Identifying business goals

In recent decades global warming has been a threatening issue all over the world which should concern everyone. There are many businesses that their production process, even final products (i.e. gas-powered cars) causes CO2 and greenhouse emission which causes the depletion of ozone layer which leads to global warming. Our project does not aim to solve the problem of a specific business, but it rather addresses a global problem which many businesses, individuals and societies can benefit, but necessarily a short-term financial profit but rather a long-term benefit. In our project, we are analyzing average temperature around the world for the last 250 years in order to see the pattern of the fluctuation to see the relationship between CO2 emission and temperature rise and also try to predict future temperature to see if the problem is getting worse. Moreover, we will also analyze the spikes in the data and try to find reasons behind the spikes.

The businesses with production methods which have excess greenhouse emission may consider finding alternative production methods which have less greenhouse emission, after seeing the contribution of their production methods to global warming. Same companies may try to find alternative products which has less greenhouse emission (i.e. electric cars) and government should introduce tax cuts and other incentive for less greenhouse emitting products in order to promote them. The ultimate goal is of course to decrease the level of greenhouse emission but even stabilizing it is considered a success.

We have temperature data for the last 250 years and we will use Python for the code part. Because the data covers a 250 years period it is possible that the data may have a few inconsistencies.

Data mining goals and criteria

We will import our data in Python in order to graph the pattern of temperature fluctuation and to build a model that predicts the future temperature and also to identify the spike periods so that we can look for cause for that specific period.

We hope that our will predict with above 98% accuracy will also be efficient.

Data Understanding:

Since our project is about climate change, the types of data that we need will be files that give us information about average temperatures and dates. In addition to that, since we also plan on comparing climate change to the CO2 and greenhouse gas emissions we will also require a file that would give us recorded data of CO2 emissions of different countries and dates that correspond to the recording. As climate change is a topic that gets talked about a lot, the data should be readily available – We already have a dataset from Kaggle for both the recorded average temperatures as well as recorded CO2 emissions. It was also pointed out to us during our presentation that we should look into finding another dataset for the average temperatures that is more recent than the one that we already have, so it is a possibility that we will look for another dataset.

In our datasets, we have many files with different data criteria, such as average global land temperature, average global sea temperature, average temperature by countries, or even average temperature by big cities. Our current plan is to make use of the data of average global land temperature and average country temperature for some bigger country(or more relevant for us, like Estonia). The temperatures are recorded with monthly intervals and start from 1750. For the CO2 emissions dataset, we have one CSV file with monthly recordings of CO2 and Greenhouse gas emissions by country since 1750. From this file we will probably add up all the countries and then get monthly global CO2 emissions.

The data that we have is CSV files provided by Kaggle. The temperature dataset is sourced to „Berkley Earth“, which is affiliated with Lawerence Berkley National Laboratory, while the CO2 emissions dataset is credited to ourworldindata.org. The data is also quite well organized and so

should be fairly easy to work with.

Exploring data:

Our datasets are quite straightforward – We don't have many different fields that we need to use, but rather we have a lot of datapoints. For both our dataset, we have one reading per month, and in the CO2 emissins dataset we have recordings for individual countries, while in the Temperature dataset, we have recordings for average global land temperature, average global sea temperature, average country/state temperature or average big city temperature. There doesn't seem to be problems with the quality of the data for the most part, the only exception being that for the CO2 dataset, technologically less advanced countries don't have recordings for the entire 270 year period, but rather their recordings start somewhere in the middle of the dataset.

In summary – We don't seem to have any problems with our data, but we will look into finding more recent data for our Temperature dataset. The problem of some missing data from the CO2 dataset also seems like it could make some problems for us since missing data will lead to bad predictions, but one possible fix we have is to simply discard all data for the time period where there is missing data, and so we would start from somwehere around year 1900 or so.

Task4)
Plan -
Things that we need to do:
1)Search for a temperature dataset that is more recent, and organize the data so that it is in a readable format, if it is not already like that.

2)Analyze the data of the temperature dataset – Make analysis of both average land temperatures of both the entire planet, as well as some of the countries(for example Estonia)

3) Analyze the data of the CO2 dataset – Make analysis of the CO2 emission dataset, plot it onto a graph, and compare it to the data of the temperature dataset – If all goes well, it should show very high correlation.

4)Find extreme points of the temperature dataset – Find out if there are any years that have been much different than the average, find out the causes - This serves no practical purpose, but it is something that can can be pointed to as an interesting fact in the final presentation, and can also point to things that have great effect on the temperature. For example, I remember reading that there was a year where due to a large volcanic eruption there was an average global temperature decrease of 1.5 degrees Celsius for a year, and in that year it was snowing in mid summer – If we can find this year in the data, it can be an example of what a change of only 1.5 degrees can do.

5)Make a predictive model – The exact method we will be using is not certain at this point, but we would like to make a model that will make predictions about the future.

6)Test the predictive model – For testing, We could test the model on an incomplete dataset, for example only data up to the year 2000, and then predict years 2001-2019, and see of accurate that prediction is by looking at the actual data.

7)Finalize the project – Put everything that we have gathered in the previous task on a A0 paper.

For the sake of this report, we will assume that each person does 50% of each task, although this will likely change.

Time planned for each task -

1)6-7 hours
2)7-10 hours
3)3-5 hours
4)5-7 hours
5)12-15 hours
6)12-15 hours
7)5-6 hours
Total:50-60 hours.