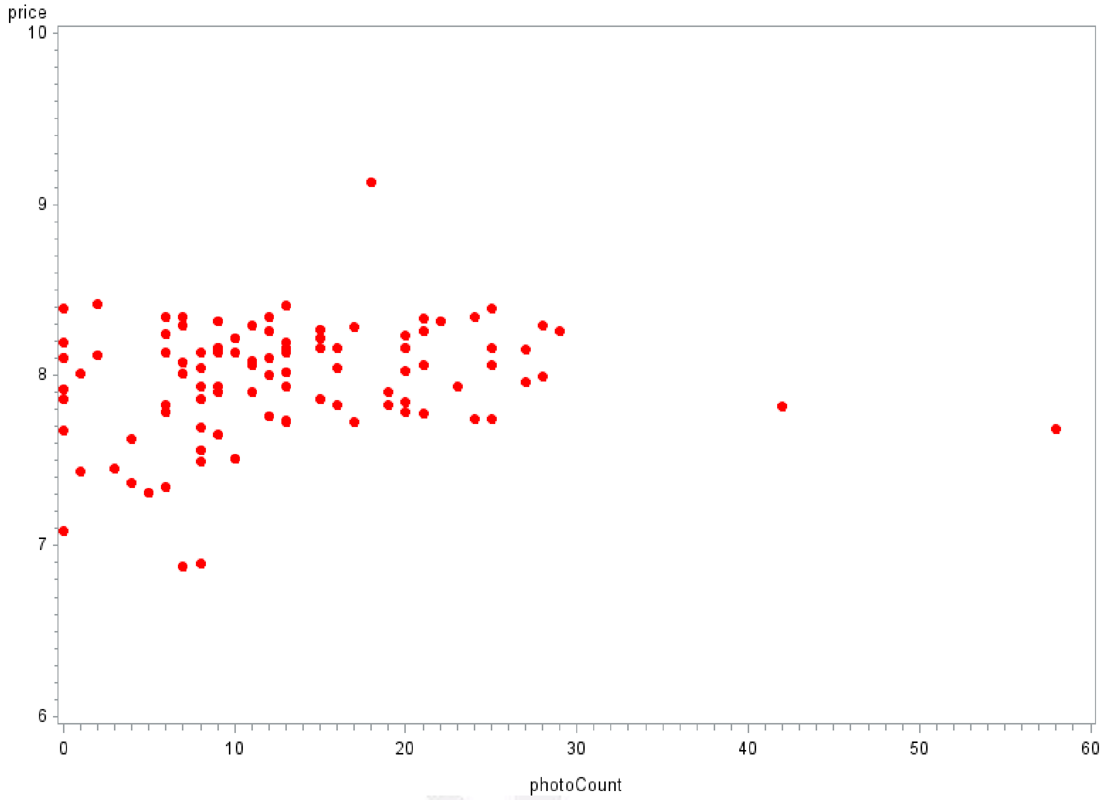
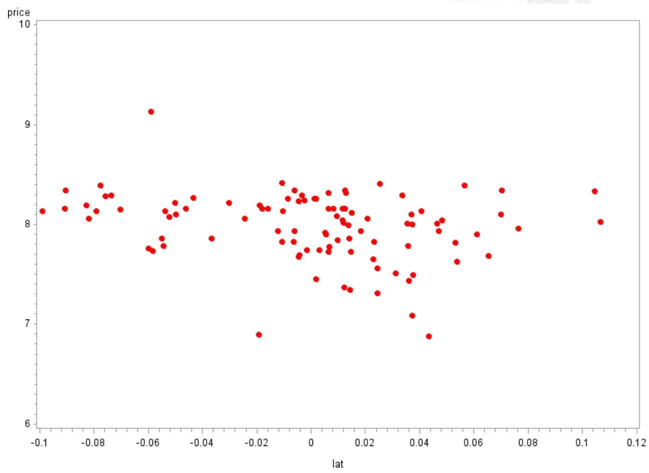


Figure 5. Scatter plot - photoCount versus price.

Figure 5 shows a positive correlation between photoCount and price, indicating that photos on the description page lead to a higher price.



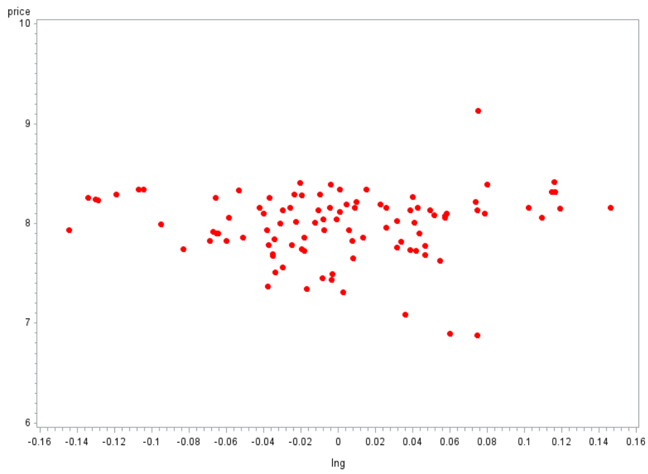


Figure 6. Scatter plot lat versus price. *Figure 7.* Scatter plot lng versus price.

Figures 6 and 7 are scatter plots demonstrating latitude versus price and longitude versus price, respectively. The data indicate no significant correlation between these two variables and price.

Figure 8 shows the relationship between distance to San Jose State University and price. The result is that being closer to campus leads to lower rent.

