Individual Goal 1: Determine which factors have the highest correlation with emissions - Divya

Colab link: https://colab.research.google.com/drive/1v24_sp21didHWpQGBAJgE5l6w1rZhZaL?usp=sharing

METHOD

I used linear regression, a supervised learning algorithm, used in this case to model the relationship between total CO_2 emissions and the global development variables in this dataset. I chose this method because it not only identified which development variables correlate with emissions, but also assigned a weight to each of these variables. These weights allowed me to see which variables have the strongest impact on CO_2 emissions.

This was run by first **selecting the top correlated variables** using the df.corr command. Then I **standardized** the data, then I split the data (80% train, 20% test) in order to test performance. Finally, I trained the linear model. The model learned a weight for each variable, allowing us to see how strongly each variable influenced CO_2 emissions. I used the **k-fold cross validation** to evaluate the model's performance, but did not have any parameters to fine tune for this model.

WRANGLING

Before fitting the model, I removed columns that contained information about other emissions, since I primarily wanted to focus on the correlation between the global development variables and CO_2 emissions.

EVALUATION PROCESS

I used the **k-fold cross validation** to evaluate the model's performance. I learned this method from the Scikit-learn documentation. This method allowed me to train and test the model 5 times (fold 1-5), and the average R^2 value of these folds was 95%, meaning this model consistently explained 95% of the variation in CO_2 emissions, which is quite strong. Also, the R^2 value for each of the folds are high and similar to each other, showing our model is stable and is not overfitting.

RESULTS:

- **Telephone infrastructure**: The number or telephone lines and mobile cellular subscriptions correlate very highly with CO_2 emissions. This suggests that telephone infrastructure may be linked with energy consumption
- **Land use**: Variables regarding land surface area, agricultural land, and land area have high and positive correlations and weights. This suggests that larger countries and countries that take part in more agricultural activities contribute to more CO_2 emissions
- **GDP**: Country GDP had a strong positive correlation and high weight. This suggests that countries that are more industrialized contribute more to CO_2 emissions.
- **Population**: Country population had a high correlation and weight, suggesting that high population leads to higher CO_2 emissions, most probably due to the increased need of energy and resources to sustain a large population.
- **Rural population variable**: This variable was interesting because it had a correlation of about 0.5 but a weight of about -132.29. This suggests that higher rural populations may actually reduce CO_2 emissions. This could be because rural areas generally have less industrial activity compared to urban or suburban areas.

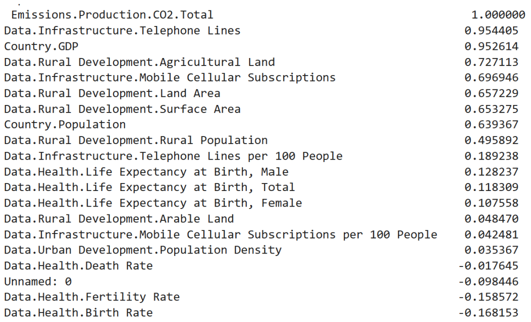
CONCLUSIONS:

This analysis allowed us to see specifically which development variables have the most impact on CO_2 emissions, which is a large part of our business goal. We know that variables with strong correlations and high, positive weights (notably the variables regarding telephone lines, rural land surface area, rural agricultural land, and country GDP) contribute most to higher CO_2 emissions.

APPENDIX AND DATA DICTIONARY:

The columns used for modeling and their correlations are listed below:

A description of each of these columns can be found here



```
 Emissions.Production.CO2.Total                                  1.000000
Data.Infrastructure.Telephone Lines                             0.954405
Country.GDP                                                     0.952614
Data.Rural Development.Agricultural Land                        0.727113
Data.Infrastructure.Mobile Cellular Subscriptions               0.696946
Data.Rural Development.Land Area                                0.657229
Data.Rural Development.Surface Area                             0.653275
Country.Population                                              0.639367
Data.Rural Development.Rural Population                         0.495892
Data.Infrastructure.Telephone Lines per 100 People              0.189238
Data.Health.Life Expectancy at Birth, Male                      0.128237
Data.Health.Life Expectancy at Birth, Total                     0.118309
Data.Health.Life Expectancy at Birth, Female                    0.107558
Data.Rural Development.Arable Land                              0.048470
Data.Infrastructure.Mobile Cellular Subscriptions per 100 People    0.042481
Data.Urban Development.Population Density                       0.035367
Data.Health.Death Rate                                         -0.017645
Unnamed: 0                                                     -0.098446
Data.Health.Fertility Rate                                     -0.158572
Data.Health.Birth Rate                                         -0.168153
```

FUTURE WORK:

This analysis can be useful to environmental policy makers as they can implement laws and regulations regarding certain activities that cause high CO_2 emissions. For example, regulations can be put into place to promote more energy efficient and sustainable agricultural practices. Urban and rural planners can also use this analysis to develop urban and rural land in a sustainable way. This can be done by promoting the use of green infrastructure.

