Income Inequality and Housing Burden in California (2010–2023)

Ryan Beavers
University of Oklahoma
LIS 5623 – Advanced Data Analytics
Dr. Glenn Hansen

Table of Contents

Introduction 2
Results: Direct Answer to the Research Question 2
Dataset Overview 3
Exploratory Data Analysis 3
Median Income Trends (2010–2023) 3
Renter Cost Burden Trends 4
Owner Cost Burden Trends 5
Renter & Owner Ability to Pay 6
Correlation and Multivariate Patterns (2023 Focus) 8
Predictive Modeling Model Viability and Predictive Insights 9
Renter Burden Models 10
Owner Burden Models 11
K-Means Clustering: Renters 12
K-Means Clustering: Owners 13
Conclusions 14
Recommendations 15
Appendix – All Code 16

Introduction

This report presents a comprehensive statistical analysis of housing affordability trends across California counties between 2010 and 2023. Drawing on public microdata and cost-burden estimates, I examine the relationships between income, rent burden, ownership burden, and derived metrics of ability-to-pay. I explore patterns among renters and homeowners separately, model the drivers of housing cost burdens, and identify structural clusters of affordability and risk using unsupervised learning.

The overarching question guiding this analysis is: **how does income inequality manifest through housing cost burdens, and what can it reveal about structural disparities across California?**

Results: Direct Answer to the Research Question

Income inequality manifests in California through persistent and widening disparities in housing cost burdens. High-income counties have shown substantial increases in median income, but those increases have not corresponded to lower housing burdens. In fact, some of the highest-income counties still display renter burdens exceeding 50%, indicating that income alone is not a sufficient buffer against housing cost pressure.

My analysis shows that in counties with lower median income, renters and owners alike face disproportionately high burdens relative to their ability to pay. These structural disparities are more pronounced among renters: many low-income counties fall into high-burden, low-affordability clusters. Even counties with similar incomes can show dramatically different burden levels depending on rent-to-income efficiency.

Through regression and clustering analysis, I uncovered that the drivers of cost burden are multi-dimensional. Predictive models explain only a fraction of the variance in burden outcomes, which implies that structural housing supply factors, regional pricing patterns, and historical inequalities may be embedded in these cost burdens. Cluster analysis revealed that some high-income counties still contain structurally vulnerable groups, while others maintain low burden despite moderate incomes.

In short, income inequality in California does not just separate the rich from the poor – it segments the population into structurally distinct affordability regimes. These regimes operate differently for renters and homeowners and cannot be reduced to simple income thresholds.

Dataset Overview

I downloaded the data from https://data.census.gov/ and used tables DP02, DP03, DP04, and DP05, with a particular focus on DP03 and DP04 for economic characteristics and housing estimates. I calculated affordability metrics differently for renters and owners: `renters_ability_to_pay` was computed as (rent * 12) / income, which functions as a burden metric (higher values indicate lower affordability), while `owners_ability_to_pay` was calculated as income / owner housing costs, where higher values indicate better affordability.

The core dataset (`housing_df`) integrates the following key metrics for all counties in California from 2010 to 2023:

- `renter_ge30`: % of renters spending 30% or more of income on rent
- `owner_ge30`: % of owners spending 30% or more of income on housing
- `median_income`: county median household income
- `renters_ability_to_pay`: calculated as (rent * 12) / income (higher = worse affordability)calculated as median_income / median_gross_rent (inverse rent burden)
- `owners_ability_to_pay`: calculated as income / monthly owner costs (higher = better affordability)calculated as median_income / monthly_owner_costs (inverse ownership burden)

After merging, filtering, and cleaning, the dataset comprises 214 county-year records.
I downloaded the data from https://data.census.gov/ and used tables DP02, DP03, DP04, and DP05, with a particular focus on DP03 and DP04 for economic characteristics and housing estimates.
The core dataset (`housing_df`) integrates the following key metrics for all counties in California from 2010 to 2023:

- `renter_ge30`: % of renters spending 30% or more of income on rent
- `owner_ge30`: % of owners spending 30% or more of income on housing
- `median_income`: county median household income
- `renters_ability_to_pay`: inverse rent burden (income / rent)
- `owners_ability_to_pay`: inverse cost burden (income / owner housing costs)

After merging, filtering, and cleaning, the dataset comprises 214 county-year records.

Exploratory Data Analysis

Median Income Trends (2010–2023)

Median income across counties has risen unevenly, with top-quartile counties doubling the income growth of bottom-quartile counties. A log-transformed density histogram (Figure 1) illustrates widening disparities, especially post-2018. This income growth has not translated into increased affordability.

