Climate Change

Group 7

Ayano Yamamoto

Summary

1. Extraction of Information
2. Integration
3. Cleaning
4. Visualisation
5. Storage Structure

1. Extraction of Information

Data we sourced

API

- World Bank: Renewable energy consumption (% of total final energy consumption)
- Twitter: 100 most recent tweets mentioning "COP26"
- Twitter: 100 most recent tweets mentioning "electric cars"

HTML

- Wikipedia: List of countries by carbon dioxide emissions
- Wikipedia: List of countries by renewable electricity production

CSV

- Kaggle: Average Temperature per country per year

Challenges sourcing the datasets

Variation of list of countries in each datasets

- Datasets contain different number of countries
- Variation of names for one country (e.g. “Cabo Verde” and “Cape Verde”)

Our solution

- Assigned ISO 3166-1 alpha-3 codes to each dataset for standardisation
- Created a function `fuzzyalpha3` to fuzzy-search country names in `pycountry` and return alpha-3 codes
- Created a function `getalpha3` to search exact country names in `pycountry` and return alpha-3 codes
- Manually matched alpha-3 codes that could not be matched using functions

Challenges sourcing the datasets

Limitations with Twitter standard search API

- Maximum of 100 tweets can be returned
- Search index has a 7-day limit. In other words, no tweets will be found for a date older than one week.

Our solution

- Used `count = 100`, `result_type = 'recent'` to search for 100 most recent tweets including mentions of 'COP26' and ‘electric cars’
- Not an accurate statistical representation of world’s population, but demonstrated our ability to work with Twitter API

Challenges sourcing the datasets

Lack of "geo-tagged" Tweets

- Twitter allows users to tweet with a specific latitude/longitude “Point” coordinate, or a Twitter “Place”.
- Most tweets are not tagged with either of these location information

Our solution

- Used user profile locations (not all data are real locations, some are null)
- Created a function `getcoordinates` that uses Geopandas `geocode` to get geolocations. Also created a function `getaddress` so we can then use Geopandas `reverse_geocode` to get standardised format addresses from geolocations including alpha 3 country codes.
- Benefit: ability to search with different languages

Challenges sourcing the datasets

Availability of change in temperature dataset

- Data containing change in temperature per country was hard to source

Our solution

- Used a dataset containing average temperature per country per year from 2000 and 2013
- Subtracted 2000 data from 2013 data, and stored the difference in a new column

Challenges sourcing the datasets

Rows that require aggregation / separation

- `tempdifference` had rows for country names Baker Island, Kingman Reef, and Palmyra Atoll. They are all part of United States Minor Outlying Islands which share the same alpha-3 code UMI.
- `co2emission` had one row for country name Serbia & Montenegro, and two values representing CO2 emission.

Our solution

- `tempdifference`: Calculated the mean value of the three countries and created a new row for United States Minor Outlying Islands
- `co2emission`: Created two separate rows for Serbia and Montenegro

2. Integration

Merging of data

- Created a dataframe of ISO 3166-1 alpha-3 codes and country names from `pycountry` as a starting dataframe
- Assigned alpha-3 codes to each dataset prior to merging
- Used left-merge to merge each dataset to the starting dataframe using alpha-3 codes

	alpha3	country	renewable_consumption	count_cop26	count_electric_cars	co2emission_incl_LUCF	renewable_production	temp_difference
0	ABW	Aruba	8.024100	NaN	NaN	NaN	148.5	0.41
1	AFG	Afghanistan	21.422701	NaN	NaN	7.59	1071.0	1.04
2	AGO	Angola	56.785500	NaN	NaN	62.93	7282.0	0.15
3	AIA	Anguilla	NaN	NaN	NaN	NaN	2.4	0.31
4	ALA	Åland Islands	NaN	NaN	NaN	NaN	NaN	-1.05

3. Data Inspection and Cleaning

Data Inspection

- DataFrame shape: 256 rows and 8 columns
- DataFrame columns: `alpha3`, `country`, `renewable_share`, `count_cop26`, `count_electric_cars`, `co2emission_incl_LUCF`, `temp_difference`, `renewable_production`
- Data types
  - `alpha3` and `country` columns as object
  - `count_cop26` and `count_electric_cars` as float64 - changed to int64
  - `co2emission_incl_LUCF` changed from object to float64 as the column contains decimal numbers
  - `renewable_consumption`, `renewable_production`, and `temp_difference` as float64
- Missing values
  - `renewable_consumption` - 37, `co2emission_incl_LUCF` - 65, `renewable_production` - 50, `temp_difference` - 20
  - `count_cop26` and `count_electric_cars` - pulling live data from 100 most recent tweets. Due to the account limitation, there is a high amount of NaN's
- 0 duplicates - pre or post data cleaning

