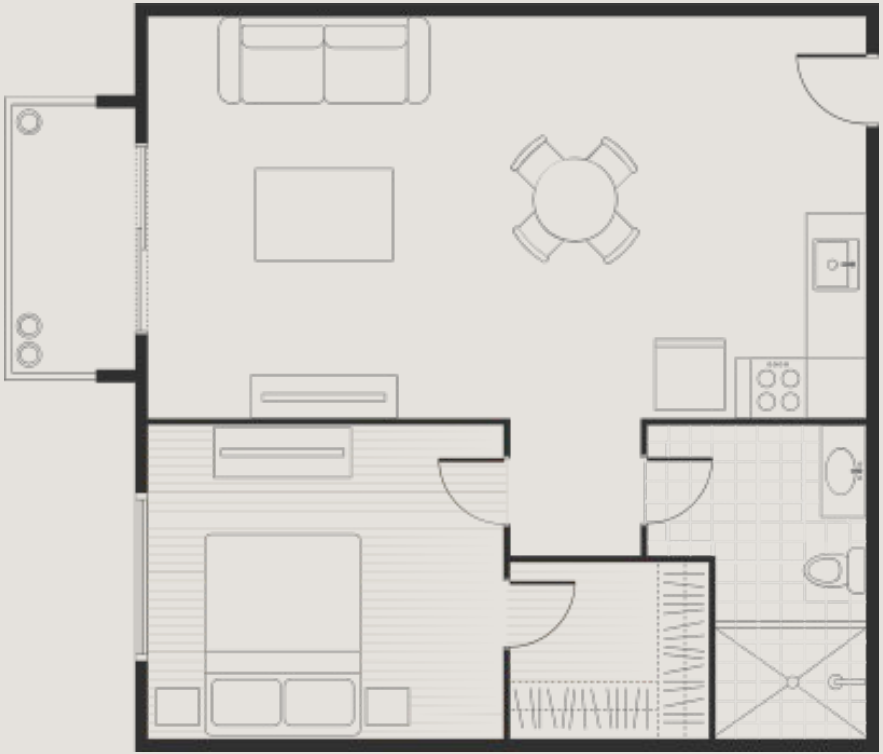
How New Construction Impacts Homeowner Affordability in U.S. Metros

Section 7
Jacob Schorr, Randy Tovar, Hannah MacDonald, Georgios Friligkos

Research Question

How has new homeowner affordability changed in metros with new construction housing from 2020-2024?



Why it matters?

- Affordability = major policy issue
- This study allows us to see if the US is succeeding in introducing more affordable housing for first time homeowners

Data Overview: Zillow

Data Source
Zillow Housing Data, U.S. metro areas

Exploratory Set
Time Frame: Jan 2020-Dec 2021

Confirmation Set
Time Frame: Jan 2022 - Dec 2024

Data Source

- Utilized data from Zillow Research focusing on U.S. metropolitan areas.
- Analyzed trends across all homes instead of individual properties.
- Although the dataset dates back to 2012, the analysis centered on the most recent 5-year period.
- Reviewed data from January 1st through December 31st for each year.
- Extracted more than 190 data points for each set
- The data is merged on a unique id, Region ID.

Variables

Variables	Types	Description
Years to Save	Outcome (Y)	Average years to save for a 20% down payment
New Construction Count	Predictor (X)	Average monthly new construction count
Monthly Payment	Predictor (Z)	Average monthly mortgage principal & interest, property taxes, and maintenance
Market Heat Index	Predictor (W)	Average market heat index (competitiveness of market)

