Identifying Your Business Goals

Background

This project, part of a university curriculum, focuses on analysing vehicular trends in Estonia's counties. It is designed to enrich students' understanding of data analysis techniques, applied in a real-world context. The automotive sector is a rich field for data exploration, offering insights into environmental impacts, economic trends, and technological adoption.

Business Goals

1. **Detailed Vehicle Analysis:** Gain a deep understanding of the variety of vehicles in each county, including make, model, age, and type.
2. **Fuel Efficiency Exploration:** Assess and compare the fuel efficiency of vehicles across counties, identifying patterns and outliers.
3. **Demographic-Vehicle Dynamics:** Investigate how demographic factors like population density, income levels, and urbanisation influence vehicle characteristics in different counties.
4. **Technical County Classification:** Develop a classification system for counties based on the technical attributes of vehicles, such as emission levels, age, and type.

Business Success Criteria

- Comprehensive data collection and accurate categorization of vehicles by type and features.
- A robust analysis of fuel efficiency trends that could potentially inform environmental policy or consumer behaviour studies.
- Establishment of clear correlations between demographic factors and vehicle types, contributing to socio-economic research.
- Development of a reliable classification model for counties based on vehicular data.

Assessing Your Situation

Inventory of Resources

- **Data Sources:** Access to national vehicle registration databases, demographic statistics from government portals, and environmental data sets.
- **Technical Resources:** Data analysis software provided by the university, cloud storage for data handling, and access to academic databases for research.
- **Human Resources:** Guidance and mentorship from university professors specializing in data science and peer support from fellow students.

Requirements, Assumptions, and Constraints

- **Requirements:** Comprehensive, accurate, and recent data; access to sufficient computational resources; faculty guidance.
- **Assumptions:** The available data accurately reflects current vehicle trends in Estonia; methods used in data analysis are appropriate for the type of data.
- **Constraints:** Limited time frame for the project, potential restrictions on data access due to privacy concerns, and the need to balance this project with other academic responsibilities.

Risks and Contingencies

- **Risks:** Incomplete or outdated data, technical challenges with software or analysis tools, difficulties in interpreting complex data sets.
- **Contingencies:** Regularly updating data sources, seeking additional technical training or support, consulting with faculty advisors on challenging aspects of the project.

Terminology

- Develop a glossary of terms related to vehicle types, demographic indicators, and data analysis methods for clear communication and understanding among team members and when presenting findings.

Costs and Benefits

- **Costs:** Time investment, potential costs for software licences or data access, opportunity costs associated with focusing on this project over other academic or personal pursuits.
- **Benefits:** Practical experience in handling and analysing real-world data sets, enhanced understanding of the automotive sector and its implications, valuable contribution to academic knowledge, potential for this project to inform future studies or policies.

Defining Your Data-Mining Goals

Data-Mining Goals

1. **In-Depth Vehicle Segmentation:** Thoroughly segment and analyze the types of vehicles in Estonian counties, using classification algorithms or statistical methods.
2. **Fuel Efficiency Pattern Identification:** Employ data mining techniques to uncover patterns and anomalies in fuel efficiency across different vehicle types and counties.

3. **Predictive Analysis of Demographics and Vehicle Data:** Use predictive modeling to explore the relationship between demographic factors and vehicle characteristics, potentially employing machine learning techniques.
4. **Advanced County Clustering Based on Vehicle Data:** Utilize clustering algorithms to group counties in novel ways based on their vehicle data, potentially revealing unforeseen insights or trends.

Data-Mining Success Criteria

- High accuracy and precision in vehicle segmentation, with clear, well-defined categories.
- Identification of meaningful and statistically significant fuel efficiency trends.
- Development of robust predictive models with high predictive accuracy and relevance.
- Effective and insightful clustering of counties, providing new perspectives on regional automotive trends.

Gathering data

For our project, we would need data that accurately describes the current condition of cars in Estonia and their locations within the country. In order to assess their condition, we'd require access to their technical inspection information, registration time and location (county-level precision would suffice). Additionally, we'd need details about the appearance of the cars and their technical specifications to compile an overview categorised by county.

