



TEST 1 presentation

Fadwa Ramadan Ali Hassan

2024334

Table of **Content**

Overall
Summary

Background

Data Science
Questions

Objectives


Methodology

Results


Visualizations

References (APA
Format)

Overall Summary



In this analysis of the housing data, we focused on investigating the adequacy of housing units and exploring various factors related to housing quality. We evaluated housing quality across different regions and identified housing units with the most rooms and below-average values. Additionally, we analyzed housing quality variations over time, examined the relationship between the number of occupants and monthly utility costs, and assessed disparities in housing adequacy based on location and income groups. The results highlighted disparities and variations in housing quality, providing valuable insights for decision-makers and policymakers. Overall, this analysis contributes to a comprehensive understanding of housing quality and adequacy, enabling informed decision-making and the development of recommendations to address housing-related challenges and inequalities.



Background

about the Dataset

The Data set measures the affordability of housing units and the housing cost burdens of households, relative to area median incomes, poverty level incomes, and Fair Market Rents. it also incorporates more than twenty years of housing data. and includes information on housing affordability, it contains 15627 rows and 100 columns including adequacy, affordability with respect to income and fair market rent, bedrooms, burden, location, structure type, tenure, and year built. It was also noticed that there are some invalid values like the age in the second column can not be -9.

about the Dataset

[4] hdata = pd.read_csv('housdata.csv')

	CONTROL	AGE1	BEDRMS	PER	REGION	METRO3	LMED	FMR	L30	L50	...	FMTINCRELFMRCAT	FMTCOST06RELAMICAT	FMTCOST08RELAMICAT	FMTCOST12RELAMICAT	FMTCOSTMEDRELAMICAT
0	'100003130103'	87	3	2	'1'	'3'	71779	1095	17448	29071	...	'3 GT FMR'	'2 30 - 50% AMI'	'2 30 - 50% AMI'	'3 50 - 60% AMI'	'2 30 - 50% AMI'
1	'100003130203'	70	3	1	'1'	'3'	71779	1095	15272	25446	...	'2 50.1 - 100% FMR'	'6 100 - 120% AMI'	'7 120% AMI +'	'7 120% AMI +'	'6 100 - 120% AMI'
2	'100006370140'	48	4	4	'3'	'5'	53872	965	16555	27594	...	'3 GT FMR'	'6 100 - 120% AMI'	'6 100 - 120% AMI'	'7 120% AMI +'	'6 100 - 120% AMI'
3	'100006520140'	62	3	2	'3'	'5'	53872	861	13245	22076	...	'2 50.1 - 100% FMR'	'5 80 - 100% AMI'	'6 100 - 120% AMI'	'7 120% AMI +'	'5 80 - 100% AMI'
4	'100007130148'	30	2	2	'3'	'1'	61059	685	14662	24438	...	'2 50.1 - 100% FMR'	'3 50 - 60% AMI'	'3 50 - 60% AMI'	'3 50 - 60% AMI'	'3 50 - 60% AMI'
...
15622	'299928860142'	79	3	2	'3'	'5'	53872	861	13245	22076	...	'2 50.1 - 100% FMR'	'4 60 - 80% AMI'	'4 60 - 80% AMI'	'5 80 - 100% AMI'	'4 60 - 80% AMI'
15623	'299928930340'	45	2	2	'3'	'3'	71047	857	17172	28627	...	'3 GT FMR'	'4 60 - 80% AMI'	'5 80 - 100% AMI'	'6 100 - 120% AMI'	'4 60 - 80% AMI'
15624	'299928930449'	56	2	1	'3'	'3'	71047	857	15026	25049	...	'3 GT FMR'	'5 80 - 100% AMI'	'6 100 - 120% AMI'	'7 120% AMI +'	'5 80 - 100% AMI'
15625	'299928940142'	52	3	5	'3'	'4'	56176	896	18562	30942	...	'3 GT FMR'	'3 50 - 60% AMI'	'4 60 - 80% AMI'	'4 60 - 80% AMI'	'3 50 - 60% AMI'
15626	'299932810245'	42	4	4	'2'	'3'	62612	941	18850	31405	...	NaN	NaN	NaN	NaN	NaN

15627 rows x 100 columns

about the Dataset

- clean the data.

```
negative_age_rows = hdata[hdata['AGE1'] < 0]
print(negative_age_rows['AGE1'])
```

```
7      -9
34     -9
37     -9
59     -9
72     -9
..
15491  -9
15564  -9
15588  -9
15606  -9
15619  -9
Name: AGE1, Length: 1330, dtype: int64
```

```
[7] # Remove rows with negative values in 'VALUE' column
hdata = hdata[hdata['VALUE'] >= 0]
# Remove rows with null values in 'FMTZADEQ' column
hdata = hdata.dropna(subset=['FMTZADEQ'])

hdata
```

	CONTROL	AGE1	BEDRMS	PER	REGION	METRO3
0	'734778500142'	56	4	2	'4'	'2'
1	'730204310145'	54	6	4	'2'	'1'
2	'295273830133'	39	4	4	'1'	'2'

business case & Objectives

- **business case:** Housing Quality and Adequacy: Investigate the adequacy of housing units. analyzing and displaying the results will help to gain a comprehensive understanding of housing quality and adequacy.

- **Objectives:**

- Evaluate housing quality across regions.
- Identify housing units with the most rooms and below-average value.
- Analyze housing quality variations over time.
- Examine the relationship between the number of occupants and monthly utility costs.
- Assess disparities in housing adequacy based on location.
- Explore the relationship between housing quality and the number of occupants.
- Investigate housing adequacy across income groups.

