**Structure**

1. Introduction: introduce the problem, why interesting

One of the most reported and observed bird species in the United States is the Northern Cardinal (*Cardinalis cardinalis*). This species is a particularly popular ‘backyard’ species and is widely known by people across the United States (1). This specific bird is especially interesting to study on a quantitative basis, as with high levels of observation frequencies, and large volumes of interactions with people and urban development, data analysis can unveil how populations of people and their behaviors impact the frequency of Northern Cardinals throughout US counties. Northern Cardinals have seemingly become dependent on certain human development features, such as frequenting specific neighborhoods and houses to feed from, as they are generally safe and easy to find food for seed-eating birds. This project will dive into the research question: how does urban development and population influxes in United States counties impact observation levels of Northern Cardinals. This research project will focus on the year 2022 and look at the County-Level Rural-Urban Continuum Codes and Urban Influence Codes, County-Level Economic Typology, Natural Rate of Change, and County Immigration Rates. As Northern Cardinals are commonly observed as ‘backyard’ bird species, this project hypothesizes that a more developed county sees a higher level of Northern Cardinal observations. This suggests a greater environmental responsibility for humans to take care of local urban environments, as some bird species are adapting to these developing environments.

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

2. Data: present the dataset, describe key features

This project contains two datasets from different sources and were merged. The first dataset comes from eBird, Cornell University Ornithology Lab’s digital bird observation collection project. Upon request, Cornell granted access to the data sets. From here, data on Northern Cardinal observations in all counties in the United States were selected. The dates selected were all observations in January, April, July, and October in 2022. This is because the data files are so large, so for efficiency, the datasets were split up. Additionally, this helps avoid seasonal observation biases, in case this species as strong seasonality shifts. A total of 982,667 observations were collected for this project.

The cleaning process involved narrowing the columns down. The data was aggregated by county level, and kept the sum of observation counts, average duration in minutes spent on the observation, and the effort distance in km that the observation took place over. These additional columns were kept seeing how features of the observation are associated with observation counts. The three primary variables maintained in the dataset all showed high levels of skewness with long tails and underwent a log-transformation as a result. The histograms can be found in the appendix. Finally, 9 states were dropped from the dataset, as they are known to have very few Northern Cardinals, and counties with less than 10 observations for the year of 2022 were dropped. There were 2,288 counties included in the Northern Cardinal observations dataset post-cleaning.

The second dataset was a county information and indicator dataset, that includes information about populations and urban development of individual counties across the United States. This data was collected from the USDA and contained 3,143 counties in the dataset. The dataset was very clean and did not need row drops for missing values. Many of the columns were dropped in the cleaning process to focus on the key columns, which included: Rural Urban Continuum Code, Urban Influence, Economic Typology Code, Natural Rate of Change, and Net Immigration Rate.

**The Rural-Urban Continuum Codes** range from 1-9. They are categorized where Code 1 is considered the most urbanized and the highest population with more than 1 million population, while Code 9 is considered the most rural with completely rural or less than 2,500 urban population, and not adjacent to a metro area. **The Urban Influence Codes** are like the Rural-Urban Continuum Codes, but have a greater range of specifications from 1-12, with 1 being the most urban developed, and 12 being the most rural. Finally, the **Economic Typology Codes** shows the primary economy of the county. There are a total of 6 codes: 0: Non-Specialized, 1: Farm Dependent, 2: Mining Dependent, 3: Manufacturing Dependent, 4: Federal/State Government Dependent, and 5: Recreation Dependent.

The net immigration rate shows the ratio of individuals moving in and out of the city, while the natural rate of change is the ratio of deaths and births within the county. These are used to analyze shifts in populations. These variables showed normal distribution and did not need transformations. All the categorical variables were transformed to dummy variables for the purpose of running regression models on Northern Cardinal observation counts. Upon merging the datasets, a total of 2,190 observations were included. Additional information regarding the detailed variable description of each code and distribution bar charts and histograms can be found in the appendix.

3. Model:

There were four sets of regression models run to test the association between urban development and populations on 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.

(a) Present the model you estimate, argue for your model choice

(b) Discuss your variables, process of feature engineering. (In words, put all needed graphs, estimation into the appendix if needed.)

(c) Show core results. Interpret what you got precisely.

4. Generalization and external validity (robustness check)

(a) Show some robustness / alternative models.

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