IRON
HACK

Data Analytics

Climate change through the eyes of data

Ferdinand Leube

June, 2022

Table of content

Introduction

Today in 2022 nearly everybody has realized that climate change is a serious problem the world will have to tackle in the near future. Every single one of us has personally experienced temperature increase over the last few summers and denying this development is not an option any longer. Having learnt the basic tools of data analysis I now want to do my own approach on analyzing data about temperature increase around the world, see if the warnings about a global temperature increase are reflected in data and as a second step look into factors which could be responsible for rising temperatures in certain countries where I have detected increasing temperatures.

My goal is to create a database that includes information about the average temperatures and average temperature increase in different countries over a certain period of time, information about different economic sectors as well as their contribution to climate change/temperature increase, and data about the number of power plants with different fuel types in different countries. With this database I want to be able to make the analysis I want to gather the insights I am looking for regarding correlation between the different factors displayed and temperature increase.

To better understand the topic of climate change I am going to start by researching which factors can have an impact on our climate and could be responsible for temperature increase so I can specify my search for data sets on those factors. After retrieving the necessary information and raw data I need for my database, I will try to receive an overview of the data I have imported, perform exploratory data analysis and then clean it regarding the result of my observations. The next challenge will be setting the foundation of my database by deciding on how to structure my database and creating an entity-relationship model. With this foundation in mind I will continue by creating my database, setting up the different tables for my entities and then importing my cleaned data into the database. At this stage I will be able to start writing queries in order to gain the insights I am hoping to attain regarding the correlation between the decided on factors and temperature increase.

Data and data sources

As for the first foundation of my data I had to find an available set of data which observed the temperature at different points around the world over a certain period of time. The dataset I have been using in this project displays daily reports of the temperature at over 150 sites in the USA and 167 other international sites. It has been created by the University of Dayton - Environmental Protection Agency and the dataset has been taken from the Average Daily Temperature Archive. The source of data collection the university has been using to create this data is the American National Climatic Data Center.

Before diving further into the data search for my project I had to decide which factors to specify on. For the first factor I wanted to check the effect of various economic sectors on climate change. As stated by the US Environmental Protection Agency the biggest reason for increasing temperatures is the increasing emission of Greenhouse gasses. Therefore I decided to base my indication of how related a certain sector is to temperature increase/climate change on the greenhouse gas emission that sector produces in a certain country.

The dataset I used for preparing data in order to analyze this factor presents internationally comparable emissions data on greenhouse gasses from 1850 to 2018 collected from multiple different data sources. It furthermore includes information about 6 different economic sectors and subsectors and classifies the specific greenhouse gasses released by this economic sector. All of the numeric data is given in Metric tons of carbon dioxide equivalent or MTCO2e. The first data source that has been used for the creation of this data set is Climate Watch. It was created in 2018 by a Washington DC based institution named: World Resources Institute. Secondly the data set I have used is based on The PRIMAP-hist national historical emissions time series (1850-2014). V. 1.1, collected and created by Gtschow Johannes, Jeffery Louise, Gieseke Robert and Gebel Ronja in 2017 as part of the GFZ Data Services institution. Lastly this data set retrieved part of its data from UNFCCC data: Greenhouse Gas Inventory Data - Detailed data by Party created by UNFCCC in 2017.

The third data set presents a long list of power plants, the country it is located in, exact location portrayed by longitude and longitude, primary fuel usage and more information describing the power plants characteristics. I have picked this data set because I wanted to see if there is a correlation between the amount of power plants a country has using a certain fuel type and temperature increase in that country. Different fuel types are made more responsible for greenhouse gas emissions than others and I wanted to analyze if there is indeed a connection.

Data collection

After defining my business case I have started my search for free API's or open data sets relating to the topics discussed in my business case. The data sets described in the section before have been republished on Keggle and I could download them on my local machine to then import them to my jupyter notebook using the pandas read csv function.

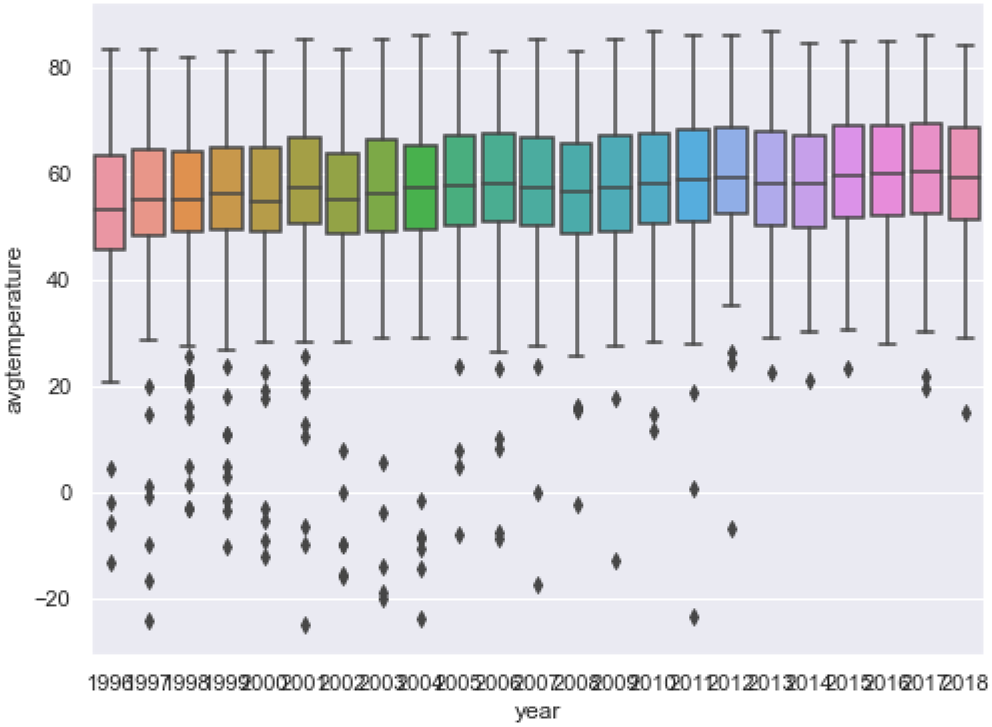
Data cleaning and Exploratory data analysis