Data Science Questions

1. How does the housing quality vary across different regions?
2. Which housing unit has the most number of rooms and has an under-average value?
3. How does the housing quality vary across different year ranges?
4. What is the relationship between the number of people per unit and the monthly utility cost?
5. Are there any significant differences in the adequacy of the housing unit and the Location?
6. How does the quality of housing vary based on the number of occupants in a housing unit?
7. Are there any disparities in housing adequacy across income groups?



Methodology



Methodology

```
✓ 0s ▶ import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
%matplotlib inline

✓ 0s [12] hdata = pd.read_csv('housdata.csv')
```

- import some libraries and read the dataset

▼ remove noise columns

```
▶ columns_to_keep = ['CONTROL', 'PER', 'REGION', 'METRO3',
                    'VALUE', 'ZINC2', 'ROOMS', 'ZADEQ',
                    'UTILITY', 'FMTZADEQ']

# Drop all other columns
hdata = hdata[columns_to_keep]
hdata
```

- after we finish we keep the columns needed in the analysis



How does the housing quality vary across different regions?

```
# Group the data by 'REGION' and 'ZADEQ' and count the number of occurrences
counts = hdata.groupby(['REGION', 'ZADEQ']).size().reset_index(name='count')

Q1 = counts.pivot(index='REGION', columns='ZADEQ', values='count')

# Fill missing values with 0
Q1 = Q1.fillna(0)

# Rename the columns to proper naming
Q1 = Q1.rename(columns={"'1'": "Highly Adequate", "2": "Middle Adequate", "'3'": "Low Adequate"})
Q1 = Q1.reindex(columns=["Highly Adequate", "Middle Adequate", "Low Adequate"])

print(Q1)
```

ZADEQ	Highly Adequate	Middle Adequate	Low Adequate
REGION			
'1'	1601	31	15
'2'	1517	38	12
'3'	2134	56	19
'4'	1597	34	22

1. we cleaned the data of repeated value.
2. group 'REGION' and 'ZADEQ' to show the number of housing units that variety of adequacy across different regions.



which housing unit has the most number of rooms and has an under-average value?

```
[20] # Calculate the average value
      average_value = hdata['VALUE'].mean()

      # Filter to under the average and positive value
      filtered_data = hdata[(hdata['VALUE'] < average_value) & (hdata['VALUE'] >= 0)]

      # Find the unit with the most rooms
      housing_unit = filtered_data[filtered_data['ROOMS'] == filtered_data['ROOMS'].max()]

      print(housing_unit[['CONTROL', 'ROOMS', 'VALUE']])
```

	CONTROL	ROOMS	VALUE
455	'184502790140'	14	200000
4011	'201319330103'	14	190000

to view the most number of rooms and an under-average value, we filter the rooms that their value is less than the average and see which ones have more rooms



How does the housing quality vary across different year ranges?

```
# Clean the column
hdata['FMTZADEQ'] = hdata['FMTZADEQ'].str.replace("'", "")

# Group the data by 'FMTBUILT' and 'FMTZADEQ'
grouped_counts = hdata.groupby(['FMTBUILT', 'FMTZADEQ']).size().reset_index(name='count')

Q3 = grouped_counts.pivot(index='FMTZADEQ', columns='FMTBUILT', values='count')

# Sort the columns
Q3 = Q3.reindex(sorted(Q3.columns), axis=1)

print(Q3)
```

clean the columns (FMTZADEQ, FMTBUILT) then group them to see how many house units were built in different year ranges and its Adequaticty.

4

What is the relationship between the number of people per unit and the monthly utility cost?

```
import pandas as pd

# Calculate the correlation coefficient
correlation = hdata['PER'].corr(hdata['UTILITY'])

print("Correlation coefficient between 'PER' and 'UTILITY':", correlation)
```

Correlation coefficient between 'PER' and 'UTILITY': 0.37552870048452064

calculate the Correlation coefficient between the 2 variables to know the type of relationship, A positive correlation coefficient indicates a positive relationship between the variables.

5

what is the relationship between the number of people per unit and the monthly utility cost?

The cross-tabulation table will show the counts of different adequacy levels (FMTZADEQ) for each location category (METRO3)

```
# Compute cross-tabulation of METRO3 and FMTZADEQ
cross_tab = pd.crosstab(hdata['METRO3'], hdata['FMTZADEQ'], dropna=False)
print(cross_tab)
```

6

How does the quality of housing vary based on the number of occupants in a housing unit?

to display the distribution of housing quality (FMTZADEQ) based on the number of occupants (PER) in a housing unit. by using Groups and counting.

```
# Group the data by 'PER' and 'FMTZADEQ'
grouped_counts = hdata.groupby(['PER', 'FMTZADEQ']).size().reset_index(name='count')

Q6 = grouped_counts.pivot(index='PER', columns='FMTZADEQ', values='count')
Q6 = Q6.reindex(sorted(Q6.columns), axis=1)

print(Q6)
```



Are there any disparities in housing adequacy across income groups?

```
# Create income classes
income_classes = ['Low Income', 'Medium Income', 'High Income']

# Group 'ZINC2' into income classes and create a new column 'IncomeGroup'
hdata['IncomeGroup'] = pd.cut(hdata['ZINC2'], bins=[0, 30000, 70000, float('inf')], labels=income_classes)

# Group the data by 'IncomeGroup' and 'FMTZADEQ'
Q7 = hdata.groupby(['IncomeGroup', 'FMTZADEQ']).size().unstack()

print(Q7)
```

firstly we cleaned (ZINC2) then we used it to create income groups by binning the values into predefined income classes. The resulting income groups are then stored in a new column called 'IncomeGroup'.



