AUTHORS

Muskaan Patel
Nazanin Ahmadi
Zachary Weachter

Crime Rates and Socioeconomic Factors in U.S. Cities

Team coolerthanicecream!

Introduction

Understanding how socioeconomic factors influence crime rates in cities is essential for creating safer and more prosperous communities. High crime rates not only affect public safety but also economic growth and social cohesion. By studying factors like income levels, population density, property crimes, and housing costs, we can uncover patterns that help inform strategies for crime prevention and urban development. Using statistical tests and machine learning, this research aims to reveal insights into urban dynamics, guiding evidence-based decision-making for building better cities.

Hypothesis

- The first hypothesis suggests that **higher median family incomes** in cities lead to **lower crime rates**, as a greater economy may result in increased investment in crime prevention, potentially deterring criminal behaviour.
- The second hypothesis examines whether **denser populations** in cities correlate with **higher crime rates**, grounded in the understanding that these areas face unique challenges.
- The third hypothesis explores the relationship between **property crime rates** and **housing costs**, suggesting that cities with higher property crime rates also tend to have higher housing costs due to the potential attraction of affluent residents.

Data

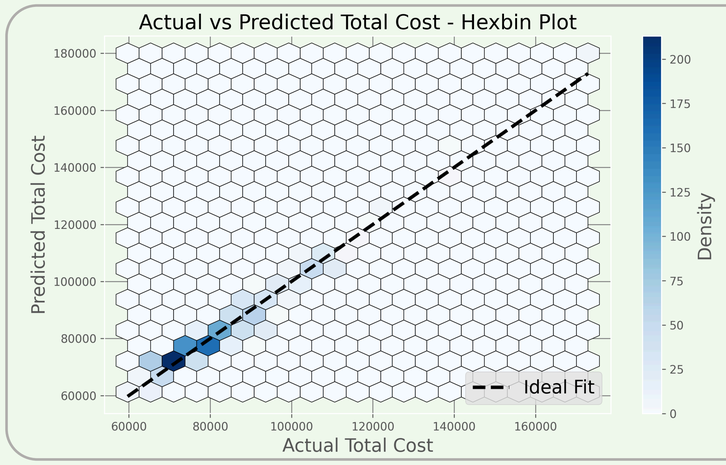
1. The US Family Budget Dataset provides detailed estimates of the cost of living for various family types across nearly 1877 counties and metro areas in the United States.
2. The United States Cities Database offers comprehensive demographic and geographical information for cities and towns, sourced from authoritative sources such as the U.S. Geological Survey and U.S. Census Bureau.
3. The 2019 crime in the United States dataset provides essential crime statistics, enabling analysis of crime rates across different urban areas.

Hypothesis Testing

- **Hypothesis 1** - The t-test did not conclusively reject the hypothesis, but showed a trend that cities with higher median family incomes tend to have lower crime rates. The Chi-squared test indicated a significant association between income levels and crime rate categories.
- **Hypothesis 2** - An ANOVA (Analysis of Variance) test was used to compare crime rates across different population density categories. ANOVA is used to compare the means of several groups. The test did not find a statistically significant relationship between density and crime rate.
- **Hypothesis 3** - A chi-squared test was used to see if there's a connection between property crime rates and housing costs, finding a significant association between the two.

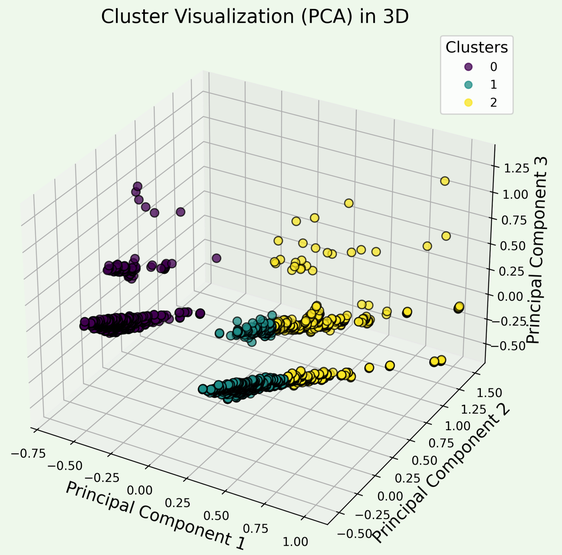
Hyp.	Test	Stat value	p-value
1	T-test	-0.7445	0.4566
1	Chi-squared	58.8418	1.71e-14
2	ANOVA	0.0619	0.99
3	Chi-squared	451.93	1.66e-96

Machine Learning Results & Visualization



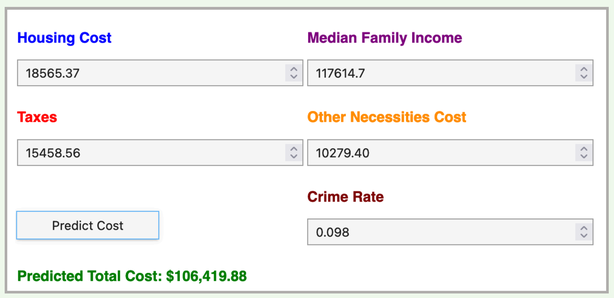
- **Linear regression** was chosen and it used data on factors like housing cost, taxes, and crime rate to predict the total cost of living in different areas.
- The resulting model achieved high R^2 values across various tests, indicating the chosen features **effectively predict total cost**.
- This analysis resulted in a model with a high R^2 value of around 0.97, indicating that approximately 97% of the variance in total cost of living can be explained by the chosen factors.
- This suggests a **strong correlation** between factors like housing cost and total cost of living.

- **K-Means clustering** was chosen to group cities with similar characteristics due to the lack of labeled data.
- The model's success was evaluated using **5-fold cross-validation** to ensure robustness.
- Challenges included selecting the optimal number of clusters (3) and dealing with slightly overlapping clusters.
- Data cleaning involved handling NaN values, adding the "Crime Rate" feature, and normalizing the remaining features.
- **Principal Component Analysis (PCA)** was applied to reduce the dimensionality of the data to 3, allowing us to visualize the clusters in a scatter plot.
- The model achieved a 0.66 average accuracy, considered good given the data's complexity (not perfectly captured by distinct clusters).
- This accuracy suggests the model **successfully grouped cities with similar attributes**, but there's room for improvement with potentially different models.



Interactive Component

- The interactive tool predicts the total cost of living using the mentioned linear regression model and dynamic inputs like housing costs, taxes, and crime rates. Integrated with Python's “ipywidgets”.
- The tool allows real-time interaction in Jupyter Notebooks. Users can modify inputs via styled FloatText widgets, and immediately see updated cost predictions, making the analysis both accessible and engaging.
- The layout optimally organizes inputs into two columns for clarity and ease of use, illustrating the model's utility and the impact of various living cost factors effectively.



Key Takeaways & Conclusions

- We found only partially significant evidence to support the claim that cities with higher median family incomes will have lower crime rates. We did, however, find significant evidence that cities with a higher number of property crimes also have higher housing costs.
- We trained a linear regression model to predict the total cost of living of a city from several other attributes of the city, and created an interactive way to update these attributes and predict how the cost of living will change.
- Using K-Means clustering, we are able to group cities into clusters and achieve a fairly good accuracy.

Challenges and Limitations

- The crime data from the FBI only contains the crimes reported from 2019. This excludes any unreported crimes and also limits our analysis by only having one year of data.
- Our original report was intended to include travel and hotel information, but we could not find relevant data for enough cities to perform an interesting analysis.