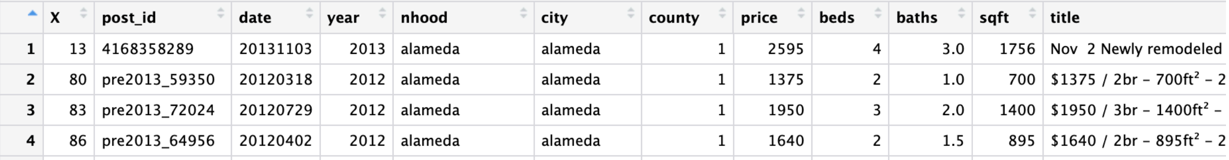
Samuel Reade and Ava Exelbirt
Statistical Learning Project 1

Goals and Introduction: We aim to explore and understand the relationships among various housing characteristics, such as price, square footage, number of bedrooms, and bathrooms, using Principal Component Analysis (PCA), Factor Analysis, and Clustering for houses in Bay Area, California. Some relationships we can predict are: predicted price by square footage, predicting price by county, and predicting county by square footage. We can use *PCA* to identify the key components that capture the most variance in housing features, such as price per square foot, price per bedroom, and price per bathroom. This will help us understand which combinations of features are most influential in describing the overall housing landscape. *Factor Analysis* can be utilized to uncover underlying latent variables or factors that may drive housing market characteristics. For example, factors might reflect "size and amenities" or "cost-efficiency," enabling a deeper interpretation of the data's structure. We can apply *clustering* to group houses into distinct segments based on their features. This analysis may reveal natural groupings such as luxury homes, budget-friendly options, or mid-range housing, offering insights into the market's composition. Overall, we can investigate the relationships among variables: we can predict house prices based on square footage, number of bedrooms, and bathrooms. We can analyze how price-per-unit metrics (e.g., price per square foot) vary by cluster. We can also explore potential regional or neighborhood-level trends. Overall, we can use PCA, FA, and the clustering results to understand how different house characteristics correlate with pricing tiers or market segmentation. On a practical level, understanding information regarding rental units can help provide insights to buyers, real estate agents, developers, and sellers on how to evaluate and compare properties based on a simplified set of meaningful variables.

Our Process: We first scraped the web for a data set that we find interesting and useful to perform analysis on. We first figured out what our research goals were in relation to the data set. After finding the Rent data set, we went into the data cleanup process. We scrapped the data, then cleaned it for outliers and NA values. After this, we thought of any new features that can be added that would be beneficial in our analysis steps. We performed feature engineering, adding 3 new features to our cleaned data set. We performed correlation analysis on the new cleaned data set and went on to perform PCA, FA, and clustering. We went back and forth, using a shared github repository to work on the same rmd markdown file. We each worked on parts and would commit and push our changes to the repository when done. Then, the other person would go in and pull the changes to work on the updated markdown file. We would meet once a week and go over our changes, giving each other feedback, and editing our work.

About the Dataset: Using a data set from TidyTuesday that contains information about rent prices for houses in the Bay Area of the Bay Area. The data set contains information regarding date of rental, location of the house (like city and county for example), information regarding size of house, and price of rent. The original data set Rent had 200,769 observations among 17 variables. However, as part of our data preprocessing we chose to drop rows that contain NA values and drop rows that contain observations with outliers, as we want our data set to have the most thorough representative data as possible. Considering there were still thousands of observations regarding rental prices in San Francisco, dropping these rows was a better choice than replacing their values, as we chose to only perform analysis on real data. The variables we chose to include in our model were **year**, **neighborhood**, **city**, **county**, **price**, **beds**, **baths**, **sqft**, and **title**. Year is the year the house was on the market, neighborhood, city, and county outline the area the house is located in the Bay Area, price is the rental price is US dollars per one month of rent, beds is the amount of bedrooms the house has, baths is the amount of bathrooms the house has, sqft is the amount of sqft the house has, and title is the title of the rental listing posting. We chose these variables because they are the most representative of houses in the Bay Area in relation to our research question.