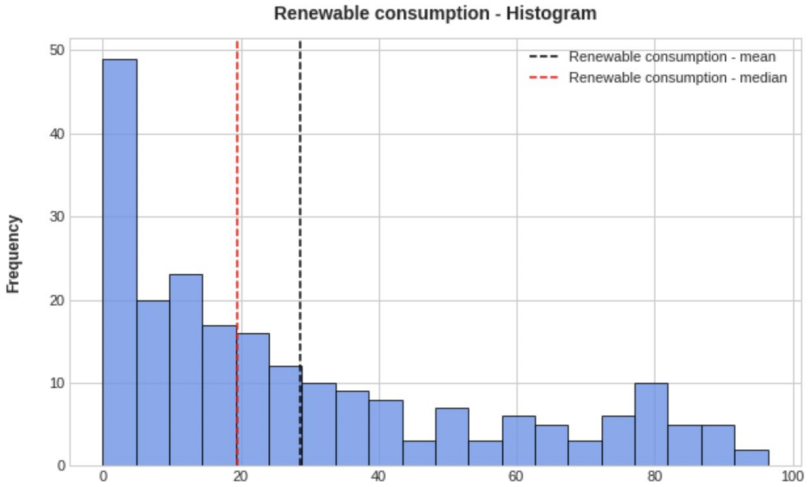
Numerical Columns Inspection

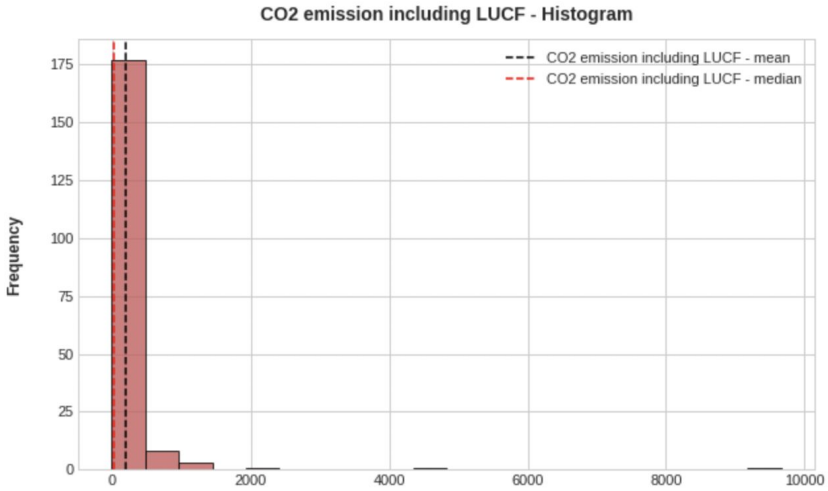
Continuous features

- `renewable_consumption`
  - Min = 0
  - Max significantly higher than Mean and Median - presence of outliers or skewed distribution
  - Mean is not representative
- `co2emission_incl_LUCF`
  - Min is a negative value - carbon negative countries
  - Max is higher than Mean and Median due to extreme outliers - countries with very high CO2 emission such as China
  - Mean is not representative
- `temp_difference`
  - Min is a negative value as we are calculating temperature differences between 2000 and 2013 for each country
  - Mean and Median have close values => close to normal distribution
  - Mean is representative
- renewable_production
  - Max is higher than Mean and Median, presence of outliers or skewed distribution
  - Mean is not representative

Continuous Features Visualisation

The data distribution for `renewable_consumption` is skewed to the right (it has a positive skewness).

