Data Analysis Overview:

To explore which factors directly correlate with rent costs, we implemented a three-pronged analytical approach.

First, we examined data at the state level to identify broad patterns and determine if regional trends had a significant impact on rent prices.

Next, we refined our focus by analyzing data at the ZIP code level, which gave us a more granular view and a larger set of data points. This allowed for a more precise evaluation of how local factors influence rent costs.

Finally, we expanded our scope to include non-economic variables to assess whether other factors outside traditional economic indicators affect rent prices. This approach helped us identify contrasting variables and strengthened our conclusions by highlighting which specific factors had a stronger correlation with rent costs compared to others.

1. State Data Analysis Report

Variables Explanation

- states: The state from which the data is collected
- avg_unemployment: Average Unemployment Rate
- industries: The dictionary that counts the number of each industry being used
- count: Number of Corporations owned Buildings in the dataset
- total: Number of Buildings in the dataset
- rate: Count / Total, Ratio of buildings occupied by corporations
- key_industry_rate: Percentage of Count occupied by core corporations, including terms engineering, technology, business, financial
- overall_rent: Annual Rent Fee per Square Foot
- cbd_rate: Percentage of buildings in the city area relative to the total number of buildings
- Population.Change.Rate: Population Growth Rate
- CPI: Consumer Price Index of each state
- Median.Household.Income: Median Household Income

We collected Data from 2018Q1 to 2024Q1. Considering the fluctuations due to the pandemic, we decided to take the average of each variable. The pandemic increases uncertainty in the data, the purpose of this report is to see what affects the rent at the national level with a macroeconomic view.
Since there are differences in the price level between each state, we added CPI value to the dataset so that the value is adjusted by price level.

There are 20 states that are extracted from the dataset.

Also, we wanted to see how expensive housing of each state is compared to their income.

External Sources:
Population.Change.Rate: **U.S. Census Bureau**
Median.Household.Rate: **Federal Reserve Bank of St. Louis**
CPI: **U.S. Bureau of Labor Statistics**

Data Explanation

We initially regressed **overall_rent** on a broad set of predictors:

Variables included:
avg_unemployment, count, total, rate, key_industry_rate, cbd_rate, populationchangerate, medianhouseholdincome, cpi

Source	SS	df	MS		
Model	2444.22785	9	271.580872	Number of obs	= 20
Residual	271.154724	10	27.1154724	F(9, 10)	= 10.02
				Prob > F	= 0.0006
Total	2715.38258	19	142.914872	R-squared	= 0.9001
				Adj R-squared	= 0.8103
				Root MSE	= 5.2073

overall_rent	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
avg_unemployment	-4.697317	2.694285	-1.74	0.112	-10.70056	1.305924
count	.0118222	.0033086	3.57	0.005	.0044503	.0191941
total	-.0005507	.0002789	-1.97	0.077	-.001172	.0000707
rate	-45.70203	24.07763	-1.90	0.087	-99.35033	7.946266
key_industry_rate	-4.806494	15.66144	-0.31	0.765	-39.70236	30.08937
cbd_rate	13.04685	9.253953	1.41	0.189	-7.572248	33.66594
populationchangerate	.2954543	.758791	0.39	0.705	-1.395238	1.986146
medianhouseholdincome	-.0001361	.0001572	-0.87	0.407	-.0004864	.0002142
cpi	.3655749	.1135039	3.22	0.009	.1126724	.6184773
_cons	-39.28945	31.60577	-1.24	0.242	-109.7115	31.13259

This model achieved **a very high R-squared** of **0.9001**, indicating strong overall explanatory power.

However, due to the **limited sample size (N = 20)**, including too many correlated variables raised concerns about **multicollinearity** and potential overfitting.

We exclude rate because rate is explained by count and total.

We visualized:

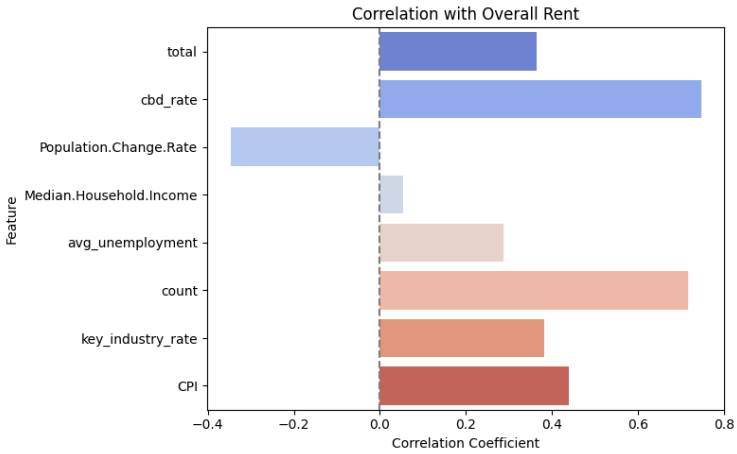
- The **correlation coefficients** between each variable and overall_rent, and
- The corresponding **p-values from the regression**

