# ORIE 4741 Midterm Report

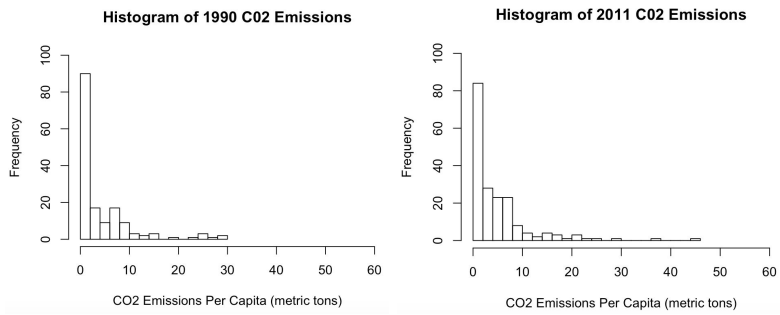
by Olia Javidi (oj33), Samantha Lee (srl84), Katie Li (yl2249)

## Introduction

As the world becomes more globalized, it is now easier for people to move to cities and other countries. In our project, we aim to analyze if changing urbanization is correlated to climate change and if so, what factors are possibly contributing to rising levels in CO2 emissions? We also aim to predict future levels of carbon emissions in different countries/regions to provide analysis on what places can benefit the most from investing in renewable power infrastructure.

## Dataset

### Initial Trend Analysis


We plotted CO2 emissions over the years by region and found that there is a general upward trend in emission levels, as seen from the red line that plots world CO2 emissions. We also see that regions are urbanizing at an increasing rate. In the histograms of C02 emissions, we see that there is a wider range of C02 emissions per capita reached in 2011 than in 1990. This is an interesting finding, because we will include year as a feature in our models to predict C02 emissions.

### Initial Filtering

Our original dataset was 5,656,458 rows, which is large and difficult to work with. We did an initial filter based on columns we assumed may be correlated with CO2 emissions. We figured indicators such as 'Life Expectancy at Birth, Female' and 'Life Expectancy at Birth, Male' could be summed up by 'Life Expectancy at Birth, Total' and that we'd rather evaluate 'Merchandise Imports and

Exports in Total', rather than breaking it down by where those imports and exports were going to. This allowed us to cut down the number of indicators from 81 to 23. Additionally, this data had entries from 1960 - 2014. We decided to focus on 1990-2011 because all CO2 emissions data entries end in 2011. Because our analysis isn't necessarily trying to analyze trends in data, but rather correlation based on different features, we don't need data that spans 50 years. This initial filtering brought our data down to 102,374 rows.

**Format**

Our initial dataset downloaded from Kaggle was formatted in a manner that made it hard to work with. The format as shown below had entries for each indicator with the value being its own column.

| | CountryName | CountryCode | IndicatorName | IndicatorCode | Year | Value |
|---|---|---|---|---|---|---|
| 0 | Arab World | ARB | Adolescent fertility rate (births per 1,000 wo... | SP.ADO.TFRT | 1960 | 1.335609e+02 |
| 1 | Arab World | ARB | Age dependency ratio (% of working-age populat... | SP.POP.DPND | 1960 | 8.779760e+01 |
| 2 | Arab World | ARB | Age dependency ratio, old (% of working-age po... | SP.POP.DPND.OL | 1960 | 6.634579e+00 |

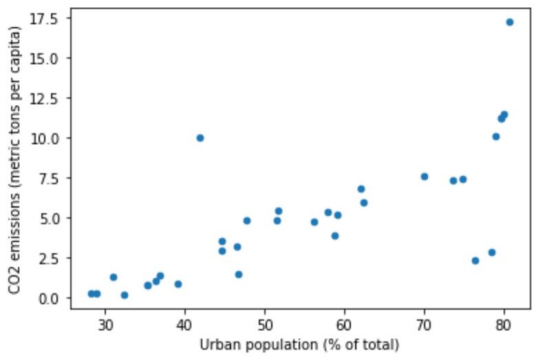
We used pandas in order to create a new table that took all the entries for a given IndicatorName, and left joined it all together based on CountryName and Year. From there we filtered out some of the columns that were unnecessary to our eventual evaluation.