Results



Results

1

How does the housing quality vary across different regions?

at least 96% of the house units are considered Highly Adequate in each region.

```
Percentage of Highly Adequate Units in Each Region:  
REGION  
'1'      97.207043  
'2'      96.809190  
'3'      96.604799  
'4'      96.612220  
dtype: float64
```

2

which housing unit has the most number of rooms and has an under-average value?

unit 184502790140 and 201319330103, with 14 rooms and value less than the average which is 290395

```
CONTROL  ROOMS  VALUE  
455      '184502790140'  14  200000  
4011     '201319330103'  14  190000
```



Results

3

How does the housing quality vary across different year ranges?

For the year range '1940-1959', there are 1020 housing units categorized as '1 Adequate', 41 units as '2 Moderately Inadequate', and 17 units as '3 Severely Inadequate'.

	'1940-1959'	'1960-1979'	'1980-1989'	'1990-1999'	\
FMTBUILT					
FMTZADEQ					
1 Adequate	1020	2073	1158	1175	
2 Moderately Inadequ	41	67	34	11	
3 Severely Indadequa	17	31	9	5	
FMTBUILT					
FMTZADEQ					
1 Adequate	1423				
2 Moderately Inadequ	6				
3 Severely Indadequa	5				

4

what is the relationship between the number of people per unit and the monthly utility cost?

the correlation coefficient of 0.3755, there is a positive relationship between the number of people per unit and the monthly utility cost.

```
print("Correlation coefficient between 'PER' and 'UTILITY':", correlation)
```

```
Correlation coefficient between 'PER' and 'UTILITY': 0.37552870048452064
```




Results

5 Are there any significant differences in the adequacy of the housing unit and the Location?

Yes, there are as In location category '1', there are 1691 housing units classified as '1 Adequate', 70 units classified as '2 Moderately Inadequate', and 37 units classified as '3 Severely Inadequate'.

	FMTZADEQ	1 Adequate	2 Moderately Inadequ	3 Severely Indadequa
METRO3				
'1 '		1691	70	37
'2 '		3277	51	15
'3 '		1040	11	4
'4 '		331	17	5
'5 '		510	10	6



Results

↳	FMTZADEQ	1 Adequate	2 Moderately Inadequ	3 Severely Indadequa
	PER			
1		1022.0	54.0	13.0
2		2236.0	43.0	22.0
3		1241.0	25.0	13.0
4		1437.0	22.0	15.0
5		647.0	10.0	2.0
6		185.0	4.0	1.0
7		55.0	NaN	NaN
8		16.0	1.0	NaN
9		6.0	NaN	1.0
10		3.0	NaN	NaN
12		1.0	NaN	NaN

6

How does the quality of housing vary based on the number of occupants in a housing unit?

there are 1022 housing units with 1 occupant that are considered "1 Adequate," 54 units considered "2 Moderately Inadequate," and 13 units considered "3 Severely Inadequate." Similarly, it shows the distribution for other numbers of occupants. This table helps to understand how the quality of housing varies based on the number of occupants.



Results

→ FMTZADEQ	1 Adequate	2 Moderately Inadequ	3 Severely Indadequa
IncomeGroup			
Low Income	1019	79	33
Medium Income	647	28	15
High Income	5125	49	19