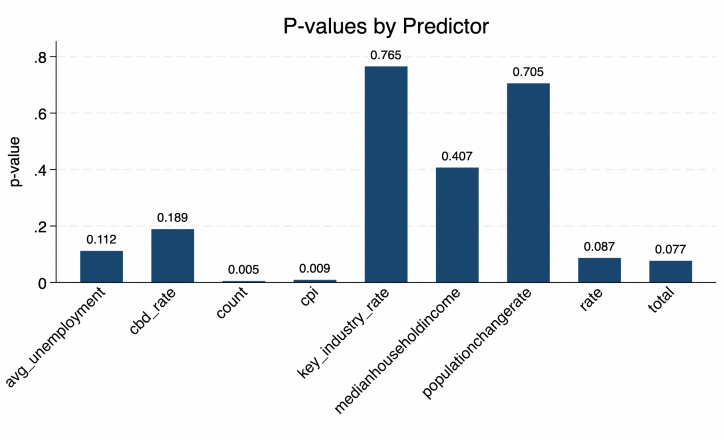
This helped us identify variables that were either:

- Weakly correlated with rent, or
- Statistically insignificant in the model ($p > 0.1$)

Based on these results, we excluded the following:

- avg_unemployment
- populationchangerate
- medianhouseholdincome





So we included those variables into the new regression model.

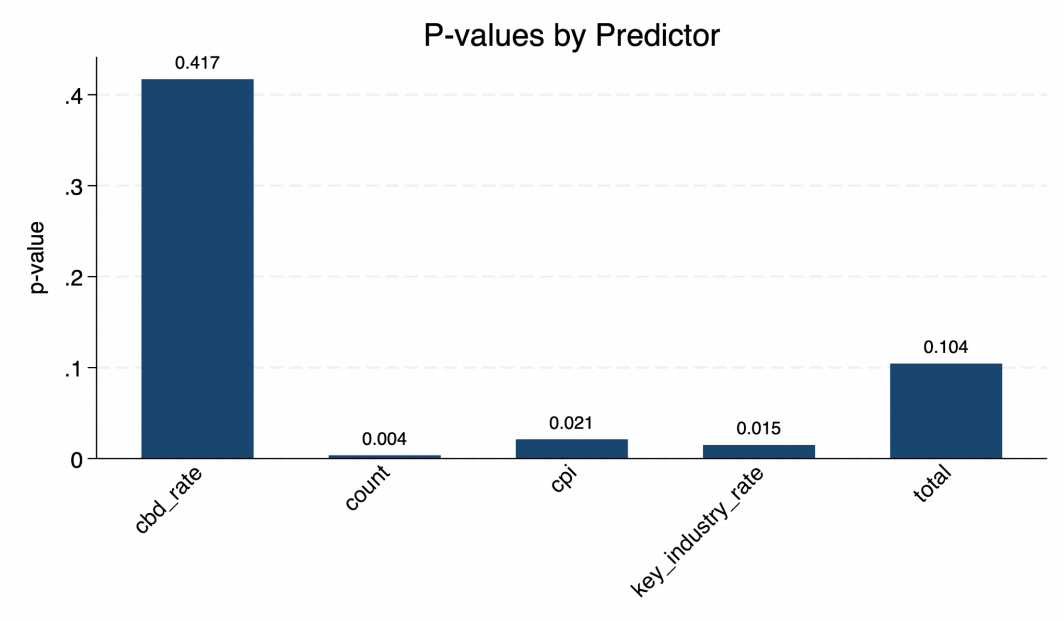
We re-estimated the model with five predictors:

Variables included:
count, total, key_industry_rate, cbd_rate, cpi

```
. reg overall_rent count total key_industry_rate cbd_rate cpi
```

Source	SS	df	MS	Number of obs	=	20
				F(5, 14)	=	13.27
Model	2242.22974	5	448.445949	Prob > F	=	0.0001
Residual	473.153016	14	33.796644	R-squared	=	0.8258
				Adj R-squared	=	0.7635
Total	2715.38276	19	142.914882	Root MSE	=	5.8135

overall_rent	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
count	.0081424	.0023247	3.50	0.004	.0031564	.0131285
total	-.0003486	.0002008	-1.74	0.104	-.0007793	.000082
key_industry_rate	22.20981	8.006914	2.77	0.015	5.036691	39.38294
cbd_rate	7.536121	9.011265	0.84	0.417	-11.79112	26.86336
cpi	.2757335	.1061646	2.60	0.021	.048033	.503434
_cons	-63.1803	29.5637	-2.14	0.051	-126.5881	.2275254