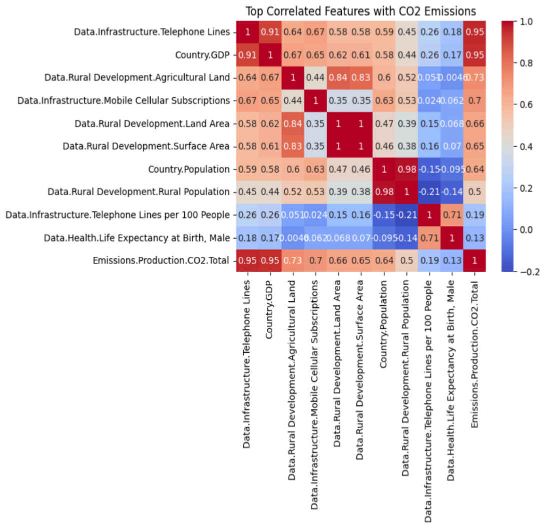
LIMITATIONS:

This dataset had many global development related variables, and we wanted to see how each one correlated with CO_2 emissions. However, certain variables like Country GDP and Country Population have some overlap which can affect the model's accuracy.

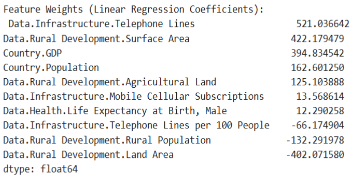
CHALLENGES:

Different variables were measured with different units (for example, GDP is measured in billions of dollars and population is in millions). In order to mitigate this issue I made sure to standardize the data so that the variables contributed equally to the model.

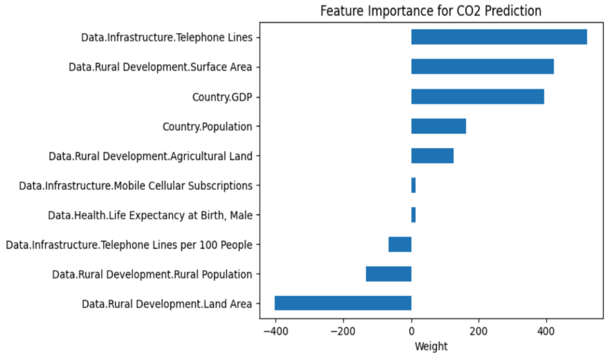
Heatmap:



Weights:



```
Feature Weights (Linear Regression Coefficients):
 Data.Infrastructure.Telephone Lines                  521.036642
Data.Rural Development.Surface Area                  422.179479
Country.GDP                                          394.834542
Country.Population                                   162.601250
Data.Rural Development.Agricultural Land             125.103888
Data.Infrastructure.Mobile Cellular Subscriptions     13.568614
Data.Health.Life Expectancy at Birth, Male            12.290258
Data.Infrastructure.Telephone Lines per 100 People   -66.174904
Data.Rural Development.Rural Population             -132.291978
Data.Rural Development.Land Area                    -402.071580
dtype: float64
```



Individual Goal 2: Development over time leads to a change in CO_2 emissions - Ashwini

Colab Link: DevelopmentStatusVSEmissions.ipynb

METHOD:

To see if there was an impact of the development status of a country on CO_2 emissions, I decided to try out a few different supervised and unsupervised learning methods to see if there was anything that could represent the dataset well. The most effective method seems to be multiple linear regression. This method was chosen because it allows us to quantify the impact of development status on the total CO_2 emissions and makes it communicable in a clear and concise way. To ensure that the code ran successfully, I decided to use the OneHotEncoder function in the Scikit library to convert the Development Status into a Numerical variable. I found out about this function while looking through the internet, and seeing their module. I also used data from the IMF to find information about the development status of these countries. Sadly, I was unable to find yearly data on this topic and had to do the analysis based on the most recent data from 2023.