Before starting my work I imported all necessary libraries I would have to work with while cleaning the data such as pandas, nump, random, seaborn for later data visualizations etc. Regarding the workflow of my data cleaning, I have performed this process with all 4 datasets in a very similar way repeatedly. The first step was always to develop an overview of the structure of the imported data by reading through the .info() function result, checking the shape of the dataframe, looking at different types of data the columns contained and observing the column names. The second repeated step was to apply my cleaning columns function which turns all the column names into lower case, strips spaces from front and end, replaces spaces with underscores and renames the columns with the “cleaned” version.

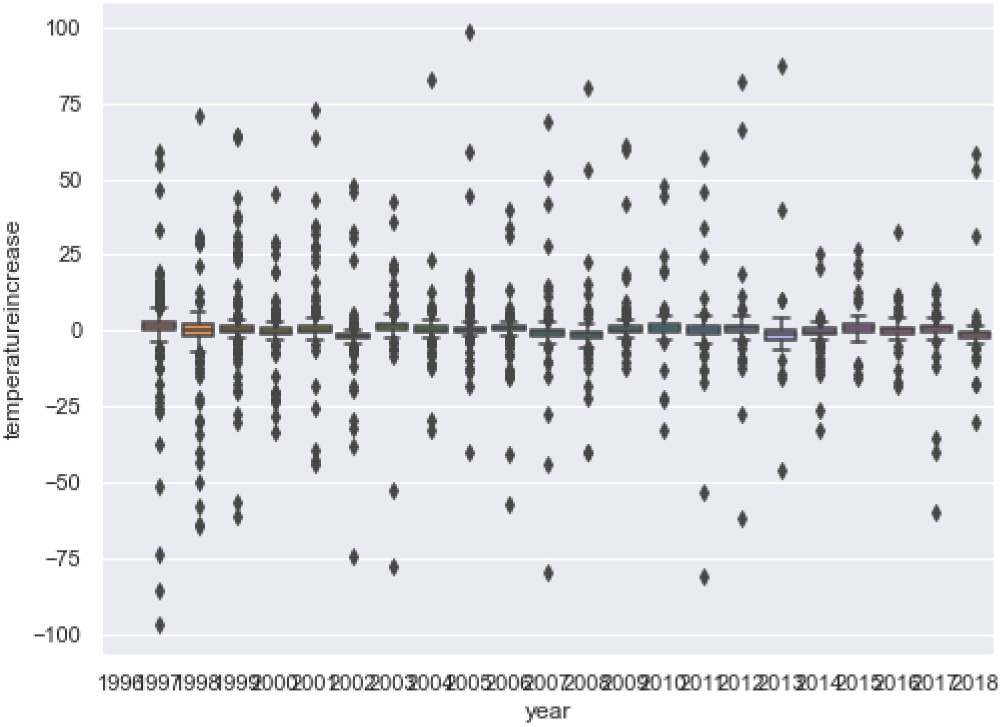
Daily recorded temperature in over 300 cities over the last 20e years data

I continued by checking the first data set for undefined values by calculating the sum of all missing values for each column. Because only the ‘state’ column contained missing values this problem was quickly resolved by simply deleting this column because of its unnecessariness for further analysis. In order to format the table in the way that I needed the data for my analysis (average Temperature per Year for every position and temperature increase from year to year) I grouped by the columns country, city, and year and aggregated the average temperature for each year at a certain position. I then reset the index to have the country, city, and year columns back as regular columns, so I could work with their values. I then checked which countries had data up until 2018 in order to be able to compare these countries' data with the other data sets(containing data until 2018 ony) by applying a filter. After checking the unique values of the filtered country column I removed all countries with non-existent data for these years from the original data frame. The next step was making sure that there was a similar amount of available data for all years so they would be equally represented in later calculations and data visualizations. A higher number of countries did not have data for the years before 1996. Additionally I checked the unique values of the column years and found the two outliers 200 and 201, which I then dropped together with the years beneath 1996.

Next I started checking the statistical summary of the data using the pandas describe function and already recognized a big difference between the mean and the maxvalue for average temperatures. Similar results were observable after plotting the statistical summary of both those columns with the seaborn boxplot function.



As you can conclude from the boxplot all values that are below the -25 degree mark are with all certainty outliers which makes sense regarding the fact that these temperatures are given in degrees fahrenheit and everything beneath -25 degrees would be unrealistic. After making sure that these outliers make up only a small portion of the data I decided to drop these rows using a filter. I then added the temperature differences by applying the pandas difference function which calculates the difference between a value and the value before in a specified column. Because the years are repeated over and over again for each position I had to set all 1996 values equal to an undefined value because there should be no difference calculation for the first recorded year of every country city combination but a difference had been created by default by the pandas difference function. I continued checking the statistics of the temperature increase column by plotting the data using a boxplot.

