Introduction

This analytic project aims to examine housing rental trends across the different states in the United States with the use of visualization, clustering and regression modelling. By examining key attributes such as the listings' location, pet-friendliness, number of amenities, we can segment the rental market and identify key patterns influencing rental prices.

The US rental housing dataset provided is rich and encompasses key attributes such as price, apartment size, number of bedrooms and bathrooms, location coordinates, pet friendliness, and types of amenities. These variables provide a strong foundation for enabling meaningful insights into US rental trends and factors affecting it. By analyzing these data points, stakeholders will be able to answer the following questions:

- How do apartment types and rental prices (square footage) vary across the different U.S regions?
- How can we segment the rental market into clusters based on price, location, amenities and pet friendliness?
- Which factors (location, amenities, pet policy) have the greatest impact on rental prices?

The insights generated from this project may be valuable for key stakeholders such as landlords and renters, investors and policy makers to better understand the influence of various attributes on rental pricing in the US.

To analyze the apartment types and rental prices in the listings, we can conduct geospatial analysis through use of heatmap for effective visualization. To segment the rental market, we can use K-means clustering to group listings based on different attributes. Additionally, we can use Classification and Regression Tree (CART) to examine which attribute has the greatest impact on influencing rental trends.

While the dataset provided is comprehensive enough for our intended analyses, we would need to transform certain variables of interest to better suit our analytical models. Furthermore, incorporating additional data such as year of listing or safety index of neighborhood will be useful in predicting rental trends and further refining our insights.

Data Understanding and Preparation

The breakdown of the dataset provided is shown in Table 1.

Table 1. Data attributes

S/N	Attribute Name	Data Type	Purpose	Data Quality Issue	Role in analysis
1	id	Nominal	Unique identifier	Not needed	N.A.
2	category	Typeless	Category of classified	Not needed	N.A.
3	title	Typeless	Title text	Not needed	N.A.
4	body	Typeless	Body text	Not needed	N.A.
5	amenities	Typeless	Description of amenities available	No issue	Transform into numeric counts and binned
6	bathrooms	Continuous	Number of bathrooms	Not needed	N.A.
7	bedrooms	Ordinal	Number of bedrooms	Wrongly classified for studio apartment type. Missing value "null"	Used to categorize apartment type
8	currency	Categorical	Currency of rent	Not needed	N.A.
9	fee	Categorical	Fee	Not needed	N.A.
10	has_photo	Categorical	Photo of apartment	Not needed	N.A.
11	pets_allowed	Typeless	Type of pets allowed	Missing value "null"	Transform into nominal
12	price	Continuous	Rental price of apartment	Extreme outlier	-
13	price_display	Typeless	Display of rental	Not needed	N.A.
14	price_type	Categorical	Monthly or weekly rent	Not needed	N.A.
15	square_feet	Continuous	Apartment size in square footage	Outlier - doesn't match apartment type	Used to calculate square footage rental
16	address	Typeless	Location (street) of apartment	Not needed	N.A.
17	cityname	Categorical	Location (city) of apartment	Not needed	N.A.
18	state	Categorical	Location (state) of apartment	Missing value "null"	Will be categorized into regions
19	latitude	Continuous	Latitude coordinate of apartment	Missing value "null"	Plot heatmap

20	longitude	Continuous	Longitude coordinate of apartment	Missing value "null"	Plot heatmap
21	source	Categorical	Origin of listing	Not needed	-
22	time	Nominal	Time listing created	Not needed	-

We will treat the data issues mentioned in Table 1 as follow:

- Using Excel, filter missing number of bedrooms. Determine if information is provided (i.e. studio/one BR) under title or body text. If unable to determine, we will omit these data points.
- Studio listings have number of bedrooms classified as 0, 1 or 2. To standardize all studio listings to 0 bedrooms, we will use grep() in R to identify studio under title description and replace wrongly classified bedrooms. Following which, check for outlier (i.e. title containing studio and two BR etc.) or wrongly replaced values. Manually correct these in excel.
- For "null" in pets_allowed, we will treat it as as none i.e. no pets allowed
- Using Excel, filter outlier data points in price and square_feet, determine if information is provided in title or body. For square_feet outlier, use median of same apartment type within the same state.
- Fill in missing data (null) for state, latitude and longitude based on address in listing description.