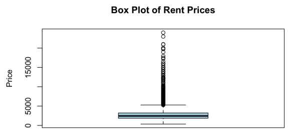
Data Preprocessing Step: We chose to remove some columns including latitude and longitude of apartment location, address of apartment, details (which are additional details), descr (description of the house), and room_in_apt. We removed these because we believe county, city, and neighborhood are better descriptors of analysis regarding location compared to latitude, longitude, and address. We also believe that descr and details will not be beneficial for our analysis as they are character descriptions. We also think that Room_in_apt can be described by bed and bath. The variables we did choose are most beneficial in answering our research questions and are best for performing factor analysis, PCA, and clustering with. After removing NA values, there are still 14,394 observations. Then, we checked for outliers. There were outliers present in price, bed, sqft, and bath variables. We chose to drop these as well, and have a final data set of 12,854 entries.

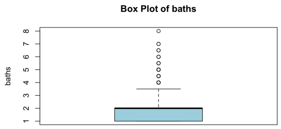


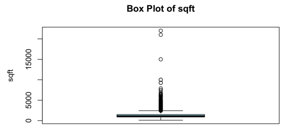
	X	post_id	date	year	nhood	city	county	price	beds	baths	sqft	title
1	13	4168358289	20131103	2013	alameda	alameda	1	2595	4	3.0	1756	Nov 2 Newly remodeled
2	80	pre2013_59350	20120318	2012	alameda	alameda	1	1375	2	1.0	700	$1375 / 2br - 700ft² - 2
3	83	pre2013_72024	20120729	2012	alameda	alameda	1	1950	3	2.0	1400	$1950 / 3br - 1400ft² -
4	86	pre2013_64956	20120402	2012	alameda	alameda	1	1640	2	1.5	895	$1640 / 2br - 895ft² - 2

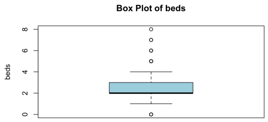
EDA: We chose to perform some EDA analysis to see the distributions of each variable. We can see the price is normally distributed with a center around $2,500/month. Sqft is also normally distributed with the center around 1,275 sqft. Beds are normally distributed with 2 beds being the most common observation, 1 being the least and 4 being the greatest number of bedrooms in an apartment. Number of baths is right skewed with the majority of houses having 1 or 2 bathrooms. Looking at the relationship between some variables in the data set, we can see that price vs. sqft has a positive almost linear relationship. We also plotted the price distribution by county to see the median prices in each county.

Side by side box plots: By visualizing the price distribution with side by side boxplots for each county it allows one to see how the prices in each county compare to each other. It is clear to see that San Francisco county has the highest rent, and Solano has the lowest. It is also very helpful to understand which counties have extreme price points. Which can be seen by the red dots on the graph.

