Data Report

Question

Does the concentration of air pollutants vary according to the population density of a country?

Data Sources

Population Density Data (www.kaggle.com)

This data source was selected because it contains exactly the data that is needed. It contains the population density of all countries worldwide. The data is presented in a consistently structured format, specifically in a a csv file. In addition to the population density of each country, the data source also includes other key figures. However, this is not a concern as the data can be filtered out in the data pipeline. The accuracy of this data is good as it reflects the real world and is correct. The data set is complete by featuring all contries. However, this data is not entirely up to date as the most recent update was in 2018. Unfortunately this was the most current available data source for the population density of each county. To download the data one must be signed in, which is why I pre-downloaded the file and placed it within the git repository. The data source is licensed under a "CC0 Public Domain" license, which means that there is no copyright on the data and there is no need to cite the source in the project, making it simple to fulfill the requirements of this license.

Pollution Data (www.public.opendatasoft.com)

This data source features a substantial amount of data on multiple pollutants across a broad range of countries. The data is consistently structured in form of a csv file, which can be downloaded without logging in. The data set encompasses not only the considered air pollutants (CO, NO2, SO2, O3) but also a multitude of other pollutants. The other pollutants will be excluded from the data in the data pipeline. The quality of the data is not always optimal. For instance, there are records where the quantity of air pollutants in the air is negative which is not possible. There are also records where the ammount is unsually high. This could be because of faulty data, but it may also be that does measuring stations are located directly near a highly polluting source and the value is therefore very high. However those extreme value deviations will be dealt with in the data pipeline. The data source is also incomplete as it only features the pollution records from 65 contries. This data source is licensed under a CC BY 4.0 License, which means that the data is free to use when appropriate credit is given, by making an attribution. I plan to give this attribution by mentioning the source of my data in this data report and also in the final-report.

Data Pipeline

pipline.sh

This shellscript deletes the daatabase, if already present, and then proceeds to execute the three parts of the data pipeline.

pipeline.jv

This first part of the pipeline is a Jayvee pipeline.

This pipeline is responsible for the download and transformation of the pollution data in order to fit the needs of my project. Within the pipline, the columns of the datasource that are needed are filtered. The type of data is also constrained to the data that should be included in the final data set, which means filtering for the records containing the analysed pollutants. Furhtermore it is ensured, that only "clean" data will be futher transformed by making sure, that all values

are inside of a resonable range. In order to make analysing the data easier, only records with the unit "µg/m³" and "ppm" are taken. This data source is updated regularly, however the structure of the data should stay, the same so that the pipeline can handle the changing data. A problem would be if the most common units change, because then most of the data would be ommited and not be considered during the analysis. However this is not to be expected.

The data set about the contries of the world is already downlaoded and therefore only is transformed inside of the pipeline. Only two columns of the table are needed which are the name of the country and its population density in the unit of "people per sqaure mile". Because the data fits nicely to the shape that I need for my project there are no more alterations that need to be done.

Both datasets are then stored in seperate tables of a sqlite database so that they can further be transformed and analysed by making SQL calls.

pipeline2.py

The second part of the pipeline is written in python. By making use of the sqlite3 library, I can create my final dataset that is then used as the result of tze data pipeline. It is first necessary to ensure that each country has a single value in order to make comparisons. To achieve this, the values in the unit of "ppm" are first converted to "µg/m³". Then records, that seem unrealistiv are removed. Afterwards a View is created, that contains the average value of each pollutant for each country. Finallly, the average values are joined with the population density in order to obtain the final data, which is then saved inside of a new table in the sqlite database.

plot.py

This python file is not really a part of the data pipeline. However it produces plots that enhance the visual representation of the data, which would be less informative when viewed in its raw, unprocessed database form. This script uses sqlite3 and pandas in order to query the database. To identify a linear trend within the data the LinearRegression module of the sklearn.linear_model library is utilized. In order to automatically generate the plots matplotlib.pyplot is used.

Result and Limitations

The output of the data pipeline is a sqlite database table with the corresponding values. It contains the country name, the type of air pollutant, the concentration of the pollutant and the population density of the country. As previously stated the values haven been plotted. The resulting plots are presented on the next page.