To further enhance our treated dataset for analysis, we will create new variables by transforming the existing data:

- price_sq_ft: To normalize rental prices across apartment sizes, we will examine price per square foot. Variable will be used as predictor in K-means and target in CART regression.
- pets_cat: Categorize whether apartments are pet-friendly 0: None/null, 1: Pets allowed for CART regression
- pets: Using Excel, bin pets_allowed into 0: null/none, 1: dogs or cats only, 2: both dogs and cats allowed for K-means clustering
- amenity_count: Count number of amenities for each listing using R
- amenity_bin: Binning amenity_count into 0: Null/none, 1: 1 to 4 amenities, 2: >=5 amenities for predictor in K-means clustering

- region: Group the 50 states and DC Washington into four geographical regions of West, Midwest, South, Northeast for meaningful visualization
- region_Midwest, region_West, region_South, region_Northeast: One hot encoding of region for K-means clustering
- apartment_type: Categorize 0 bedrooms as “studio”, 1 bedroom as “1-BR”, 2 bedrooms as “2-BR” and >=3 bedrooms as “3+ BR” for visualization

Table 2. Example of untreated dataset with data quality issues (yellow)

title	body	amenities	bathrooms	bedrooms	pets_allowed	price	square_feet	state	latitude	longitude
One BR Leewa	This unit is located at Leeward	null	1	1	None	525	200	null	null	null
One BR Mullic	This unit is located at Mullica	Pool	1	1	None	750	219	null	null	null
One BR Se Ash	This unit is located at Se Ash S	null	1	1	None	850	400	null	null	null
One BR 2530 N	This unit is located at 2530 M	null	1	1	Cats,Dogs	705	464	null	39.8163	-98.5576
One BR 1260 H	This unit is located at 1260 Ho	Parking,Refriger	1	1	Cats,Dogs	2295	500	null	39.8163	-98.5576
Studio apartm	This unit is located at 178-60	Elevator,Parkin	1	2	None	1599	400	null	39.8163	-98.5576
Studio apartm	This unit is located at 545 Ged	null	1	1	None	950	200	CA	38.1172	-122.2313
Studio Cottage	New Bern Studio includes : 1 b	AC,Basketball,C	1	1	Cats,Dogs	1560	200	NC	35.0847	-77.0609
A-P-T Suites La	A-P-T Suites is your next Exter	Cable or Satelli	null		Cats,Dogs	275	300	FL	28.0451	-81.9689
One BR in New	Monthly Rent$4,605 -to $4,79	Basketball,Cabl	null	1	null	4790	40000	NY	40.7716	-73.9876
Studio apartm	Barstow Its 14/18ft. studio ap	AC,Cable or Sat	1	0	null	52500	1418	CA	34.887	-117.035
5115 N 40th St	all utilities included avail 12/2	Cable or Satelli	1		null	849	405	AZ	33.4993	-111.9838
bedroom in M	Medford Walk-In Store Front (	null	1		None	1200	550	MA	42.4194	-71.111

Table 3. Treated dataset with resolved data (orange) and omitted data (grey)

title	body	amenities	bathrooms	bedrooms	pets_allowed	price	square_feet	state	latitude	longitude
One BR Leewa	This unit is located at Leeward	null	1	1	None	525	200	FL	30.08297356	-81.701686
One BR Mullic	This unit is located at Mullica	Pool	1	1	None	750	219	NJ	39.54393385	-74.662663
One BR Se Ash	This unit is located at Se Ash S	null	1	1	None	850	400	OR	45.52164323	-122.60325
One BR 2530 N	This unit is located at 2530 M	null	1	1	Cats,Dogs	705	464	VA	37.4382188	-77.437188
One BR 1260 H	This unit is located at 1260 Ho	Parking,Refriger	1	1	Cats,Dogs	2295	500	CA	37.87771902	-122.28843
Studio apartm	This unit is located at 178-60	Elevator,Parkin	1	0	None	1599	400	NY	40.71359541	-73.783613
Studio apartm	This unit is located at 545 Ged	null	1	0	None	950	200	CA	38.1172	-122.2313
Studio Cottage	New Bern Studio includes : 1 b	AC,Basketball,C	1	0	Cats,Dogs	1560	200	NC	35.0847	-77.0609
A-P-T Suites La	A-P-T Suites is your next Exter	Cable or Satelli	null		Cats,Dogs	275	300	FL	28.0451	-81.9689
One BR in New	Monthly Rent$4,605 -to $4,79	Basketball,Cabl	null	1	null	4790	800	NY	40.7716	-73.9876
Studio apartm	Barstow Its 14/18ft. studio ap	AC,Cable or Sat	1	0	null	500	252	CA	34.887	-117.035
5115 N 40th St	all utilities included avail 12/2	Cable or Satelli	1		null	849	405	AZ	33.4993	-111.9838
bedroom in M	Medford Walk-In Store Front (	null	1	0	None	1200	550	MA	42.4194	-71.111