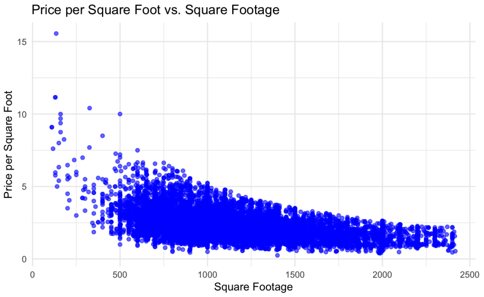


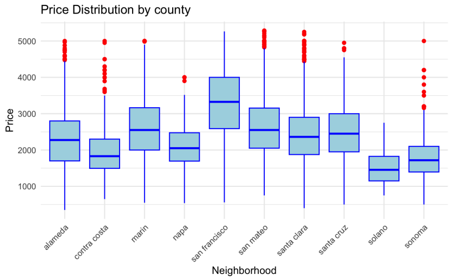


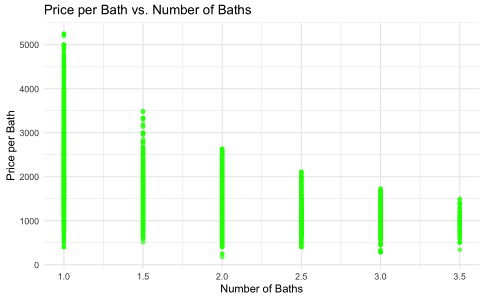


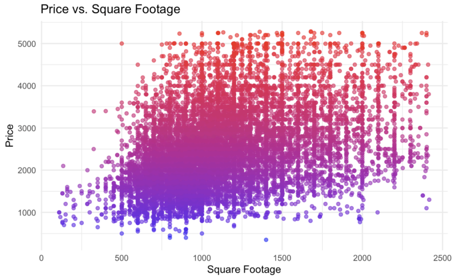


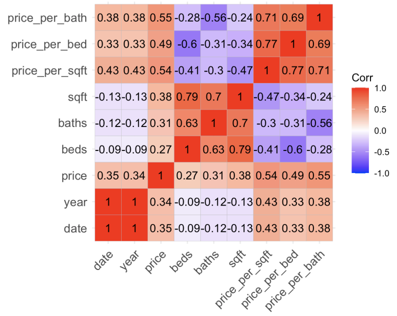
Feature Engineering: We added some new variables as well: **price_per_sqft** was calculated by taking the per month rental price divided by the house's sqft. It is right skewed with the center at about 2.5$/sqft. **Price_per_bath** is the rental price divided by the number of baths. It is right skewed with the center at about $1,500/bed. **Price_per_bed** is the rental price divided by the number of beds. It is right skewed with the center at about $1,000/bed. We then scaled the data set to later use for our analysis. We furthermore conducted correlation analysis with this new data set to see which variables are related.







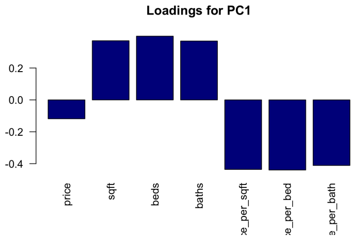


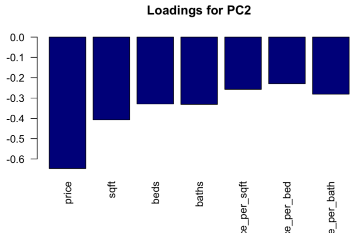


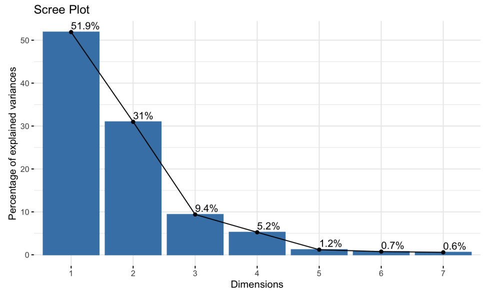
	date	year	price	beds	baths	sqft	price_per_sqft	price_per_bed	price_per_bath
price_per_bath	0.38	0.38	0.55	-0.28	-0.56	-0.24	0.71	0.69	1
price_per_bed	0.33	0.33	0.49	-0.6	-0.31	-0.34	0.77	1	0.69
price_per_sqft	0.43	0.43	0.54	-0.41	-0.3	-0.47	1	0.77	0.71
sqft	-0.13	-0.13	0.38	0.79	0.7	1	-0.47	-0.34	-0.24
baths	-0.12	-0.12	0.31	0.63	1	0.7	-0.3	-0.31	-0.56
beds	-0.09	-0.09	0.27	1	0.63	0.79	-0.41	-0.6	-0.28
price	0.35	0.34	1	0.27	0.31	0.38	0.54	0.49	0.55
year	1	1	0.34	-0.09	-0.12	-0.13	0.43	0.33	0.38
date	1	1	0.35	-0.09	-0.12	-0.13	0.43	0.33	0.38

Correlation plot: The correlation plot [above] helps one understand how each feature impacts other features in the data set. From this correlation plot it is clear that multiple features correlate with each strongly in positive and negative ways. The square footage has a high 0.79 correlation with the number of beds. The number of baths and number of beds also have a high positive correlation. The price is also heavily correlated with square footage, beds, and baths. There are also strong negative correlations between the price per bed, bath, and square foot, with their respective elements. However, they all have strong positive correlations with each other.

