ALY-6040-MOD6-FINAL PRESENTATION

Submitted To: Hema Seshadri

Submitted By: Group 9

Srishti Singh

Table Of Contents

- Introduction
- Business Questions
- EDA
- Decision Tree Model Performance
- XgBoost Model Performance
- Conclusion
- References

INTRODUCTIO N

- Child poverty remains an enduring global challenge with far-reaching consequences for the well-being of children, including their health and education. However, accurately measuring child poverty rates can be difficult, especially in low-income countries or regions with limited resources. In such cases, alternative indicators such as income levels can provide valuable insights into estimating child poverty rates and understanding the scale of the issue.
- The US Census dataset is an invaluable source of comprehensive information on the population and demographics of the country. It is widely utilized by government agencies, researchers, businesses, and community organizations to shape policy decisions, planning processes, and community development initiatives. The concept of census tracts ensures that data collected is representative and reliable by designing geographic regions with similar population sizes.

Business Question

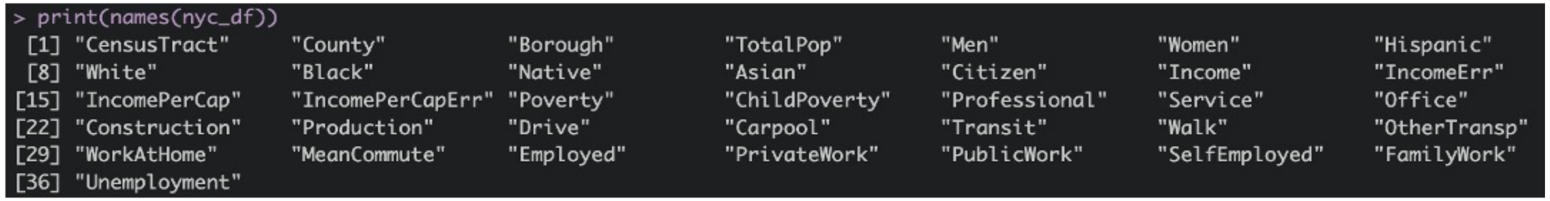
What are the variations in child poverty rates across different regions in New York? Can we identify areas with notably high or low child poverty rates? What would be predicted Child poverty rate based on different predictors.

Which factors are associated with higher levels of child poverty? For instance, is there a correlation between lower Income or Gender difference or Transportation medium difference and increased child poverty rates in specific areas?

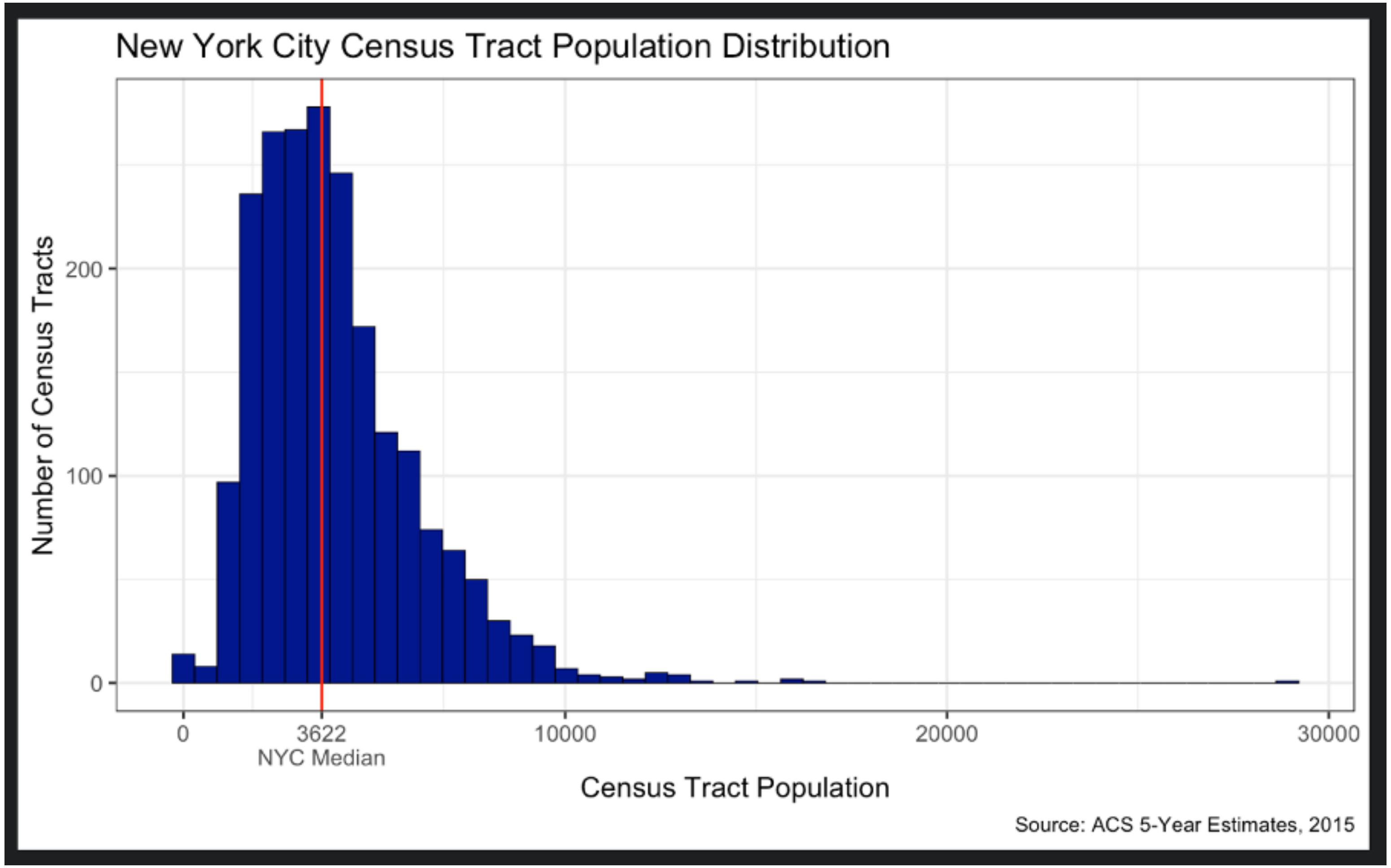
How has child poverty evolved over time in the United States? Are there specific time periods or policy changes that have significantly influenced child poverty rates?