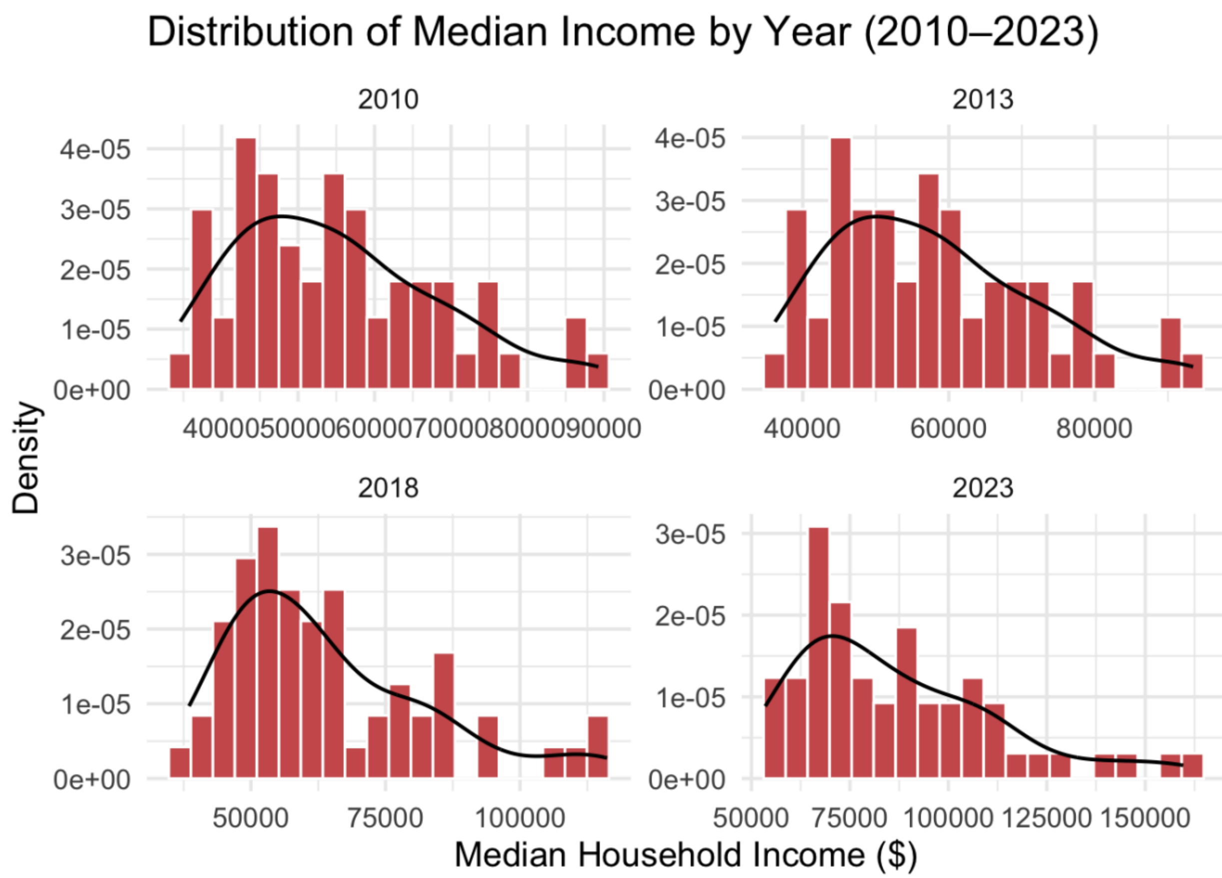


Figure 1: Density + Histogram of Log Median Income by Year

Renter Cost Burden Trends

Renters face consistently high housing burdens. Across all counties and years:

- **Mean renter burden** = 53.6%
- **Max** = 68.5%, **Min** = 6.8%
- Post-2018 shows upward skew, indicating growing strain

Boxplots (Figure 2) show a mild rightward shift in central tendency, but the most revealing insight is the lack of progress in reducing burden.

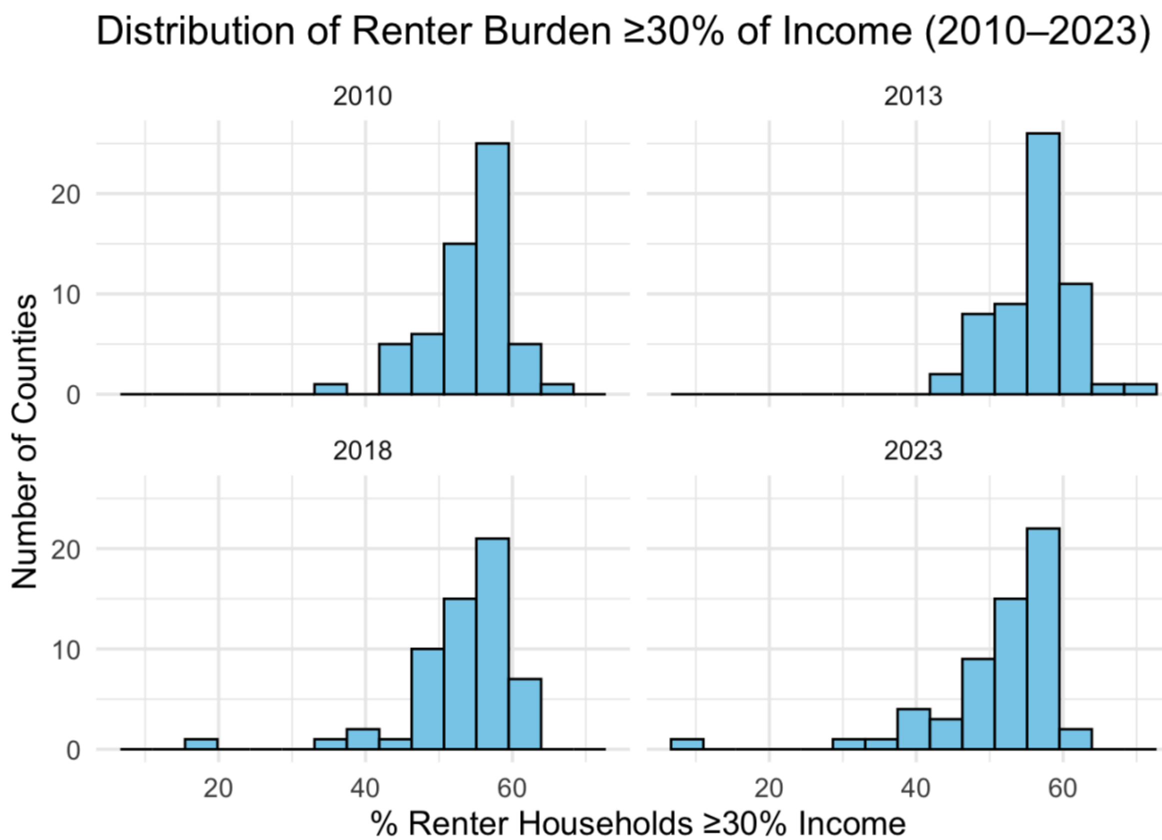


Figure 2: Renter Cost Burden by Year

Owner Cost Burden Trends

Homeowners fare better than renters in aggregate. However, a substantial proportion still face cost burdens >30%. Boxplots of owner burden (Figure 3) suggest less variation across years, with a tighter central distribution.