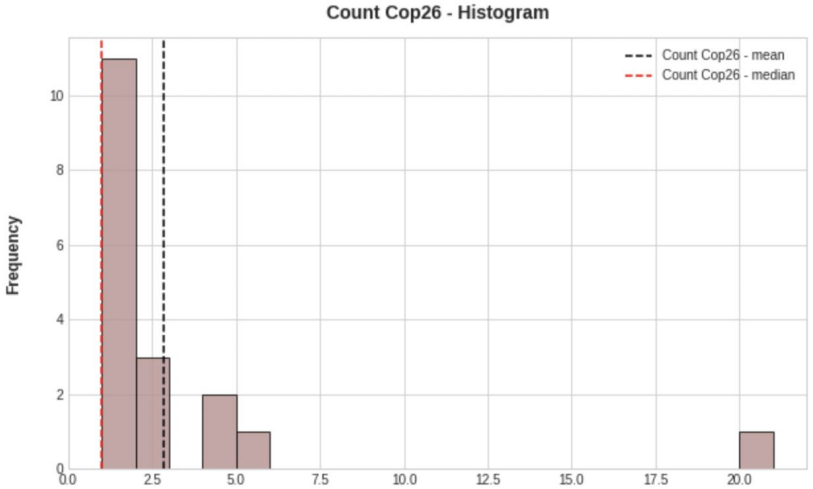


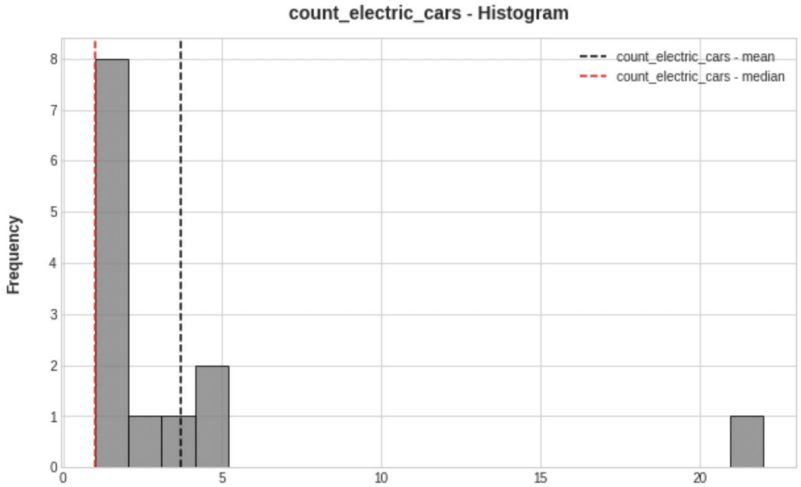


Data distribution is not recognisable for `co2_emission_incl_LUCF` due to the presence of extreme outliers (China - very high CO2 emissions).

Continuous Features Visualisation

Data distribution for `count_cop26` is undefined due to the high amount of missing values. It will change every time the code is run.

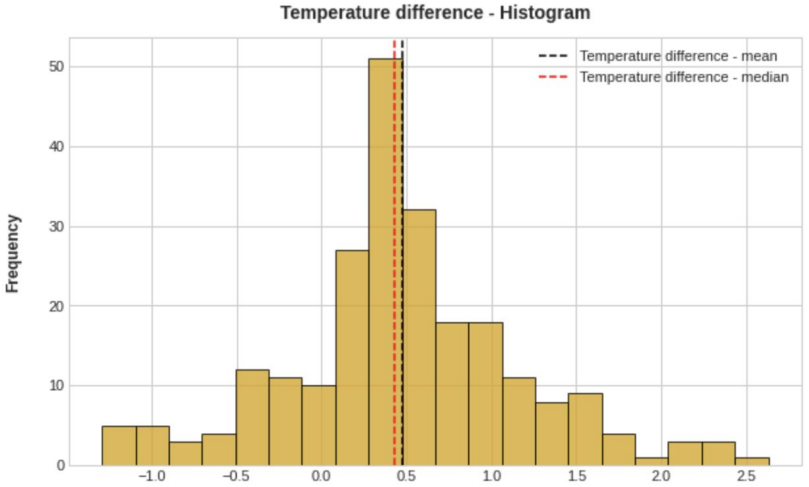


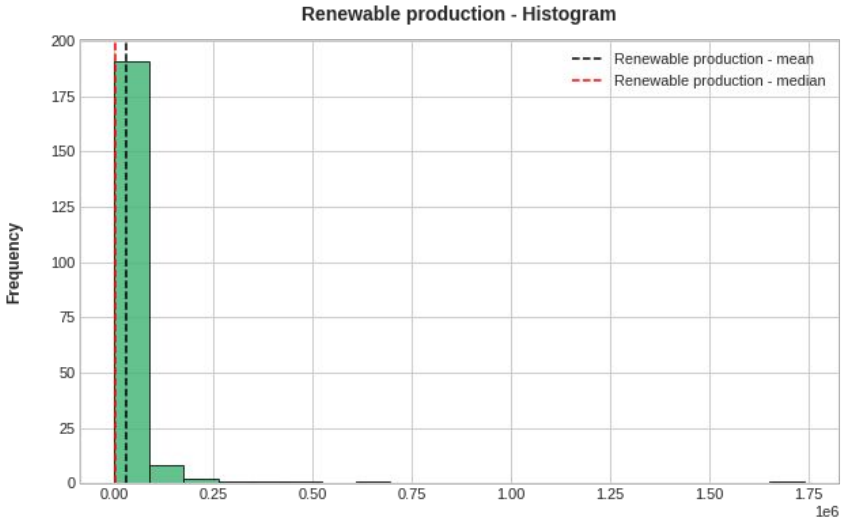


The same for `count_electric_cars`, the data distribution is undefined due to the high amount of missing values. It will change every time the code is run.

Continuous Features Visualisation

The data distribution for `temp_difference` has a normal distribution with a high peak.





Data distribution is not recognisable for `renewable_production` due to the presence of extreme outliers (China - very high production of renewable electricity).