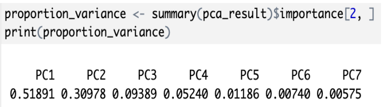
PCA: For finding Principal components we used all seven of the numeric fields in the data set scaled. We first standardized the data, then performed PCA, and viewed the PCA loadings. The PCA loadings model the relationship between original variables and principal components. This will help understand how each feature contributes to each principal component. To visualize how each principal component explains the variance of the 7 variables we used a Scree plot. It shows PC1 explains 51.9%, and PC2 31% of the variance. To visualize the contribution each variable had to the construction of the PCs we created 2 bar plots. For PC1, 3 variables had a positive contribution while the rest had a negative effect. For PC2 the loadings show all variables had a negative contribution, with the price variable having a very large negative effect.



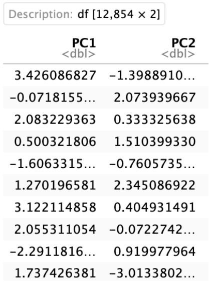




We calculated the proportion of variance for each PC tells you how much of the variability in the data is explained by each component

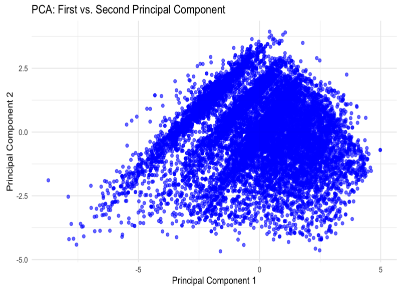


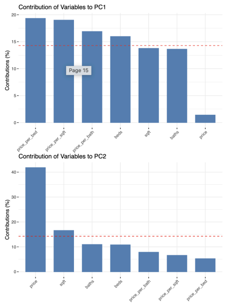
```
proportion_variance <- summary(pca_result)$importance[2, ]
print(proportion_variance)

    PC1     PC2     PC3     PC4     PC5     PC6     PC7 
0.51891 0.30978 0.09389 0.05240 0.01186 0.00740 0.00575 
```



Description: df [12,854 × 2]

PC1 <dbl>	PC2 <dbl>
3.426086827	-1.3988910...
-0.0718155...	2.073939667
2.083229363	0.333325638
0.500321806	1.510399330
-1.6063315...	-0.7605735...
1.270196581	2.345086922
3.122114858	0.404931491
2.055311054	-0.0722742...
-2.2911816...	0.919977964
1.737426381	-3.0133802...

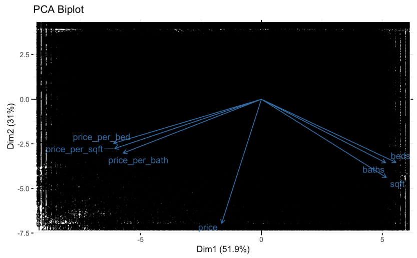




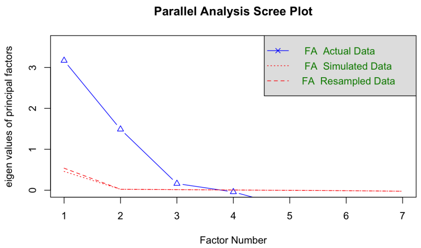
We can see that PC1 explained the most variability as compared to the other PCs.

Contribution of Variables to PCs: This shows how much each variable is contributing to the two PCs. The bar heights indicate how much each variable contributes to the variance captured by PC1 or PC2. A high contribution means that the variable is important for explaining the variation along PC1 or PC2. For PC1, price_per_bed is the most important for explaining the variance captured by PC1, followed by price_per_sqft. For PC2, price is the most important for explaining the variance captured by PC2, with all other variables at least 20% lower in explaining variance.

PCA Biplot: Use the Biplot to visualize both the scores (the transformed data points) and the loadings (the contributions of the original variables) on the same plot. It helps understand how the original variables relate to each principal component and how the data points are distributed in the reduced-dimensional space.



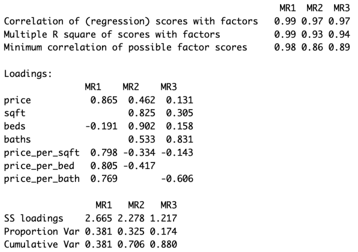
Factor Analysis: We will use the Rent_analysis data frame that has the variables price_per_sqft, price_per_bed, and price_per_bath, price, sqft, beds, and bath.



