# Estonian car statistics exploration.

# Project goals and planning.

*GitHub repository:* [*https://github.com/SilverKolde/IDS2020-group-project*](https://github.com/SilverKolde/IDS2020-group-project)

Collaborators:

* Taavi Karelson
* Silver Kolde
* Marten Kuusmann

## Business understanding

**Identifying business goals**

**Background**

The idea of this project started when thinking about recent fuel excise lowering in Estonia. The government’s main goal was to revive the transportation sector by lowering diesel excise. We began to question if it has any effect on private citizens also when buying passenger cars. After some digging into the idea, we discovered a large dataset from Estonian Road Administration that holds all the passenger cars (M1 and M1G category). That spurred our curiosity even further to find out different trends over the years – what features are relevant to Estonians when buying new cars? Have the most important features changed over the years? Does it differ by location (cities, countryside areas)?

So let it be stated here that this project is grown out of curiosity rather than a business need or requirement.

Business goals

One of the main goals of this project is to find out if Estonian’s new car buying preferences change after fuel excise has been altered.

The other one is to discover trends – are there any features that are becoming increasingly important and does it differ by region.

Business success criteria

What we hope to see is that fuel excise has affected the buying habits of Estonians over the years, but we are very aware of the fact that Estonia is small and we can’t make fundamental conclusions without investigating global trends. So the project is considered successful when all members have learned something new solidified basic concepts of data science/data mining.

**Situation assessment**

Inventory of resources

We have a dataset of all cars registered in Estonia (about 800 000 entries). We found some data about fuel excise from tax administration, but it may not be enough. We are still gathering information and waiting for a reply from the tax administration.

Requirements, assumptions, and constraints

The deadline for making the introduction video about the project 12:00, Monday, Dec 14.

Risks and contingencies

It may happen that we can’t get enough useful data (soon enough) about fuel excise and we fail to deliver one of the main goals of the project.

Terminology

At this point, we don’t see anything that needs extra clarification.

*~~Costs and benefits~~ irrelevant*

**Data-mining goals**

Goals

We aim to discover the trends in buying a car (instead of people buying them randomly) and display them in graphs.

Data-mining success criteria

If we discover that fuel excise directly affects people’s buying decision, we consider it a major success. And if we only manage to find some features that have become increasingly important over the years, we consider it good enough.

# Data understanding

## Gathering data

To test our hypothesis, we need data regarding total registered cars and newly registered cars for past 10 years.

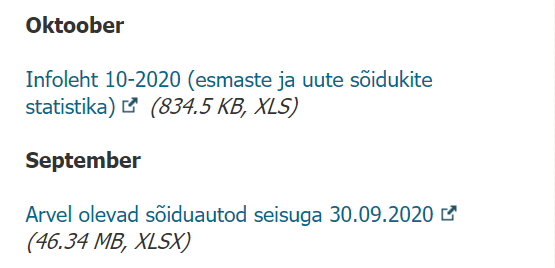
Additionally, to that we need to have dates when the fuel excise has been changed in the past 10 years. But this is not really a dataset.

### Outline data requirements

To achieve set goals, we need two datasets from Republic of Estonia Road Administration's website. Firstly, we need dataset of all of the M1 and M1G type cars registered in Estonia. Minimum required fields that are needed are: body type, engine capacity, horsepower’s, county of registration, city of registration and year of registration. And secondly we need a dataset of all newly registered M1 and M1G type cars in Estonia for the past 10 years. Minimum requirements for the second dataset are: date of registration (MM-yyyy), fuel type, engine capacity, number of cars and type or registration (private person or not).

### Verify data availability

All of the required data is public under Republic of Estonia Road Administration's website statistics page at: <https://www.mnt.ee/et/ametist/statistika/soidukid/soidukitega-tehtud-toimingute-statistika>.

The list of all registered cars is updated quarterly, so we use dataset as of 30.09.2020 (“Arvel olevad sõiduautod seisuga 30.09.2020”). With over 640000 rows and total of over 0,8 million cars in total. All of the required fields are available in the dataset, but there is some additional information also available, like color, CO2 etc.

The list of newly registered vehicles data is submitted monthly, so we need to gather this data for every month separately and then concatenate those sets into one big dataset. As an example we have October 10-2020 first and new registrations statistics file “Infoleht 10-2020(esmaste ja uute sõidukite statistika)”. In there we have all of the required fields and a lot of additional information. If we take October as an example, then the total rows in the final dataset should be approximately 1223 (new cars registered in October) \* 12 (months in year) \* 10 (total years) = ~145000 cars.

### Define selected criteria

Ei tea, mida siia kirjutada

## Describe data

The data for both datasets is available as Windows Excel file format “.xlsx”. First dataset can be used as it is. Second dataset files have 11 subpages, so the correct subpages have to be picked. And then only relevant M1 and M1G cars to be added.

## Exploring data

Electric cars have 0 capacity engines, so those should not be included into datasets where that feature is compared.

## Verifying data quality

In the first dataset of total registered cars, there are 11 cars without first registration year, 27 cars do not have engine capacity, 26 do not have engine power values but all of those are marked as inactive cars, so those missing values are not relevant.