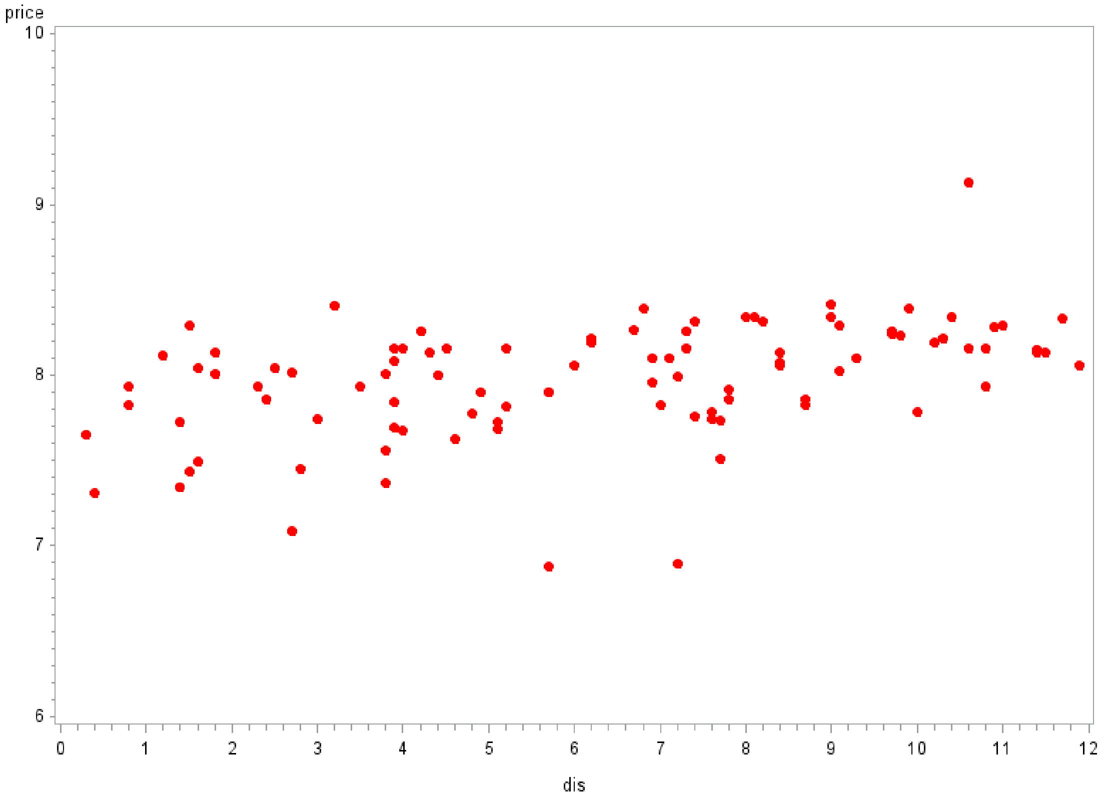
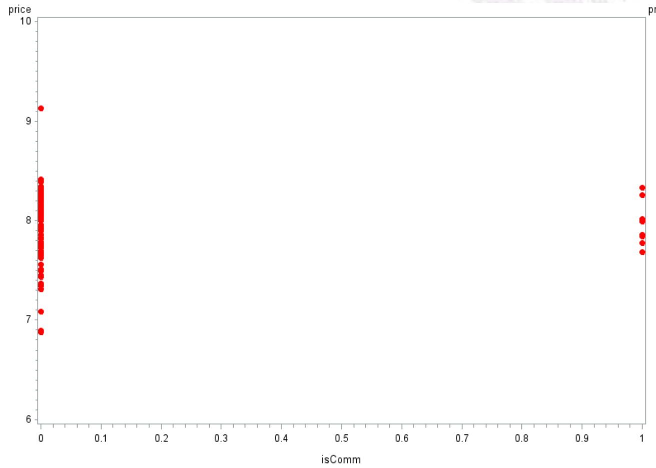


Figure 8. Scatter plot - distance versus price.



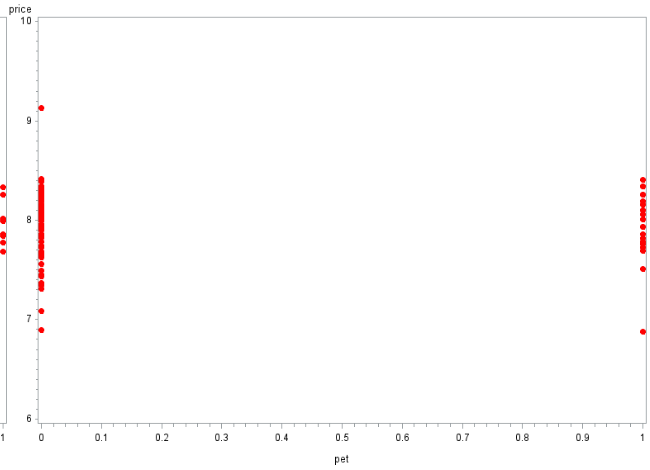


Figure 9. Scatter plot - isComm versus price.

Figure 10. Scatter plot - pet versus price.

Figures 9 and 10 are scatter plots of two dummy variables: isComm and pet. The results indicate that housing located in a community and pet-friendly housing do not have an impact on rent.