- This reduced model maintained a **strong R-squared of 0.8258** and **Adjusted R-squared of 0.7635**
- Most variables remained statistically significant, except cbd_rate ($p = 0.417$)

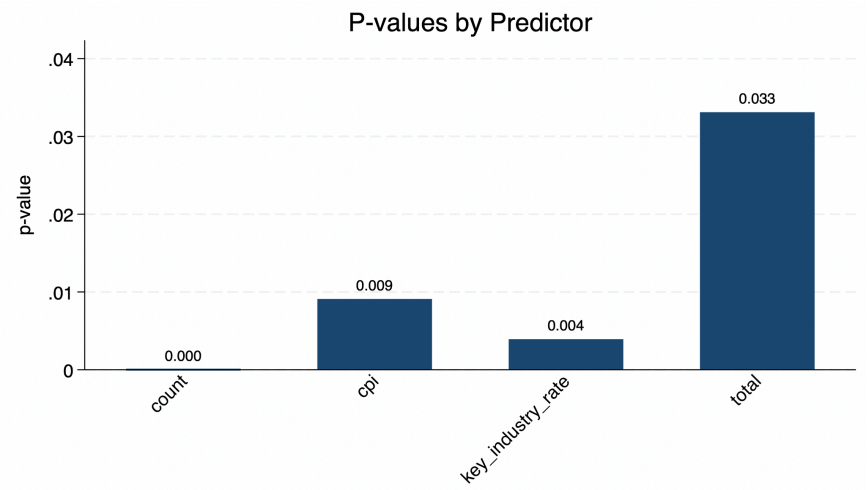
Although cbd_rate showed a **strong positive correlation** with overall_rent, its high p-value suggests that **its effect may be captured by other variables**, particularly count, which reflects overall leasing activity.

The final model regressed overall_rent without cbd_rate.
Variables Included:
Count, total, key_indutsry_rate, cpi

```
. reg overall_rent count total key_industry_rate  cpi
```

Source	SS	df	MS	Number of obs	=	20
				F(4, 15)	=	16.75
Model	2218.59245	4	554.648112	Prob > F	=	0.0000
Residual	496.790312	15	33.1193541	R-squared	=	0.8170
				Adj R-squared	=	0.7683
Total	2715.38276	19	142.914882	Root MSE	=	5.7549

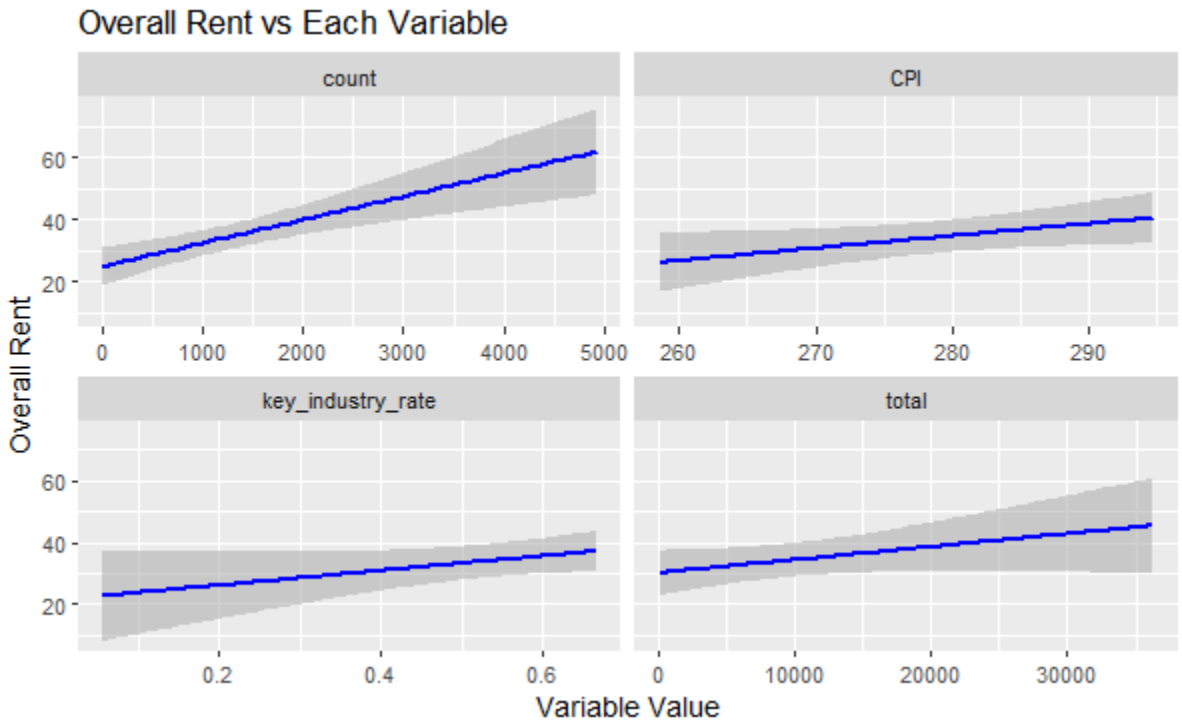
overall_rent	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
count	.009555	.0015812	6.04	0.000	.0061846	.0129253
total	-.0004209	.0001794	-2.35	0.033	-.0008033	-.0000386
key_industry_rate	24.83145	7.293549	3.40	0.004	9.285614	40.37728
cpi	.3012977	.1006448	2.99	0.009	.0867784	.5158169
_cons	-70.95995	27.7794	-2.55	0.022	-130.1703	-11.74956



