**Structure**

1. Introduction: introduce the problem, why interesting

The Northern Cardinal (Cardinalis cardinalis) is one of the most observed bird species in the U.S., commonly found in backyards and well-recognized nationwide. This study quantitatively investigates how human populations and urban development affect Northern Cardinal observation frequencies in U.S. counties. Given their adaptation to certain human-developed areas for food sourcing, this research examines the impact of urban development and population changes on these birds' observation levels. Focusing on 2022 data, the study analyzes County-Level Rural-Urban Continuum Codes, Urban Influence Codes, Economic Typology, Natural Rate of Change, and Immigration Rates. It hypothesizes that more developed counties will record higher Northern Cardinal observations, suggesting a growing environmental responsibility for humans to maintain urban habitats conducive to certain bird species adapting to these environments.

(1)<https://birdwatchinghq.com/common-birds-in-the-united-states/>

2. Data: present the dataset, describe key features

This project merged two datasets. The first, from Cornell University’s Ornithology Lab’s eBird project, provided Northern Cardinal observations in U.S. counties. Access was granted upon request, and data from January, April, July, and October 2022 were chosen to manage large file sizes and minimize seasonal biases, totalling 982,667 observations. Data cleaning involved aggregating data at the county level, retaining total observation counts, average observation duration, and effort distance. These variables, showing significant skewness, underwent log-transformation, including the key observation count variable for regression analysis. Appendix includes histograms of these variables. Nine states with minimal Northern Cardinal presence and counties with fewer than 10 observations in 2022 were excluded, resulting in 2,288 counties in the cleaned dataset.

The second dataset, sourced from the USDA, contained county-specific information on population and urban development across the United States, including 3,143 counties. It required minimal cleaning, with no need for row eliminations due to missing values. The cleaning process primarily involved narrowing down to key columns: Rural Urban Continuum Code, Urban Influence, Economic Typology Code, Natural Rate of Change, and Net Immigration Rate.

The Rural-Urban Continuum Codes classify counties on a 1-9 scale, with Code 1 representing highly urbanized areas (over 1 million population) and Code 9 indicating extremely rural areas (under 2,500 population, not metro-adjacent). Urban Influence Codes extend this categorization with a 1-12 range, where 1 is the most urban and 12 the most rural. Economic Typology Codes, totalling six, identify the primary economy of a county: 0 for non-specialized, 1 for Farm, 2 for Mining, 3 for Manufacturing, 4 for Federal/State Government, and 5 for Recreation Dependent. The net immigration rate measures population movement in and out of counties, while the natural rate of change compares births and deaths. These rates, showing normal distribution, along with categorical variables converted to dummy variables, were used in regression models analysing Northern Cardinal observation counts. After merging the datasets, the study included 2,190 observations. The appendix provides detailed code descriptions and distribution charts.

3. Model:

There were four sets of regression models run to test the association between urban development and populations on the log of Northern Cardinal observations. As the first three variables were categorical variables, they were run separately in groups of binary variables to avoid multicollinearity and overfitting by too many variables.

***OLS Regression Models of Urban Influence and Rural-Urban Continuum Codes on Log of Observation Count***

In this research, two distinct sets of Ordinary Least Squares (OLS) models were executed to examine the association of Rural-Urban Continuum Codes and Urban Influence Codes on the log counts of Northern Cardinal observations. The selection of the OLS methodology was based on the hypothesized linear association between the urban development status and the log of Northern Cardinal observation frequencies. Given that both Rural-Urban Continuum and Urban Influence Codes provide analogous insights regarding a county's characteristics, they were analysed independently to corroborate and enhance the robustness of the findings. In each set of models, the initial code (Code 1 for both sets) was designated as the reference category and consequently excluded from the regression analysis. Subsequently, a series of OLS regressions were conducted for each code set, progressively incorporating additional code variables as explanatory factors (Rural-Urban Continuum Codes 2 through 9 and Urban Influence Codes 2 through 12, respectively). To account for potential heteroscedasticity, all models were estimated using heteroscedasticity-consistent standard errors of type HC1.