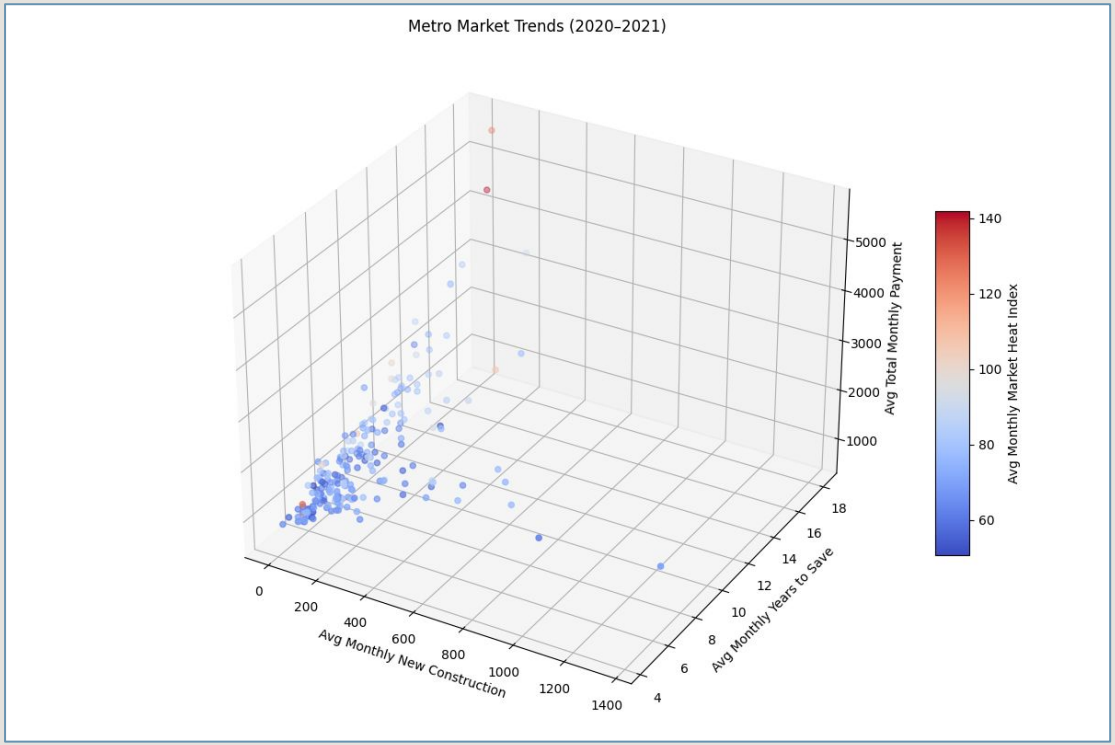
Sample of Aggregated Data

	RegionName <chr>	avg_monthly_new_construction <dbl>	avg_monthly_years_to_save <dbl>	avg_total_monthly_payment <dbl>	avg_monthly_market_heat_index <dbl>
1	New York, NY	506.7917	12.470567	3005.665	79.50000
2	Los Angeles, CA	208.4167	17.991127	3366.633	89.41667
3	Chicago, IL	375.2917	6.427921	1488.451	61.79167
4	Dallas, TX	759.0000	7.414214	1609.409	78.79167
5	Houston, TX	904.3750	6.806715	1336.074	62.62500
6	Washington, DC	398.3333	8.683215	2270.399	78.04167

Confirmation set aggregated data sample:

- Averaged monthly values across 5 years to represent one record per metro
- Removed “United States” row because it was aggregated data for all metros

Exploratory Analysis

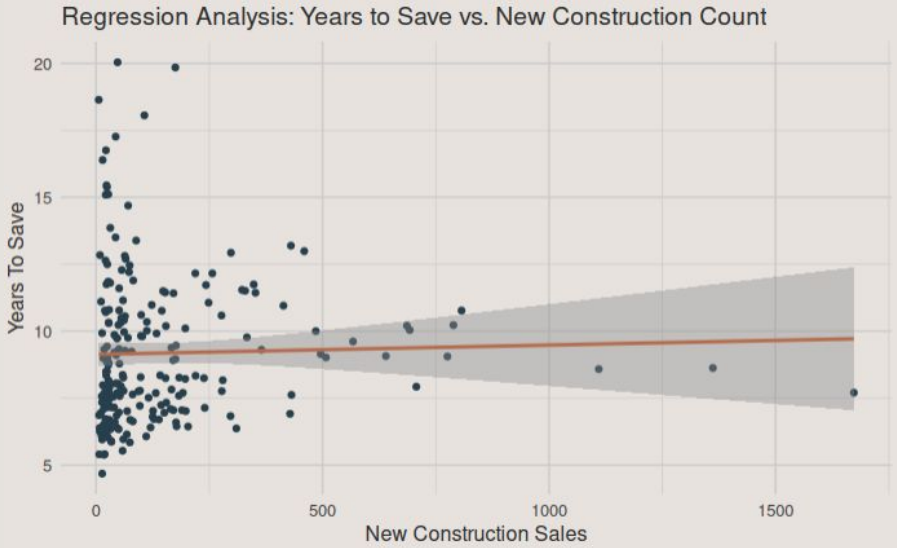


Simple 4-D Plot Graph

- Avg # of New Construction Homes Sold (Monthly)
- Avg # of Years to Save (tracked monthly)
- Avg Total Monthly Payment (Mortgage P&I, Property Taxes & Maintenance)
- Avg Market Heat Index (tracked monthly)

Simple Linear Regression Model

- Data points spread along the y-axis, seem to have more of a symmetrically distribution.
- A majority of the data points cluster toward the lower end of the distribution, suggest while most observations register lower sales counts, a few outliers exhibit significantly higher values.
- As new construction sales increase the shaded region (confidence interval) is relatively broad, signifying uncertainty.



