Introduction

Cities play a significant role in the global struggle against climate change since they house the majority of the global population. By 2030, urban areas are projected to house 60% of people globally[1]. Moreover, a large part of the economic activity is concentrated in cities[2]. Limiting the emissions of greenhouse gases (GHG) is key to achieving a global warming limited to 1.5 or 2°C, as captured in the Paris agreement of 2015[3]. What are the actions cities can undertake to limit their GHG emissions? In order to answer this question, a first step is to get an idea over which factors are connected to city-wide GHG emission quantities and emission increases and decreases over time. The Climate Disclosure Panel (CDP) datasets on cities and greenhouse gases are a good starting point for this research. By analyzing these data and complementing them with additional sources, the variables that are related to reported GHG emission quantities can be identified. This is supplemented with a classification algorithm that aims to predict whether a city will report an increase or decrease in emissions compared to the previous year, based on a number of socio-economic and environmental changes between the years concerned. In these analyses, the main focus lies on the year 2017. Lastly, a data science app is developed to visually render the analysis of reasons for GHG emission increases and decreases that are reported by the cities, over different years and regions.

Part 1: Data quality check

The dataset on city-wide emissions was extracted from the CDP website. First of all the completeness and correctness of the dataset were assessed. Some figures were made using the missingno package showing that a rather large amount of data is missing throughout this dataset. More detailed analysis of missing data was performed for the columns that are of importance for the planned analyses, using the packages pandas, matplotlib and seaborn. It is shown that some countries have up to approximately 30% of these valuable data missing. A boxplot of the missing fractions for cities, grouped per region, indicates that there are no regions with a missing data fraction that is out of proportion in size.
The correctness of the data can be assessed as follows: for the 2017 dataset, the reported increase or decrease was checked by comparing the emission quantity in 2017 with the one in 2016. However, this can only be done in cases where both years include the same scopes and the same gases. A sunburst diagram was made with the plotly package to visually represent the amount of incorrect data per region. The subsequent layers in this diagram represent the amount of data per region, the proportion of checkable versus non-checkable data and lastly, the proportion of correct versus incorrect data for the checkable portion.

Part 2: Analysis of reasons for decreases/increases in emissions

In the 2017 dataset, a column is present where the cities can report the reason for the increase or decrease in emissions that is measured in that particular city. This natural language can provide useful insights into important phenomena that influence emission quantities. To analyse these unstructured data, various packages, including textblob, sklearn, matplotlib, networkx, are used. A first step is text pre-processing. The text is converted into lowercase, lemmatized and stop words are removed. Next, to get a general idea over the topic discussed in these reasons, two approaches are compared. A first approach uses term frequency - inverse document frequency (TF-IDF) with the aim of finding words that are important for particular reasons without being too common . A ranking of the different words can be obtained by

[1] 2016). The world's cities in 2016. United Nations.

[2] Pierre-Alexandre Balland, C.J.-F., Sergio Petralia, Mathieu Steijn, David Rigby, Cesar A. Hidalgo (Complex Economic Activities Concentrate in Large Cities.

[3] Streck, C., Keenlyside, P. von Unger, M. (2016). The Paris Agreement: A New Beginning. Journal for European Environmental Planning Law, 13(1), 3-29.

summing the TF-IDF for each word. However, it turns out that the penalty given to frequently occurring words is not enough: words like 'emission', 'increase' and 'decrease' are still evaluated to be among the most informative terms.
To solve this problem, a different approach is followed: a set of collection words is defined, which contains the previously discussed predominant but less valuable words. In addition to stop words, also these collection words are removed, followed by analysis of the normal term frequencies. In order to draw conclusions over reasons for decreased and increased emissions separately, the frequencies are calculated independently over both categories. Baseline frequencies are calculated, which are essential in discovering words that appear more often in a particular category than is expected based on random distribution of the occurring terms. The result is displayed in a bar plot.
Even though this bar plot already provides some useful information for interpreting the reasons, its value is somewhat limited due to a lack of context that complicates interpretation of the terms. Therefore, a next step is performed where some context is included in the data analysis. For every reported reason, bigrams are counted. The bigrams are scored based on their overall occurrence, and the highest scoring ones are visualized in a network plot. This provides a lot of added value to the simple bar plots obtained before: whereas the word 'sector' was previously identified to be mostly related to reasons for increased emissions, the additional context that is obtained now (residential, commercial, industrial sector) provides a lot of clarification.