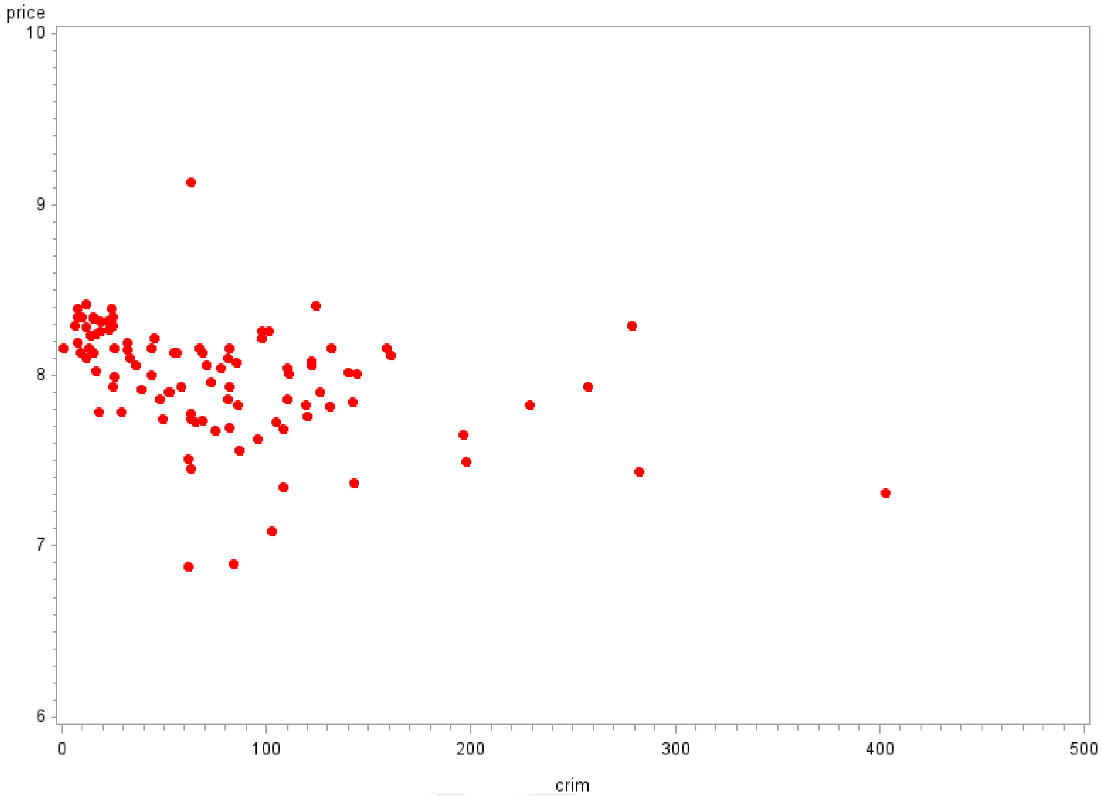


Figure 11. Scatter plot - crim versus price.

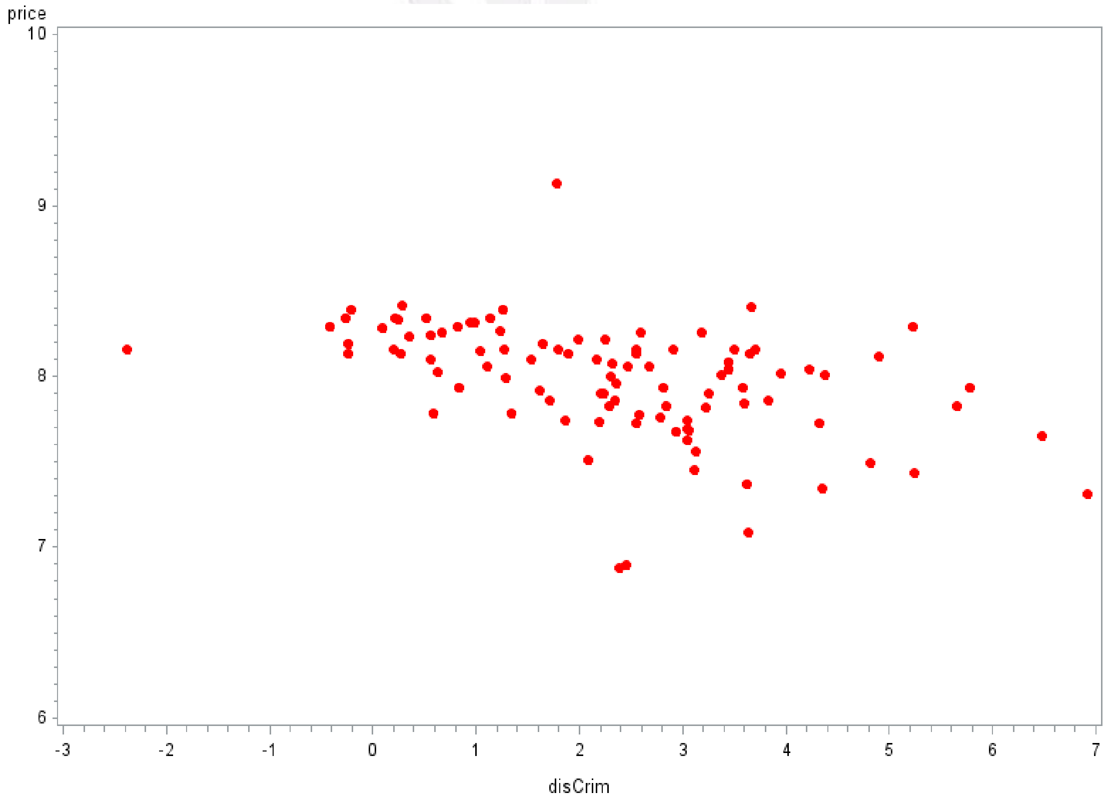


Figure 12. Scatter plot disCrim versus price.

Figure 11 demonstrate a negative correlation between the number of crimes around housing and rent. A higher number of crimes leads to lower rent.

Figure 12 shows a negative correlation between disCrim and price.