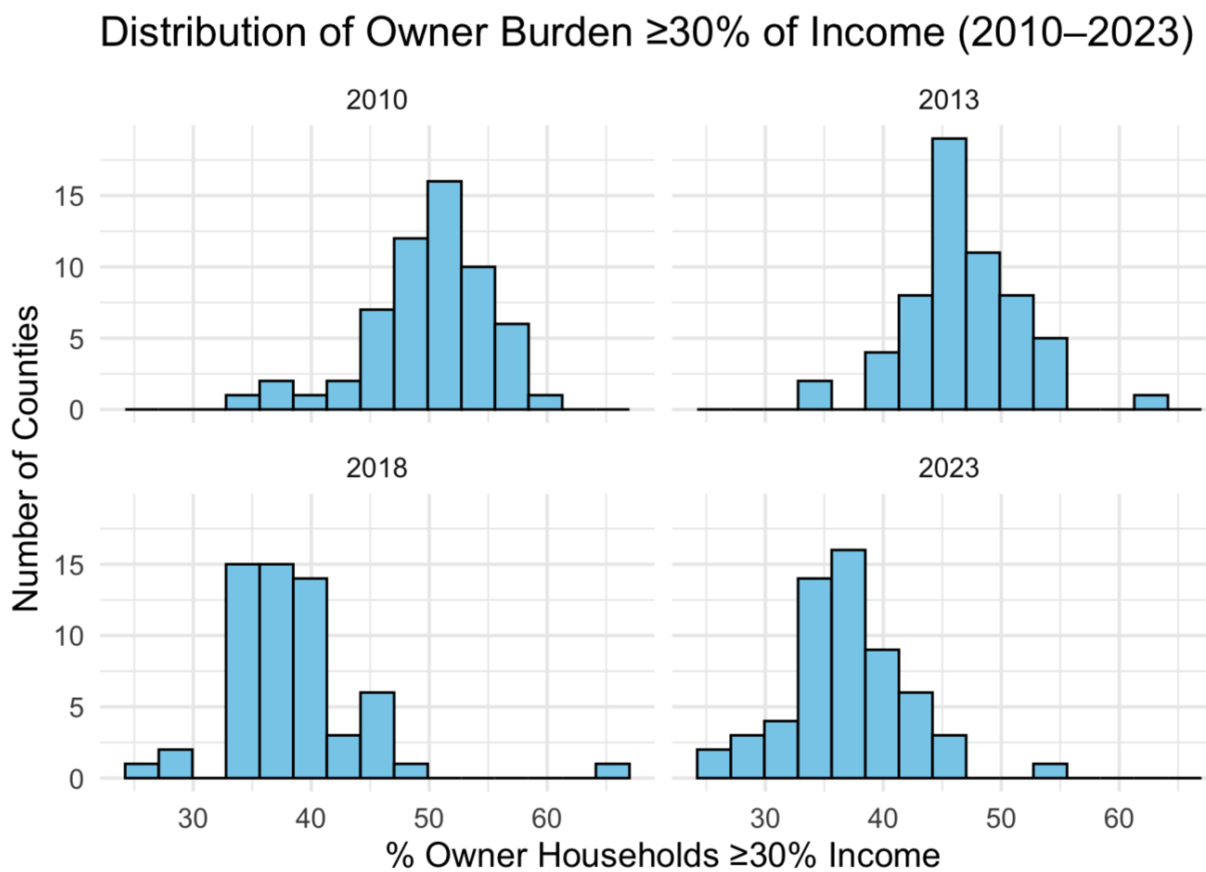


Figure 3: Owner Cost Burden by Year

Renter & Owner Ability to Pay

"Ability to Pay" is an inverse burden metric: higher values indicate more breathing room. For renters:

- **Mean renters_ability_to_pay** = 0.229
- Counties below 0.18 show consistent high burden

For owners:

- **Mean owners_ability_to_pay** = 7.63

  A handful of counties exceed 11.0, indicating structural advantage

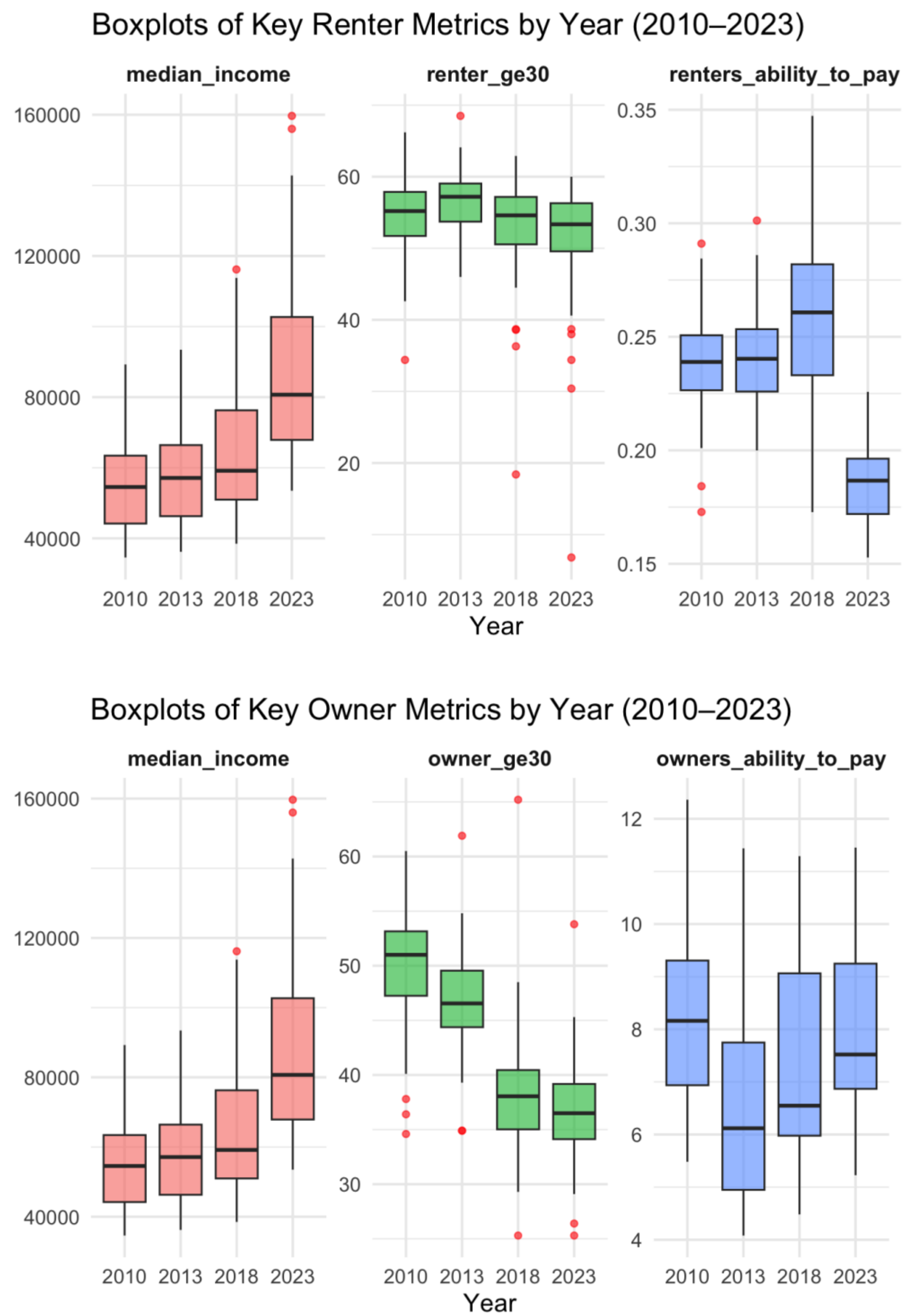


Figure 4: Boxplots of Ability to Pay for Renters and Owners

Correlation and Multivariate Patterns (2023 Focus)