Based on these findings, we constructed a final regression model including only variables with strong theoretical relevance and statistical significance.

This specification balances:

- **Model simplicity**
- **Statistical robustness**
- **Economic interpretability**



Looking at the graph, we can see there is a positive and obvious correlation between the rent and each variable used for the regression. Also, SE, explained by the shaded area around the lines, is not wide, showing the accuracy of the model.

Conclusion

We conclude that the rent increases if the state has more buildings owned by corporations, especially corporations that belong to one of the core industries(business, technology, engineering, and financial services), and a higher price level. We are going to develop this model to predict further specification.

Explanation
Key_industry_rate:
Count:
State-level observation
How to get the Regression model - p-value + correlation + multicolinear + simplifying the model maintaining R squared
External Source

2. Zip Code Data Analysis Report - Economic variable focused

Variables Explanation

- zip: the median ZIP code of the geographical cluster
- industries: The dictionary that counts the number of each industry being used
- count: Number of Corporations owned Buildings in the dataset
- total: Number of Buildings in the dataset
- overall_rent: Annual Rent Fee per Square Foot
- rate: Count / Total, Ratio of buildings occupied by corporations
- key_industry_rate: Percentage of Count occupied by core corporations, including terms engineering, technology, business, financial
- overall_rent: Annual Rent Fee per Square Foot

This report narrowed down the observation from the first report, which was at the state level. The observations are organized by zip code. The buildings with similar zip codes are geographically clustered. This allows for a closer look into each area's characteristics, specifically with a focus on how developed the area is.

First, we made a dictionary that counts the number of each industry being used, putting that under industries variable.
That way we can determine the density of core industry in each zip code interval.
In order to see the relationship between technological advance and the overall rent, we decided to remove the other variables unrelated to core industry indexes.

We collected *count, total,* and *overall_rent rate, key_industry_rate, industries* from **the provided csv file.**
We collected *zip* from **the provided csv file** and organized it with an interval of 200.

External Sources:
CPI: **U.S. Bureau of Labor Statistics**

Data Explanation

We began by regressing **overall_rent** on the following predictors using a larger dataset with finer regional granularity (e.g., by ZIP or smaller metro areas):