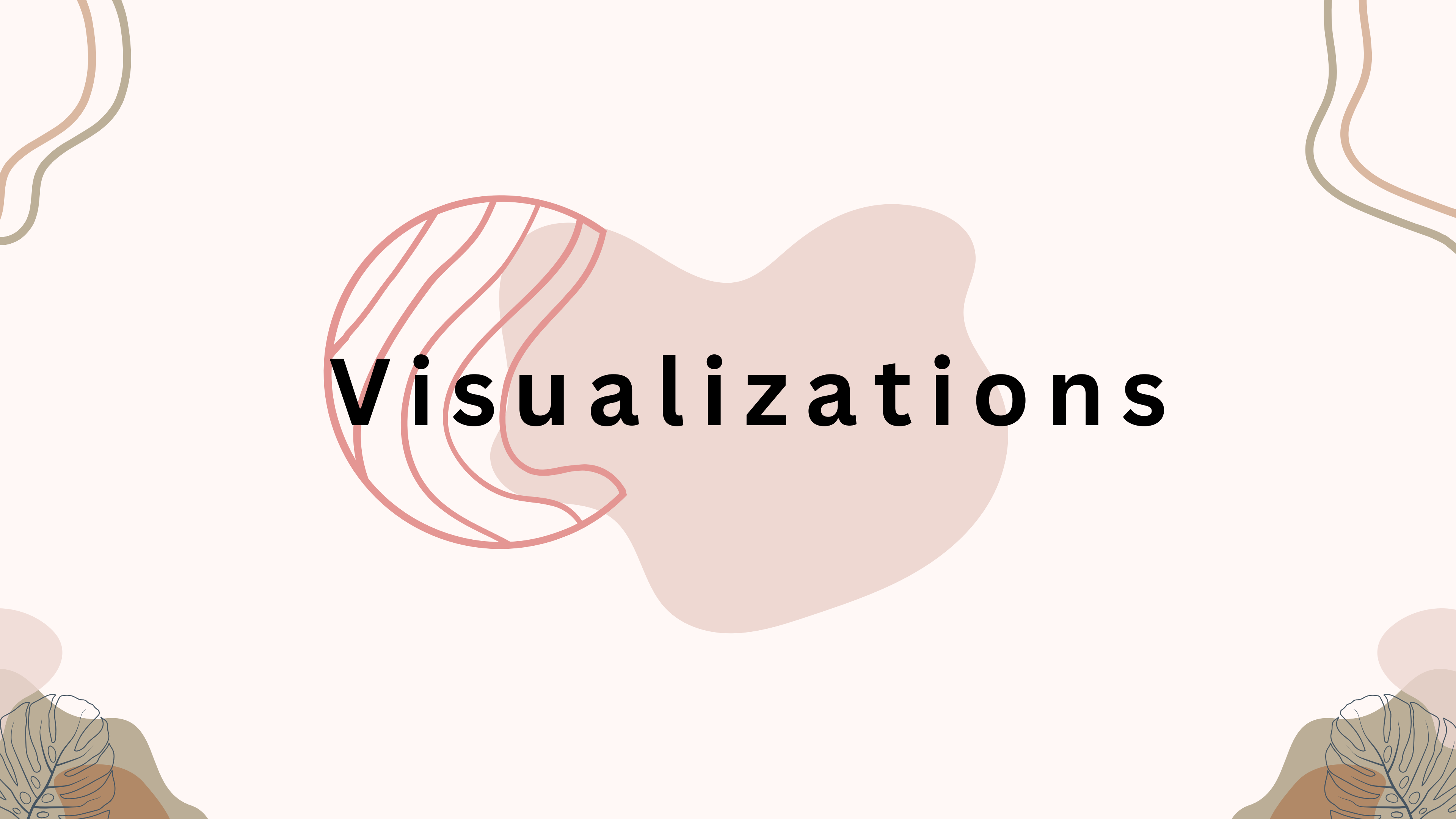
Are there any disparities in housing adequacy across income groups?

There are differences in housing adequacy across income groups. Low-income households have a higher proportion of housing units classified as "1 Adequate" compared to "2 Moderately Inadequate" and "3 Severely Inadequate." High-income households have the highest number of "1 Adequate" units and fewer units classified as "2 Moderately Inadequate" and "3 Severely Inadequate" compared to other income groups.