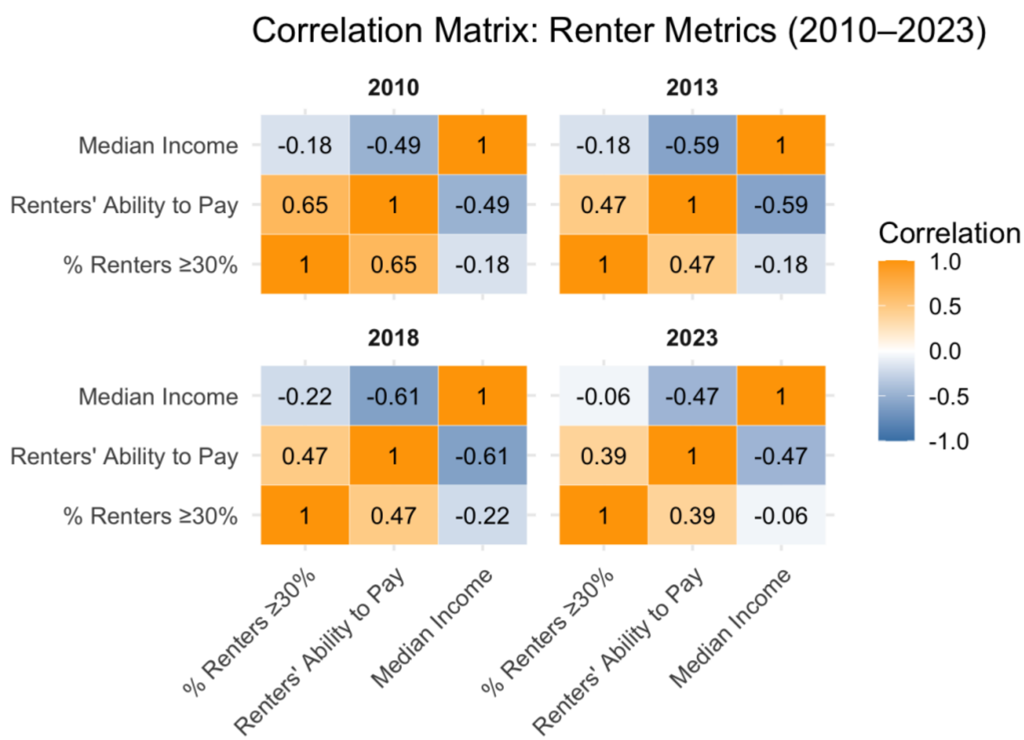
I filtered for 2023 and computed Pearson correlations:

Renters:

- renter_ge30 vs median_income: −.47
- renter_ge30 vs ability_to_pay: −.39
- renter_ge30 vs owner_ge30: 0.27

Owners:

- owner_ge30 vs median_income: −0.11
- owner_ge30 vs ability_to_pay: −0.43



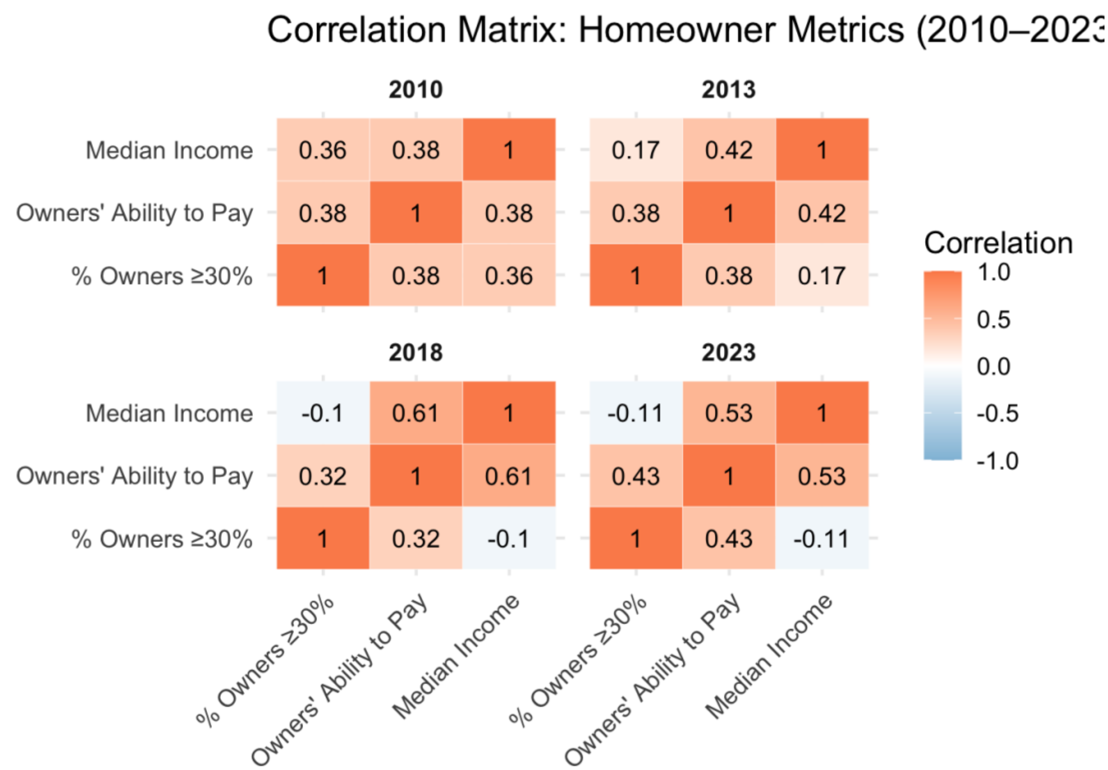


Figure 5: Correlation Heatmaps for Owners and Renters

Conclusion: Income and ability-to-pay are inversely related to burden, but the moderate correlations suggest other latent drivers.

Predictive Modeling

Model Viability and Predictive Insights

I tested three modeling approaches to estimate housing cost burden: linear regression and random forest for both renters and owners. While these models used the same predictors—median income, cost burden ratios, and ability-to-pay—their performance revealed important differences in predictive power and interpretability.

