Project Report

VEHICLE REGISTRY ANALYSIS

Allan Loo, Karl Rapur
https://github.com/Rapur101/IDS2020Project

Business understanding

Identifying your business goals

Background

Estonia's vehicle registry contains a significant amount of data about transportation within the country, but goes largely unutilized. People's choice of vehicle in a given region can point towards a number of different trends and reasonings behind them, as can the rest of the data found in the registry.

Business goals

By analyzing the vehicle registry of Estonia, we can point out different trends by region and based on other aspects. Such knowledge is useful when determining the quantity of different vehicles to be imported for sale within the country. Our goal is also to provide insight on the longevity of different vehicles and how people's taste in cars has changed over time. As a bonus, we are planning to train classifiers, which will predict car emissions limit, this is how future car models can show if they meet the expectations/predictions.

Business success criteria

We will consider the project a success, if we can uncover the most popular and unpopular brands and models, provide info on vehicle longevity, show data about vehicle popularity per county and show popularity trends over time.

Assessing your situation

Inventory of resources

- Estonia’s M1 vehicle registry data from Republic of Estonia’s Road Administration: arvel_m1-300920.xlsx (47,457 KB)
- Tableau data visualization software
- Jupyter Notebook

- External Python libraries (Numpy, Pandas etc.)
- Researchers personal computer hardware

Requirements, assumptions, and constraints

The project will have to be completed by December 17, 2020 when the results will be presented to the public. Introductive video has to be completed three days earlier. Each team member is required to contribute at least 30 hours of labor into the project. The results of the project will be limited by the amount of data available to the public.

Risks and contingencies

If an event were to occur from the result of which further progress on the project can not be made, either the work is moved to a different location or the progress up to that point is consolidated into something presentable. In case of too few labor hours at early stages, we might add goals.

Terminology

Vehicle category - A vehicle category classifies a land vehicle or trailer for regulatory purposes. (M1 for regular passenger cars)

Manual transmission - Type of transmission where gears are shifted manually by the driver.

Automatic transmission - Type of transmission where the gears are shifted automatically by the gearbox assembly.

CVT transmission - An automatic transmission that can change seamlessly through a continuous range of gear ratios.

Kerb mass - is the total mass of a vehicle with standard equipment and all necessary operating consumables such as motor oil, transmission oil, brake fluid, coolant, air conditioning refrigerant, and sometimes a full tank of fuel, while not loaded with either passengers or cargo.

Car axle - is a central shaft for a rotating wheel or gear.

Without active registration - Vehicles that do not possess a legal right to be driven on public roads. Is removed from the registry if the vehicle is sold to a different county, the owner moves to a different country, the vehicle has no active insurance and technical inspection, the owner has removed the registration for upto 24 months.

Costs and benefits

This is a non-profit student project analysis, there are not any profits expected from the completion of the project and all the costs associated will be covered by the participating researchers. The Republic of Estonia's Road Administration isn't behind this project and won't benefit from it. Main benefits for us are experience, a bigger portfolio and better understanding about Estonian's car market.

Defining your data-mining goals

Data-mining goals

We will deliver a trimmed down dataset, that has all the unnecessary data removed from the dataset. Also we will deliver a number of different graphs visualising the results in a more informative and understandable manner. Also a poster and a video will be made to sum up the most important results in an easy to grasp presentation. As a bonus and extension of labour hours, we are planning to train different emissions classifiers, to set future goals for new cars.

Data-mining success criteria

In order to support the business success criteria, we must deploy different data analysis methodology. Data-mining must provide the necessary data required to fulfill the business success criteria.

Data understanding

Gathering data

Outline of data requirements

We will only require the Estonian Road Administrations M1 vehicle registry info to conduct this project.

Data availability verification

The data used is publicly available on the Estonian Road Administration website and thus pose no issue for availability.

Selection criteria

We will be only looking at vehicles with a first registration date higher than 1996, because registrations prior to that do not have the vehicle transmission type specified. We will be using all the columns except the ones detailing axle count and CO2 emissions, for those are not within the scope of this project.

Data description

The used dataset is the Estonian Road Administrations M1 vehicle registry from 20.09.2020. It includes every registered M1 category vehicles registration status, category, brand, model, body type, year of first registration, colour, engine type, engine output in KW ,CO2 emissions, engine capacity, kerb mass, transmission type, axle count, county of registration, city(if registered in one), vehicle count. In total, there are 805 924 vehicles listed in the data.

There are cases where fields are left empty. Most of the time this is due to the fact that, at the time of first registration the amount of data gathered differed from what is gathered today.