2.4 Correlation Matrix

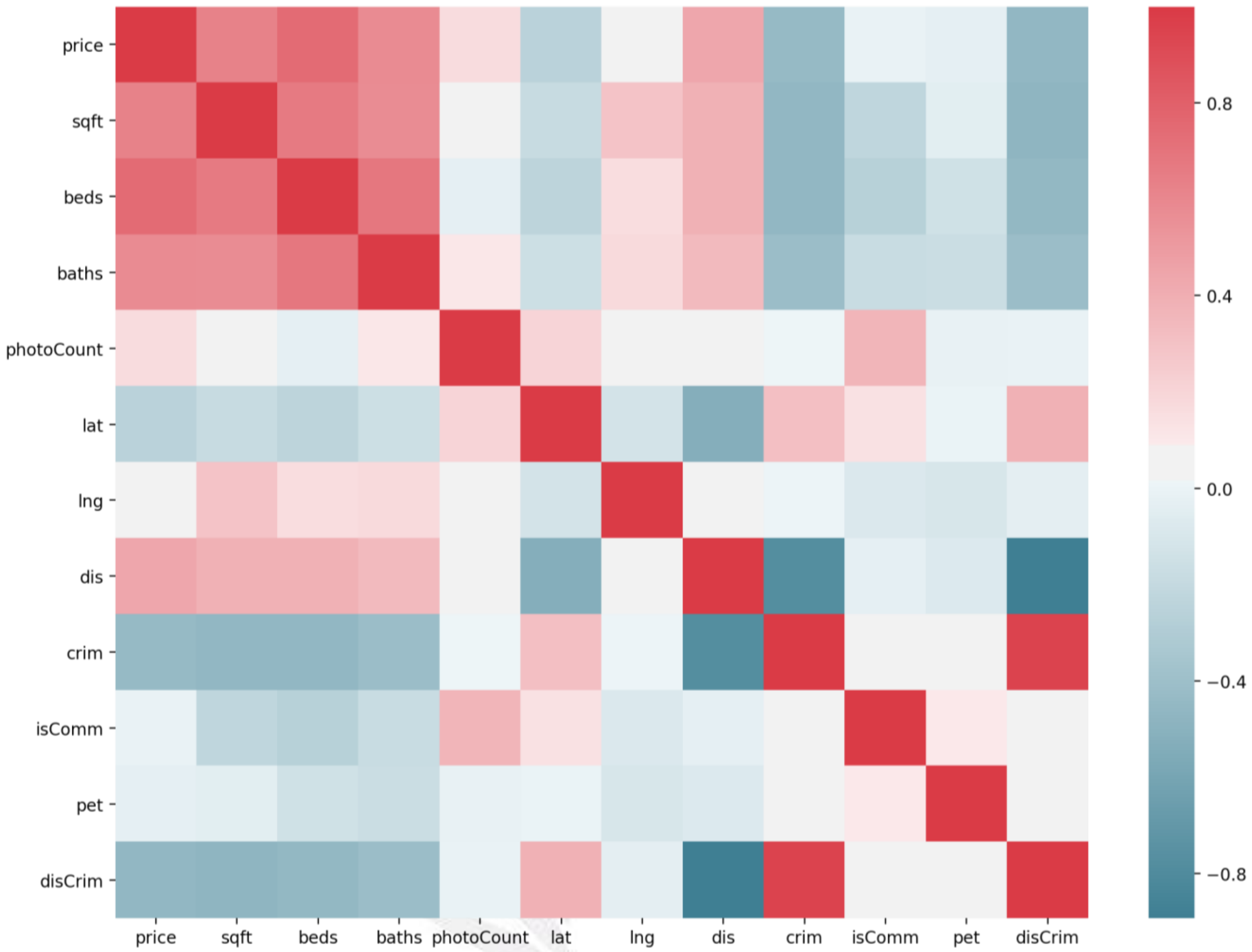


Figure 13. Correlation matrix of variables. The red block indicates a positive correlation, while the blue block indicates a negative correlation.

Correlation quantifies the "strength of the linear association between two variables" (Barron & Kim, 1997), and a correlation matrix demonstrates correlations between a set of variables. The correlation matrix in Figure 13 reflects that the size of the housing, the number of bedrooms, the number of bathrooms, and distance to San Jose State University have a significantly positive correlation with price, thus supporting that an increase in these components lead to higher rent. In contrast, the number of crimes around housing has a strongly negative correlation with price, thereby indicating that more crimes in an area lead to lower rent. Another important finding is that there is a large negative correlation between distance to San Jose State University and number of crimes around housing, indicating that more crimes take place around San Jose State University.

2.5 Full Model

Variable	Parameter estimated	Standard error	t Value	Pr > |t|	Variance inflation
Intercept	5.787	0.501	11.560	<.0001	0.000
sqft	0.200	0.069	2.900	0.005	2.251
beds	0.191	0.031	6.050	<.0001	2.584
baths	0.033	0.038	0.870	0.389	2.102
photoCount	0.005	0.002	1.930	0.057	1.304
lat	-0.693	0.582	-1.190	0.237	1.603
lng	-0.984	0.359	-2.740	0.008	1.174
dis	0.011	0.019	0.590	0.560	8.610
isComm	0.195	0.084	2.340	0.022	1.303
pet	0.037	0.049	0.760	0.449	1.070
disCrim	-0.019	0.070	-0.270	0.789	32.395
crim	0.015	0.022	0.670	0.505	16.280

Table 3. Parameter estimates for the full model.

Using the multiple linear regression, the full model is:

$$ln(\widehat{price}) = 0.200ln(sqft) + 0.191beds + 0.033baths + 0.05photoCount$$

$$-0.693lat - 0.984lng + 0.011ln(dis) + 0.195isComm + 0.037pet$$

$$-0.019ln(disCrim) + 0.015\sqrt{crim}, \widehat{\sigma^2} = 0.040.$$

In order to use multiple linear regression, the assumption is that the distribution of the error term is normal and with zero mean; furthermore, the variance σ^2 cannot be volatile (The Pennsylvania State University, 2018a). In this model, the value for Adjust - R^2 is 0.655, which means that the model can explain 65.5% of the variance. We use Adjust - R^2 instead of R^2, because it is an "unbiased estimator that corrects for the sample size and numbers of coefficients estimated" (Nau, 2017). For this model,

$\widehat{\sigma^2}$ is 0.040, and it is the estimate of σ^2, which "quantifies how much the responses vary around the mean population regression line" (The Pennsylvania State University, 2018b).

