Predicting Urban Population As Part of Global Development: Final Report

Ashton Reed

Problem Statement:

People moving to cities in large numbers can strain the existing infrastructure, and all too often, short-term solutions are imposed with little regard to long-term consequences or scaling. Organizations, such as NGOs, can use urban population information in order to determine where to concentrate resources and to 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 sustainable in the long-term and keep an eye on areas that are expanding rapidly and might experience problems often associated with rapid urban growth.

The UN uses indicators to track the progress of development in each country. The goal of this machine learning project is to predict the urban population five years into the future. Given the UN's dataset, I have chosen to build a model to predict the "Urban Population" indicator 5 years into the future. This is important in analyzing the growth and development of countries and predicting their infrastructure needs in urban settings.

As this is my first big project, it is only the first iteration, so there are several things to note. For simplicity, I am using the "World bank, 4 income groups 2017" categorical feature provided in an additional dataset from Gapminder. In reality, countries can--and did--move from one income group to another during the timespan of our dataset, so the income group in which each country found itself in 2017 might or might not be representative of the income group of that country from 1972 to 2002. In a future iteration of this project, I could instead pull the information regarding how each country was categorized for each year in the dataset. However, at the moment, that is outside the scope of this iteration.

In some cases, a more complex and "better" method might be replaced by a simpler one that serves the same function but might result in predictions that are slightly less accurate. For example, in the data wrangling file, I have chosen to backfill missing values instead of interpolating based on the data I do have for a specific country. That is outside the scope of this first iteration. Because this dataset has a lot of missing values, I opted to sort by the top 10 most-complete series--excluding 2 dealing with land area. And because 8 of the remaining 10 deal with population and only one of them is the target variable, this slimmed down dataset does not meet the IID (independent and identically distributed) requirement as it does not meet the first "i"

criteria. However, I am only making a prediction for a single feature. In a future iteration of this project, I can analyze more series/indicators from the original UN dataset. At this time, I am focusing on exploring the interdependence between the target variable and other population-related series, along with the relationship between urban population and the region and income group by which countries are categorized in an additional dataset.

Using supervised learning to uncover the relationships between the urban population and the other indicators in the narrowed dataset, this project will predict the urban population five years into the future, given data from 1972 to 2002. I will then use supervised learning to predict the change in urban population 5 years in the future. The training dataset used (1972-2007 indicators) is publicly available for download at https://www.drivendata.org/competitions/1/united-nations-millennium-development-goals/

The deliverables for this project will be the machine learning model development along with a final presentation and report. The first will outline the process of building the model to predict the change in indicator five years into the future. The final two will detail population trends and predict the urban population five years into the future.

Data Wrangling:

I began the process of data cleaning by inspecting the dataset. The first column appeared to be an index that was missing several numbers and did not appear to be very useful for analysis, so I dropped it. I then determined the number of unique years, countries, and series (both codes and names) to determine how many data points should exist in the dataset. With 36 years, 214 countries, and 1305 series, I determined there would be 10,053,720 data points if none were missing. Next, I cleaned the “Years” columns and added the countries and series to the new “Years” columns.

In the initial dataframe, the years were columns, and the series were in rows, so I melted the dataframe to change the years to rows. I then labeled the year column and changed the series to columns by unstacking the dataframe and setting the index. This way, each observation would be tied to a country and a year. In making these changes, the dataframe was changed from 195402 rows × 39 columns to 7704 rows × 1307 columns.

Because I noticed a lot of NaNs, I decided to rank each of the series (columns) by the number of null values in each. Using a “for” loop, I dropped all series that were not at least half full. As a result, 867 series were dropped from the original dataset and only 438 remained. I then chose to look at the first country in the list, Afghanistan, as an example of what the remaining series information or graphs may look like. Several series were missing data for all 36 years in Afghanistan’s dataset. This led to only one

graph of the first four for Afghanistan showing any data. Two series had no data points, and the third series only had one.

I reviewed the dataframe again and decided to more closely inspect the "Telephone lines (out of 100 people)" out of curiosity. If there are 36 years in the dataset, then a series that has at least half of the data for each of the 214 countries should have at least 18 data points for each country. For this series regarding telephone lines, we see that 17 countries are missing at least half of the data points for the "Telephone lines (out of 100 people)" series. While having *at least* a half-full dataset for 92% of the countries is not bad for a series, I ultimately decided to only look at the top 12 most-populated series in terms of overall rows (out of 7704 possible) in the dataset to see if any of them had an observable relationship with the urban population.

To prepare for the next portions of the project, I have added a column containing the urban population value 5 years into the future. The years 2003-2007 have been removed as they do not have a value in this 5-year column and cannot be used in the training set. Missing values were backfilled. Additionally, I imported a dataset from Gapminder in order to group the countries by region and income for better visualization and analysis. As pointed out in the proposal, the Gapminder data reveals the region and income group in which each country was categorized in 2017. In reality, several countries moved from one income group to another over the timespan of our dataset, which ends in 2002. Therefore, the income group may not be as accurate of a depiction for some countries in the dataset.