Exploratory Data Analysis

Data Distribution



```
> print(names(nyc_df))
 [1] "CensusTract"     "County"          "Borough"         "TotalPop"        "Men"             "Women"           "Hispanic"
 [8] "White"           "Black"           "Native"          "Asian"           "Citizen"         "Income"          "IncomeErr"
[15] "IncomePerCap"    "IncomePerCapErr" "Poverty"         "ChildPoverty"    "Professional"    "Service"         "Office"
[22] "Construction"    "Production"      "Drive"           "Carpool"         "Transit"         "Walk"            "OtherTransp"
[29] "WorkAtHome"      "MeanCommute"     "Employed"        "PrivateWork"     "PublicWork"      "SelfEmployed"    "FamilyWork"
[36] "Unemployment"
```

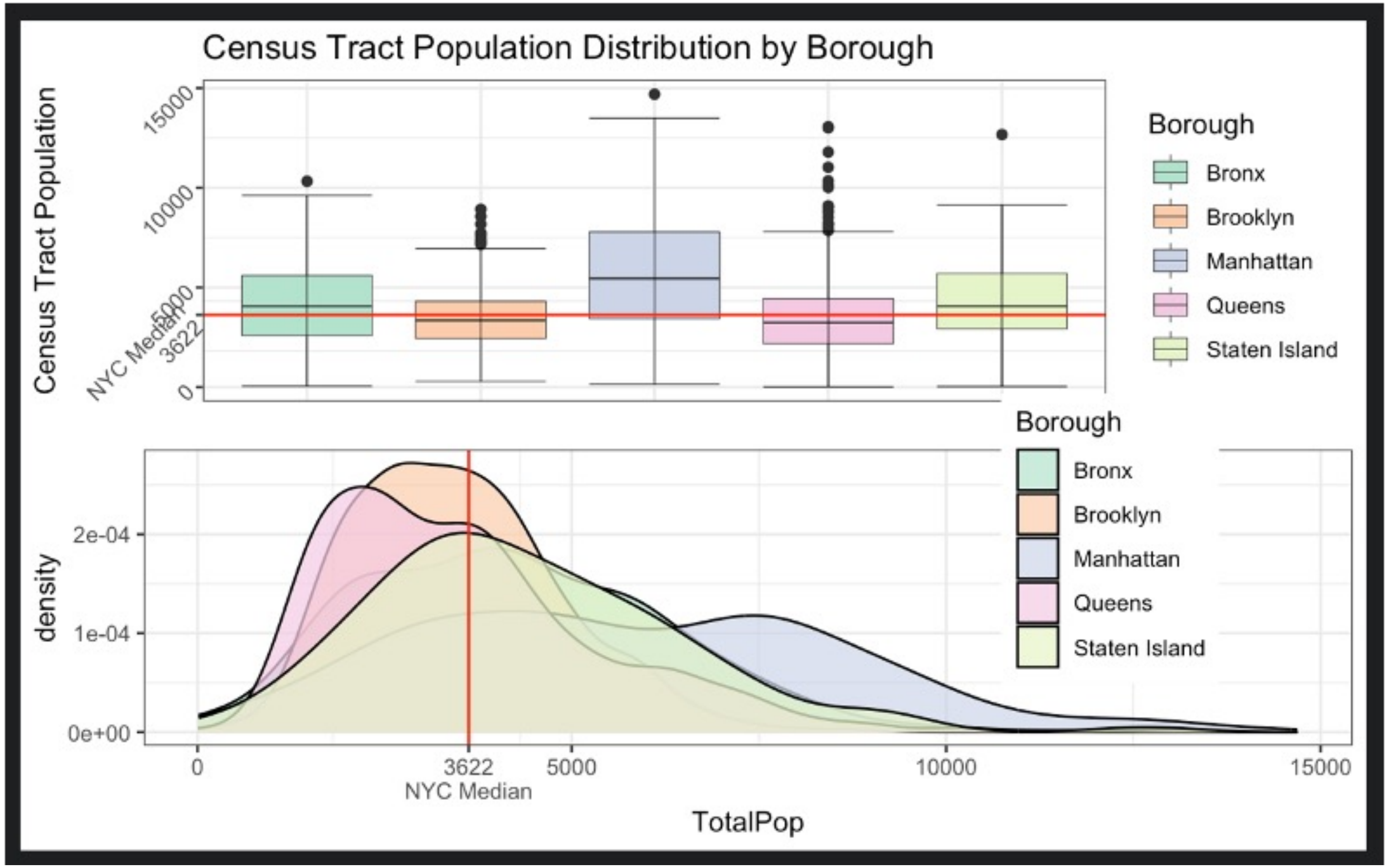


New York City Census Tract Population Distribution

Source: ACS 5-Year Estimates, 2015

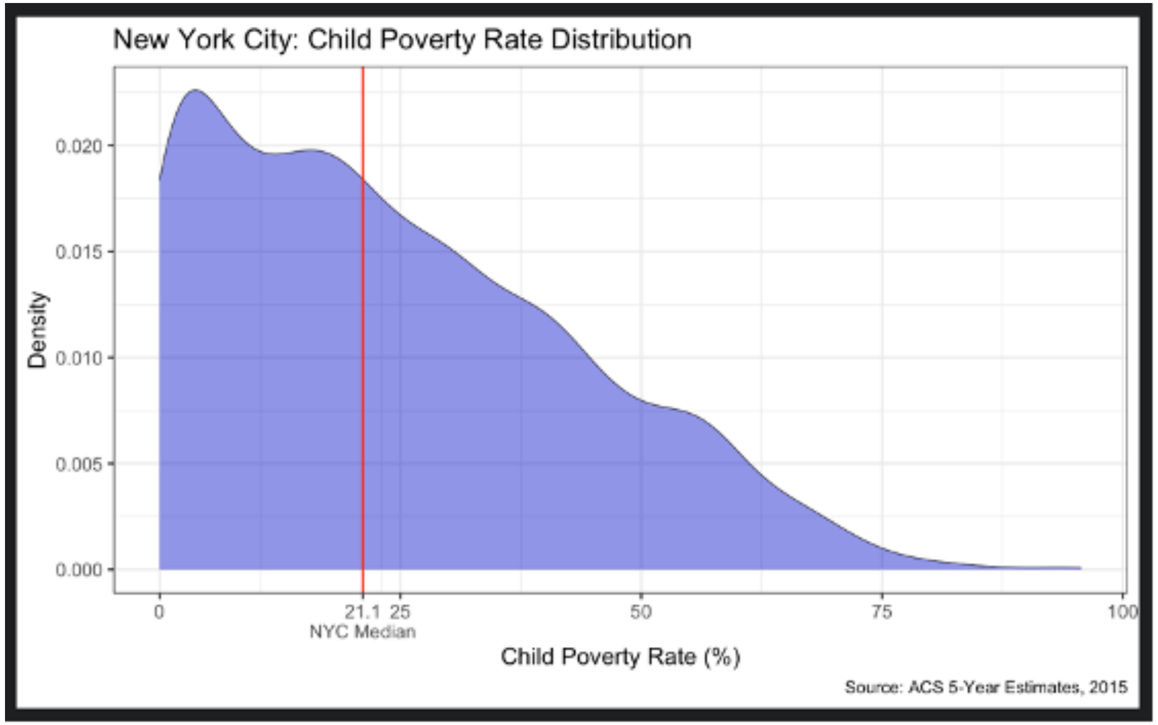
Population by NYC Areas

- The visualization reveals that Brooklyn is relatively close to the median population distribution of NYC, while other boroughs like Manhattan may exhibit different patterns.
- By identifying areas with unique population characteristics and needs, decision-makers can allocate resources and support more effectively, ensuring that the diverse needs of different communities are adequately addressed