Data Source

- Simple Regression model calculated on the confirmation set from June 2022 -> Dec 2024
- Displays individual data points comparing Years to Save (y-axis) with New Construction Sales (x-axis).

Interpretation

Independent Variable

- The coefficient is 0.0003, suggesting a negligible relationship.

Model Fit

- The Multiple R-squared is 0.0007, meaning that less than 0.1% of the variation in Years to Save is explained by New Construction.

Overall Significance:

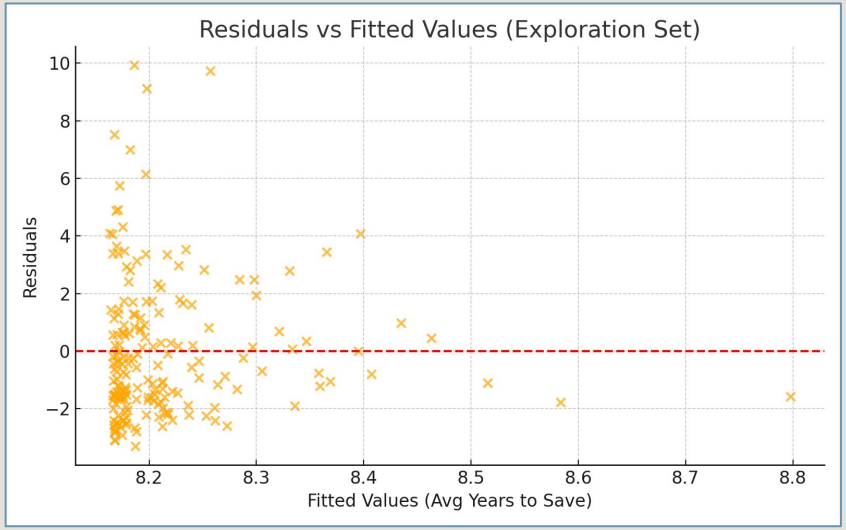
- With a p-value of 0.6923, the model as a whole is not statistically significant.

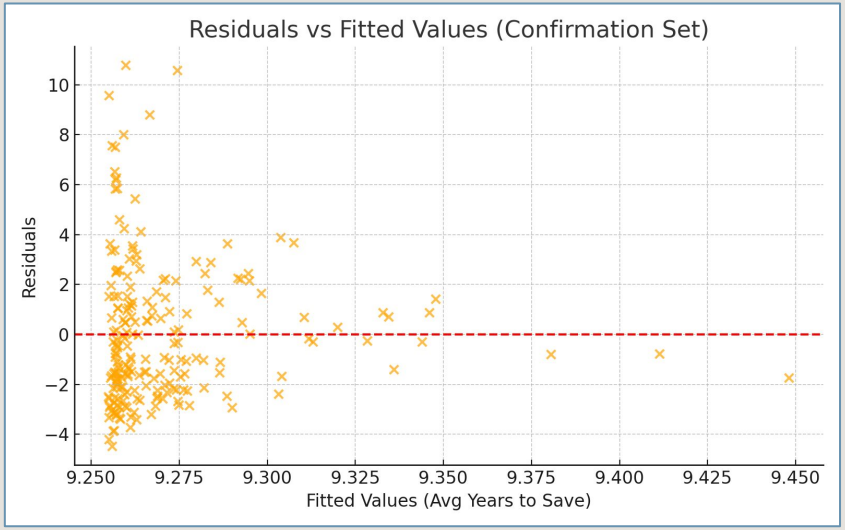
	Dependent variable:
	Years to Save
New Construction	0.0003
	(0.0009)
Constant	9.1343***
	(0.2286)
Observations	214
R2	0.0007
Adjusted R2	-0.0040
Note:	*p<0.1; **p<0.05; ***p<0.01

```
Residual standard error: 2.792 on 212 degrees of freedom
Multiple R-squared:  0.0007403, Adjusted R-squared:  -0.003973
F-statistic: 0.1571 on 1 and 212 DF,  p-value: 0.6923
```