The results of my pipeline is structured data in the form of a sqlite database table. By always downloading the current pollution data it is also mostly up to date. The population density of a country usually is not very fluctuating therefore it is acceptable that this data is not always 100% up to date. The accuracy of my results is not optimal as for calculating the the amount of air pollutants, the average of each station is taken. This however is quite obiously not the real average amount that is in the air of that county, but it should be at least representative and comparable to the "real" average. The data is also inccomplete as it does not feature of the data for each county, caused by the incompleteness of the second data source. However it features a representative amount of about 60 countries. The resulting data is highly relevant to the question of this project as it is necessary to answer it.

Attributions

Thanks to OpenAQ for the datasource air pollution.

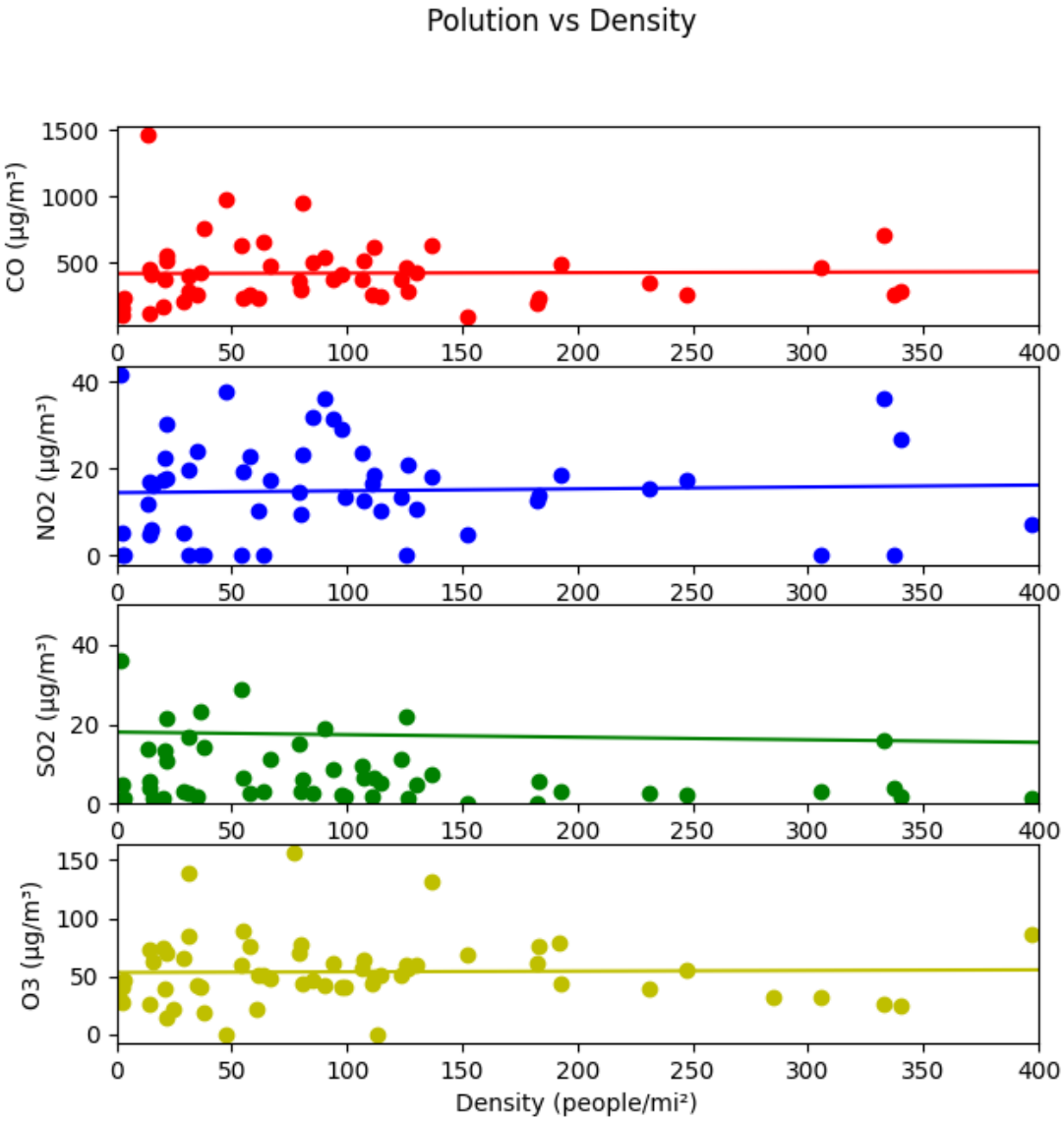

Figure 1: plots