| | CountryName | CountryCode | Year | Adolescent fertility rate (births per 1,000 women ages 15-19) | Population density (people per sq. km of land area) | GDP at market prices (current US$) | CO2 emissions (metric tons per capita) | Arms exports (SIPRI trend indicator values) | Birth rate, crude (per 1,000 people) | Death rate, crude (per 1,000 people) |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Arab World | ARB | 2000 | 53.829472 | 20.651464 | 7.307833e+11 | 3.717403 | NaN | 28.282837 | 6.465841 |
| 1 | Caribbean small states | CSS | 2000 | 71.619280 | 15.943475 | 3.220670e+10 | 6.617366 | NaN | 20.044584 | 7.146089 |

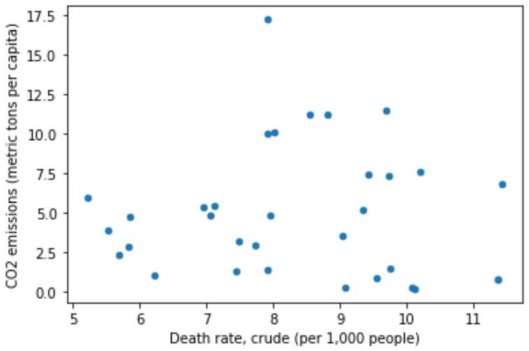
Our dataset is also organized such that each year the first 33 entries are regions (i.e. Arab World, European Union, East Asia) or descriptors (i.e. Low and middle income, Fragile and conflict affected situations, Heavily indebted poor countries) and the next 194 entries are countries. For this reason we split our dataset into these two categories because the regions and characteristics are helpful for understanding what's happening geographically, but are dependent on values in the rows of countries, therefore making our data interdependent. When we move to further analysis, we will use the regional and characteristic data in order to identify countries we should focus on as our findings dictate.

**Filtering Columns**

We then picked to hold a year constant and evaluate the correlation between each indicator and CO2 emissions. For this initial filter we were looking for data that was even somewhat correlated, and then validated it by filtering based on a different year and looking to see if correlations held steady. We did this once for individual countries and once for



characteristics/regions. Some indicators showed definite patterns while others did not.




Here, we can see that the urban population as a % of total population shows a linear relationship with CO2 emissions, likely making it a good predictor for later. On the other hand, death rate doesn't seem to have a strong relationship with CO2 emissions, and thus will probably not help in our analysis and may lead to overfitting.

After evaluating each plot, we narrowed it down to 8 features: Adolescent fertility rate (births per 1,000 women ages 15-19), Birth rate, crude (per 1,000 people), Fertility rate, total (births per woman), Life expectancy at birth, total (years), Mobile cellular subscriptions (per 100 people), Population, ages 0-14 (% of total), Population, ages 15-64 (% of total), and Urban population (% of total).

| | Only filtering by the 23 indicators | countries | Regions & characteristics |
|---|---|---|---|
| NaN | 40,816 | 501 | 14 |
| Total Rows | 255,568 | 4,268 | 726 |

## Preliminary Model

Through our correlation analysis, we created a preliminary linear regression analysis to test the significance of the 8 features we selected, along with year, against the dependent variable CO2 emissions (metric tons per capita). Before the analysis, we assumed that all 8 of our selected variables to be significant. In the analysis, we found that all features except for adolescent fertility rate were significant.

```
Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        2.555e+02  4.545e+01   5.621 2.10e-08 ***
Year                                              -1.774e-01  2.277e-02  -7.792 9.43e-15 ***
Adolescent.fertility.rate..births.per.1.000.women.ages.15.19. -5.192e-04  3.850e-03  -0.135 0.892754
Birth.rate..crude..per.1.000.people.              -6.001e-01  6.793e-02  -8.835  < 2e-16 ***
Fertility.rate..total..births.per.woman.           6.067e+00  3.380e-01  17.949  < 2e-16 ***
Life.expectancy.at.birth..total..years.            8.315e-02  2.109e-02   3.942 8.28e-05 ***
Mobile.cellular.subscriptions..per.100.people.     1.374e-02  4.031e-03   3.408 0.000665 ***
Population..ages.0.14....of.total.                 3.552e-01  4.134e-02   8.592  < 2e-16 ***
Population..ages.15.64....of.total.                1.291e+00  4.412e-02  29.267  < 2e-16 ***
Urban.population....of.total.                      5.535e-02  5.718e-03   9.679  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

## Next Steps

Our next steps are to create more predictive models, especially to expand beyond linear relationships between features and carbon emissions. We will also look at non-linear models and interaction terms. As we move into non-linear models, we will validate our models through k-fold cross validation and determine what order of polynomial we might use by plotting test MSE for a range of degree polynomials. Our predictive model was only for countries, and as we described earlier, we will dig deeper into descriptors which further describe regional dynamics.