Assumptions

Linearity and Homoskedasticity

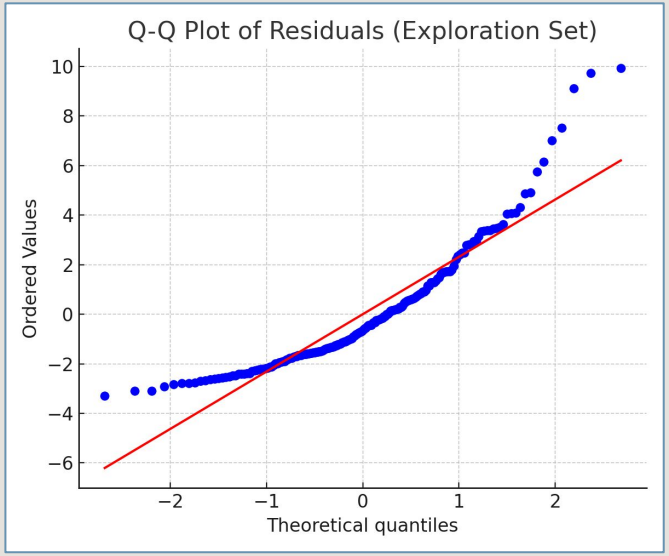


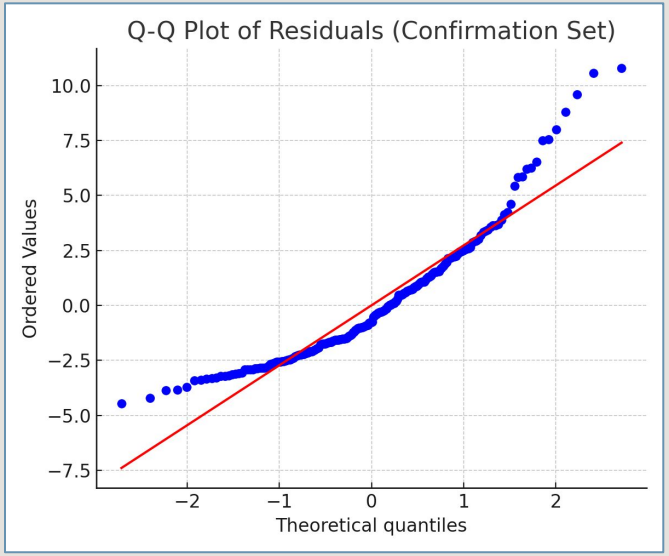


- Residuals vs Fitted plots reveal clustering at lower fitted values
- Residuals are not evenly spread which suggests non-constant variance
- Some pattern or trend visible which indicate possible non-linearity
- Mild violation of linearity and homoscedasticity assumptions

Assumptions

Normality and Independence of Residuals





- Residuals are approximately normal with slight deviations(especially the upper End) in the tails. This is not a major concern and the assumption is reasonably satisfied for both datasets
- The independence assumption is inherently satisfied by our data design, since there are no repeated time-based observations within each metro
- Multicollinearity: Not applicable, only one predictor used

Multivariate Linear Regression Model

Model:
Years to Save ~ New Construction + Monthly Payment + Market Heat Index

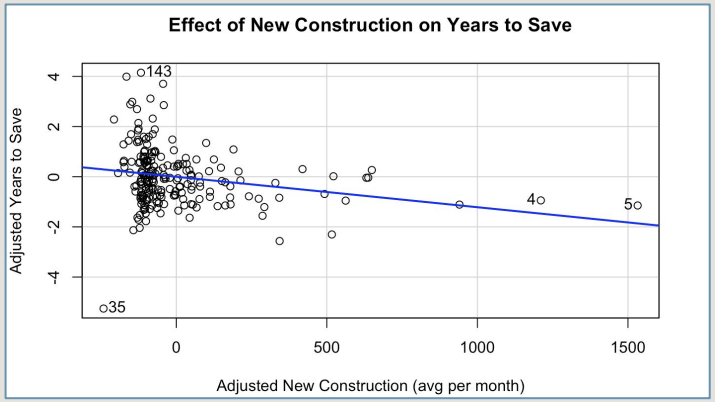
Adjusted R-squared:
84% of the variation on affordability is explained by this model