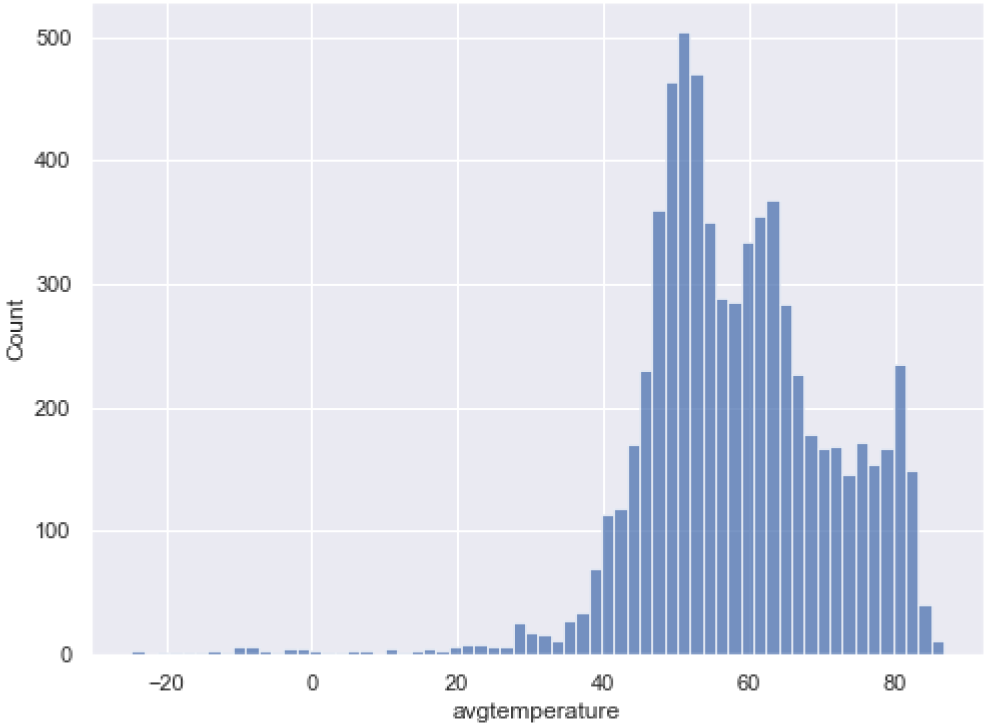


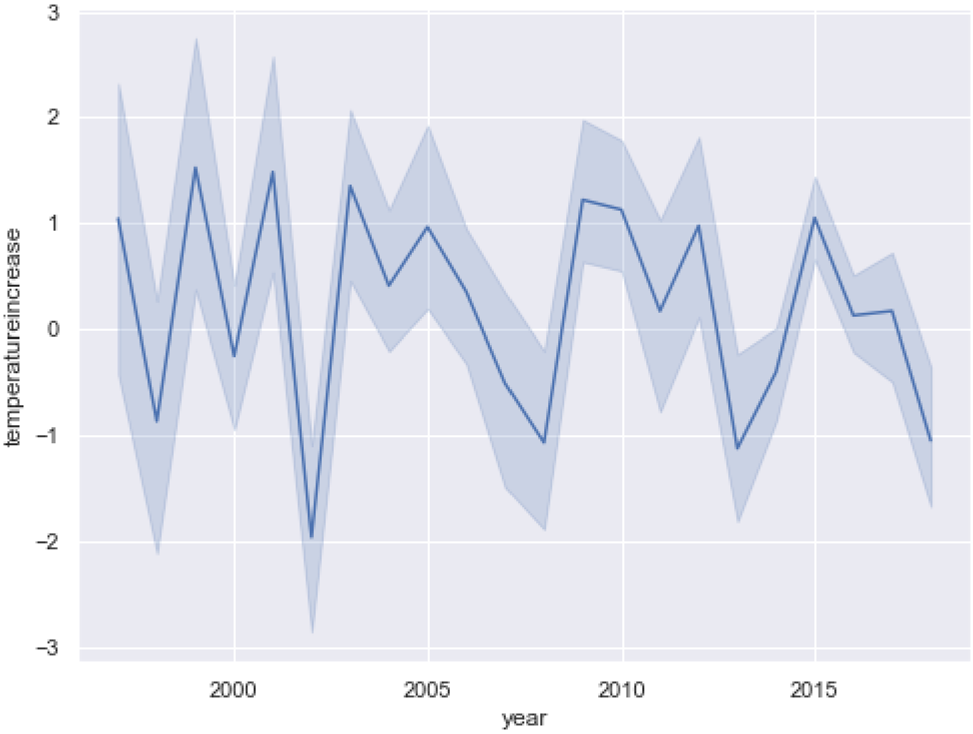
Once again you can see that also for the increasing temperature column there are many outliers but I decided to not delete the outliers in this case because when regarding minimum and maximum of the average temperatures these temperature increases or decreases are theoretically possible.

Further exploratory data analysis

The first figure shows the frequency distribution of the average temperature column with bins equal to twenty. You can see that the data has still a little left skewness but much less after we have eliminated the outliers. Normalizing the data even further would require deleting or replacing realistic measurements which I did not want to do.

The second figure is a line plot of the average temperature increase development over time. The data here looks tremendously different from the box plot about temperature increase above. Here you can clearly see that most of the data (25%-75% displayed through shadow around the average line) have only very slight temperature changes between +- 3 degrees. The given chart indicates no concrete development over the last couple of years, rather a decrease of temperature at the end of the observed time and therefore displays results very different than expected by the business case hypothesis.