Variance inflation (VIF) can detect the existence of multicollinearity; i.e. how strong a correlation is between predictors. Multicollinearity exists if VIF is greater than ten and can lead to insignificance of the individual Bata even if R^2 is large (Chen, 2017a). As demonstrated in Table 3, VIFs for the variables of disCrim and crim are greater than ten, which means that multicollinearity exists in the full model. We later use variable selections to resolve this problem by selecting a suitable subset of predictors.

2.6 Outliers and Influential Point Detection

Outliers are "extreme observations in the response variables" and generally have extreme values in the response variable. In contrast, leverage points have extreme values in the independent variable. Furthermore, outliers can substantially influence regression estimates, and if the outlier values have high leverage, then they denote 'influential points' (Refer Freund, Wilson, & Sa, 2006).

We utilized the student residual (r_i) to detect outliers, with an observation considered as an outlier if $|r_i| > 3$. There are three outliers in the dataset, and the locations of these outliers are 55, 60, and 78, respectively. We delete all outliers immediately upon detection and locate influential points using Cook's distance (Cook's D). We then define an observation as a possible influential point if Cook's D is greater than 0.5. In the dataset, the Cook's D for each observation is less than 0.5 and therefore indicates that there are no influential points in the dataset. While influential points do not appear in the dataset, outliers still have a significant influence on estimation. For instance, the scatter plot on studentized residuals versus fit plot shows a significant pattern prior to deleting the outlier, indicating that the residual's variance is not constant due to the existence of outliers.

2.7 Model Selection

This research uses five different methods to select the best model: backward elimination, forward selection, stepwise regression, adjust-R^2 selection, and Cp selection.

Backward elimination begins with the full model, and thereafter we calculated the p-value for every predictor in the model. We remove from the model any predictor with a p-value larger than the significant level (α) and the largest from this set, halting elimination only after every in-model independent variable's p-value is smaller than α. Forward selection starts with an empty model and then computes the p-value for every predictor that is not in the model. Only the predictor with the smallest p-value and the p-value fulfilling the criterion α can enter the model. The selection continues until there are no external independent variable p-values less than α. Stepwise regression is a combination of backward elimination and forward selection and has two criteria: α_{in} and α_{out}. It starts with an empty model and uses the same procedures as in forward selection. However, after the independent variable joins the model, it re-evaluates all independent variables in the model using the backward elimination process. Stepwise regression stops when the p-values of the model's independent variables meet both criteria (α_{in} and α_{out}). Finally, the Adjust - R^2 selection uses the model with the highest Adjust - R^2, while the Cp selection uses the model with the lowest Cp (Chen, 2017b).

Variable	Backword	Forward	Stepwise	Adjust - R^2	Cp
price					
sqft	✓	✓	✓	✓	✓
beds	✓	✓	✓	✓	✓
baths				✓	
photoCount	✓	✓	✓	✓	✓
lat					
lng	✓	✓	✓	✓	✓
dis	✓	✓	✓	✓	✓
isComm	✓	✓	✓	✓	✓
pet	✓	✓	✓	✓	✓
disCrim					
crim					

Table 4. Models based on different selection methods. The significant level for backward elimination, forward selection, and stepwise regression is 0.15. "✓" indicates variables selected by the relevant method.

Table 4 demonstrates backward elimination, forward selection, stepwise regression, and Cp selection methods using the same model. We use the model selected by these four methods. The model is:

$$ln(\widehat{price}) = 5.190 + 0.334ln(sqft) + 0.117beds + 0.003photoCount - 0.663lng$$

$$+0.013dis + 0.162isComm + 0.061pet, \widehat{\sigma^2} = 0.017.$$

Adjust - R^2 for this model is 0.790, and $\widehat{\sigma^2}$ is 0.017. In Table 5, every variable's VIF in this model is less than 10. Therefore, multicollinearity does not exist in the selected model.

Variable	Parameter estimated	Standard error	t Value	Pr > |t|	Variance inflation
Intercept	5.190	0.286	18.160	<.0001	0.000
sqft	0.334	0.045	7.340	<.0001	2.229
beds	0.117	0.020	5.940	<.0001	2.247
photoCount	0.003	0.002	1.680	0.098	1.189
lng	-0.663	0.230	-2.890	0.005	1.092
dis	0.013	0.005	2.850	0.005	1.226
isComm	0.162	0.054	3.000	0.004	1.288
pet	0.061	0.032	1.920	0.059	1.038

Table 5. Parameter estimates for the selected model.

2.8 Additional Treatment

The performance of the model increases significantly after the model selection, but a new outlier appears. The location of this outlier is 98 with a student residual equal to 5.045. Upon detection of this outlier, we delete it. Furthermore, we repeat model selection and use the model nominated through stepwise selection, which is:

$$\ln(\widehat{price}) = 4.923 + 0.389ln(sqft) + 0.089beds - 0.578lng + 0.011dis$$

$$+0.180isComm + 0.046pet, \widehat{\sigma^2} = 0.012.$$

Variable	Parameter estimated	Standard error	t Value	Pr > |t|	Variance inflation
Intercept	4.923	0.245	20.120	<.0001	0.000
sqft	0.389	0.039	9.990	<.0001	2.282
beds	0.089	0.017	5.170	<.0001	2.323
lng	-0.578	0.195	-2.970	0.004	1.086
dis	0.011	0.004	2.850	0.005	1.219
isComm	0.180	0.043	4.240	<.0001	1.109
pet	0.046	0.027	1.680	0.096	1.039

Table 6. Parameter estimates for the selected model after deleting the outlier.

Adjust - R^2 for this model is 0.832, and $\widehat{\sigma^2}$ is 0.012. As Table 6 demonstrates, every variable's VIF for this model is still less than 10. Thus, we use the same model throughout the rest of the study.

2.9 Verify Four Assumptions

There are four assumptions for error terms: zero mean, equality of variance, independence, and normality (Bansal, n.d.).