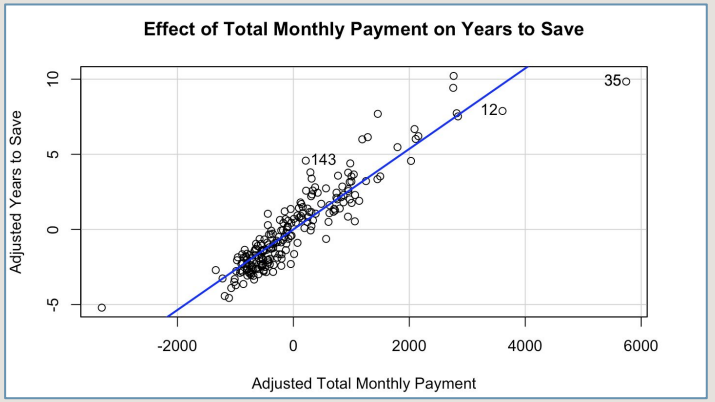
```
Residuals:
    Min      1Q  Median      3Q     Max
-5.5495 -0.7411 -0.0657  0.5482  4.0039

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   6.744e+00  3.956e-01  17.046  < 2e-16 ***
avg_monthly_new_construction -1.215e-03  3.679e-04  -3.302  0.00113 **
avg_total_monthly_payment     2.680e-03  8.304e-05  32.279  < 2e-16 ***
avg_monthly_market_heat_index -5.644e-02  6.742e-03  -8.370 8.74e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.156 on 206 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.8367,    Adjusted R-squared:  0.8343
F-statistic: 351.8 on 3 and 206 DF,  p-value: < 2.2e-16
```

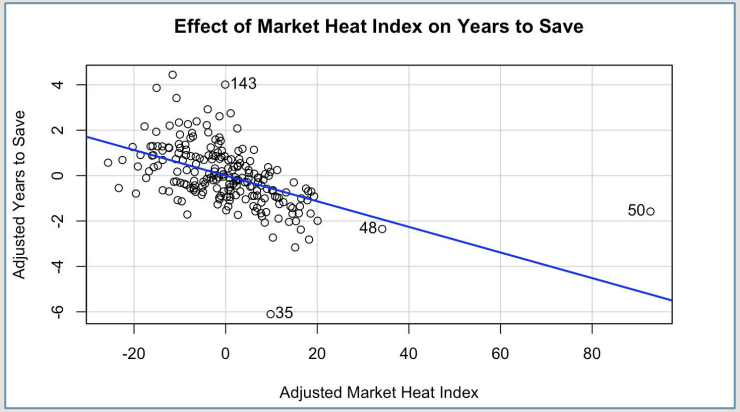
At first glance after fitting our model we see:

- All predictors are incredibly significant
- The direction aligns with expectations:
    - Housing affordability gets worse with higher payments and better with more supply

