Homework 10 D3_report

- Team number : D3
- Rait Pommer, Cardo Kenten Ross, Liis Reimand
- Project title : Estonian vehicles
- Github

Task 2

The project's database from *Eesti avaandmed* includes information about all the M1 and M1G category vehicles in Estonia, mainly their technical aspects such as gearbox, engine type and capacity.

- M1 category vehicles do not have more than eight seats in addition to the driver's seat
- M1G category vehicles are basically off-road vehicles whose full weight does not exceed two tons.

Based on the given data about technical specifications, good statistics will be created about the overall condition and technical package of Estonian cars, dividing them into categories by county, brand and fuel type. Database also has information about CO_2 NEDC (*New European Driving Cycle*) and emission standards (EURO) of specific cars, which gives an opportunity to predict CO_2 emissions by car brand and also give an overview of current situations by fuel, engine and brand .

- NEDC score - shows the emission level of the engine
- EURO standard - specifies which pollution standard the car classifies to (EURO4, EURO 6 etc).

Our project's purpose is not financial gain, which is why we can just bring out who will benefit from given work. Since climate policy and the level of pollution caused by cars are very relevant topics nowadays, the project will be mostly beneficial for organizations responsible for meeting climate requirements, which concern the emission level of cars. The project gives a good overview of the current situation in Estonia - which cars have significantly higher emission levels, in which counties people mainly drive cars that already meet the necessary euro requirements and also have a predicted scenario of how the emission levels will change in the future. This helps to set new requirements of vehicles and draw conclusions towards the solution of the given problem. The project will also benefit technical inspection and service shop companies, because it pinpoints exactly where there is a greater need of technical assistance. Car dealers would get information about people's preference in different areas, therefore it is easier to offer specific cars to the right target group.

The source of the data is the Eesti Avaandmed "Vehicle statuses in Estonia" database (138.1MB). The project will be carried on by all team members using coding language Python and platform Jupyter Notebook. The result of the project will be presented on 15th of

December and final work will be submitted on 12th of December. The only contingency could be a failure of our technical equipment or errors in the Jupyter Notebook system.

Terminology:

- Jupyter Notebook - web-based interactive computing platform
- Technical specifications - engine, gearbox, fuel etc
- Capacity of engine / engine size - the volume of fuel and air that can be pushed through a car's cylinders and is measured in cubic centimeters
- Engine type - power source for example as a fuel, electricity, gas etc
- Fuel type - petrol, diesel, lpg, cng, electricity
- CO2 emission level - emission stemming from the burning of fossil fuels (for this project)
- Brand - BMW, Audi, Ford
- Model - refers to the name of a car product and sometimes a range of products / manufacturer's range or series of cars
- Registered weight - the weight the car has been registered into the system
- Full weight - the maximal weight of the car (set by manufacturer)
- Empty weight - weight of vehicle, which does not include passengers or cargo
- Axles - a rod or shaft that connects a pair of wheels to propel them and retain the position of the wheels to one another.
- Body type - categorisation of a vehicle based on its design, shape and space (universal, sedan, coupe, SUV)
- CVT gearbox -continuously variable transmission, does not use any gears at all
- Manual gearbox - gear changes require the driver to manually select the gears by operating a gear stick and clutch
- Automatic gearbox - does not require any input from the driver to change forward gears under normal driving conditions
- Catalytic converter - uses a catalyst chamber to change the harmful compounds from an engine's emissions into safe gasses, like steam.
- NOVC-HEV - not off-vehicle charging hybrid electric vehicle

Data-mining goals

- Give an overview of the technical specs of registered Estonian M1 and M1G category vehicles by county, brand and fuel type.
- Establish connections between technical aspects and pollution levels of vehicles.
- Visually represent the current percentages of CO2 emission by counties on an Estonian map.
- Build a model using available data to predict CO2 emission levels and represent them visually divided by a car brand.
- Present the results of the data analysis on a poster.

Data-mining success criteria

- The accuracy of the predicting model is higher than 0.8.
- As the result of data analysis connections between different aspects become apparent.

Task 3

The database “Estonian vehicle statuses” is from Estonian open data and is in Estonian, which means that it needs to be translated into English. It contains information about 845952 vehicles, both registered and with suspended registration. Database is last updated on the 1st of January 2022. We do not include any other databases, since the information is adequate and appropriate for our project.

Data covers information about technical specifications about vehicles. Gearbox column covers basic categorical values - manual, automatic and CVT. Fuel type shows which fossil fuels (petrol, diesel) or other energy source (electricity, cng, lpg) car uses. Engine type is mostly identical to the fuel type except for the petrol fuel engines, which have the basic petrol and catalyst engine versions. Engine size is represented as numerical value, the unit of measurement is the cubic centimeter. Hybrid car type aspect is always filled with NOVC-HEV, because this is the only type right-now recognized in the Estonian registration system. Fuel combination is identical to the fuel type except for the hybrid cars, which have different combinations of power sources like electricity and petrol, cng and petrol. Car category has two categorical values (M1, M1G) since the database only covers two of them. Brand column covers almost all car brands existing, model values could be numerical or alphabetical/string values depending on a brand. All weight categories - registered, full and empty, are expressed in kilograms. The color, body name and county have categorical/string values, which all three are very important aspects for our project. Body type column has identical information to the body name column, but the aspects are presented in different ways - body type has two-letter values symbolizing different body names like sedan, universal, coupe etc.

It is inefficient to include the number of axles and quantity since it has the same value for each car. The same problem is with hybrid type column, which always has only one category (NOVC-HEV). It is really important to divide cars by statuses (registered and suspended registration), because our goal is to get an overview of street-legal cars and not registered cars will make statistical conclusions not truthful . It should also be noted that cars registered in a certain county may not actually drive in that location, which makes the results not hundred percent relevant. Some of the brand names have been changed during the years, so it is important to consider it when including brand names in analysis. In the model column there are some values that for example have the rear drive specification, which should be ignored in this case.

Task 4

- TM- team member (columns show time cost of task for every team member in hours)
- We plan on doing some of the tasks together, because it is easier for us to complete the tasks like that since it prevents us from making mistakes and not following previously set goals.

Task	TM1	TM2	TM3	Comments
Selecting data - deleting useless data (columns - andmed seisuga, telgede arv, arv, keretüüp) All other columns are useful for our project.	0.5	0.5	0.5	Python Jupyter Notebook We will make a copy of the original database and delete the columns from that to avoid problems, when we need some of the data during the project that we previously have deleted.
Translating data - since the original data is in Estonian and the project must be performed in English, the translating process is needed	1	1	1	Dictionary.com Google translate This task is time-consuming, because vehicle technical specs are often expressed using specific vocabulary, which is why it is more difficult to translate it.
Cleaning data - making error corrections, changing data to another form if necessary, disassembling/reassembling columns	3	3	3	Reviewing the data in great detail and getting it into shape to start analyzing and processing it
Data analysis and statistics - finding percentages of different technical aspects, building the predicting model, finding correlations between different aspects	13	13	13	Coding part - Python
Visual representation - visually represent the project results obtained on poster	10	10	10	Jupyter Notebook - graphs Other visual representation programs
Presentation (poster session) - PowerPoint and last corrections of poster design	3	3	3	Academic PowerPoint + presentation Submitting poster