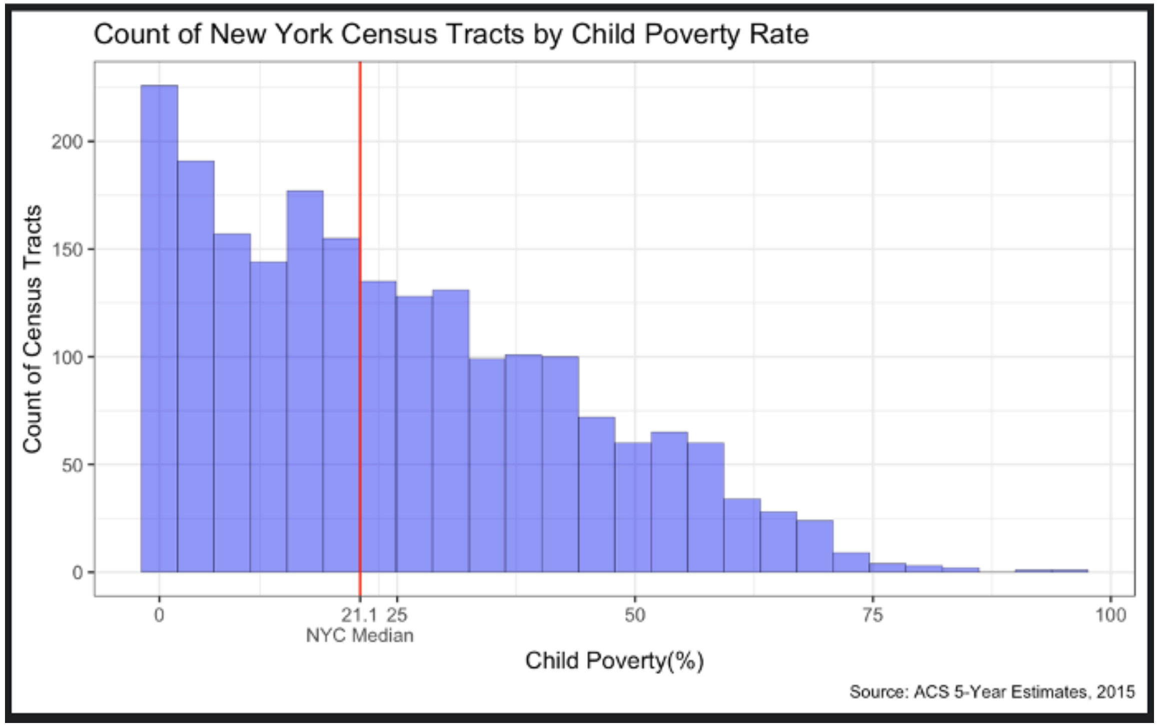


Child Poverty Distribution

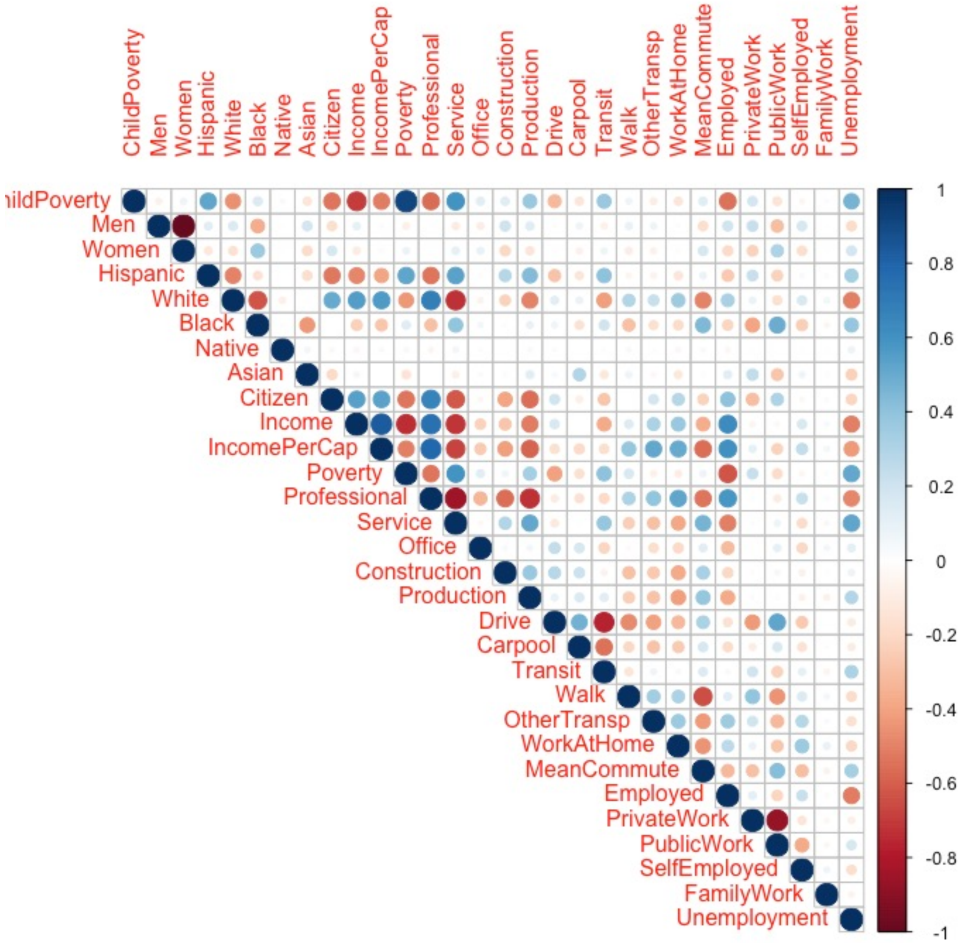
- The distribution of child poverty rates closely resembles a negative linear relationship. The frequency of child poverty rates decreases consistently in accordance with their size.