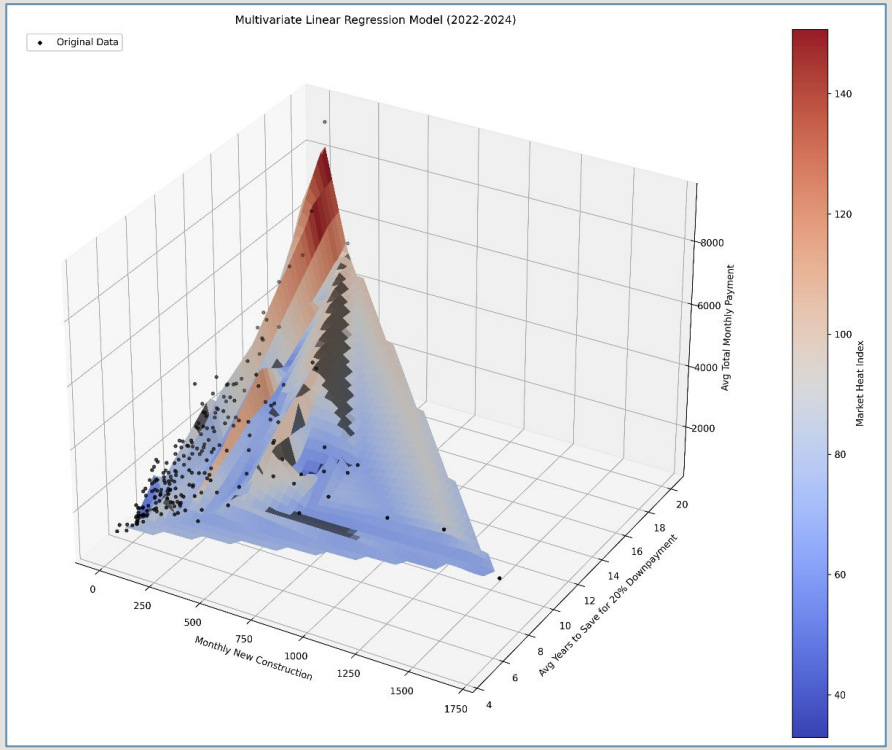
Unique Contribution of Each Predictor







Multivariate Linear Regression Model Findings



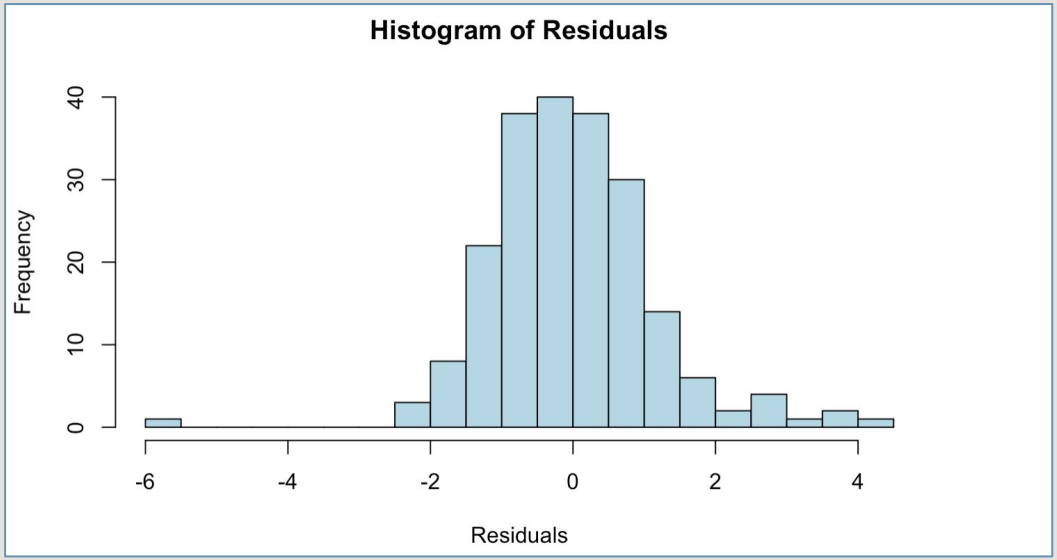
After reviewing our model, we discover it:

- Captures systemic market signals
- Fits well, despite the messiness of real world data
- Highlights structural inequities
  - Longer savings times and high payments cluster together in high-heat, low-construction markets

Assumptions

Normality of Residuals:

- Roughly symmetric and bell-shaped, allows us to begin assuming the assumption is met



Independent and Identically Distributed:

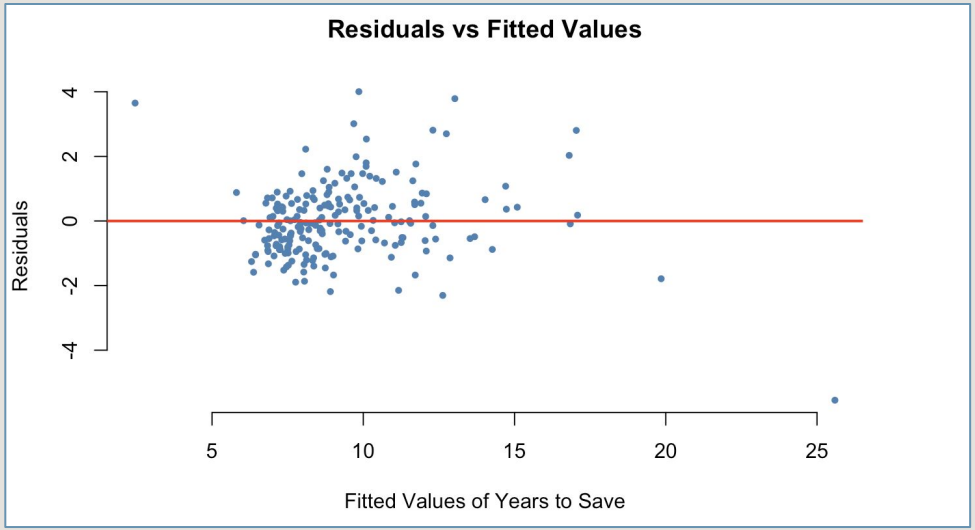
- Since we aggregated the data to contain one summary record per metro area, it should be approximately IID
- We verified this with the Durbin-Watson test

```
        Shapiro-Wilk normality test

data:  resid(model2)
W = 0.95006, p-value = 1.122e-06
```