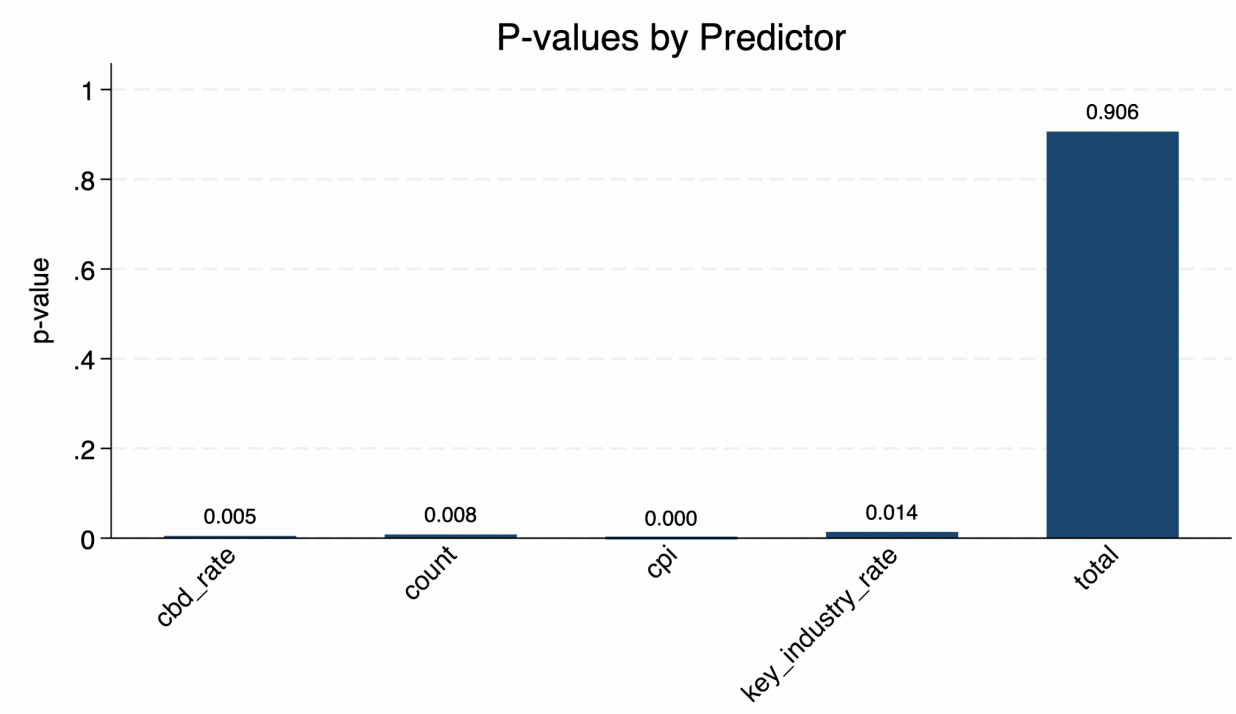
Variables included:

count, total, key_industry_rate, cbd_rate, cpi

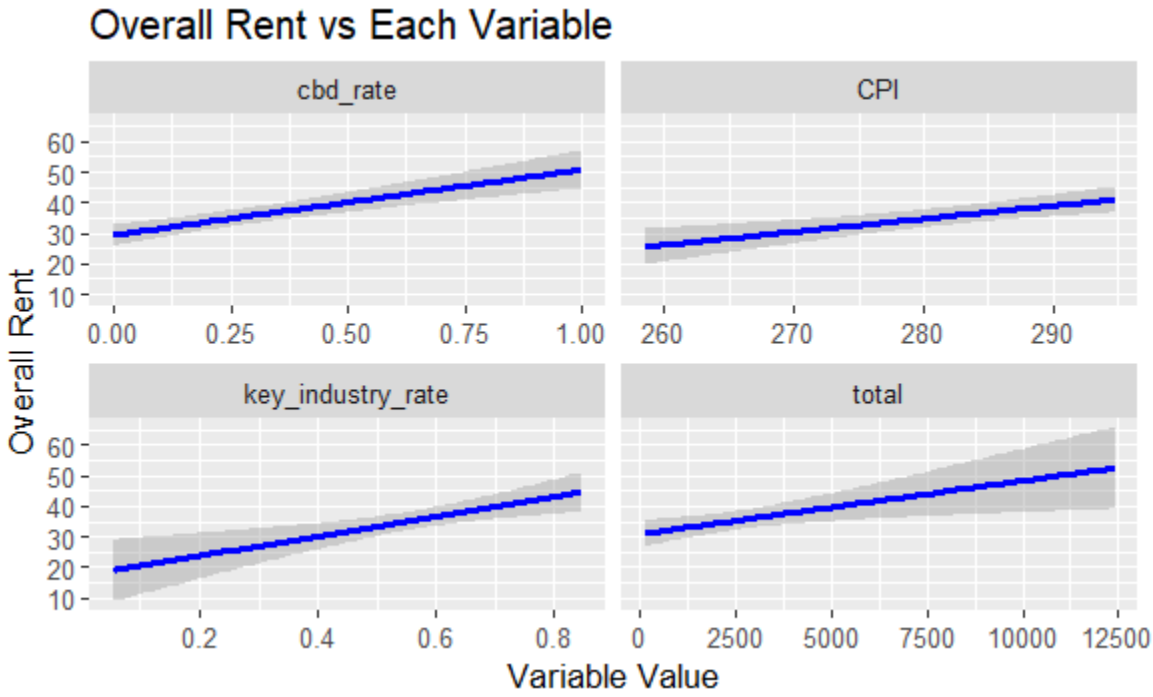
```
. reg overall_rent count total key_industry_rate cbd_rate cpi
```