Source: ACS 5-Year Estimates, 2015

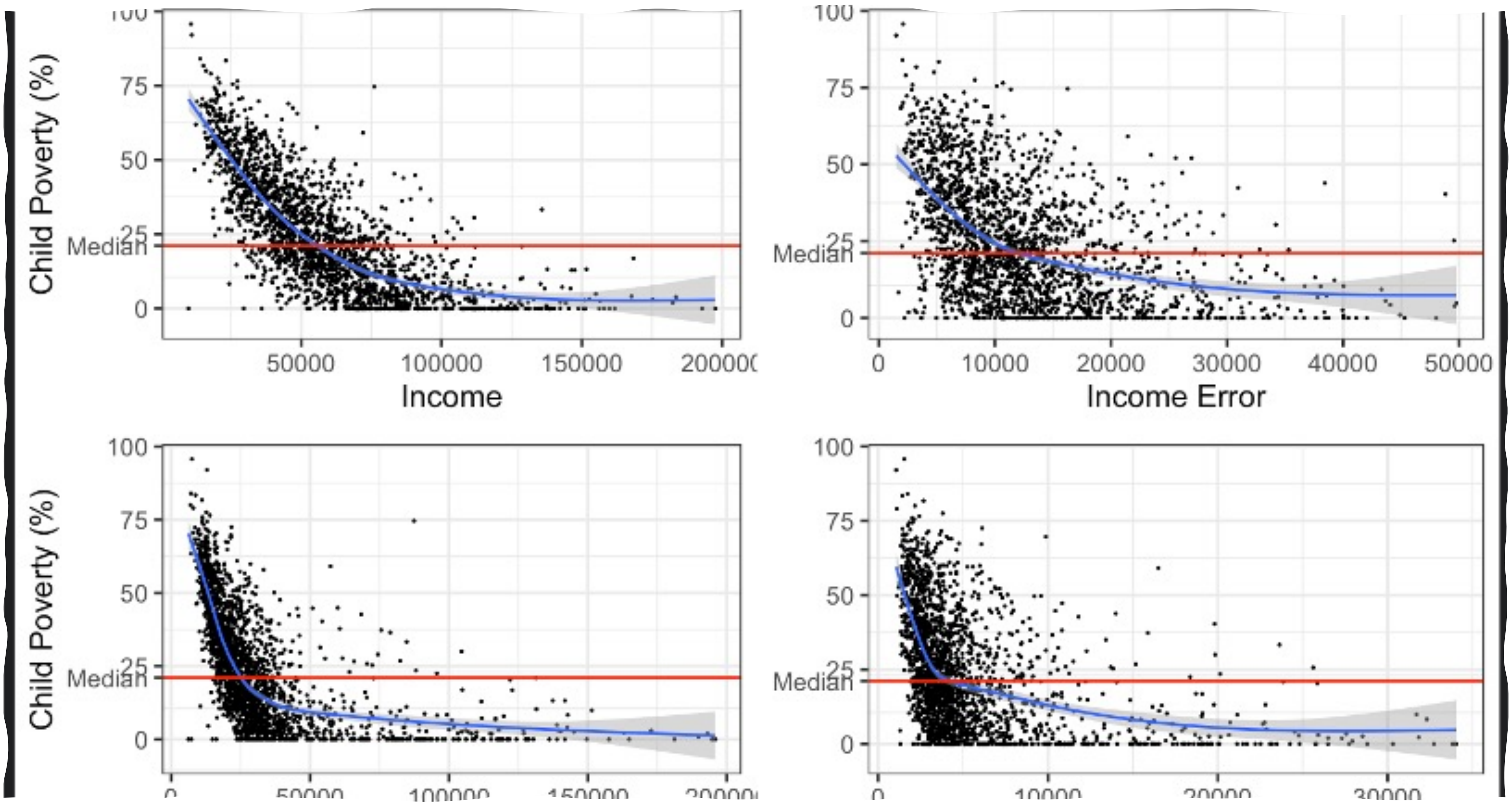


Source: ACS 5-Year Estimates, 2015

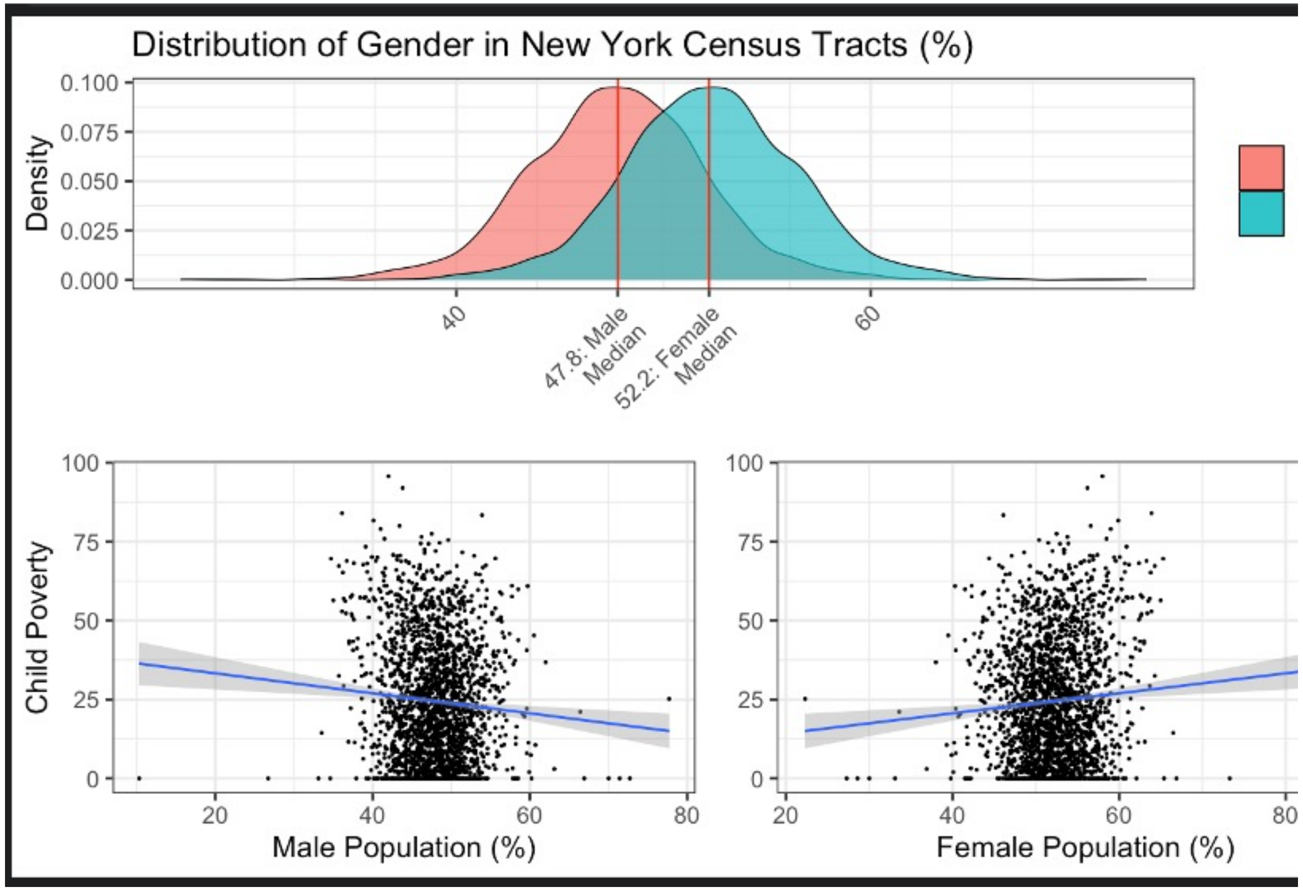


Correlation Analysis Between Variables

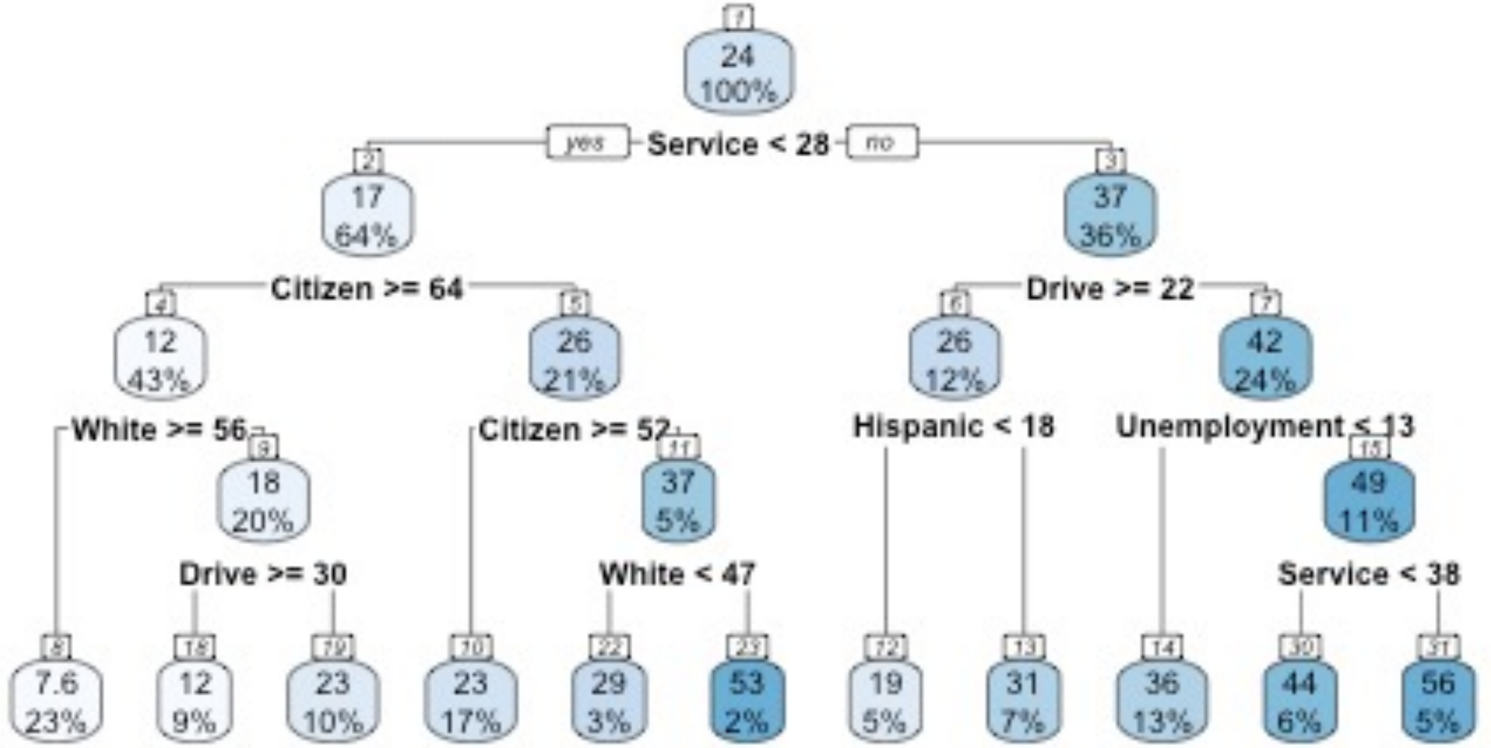
- The correlation plot reveals that ChildPoverty exhibits a strong negative correlation with Income, indicating that higher levels of income are associated with lower child poverty rates.