Determine Optimal Number of Factors: Using fa.parallel and a scree plot we can determine the number of factors to use in a factor analysis by comparing the eigenvalues of the data's correlation matrix with those of randomly generated data. Factors with eigenvalues greater than those from random data are considered meaningful and should be retained. Factors with eigenvalues smaller than those from random data are considered to be noise and are discarded.

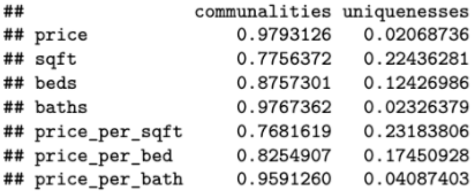
Visualize Factor Analysis Structure: Use fa.diagram to show the factors as MR1/2/3. The arrows show the factor loadings, how much each variable is associated with each factor. We also visualized factor scores over time from 2000 to 2018.

Model Interpretations: The factor loadings tell you how strongly each variable is related to the factors. A higher absolute value indicates a stronger relationship between the variable and the factor. For factor 1, baths have the strongest relationship between the variable and the factor. For factor 2, price is the strongest, and for factor 3, beds are the strongest relationship.

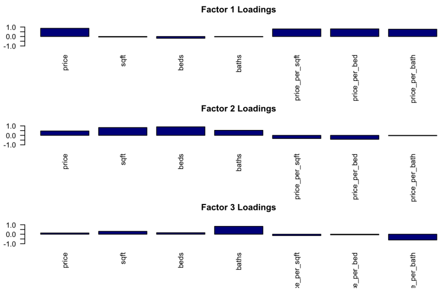


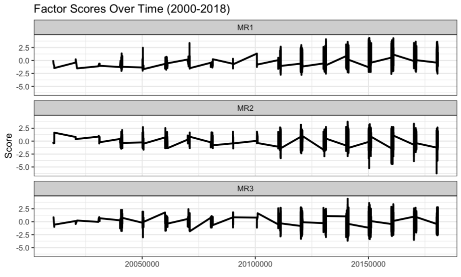
```
                                                  MR1  MR2  MR3
Correlation of (regression) scores with factors   0.99 0.97 0.97
Multiple R square of scores with factors          0.99 0.93 0.94
Minimum correlation of possible factor scores     0.98 0.86 0.89

Loadings:
               MR1    MR2    MR3   
price           0.865  0.462  0.131
sqft                   0.825  0.305
beds           -0.191  0.902  0.158
baths                  0.533  0.831
price_per_sqft  0.798 -0.334 -0.143
price_per_bed   0.805 -0.417       
price_per_bath  0.769        -0.606

                 MR1   MR2   MR3
SS loadings    2.665 2.278 1.217
Proportion Var 0.381 0.325 0.174
Cumulative Var 0.381 0.706 0.880
```

Find Communalities and Uniqenesses: This function calculates the sum of squared factor loadings for each variable across all factors. The result, communalities, represents the total proportion of variance for each variable that is explained by the factors. Uniqueness is also calculated, which is the amount of variance that is not explained by the factors. Uniquenesses are calculated as 1 - communalities. This value indicates how much of each variable's variance is left unexplained after the factors are extracted.

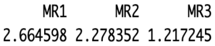


```
##                communalities uniquenesses
## price              0.9793126   0.02068736
## sqft               0.7756372   0.22436281
## beds               0.8757301   0.12426986
## baths              0.9767362   0.02326379
## price_per_sqft     0.7681619   0.23183806
## price_per_bed      0.8254907   0.17450928
## price_per_bath     0.9591260   0.04087403
```