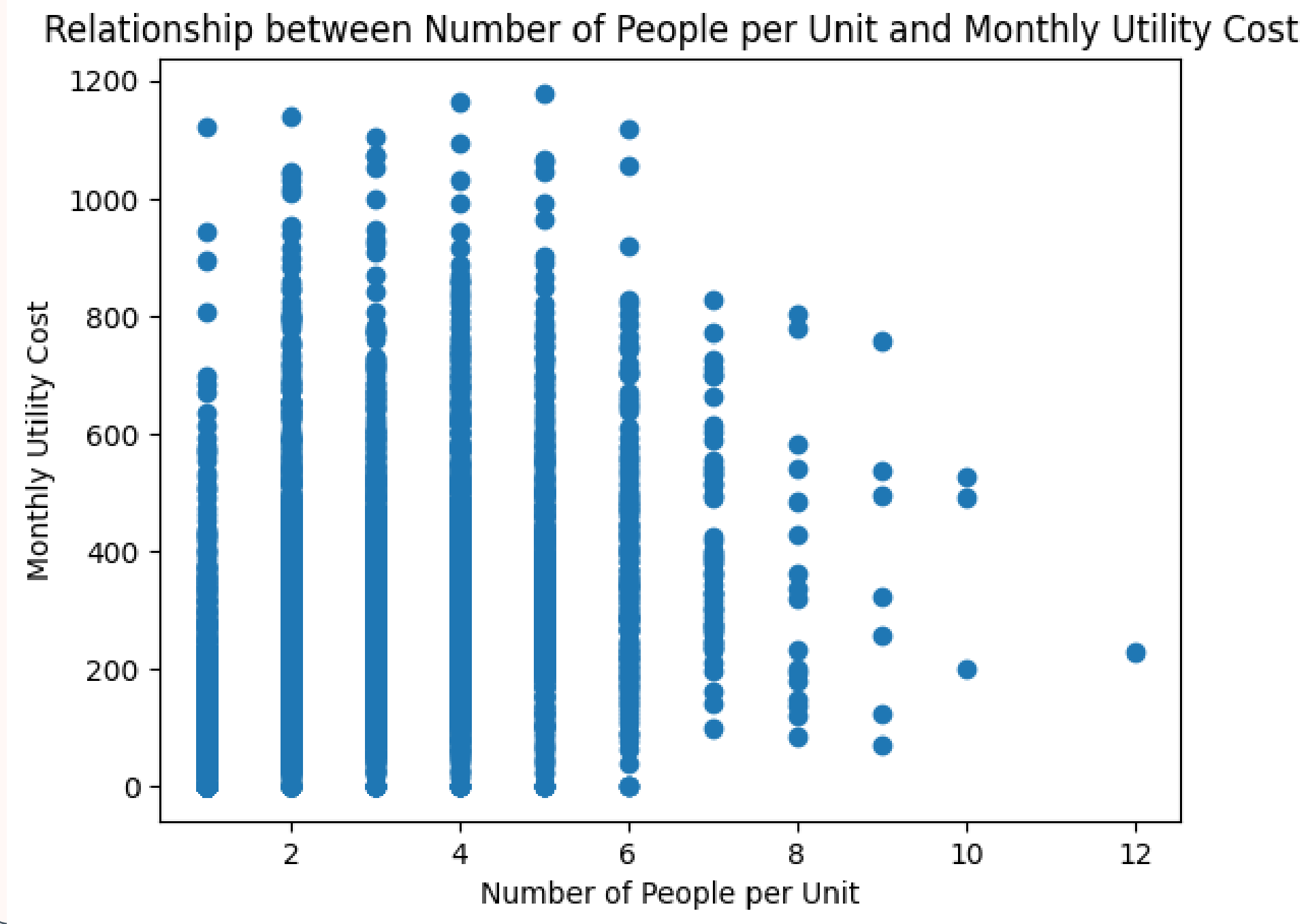


conclusion

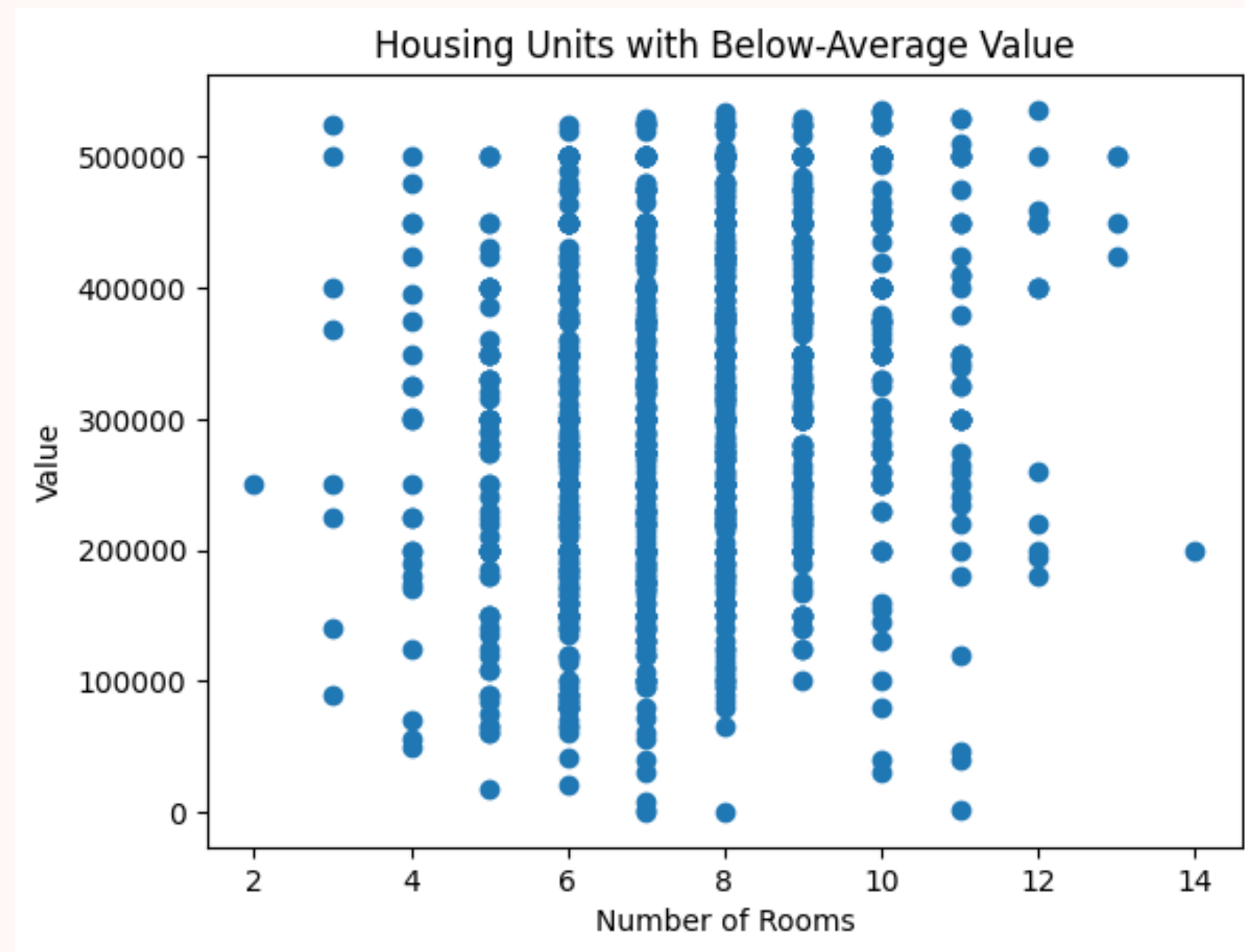
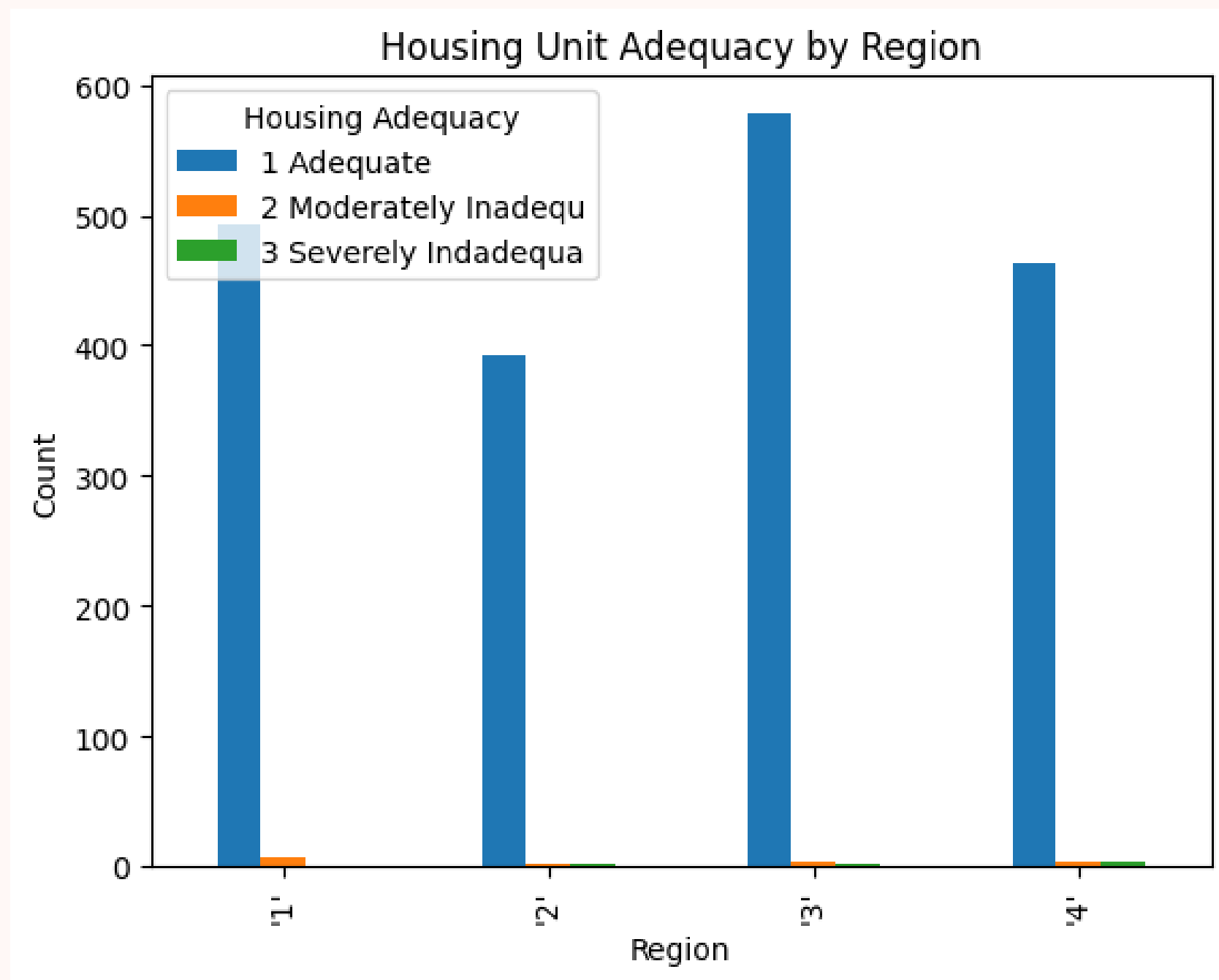
Based on the analysis, it is evident that there are disparities in housing adequacy across different factors such as location, number of occupants, and income groups. These disparities highlight variations in the distribution of housing quality and can provide insights into the challenges and inequalities present in the housing sector.