Part 3: Linear regression of emissions

Now that some knowledge is collected over what cities themselves report to be crucial factors for changes in emissions, a next step is to construct a model of the emissions in 2017 in function of the available parameters in the dataset and see if some of the reported factors can indeed be found to show a significant correlation to the emissions. For this purpose, a multiple linear regression is performed on the dataset, where the total emissions are considered as response variable, and the region, scopes included, population, GDP, average altitude, land area, longitude and latitude are considered as potential predictors. The statsmodels and sklearn packages are used to perform the analysis. First, some preprocessing is necessary: longitude and latitude need to be extracted from the reported city location, and the GDP needs to be converted into a consensus currency in order to be comparable (USD is chosen).
In order to select the relevant predictors from the set of possible predictors, a forward stepwise selection procedure is followed, which results in a set of possible models, containing different numbers of predictors. The best model from this set is chosen, taking into account different evaluation criteria: Mallow's Cp, AIC and the adjusted R2 value. The model containing three predictors is ultimately chosen as the best linear regression model. These predictors are population, region East Asia and region North America. In this final model, population has a significant positive effect on the emissions. The fact that the two binary factors indicating the regions East Asia and North America are included in the final model and have a significant positive effect suggests that the average emissions in these regions are higher than in other regions.
Even though a lot of the proposed predictors are not included in the final model, the result implying that the population size of a city has an effect on its emissions is supported by the non-structured data analysis part, where the bigram 'population growth' was among the top 20 most frequently occurring bigrams. The insight that cities in the regions North America and East Asia perform badly in terms of emissions can be valuable in world-wide efforts to decrease these global emissions. The visualisation of the final model helps to better understand the effects of the different predictors.

Part 4: Classification algorithm for predicting decrease/increase in emissions

Making predictions of whether the greenhouse gas emission rate of a city will increase or decrease in

the (near) future can be a meaningful tool for decision making in terms of environmental protection. For this, classification models were generated for the changes in city greenhouse gas emission rate in 2017 compared to 2016.
In the data preprocessing, we incorporated extra sets of population and GDP data by country as a representation of the population and GDP changes in the cities that reported emission increase/decrease in 2017. Meanwhile, the cities' average temperature in summer and in winter were processed from different datasets and incorporated in our dataframe. An IterativeImputer was used to impute the missing data (which were mainly temperature data) based on the entries available.
Afterwards, to select meaningful predictors and reduce the dimensionality of the model, we decided to select 4 features: population growth, GDP growth, temperature change in summer and temperature change in winter as predictors for classification based on the feature importance in a tree-based feature selection, and the region variable was abandoned. A classification pipeline was built and applied on multiple classifiers separately: logistic regression, linear discriminant analysis, K-nearest neighbour, gaussian naive bayes, gradient boosting, decision tree, random forest, and support vector machine classifier. The training data was used to fit the classifiers, and the models were compared based on accuracy obtained by 10-fold cross-validation. Furthermore, it was made possible to compute AUC (area under the curve) that measures the quality of the model's predictions, and to plot ROC curves for every model. As a result, logistic regression, gradient boosting, decision tree and random forest provided somewhat acceptable models, although their accuracies are not great. The best model generated was the random forest model, which has an AUC of 0.79 and achieved an accuracy of 0.70. Therefore, it can be used to make predictions of future gas emission rates, although the prediction result cannot be perfectly reliable. For the rest of the models, the results are considered bad in terms of the low AUCs, and the accuracy scores are not high either. It can be assumed that the reliability of the models could be improved if higher-quality datasets with less missing values were available in this study, and a more desirable model might be generated if other predictors correlated with the emission could be incorporated.

Part 5: Data science app

The analysis of the reasons for increased and decreased emissions described in part 2 is extended and integrated into an app. Users can indicate the year and region for which they want a visualization, and the amount of detail they want to receive. The bar plot and bigram network plot (as explained in part 2) specific to their choice of parameters is then calculated and displayed. This app can be used to get an idea over what the most important issues are for certain regions and timeframes in their efforts toward a more sustainable economy with lower emissions. The app can be viewed at: `https://mda-somalia.herokuapp.com/`

Conclusion

The main difficulties encountered when performing data-analysis on the CDC datasets were data incompleteness and incorrectness. Mostly for the classification algorithm, these issues turned out to be limiting factors, despite efforts to improve the data quality and handle missing data. However, the performed data analyses still allow for a number of conclusions. The linear regression of emissions with variable selection shows that population is a key determining factor for city-wide emissions. It also revealed two regions that had relatively high emissions in the year 2017. The classification algorithm that was constructed, predicts reasonably well whether increased or decreased emissions will be measured for a city compared to the previous year, based on population growth, GDP growth and changes of average temperature in summer and winter. Lastly, natural language processing performed on the reasons that cities report for their changed emissions also provides insights on important factors in this matter. The data-science app that was constructed, visualizes this in a temporal and spatial dimension.