TOPSECTOR LOGISTIEK

# Analyse opdracht logistieke tijdseries

Noder

September 2018

1. Executive summary

This report covers the results of an exploratory study that investigates the feasibility of combining different sources, containing transportation, environmental, geographical and demographical data, into a visualization tool. This tool should be able to visualize development of different statistics over time, e.g. a visualization of the development of *goods transport by road* in relation to *CO_2 emissions* over the last 10 years for different countries within the EU.

The process of building such a tool consists of four steps.

1. Finding the right data sources;

2. Pre-processing the data to allow for combination;

3. Make the data accessible from a combined storage;

4. Build visualization on top of this combined storage.

| Global data sources | Processed data | Combined storage | End-user visualizations |
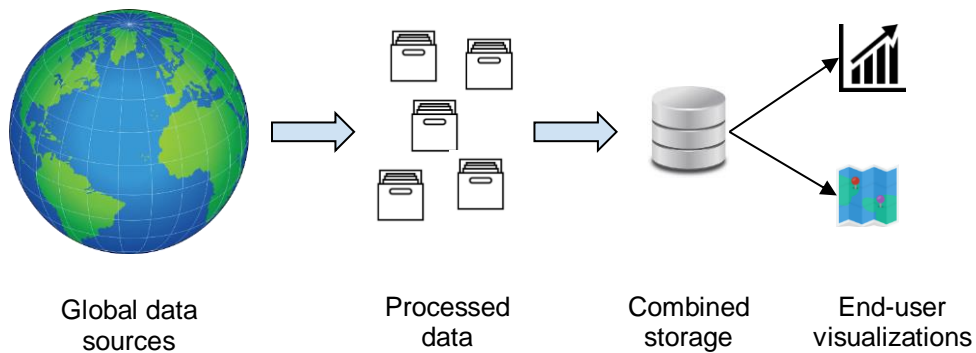
2. Data sources

The search for relevant data sources concentrated primarily on publicly available data provided by renowned national and international organizations and institutions. Data sources investigated for the purpose of this report included i.a. Statista, Eurostat, OECD, the Bureau of Transportation Statistics (United States) and the European Data Portal.

These data sources can be qualified in two categories, "direct data sources" (actual data) and "data source indexers" (referral sites to data). All the data sources below were reviewed for usefulness in this exploratory study.

Direct data sources:

1. **Eurostat - The statistical office of the European Union**
   Accessible through: https://ec.europa.eu/eurostat/
   Coverage: International/EU
   Relevant datasets:
   a. Transport in modal split (train, road, waterways)
   b. Transport safety
   c. Road transport infrastructure (length in km)
   d. Employment in transportation industry
   e. Total freight road transport (EU total/national)
   f. International annual road freight transport - goods unloaded in reporting country, by group of goods and type of transport
   Dataformat: data presented in excel/csv
   Used for this study: Yes

2. **OECD - The Organisation for Economic Co-operation and Development**
   Accessible through: https://stats.oecd.org/
   Coverage: International/OECD
   Relevant datasets:
   a. Transport infrastructure investment and maintenance spending
   b. Inland freight transport
   c. Road haulage charges and taxes
   d. Performance indicators
   Dataformat: data presented in excel/csv
   Used for this study: Yes

3. **Dataportaal van de Nederlandse overheid**
   Accessible through: https://data.overheid.nl/data/dataset
   Coverage: Netherlands / Caribbean Netherlands
   Relevant datasets:
   a. Transport and storage turnover
   b. Traffic participants
   c. Transport, storage and communications; finance
   d. Transport and storage; turnover change, index
   Dataformat: data presented in XML

4. **BTS - Bureau of transportation statistics**
   Accessible through: https://www.bts.gov/topics/freight-transportation
   Coverage: USA
   Dataformat: data presented in XLS
   Used for this study: No (scoped the study to Europe)

5. **Canada open government**
   Accessible through: https://open.canada.ca/data/dataset/0aab8f62-2e87-4ce5-b56e-5023f8f9e457
   Coverage: Canada
   Dataformat: data presented in CSV, XML
   Used for this study: No (scoped the study to Europe)

