**Task 3. Data understanding**

* Gathering data

For our project ’Factors that effect CO2 emissions’ our team collects data from World Bank website databank.worldbank.org.

* + Outline data requirements

We prefer that the data we collect is in csv-format and our source satsifies that condition. We try to collect data about 264 countries and the time range is from 1960 till 2019.

* + Verify data availability

If data turn out to be incomplete for certain countries or for a certain periood of time, we will narrow down the country choices and the periood.

* + Define selection criteria

We will collect data from the website databank.worldbank.org. Our goal is to find features that correlate best with CO2 emissions on diferent countries through years and find the best models within the set of features to predict CO2 emissions.

* Describing data

We collect data about CO2 emissions and three main topics: world develepment indicators, education statistics, poverty and equity. For each topic we search data from about 5-7 series.

For example about World Develepment Indicators we suppose that CO2 emission is correlated with 1) GDP per capita 2) households final expenditure 3) population 4) industry (% of GDP) 5) electricity production from oil, gas and coal sources 6) goods export.

Our data has three dimensions: country, year and values from diferent series. We put values to the columns (approximately 15-21 + 1 columns) and country/year together to the rows (maximum ca 15500 rows).

Most of the data will be numerical - continuos data. We will have categorical - ordinal data, also (for example data about education). We will use different methods to preprocess data. For example different education levels we will put on the scale: 0 means no education, 5 means high education and 5 is better than other values till 0. Numeric continiuos data we will categorize. We will normalize all features so that we place values between 0 and 1 separately for each feature.

* Exploring data

Having read the data more closely we conclude that the main probleem is missing data. Many countries has missing data for following reasons: countries has a short or fragmentary history – for example countries of the former Yugoslavia or Soviet Union; countries are or have been at war for some years. If country has no data for longer periood we will exclude the country. If there will be missing data for some years only, we will replace gaps with averages.

* Verifying data quality

It seems that data quality from World Bank is quite good – missing data is marked in the same way only (no zeros, Nan’s and spaces together). It’s possible to layout the data in the necessary way (choose columns and rows) on the webpage, already. Exporting data in csv-format is also possible.