Predicting Urban Population

Ashton Reed
Springboard's Data Science Track
Capstone 1

Problem

- People moving to cities in large numbers can strain the existing infrastructure
  - Short-term solutions often imposed
  - Little regard for long-term consequences or scaling
- Organizations like the United Nations need a way to track global development and growth

Solution: Predicting Urban Population

- Can use urban population information in order to determine where to concentrate resources
- Can keep track of urban development in each country
- By being able to predict the growth of urban population numbers, city planners can implement urban infrastructure that is more sustainable in the long-term
- Can keep an eye on areas that are expanding rapidly and might experience problems often associated with rapid urban growth
- Organizations such as the United Nations can evaluate growth and development of a country

Who Benefits from Predicting Urban Population?

- Organizations such as NGOs
- IGOs such as the United Nations, World Bank, or European Union
- National committees concerned with infrastructure
- National committees in charge of budgets and deciding how to allocate
- City planners
- Citizens in these urban areas to which resources are allocated as a result

Dataset

- United Nations dataset from https://www.drivendata.org/competitions/1/united-nations-millennium-development-goals/
- Gapminder dataset from Gapminder data source: https://docs.google.com/spreadsheets/d/1qHalit8sXC0R8oVXibc2wa2gY7bkwGzOybEMTWp-08o/edit#gid=1597424158
- CSV file with region and income group copied from https://data.worldbank.org/

Approach

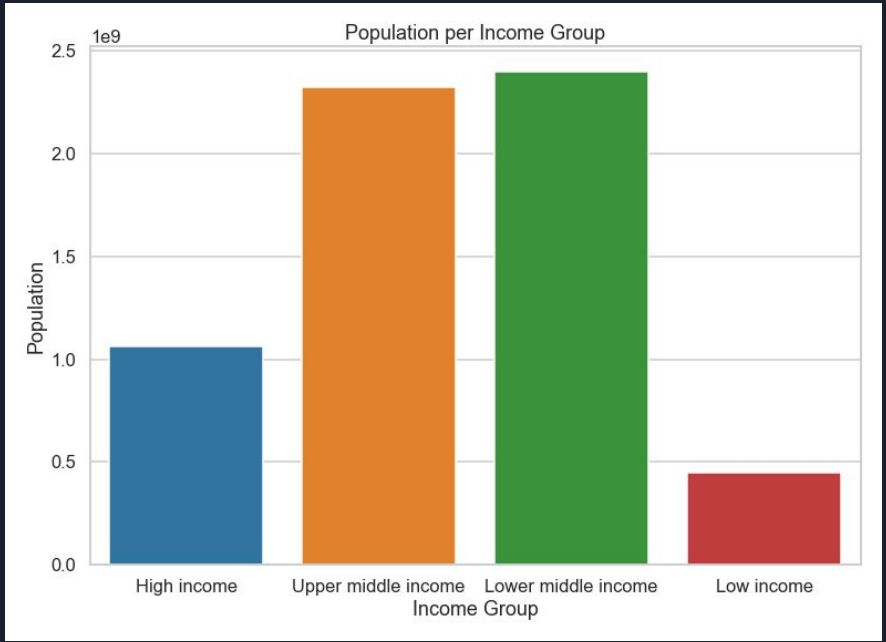
- Created a tidy dataframe from UN dataset
- Backfilled missing values
- Dropped all but top 10 most populated features (excluding those related to land area)
- Created a new urban population column (“Urban population 5 years in the future”) by copying “Urban population” column and shifting it by 5
- Dropped observations for 2003-2007 since these did not have a known value for the new column
- Merged the Gapminder dataset into the UN dataset in order to add income group and region
- Used World Bank site to create a csv file with income group and region for countries missing from Gapminder dataset