Which model performed best?

The best-performing model, by marginal advantage, was the **Random Forest model for owner cost burden**. It yielded the highest R^2 (0.291) and the lowest RMSE (6.1) among all models. Random forests are better at capturing non-linear interactions and subtle patterns in the data. For owners, the burden dynamics are less chaotic and more explainable through structured predictors like income and affordability ratios.

Which model performed worst?

The **Linear Regression model for owner cost burden** performed the worst, with an R^2 of only 0.234. This suggests that linear models struggle with capturing the true complexity of owner burden dynamics, possibly due to regional variance or hidden variables like mortgage terms, tax treatment, or historical purchase prices that do not scale linearly.

What do the models predict?

Each model predicts the percentage of renters or owners in a given county and year who spend 30% or more of their income on housing. This is a critical policy metric used to identify housing stress. However, the modest R^2 values indicate that these models explain only 23–29% of the variance in housing cost burden, meaning that most of the predictive signal lies in factors not included in the model.

How could these models be improved?

To improve predictive accuracy, I would incorporate additional variables such as:

- **Rental vacancy rates**
- **Housing supply indicators**
- **Zoning or permit restrictions**
- **Commuting costs and job access metrics**
- **Demographic controls (e.g., household size, age)**

Moreover, time-series modeling or hierarchical Bayesian models could help account for temporal shifts and county-level random effects. The addition of real rent or mortgage cost medians instead of estimates could also strengthen the signal. The current predictors do not sufficiently capture supply-side constraints, wealth effects, or credit access, all of which likely drive the remaining unexplained variance.

Renter Burden Models

I trained linear regression and random forest models using median income, renter ability, and owner burden as predictors.

Model	RMSE	R^2
Linear Regression	5.0	0.269
Random Forest	5.4	0.290

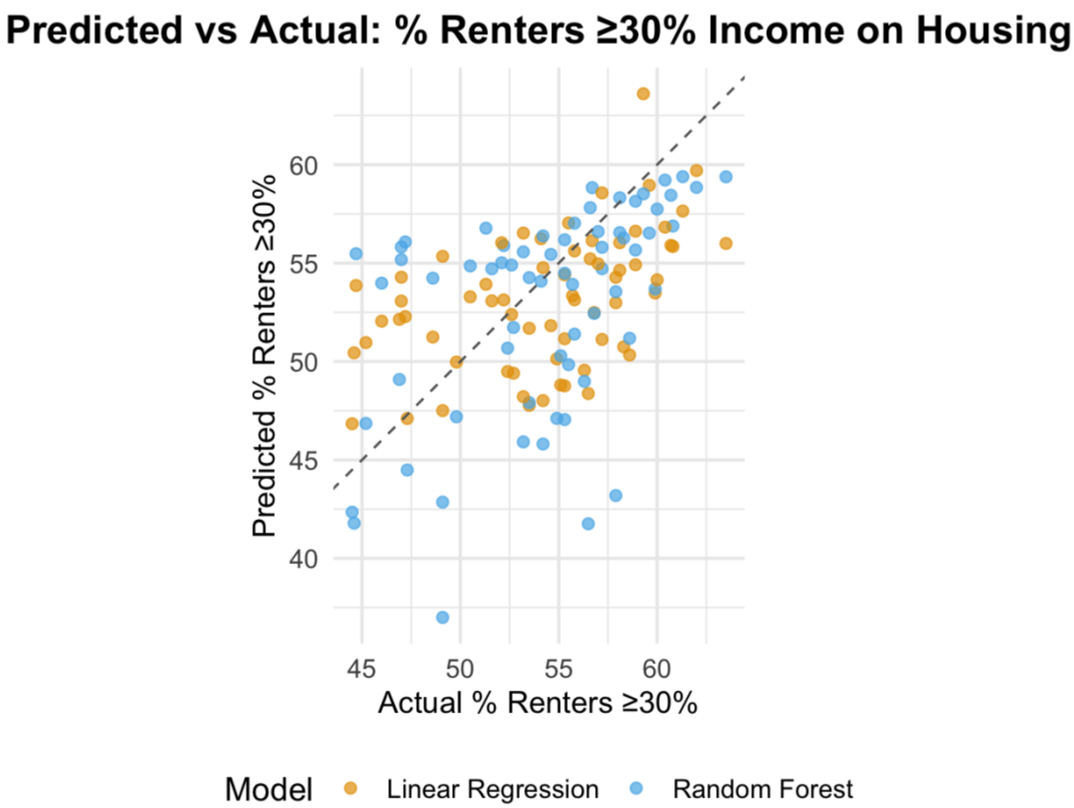


Figure 6: Predicted vs. Actual Scatter for Renters

Takeaway: These predictors explain only ~27–29% of the variance, underscoring complexity. RF performs marginally better.