Table 4. Example of newly transformed variables (orange)

pets	pets_cat	price_sq_ft	latitude	longitude	amenity_count	amenity_bin	apartment_category	region	region_Midwest	region_Northeast	region_South	region_West
0	0	12.9906542	38.891	-77.0816	0	0	Studio	South	0	0	1	0
0	0	7.97413793	47.616	-122.328	0	0	Studio	West	0	0	0	1
0	0	6.14143921	40.7629	-73.9885	5	2	Studio	Northeast	0	1	0	0
0	0	10.8333333	37.7599	-122.438	1	1	Studio	West	0	0	0	1
0	0	8.92105263	37.7599	-122.438	1	1	Studio	West	0	0	0	1
2	1	7.8	35.0847	-77.0609	8	2	Studio	South	0	0	1	0
2	1	7.8	35.096	-77.0272	8	2	Studio	South	0	0	1	0
0	0	5	30.0871	-95.4685	0	0	1-BR	South	0	0	1	0
0	0	4.75	38.1172	-122.231	0	0	Studio	West	0	0	0	1
0	0	3.125	33.9649	-84.5107	1	1	1-BR	South	0	0	1	0
0	0	3	35.2016	-80.8124	0	0	1-BR	South	0	0	1	0

Data analysis

Data analysis will be performed using R (visualizations) and SPSS modeler (clustering and CART).