Mean of residuals equals zero, $E(\epsilon_i) = 0$. The hypothesis to test this assumption is:

$$E(\epsilon_i) = 0$$

$$E(\epsilon_i) \neq 0.$$

Based on the student's t test, sign test, and signed rank test in Table 7, the p-values for these three tests are all greater than 0.05, therefore failing to reject H_0. The mean of the residuals is hence equal to zero.

Test	Statistics	p Value
Student's t	t = 0.851	0.397
Sign	M = -0.5	1.000
Signed rank	S = 123	0.650

Table 7. Test for location: $\mu = 0$

Variance of residuals is constant, $Var(\epsilon_i) = \sigma^2$. Figure 14 exhibits the residual plot (studentized residual versus fitted plot). An observation of the plot indicates that residuals in the plot locate randomly around zero and do not have significant patterns. Thus, the conclusion drawn is that the variance of residuals is constant.

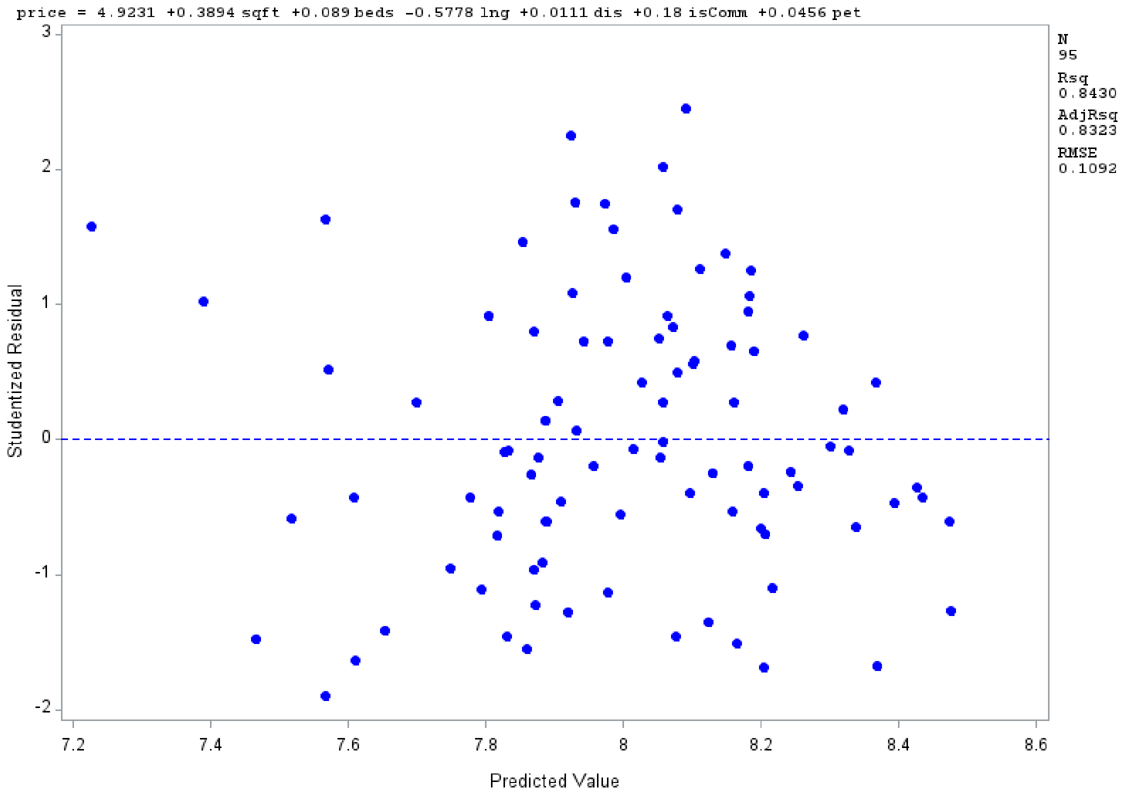


Figure 14. Studentized residuals versus fit plot

Order	DW	Pr < DW	Pr > DW
1	1.744	0.088	0.913

Table 8. Durbin-Watson statistics.

Independence of residuals, $\boldsymbol{Cov(\epsilon_i, \epsilon_j) = 0\ \forall i \neq j}$. The hypotheses for this assumption are:

$$H_0: \rho = 0$$

$$H_a: \rho > 0,$$

and

$$H_0: \rho = 0$$

$$H_a: \rho < 0$$

The first hypothesis is a test for positive autocorrelation, while the second one is a test for negative autocorrelation. This study utilizes the Durbin-Watson statistic (DW) to check for the existence of autocorrelation. Here, DW is between zero and four. A value equal to two means there is no autocorrelation, while values leaning toward zero reflect negative autocorrelation and values closer to four represent positive autocorrelation. In Table 8, although DW is not equal to two, the p-value for testing positive autocorrelation (Pr < DW) is 0.088, which is greater than 0.05. Thus, it fails to reject H_0 in the first hypothesis, and the failure results in no positive autocorrelation for residuals. The p-value for testing negative autocorrelation is 0.913, which is also greater than 0.05 and thus a failure to reject H_0 in the second hypothesis results in no negative autocorrelation for residuals.

Test	Statistics	p Value
Shapiro-Wilk	W = 0.983	0.244
Kolmogorov-Smirnov	D = 0.067	>0.1500
Cramer-von Mises	W-Sq = 0.085	0.181
Anderson-Darling	A - Sq = 0.556	0.151

Table 9. Test for normality.

Normal distribution of residuals, $\boldsymbol{\epsilon_i \sim N(0, \sigma^2)}$. The hypothesis to test this assumption is:

$$H_0: \epsilon_i \sim N(0, \sigma^2)$$

$$H_a: \epsilon_i \nsim N(0, \sigma^2)$$

The four tests to calculate the p-value demonstrated in Table 9 reveal values that are all greater than 0.05 and therefore do not reject H_0, indicating that the distribution of the residual is normal.