Owner Burden Models

Model	RMSE	R^2
Linear Regression	6.3	0.234
Random Forest	6.1	0.291

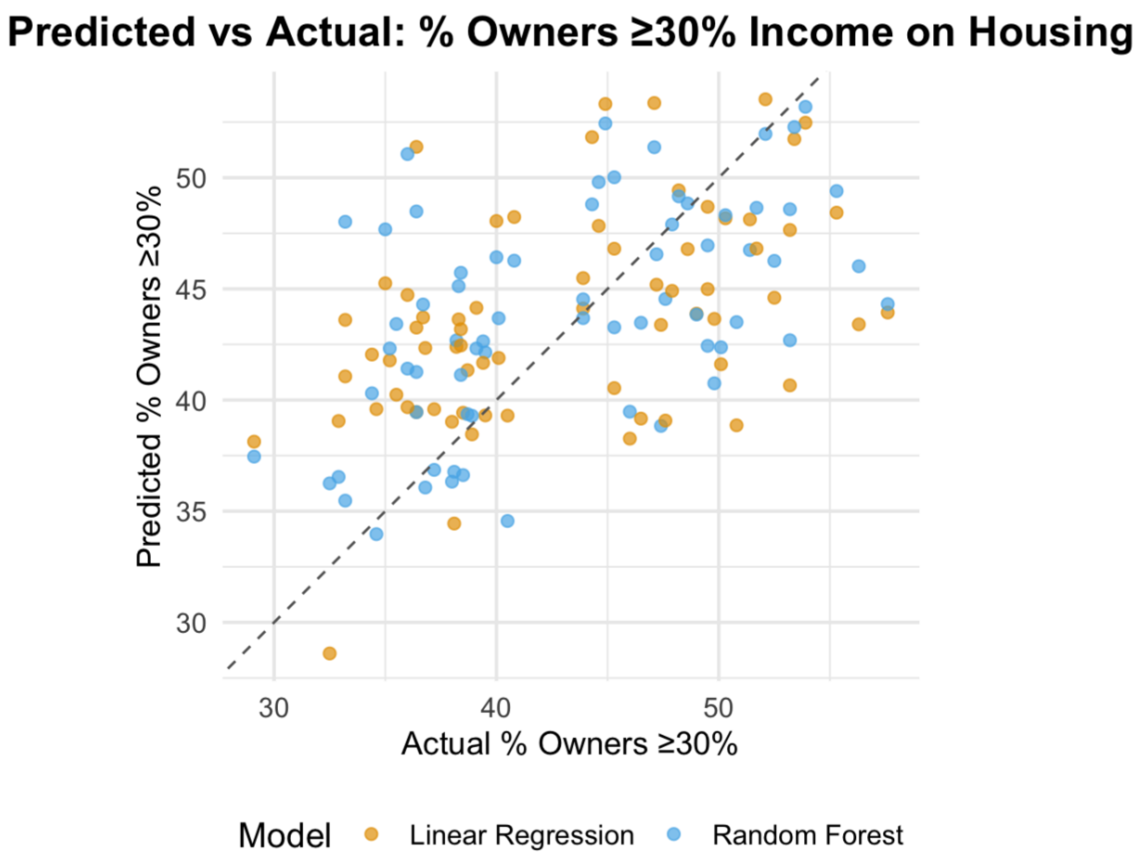


Figure 7: Predicted vs. Actual Scatter for Owners

Conclusion: Similar dynamics apply. Slight performance increase from RF, but both suggest low predictive power from income and affordability alone.

K-Means Clustering: Renters

Three clusters emerged from renter-based clustering:

Cluster	Count	Mean Burden	Ability	Income
1	124	57.1%	0.257	$54,445
2	64	53.3%	0.194	$93,242
3	44	44.5%	0.207	$59,632

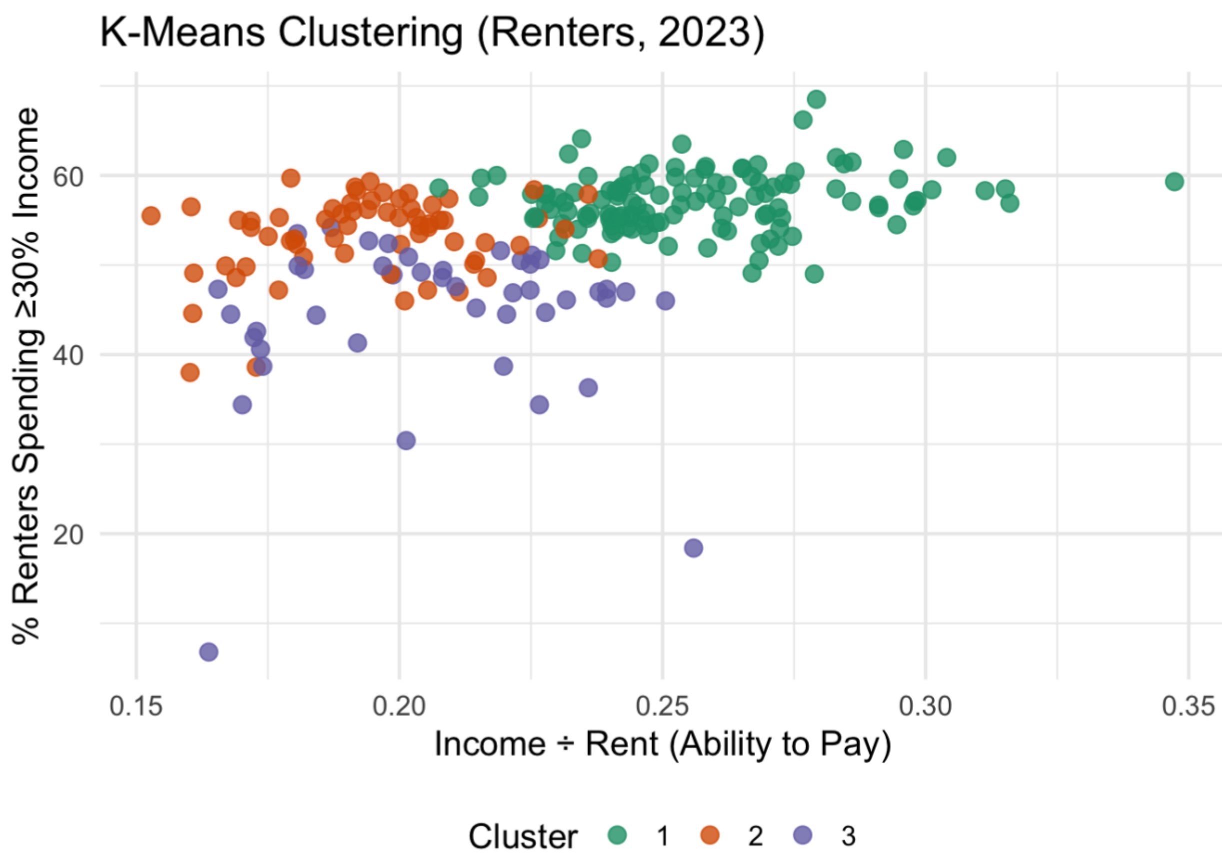


Figure 8: Renter Cluster Scatterplot and Cluster Summary Table

- **Cluster 1**: High burden, low income = most vulnerable
- **Cluster 2**: High income, moderate burden = "income rich, rent poor"
- **Cluster 3**: Lowest burden, mixed income = stability

K-Means Clustering: Owners

Owner clusters revealed a different pattern:

Cluster	Count	Mean Burden	Ability	Income
1	45	38.0%	9.09	$101,337
2	105	39.6%	6.04	$57,188
3	82	50.4%	8.55	$58,263