Geospatial patterns

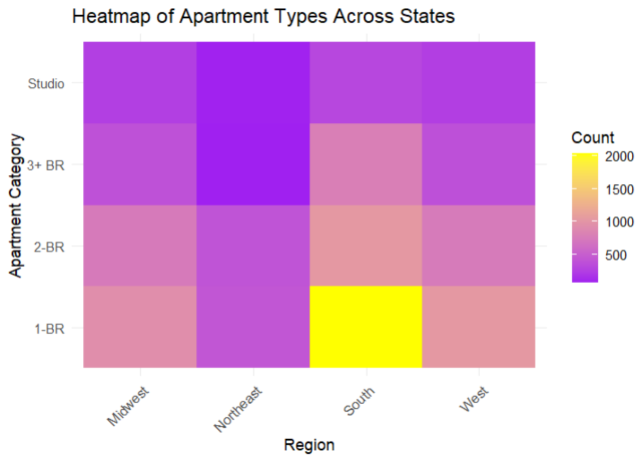


Figure 1. Heatmap of apartment types across U.S regions

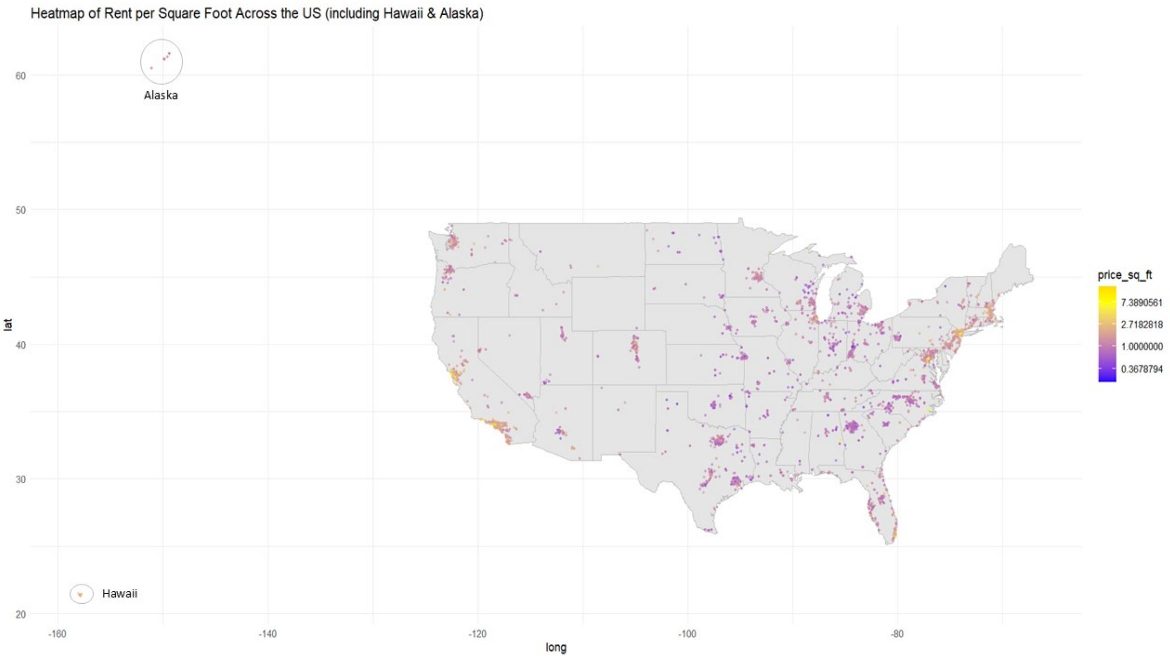


Figure 2. Heatmap of rental prices across US (including Hawaii and Alaska)

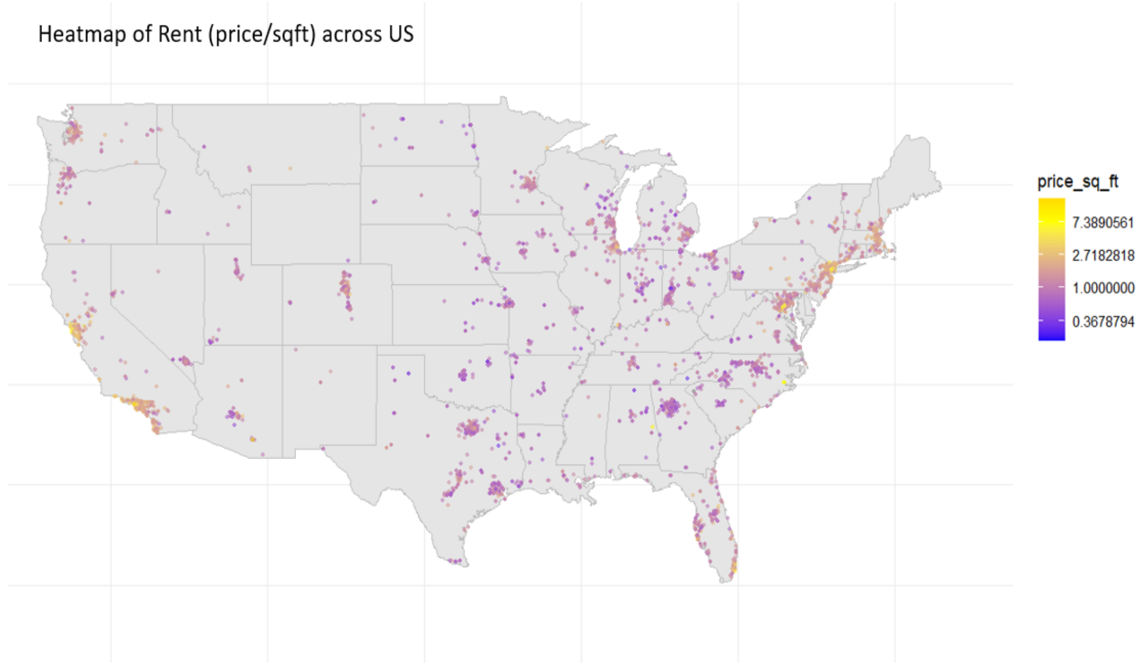


Figure 3. Zoomed in US mainland heatmap

In Figure 1, we observe that the distribution of apartment types differs across the 4 regions, with Northeastern states having the least number of listings and South having the most. To visualize the square footage rental price across the different US states, we will plot a heatmap in R using price_sq_ft, longitude and latitude. From the heatmap (Figure 2 & 3), we observe that higher rent prices are concentrated in the Western (including Hawaii) and Northeastern states. Listings with lower rental prices are mainly situated in the Midwestern and Southern states with the exclusion of Florida. Taken together, these heatmaps allow us to have an effective visual comparison of the types of apartments and rental trends across the different regions.