500
400
300
200
100
0
Count
−20
0
20
40
60
80
avgtemperature

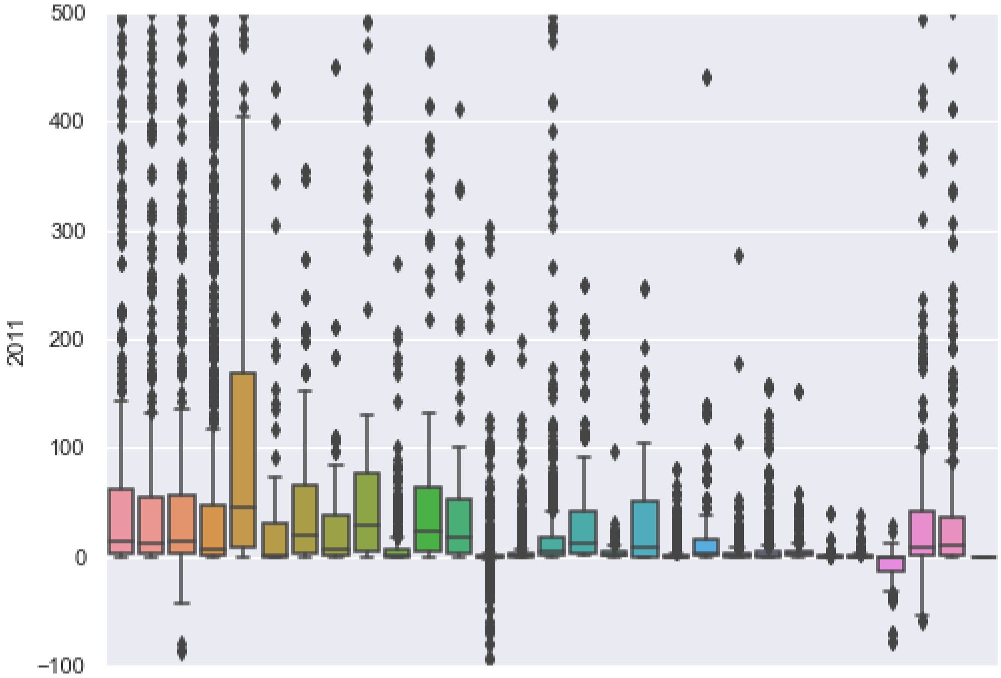

3
2
1
0
−1
−2
−3
temperatureincrease
2000
2005
2010
2015
year

Data about 12 different economic sectors and their greenhouse gas emissions in various countries

After completing the repetitive steps of initializing my data cleaning process mentioned above, I started cleaning the data to my personal needs, meaning I had to update it to only consist of the data that I could compare with the other datasets. I started by excluding all columns with information about years that are not covered in other datasets by looking for the index of the 1996 column and dropping all others following that index. Next I checked the unique values of the countries column. Both for this and the previous data set I turned all names into lowercase and checked their general structure to be sure that ‘Germany’ and ‘germany ‘ would not be seen as two different values. Following this analysis I created a list with all countries which are present in this but not in the previous data set and removed the information about these countries from my data set. I continued my data cleaning by checking the unique values of the other categorical data columns and found that all values followed the same format and there were no unwanted duplicates as e.g. the example given above.

Same as before, the following challenge was to look for and decide on how to deal with missing values. The sum of missing values was quite high and percentage wise made up a big part of the column so I could not simply replace them by random. I started by selecting only columns with numerical data and then transposed the data frame in order to calculate the average, standard deviation etc. for every country - economic sector combination column wise. I then wrote a function which takes a column series as an input, drops all the missing values, calculates the mean and standard deviation of the top five values (if five values were existent) and returns an array of the size of all missing values with random values in between +- one time the standard deviation from the mean of the most recent values to realistically update the rows (this area representing 68% of the values). I applied this step to base the missing values on the most recent values present in the dataset. Before applying the function and replacing all missing values with a random choice output of the arrays created above by looping through the columns, I had to get rid of all columns consisting only of missing values. The function works only when at least one numerical value is given and columns that don’t contain any data are not valid for my analysis. After checking for leftover missing values I concatenated the numeric data frame back with data frame consisting only of categorical values and transposed it back to the original format.

I started my exploratory analysis by creating boxplots for all the different years showing the statistical summary of all values for each sector. Exemplary boxplot is given below for the year 2010:

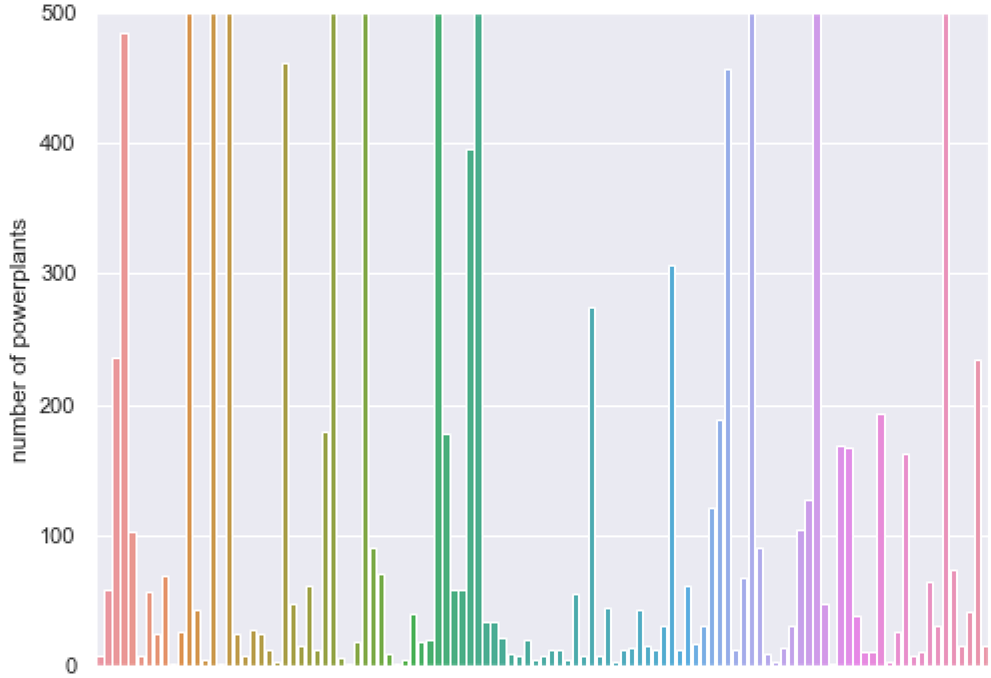


After analyzing all the different boxplots I decided to define all values below 0(displaying emissions here, values beneath 0 have no logical sense, you cannot produce negative emissions) as outliers and drop the data with such values. I then continued my analysis with the statistical outliers that are shown in the box plot and checked the percentage of data seen as a statistical outlier for every column. Because the amount of data that is defined as an statistical outlier makes up a big number of the entire data set, therefore simply dropping those columns is not an option. Furthermore, just because a data value is defined as a statistical outlier, does not mean it is unvalid data. For this dataset for example certain sectors in certain countries can have much higher emissions than that sector in a different country in a specific year. I therefore decided not to exclude those values from my data set.