Gapminder data source:
https://docs.google.com/spreadsheets/d/1qHalit8sXC0R8oVXibc2wa2gY7bkwGzOybEMTWp-08o/edit#gid=1597424158

Exploratory Analysis:

Going into this portion of the project, I decided to analyze the total population, urban population, and urban population percentage of countries in relation to income group and region. In grouping countries by income level and region, several underlying patterns began to emerge.

Urban population across every region experienced positive growth. In 2002, the "Population growth (annual percentage)" figure was only larger in North America and Sub-Saharan Africa, meaning that population growth rate was increasing in these two regions and decreasing in the other 5.

The Middle East and North Africa experienced the greatest percentage increase in population density over our 31-year timespan with 102.8% change, followed closely by South Asia (101.5%) and Sub-Saharan Africa (97.3%). North America exhibited the

lowest percentage increase in population density (16.6%), with Europe & Central Asia experiencing a 28.8% increase. Latin America & Caribbean clocked in at 43.8%, and East Asia & Pacific ending up in the middle of the group with a 71.2% increase in population density.

North America experienced the lowest increase in population density, but not the lowest increase in overall population (37.6%)--it was Europe & Central Asia that experienced the lowest increase in total population (53.7%). Sub-Saharan Africa experienced the greatest percent change in population, followed by Middle East & North Africa (108.6%) then South Asia (90.8%).

Between 1972 and 2002, the rural population decreased across all regions, and the urban population increased. Europe & Central Asia experienced the lowest percentage increase in urban population (27.1%), with North America experiencing a slightly larger change at 48.6%. Sub-Saharan Africa saw a 270.8% change in urban population, followed by Middle East & North Africa (180.67%), South Asia (176.4%), East Asia & Pacific (165.2%), and Latin America & Caribbean above North America at 133.4%. Of all 7 regions that experienced a positive percentage change in growth, only North America had a higher “Urban population growth (annual %)” value and “Population growth (annual %)” value in 2002 than in 1972. Sub-Saharan Africa also saw a positive change in “Population growth (annual %)” when comparing its value in 1972 and 2002.

From finding the averages of each income group over our 31-year timespan, we notice several things. Overall, the average total population is highest in "upper middle income" countries, followed by "lower middle income", "high income", and "low income" countries at the very bottom. Income and population growth (as a percentage) are inversely related with high income countries having the least growth and low income countries having the most. Population density is directly related to income, with high income countries being the most densely populated. Urban population percentage numbers have a similar correlation, with the rural population percentage being inversely related to income. Lower income countries have the highest urban population growth (as a percentage) while the already densely-populated higher income countries exhibit less urban growth.

From plotting the empirical cumulative distribution function of the total population, I discovered that there were two most extreme outliers (China and India), which would skew the mean so that both the urban population and total population of more than 80% of the countries in the dataset fell below the overall mean for both 1972 and 2002. This means they will also skew the mean for their respective income group and region. Because these outliers are very large but not suspicious, we will retain them in the dataset, and we will also investigate how each of the dataset features change over time in these countries and compare them to the change in features of the countries with the smallest total populations.

For this storytelling/exploratory data analysis portion of the project, I utilized bar graphs to compare how many countries or people are in a region/income group, and histograms were utilized for visualizing distributions of population by region/group. Line graphs were used for visualizing trends, and scatter plots allowed for plotting two features against each other for a particular year.

From looking at the histogram of urban population in 2002 broken down by region, I saw several outliers. After accounting for this by only plotting urban populations less than 4 million, it appears that only Europe & Central Asia and Latin America & Caribbean regions have a somewhat similar distribution curve, but that of Europe & Central Asia is shifted slightly to the right. Otherwise, each group appears to have its own unique distribution. When we run a t-test on every possible combination of regions, we see that the largest p-value is only 2%, which is less than our goal of five percent. Therefore, we conclude that these regions are distinct populations.