6. **Statista**
   Accessible through: https://de.statista.com/themen/1438/logistikunternehmen/
   Coverage: International
   Dataformat: XLS
   Reason not to include: Detailed data is only available commercially

Data source indexers:

7. **European data portal**
   Accessible through: https://www.europeandataportal.eu/data/en/dataset
   Coverage: International/EU
   No data taken from here. (Taken from the sources above directly)
   Efficient and effective way of finding sources of data

8. **Google datasearch**
   Accessible through: https://toolbox.google.com/datasetsearch (use Google Chrome)
   Coverage: International
   No data taken from here. (Taken from the sources above directly)
   Efficient and effective way of finding sources of data

An overview of and further information on explored data sources can be found in the Excel file *Data_Sources*.

3. Data pre-processing

An integral part of combining different data sources, is the preprocessing of these data sources. Different data sources don't per se adhere to the same standards (e.g. type of country codes used) or Units of Measure (UoM).

The steps that had to be undertaken for each dataset were:

1. **Harmonize country naming**
   All datasets used in testing were harmonized to use three letter country abbreviation according to ISO-3166 Alpha-3. Eg. The Netherlands → NLD

2. **Unpivot / de-normalize datasets**
   Most datasets come in a pivoted format. For example:

Year Country	2001	2002	2003
Country A	1234	1235	1236
Country B	1237	1238	1239
Country C	1240	1241	1242

   For this data to be easily combined and used by visualization tools, the data was unpivoted / normalized to the following format:

Country	Year	Value
Country A	2001	1234
Country A	2002	1235
Country A	2003	1236
Country B	2001	1237
Country B	2002	1238
Country B	2003	1239
Country C	2001	1240
Country C	2002	1241
Country C	2003	1242

3. **Harmonize the Unit of Measure (UoM)**
   Not all datasets use the same UoM. E.g. thousand tonnes vs. tonnes. These had to be harmonized to one format for datasets to be combined.

These pre-processing steps can be done manually, however given the amount of preprocessing required for each table, a preprocessing script in Python that would automate this procedure to a large extent was developed.

While preprocessing the data in an automated fashion, there were some obstacles to be taken into account in processing future data:
- Excel downloads by the OECD turned out to be "disguised' xml files. Manual changes (deleting the first hidden line and saving as .xls) to each file were necessary to enable further processing
- Country codes needed to be added to all excel files to be able to display country-related values in both Gapminder and Tableau
- Country names (e.g., "Bosnia-Herzegovina" vs. "Bosnia and Herzegovina") were not consistent across tables → manual changes were necessary
- The table structure was not consistent across tables → manual changes were necessary

The preprocessed files and the Jupyter notebooks written in Python that automated a large amount of the preprocessing are *Create_compatible_xls_OECD* and *Create_compatible_xls_Eurostat* are delivered as part of this report.

4. Data Access

All the data sources that have been investigated offer data in either XLS, XML or CSV format. Whilst these types of files lend themselves very well for one-off analysis, they are somewhat limited when it comes to convenient re-use to create different visualizations over time.

Therefore this research investigated the creation of a single repository in which different data sources can be combined. In this example, five different datasets were obtained in separate files:

1. Amount of goods transported by road
2. Amount of goods transported by rail
3. Amount of goods transported by inland waterways
4. Greenhouse gas emissions by transport over road
5. Country names in different formats

These separate datasets were pre-processes as described in section 3. Based on a preliminary selection of visualization tools, two options for a single repository were considered:

1. **Distributed Data Framework (DDF)**
   DDF is a data model, meaning it describes a way to organize data and to define how pieces of data relate to each other. DDF can be used for data harmonization, meaning it can combine data from different sources into integrated, consistent and unambiguous data sets.[1]
   a. <u>Pros</u>: Purpose built for data-analysis, common data model in open number initiative. Toolset available in common programming languages.
   b. <u>Cons</u>: No standard interface available in most visualization tools, unknown to participants in this project and therefore requires big time investment.