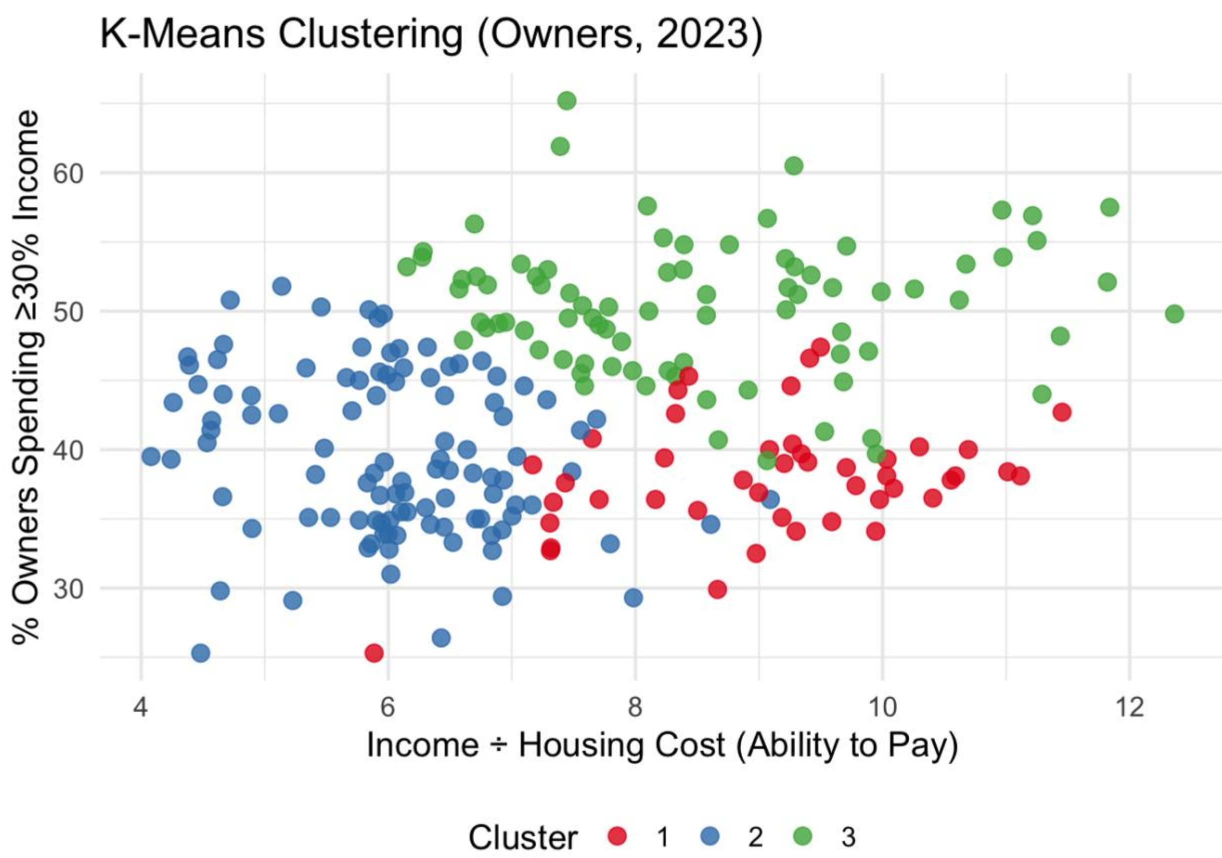


Figure 9: Owner Cluster Scatterplot and Cluster Summary Table

- **Cluster 3**: Most burdened, moderate income
- **Cluster 1**: Highest income, lowest burden = elite ownership
- **Cluster 2**: Marginal owners, vulnerable to economic shocks

---

Conclusions

- **Renters face higher cost burdens** and lower ability-to-pay, with many counties exceeding 55% burden.
- **Owners experience milder burden**, but a sizable minority are still cost-challenged.
- **Income alone is not a sufficient predictor** of housing stress. Both renter and owner models show limited R^2.
- **Clustering reveals hidden structures**: high income does not guarantee affordability; some mid-income counties carry significant stress.
- **Policy must distinguish between structural affordability and surface-level income measures.**

Recommendations

- Target assistance to **Cluster 1 renters and Cluster 3 owners**.
- Develop affordability metrics that go **beyond income**: include cost-burden ratios and ability-to-pay estimates.
- Integrate housing burden analytics into **regional planning tools**.
- Consider **predictive early-warning indicators** based on ability-to-pay shifts.