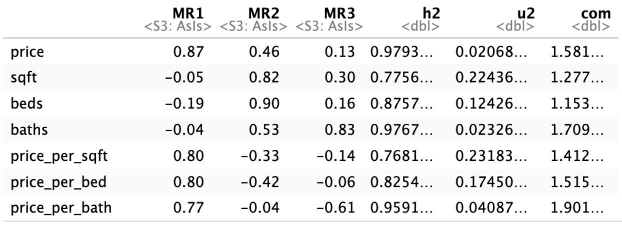




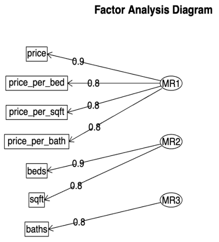
Evaluate Variance with Sum of Squared Loadings: The first factor explains 2.664598 units of variance in the dataset. This is the most important factor and explains the largest portion of the total variance. The second factor explains 2.278352 units of variance. This is the second most important factor, but it explains slightly less variance than the first factor. The third factor explains 1.217245 units of variance. This factor explains less variance compared to the first two.



```
     MR1      MR2      MR3 
2.664598 2.278352 1.217245 
```

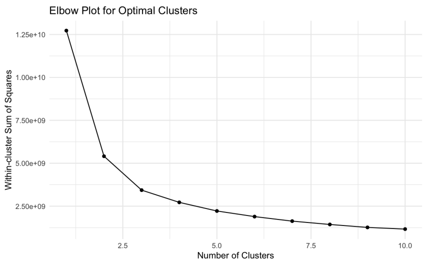


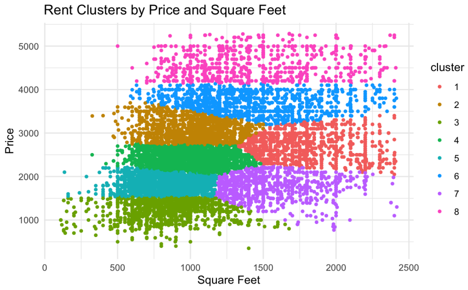
	MR1	MR2	MR3	h2	u2	com
	<S3: AsIs>	<S3: AsIs>	<S3: AsIs>	<dbl>	<dbl>	<dbl>
price	0.87	0.46	0.13	0.9793…	0.02068…	1.581…
sqft	−0.05	0.82	0.30	0.7756…	0.22436…	1.277…
beds	−0.19	0.90	0.16	0.8757…	0.12426…	1.153…
baths	−0.04	0.53	0.83	0.9767…	0.02326…	1.709…
price_per_sqft	0.80	−0.33	−0.14	0.7681…	0.23183…	1.412…
price_per_bed	0.80	−0.42	−0.06	0.8254…	0.17450…	1.515…
price_per_bath	0.77	−0.04	−0.61	0.9591…	0.04087…	1.901…



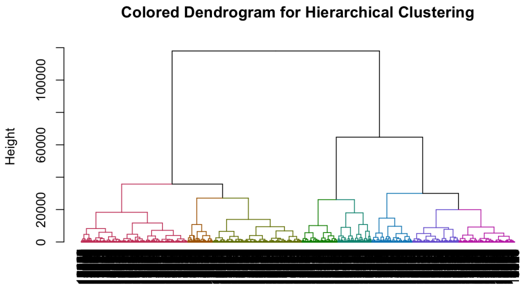
Clustering Methods

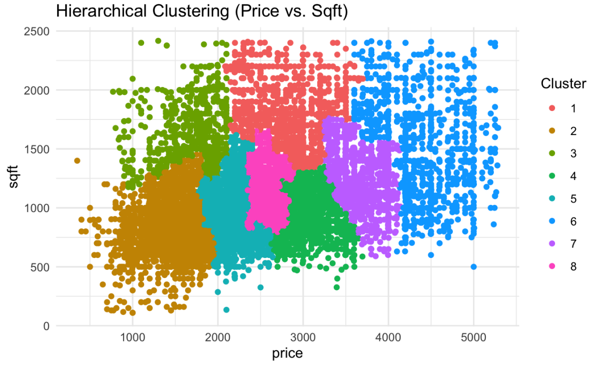
Clustering: Price and sqft : To start off the clustering we clustered price and square footage. We used the elbow plot to find the ideal number of clusters, which was eight. The clustering returned very good clusters which represents the luxury of the properties. The properties in the top right of the graph represent large and expensive. The properties in the bottom left represent small and cheap. The clusters are organized in a similar way, as they change from one to another from left to right and down to up.



