Global Carbon Emission Comparisons

Group 15 - Ruchita, Avneesh, Tarini, Kobe, Mikhail

Topic Description

- Alarming rates of carbon emissions
- Study major factors affecting climate conditions
- Analyze correlations between energy usage and GHG emissions through transportation
- Determine the recent impact of alternative energy sources
- Predict future impact of higher alternative energy adoption

Group Roles

- X - Avneesh
- Triangle - Tarini
- Circle - Mikhail
- Square - Ruchita & Kobe

Data Description

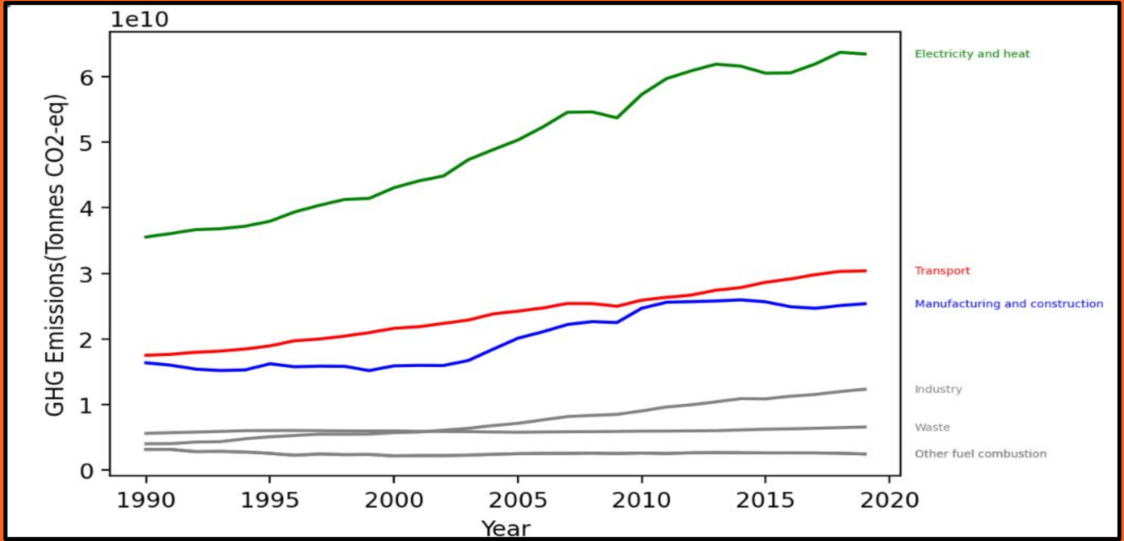
- Annual worldwide vehicle mix by different energy types (Diesel, Electric, Hybrid, etc)
- Overall GHG emissions by sector
- Market demand of all vehicles types
- Energy production
- Gathered data to determine our main goal of proving that the introduction of EV will decrease annual gas emissions on a global scale

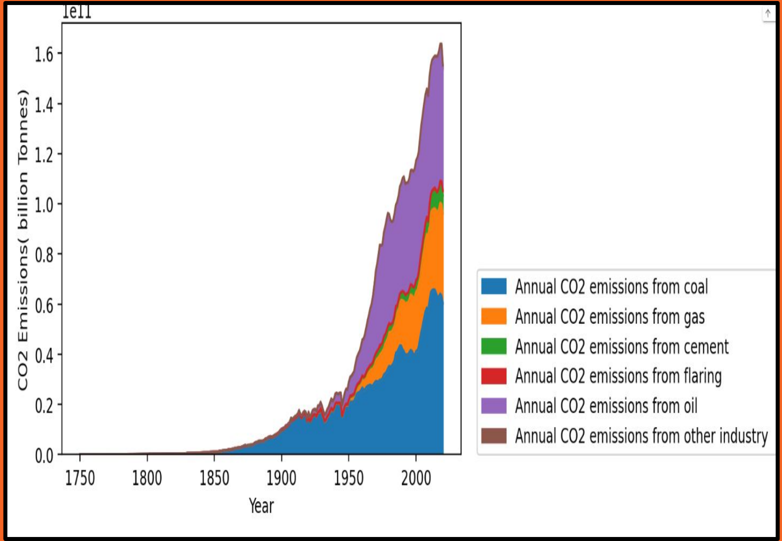
Machine Learning Preliminary Process

Conducted Data Exploration to cover the following:

- All datasets were merged based on country and year
- Extracted EV demand dataset from original sheet
- Remove non-value variables and refined the column names to match across all merged datasets
- Loaded the GHG emissions and EV demand datasets as pandas dataframe
- Defined the merge logic to join both the datasets

Data Analysis Phase



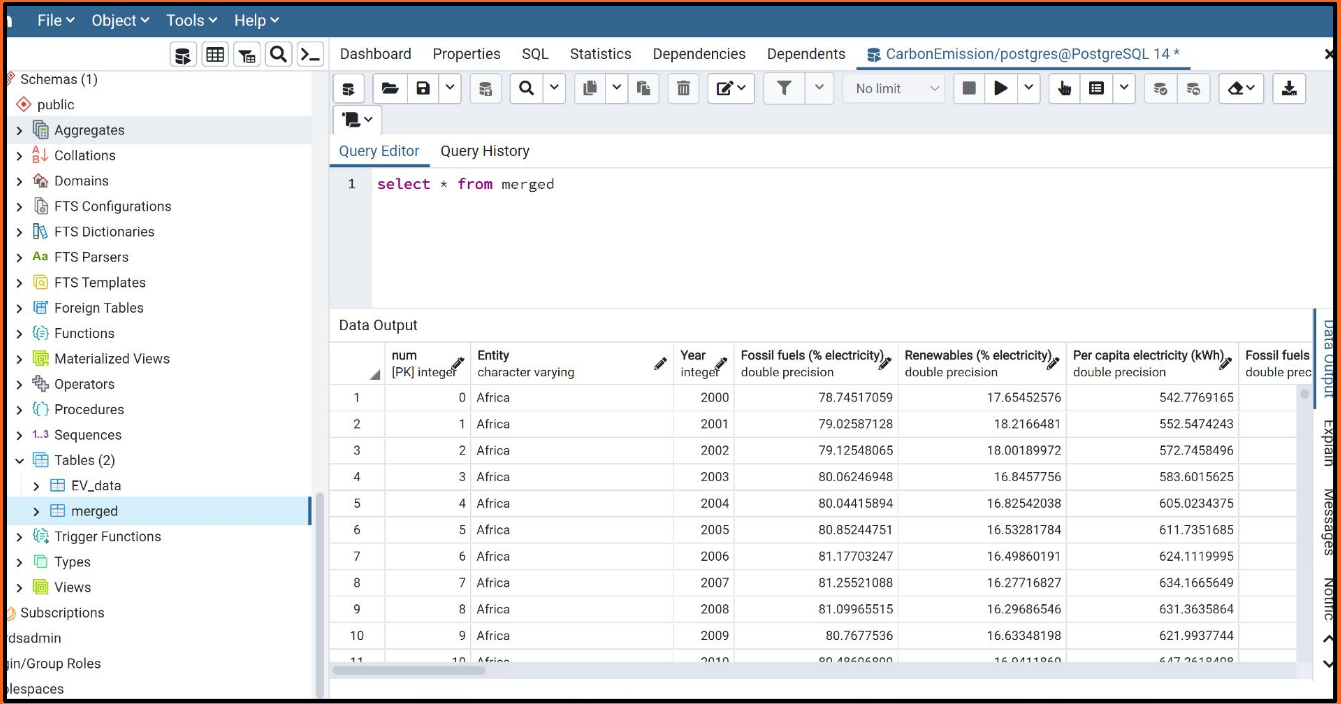


- The major focus was to analyze how much CO_2 emissions were saved when introducing EV variable into the transport sector.
- Out of all factors, **electricity & heat**, **transport**, and **manufacturing & construction**.