Appendix – All Code

```
# ──────────────────────────────────────────────────────────────
# Load Core Yearly Data: 2010, 2013, 2018, 2023
# ──────────────────────────────────────────────────────────────

# Load housing cost-burden metrics (renter and owner)
dp04 <- read_csv(
  "Data/dp04_2010_2023_housing_with_renter_share.csv",
  show_col_types = FALSE
) %>%
  mutate(
    owner_ge30 = selected_monthly_owner_costs_as_a_percentage_of_household_in
come_smocapi_30_to_34_9_percent +
                 selected_monthly_owner_costs_as_a_percentage_of_household_in
come_smocapi_35_percent_or_more
  ) %>%
  select(
    county,
    year,
    renter_ge30 = renter_ge30_pct,
    owner_ge30
  )



# Load ability-to-pay estimates
ability <- read_csv("/Users/meganryan/Documents/University_Oklahoma/R Project
s/Income Inequality Project/Data/ability_to_pay_estimated_2010_2023.csv", sho
w_col_types = FALSE) %>%
  select(county, year, renters_ability_to_pay, owners_ability_to_pay, median_
income)

# ──────────────────────────────────────────────────────────────
# Merge all into a single long-form dataset (all years)
# ──────────────────────────────────────────────────────────────
housing_df <- dp04 %>%
  left_join(ability, by = c("county", "year")) %>%
  drop_na()

# Preview

# Renter clustering (k-means)
cluster_renters <- housing_df %>%
  select(county, renter_ge30, renters_ability_to_pay, median_income) %>%
  drop_na()

scaled_renters <- scale(cluster_renters %>% select(-county))
```

```r
set.seed(2025)
k_renters <- kmeans(scaled_renters, centers = 3, nstart = 25)

cluster_renters_labeled <- cluster_renters %>%
  mutate(Cluster = factor(k_renters$cluster))

# Owner clustering (k-means)
cluster_owners <- housing_df %>%
  select(county, owner_ge30, owners_ability_to_pay, median_income) %>%
  drop_na()

scaled_owners <- scale(cluster_owners %>% select(-county))

set.seed(2025)
k_owners <- kmeans(scaled_owners, centers = 3, nstart = 25)

cluster_owners_labeled <- cluster_owners %>%
  mutate(Cluster = factor(k_owners$cluster))

# Renters clustering (no filtering)
cluster_renters <- housing_df %>%
  select(county, renter_ge30, renters_ability_to_pay, median_income) %>%
  drop_na()

# Scale the numeric variables
scaled_renters <- scale(cluster_renters %>% select(-county))

# K-means
set.seed(2025)
k_renters <- kmeans(scaled_renters, centers = 3, nstart = 25)

# Add cluster labels to housing_df — using row numbers as anchor
housing_df$RenterCluster <- NA
housing_df$RenterCluster[as.numeric(rownames(cluster_renters))] <- k_renters$
cluster
housing_df$RenterCluster <- factor(housing_df$RenterCluster)

# Plot clusters
ggplot(cluster_renters_labeled, aes(x = renters_ability_to_pay, y = renter_ge
30, color = Cluster)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(
    title = "K-Means Clustering (Renters, 2023)",
    x = "Income ÷ Rent (Ability to Pay)",
    y = "% Renters Spending ≥30% Income"
  ) +
  scale_color_brewer(palette = "Dark2") +
  theme_minimal(base_size = 14) +
  theme(legend.position = "bottom")
```

```
# Owners clustering (no filtering)
cluster_owners <- housing_df %>%
  select(county, owner_ge30, owners_ability_to_pay, median_income) %>%
  drop_na()

# Scale the numeric variables
scaled_owners <- scale(cluster_owners %>% select(-county))

# K-means
set.seed(2025)
k_owners <- kmeans(scaled_owners, centers = 3, nstart = 25)

# Add cluster labels to housing_df
housing_df$OwnerCluster <- NA
housing_df$OwnerCluster[as.numeric(rownames(cluster_owners))] <- k_owners$clu
ster
housing_df$OwnerCluster <- factor(housing_df$OwnerCluster)


# Plot clusters
ggplot(cluster_owners_labeled, aes(x = owners_ability_to_pay, y = owner_ge30,
color = Cluster)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(
    title = "K-Means Clustering (Owners, 2023)",
    x = "Income ÷ Housing Cost (Ability to Pay)",
    y = "% Owners Spending ≥30% Income"
  ) +
  scale_color_brewer(palette = "Set1") +
  theme_minimal(base_size = 14) +
  theme(legend.position = "bottom")

# Print renter cluster summary
renter_cluster_summary <- housing_df %>%
  filter(!is.na(RenterCluster)) %>%
  group_by(RenterCluster) %>%
  summarise(
    Count = n(),
    Mean_Renter_GE30 = round(mean(renter_ge30, na.rm = TRUE), 1),
    Mean_Ability = round(mean(renters_ability_to_pay, na.rm = TRUE), 3),
    Mean_Income = round(mean(median_income, na.rm = TRUE), 0)
  )

print(renter_cluster_summary)

# Print owner cluster summary
owner_cluster_summary <- housing_df %>%
  filter(!is.na(OwnerCluster)) %>%
  group_by(OwnerCluster) %>%
  summarise(
```

```
    Count = n(),
    Mean_Owner_GE30 = round(mean(owner_ge30, na.rm = TRUE), 1),
    Mean_Ability = round(mean(owners_ability_to_pay, na.rm = TRUE), 3),
    Mean_Income = round(mean(median_income, na.rm = TRUE), 0)
  )

library(knitr)

kable(renter_cluster_summary, caption = "Renter Cluster Summary (K-Means)")
kable(owner_cluster_summary, caption = "Owner Cluster Summary (K-Means)")

# Cluster summary table (already created)
renter_cluster_summary <- housing_df %>%
  filter(!is.na(RenterCluster)) %>%
  group_by(RenterCluster) %>%
  summarise(
    Count = n(),
    Mean_Renter_GE30 = round(mean(renter_ge30, na.rm = TRUE), 1),
    Mean_Ability = round(mean(renters_ability_to_pay, na.rm = TRUE), 3),
    Mean_Income = round(mean(median_income, na.rm = TRUE), 0),
    .groups = "drop"
  )

# Melt to long format for plotting
renter_cluster_long <- renter_cluster_summary %>%
  pivot_longer(cols = starts_with("Mean_"), names_to = "Metric", values_to =
"Value") %>%
  mutate(
    Metric = recode(Metric,
                    "Mean_Renter_GE30" = "% Renters ≥30%",
                    "Mean_Ability" = "Ability to Pay",
                    "Mean_Income" = "Median Income")
  )

# Plot
ggplot(renter_cluster_long, aes(x = RenterCluster, y = Value, fill = Metric))
+
  geom_col(position = position_dodge(width = 0.7), width = 0.6) +
  facet_wrap(~ Metric, scales = "free_y") +
  scale_fill_brewer(palette = "Set2") +
  labs(
    title = "Renter Clusters: Mean Values by Cluster (K-Means)",
    x = "Renter Cluster", y = NULL
  ) +
  theme_minimal(base_size = 14) +
  theme(
    legend.position = "none",
    plot.title = element_text(face = "bold", hjust = 0.5)
  )
```

```r
# Cluster summary
owner_cluster_summary <- housing_df %>%
  filter(!is.na(OwnerCluster)) %>%
  group_by(OwnerCluster) %>%
  summarise(
    Count = n(),
    Mean_Owner_GE30 = round(mean(owner_ge30, na.rm = TRUE), 1),
    Mean_Ability = round(mean(owners_ability_to_pay, na.rm = TRUE), 3),
    Mean_Income = round(mean(median_income, na.rm = TRUE), 0),
    .groups = "drop"
  )

# Long format
owner_cluster_long <- owner_cluster_summary %>%
  pivot_longer(cols = starts_with("Mean_"), names_to = "Metric", values_to =
"Value") %>%
  mutate(
    Metric = recode(Metric,
                    "Mean_Owner_GE30" = "% Owners ≥30%",
                    "Mean_Ability" = "Ability to Pay",
                    "Mean_Income" = "Median Income")
  )

# Plot
ggplot(owner_cluster_long, aes(x = OwnerCluster, y = Value, fill = Metric)) +
  geom_col(position = position_dodge(width = 0.7), width = 0.6) +
  facet_wrap(~ Metric, scales = "free_y") +
  scale_fill_brewer(palette = "Set1") +
  labs(
    title = "Owner Clusters: Mean Values by Cluster (K-Means)",
    x = "Owner Cluster", y = NULL
  ) +
  theme_minimal(base_size = 14) +
  theme(
    legend.position = "none",
    plot.title = element_text(face = "bold", hjust = 0.5)
  )
```