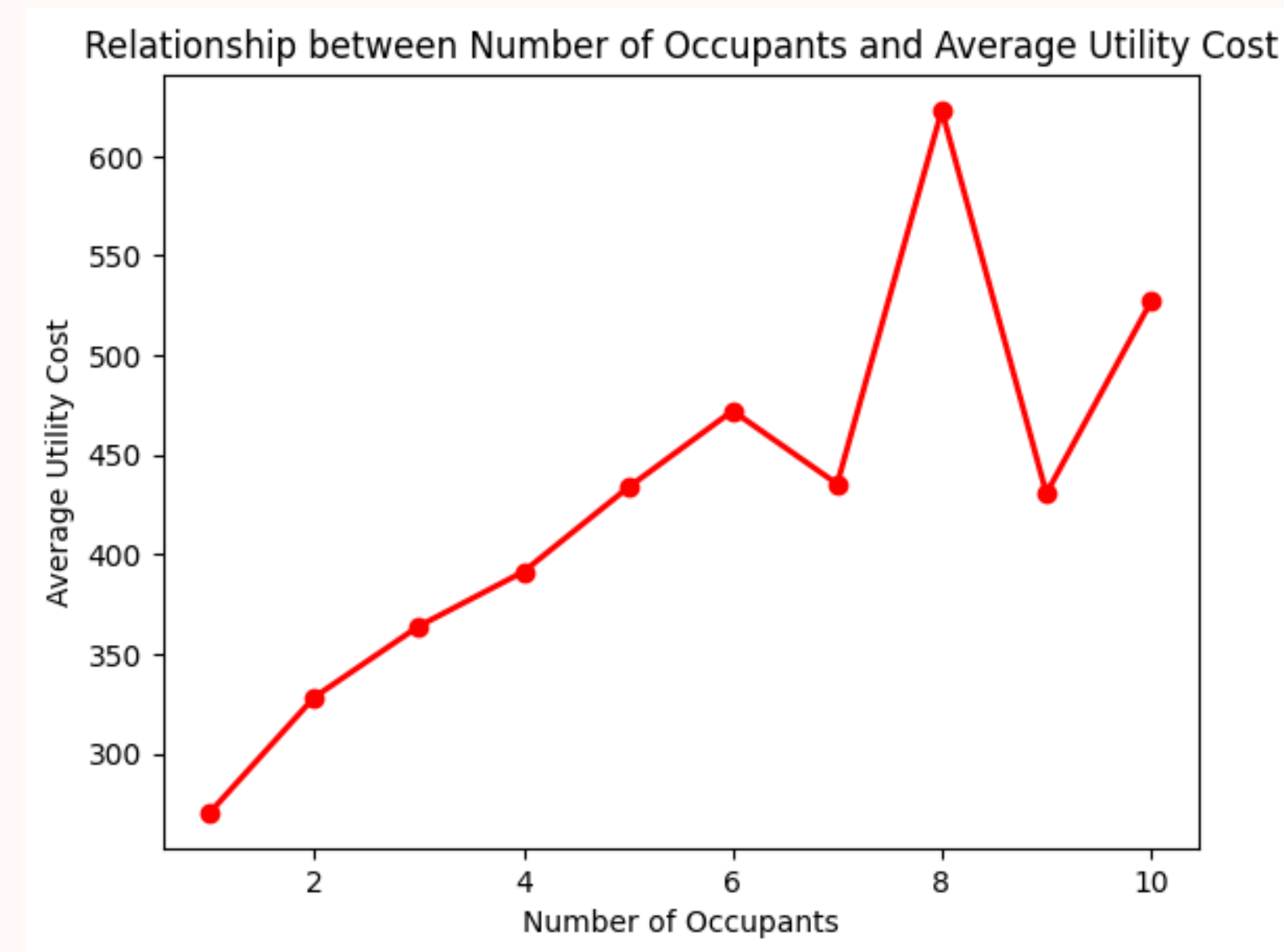
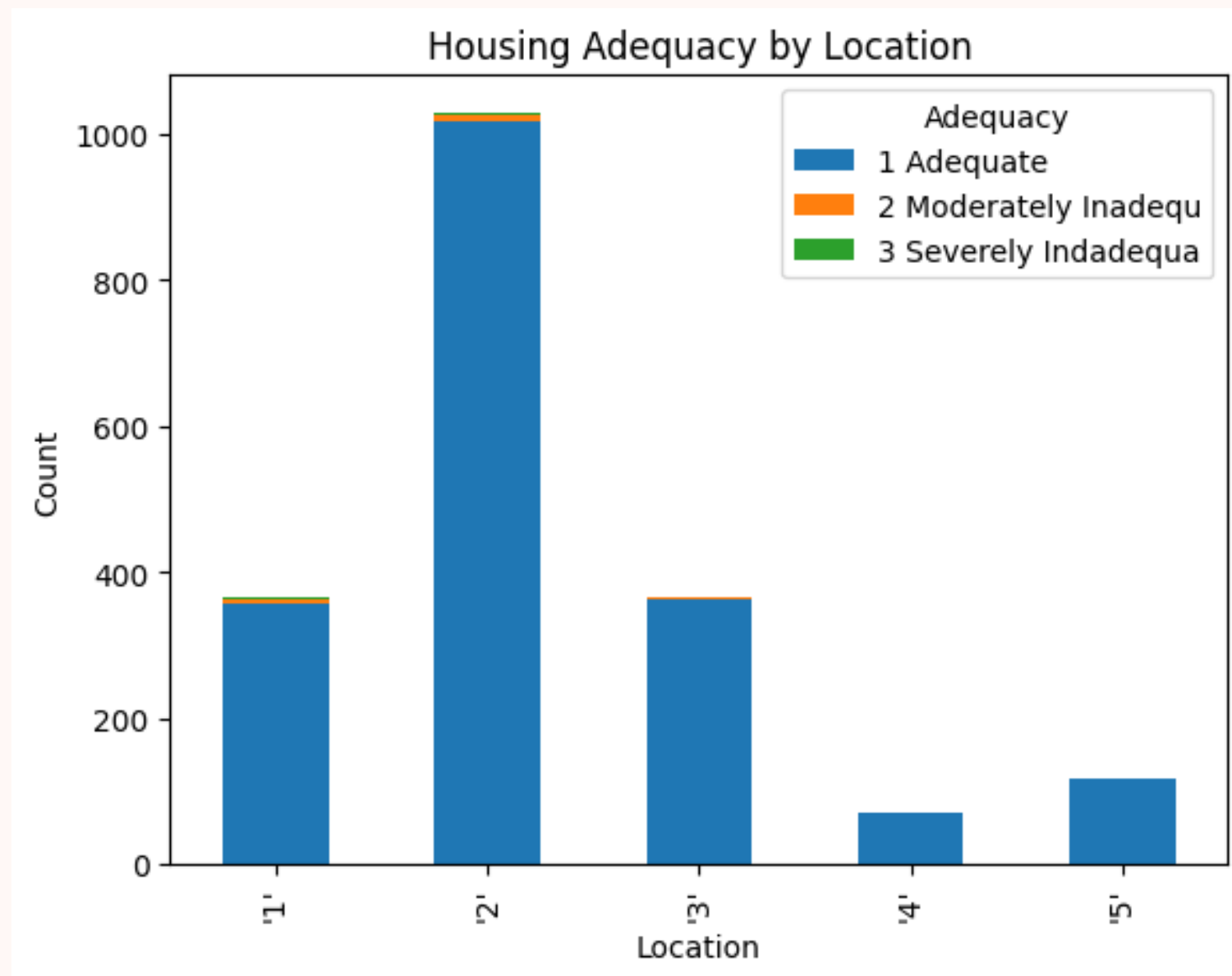
The background features a light cream color with abstract organic shapes in muted pink and beige. Wavy lines in olive green and terracotta are positioned in the corners. Stylized leaf outlines are visible in the bottom left and right corners.