WRANGLING:

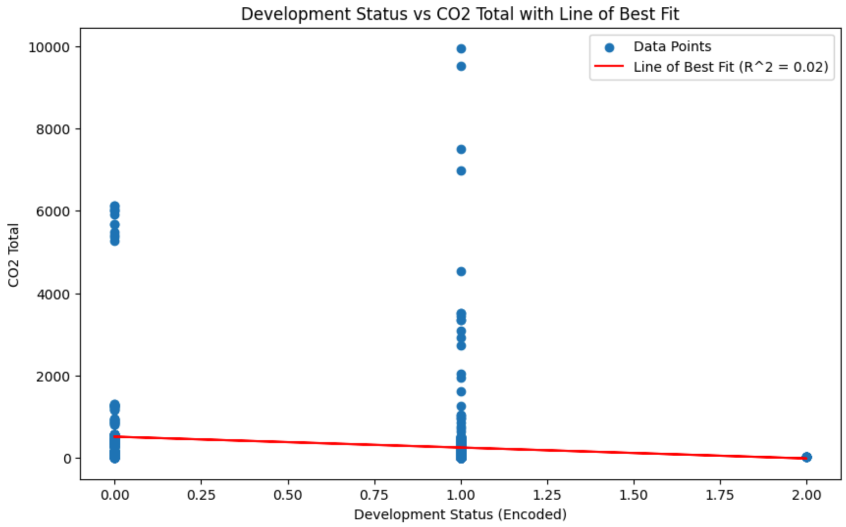
The most important columns for this analysis are the Development Status column and the CO_2 emissions column since they are necessary for us to provide the analysis.

EVALUATION PROCESS:

I used R-squared to see the 'Goodness of Fit' to find the proportion of variance in our target variable which is total CO_2 emissions in this case, and the MSE to get the average of the squared prediction errors which helps us see how much the model's predictions are off by on average.

RESULTS:

After running the code, I found the R^2 to be -0.028 which implies a negative correlation between the two variables of interest. We also get the MSE to be 1398355.08 which helps us see that the model's predictions are off by a really high amount. These results imply that our model is not well fitted, and does not have a significant impact on each other. I think the fact that the development statuses are not historically accurate could be to blame for this, however, it is not something that can be confirmed without additional data and would require further research.



CONCLUSIONS:

Based on the data that we currently have, and our Business Goal to determine how different factors of global development contribute to emissions, we can conclude that there is no strong evidence of any correlation between CO_2 emissions and the development status of a country.

APPENDIX AND DATA DICTIONARY:

The columns used and created during the modeling:

- 'Development Status': The development status of the country in the year 2023
- 'Emissions.Production.CO2.Total': The total amount of CO_2 produced (kilotons)

Information about the other columns can be found on the Github wiki: https://github.com/zpenguin19/Intro-to-Data-Science/wiki/Merged-Dataset-Data-Dictionary

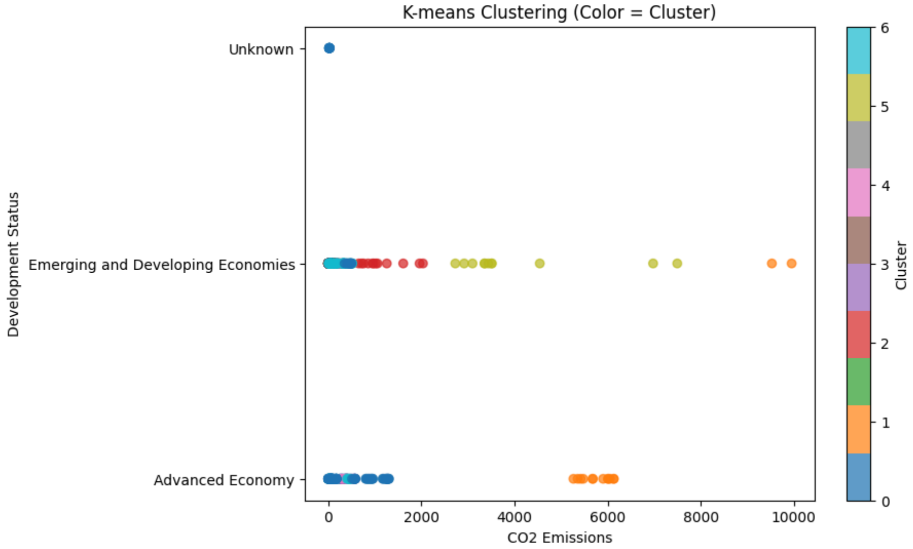
FUTURE WORK:

Should these results have turned out as expected, there are a lot of groups that could use these results, such as:

- Policymakers to shape climate policies and sustainability initiative
- Environmental agencies to track and report how development contributes to emissions globally
- Urban planners to plan smarter cities and balance urban growth with emissions
- NGOs and Advocacy Groups to push for targeted reforms in high-emission regions
- International Development organizations to design funding programs that address both development and sustainability
- Others like researchers, investors, and economists can use this to assess environmental risks tied to the development levels of countries.

CHALLENGES & LIMITATIONS:

The biggest challenge in answering this question was finding data consistent and accurate data about the development statuses of these countries. I had to do a lot of looking, and there were no sources that had consistent data, which resulted in me finally just choosing to use the most updated and consistent data I could find - which happened to be from 2023- a year ten years ahead of our dataset. I tried representing the dataset with K-means Clustering but was unable to get legitimately useful results due to the fact that the development status was consistent across the years.



Individual Goal 3: Create a model to predict the amount of CO_2 emissions in a future year for a given country - Zoe
Colab Link: EmissionTrendPredictions.ipynb
METHOD:

- Stepwise regression combines the advantages of both forward and backward regression algorithms. I chose to use this method to determine which variables would be the most important to include in my model since the algorithm allows for both adding and removing variables based on their significance. Variables that seem significant at first but change after adding more variables can be removed if necessary, but can be added back if they are shown to be important for the model.
- I ran the following code for the stepwise regression model:

  ```
  sfs_step = sfs(lreg, k_features=6, forward=True, floating= True, verbose=2,
  scoring='r2', cv= 5)
  ```
- The only parameter I changed from the code given in class was `k_features=6`, since I wanted to use a few more than 4 variables in my model but not more than necessary.
- The particular parameter I changed was pretty straightforward, and an explanation for it was provided in the Google Colab given in class at this link: Regression.ipynb

WRANGLING:
Before determining the features that would be used in the model, I removed several columns that I knew I did not want to include. These columns are:

- `Country.Code` - the name of the country is given as a part of `Country_Year`
- `Emissions.Production.CH4` and `Emissions.Production.N2O` - our goal only regards CO_2 emissions, not emissions for any other chemical
- `Data.Rural Development.Agricultural Land Percent`, `Data.Rural Development.Arable Land Percent`, and `Data.Urban Development.Urban Population Percent` - for this model, I only care about the total agricultural, arable, and urban land / population to predict the total number of kilotons of CO_2 emitted
- All columns that begin with `Emissions.Global Share` - I only want to predict the number of kilotons of CO_2 a country will produce rather than the global percentage

After determining the features I would be using in the model, I removed all columns except for `Country_Year`, the six explanatory variables, and `CO2Total`.

Since I wanted a model that would predict future CO_2 emissions for a particular country, I removed all rows that were not for that country once I had selected a country to test the model for.

EVALUATION PROCESS:
I chose one country and one year to focus on when evaluating the model. The code in the Google Colab can be changed to be run for any country in the dataset and for any year beyond what is included, but I chose to test Canada in 2020 (which is a future year as our dataset only contains data up to about 2013).

For each of the six explanatory variables, I determined a line of best fit for a linear model, a quadratic model, and an exponential model. I also got the R^2 values for each to determine which of those three regression models would fit the best for each variable. Once I chose the best model, I used the equation provided by that model, plugging in the year I wanted to predict for, to determine the predicted value of that explanatory variable. After I determined the predicted values for each of the six explanatory variables, I was able to use the stepwise regression model to predict the number of kilotons of CO_2 emitted in the selected country in the desired year.

RESULTS:
The six features that were selected by stepwise regression were `TelephoneLines`, `GDP`, `AgriLand`, `RuralPop`, `ArableLand`, and `PopDensity` (each defined below in the Data Dictionary section). The mean squared error was 62791.7879, which represents the average squared difference between the actual number of kilotons of CO_2 emitted in a

particular country during a single given year (found in the column `CO2Total`) and the predicted number of kilotons from the stepwise regression model. The absolute value of the difference between actual and predicted is 250.58 kilotons.