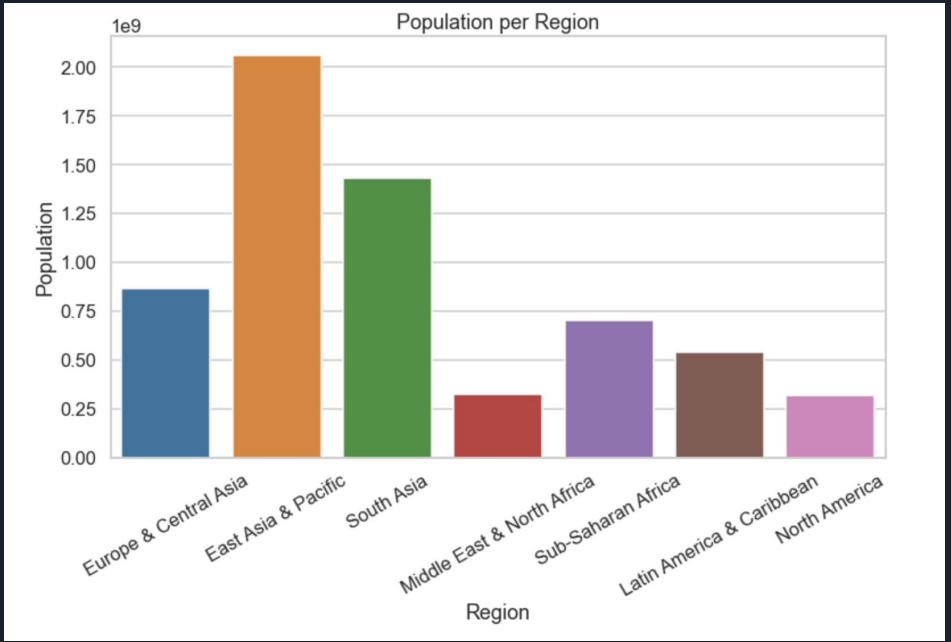
Insights

- There is a clear divide when countries are grouped by income level or region
- China and India are the two most extreme outliers regarding total population and urban population
- Over 80% of the countries in the dataset had a population less than the mean value for both 1972 and 2002--in cases like this, the median is more representative of the dataset
- Higher income countries have a higher proportion of urban population
- Lower income countries are becoming more urban at a faster rate

Insights (cont.)

- Neither countries nor population are evenly distributed across income levels or regions





Modeling

score	Lasso	Basic Linear Regression (Model 0)	LinReg with Feature Selection & Scaling
R_squared	0.99762	0.99937	0.99935
RMSE	-1750181.85597	896616.74348	915656.90161

- Created 3 models
  - Lasso
  - Basic Linear Regression
  - Linear Regression with Feature Selection and scaling
- Used cross validation and test-train-split for better predictions
- Basic Linear Regression performed best

- Lasso regression for feature selection
  - Income groups
  - Adjusted savings: mineral depletion (current US$)
  - Population (Total)
  - Rural population (% of total population)
  - Urban population (% of tital)
  - Rural population
  - Urban population
  - Urban population growth (annual %)
  - Country
  - Year

Limitations and Future Iterations

- Using “World bank, 4 income groups 2017”--in a future iteration could instead pull the information regarding how each country was categorized for each year in the dataset.
- Columns with a lot of missing data were dropped, limiting features to 10 from the original dataset
- Missing data in remaining columns was backfilled instead of interpolated
- Dataset does not meet the IID requirement for majority of numerical features--however, only predicting a single feature.
- In a future iteration could analyze more series/indicators from the original UN dataset to see if I dropped any important ones
- Since we trained and fit the models across all countries, predictions for some countries are more accurate than for others. One way to fix this would be to customize the model for each country.
-

Limitations and Future Iterations

- If we want to work on optimizing the models, we could look into the percentage increase in urban populations from previous years to determine just how feasible our predictions are.
- When comparing the Lasso prediction and basic Linear Regression predictions for each country, we see that from the 10 rows we can see above, the LinReg model's prediction is typically closer to the population in 2007 than the Lasso model's.
- The first Linear Regression model was trained across all features while the second was trained on a slightly smaller set of features. The features were also scaled in the second model. It is also possible to continue looking for a better Linear Regression model by creating one with scaled features and the full dataset and creating another without scaling and with fewer features. These models would be created similarly to the second linear regression model above.
- It is also possible to use time series analysis to predict population. However, that is currently out of the scope of this project.

Questions?