Exploring data

Registration status - Shows whether or not the vehicle in question possesses an active registration. Not possessing one, does not necessarily mean anything about the vehicle's condition, it simply means that it is either temporarily or permanently suspended from legal road use until registration is reactivated.

Category - Shows the vehicle category. In this dataset every vehicle belongs in either M1 or M1G category. The G marks offroad capable vehicles, they need to have more than one driving axle, have atleast one of them be lockable and not weigh over 2-tons.

Brand - The manufacturer's brand the car belongs to. The most numerous brand is Volkswagen with approximately 90 000 vehicles present in the registry. Not far behind are Ford and Audi with 51 000 and 55 000 vehicles registered respectively.
There are some offshoots of brands listed as unique brands, when in reality they should not be listed like that. Data will be cleaned up to reflect that.

Model - Name given to a specific variant of a released vehicle. In most cases contains an accurate model name, but in some other cases simply contains info about the car's engine capacity, whether or not is self made or is left empty entirely.

Body type - Has vehicles listed in one of 15 categories, convertible, hatchback, sedan, coupe, estate, station wagon, bus, ambulance, van, caravan, limousine, armored vehicle, for intended purpose, competition cars.

Year of first registration - Shows the year the vehicle was first registered. The vehicle could have been registered in any country. Registrations start as far back as 1912. We will most likely be only looking at vehicles first registered after 1995, so that the large majority of vehicles analyzed have the same traits listed, older vehicles have relatively small amounts of data about them.

Colour - The colour that vehicle was at the time of registration check up. Almost always accurately reflects the vehicle's current colour as well so there is no reason to place doubt on the accuracy of finding based on this feature.

Engine type - Shows the type of fuel the vehicle runs on and in the case of gasoline also shows whether or not it possesses a catalytic converter. Possible categories: Gasoline, gasoline-catalitic-converter, diesel, CNG, electric, gasoline-hybrid, diesel-hybrid, LPG.

Engine output - Reflects the vehicle's engines maximum output in kilowatts.

Kerb mass - Look for definition under terminology. In the large majority of cases it falls under 2000 kilograms.

Transmission type - Reflects the type of transmission the vehicle uses. Possible entries are automatic, manual, CVT. Most vehicles older than 1995 do not have the transmission type specified and the field is left empty.

Axel count - The number of axles a vehicle possesses. This value is either 3 or 2, large majority 2.

County - Shows the county the vehicle is registered in. Value is the name of one of Estonias 15 different counties.

City - If the vehicle is registered in a city, the name is given otherwise the “Määramata” is written.

Count - Some vehicles registered can have identical features, in that case instead of having duplicate lines, the vehicle count is increased.

Data quality verification

The data qualifies for this project and allows us to reach the goals set. More conclusions could be reached if more features were present in the data, but for this project the data present will suffice. Since the data is publicly available, access is not an issue, also none of the quality concerns are severe enough to significantly hinder the project's completion. Older cars have many missing features, but after trimming we won’t look at old cars, which have these missing values.

The Plan

- Thinking about an idea and project consultations.
    - Asignee: Allan, Karl
    - Estimated completion time: 4h
- Making a report
    - Asignee: Allan, Karl
    - Estimated completion time: 8h
- Data preparation
    - Asignee: Karl
    - Estimated completion time: 6h
    - Comments: Trim the data, and count columns.
- Data-catagorizing:
    - Asignee: Allan, Karl
    - Estimated completion time: 4h
    - Comments: categorize data for future work and analysis.
- Data-mining and analysis
    - Asignee: Allan, Karl
    - Estimated completion time: 16h
    - Comments: Relations of attributes, correlations, frequent pattern mining, computational statistics
- OPTIONAL: Train basic CO2 prediction model
    - Asignee: Allan
    - Estimated completion time: 6h
    - Comments: Have to try different classifiers approaches. Not sure if gonna work.
- Repository cleanup and refactoring
    - Asignee: Allan, Karl
    - Estimated completion time: 2-4h
- Create a 3 minute video
    - Asignee: Karl
    - Estimated completion time: 2h
- Create a visual graphs for poster with Tableau
    - Asignee: Karl
    - Estimated completion time: 5h
    - Comments: Have to learn how to use Tableau at first
- Create a poster visualizing the most important findings
    - Asignee: Allan, Karl
    - Estimated completion time: 12h
    - Comments. Will be done in vector graphics.
- Poster session
    - Asignee: Alla, Karl
    - Estimated completion time: 3h

List of methods and tools:

- Python3
- Tableau data visualization software
- Jupyter Notebook
- Matplotlib library
- Seaborn library
- Numpy library
- Pandas library
- Sklearn libary