CONCLUSIONS:

The variables chosen by stepwise regression tend to have relatively strong correlations with total CO_2 emissions. The MSE of the model ended up being pretty high given the context, meaning the model does not fit especially well with the data we had. This may be due to wide discrepancies between countries that emit thousands of kilotons of CO_2 per year and countries that emit fewer than ten kilotons of CO_2 per year.

APPENDIX AND DATA DICTIONARY:

Information about the other columns can be found on the Github wiki:
https://github.com/zpenguin19/Intro-to-Data-Science/wiki/Merged-Dataset-Data-Dictionary

These are the columns I used:

- `Country_Year` - the name of the country and the year the data was collected
- `Country.GDP` - a country's gross domestic product for a given year
- `Emissions.Production.CO2.Total` - The total amount of CO_2 produced (kilotons)
- `Data.Infrastructure.Telephone Lines` - the total number of telephone lines
- `Data.Rural Development.Agricultural Land` - total area of agricultural land
- `Data.Rural Development.Arable Land` - total area of arable land
- `Data.Rural Development.Rural Population` - number of people living in rural areas
- `Data.Urban Development.Population Density` - density of people living in urban areas

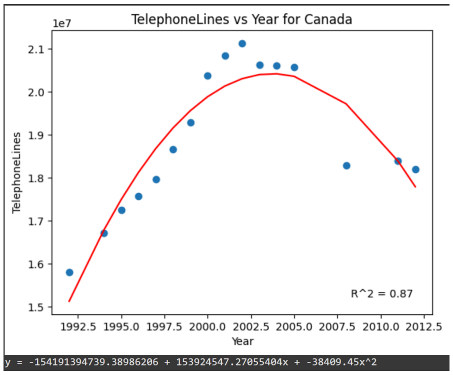
FUTURE WORK:

This analysis can be useful to environmental lawmakers as they can use the prediction model to determine how some of the most correlated variables will change in upcoming years as well as how these changes can affect CO_2 emissions. With the predictions provided, lawmakers could put policies in place to mitigate some of the changes.

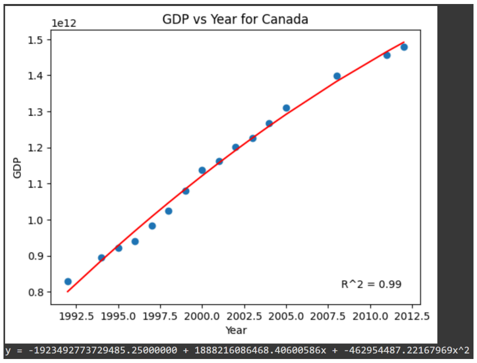
CHALLENGES AND LIMITATIONS:

One challenge I had was predicting values for certain explanatory variables. For example, at one point I got a predicted value of around 1600 for arable land in Canada when previous values were between 1 and 2 and were decreasing over time. I even got a negative result for population density at one point. I realized that this issue was due to the fact that the regression equation I had stored was only using coefficients with two decimal places, and rounding the coefficients to this degree was significantly decreasing the accuracy of prediction, so I increased the number of decimal places to 8 for each coefficient to make sure the regression equation made sense and provided values that were reasonable.

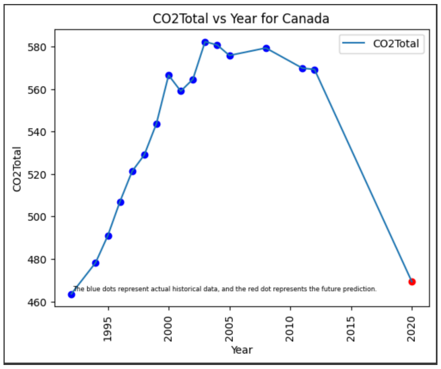
The dataset does not provide information past 2013, and does not include data from every country. It likely will not be able to be used to predict future years well past 2013, as there are many variables that need to be considered and each one needs to be extrapolated to determine a prediction for CO_2 emissions.



```
y = -154191394739.38986206 + 153924547.27055404x + -38409.45x^2
```



```
y = -1923492773729485.25000000 + 1888216086468.40600586x + -462954487.22167969x^2
```



Business Goal: Determine how different factors of global development contribute to emissions

Colab Link: HW7_BusinessGoal.ipynb

METHOD:

To see how different factors of global development contributed to emissions, we decided to create a heatmap comparing the different columns to each other to see if there was any correlation. This was one of the best methods to use to answer our business goal since it helped us see the correlations of all the columns vs CO_2 emissions efficiently. To ensure that the model ran successfully, we had to convert the Development Status column into multiple different columns for each status, where 1.0 means true, and 0.0 means false. We accomplished this using the OneHotEncoder function that was used in one of the individual goals above after it was found in the SciKit library.