Data Cleaning

Replacing NaN

- `temp_difference`: replaced with mean since mean is representative
- `renewable_consumption`, `co2emission_incl_LUCF`, `renewable_production`: replaced with median since mean is not representative
- `count_cop26`, `count_electric_cars`: replaced with the global constant 0 and changed the data type to int64

Data normalisation

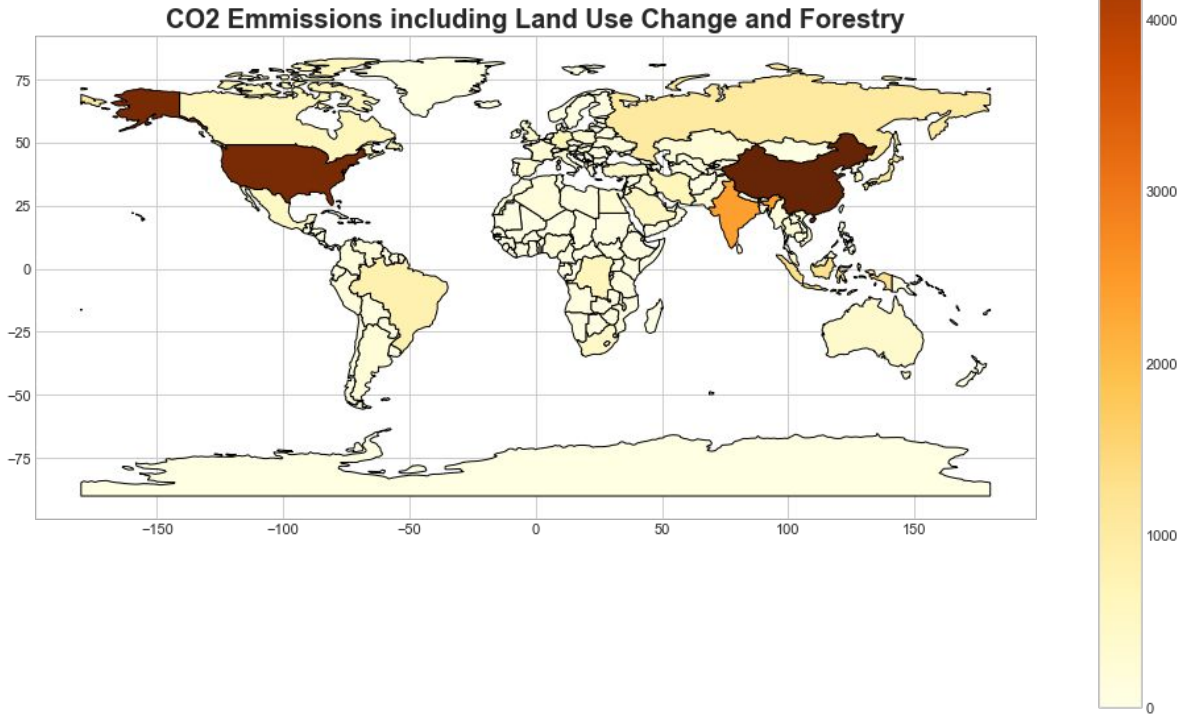
- Used `MinMaxScaler` from scikit-learn

4. Visualisation

Visualisations

- Comparing countries with the highest CO2 emissions
- Locations of most recent tweets including ‘COP26’
- Countries with the highest temperature difference
- Countries with the highest amount of renewable electricity production
- Locations of most recent tweets including ‘electric cars’
- Countries with the highest renewable energy consumption