2. **MySQL**
   MySQL is the world's most popular open source database. With its proven performance, reliability and ease-of-use, MySQL has become the leading database choice for web-based applications, used by high profile web properties including Facebook, Twitter, YouTube, Yahoo! and many more.[2]
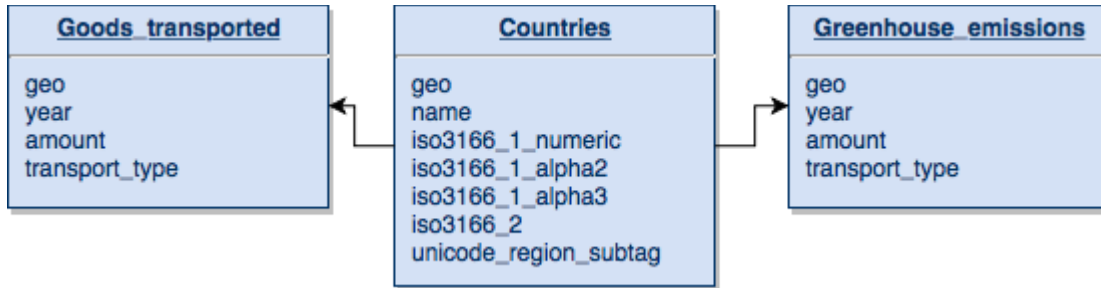   a. <u>Pros</u>: One of the most widely used relational database solutions, with lots of supporting tools that allow for easy usage (e.g. easy XLS to MySQL conversion). Known technology for participants in this project. Standard interfacing with most visualization tools available
   b. <u>Cons</u>: Good solution for data storage, not for modifications on large data-sets.

Based on the pros and cons it has been decided to use a MySQL database in this example, as it was not preferable to invest too much of the available time in getting familiar with new technology. The resulting MySQL database looked as follows:

---

[1] https://open-numbers.github.io/ddf.html
[2] https://www.mysql.com/about/

Goods_transported

geo
year
amount
transport_type

Countries

geo
name
iso3166_1_numeric
iso3166_1_alpha2
iso3166_1_alpha3
iso3166_2
unicode_region_subtag

Greenhouse_emissions

geo
year
amount
transport_type

This database allows a visualization tool to get access to all the datasets listed above through a single connection:

1. Good_transported contains
   a. Amount of goods transported by road
   b. Amount of goods transported by rail
   c. Amount of goods transported by inland waterways
2. Countries, contains:
   a. Country names in different formats
3. Greenhouse_emissions
   a. Greenhouse gas emissions by transport over road

This structure can easily be expanded with different data sources by adding tables to the database.

5. Visualization tools

With respect to visualization tools, a short review was done of:
- Gapminder Tools Offline v.3.4.0, provided by the Gapminder Foundation (https://www.gapminder.org/),

as well as:
- Tableau Desktop 2018.2 (https://www.tableau.com/products/desktop).

Gapminder

Cost: Free

Flexibility in data sources: Low. Limited to csv.

Flexibility in visualization: Limited. Few different visualizations, and limited customization per visualization.

Tableau

Cost: $70/user/mth

Flexibility in data sources: High.

Flexibility in visualizations: High.

Visualization in Tableau

After a preliminary test with both tools, and due to the time constraint for this exploratory study, the decision was made to make the visualizations in Tableau. These visualizations are based on the MySQL dataset created as described in section 4.