Source	SS	df	MS		
				Number of obs	= 60
				F(5, 54)	= 15.75
Model	5769.97642	5	1153.99528	Prob > F	= 0.0000
Residual	3955.31417	54	73.2465587	R-squared	= 0.5933
				Adj R-squared	= 0.5556
Total	9725.29059	59	164.835434	Root MSE	= 8.5584

overall_rent	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
count	.0098412	.0035872	2.74	0.008	.0026493	.017033
total	-.0000836	.000705	-0.12	0.906	-.0014969	.0013298
key_industry_rate	21.23604	8.330908	2.55	0.014	4.533578	37.93851
cbd_rate	11.10938	3.783527	2.94	0.005	3.523868	18.6949
cpi	.3370306	.0903194	3.73	0.000	.1559512	.51811
_cons	-77.55115	24.73094	-3.14	0.003	-127.1337	-27.9686



The model achieved a **moderate R-squared of 0.5933**, indicating decent explanatory power.
However, the **p-value for total was 0.906**, suggesting it was not statistically significant in this new context.
Conversely, **cbd_rate now became highly significant (p = 0.005)**, indicating its relevance at more granular spatial scales.



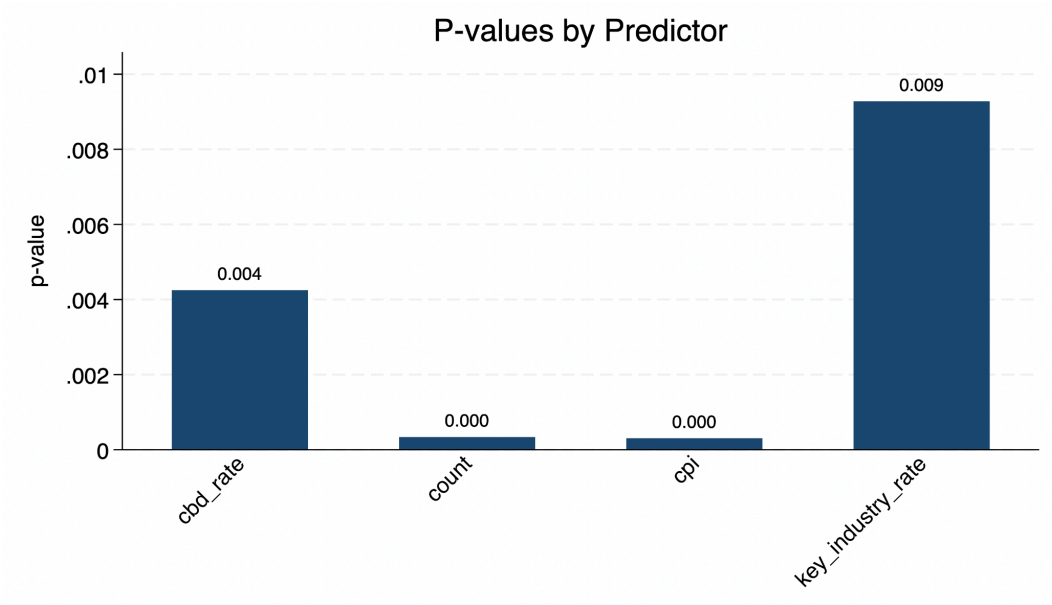
We removed the Total variable and regressed it again.

Variables included:
count, key_industry_rate, cbd_rate, cpi

```
. reg overall_rent count key_industry_rate cbd_rate cpi
```

Source	SS	df	MS	Number of obs	=	60
				F(4, 55)	=	20.05
Model	5768.94724	4	1442.23681	Prob > F	=	0.0000
Residual	3956.34335	55	71.9335154	R-squared	=	0.5932
				Adj R-squared	=	0.5636
Total	9725.29059	59	164.835434	Root MSE	=	8.4814

overall_rent	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
count	.0095382	.0024941	3.82	0.000	.00454	.0145364
key_industry_rate	20.89518	7.748501	2.70	0.009	5.366842	36.42353
cbd_rate	11.14658	3.736544	2.98	0.004	3.658379	18.63478
cpi	.3390355	.0879228	3.86	0.000	.1628343	.5152367
_cons	-78.03058	24.17828	-3.23	0.002	-126.4849	-29.57623



We re-estimated the model **without total**, which led to:

- An essentially **unchanged R-squared (0.5932)**
- Improved **adjusted R-squared (0.5636)** due to reduced noise
- **All remaining predictors became statistically significant at the 1% level or better.**

Conclusion

In our initial model using state-level data, total — the total number of buildings — appeared to play a key role in explaining overall rent. However, as we moved to a more granular ZIP-code-level analysis, total lost its significance, while cbd_rate became a much stronger predictor. This shift highlights how **the importance of variables can change depending on the spatial scale of analysis**: supply factors dominate at broader levels, while centrality and urban density become more influential locally.