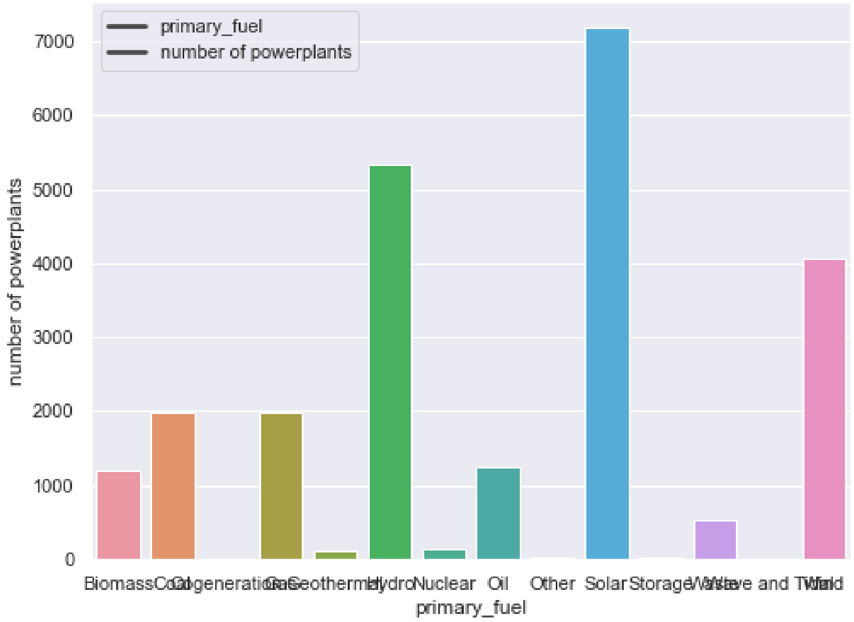
Data about power plants, their location capacity and primary fuel usage

I began with the same procedure used for the two previous data sets and cleaned the data to my needs by excluding all data regarding countries that are not in my first data set. While looking at the sum of missing values for the different columns I observed that only columns such as secondary fuel etc. had missing values. As stated in the data description above, this raw data set had a lot of data I would not need for my later analysis so I dropped all columns beside the power plants name, country, country code, capacity, position in longitude longitude and, the most important, primary fuel. I continued with checking the unique values for each column and checked for any duplicates of power plants names to remove duplicate information about the same power plant in case of existence.

I began with my exploratory analysis by checking the frequency distribution of the amount of power plants for each country.

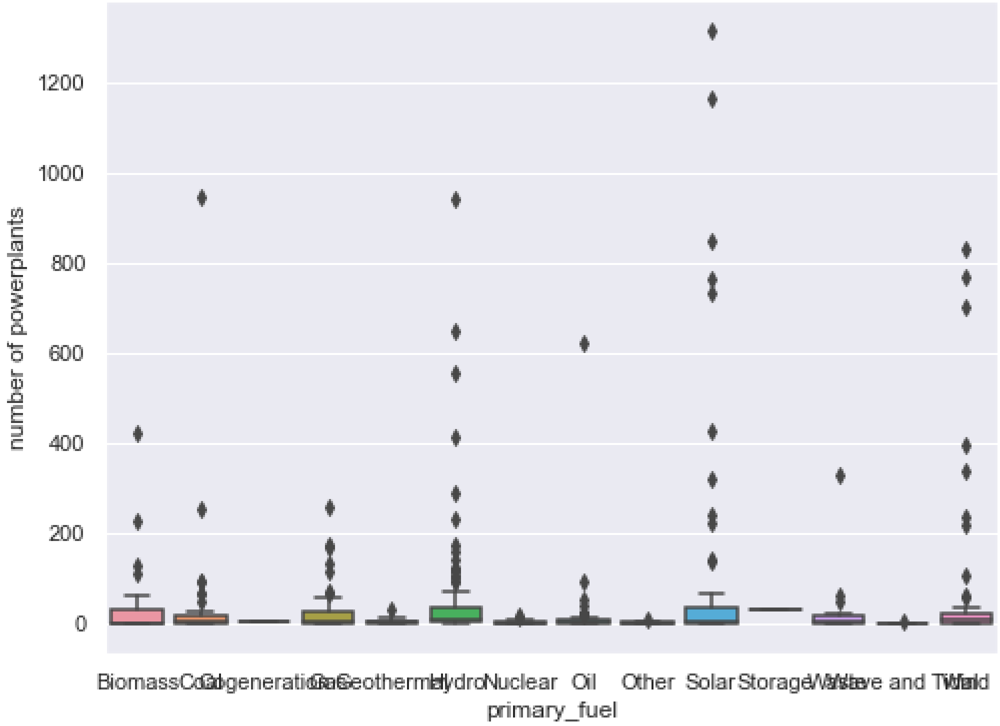


Clearly there is no equal distribution of power plants which can be logically explained by the different sizes of countries and their economic capacity/infrastructure. Next I wanted to check the frequency distribution of power plants per primary fuel.



From the power plants in my data set, the most frequent primary fuel used is solar, followed by hydro energy.

In the last chart I wanted to check the statistical summary of the frequency distribution for primary fuels using a box plot. Even though the box plot displays a few outliers I decided not to remove those from the dataset because it is indeed possible that bigger countries have a high number of power plants with the primary fuel being solar energy.