6. Results: examples

1. **The amount of goods transported by road (per country)**
   Description: This visualization shows the amounts of goods transported by EU countries from 2006 - 2016.
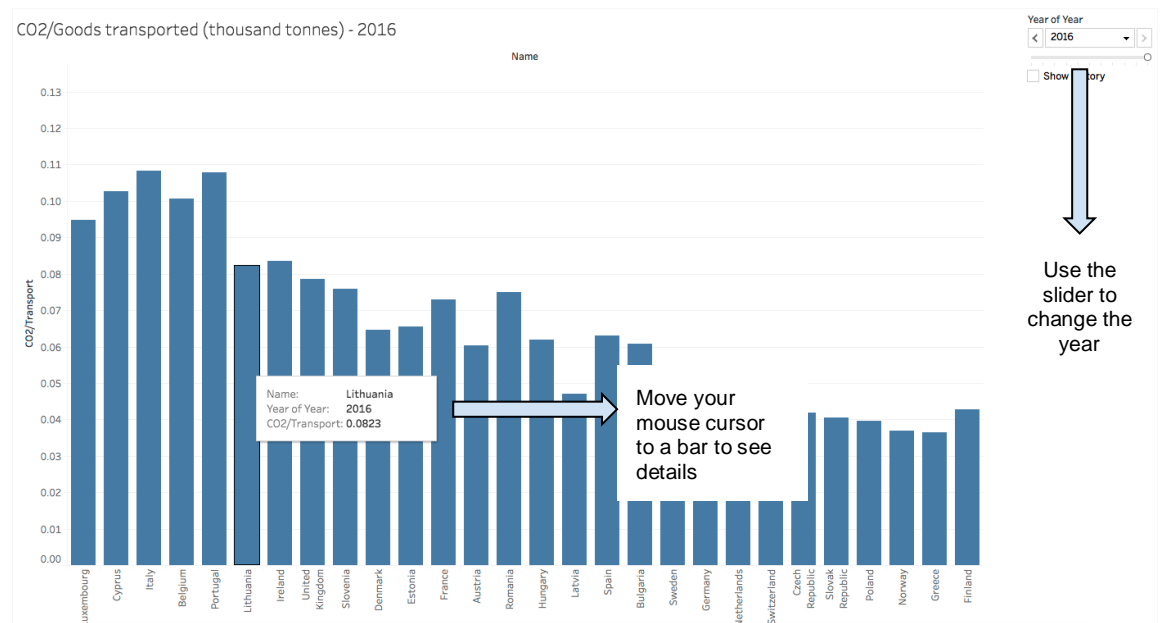   Link: Tableau link
   Preview:
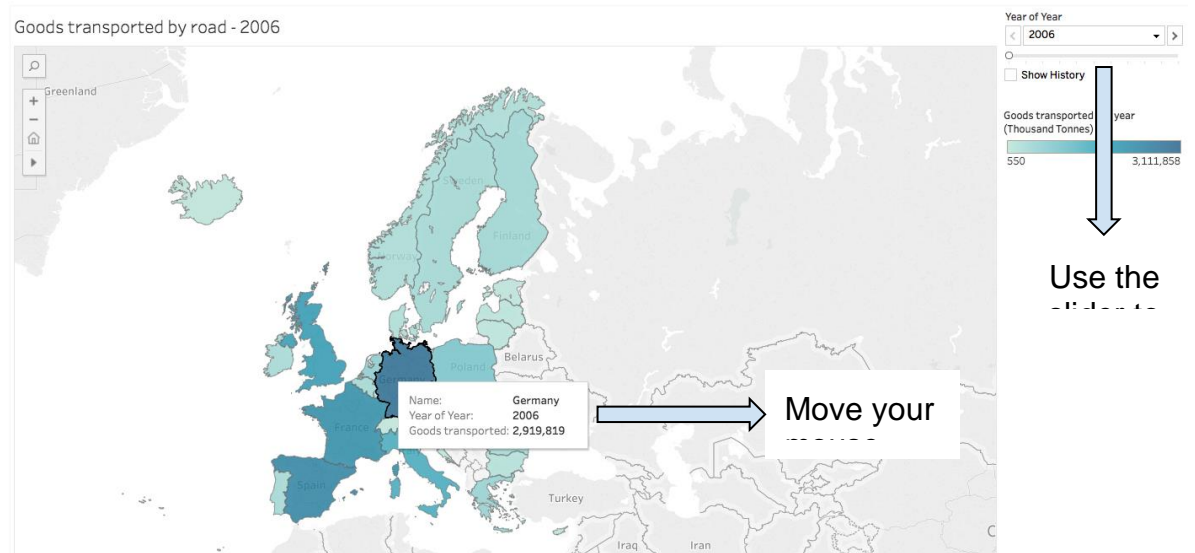
2. **The amount of greenhouse gasses emitted per goods transported by road: by country.**
   Description: This visualization shows factor of greenhouse gasses emitted per goods transported by road. The data is shown per country, per year.
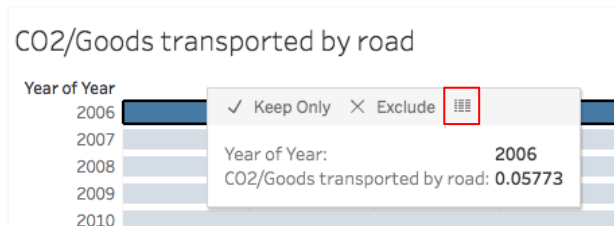   Link: Tableau link
   Preview:

3. **The amount of greenhouse gasses emitted per goods transported by road: European Aggregates**

Description: This visualization shows factor of greenhouse gasses emitted per goods transported by road. The data is shown aggregated for multiple countries, per year.