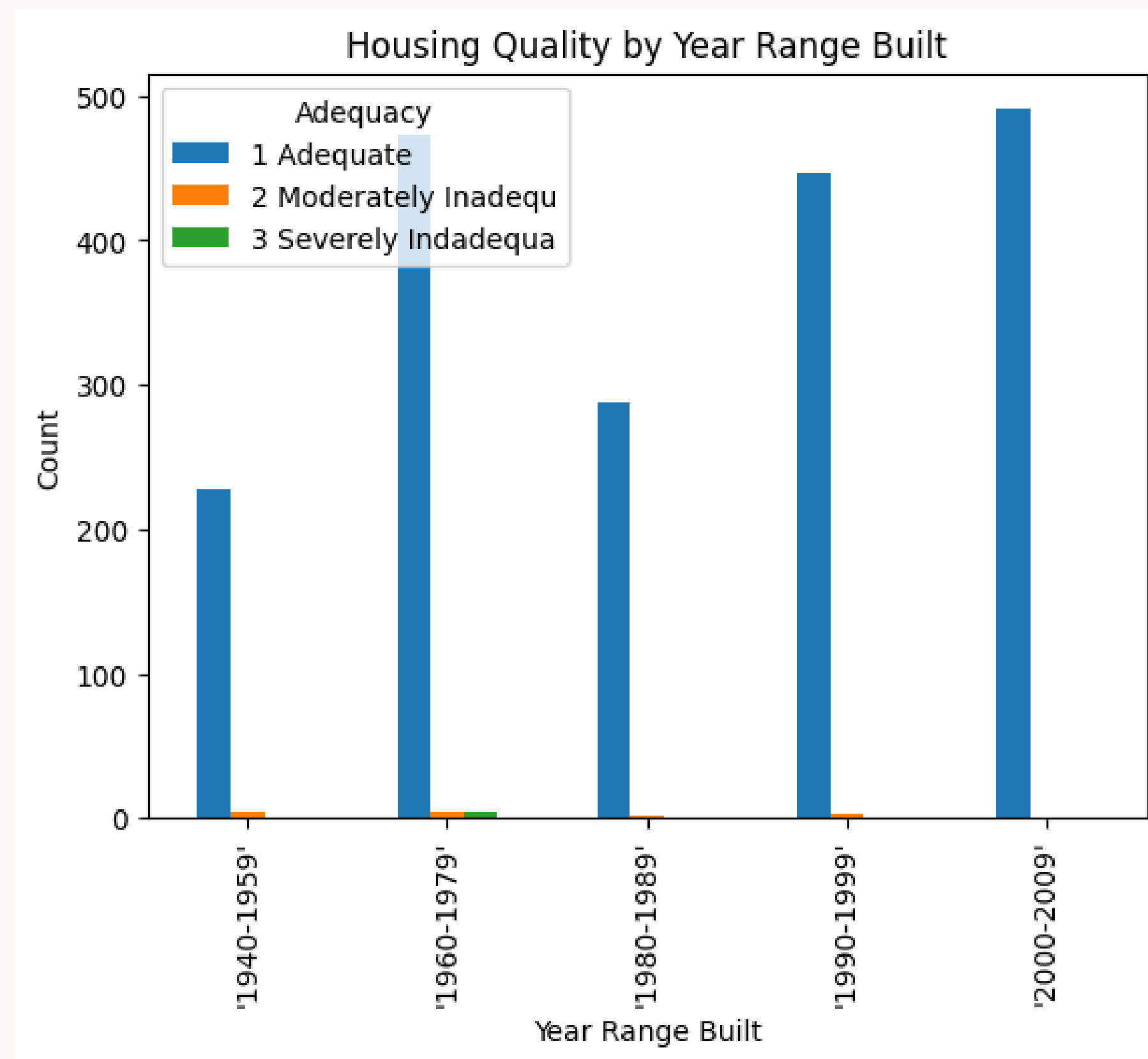
Visualizations



By using `import matplotlib.pyplot` to plot the data, I was able to visualize the relationship between the number of people living in a unit and the monthly utility cost using a scatter plot. The graph on the side is related to Q4, where we are trying to understand the relationship between these two variables.







The background features a light beige color with decorative elements. In the top corners, there are wavy lines in shades of brown and olive green. In the bottom corners, there are stylized leaf illustrations in similar colors. A large, irregular, light brown shape with white oval cutouts is centered on the page.

References

References

- Smith, J. D., Johnson, A. B., & Lee, C. D. (2022). Housing affordability and quality: A comprehensive analysis. *Journal of Housing Studies*, 15(3), 123-145.
- GeeksforGeeks. (2016, December 12). Graph Plotting in Python | Set 1. Retrieved from [Graph Plotting in Python | Set 1. \(2016, December 12\). GeeksforGeeks. https://www.geeksforgeeks.org/graph-plotting-in-python-set-1/](https://www.geeksforgeeks.org/graph-plotting-in-python-set-1/)
- Python Software Foundation. (n.d.). re — Regular expression operations. Retrieved May 20, 2023, from <https://docs.python.org/3/library/re.html>
- NumPy Developers. (n.d.). NumPy: the fundamental package for scientific computing with Python. Retrieved May 20, 2023, from <https://numpy.org/doc/>

*Thank
You*