WRANGLING:

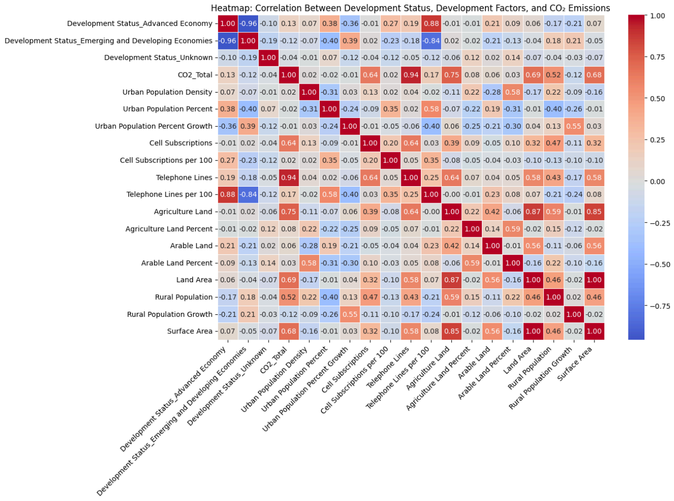
Before any analysis could be performed, we had to drop all unnecessary columns, leaving us with: 'Development Status', 'CO2_Total', 'Urban Population Density', 'Urban Population Percent', 'Urban Population Percent Growth', 'Cell Subscriptions', 'Cell Subscriptions per 100', 'Telephone Lines', 'Telephone Lines per 100', 'Agriculture Land', 'Agriculture Land Percent', 'Arable Land', 'Arable Land Percent', 'Land Area', 'Rural Population', 'Rural Population Growth', 'Surface Area'. We also had to convert Development Status into three different numerical columns: 'Development Status_Advanced Economy', 'Development Status_Emerging and Developing Economies', and 'Development Status_Unknown', to allow for analysis.

EVALUATION PROCESS:

We used the k-fold cross validation to evaluate the model's performance. We learned this method from the Scikit-learn documentation. This method allowed us to train and test the model 5 times (fold 1-5), and the average R^2 value of these folds was 88%. This means the model consistently explained 88% of the variation in CO_2 emissions, which is strong. We also see that the R^2 values of each fold are quite similar, meaning that our model is not overfitting.

RESULTS:

Telephone Lines, Agriculture Land, Land Area, Surface Area, Cell Subscriptions, and Rural Population seem to be the variables with the highest correlation with CO_2 emissions. This implies that countries with more land/surface area and infrastructure(Telephone lines, cell subscriptions) seem to have higher energy use and thus more emissions. Agricultural land also contributes heavily to emissions due to deforestation and heavy machinery usage. Rural Populations also have a high correlation with CO_2 emissions due to high transportation emissions, and agricultural emissions.



Heatmap: Correlation Between Development Status, Development Factors, and CO_2 Emissions