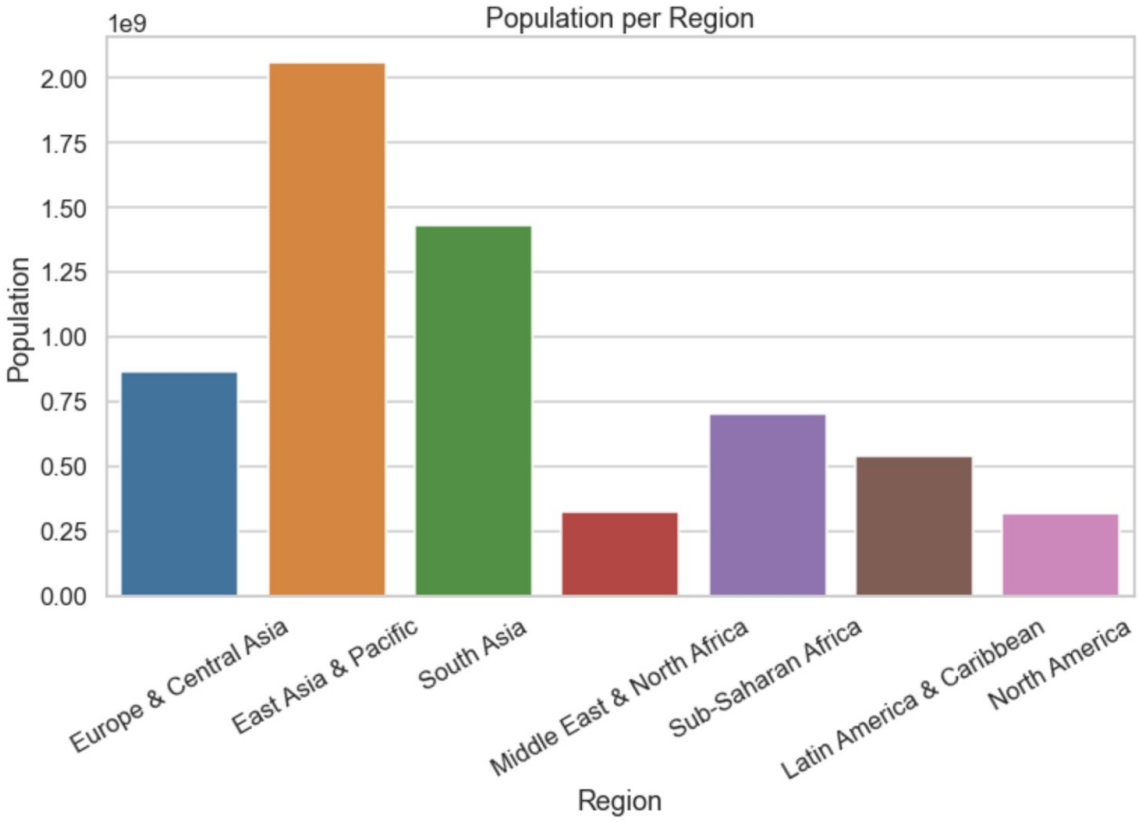
When we look at a graph of the average urban and total populations over the years, Europe & Central Asia appears to be the most stagnant region. South Asia has a greater total population than North America for every year on the graph, but North America maintains a higher urban population for every year in the 31-year timespan. East Asia & Pacific is ranked third in terms of both total and urban populations each year. Sub-saharan Africa, Middle East & North Africa, and Latin America & Caribbean appear to have similar values and a slight upward trend for both total and urban populations. We should also keep in mind that some regions have fewer countries (i.e. North America only contains three countries).

When we look at the data by income group, we again see a distinction between groups. By overlaying histograms of each income group's urban population distribution, we see that many of the observations fall into the first bin. While the low income group is still not normally distributed, it does not seem to have the problem of so many observations in the first bin, something that the other three income groups exhibit. It also appears to be the only income group that does not have an urban population greater than 1,000,000. When performing a t-test of all the possible combinations of income groups, we see that high income and lower middle income groups have a p-value of 0.52, or 52%. There is not necessarily a clear distinction in these two groups. None of the regional groups seemed to intersect, but when we plot both urban and total populations over the years, we see that the lower middle income group and the high income group averages intersect in 1985. This explains the high p-value between these two groups. If we look at the graph of total population broken down by group over the years, we see an intersection between the lower middle income and upper middle income groups in 1990. If we analyzed the total population by group and performed a t-test, we could expect a high p-value for the lower middle and upper middle groups because they intersect.

Since groups have different numbers of countries and people in each, I have provided the breakdown of countries/people per region/income group below.

Countries per region:
East Asia & Pacific: 36
South Asia: 8
Europe & Central Asia: 56
Sub-Saharan Africa: 48
Latin America & Caribbean: 38
Middle East & North Africa: 21
North America: 3

People per region:
East Asia & Pacific: 2,057,032,608 (33% of the world's population)
South Asia: 1,429,513,552 (23% of the world's population)
Europe & Central Asia: 863,676,818 (14% of the world's population)
Sub-Saharan Africa: 700,219,447 (11% of the world's population)
Latin America & Caribbean: 540,028,983 (9% of the world's population)
Middle East & North Africa: 323,928,401 (5% of the world's population)
North America: 319,050,105 (5% of the world's population)