Rental market segmentation

To segment the rental housing market, we can use unsupervised K-means clustering in SPSS modeler to identify patterns and group similar listings based on multiple attributes inputted such as square footage rent, amenities, pet-friendliness, and location. We will set the number of clusters to 3.

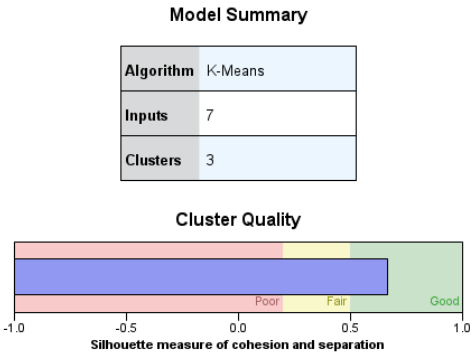
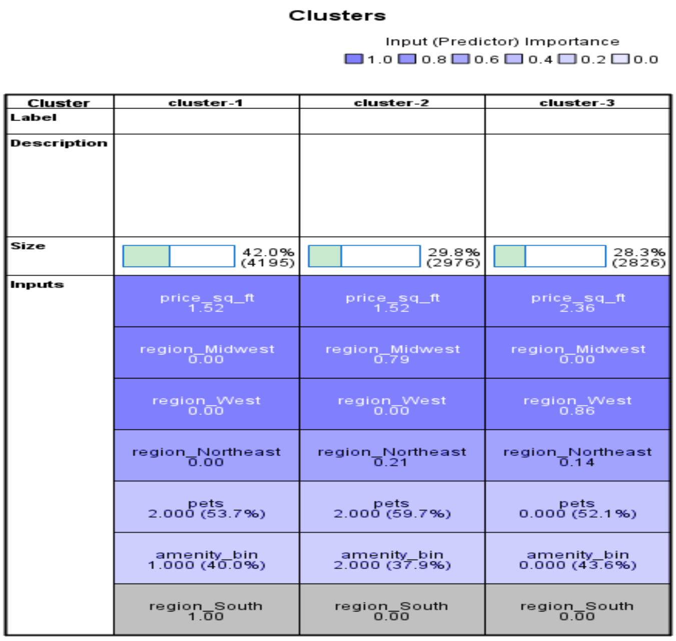


Figure 4. Model summary indicating silhouette score of 0.7 indicating strong cluster separation



Cluster	cluster-1	cluster-2	cluster-3
Label			
Description			
Size	42.0% (4195)	29.8% (2976)	28.3% (2826)
Inputs	price_sq_ft 1.52	price_sq_ft 1.52	price_sq_ft 2.36
	region_Midwest 0.00	region_Midwest 0.79	region_Midwest 0.00
	region_West 0.00	region_West 0.00	region_West 0.86
	region_Northeast 0.00	region_Northeast 0.21	region_Northeast 0.14
	pets 2.000 (53.7%)	pets 2.000 (59.7%)	pets 0.000 (52.1%)
	amenity_bin 1.000 (40.0%)	amenity_bin 2.000 (37.9%)	amenity_bin 0.000 (43.6%)
	region_South 1.00	region_South 0.00	region_South 0.00

Figure 5. K-means clusters

From Figure 5, we observe differences in pricing, regions, amenities and pet-friendliness across the three clusters. Cluster 1 and 2 are priced the same (1.52 psf), however, Cluster 2 has more amenities provided and has a higher number of pet-friendly listings. The two clusters also differ in terms of region – Cluster 1 are all in the South whereas Cluster 2 are predominantly in Midwest with some in Northeast. In contrast, Cluster 3 is priced the highest (2.36 psf) with majority of listings having no amenities (43.6%) provided and are not pet-friendly (52.1%).