3. Results

The dataset contains five outliers and has no influential point. Variables in the final model are all significant at the 0.05 level, except for the variable dealing with pet-friendly housing.

The results suggest that, in addition to structural factors, geographical and safety factors also influence a property's rent. Houses within a community have a higher rent, and an increase in the following variables can also increase the rent: size of the house; number of bedrooms in the house; and distance between the housing to San Jose State University. However, in contrast, an increase in the longitude can decrease the rent value for the housing.

4. Discussion

4.1 Student Safety

Although the variable ‘crime’ is not selected for the final model, it is still an important factor that students need to be concerned about. Figure 14 is the scatter plot demonstrating ‘distance versus crimes’ for a total of 846 observations. As the plot shows, crime rates are fairly high in the areas located between 0 to 2.5 miles away from San Jose State University. Therefore, students may want to find rental property located 2.5 miles away from the campus in order to have a safer environment.

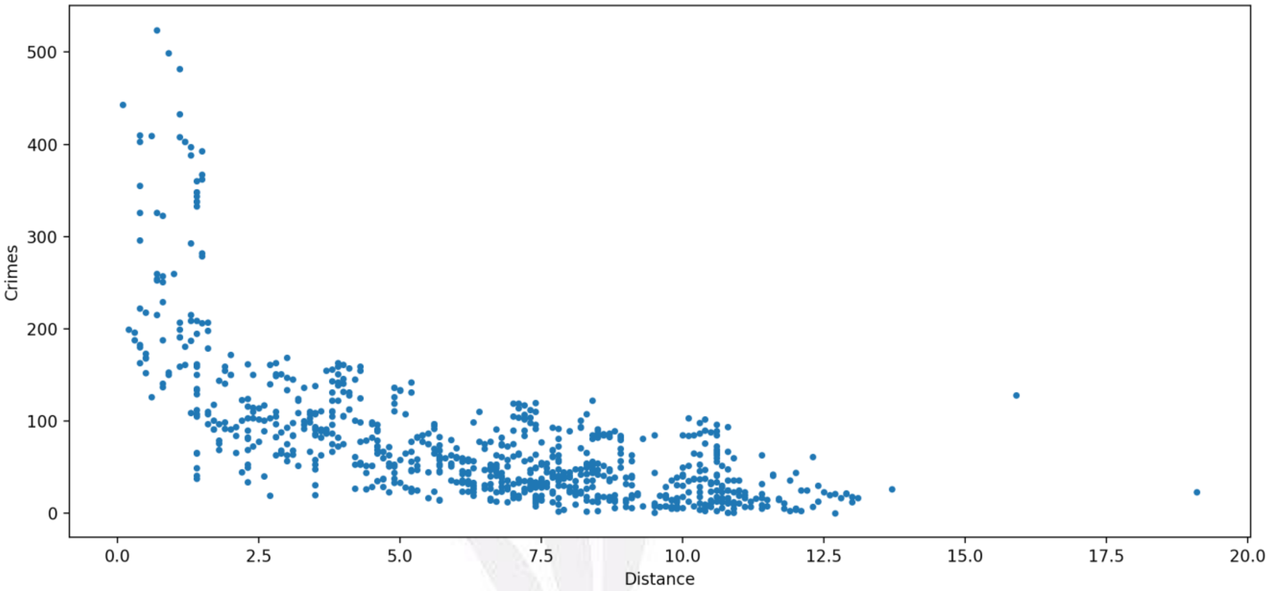


Figure 15. Distance versus crime plots.

4.2 Amount Saved by Living Off-campus

There are four dormitories in San Jose State University: building A, building B, building C, and building 2 (San Jose State University, 2017). In Figure 16, the y-axis is the average savings per month for students who live off-campus. These values are calculated as the average cost of living in a dormitory (without a meal plan) per month minus the average cost of living off-campus per month. On average, a student can save US$1,000 by living off campus for seven months. However, for those students who really want to save money, building 2 is the perfect option, as living in this building is more economical than living off-campus.

It is necessary to highlight that the rental housing market changes over time, and these estimates can be different in the future. Furthermore, the number of crimes around housing may fluctuate across different seasons.

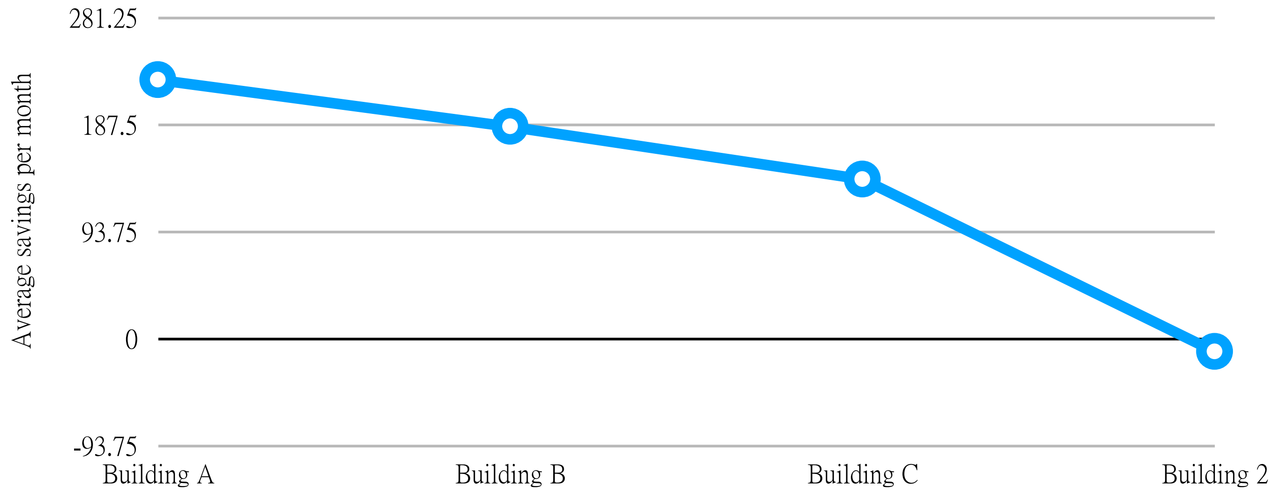


Figure 16. Average savings per month by living off-campus.