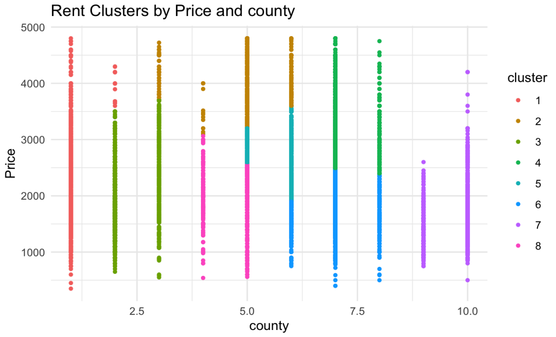


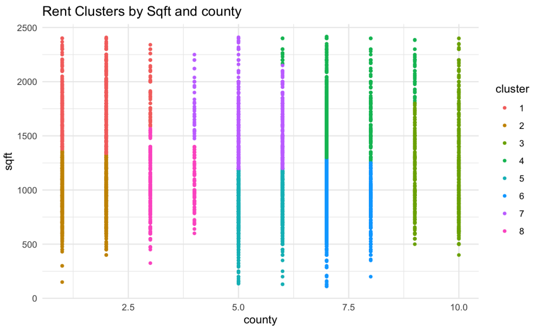
Clustering: Hierarchical (price vs sqft) Another clustering method we used was hierarchical clustering. This started with one cluster and organized each observation into 8 different clusters. The colored dendrogram shows the clustering process. The groups continued to be defined at each stage in the hierarchical tree, until all observations belonged to one of the 8 clusters. To investigate the clusters further we created a scatter plot to visualize each group.



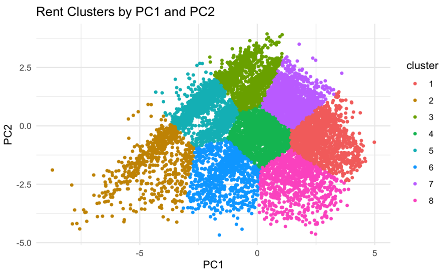


Clustering: Rent and price by encoded counties: To see where properties in each county would be grouped we encoded the counties into dummy variables, and then clustered by price and square footage. It is evident that in each county there are discrepancies for where each observation would be grouped. There are many observations in each county that belong to different clusters. However, it is clear that each county has a trend where the properties belong to a few clusters for each one.

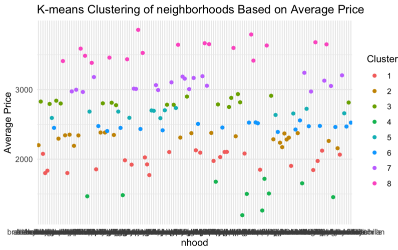


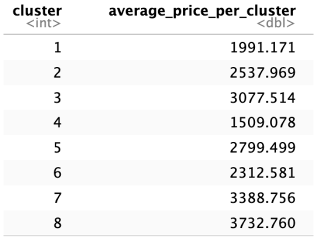


Clustering: PC1 and PC2: In order to see how all the features affect clustering we clustered by the two PCs that explain the variance the most. The output shows 8 very well defined clusters. These clusters can be interpreted as high end properties compared to mid range, and low end properties. With the higher end ones being towards the upper right, and the low end ones towards the lower left.

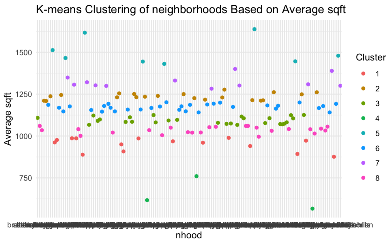


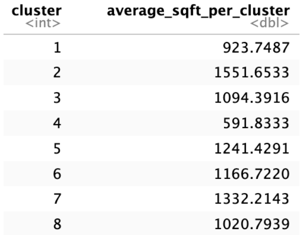
Clustering: Avg price and sqft: We first created two new data frames with each neighborhood and the average price and square footage. To establish the ideal number of clusters we checked with elbow plots. We then clustered each neighborhood by the averages. These clusters were very informative because they showed which neighborhoods contained high average values of price and square footage, with the clusters higher on the graph indicating higher price and square footage for their respective graphs.