Impact of location, pet-friendliness, amenities

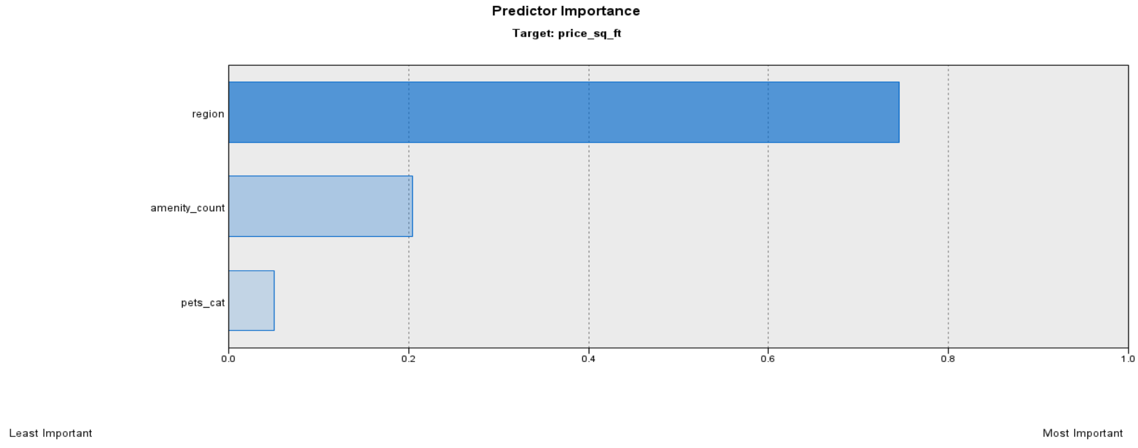


Figure 6. Predictor importance for CART.