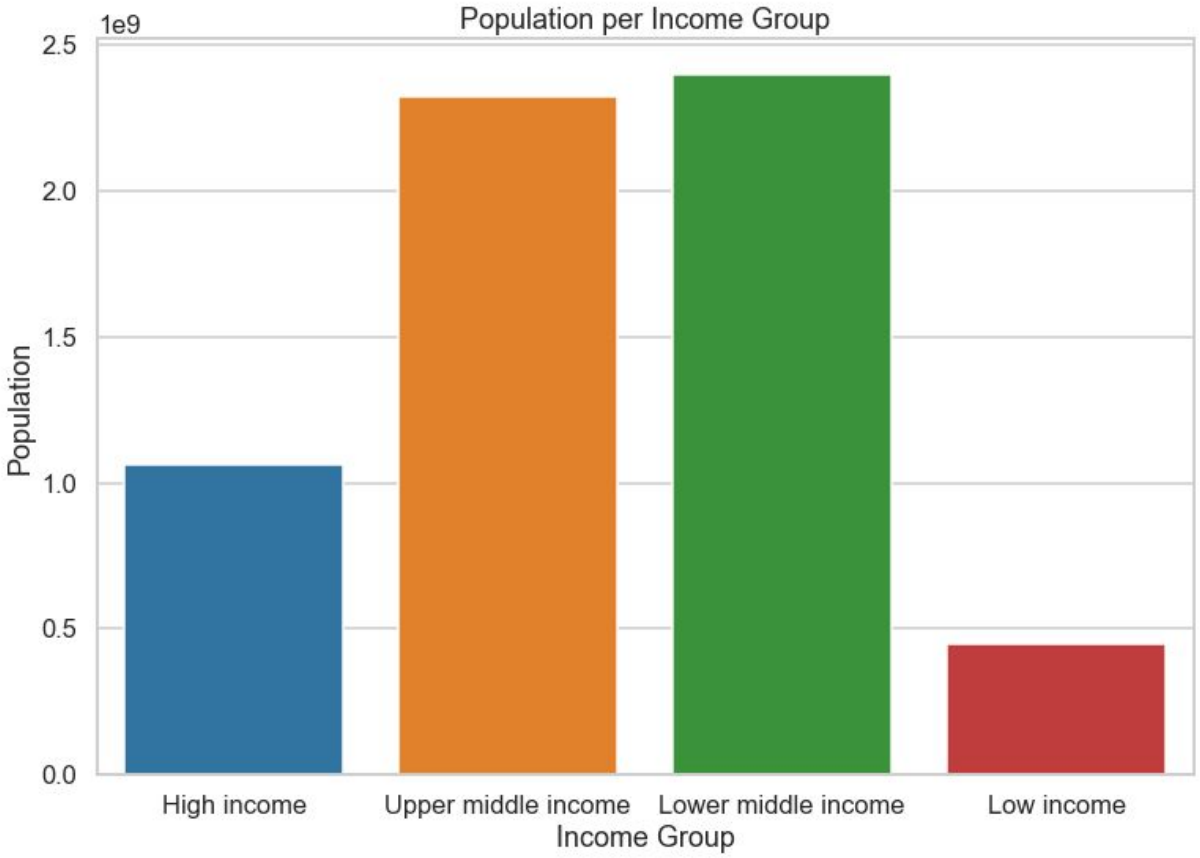


Countries per income group:

- High income: 72
- Upper middle income: 55
- Lower middle income: 52
- Low income: 31

People per income group:

- High income: 1,061,860,535 (17% of the world's population)
- Upper middle income: 2,323,809,235 (37% of the world's population)
- Lower middle income: 2,401,753,297 (39% of the world's population)
- Low income: 446,026,847 (7% of the world's population)



Machine Learning:

For the machine learning portion of the project, I used Lasso regression to determine which features were the best predictors of the target variable. I learned that the best predictors of the urban population five years in the future were the 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 %), and of course Year and Country.

Instead of using something such as time series forecasting for the model, I used linear regression. Without using cross validation, we get a score of 0.9994 for our linear regression. When using cross-validation, we get a minimum score of 0.9990, a maximum score of 0.9996, and an average score of 0.9993. Because the accuracy of a model can be affected by something as arbitrary as how the data is split into testing and training groups, we use cross validation to see a range of estimated scores. Because the linear regression model had such a high score, the next step would be feature engineering and to check for overfitting. Using Lasso Regression, we created a model that chose 41 features (including several one-hot-encoded features). From that,

we built a linear model in which we both scaled the data and used only the features used by the Lasso model. Ultimately, the basic linear regression model had a better r-squared score and a root mean squared error score, too. Therefore, we conclude that in this case, the basic linear regression model is the best choice.

Presentation:

https://docs.google.com/presentation/d/1mQEeIGmNSqkDK_2JB7N8m0YgFX0Ep5WsRVRNjC0uer4/edit?usp=sharing