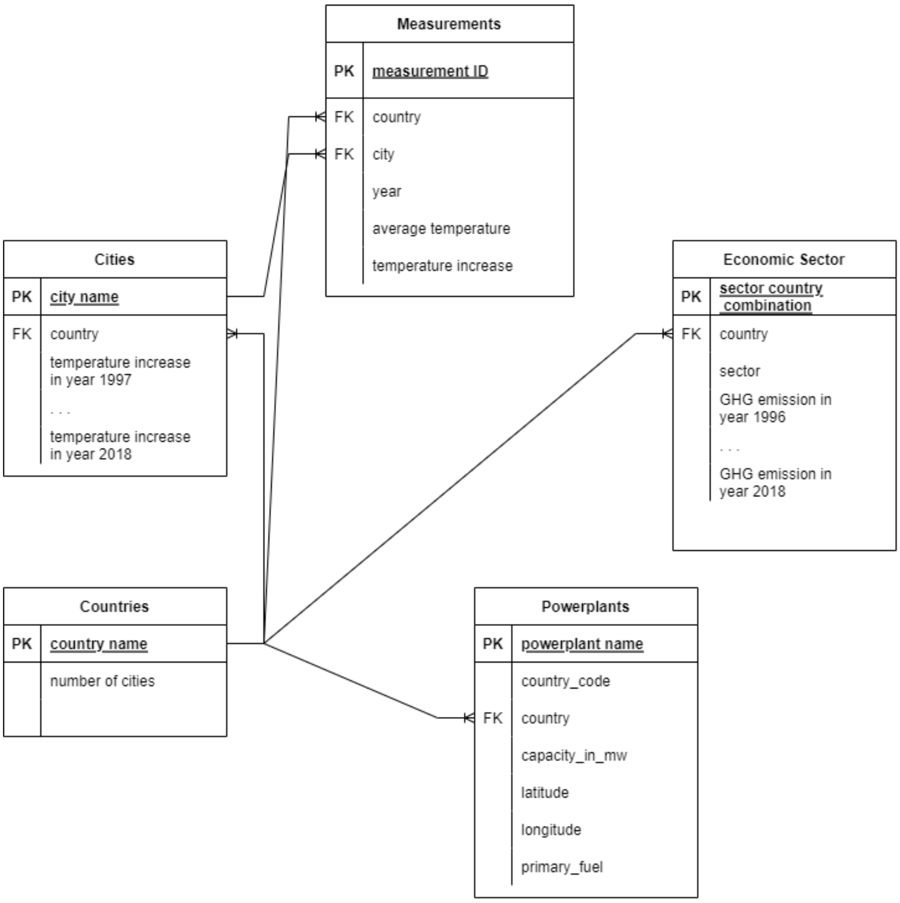


Deciding on what type of database to use

SQL	NoSQL
can work with smaller amounts of data	can work with big amount of data
supports transactions on cell level	does not support atomic transactions
slower querying, but faster updating	slower update, faster querying
Structured data	unstructured or semi structured data
OOP unfriendly (object oriented programming)	OOP friendly

Following my plan, the next step is creating my database. I decided to use SQL because this type of database allows me to use my cleaned and converted dataframes(structured data) and perform queries on a small level. Furthermore I can link the different tables with each other using primary and foreign keys which allows me to quickly make queries and join different information from various tables. In my situation I also do not care about how friendly the system is towards object oriented programming and writing longer queries is acceptable. The most important factor is to be able to work with the output of my data cleaning process: structured data. The relational database I will use is MySQL. In order to prepare the foundation for the creation of my database I created the following entity relationship model to clarify the different entity tables I will use and specify their relationships.

Entities. ERD

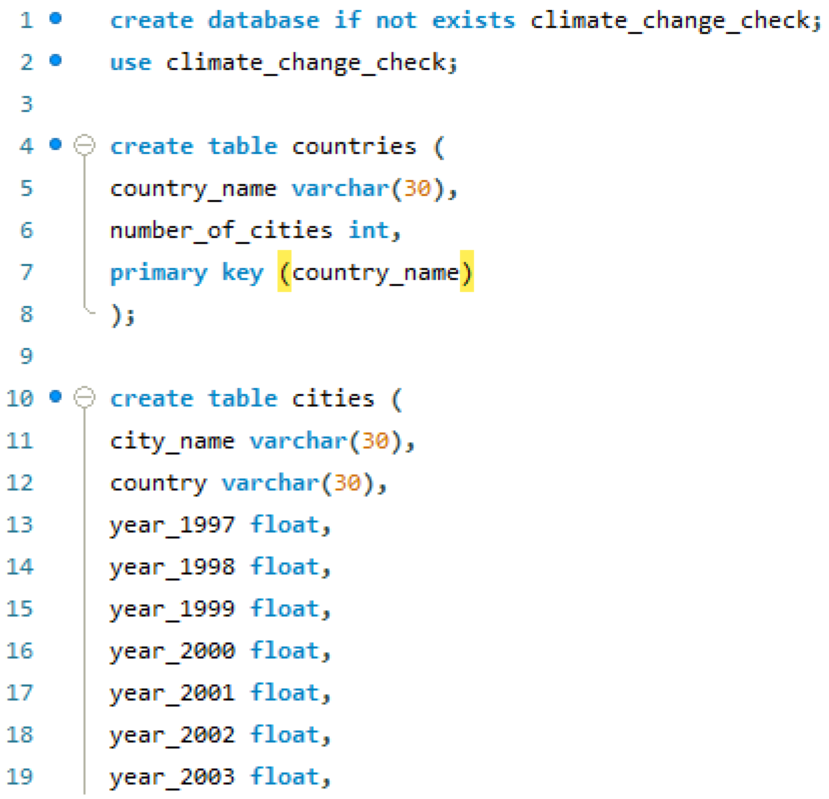


The following entities represent the results of my data cleaning process. First of all we have the measurement entity with all of the left over data from the temperature observation data set. Each measurement has its unique measurement ID, as well as all the information about where the measurement has been made, observation time, and results for average temperature and temperature increase. Second we have the country entity data with all countries and the number of their cities present in the temperature observation data. Then I created a more detailed entity with the unique key here being all the different cities. Information added is the country of the city, the average temperature in years 1997 through 2018 and their temperature increase for every year to the year before. The economic sector entity is the result of the greenhouse gas emission data. Each sector-country combination is unique and represents the greenhouse emission for a certain sector in a certain country from

1996 to 2018. The last entity displays all of the powerplants. Every power plant here is unique and therefore takes the place of the primary key in this entity. The entity also includes the country for each powerplant, it's primary fuel type etc. All of the entities have country as their foreign key referencing the country entity. This relationship between all entities will later allow me to perform various queries on correlations between the data of the different entities.