Countries with Highest CO2 Emission in 2018

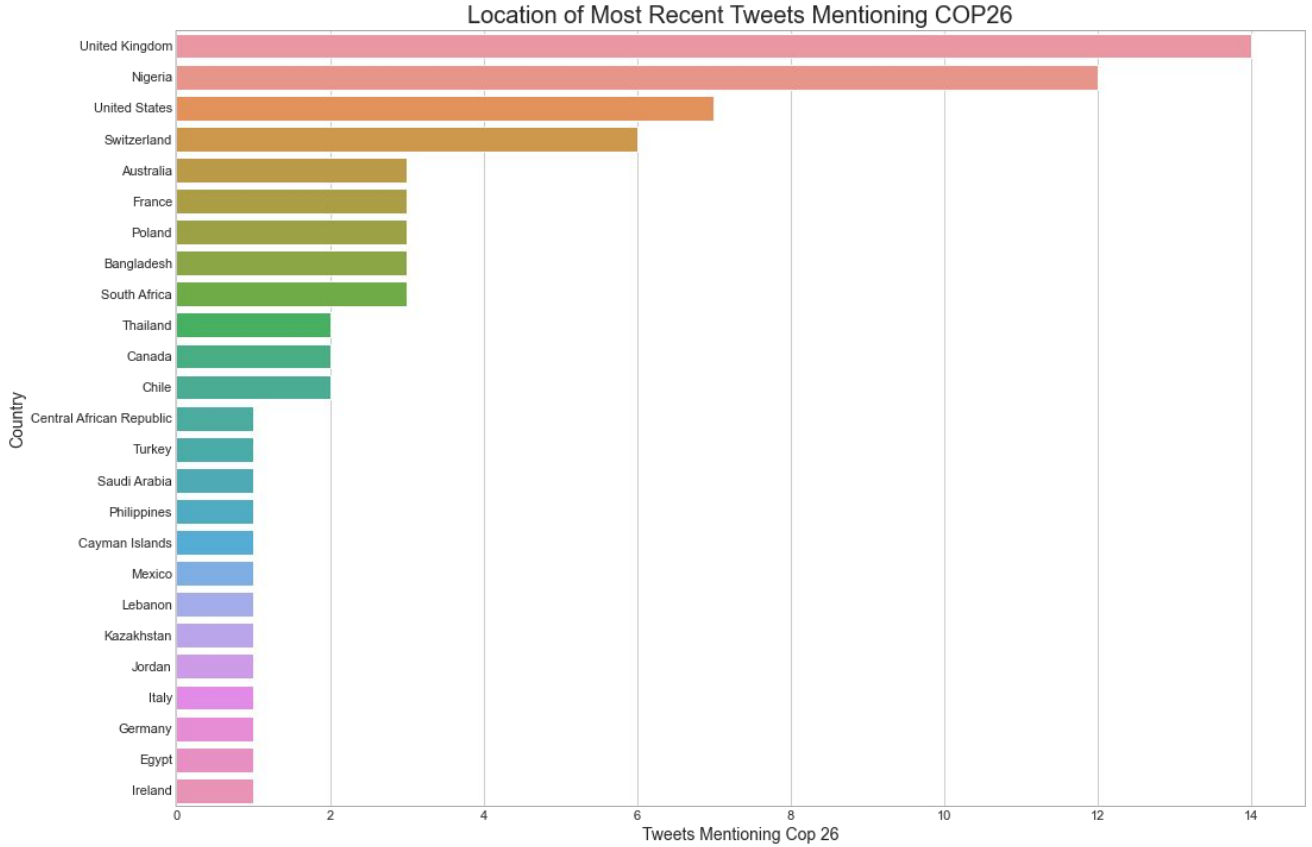


Countries with Highest CO2 Emission in 2018

Insights:

- Pie chart (left): 10 countries with the highest CO2 emission in 2018, and their % share among the 10.
- Map (right): All countries visualised based on their CO2 emission.
- China had the highest CO2 emission in 2018 of 9,663.36 metric tonnes, followed by United States (4,749.57), India (2,400.25), and Indonesia (1,269.55).
- China's value of nealy 10,000 metric tonnes dwarfs other countries' visualisations. To resolve this, `vmax` was set to `5000`.

Location of most Recent Tweets featuring 'COP26'