The analysis of both model sets yielded similar outcomes. In each case, the comprehensive model, encompassing all codes except the benchmark, demonstrated superior model efficacy, as evidenced by the highest R-Squared value. Consequently, these models were selected for the analysis. The R-Squared for the Urban Influence Code OLS regression was determined to be 32.5%, while 34.1% for the Rural-Urban Continuum Code OLS regression. This indicates that approximately 33% of the variability in the log observation counts is explicable by each model.

A consistent pattern emerged across both the Rural-Urban Continuum Codes and Urban Influence Codes. All codes exhibited statistically significant correlation coefficients at the 1% significance level. This uniform significance is likely due to the substantial data volume. Notably, in each model, the correlation coefficients increasingly deviated from zero in a negative trajectory. This trend suggests an inverse relationship between the degree of urbanization and the logarithm of observation counts, a phenomenon observed in both the Rural-Urban Continuum Codes (Codes 2-9) and Urban Influence Codes (Codes 2-12). Therefore, the data provides evidence in support of a positive association between both Rural-Urban Continuum Codes and Urban Influence Codes, and the log counts of Northern Cardinal observations in US counties.

***OLS Regression Models of County Economic Typology on Log Counts of Northern Cardinal Observations***

OLS models were executed to examine the association County Typology Codes on the log counts of Northern Cardinal observations. The OLS model was selected for the hypothesized linear association between the urban development status and the log of Northern Cardinal observation frequencies. Economic typology represents the urban development status of a county from a different. In the model, the code 0 (non-specialized counties) was designated as the reference category and consequently excluded from the regression analysis. Subsequently, a series of OLS regressions were conducted for each code set, progressively incorporating additional code variables as explanatory factors (Code 1: Economically Farming Dependent, etc.). To account for potential heteroscedasticity, all models were estimated using heteroscedasticity-consistent standard errors of type HC1.

The analysis revealed that the regression model with all economic types was the best fit, with an R-Squared explaining 11.8% of the variance in log observation counts. Farming, mining, and manufacturing dependent resulted in statistically significant negative coefficients at the 1% threshold, while government dependent and recreation dependent were not statistically significant. This would suggest that farming, mining, or manufacturing dependent counties are associated with less Northern Cardinal observation counts.

***OLS Regression Models for Natural Change Rate, Net Immigration Rate, Observation Duration, and Effort Distance on Log Observation Counts***

The final OLS models examined the associations between population changes, observation methods, and Northern Cardinal observation counts. These models incorporated natural change and net immigration rates per county to explore how population dynamics relate to observation counts. Additionally, log-transformed variables for the duration and distance of observations were included to examine potential associations with observation counts. The chosen model accounted for 20.2% of the variance in observation counts. Both natural change and net migration rates exhibited small, yet significant, positive associations with observation counts at the 1% significance level. For observation methods, the model included variables such as the log duration and log distance of observation. The log duration of observation displayed a positive association at the 1% significance threshold, indicating an association between longer observation times and higher counts. Conversely, the log distance travelled for observation demonstrated a negative association at the same significance threshold, indicating an association between greater travel distances for observations and lower counts.

4. Generalization and external validity (robustness check)

Predictive models were applied to the OLS models' variables to validate the associations with Log Northern Cardinal observation counts. These models further validated the negative association between urban influence, rural-urban continuum codes, and observation counts, suggesting lower counts in more rural counties. Economic typology modelling indicated non-specialized counties and recreation dependent counties have the highest observation counts, with government, mining, and manufacturing dependent counties showing lower counts. Predictive models for natural change rate, net immigration rate, observation duration, and effort distance yielded results analogous to the OLS models: negligible associations with natural change and net immigration rates, a positive association with longer observation durations, and a negative association with greater observation distances. These findings suggest that observation counts of backyard bird species are inversely related to urbanization levels, with recreational areas possibly observing more species. Observation duration positively correlates with counts, whereas wider observation effort distances may lead to fewer observations. This underscores the potential influence of observation methods and urban development on bird observation frequencies.

5. Causal interpretation / main summary

(a) Summarize your findings. Discuss room for a causal interpretation.

6. Conclusion

(a) Conclude and make business / policy comments / recommendations