Creation of the database and data importation

After deciding on which type of database to use I began creating my relational database on MySQL workbench. After creating a new database I set up the tables for the entities, defined the different columns, the data types of their inputted values as well as primary and foreign keys to set up the different relations between the entities.

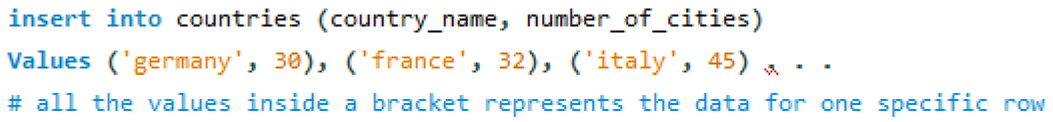


```sql
create database if not exists climate_change_check;
use climate_change_check;

create table countries (
country_name varchar(30),
number_of_cities int,
primary key (country_name)
);

create table cities (
city_name varchar(30),
country varchar(30),
year_1997 float,
year_1998 float,
year_1999 float,
year_2000 float,
year_2001 float,
year_2002 float,
year_2003 float,
```

Now filling those tables with my cleaned data was a challenge. Usually there is following syntax to import data into a table:



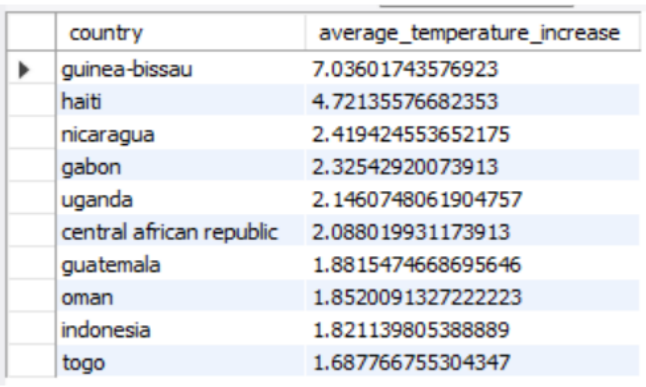
```sql
insert into countries (country_name, number_of_cities)
Values ('germany', 30), ('france', 32), ('italy', 45) , . .
# all the values inside a bracket represents the data for one specific row
```

Regarding the size of my data sets this is a very ineffective and outdated method. I therefore used the table data import wizard method to import my data into the tables.

For two of these tables this method did not work because the datasets were too large. To deal with this problem I imported the pymysql and sqlalchemy library import to set up a connection between MySQL local server and my jupyter notebook and import the data into my database with this engine.

Insights

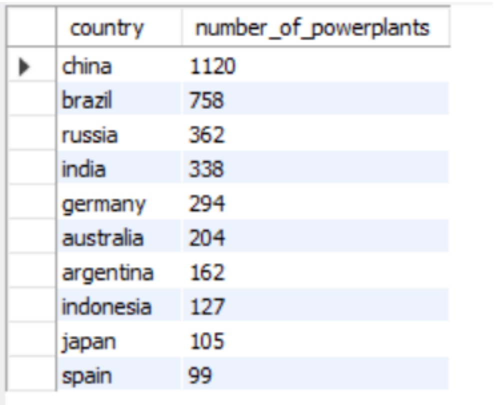
Having finished setting up my database I started writing queries in order to receive various insights from my data. I began with checking the baseline question of my business case: does the data reflect increasing temperatures over the last couple of years and which countries had the highest average temperature increase. For this insight I used the measurement table, grouped by all of the countries, aggregated the average temperature increase for each country and limited the output to the top 10 results.



country	average_temperature_increase
guinea-bissau	7.03601743576923
haiti	4.72135576682353
nicaragua	2.419424553652175
gabon	2.32542920073913
uganda	2.1460748061904757
central african republic	2.088019931173913
guatemala	1.8815474668695646
oman	1.8520091327222223
indonesia	1.821139805388889
togo	1.687766755304347