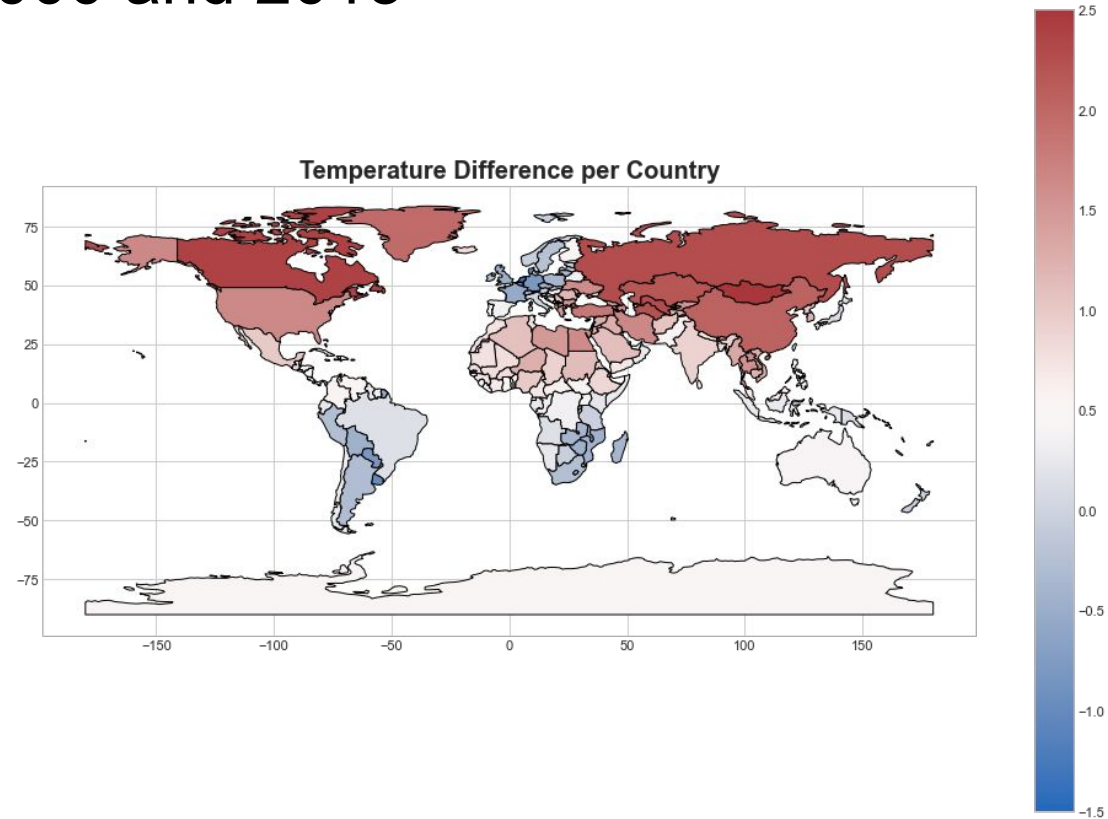


Location of most Recent Tweets mentioning 'COP26'

Insights:

- Countries displayed in user profile locations from 100 most recent tweets mentioning the keyword ‘COP26’.
- ‘COP26’ is a commonly used name for the 2021 United Nations Climate Change Conference held in October - November 2021.
- Result changes every time the search is performed.
- At the time of visualisation, United Kingdom had the highest number of tweets matching the search with 12 tweets, followed by United States (9), India (7), and Australia (5).
- This search does not consider the context of the tweets.

Countries with the Highest Increase in Temperatures between 2000 and 2013

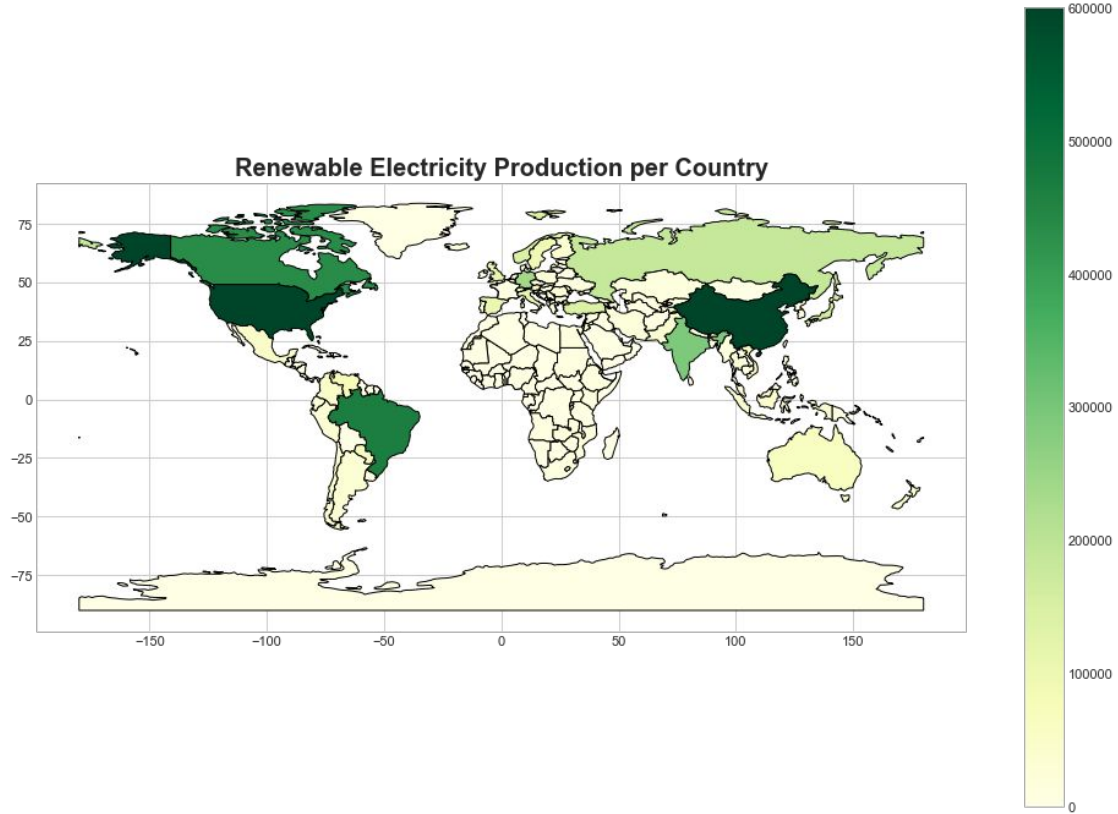


Countries with the Highest Increase in Temperatures between 2000 and 2013

Insights:

- All countries visualised based on their temperature difference between 2000 and 2013.
- Mongolia had the highest increase in temperature with 2.63 degrees celsius, followed by Canada (2.39), Russia (2.28), and Uzbekistan (2.25).
- In general, countries with higher latitudes are experiencing larger increase in temperature.