Link: Tableau link

Preview:

1. To review the detailed data that was used in this analysis, move your mouse cursor to a specific bar, click on the bar and hold your cursos there until the box depicted below is shown:
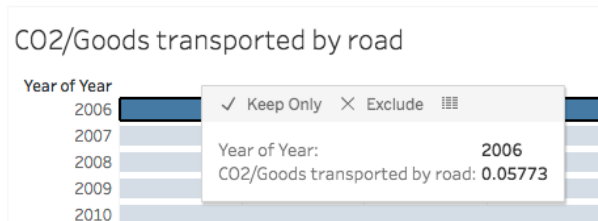
2. Click on the icon highlighted in red

3. Click on the tab "Full data"

4. Check the "show all columns" checkbox

5. All the source data is shown as below

6. Recommendations

1. **The Top Sector could play an active role in the collection of transportation data.**
   To improve the gathering and aggregation of data on the state of the transportation in the Netherlands, the Top Sector could play an active role. We see an opportunity to steer the data collection process and that way improve the quality and relevance of the available data.

2. **Another, open source, BI tool can be used to achieve the same results.**
   Tableau is a very strong BI tool, however for the purposes of the of this investigation, it might be an overkill. To be more effective, another BI tool can be employed. Open source tools, such as Metabase can be considered as these tools are functionally comparable to Tableau.

3. **Co-operate with data providers to reduce the need for pre-processing**
   The current datasets are delivered in varying formats, when it comes to lay-outs, UoM or Country naming. The Topsector might consider to approach data providers to deliver data in a uniform format.

4. **Dataportaal van de Nederlandse overheid**
   In further research, the Dataportaal van de Nederlandse overheid could be taken into account. This data source was not included in the current research, due to formatting issues and time constraints.

7. Annexes

These datasets are available in a separate ZIP/RAR:

Folder/File	Content
gapminder-geo-id-mapping.xlsx	Mapping between country name and country codes/abbreviations (iso3166_1_alpha2, iso3166_1_alpha3, iso3166_1_numeric, iso3166_2, unicode_region_subtag)
Eurostat	
Air_transport_of_freight.xls	Air transport of freight by NUTS 2 regions
Air_transport_of_goods.xls	Air transport of goods (in tonnes)
Goods_transport_by_inland_waterways.xls	Goods transport by inland waterways (in thousand tonnes)
Goods_transport_by_rail.xls	Goods transport by rail (in thousand tonnes)
Goods_transport_by_road.xls	Goods transport by road (in thousand tonnes)
Maritime_transport_of_freight.xls	Maritime transport of freight by NUTS 2 regions
Modal_split_of_freight_transport.xls	Modal split of freight transport (% in total inland freight tonne-km) Railways, inland waterways - sum of available data
Sea_transport_of_goods.xls	Sea transport of goods (thousand tonnes)
Total_length_of_motorways.xls	Total length of motorways (in KMs)
Volume_of_freight_transport_relative_to_GDP.xls	Volume of freight transport relative to GDP index (2005 = 100)
Eurostat_xls_Gapminder_compatible	
Air_transport_of_freight_by_NUTS_2_regions_Gapminder_compatible.xlsx	processed with *Create_Gapminder_compatible_xls_Eurostat.py* based on *Air_transport_of_freight.xls*
Air_transport_of_goods_Gapminder_compatible.xlsx	processed with *Create_Gapminder_compatible_xls_Eurostat.py* based on *Air_transport_of_freight.xls*
Goods_transport_by_inland_waterways_Gapminder_compatible.xlsx	processed with *Create_Gapminder_compatible_xls_Eurostat.py* based on Goods_transport_by_inland_waterways.xls
Goods_transport_by_rail_Gapminder_compatible.xlsx	processed with *Create_Gapminder_compatible_xls_Eurostat.py* based on Goods_transport_by_rail.xls
Goods_transport_by_road_Gapminder_compatible.xlsx	processed with *Create_Gapminder_compatible_xls_Eurostat.py* based on Goods_transport_by_road.xls
Maritime_transport_of_freight_by_NUTS_2_regions_Gapminder_compatible.xlsx	processed with *Create_Gapminder_compatible_xls_Eurostat.py* based on Maritime_transport_of_freight.xls
Modal_split_of_freight_transport_Gapminder_compatible.xlsx	processed with *Create_Gapminder_compatible_xls_Eurostat.py* based on Modal_split_of_freight_transport.xls
Sea_transport_of_goods_Gapminder_compatible.xlsx	processed with *Create_Gapminder_compatible_xls_Eurostat.py* based on Sea_transport_of_goods.xls