	Development Status_Advanced Economy	Development Status_Emerging and Developing Economies	Development Status_Unknown	CO2_Total	Urban Population Density	Urban Population Percent	Urban Population Percent Growth	Cell Subscriptions	Cell Subscriptions per 100	Telephone Lines	Telephone Lines per 100	Agriculture Land	Agriculture Land Percent	Arable Land	Arable Land Percent	Land Area	Rural Population	Rural Population Growth	Surface Area
Development Status_Advanced Economy	1.00	-0.96	-0.10	0.13	0.07	0.38	-0.36	-0.01	0.27	0.19	0.88	-0.01	-0.01	0.21	0.09	0.06	-0.17	-0.21	0.07
Development Status_Emerging and Developing Economies	-0.96	1.00	-0.19	-0.12	-0.07	-0.40	0.39	0.02	-0.23	-0.18	-0.84	0.02	-0.02	-0.21	-0.13	-0.04	0.18	0.21	-0.05
Development Status_Unknown	-0.10	-0.19	1.00	-0.04	-0.01	0.07	-0.12	-0.04	-0.12	-0.05	-0.12	-0.06	0.12	0.02	0.14	-0.07	-0.04	-0.03	-0.07
CO2_Total	0.13	-0.12	-0.04	1.00	0.02	-0.02	-0.01	0.64	0.02	0.94	0.17	0.75	0.08	0.06	0.03	0.69	0.52	-0.12	0.68
Urban Population Density	0.07	-0.07	-0.01	0.02	1.00	-0.31	0.03	0.13	0.02	0.04	-0.02	-0.11	0.22	-0.28	0.58	-0.17	0.22	-0.09	-0.16
Urban Population Percent	0.38	-0.40	0.07	-0.02	-0.31	1.00	-0.24	-0.09	0.35	0.02	0.58	-0.07	-0.22	0.19	-0.31	-0.01	-0.40	-0.26	-0.01
Urban Population Percent Growth	-0.36	0.39	-0.12	-0.01	0.03	-0.24	1.00	-0.01	-0.05	-0.06	-0.40	0.06	-0.25	-0.21	-0.30	0.04	0.13	0.55	0.03
Cell Subscriptions	-0.01	0.02	-0.04	0.64	0.13	-0.09	-0.01	1.00	0.20	0.64	0.03	0.39	0.09	-0.05	0.10	0.32	0.47	-0.11	0.32
Cell Subscriptions per 100	0.27	-0.23	-0.12	0.02	0.02	0.35	-0.05	0.20	1.00	0.05	0.35	-0.08	-0.05	-0.04	-0.03	-0.10	-0.13	-0.10	-0.10
Telephone Lines	0.19	-0.18	-0.05	0.94	0.04	0.02	-0.06	0.64	0.05	1.00	0.25	0.64	0.07	0.04	0.05	0.58	0.43	-0.17	0.58
Telephone Lines per 100	0.88	-0.84	-0.12	0.17	-0.02	0.58	-0.40	0.03	0.35	0.25	1.00	-0.00	-0.01	0.23	0.08	0.07	-0.21	-0.24	0.08
Agriculture Land	-0.01	0.02	-0.06	0.75	-0.11	-0.07	0.06	0.39	-0.08	0.64	-0.00	1.00	0.22	0.42	-0.06	0.87	0.59	-0.01	0.85
Agriculture Land Percent	-0.01	-0.02	0.12	0.08	0.22	-0.22	-0.25	0.09	-0.05	0.07	-0.01	0.22	1.00	0.14	0.59	-0.02	0.15	-0.12	-0.02
Arable Land	0.21	-0.21	0.02	0.06	-0.28	0.19	-0.21	-0.05	-0.04	0.04	0.23	0.42	0.14	1.00	-0.01	0.56	-0.11	-0.06	0.56
Arable Land Percent	0.09	-0.13	0.14	0.03	0.58	-0.31	-0.30	0.10	-0.03	0.05	0.08	-0.06	0.59	-0.01	1.00	-0.16	0.22	-0.10	-0.16
Land Area	0.06	-0.04	-0.07	0.69	-0.17	-0.01	0.04	0.32	-0.10	0.58	0.07	0.87	-0.02	0.56	-0.16	1.00	0.46	-0.02	1.00
Rural Population	-0.17	0.18	-0.04	0.52	0.22	-0.40	0.13	0.47	-0.13	0.43	-0.21	0.59	0.15	-0.11	0.22	0.46	1.00	0.02	0.46
Rural Population Growth	-0.21	0.21	-0.03	-0.12	-0.09	-0.26	0.55	-0.11	-0.10	-0.17	-0.24	-0.01	-0.12	-0.06	-0.10	-0.02	0.02	1.00	-0.02
Surface Area	0.07	-0.05	-0.07	0.68	-0.16	-0.01	0.03	0.32	-0.10	0.58	0.08	0.85	-0.02	0.56	-0.16	1.00	0.46	-0.02	1.00

CONCLUSIONS:

Looking at the heatmap, we can see that some variables have a high correlation with CO_2 Emissions. Individual goal 1 looked to determine which factors have the highest correlation with emissions and found that the variables regarding telephone lines, rural land surface area, rural agricultural land, and country GDP contribute most to higher CO_2 emissions. Individual goal 2 looked to see if the development over time leads to a change in CO_2 emissions, and found that there were no significant correlations between the two. Individual goal 3 created a model to predict the amount of CO_2 emissions in a future year for a given country and found that similar variables have the highest impact on CO_2 emissions.