Database

- Downloaded datasets related to co2 emissions and GHG emissions
  - Source - https://ourworldindata.org/
- Tried to find the unique key to merge the data in this case it is the country name and year
- Collected the datasets for Transport Vehicle demand country wise and year wise, merged it with the GHG emissions database
- **WIP: Two tables, one join (not including pandas), one connection string**

Database Sample



Machine Learning Model Summary

OLS Regression Results

Dep. Variable:	ev_savings	**R-squared (uncentered):**	0.507
Model:	OLS	**Adj. R-squared (uncentered):**	0.493
Method:	Least Squares	**F-statistic:**	36.02
Date:	Thu, 01 Sep 2022	**Prob (F-statistic):**	2.36e-75
Time:	03:50:38	**Log-Likelihood:**	-8982.0
No. Observations:	576	**AIC:**	1.800e+04
Df Residuals:	560	**BIC:**	1.807e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Fossil fuels (% equivalent primary energy)	7950.8993	5231.081	1.520	0.129	-2324.038	1.82e+04
Transport_x	0.1986	0.161	1.234	0.218	-0.118	0.515
Agriculture	0.0011	0.001	1.013	0.312	-0.001	0.003
Land-use change and forestry	0.0015	0.001	2.731	0.007	0.000	0.003
Waste	0.0082	0.005	1.731	0.084	-0.001	0.018
Industry	0.0584	0.005	10.943	0.000	0.048	0.069
Manufacturing and construction	-0.0240	0.002	-11.381	0.000	-0.028	-0.020
Transport_y	-0.1989	0.155	-1.280	0.201	-0.504	0.106
Electricity and heat	0.0080	0.001	7.096	0.000	0.006	0.010
Buildings	-0.0098	0.003	-3.043	0.002	-0.016	-0.003
Fugitive emissions	-0.0062	0.002	-3.892	0.000	-0.009	-0.003
Other fuel combustion	0.0040	0.016	0.246	0.805	-0.028	0.036
Aviation and shipping	-0.0209	0.006	-3.501	0.001	-0.033	-0.009
Renewables (% equivalent primary energy)	5839.8990	1.32e+04	0.444	0.657	-2e+04	3.17e+04
Fossil fuels (% electricity)	-1.103e+04	5700.997	-1.935	0.054	-2.22e+04	168.776
Renewables (% electricity)	-4951.9647	9257.701	-0.535	0.593	-2.31e+04	1.32e+04

Renewables (% equivalent primary energy)	5839.8990	1.32e+04	0.444	0.657	-2e+04	3.17e+04
Fossil fuels (% electricity)	-1.103e+04	5700.997	-1.935	0.054	-2.22e+04	168.776
Renewables (% electricity)	-4951.9647	9257.701	-0.535	0.593	-2.31e+04	1.32e+04

Omnibus:	416.999	**Durbin-Watson:**	1.018
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	21995.226
Skew:	2.579	**Prob(JB):**	0.00
Kurtosis:	32.831	**Cond. No.**	2.41e+08

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 2.41e+08. This might indicate that there are
strong multicollinearity or other numerical problems.

Machine Learning Model - Data columns

```
Index([                    ('Petrol', 'Econ'),
                            ('Petrol', 'Mid'),
                            ('Petrol', 'Lux'),
                       ('Adv Petrol', 'Econ'),
                        ('Adv Petrol', 'Mid'),
                        ('Adv Petrol', 'Lux'),
                           ('Diesel', 'Econ'),
                            ('Diesel', 'Mid'),
                            ('Diesel', 'Lux'),
                       ('Adv Diesel', 'Econ'),
                        ('Adv Diesel', 'Mid'),
                        ('Adv Diesel', 'Lux'),
                              ('CNG', 'Econ'),
                               ('CNG', 'Mid'),
                               ('CNG', 'Lux'),
                           ('Hybrid', 'Econ'),
                            ('Hybrid', 'Mid'),
                            ('Hybrid', 'Lux'),
                         ('Electric', 'Econ'),
                          ('Electric', 'Mid'),
                          ('Electric', 'Lux'),
                            ('Bikes', 'Econ'),
                             ('Bikes', 'Lux'),
               ('Electric Bikes', 'Adv Econ'),
                ('Electric Bikes', 'Adv Lux'),
 'Fossil fuels (% equivalent primary energy)',
                                'Transport_x',
                                'Agriculture',
                  'Land-use change and forestry',
                                      'Waste',
                                   'Industry',
             'Manufacturing and construction',
                                'Transport_y',
                       'Electricity and heat',
                                  'Buildings',
                          'Fugitive emissions',
                         'Other fuel combustion',
                         'Aviation and shipping',
   'Renewables (% equivalent primary energy)',
              'Fossil fuels (% electricity)',
                'Renewables (% electricity)',
                              'fossil_demand',
                                  'ev_demand',
```