Total_length_of_motorways_Gapminder_compatible.xlsx	processed with *Create_Gapminder_compatible_xls_Eurostat.py* based on Total_length_of_motorways.xls
Volume_of_freight_transport_relative_to_GDP_Gapminder_compatible.xlsx	processed with *Create_Gapminder_compatible_xls_Eurostat.py* based on Volume_of_freight_transport_relative_to_GDP.xls
Eurostat_xls_Tableau_compatible	
Air_transport_of_freight_by_NUTS_2_regions_Tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_Eurostat.py* based on *Air_transport_of_freight.xls*
Air_transport_of_goods_Tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_Eurostat.py* based on *Air_transport_of_goods.xls*
Goods_transport_by_inland_waterways_Tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_Eurostat.py* based on *Goods_transport_by_inland_waterways.xls*
Goods_transport_by_rail_Tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_Eurostat.py* based on *Goods_transport_by_rail.xls*
Goods_transport_by_road_Tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_Eurostat.py* based on *Goods_transport_by_road.xls*
Maritime_transport_of_freight_by_NUTS_2_regions_Tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_Eurostat.py* based on *Maritime_transport_of_freight.xls*
Modal_split_of_freight_transport_Tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_Eurostat.py* based on *Modal_split_of_freight_transport.xls*
Sea_transport_of_goods_Tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_Eurostat.py* based on *Sea_transport_of_goods.xls*
Total_length_of_motorways_Tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_Eurostat.py* based on *Total_length_of_motorways.xls*
Volume_of_freight_transport_relative_to_GDP_Tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_Eurostat.py* based on *Volume_of_freight_transport_relative_to_GDP.xls*
OECD	
Road_injury_accidents_overview.xls	Road injury accidents (number)
Charges_and_taxes_by_type.xls	Dataset: Road haulage charges and taxes (Vehicle taxes (EUR/year)
Net_charges_per_domestic_haul.xls	Dataset: Road haulage charges and taxes (Vehicle taxes (EUR/year))
Net_charges_per_domestical_haul_by_type.xls	Dataset: Road haulage charges and taxes
Economic_and_social.xls	Share of household expenditure for purchase of vehicles in total household expenditure for transport
Energy_and_environment.xls	CO_2 emissions from transport in tonnes per one million units of current USD GDP
Safety.xls	Road fatalities per one million inhabitants
Traffic.xls	Road traffic in vehicle-km per one thousand units of current USD GDP
Transport_equipment.xls	Road motor vehicles per one million units of current USD GDP