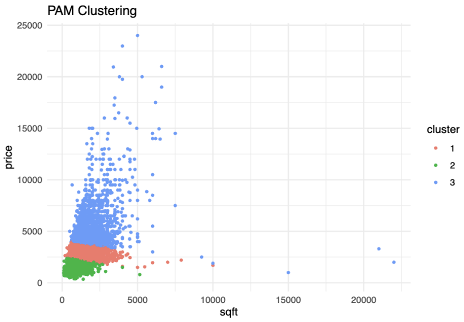
cluster <int>	average_price_per_cluster <dbl>
1	1991.171
2	2537.969
3	3077.514
4	1509.078
5	2799.499
6	2312.581
7	3388.756
8	3732.760



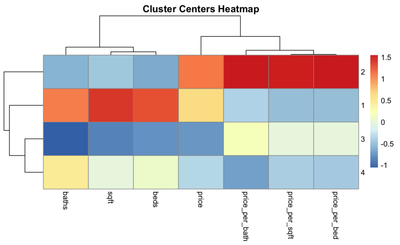


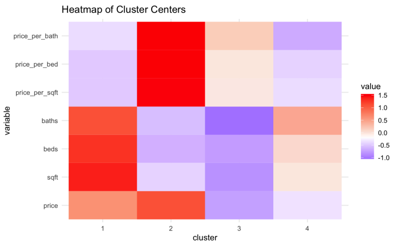
cluster <int>	average_sqft_per_cluster <dbl>
1	923.7487
2	1551.6533
3	1094.3916
4	591.8333
5	1241.4291
6	1166.7220
7	1332.2143
8	1020.7939

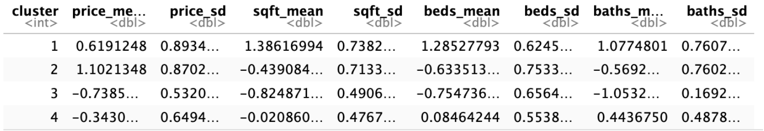
Clustering PAM: This algorithm groups the data by using the median instead of the mean, like how k-means establishes groups. We decided to use PAM to see what would happen if we included the outliers in the clustering. We decided to use a smaller amount of groups to ensure well defined clusters.



Clustering by all seven features scaled: Clustering by all seven of the variables to see how each variable would be represented in the clusters. The heatmap displays how each feature compares to the expected averages for the whole data set. If a shade is darker red for the feature it is higher than its expected mean. If it was darker blue it was less than its expected mean. The darker the color means it was more extreme, higher or lower.







cluster <int>	price_me… <dbl>	price_sd <dbl>	sqft_mean <dbl>	sqft_sd <dbl>	beds_mean <dbl>	beds_sd <dbl>	baths_m… <dbl>	baths_sd <dbl>
1	0.6191248	0.8934…	1.38616994	0.7382…	1.28527793	0.6245…	1.0774801	0.7607…
2	1.1021348	0.8702…	−0.439084…	0.7133…	−0.633513…	0.7533…	−0.5692…	0.7602…
3	−0.7385…	0.5320…	−0.824871…	0.4906…	−0.754736…	0.6564…	−1.0532…	0.1692…
4	−0.3430…	0.6494…	−0.020860…	0.4767…	0.08464244	0.5538…	0.4436750	0.4878…

Citations

1. Session 1 and 2 Lab: Data Preprocessing and Visualization Notebook
2. Session 4 Lab: PCA_CaseStudy_NBA_students Notebook
3. Session 5 Lab: 2024_FA_CaseStudy_InterestRates_students Notebook
4. Session 6 Lab: Clustering_WHO_2024_students Notebook
5. Pennington, Kate (2018). Bay Area Craigslist Rental Housing Posts, 2000-2018. Retrieved from https://github.com/katepennington/historic_bay_area_craigslist_housing_posts/blob/master/clean_2000_2018.csv.zip.