Assumptions

Linearity:

- No strong trend or curve -> no major nonlinearity jumps out
- Spread fanning out as fitted values increase



Homoskedasticity:

- BP test indicates heteroskedasticity ($p < 0.05$)
  - Violates this assumption
- Used robust standard errors and robust F-test (Wald test) to re-test overall model significance and get reliable p-values for predictors

```
        studentized Breusch-Pagan test

data:  model2
BP = 46.023, df = 3, p-value = 5.609e-10
```

Interpretation

- All predictors are strongly associated with monthly housing costs — and the relationships are statistically robust, even after correcting for non-normality.
- These predictors act as reliable signals of affordability pressure — their values track consistently with changes in monthly payments across metros.
- Effect sizes are practically meaningful: shifts in supply, demand, or savings dynamics are tied to noticeable changes in monthly cost burden.
- While not strictly causal, the results point to clear, observable pressure points that align with affordability outcomes — and many are policy-responsive.

Feature	Test	Robust p-value
Avg Monthly New Construction	Robust t-test	0.00010
Avg Monthly Years to Save	Robust t-test	< 2.2e-16
Avg Monthly Market Heat Index	Robust t-test	0.000017
Model vs. Null (Intercept Only)	Robust Wald F-test	< 2.2e-16

Conclusion

Housing affordability is measurable and patterned

Affordability patterns align with metro-level indicators. The market heat, payment burden, and supply offer reliable metrics.

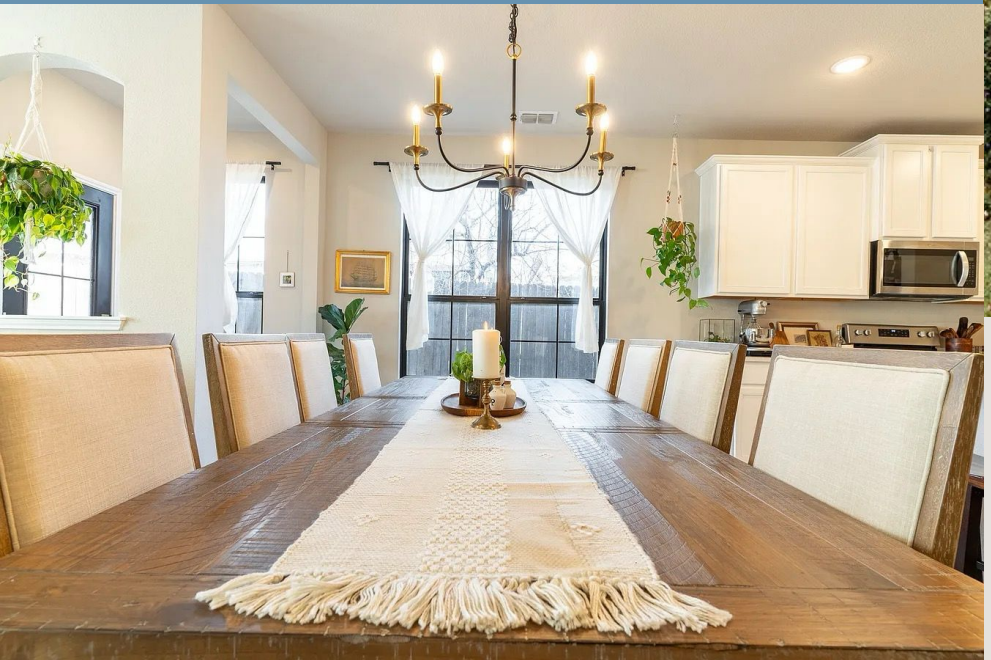
Multivariate modeling reveals meaningful structure

The three predictor values fitted together account for 84% of variation. Relationships are statistically and practically significant.

Limitations & Further Opportunities

This descriptive model sets the stage for future modeling. Next steps could be causal inference or time-based forecasting.

Thank you!



Any Questions?