Transport_infrastructure.xls	Total inland transport infrastructure investment per GDP
Transport_measurement.xls	Total inland freight transport in tonne-km per one thousand units of current USD GDP
Inland_waterways.xls	Total inland waterways freight transport (Tonnes-kilometres, Millions)
Short_term_indicators_Overview.xls	Total rail freight transport (Tonnes-kilometres, Millions)
Road.xls	Total road freight transport (Tonnes-kilometres, Millions)
Transport_infrastructure_investment_and_maintenance_spending_Capital_Value.xls	Total inland transport infrastructure capital value (euro)
Transport_infrastructure_investment_and_maintenance_spending_Investment_Spending.xls	Total inland transport infrastructure investment (euro)
Transport_infrastructure_investment_and_maintenance_spending_Maintenance_Spending.xls	Transport infrastructure investment and maintenance spending (euro)
Freight_transport_Coastal_shipping.xls	Coastal shipping (national transport) (Tonnes-kilometres, Millions)
Containers_transport.xls	Rail containers transport (TEU)
Freight_transport_Inland_freight_transport.xls	Total inland freight transport (Tonnes-kilometres, Millions)
Inland_passenger_transport.xls	Total inland passenger transport (Passenger-kilometres, Millions)
Road_casualties.xls	Road casualties (injured + killed) (Persons)
Road_injury_accidents.xls	Road injury accidents (Number)
OECD_xls_Tableau_compatible	
2012_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Net_charges_per_domestical_haul_by_type.xls*
CO2_emissions_from_transport_in_tonnes_per_one_million_units_of_current_USD_GDP_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Energy_and_environment.xls*
Coastal_shipping_(national_transport)_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Freight_transport_Coastal_shipping.xls*
Net_charges_per_domestic_haul_(Euro)_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Net_charges_per_domestic_haul.xls*
Rail_containers_transport_(TEU)_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Containers_transport.xls*
Road_casualties_(injured_+_killed)_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Road_casualties.xls*
Road_fatalities_per_one_million_inhabitants_tableau_	processed with *Create_Tableau_compatible_xls_OECD.py*

compatible.xlsx	based on *Safety.xls*
Road_infrastructure_maintenance_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Transport_infrastructure_investment_and_maintenance_spending_Maintenance_Spending.xls*
Road_injury_accidents_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Road_injury_accidents.xls*
Road_motor_vehicles_per_one_million_units_of_current_USD_GDP_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Transport_equipment.xls*
Road_traffic_in_vehicle-km_per_one_thousand_units_of_current_USD_GDP_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Traffic.xls*
Share_of_household_expenditure_for_purchase_of_vehicles_in_total_household_expenditure_for_transport_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Economic_and_social.xls*
Total_inland_freight_transport_in_tonne-km_per_one_thousand_units_of_current_USD_GDP_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Transport_measurement.xls*
Total_inland_freight_transport_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Overview_Freight_transport.xls*
Total_inland_passenger_transport_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Inland_passenger_transport.xls*
Total_inland_transport_infrastructure_capital_value_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Transport_infrastructure_investment_and_maintenance_spending_Capital_Value.xls*
Total_inland_transport_infrastructure_investment_per_GDP_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Transport_infrastructure.xls*
Total_inland_transport_infrastructure_investment_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Transport_infrastructure_investment_and_maintenance_spending_Investment_Spending.xls*
Vehicle_taxes_(EURyear)_tableau_compatible.xlsx	processed with *Create_Tableau_compatible_xls_OECD.py* based on *Charges_and_taxes_by_type.xls*
StatCanada	
23100230.csv	Canada: Trucking industry, fleet and equipment statistics, by province and territory
23100230_MetaData.csv	Meta-data with respect to *23100230.csv*

The following scripts were used:

File	Content	Dependencies	Remarks
Create_Tableau_compatible _xls_OECD.py	The py-file reads and preprocesses the xls-files in all subdirectories of the directory *OECD.* It outputs preprocessed xls-files that are compatible with MySQL/Tableau. These are saved in the folder *OECD_xls_Tableau_compatible.*	Python libraries: Pandas 0.23.0 Numpy 1.14.4 gapminder-geo-id-mapping.xlsx	Please note: The py-files automate the preprocessing to a large extent. Manual checks are however necessary <u>for all files</u>, given that the structures of the files may differ in unpredictable ways.
Create_Tableau_compatible _xls_Eurostat.py	The py-file reads and preprocesses the xls-files in all subdirectories of the directory *Eurostat.* It outputs preprocessed xls-files that are compatible with MySQL/Tableau. These are saved in the folder *Eurostat_xls_Tableau_compatible.*	Python libraries: Pandas 0.23.0 Numpy 1.14.4 gapminder-geo-id-mapping.xlsx	
Create_Gapminder_compatible_xls_Eurostat.py	The py-file reads and preprocesses the xls-files in all subdirectories of the directory *Eurostat.* It outputs preprocessed xls-files that are compatible with Gapminder. These are saved in the folder *Eurostat_xls_Gapminder_compatible.*	Python libraries: Pandas 0.23.0 Numpy 1.14.4 gapminder-geo-id-mapping.xlsx	