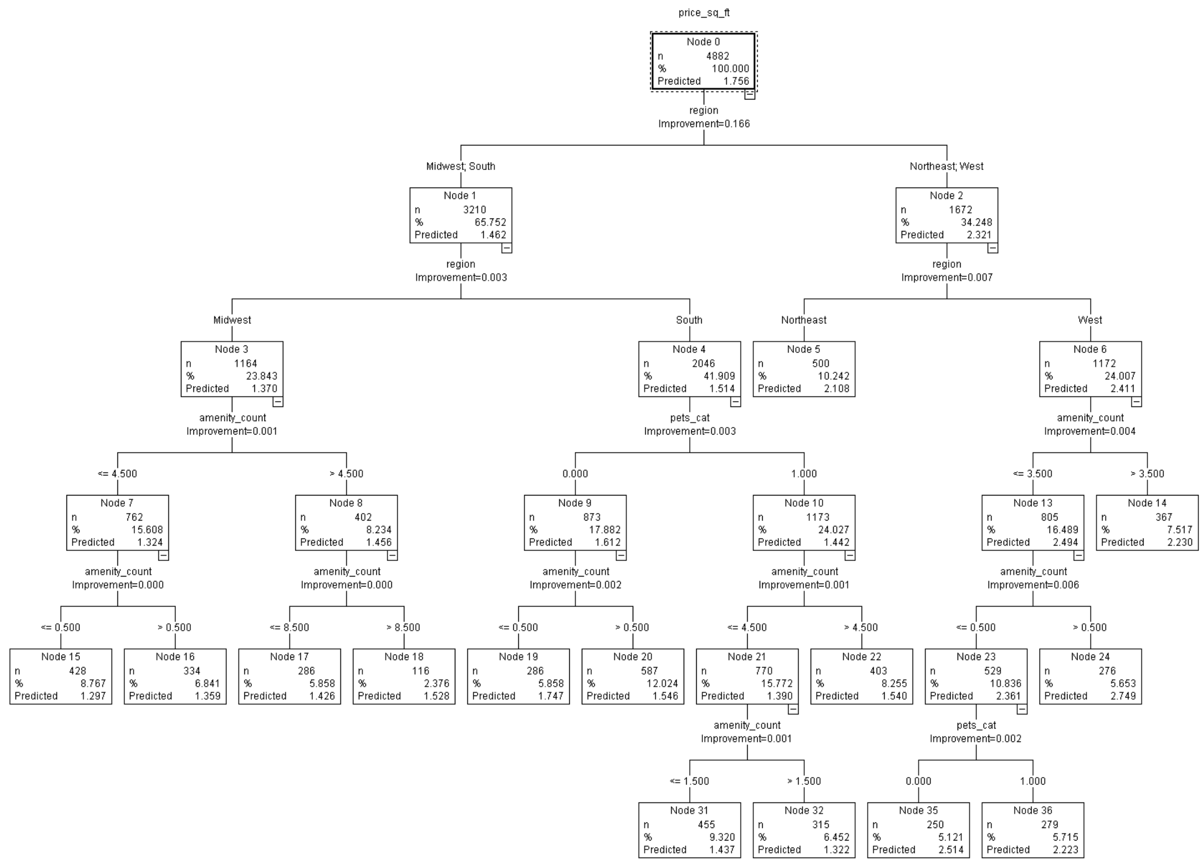


Figure 7. CART results

To further determine the impact of location, pet-policy and amenities on rent prices, we will perform CART modelling using price_sq_ft as target, and region, pets_cat and amenity_count as predictor inputs. We partitioned the data set into 70% training and 30% testing and used default seed setting. In our CART model, region is the most important predictor followed by number of amenities and pet-friendliness (Figure 7). While number of amenities is used to split the tree nodes for every region, pet-friendliness is only involved in splitting the nodes in South (Node 4) and in the West (Node 6). In the Midwest, listings with more amenities are priced higher (Node 7 and 8). On the contrary, in the South and West – listings with lesser amenities are priced higher.

Evaluation and Discussion

Based on our analyses, a key pattern driving differences in rent prices is the location, with the highest rent found in West followed by the Northeast. This aligns with findings by Boeing and Waddell (2016) who reported higher rent prices being concentrated along the Californian coast and the Boston-Washington corridor.

Based on our CART model (Figure 7), we found it interesting that listings that do not allow pets have higher rents in comparison to pet-friendly ones. While this may seem advantageous for pet owners, the reality with pet-friendly listings is more nuanced. A Texan study by Applebaum et al. (2021) had shown that pet-friendly listings often have pet fees, suggesting that our analysis may not be reflective of actual market situation, as these extra charges are often not indicated in the listing prices, undermining the true rental cost for a pet owner. These findings have important implications on renters who are pet owners, especially those seeking to rent in the South and West where hidden costs may be a challenge, in addition to the high rent.

The CART model (Figure 7) also shows that in the Midwest, listings with more amenities are priced higher, aligning with our expectations. Yet interestingly, the reverse is true for listings in the South and West where listings with fewer amenities were being priced higher. This discrepancy may be due to differing economic conditions (Harvard Joint Center for Housing Studies, 2022) and demand differences across the various regions. Rental demand may be lower in the Midwest – urging landlords to enhance listings with amenities to attract renters. Conversely, listings in the South and West may be situated in highly desirables neighborhoods where demand is strong hence reducing the need for landlords to justify rent prices with many amenities. This insight suggests that landlords should consider both location and amenities in order to achieve strategic pricings for their rentals.

Table 5. Evaluation metrics for CART model

'Partition'	1_Training	2_Testing
Minimum Error	-2.153	-2.067
Maximum Error	11.244	8.085
Mean Error	-0.004	-0.001
Mean Absolute Error	0.615	0.61
Standard Deviation	0.914	0.897
Linear Correlation	0.43	0.442
Occurrences	6,969	3,028

One limitation of our CART model is that our predictors only have moderate predictive power in predicting rent prices – as indicated by linear correlation of 0.44 (Table 5), this suggests that other variables are involved in explaining the price variance. Another limitation is using the absolute count of amenities as predictor in CART as this assumes linearity between number of amenities and price. However, some amenities may be more highly valued than others, resulting in an unequal impact on price. For instance, Zillow (2024) revealed that off-street parking and in-unit laundry were in highest demand as compared to other amenities. These limitations underlined the need to adopt a more nuanced modelling approach by including additional variables and weighing amenities importance based on actual demand to improve predictive accuracy.

Conclusion

In conclusion, our analyses have revealed location, amenities and pet-friendliness of listings to be key drivers of rent prices across the US. While rent prices are highest in the West and Northeast; this could be attributed to high demands, limited supply and popularity of the area. In contrast, the Midwest, with the lowest rent price amongst the regions, sees an increase in rent prices with more amenities. Our analysis of pet-friendly listings also suggests hidden costs such as pet fees which may affect overall rent prices. Findings from this project have implications on stakeholders like renters and landlords, underlining the importance of strategic rental pricing based on location and listings' features.