- The correlation plot reveals that ChildPoverty exhibits a strong postive correlation with Poverty, which aligns with our intuitive understanding of the relationship between poverty and child poverty rates


Child Poverty (%)
100
75
50
25
Median
0
50000
100000
150000
200000
Income
100
75
50
25
Median
0
0
10000
20000
30000
40000
50000
Income Error
Child Poverty (%)
100
75
50
25
Median
0
0
50000
100000
150000
200000
100
75
50
25
Median
0
0
10000
20000
30000

- New York as a whole has a slight gender imbalance with the average ratio of men to women being 47.8:52.2 across census tracts. There is a negative correlation between child poverty rates and the percentage of men, and the reverse for women although it is extremely weak in both cases.

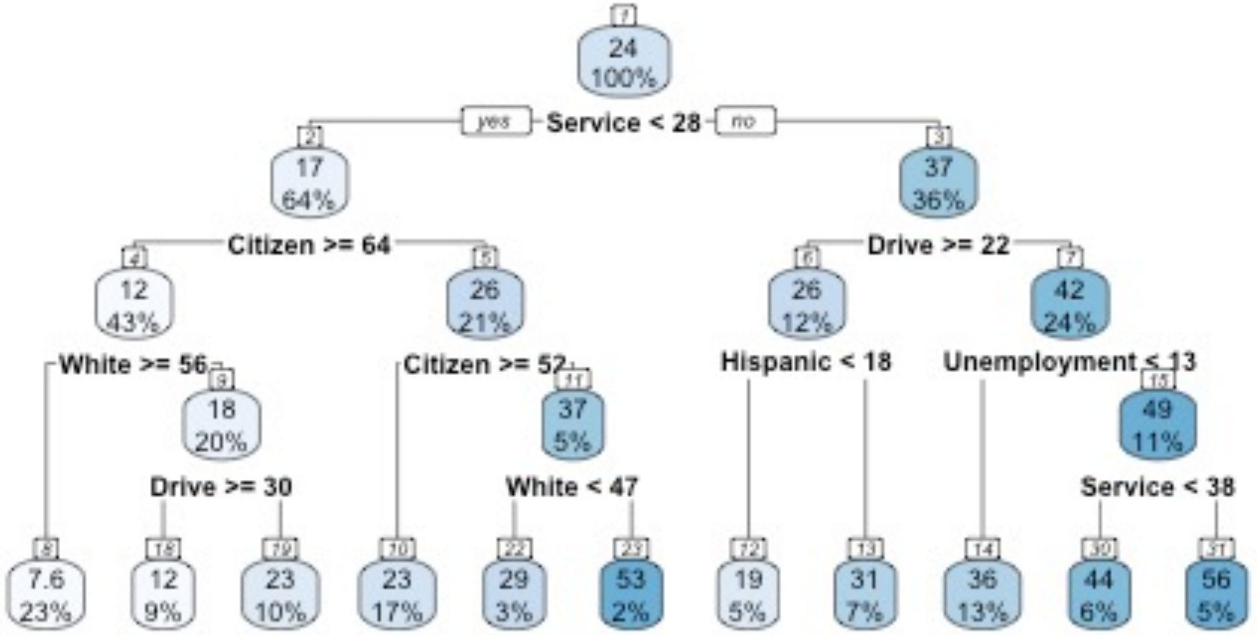


Decision Tree Model on Training dataset 1(with NAs)