5. Conclusion

Various factors influence rental property in San Jose, including those that are structural, geographical, attitudinal, and safety. The fact that the number of crimes is not accounted for in the final model is a surprising observation, but due to the rapid nature of changes in San Jose, it does not significantly affect rent. The following conclusion offers certain suggestions for landlords and students.

In the case of property owners in San Jose, there is a high opportunity cost associated with placing that focuses on uploading images of housing to the relevant description page, because this action does not have a significant influence on rent. Since it is almost impossible for owners to change the geographic factor of the property, one recommendation is for them to add a reasonable number of bedrooms to increase rent. Moreover, allowing tenants to have pets in the house can also facilitate higher rents for landlords.

In the case of students, in most situations it is cheaper to live off-campus than to live on-campus. Rent is lower on the north side of San Jose State University in comparison to properties to the south of campus. Furthermore, the number of crimes holds at a steady rate in areas located 2.5 miles away from campus. As Figure 17 demonstrates, one recommendation for students who want to live off-campus is to live north of the university and at least 2.5 miles away from it, as the rent is relatively cheap and the surroundings are safe.

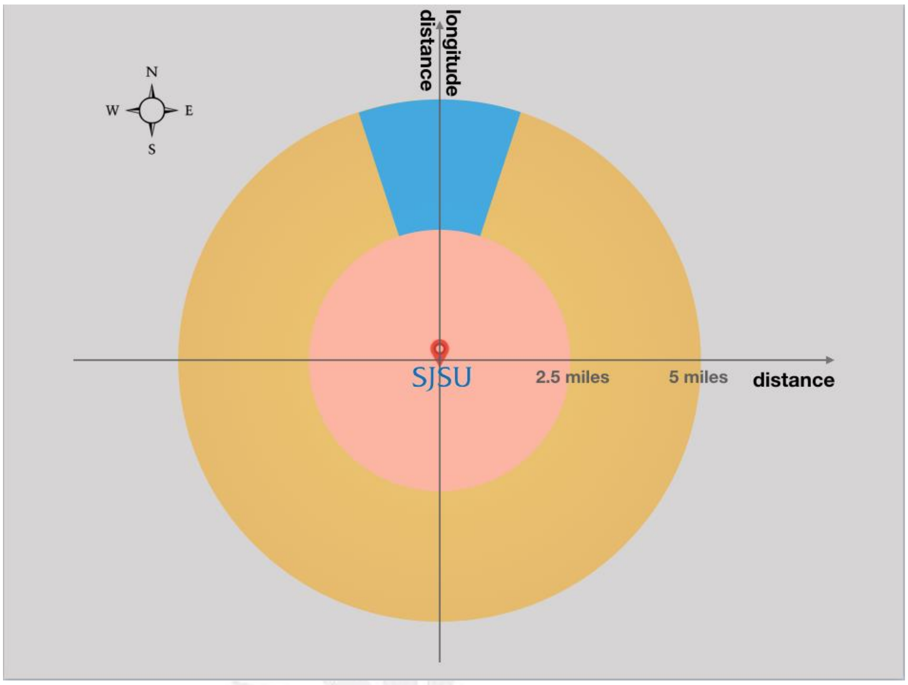


Figure 17. Recommended rent area (blue area). Red area indicates high crime area, while yellow area indicates normal area.

References

Barron, A., & Kim, J. (1997). Introduction to Statistics: Life Sciences. *Correlation*. Retrieved from: http://www.stat.yale.edu/Courses/1997-98/101/correl.htm

Bansal, G. (n.d.). What are the four assumptions of linear regression? Retrieved from: http://blog.uwgb.edu/bansalg/statistics-data-analytics/linear-regression/what-are-the-four-assumptions-of-linear-regression/

Chen, W. S. (2017a). Introduction to Data Analytics [PDF file]. *Lecture 5*. Retrieved from: http://ilearn2.fcu.edu.tw/pluginfile.php/688395/mod_folder/content/0/reg17_lec_5_multicollinearity.pdf?forcedownload=1

Chen, W. S. (2017b). Introduction to Data Analytics [PDF file]. *Lecture 6*. Retrieved from: http://ilearn2.fcu.edu.tw/pluginfile.php/694361/mod_folder/content/0/reg17_lec_selection.pdf?forcedownload=1

Freund, R. J., Wilson, W. J., & Sa, P. (2006). *Regression analysis* (2nd ed.). Cambridge, JHY: Academic.

Nau, R. (2017). Linear regression models. *What's a good value for R-squared?* Retrieved from: https://people.duke.edu/~rnau/rsquared.htm

San Jose State University. (2017). 2017-2018 Academic Year Rates. Retrieved from: http://www.housing.sjsu.edu/rates/2017-2018-academic-year-rates/index.html

The Pennsylvania State University. (2018a). Lesson 5: Multiple Linear Regression. *5.3 – The Multiple Linear Regression Model*. Retrieved from: https://onlinecourses.science.psu.edu/stat501/node/311

The Pennsylvania State University. (2018b). Lesson 1: Simple Linear Regression. *1.4 - What is The Common Error Variance?* Retrieved from: https://onlinecourses.science.psu.edu/stat501/node/254

U.S. News & World Report. (n.d.). San Jose State University. *Housing*. Retrieved from: https://www.usnews.com/best-colleges/san-jose-state-1155/student-life#