References

Applebaum, J. W., Horecka, K., Loney, L., & Graham, T. M. (2021). Pet-Friendly for Whom? An Analysis of Pet Fees in Texas Rental Housing. *Frontiers in veterinary science*, *8*, 767149. https://doi.org/10.3389/fvets.2021.767149

Boeing, G., & Waddell, P. (2017). New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings. *Journal of Planning Education and Research*, 37(4), 457-476. https://doi.org/10.1177/0739456X16664789

Harvard Joint Center for Housing Studies. (2022). *America's rental housing 2022*. Harvard Graduate School of Design & Harvard Kennedy School. https://www.jchs.harvard.edu/americas-rental-housing-2022

Zillow. (2024). Renters are Looking for Perks Like Pet Areas and Happy Hours Over Gyms and Pools. Zillow. https://www.zillow.com/research/listing-features-rent-34408/

Appendix

```r
# Read CSV file with semicolon as separator
drent <- read.csv("C:/Users/lynnett/Downloads/apartments_for_rent_classified_10K.csv",
                  sep = ";", stringsAsFactors = FALSE)
```

R code 1. Tidying CSV file

```r
# Correct bedrooms for listings that mention "studio" in the title
drent <- drent %>%
  mutate(bedrooms =
           ifelse(grepl("studio", title, ignore.case = TRUE), 0, bedrooms))
```

R code 2. Correcting bedrooms for studio apartments

```r
# Compute amenity count from amenities
drent <- drent %>%
  mutate(amenity_count = ifelse(is.na(amenities) |
                                  amenities == "", 0,
                                str_count(amenities, ",") + 1))
```

R code 3. Counting amenities from description

```r
#Heatmap of rent across US
    ggplot() +

    geom_polygon(data = usa_map, aes(x = long, y = lat, group = group),
                 fill = "gray90", color = "gray", size = 0.1) +


    geom_point(data = drent, aes(x = longitude, y = latitude, color = price_sq_ft),
               alpha = 0.5, size = 0.8) +


    scale_color_gradient2(low = "blue", mid = "yellow", high = "red",
                          midpoint = 2, limits = c(0.2, 15), trans="log") +
    labs(title = "Heatmap of Rent per Square Foot Across the US (including Hawaii & Alaska)") +
    theme_minimal() +
    theme(legend.position = "right")
```

R code 4. Plotting geospatial heatmap

Field	Measurement	Values	Missing	Check	Role
pets	Nominal	0.0,1.0,2....		None	Input
price_sq_ft	Continuous	[0.194218...		None	Input
amenity_bin	Nominal	0.0,1.0,2.0		None	Input
region_Midw...	Continuous	[0.0,1.0]		None	Input
region_North...	Continuous	[0.0,1.0]		None	Input
region_South	Continuous	[0.0,1.0]		None	Input
region_West	Continuous	[0.0,1.0]		None	Input

Table 1. Variables used in K-means clustering

Field	Measurement	Values	Missing	Check	Role
pets_cat	Flag	1.0/0.0		None	Input
price_sq_ft	Continuous	[0.194218...		None	Target
amenity_count	Continuous	[0.0,18.0]		None	Input
region	Nominal	Midwest,...		None	Input

Table 2. Variables used in CART modelling

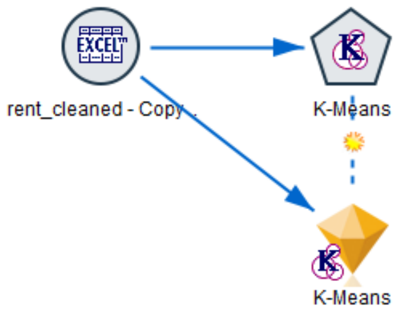


Figure 1. K-means SPSS stream

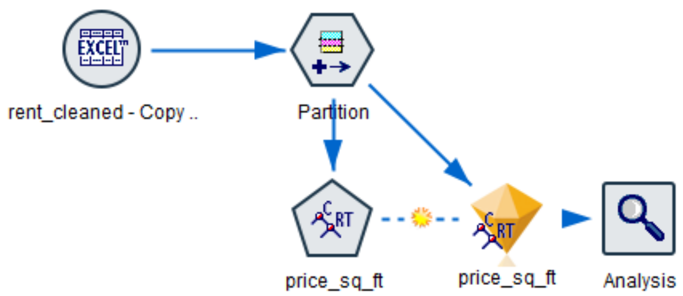


Figure 2. CART SPSS stream