Countries with the Highest Amount of Renewable Electricity Production in 2016

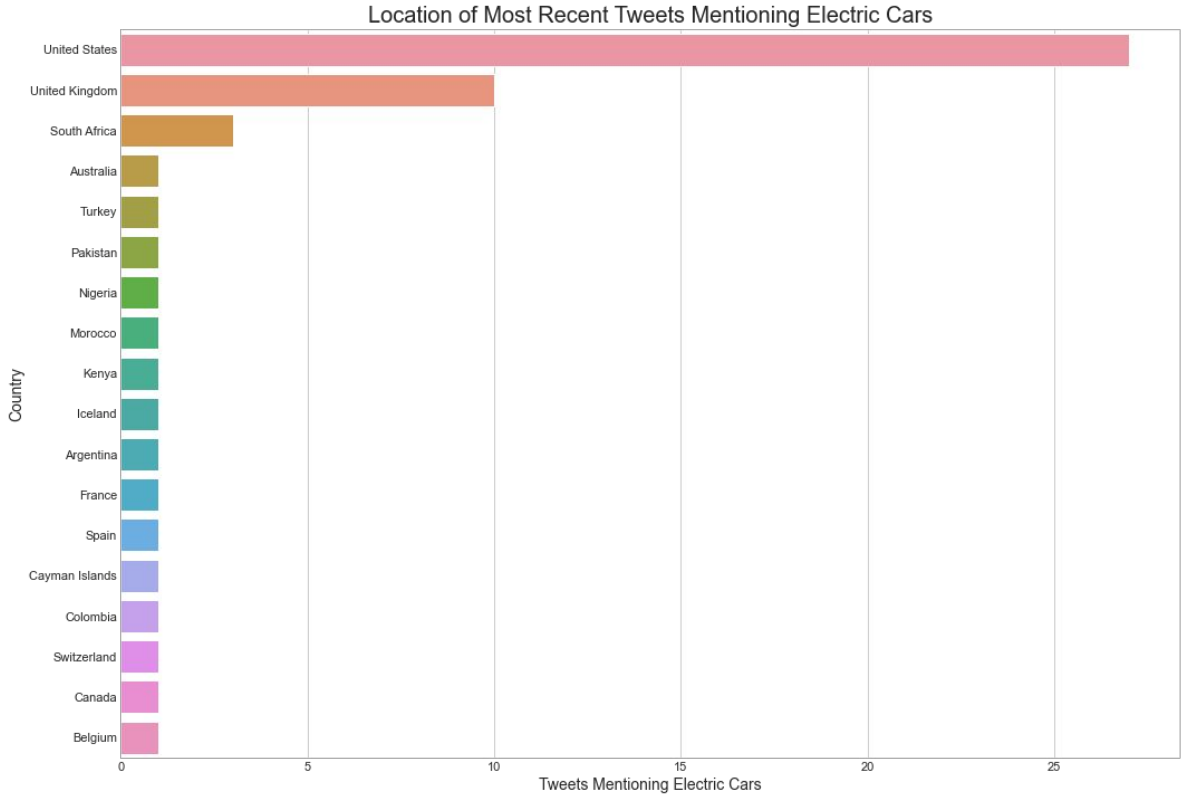


Countries with the Highest Amount of Renewable Electricity Production in 2016

Insights:

- All countries visualised based on their renewable electricity production in 2016.
- Data includes hydropower, wind power, biomass, solar power, and geothermal electricity productions, and measured in gigawatt hours.
- China had the highest amount of renewable electricity production with 1,739,400 gigawatt hours, followed by United States (637,076 Gwh), Brazil (465,579 Gwh), and Canada (433,597 Gwh).
- China's data was another high outlier here. To resolve this, `vmax` was set to `600,000`.