We know that such data exists and is available in public information systems, but such information for everyone (non-anonymously) is not publicly disclosed. Therefore, we probably need to work with anonymized data. Since it is known that such data exists, we can assume we'll find it somewhere.

We found that we can use data from the Estonian Open Data portal. These datasets have been released by the Transport Authority (Transpordiamet, did not find a better name in english). The first dataset contains vehicle statuses, and the second one provides technical inspection information. Additionally, we'll utilise the accompanying context table for fault codes. These are Status of vehicles in Estonia and Roadworthiness tests of land vehicles in Estonia.

Describing data

In the data we found, a significant portion of the necessary information is already available. In the dataset of statuses, there are vehicle statuses as of December 1, 2023, clearly reflecting the current state. Additionally, it provides information on whether the car is in use or not (in a legal sense, not practical). It also includes details about registration times and the county where the vehicle was registered. The technical details are minimal there, including basic technical data such as engine and fuel type, engine capacity and power, general appearance descriptors like colour, number of axles, body type, and seating capacity.

Unfortunately, it's not directly possible to match vehicles between the dataset of technical inspections and their statuses. However, it contains registration information, model, and category, which is sufficient for making generalisations. Of course, there's inspection information as well as fault codes in case of issues. It also specifies where the inspection was conducted. Regarding technical inspections, there are 600 thousand rows of data, and for statuses, there are 1.2 million rows, although there are vehicles listed that are no longer in use.

Exploring data

First, I'm reviewing the dataset of statuses as it provides the most information. Most of the data is either categorical, dates, or numerical. The status can be either registered or suspended. The initial registration date for older cars is typically the first day of the respective year, while newer ones have precise registration dates. The gearbox type is categorical, likely including information about non-modern vehicles or unconventional ones like tractors. The fuel type is also categorical, encompassing common fuels (with general names like gasoline, diesel, etc.). The engine type is similar to fuels, but for hybrids, it includes an additional suffix. Engine power and capacity express the respective attributes of the vehicle. The fuel combination field indicates which fuels the vehicle can use. The category denotes the primary use and partly the size of the car. While there are many different categories in the dataset, the project focuses more on M1 and M1G categories, primarily representing passenger cars. The dataset also includes appearance information, mostly categorical. Second dataset is self explanatory and has been described in some sort in previous parts.

Our assumptions at this point are that most of the newer cars in use are some form of hybrids and that the county location has some sort of influence on the technical data of the car.

Verifying data quality

Most of the data is accurate, although there might be input errors in the dates; for instance, a year like 0987 likely didn't have vehicle registrations yet. Some data is missing for older cars because at that time, the publication of such indicators wasn’t mandatory or necessary, such as CO2 emissions. In the dataset of technical inspections, the fault codes indicate issues, but they seem manually written, making categorization harder, if possible at all. Otherwise, both datasets have sufficient accurate data, so there's no need to search for additional information. Therefore, the data integrity is sound.

Planning

Tasks:

1. Find datasets that would be useful **(2h Gregor)**
   Find datasets that would be useful and have meaningful data.
2. Get data and load it into Python **(1h Gregor)**
   Load data into Python and clean it (remove invalid values, etc)

3. Decide what data to use in analysis **(2h Gregor)**
   Remove data that would be not useful and rows that would not have significant info in them (ig vehicles like tractor, combine …, vehicle that have been removed from usage)
4. Explore data and find interesting statistics, and correlations **(24h, 12h each)**
   Try to find interesting statistics and correlations and also make comparisons by county.
5. Visualise findings and general statistics **(8h, 4h each)**
6. Try to create a model (with various method) to classify road vehicles based on county **(10h Markus)**
   Create, test, and evaluate model, to see if it is possible to guess county by seeing some cars in it.
7. Create text and descriptions that would explain findings **(8h, 4h each)**
8. Make it into poster **(6h Gregor)**