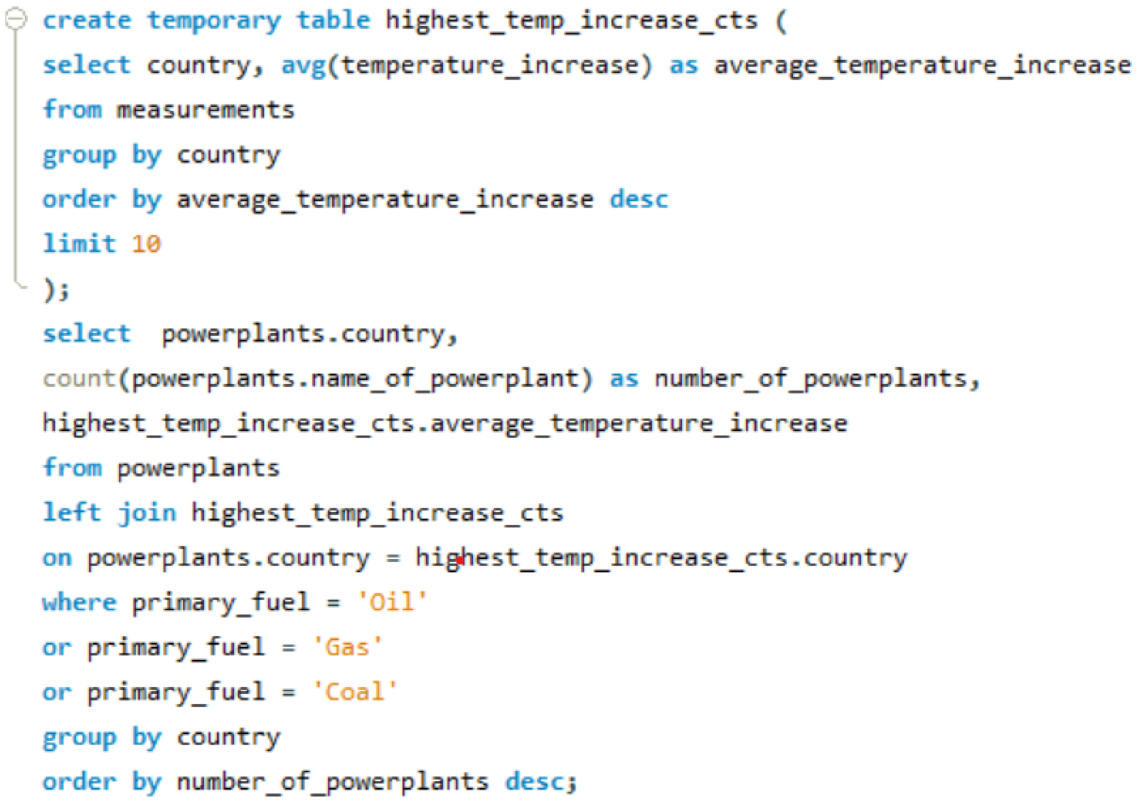
Next I wanted to look for correlations between those temperature increases and the factors I have decided to further analyze. It is said that greenhouse gas emissions and power plants using carbon based fuel types have a negative impact on our climate and I wanted to see whether these characteristics are very pronounced in these countries.

I started by using the power plants table, grouping by countries and aggregating the number of power plants whose primary fuel type is carbon based such as e.g. natural oil or coal. Below are the top 10 countries with the highest number of such power plants.



country	number_of_powerplants
china	1120
brazil	758
russia	362
india	338
germany	294
australia	204
argentina	162
indonesia	127
japan	105
spain	99

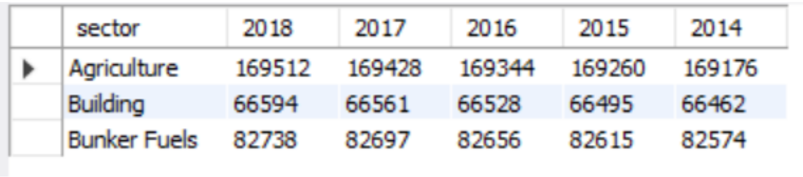
Using a left join I aggregated the results of this query and the previous query about looking for the top 10 countries with highest temperature increases to see if those countries are among the countries which use carbon based fuels powerplants.



```
create temporary table highest_temp_increase_cts (
select country, avg(temperature_increase) as average_temperature_increase
from measurements
group by country
order by average_temperature_increase desc
limit 10
);
select  powerplants.country,
count(powerplants.name_of_powerplant) as number_of_powerplants,
highest_temp_increase_cts.average_temperature_increase
from powerplants
left join highest_temp_increase_cts
on powerplants.country = highest_temp_increase_cts.country
where primary_fuel = 'Oil'
or primary_fuel = 'Gas'
or primary_fuel = 'Coal'
group by country
order by number_of_powerplants desc;
```

The results showed that 9 of the top ten temperature increase countries were among those, indonesia for example even being among the top ten countries with the highest number of power plants using carbon based fuels and we can conclude from the given data that there might be indeed a connection between temperature increase and the number of power plants using carbon based fuels.

Next I wanted to check my second factor: is there a correlation between how dominant economic sectors with high greenhouse gas emissions are represented in countries and their temperature increase. I began by looking at which economic sectors have the highest greenhouse gas emissions, using the sector country combo table, grouping by sector and summing the values for every sector in the last 5 years. The results are given below.



	sector	2018	2017	2016	2015	2014
▶	Agriculture	169512	169428	169344	169260	169176
	Building	66594	66561	66528	66495	66462
	Bunker Fuels	82738	82697	82656	82615	82574

I continued by using the same table, this time grouping by country, summing all of the values that have been collected for the emissions of those 3 sectors shown above for the last five years and ordering those values descendingly. To check if the top ten temperature-increasing countries are among those having data on these 3 sectors, I joined a temporary table displaying the top ten countries mentioned above. Once again all of them were among these countries, 4 of them even being among the top 40 countries with the highest amount of greenhouse emissions for the 3 specific sectors. From the given data we can assume a connection between the rising temperatures and the country's greenhouse gas emissions from the specified sectors.

Conclusion

Looking at climate change in the form of concrete data about temperature measurements over a longer period of time has been a very interesting project. Against my expectations my exploratory analysis has revealed that there has not been a general temperature increase over the past years but rather a form of up and down cycle with temperatures falling and rising every few years. Nonetheless on the country level there has been temperature increases reaching up to 7 degrees. Furthermore I found correlations between the factors defined in my business case I wanted to look into more deeply. Based on the given data, the queries done on MySQL have led to the conclusion that there might indeed be a connection between those factors and the increasing temperatures of the countries with the highest temperature increases. One has to keep in mind that these conclusions are all based on only the given data sets I have used. For example the fact that over 150 cities recorded in the data were based in the US might have affected the output of my analysis regarding the general temperature increase over the years. If I would have had more time, further steps could have been taking into account more data measurements about different countries so there would be an equal amount of data for each country proportionally.

Links

Github repository:
https://github.com/ferdi-leube/RNCP_Project

Jira project planning:
https://leubef.atlassian.net/jira/software/projects/RP/boards/2/roadmap?selectedIssue=RP-18