100 most recent tweets mentioning "electric cars" from Twitter

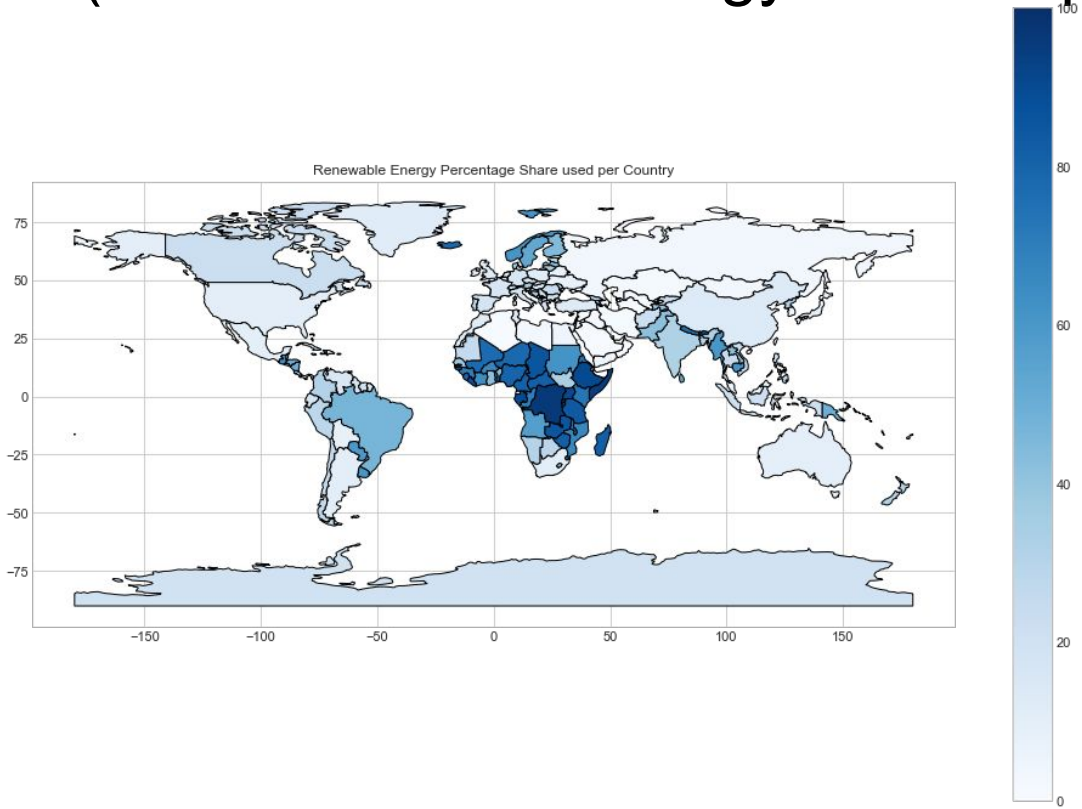


100 most recent tweets mentioning "electric cars" from Twitter

Insights:

- Countries displayed in user profile locations from 100 most recent tweets mentioning the keyword ‘electric cars.’
- Result changes every time the search is performed.
- At the time of visualisation, United States had the highest number of tweets matching the search with 28 tweets, followed by United Kingdom (17), Brazil (2), and Australia (2).
- This search does not consider the context of the tweets. For example, some of these tweets may be about people discussing usage of electric vehicles, others may be tweets about Elon Musk.

Countries with the Highest Renewable Energy Consumption (% of Total Final Energy Consumption) in 2018



Countries with the Highest Renewable Energy Consumption (% of Total Final Energy Consumption) in 2018

Insights:

- All countries visualised based on their renewable electricity consumption as percentage of their total final energy consumption in 2018.
- The Democratic Republic of Congo had the highest percentage of renewable electricity consumption with 96.38% of total final energy consumption, followed by Somalia (94.88%), Uganda (90.33%), and Ethiopia (89.92%).
- In general, countries with higher potential to generate renewable energy with lower total final energy consumption ranked high.

Challenges Creating Visualisations

Mapping

- Plotting data on to a map

Our solution

- Use basemap dataset from `Geopandas` to allow multiple layers be plotted
- Identify what countries from the basemap are present in our dataframe using alpha 3 codes
- Merge dataframes
- Plot data on to basemap
- Adjusted `vmax` to account for outliers

5. Storage

Storing of data

Performed 5 different types of data storage

- Used `to_csv` to write a dataframe to a comma-separated file (Section 3.4)
- Used `to_excel` to write a dataframe to an excel file
- Used `to_json` to export a dataframe to a JSON file
- Used `sqlite3` to save a dataframe to a relational database
- Used `pymongo` (`MongoClient`) and `certifi` to save the dataframe to a non-relational database

Storing of data

Performed retrieval of data from MongoDB

- Read data from MongoDB into a dataframe
- Performed queries from MongoDB
  - Countries with renewable energy consumption less than 0.1% share of their total consumption
  - Countries with CO2 emission greater than or equals to 1,000 metric tonnes
  - Countries with temperature difference rise of 0.39 degrees celsius between 2000 and 2013
  - Country name ‘Ireland’
  - Country names that start with letter ’A’
- Performed CRUD operations

Thank you