- To calculate CO2 emissions saved by electric vehicles year by year
- Features given in the screenshot are taken into consideration

Machine Learning Model Output

```
[121] X = df_ev_ghg_merged[[
          'Fossil fuels (% equivalent primary energy)',
      'Transport_x',
      'Agriculture',
      'Land-use change and forestry',
      'Waste', 'Industry', 'Manufacturing and construction', 'Transport_y',
      'Electricity and heat',
      'Buildings', 'Fugitive emissions', 'Other fuel combustion', 'Aviation and shipping', 'Renewables (% equivalent primary energy)', 'Fossil fuels (% electricity)',
      'Renewables (% electricity)']]

[122] X

      y = df_ev_ghg_merged[['ev_savings']]

[124] lm = linear_model.LinearRegression()
      model = lm.fit(X,y)

      y_predictions = lm.predict(X)

[126] lm.score(X,y)

      0.4954363469172296
```

Machine Learning Model Output

```
[116] lm = linear_model.LinearRegression()
      model = lm.fit(X_train,y_train)

[117] y_predictions = lm.predict(X_test)

[118] from sklearn.metrics import r2_score

[119]  r2_score(y_test, y_predictions)

      0.727704844938476
```

- Coefficient of determination(r square) = 0.72 or 70%
- This indicates that 72% variability is explained by our model.

Machine Learning Data Split

- Split the original dataset into training and test split using scikit-learn and train_test_split module
- We train the model on train split then we test the model on X values of test split

Split the Data into Training and Testing

```python
# Create our features
X = df_ev_ghg_merged[[
'Transport_x',
'Agriculture',
'Land-use change and forestry',
'Waste', 'Industry', 'Manufacturing and construction', 'Transport_y',
'Electricity and heat',
'Buildings', 'Fugitive emissions', 'Other fuel combustion', 'Aviation and shipping', 'Renewables (% equivalent primary energy)',
'Renewables (% electricity)']]

# Create our target
y = df_ev_ghg_merged[['ev_savings']]
```

```python
[115] from sklearn.model_selection import train_test_split
      X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1)
```

```python
[116] lm = linear_model.LinearRegression()
      model = lm.fit(X_train,y_train)
```

```python
[117] y_predictions = lm.predict(X_test)
```

```python
[118] from sklearn.metrics import r2_score
```

```python
[119] r2_score(y_test, y_predictions)

      0.727704844938476
```

Machine Learning Model Choice

Linear Regression Model

Limitations:

- Accuracy was initially low when including electric bike data
- For some of the features, standard error was really high, which required dropping columns

Benefits:

- Accuracy increased by 20%
- R-squared variance also increased by 20%

Dashboard Blueprint

Tools:

- PostgresSQL
- Scikit-Learn
- Jupyter Notebook
- Python
    - Matplotlib & pandas

Findings & Recommendations

- Electricity and transport are the highest drivers of carbon/GHG emissions
- Selecting a country/region to determine patterns of increased usage of alternate energy
- Cleaner energy sources to aid EV charging capabilities

Interactive Elements:

- Graphical visualization of our findings and key takeaways.

Next Steps...

- Include list of technologies, languages, tools, and algorithm used throughout the project
- Update README.md
- Describe how the model has been trained and the current accuracy score
- Finalize the dashboard with images, ML data, and interactive element(s)