Decision Tree Model on Training dataset 2(with using MICE)

Using confusion matrix, we determined the Accuracy of Predicting ChildPoverty through Deci -sion Tree Model is equal to 71.2%



XGBoost Model

During the model construction process, I utilized XGBoost (eXtreme Gradient Boosting), which is known for its complexity, adaptability, and effectiveness in model-building and prediction. XGBoost has been successful in various contests and offers several advantages over other packages, such as parallel data processing and the flexibility to handle missing values.

Unlike most Random Forest packages, XGBoost continues to split nodes until reaching the specified maximum level and then backtracks and prunes if a larger positive loss follows a negative loss. This flexibility and customizability make XGBoost a valuable tool for machine learning, providing more transparency.

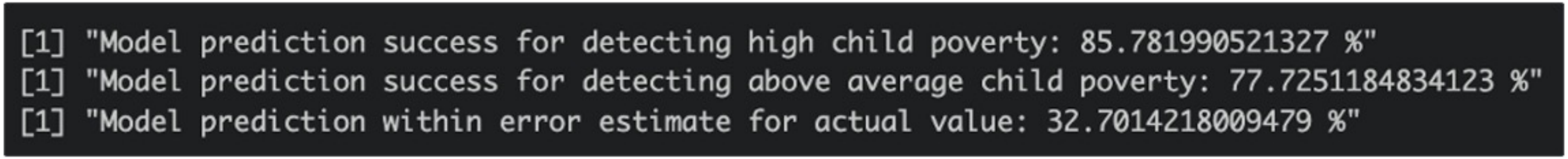
Hyperparameter tuning using K Fold Cross Validation

In order to flag problems like overfitting or selection bias we are running a K Fold cross-validation process first.

In cross-validation instead of evaluating the training data against the test data, the training data is itself repeatedly split up into training and test data, and tested against itself until the whole of it has at some point been tested on.

For example in 5-fold cross validation, 80% of the training data would be used to train a model and tested on the remaining 20%.

This process would be repeated four more times until all the training data has been tested on, the model being refined at each stage. This process can be extremely time-consuming and is generally reserved for smaller datasets as a result.



```
[1] "Model prediction success for detecting high child poverty: 85.781990521327 %"
[1] "Model prediction success for detecting above average child poverty: 77.725118483412 3 %"
[1] "Model prediction within error estimate for actual value: 32.7014218009479 %"
```

Model Performance

- The model's accuracy achieved was 84.36% at predicting whether a census tract will exhibit high child poverty and 80.09% at predicting whether it has above average child poverty. In terms of explicit accuracy, the model estimate fell withing the error boundary of the actual child poverty rate only 32.94% of the time, indicating that without information on household income it is extremely difficult to predict specific child poverty rates.

CONCLUSION

- There is a strong inverse relationship between household income and child poverty, where child poverty rates sharply decrease when household income reaches around $45,000 and per capita income reaches around $30,000. The decline in child poverty becomes more gradual thereafter until it approaches 0%.
- It's important to note that the correlation between child poverty rates and overall poverty rates is very strong, but the relative increase in each variable varies due to differences in average household composition.
- Demographics: Child poverty tends to increase with a higher proportion of women in the population and decrease with a higher proportion of males, although the correlation with gender is weak.
- Via multiple imputation and boosted regression it was possible to model the estimated child poverty rates for census tracts in New York City without any household or personal income data.
- Decision Tree Model fails to deliver the prediction in Child Poverty due to missing values in the Dataset.
- The model successfully predicted the condition of High Child Poverty (child poverty rates in the upper quartile of the city as a whole) with a minimum of 80% accuracy.
- The model successfully predicted the condition of Above Average Child Poverty (child poverty rates above the median of the city as a whole) with a minimum of 75% accuracy.

References

- 1. What is XGBoost? (n.d.). NVIDIA Data Science Glossary. https://www.nvidia.com/en-us/glossary/data-science/xgboost/#:~:text=XGBoost%2C%20which%20stands%20for%20Extreme,%2C%20classification%2C%20and%20ranking%20problems

- 2. Brownlee, J. (2021). A Gentle Introduction to XGBoost for Applied Machine Learning. MachineLearningMastery.com. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

- Dataset: New York City Census Data. (2017, August 4). Kaggle. https://www.kaggle.com/datasets/muonneutrino/new-york-city-census-data

Question ??

Thank you