TIMELINE:

Planned timeline (from HW6):

- Before our next meeting, we will each do additional EDA as necessary to determine correlations between necessary variables
- Meet on **April 7th** to discuss our findings and discuss the final week of the project
- On **April 12th**, we will discuss our individual portions of the project and begin work on the group portion
- By **April 13th**, we will have finished writing our presentation to be submitted, and do any final reviews
- Finalize the presentation on **April 14th** and submit the assignment

Executed timeline:

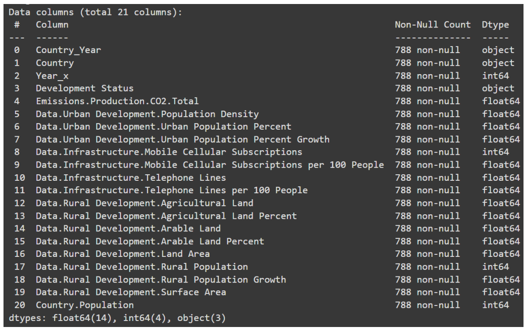
- Met up on **April 8th** to allocate individual goals to each team member
- Work on individual goals and analysis over the course of the rest of the week
- Met up on the morning of **April 13th** (rather than **April 12th** as expected) to discuss individual methods and findings, and work on the business goal; we are also working on the presentation slideshow today

RELEVANT LINKS:

GitHub link: https://github.com/zpenguin19/Intro-to-Data-Science

APPENDIX AND DATA DICTIONARY:

The columns used and created during the modeling:



```
Data columns (total 21 columns):
 #   Column                                                    Non-Null Count  Dtype  
---  ------                                                    --------------  -----  
 0   Country_Year                                              788 non-null    object 
 1   Country                                                   788 non-null    object 
 2   Year_x                                                    788 non-null    int64  
 3   Development Status                                        788 non-null    object 
 4   Emissions.Production.CO2.Total                            788 non-null    float64
 5   Data.Urban Development.Population Density                 788 non-null    float64
 6   Data.Urban Development.Urban Population Percent           788 non-null    float64
 7   Data.Urban Development.Urban Population Percent Growth    788 non-null    float64
 8   Data.Infrastructure.Mobile Cellular Subscriptions         788 non-null    int64  
 9   Data.Infrastructure.Mobile Cellular Subscriptions per 100 People  788 non-null    float64
 10  Data.Infrastructure.Telephone Lines                       788 non-null    float64
 11  Data.Infrastructure.Telephone Lines per 100 People        788 non-null    float64
 12  Data.Rural Development.Agricultural Land                  788 non-null    float64
 13  Data.Rural Development.Agricultural Land Percent          788 non-null    float64
 14  Data.Rural Development.Arable Land                        788 non-null    float64
 15  Data.Rural Development.Arable Land Percent                788 non-null    float64
 16  Data.Rural Development.Land Area                          788 non-null    float64
 17  Data.Rural Development.Rural Population                   788 non-null    int64  
 18  Data.Rural Development.Rural Population Growth            788 non-null    float64
 19  Data.Rural Development.Surface Area                       788 non-null    float64
 20  Country.Population                                        788 non-null    int64  
dtypes: float64(14), int64(4), object(3)
```

Information about the columns can be found here.

FUTURE WORK:

These results can be used by:

- Governments: To create targeted policy design around emissions and land/urban development
- UN/World Bank/IMF: To have data-driven funding and planning of sustainable development initiatives
- Scientists/Researchers: To model design, simulation support, and feature validation
- Urban Planners: To create smarter infrastructure planning with climate impact in mind
- NGOs/Environmental Advocates: To create visuals and arguments for reform and public awareness
- ESG Analysts/Investors: To evaluate country-level sustainability and emissions risk

LIMITATIONS:

Although our dataset has many global development variables, having information on more topics could allow us to see a bigger picture. For example, the number of vehicles used per country and the fuel type for these vehicles could make our analysis more in-depth, especially since we know that these factors play a large part in CO_2 emissions.

CHALLENGES:

A certain variable may have a high correlation with emissions, but that does not necessarily mean that they hold a high weight. This is why we calculated the weights of each of the highly correlated variables, this allowed us to get a better understanding on which variables had more of an impact on high CO_2 emissions.