Our ultimate objective was to identify which variables have the **strongest and most direct impact on rent**. Based on statistical significance and explanatory power, we conclude that **count, key_industry_rate, cbd_rate, and cpi** are the most robust predictors in our final model. These findings provide valuable insights into **what truly drives rent variation**, helping individuals, businesses, and policymakers make better location-based decisions.

3. Quality of Life vs Overall_rent - non economic value

Variable Explanations:

- cbd_rate: Percentage of buildings in the city area relative to the total number of buildings
- class_A_rate: The quality of buildings
- crime.score: Violent crime rate = (A state's number of crimes (Homicide, Rape, Robbery, Aggravated Assaulted) / population) * 100,000
- education: Education score
- health.value: Health score
- overall_rent: Annual Rent Fee per Square Foot
- Median.Household.Income: Median Household Income
- pop_density: The number of people per square mile
- Population.Change.Rate: Population Growth Rate
- retail.score: The number of each state's major retailers (Costco, Target, Walmart) / state's area

These 9 factors were used to determine the quality of life and we used linear regression to determine how much the quality of life of an area correlates to the overall rent. This allows for a closer look into other factors that may affect the overall rent. This shows that not all factors affect the overall rent and that the economic factors have a stronger and more important impact on the overall rent.

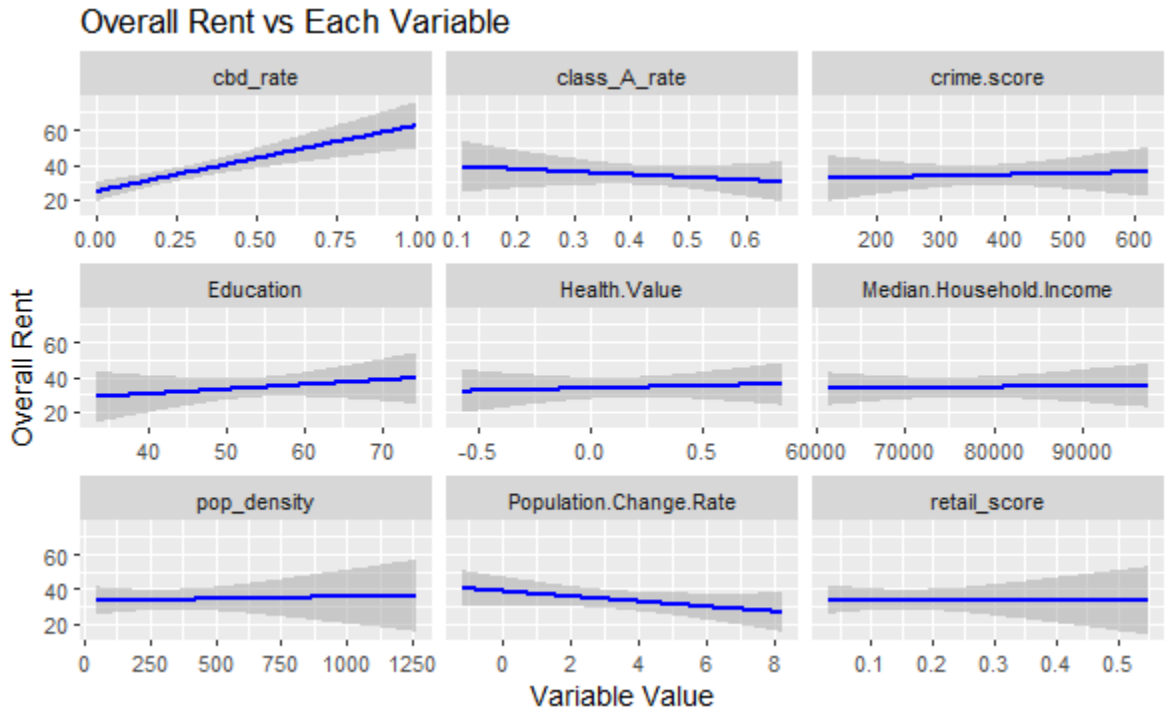
External Sources:

Retail.score - Costco, Walmart, and Targets: **Kaggle**
health.value: **America's Health Rankings**
education: **WalletHub**
pop_density: **statsamerica(populations), statesymbolsusa(states by square mile)**
Crime.score: **FBI**
Population.Change.Rate: **U.S. Census Bureau**
Median.Household.Rate: **Federal Reserve Bank of St. Louis**

```
. reg overall_rent populationchangerate medianhouseholdincome class_a_rate education crimescore healthvalue retail_score pop_density cbd_rate
```

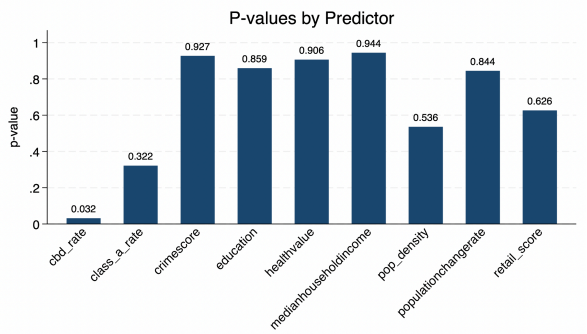
Source	SS	df	MS		
				Number of obs	= 20
				F(9, 10)	= 1.87
Model	1702.79816	9	189.199795	Prob > F	= 0.1720
Residual	1012.5846	10	101.25846	R-squared	= 0.6271
				Adj R-squared	= 0.2915
Total	2715.38276	19	142.914882	Root MSE	= 10.063

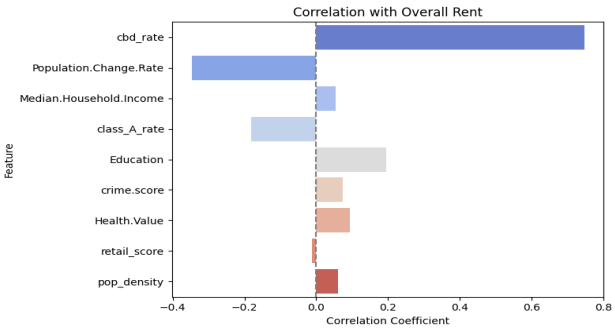
overall_rent	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
populationchangerate	-.2248374	1.115737	-0.20	0.844	-2.710855	2.261181
medianhouseholdincome	-.000036	.0005019	-0.07	0.944	-.0011542	.0010823
class_a_rate	-21.59209	20.71422	-1.04	0.322	-67.74626	24.56207
education	.0927095	.5094195	0.18	0.859	-1.042348	1.227767
crimescore	-.0029696	.0316221	-0.09	0.927	-.0734281	.0674889
healthvalue	2.001869	16.58232	0.12	0.906	-34.94585	38.94959
retail_score	-46.93961	93.35578	-0.50	0.626	-254.9493	161.07
pop_density	.0233028	.0363427	0.64	0.536	-.0576738	.1042794
cbd_rate	34.69586	13.88644	2.50	0.032	3.754944	65.63677
_cons	34.06027	39.68521	0.86	0.411	-54.36388	122.4844



As seen in these graphs, except for cbd_rate and education, all of the graphs have a slope very close to 0 (close to horizontal lines). This explains how the factors (quality of life) considered do not have a strong correlation with the overall rent. So every factor, except

for cbd_rate and education, don't explain overall rent very well. So we decided to find the correlation between overall rent and each of the variables to examine the exact correlation between the variables and overall rent.





As examined earlier the new correlation tests, except for cbd_rate and education, explain that the other variables don't majorly affect the overall rent.

From these observations made, even though we used several features that explained quality of life to determine if quality of life as a whole had an effect on overall rent the observations show that there is no major correlation between quality of life and overall rent. So, quality of life has little to no effect on overall rent.

Factors that affect overall rent are economic factors like cbd_rate, not quality of life factors.

External Sources

Population.Change.Rate:
https://www.census.gov/library/visualizations/2023/comm/percent-change-state-population.html

Median.Household.Rate: https://fred.stlouisfed.org/series/MEHOINUSA646N

CPI: https://www.bls.gov/regions/subjects/consumer-price-indexes.htm

Retail.score
Costco: https://www.kaggle.com/datasets/polartech/complete-store-locations-of-costco
Walmart: https://www.kaggle.com/datasets/jackogozaly/us-walmart-store-locations
Targets: https://www.kaggle.com/datasets/saejinmahlauheinert/target-store-locations

health.value: https://www.americashealthrankings.org/explore/measures/Overall

education: https://wallethub.com/edu/e/states-with-the-best-schools/5335

pop_density:
Population: https://www.statsamerica.org/sip/rank_list.aspx?rank_label=pop1&ct=S18,
Size of States:
https://statesymbolsusa.org/symbol-official-item/national-us/uncategorized/states-size

Crime.score: https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/home