VILNIUS UNIVERSITY
FACULTY OF MATHEMATICS AND INFORMATICS
INSTITUTE OF COMPUTER SCIENCE
INFORMATION TECHNOLOGIES STUDY PROGRAM

Project work

Application of graphic methods in data analysis

Done by:
Yurii Dmytruk

Vilnius
2025

Contents

Concepts

Since this work's primary focus is to analyse driving licence-related data, it would be relevant to introduce readers to the driving licence categories and types of vehicles that can be operated by a holder of those categories.

Category	Vehicles allowed
AM	Mopeds and light quadricycles, maximum speed up to 45 km/h.
A1	Light motorcycles (up to 125 cc and less than 11 kW).
A2	Motorcycles (under 35 kW).
A	Heavy motorcycles (no power restrictions).
B1	Vehicles up to 550 kg mass.
B	Passenger vehicles (up to 3,500 kg, maximum eight passengers).
BE	Category B vehicle towing a heavy trailer under 3,500 kg in total.
C	Trucks over 3500 kg.
CE	Trucks over 3500 kg with trailers.
C1	Light trucks (between 3500 kg and 7500 kg).
C1E	Light trucks with trailers (under 12 000 kg in total).
D	Buses with more than 8 passenger seats with a trailer up to 750 kg.
DE	Buses with more than 8 passenger seats with an over 750-kg trailer.
D1	Minibuses with up to 16 passenger seats.
D1E	Minibuses with up to 16 passenger seats with a trailer, up to 12 000 kg total.
T	Trolleybuses

Table 1. Licence category types

Additionally, some technical terminology:

- **R** - open-source software for working with data analysis, visualization and machine learning.
- **CSV** - data storage format widely used in sciences and computing. In .csv files, values are separated on previously agreed delimiter (usually a comma).
- **Shiny** - package for R and Python that allows to build interactive web-based applications.

Introduction

Topic

Demographic and Categorical Analysis of Lithuanian Driver’s licence Holders

Objective

Learn category, gender, and age-related characteristics of Lithuanian drivers.

Tasks

- Investigate the distribution of driving licence categories popularity;
- Analyse the gender distribution across different licence categories;
- Examine trends in drivers’ age at licence acquisition date;
- Select plot types to display obtained information and build them;
- Analyse data and form conclusions;

Data

This paper uses the 1.4 million data-point CSV collection [2] of valid driving licences issued by the Lithuanian State Enterprise *Regitra*. Multidimensional data includes many properties such as licence issue date, categories, and owners’ birth date, gender, country of origin, etc. The dataset was preprocessed to keep relevant columns only and reduce the file size.

1 Work planning

1. Define the project's objective and tasks;
2. Work with literature related to data visualization;
3. Find and preprocess the data;
4. Choose suitable plot types and justify the choice for selected tasks;
5. Implement data plotting;
6. Analyse, find observations and conclude;

2 Execution

2.1 General information regarding the data sample

When last accessed, the sample contained more than 1.4 million records about valid driving licences. According to the Wikipedia web page of Lithuania [4], the current country's population is 2.89 million. Taking that into consideration, we may assume that roughly **50%** of the Lithuanian population is permitted to operate some kind of motor vehicle.

Firstly, the dataset contained loads of encrypted drivers' personal data, so a clean-up was performed to reduce file size, increase processing speed, by omitting unnecessary attributes. That way, only 4 values remained in each row:

- Licence issuance date
- Driver's birth date
- Driver's gender
- Licenced categories (Table 1)

Secondly, all the rows lacking data in any of those attributes (N/A) were dropped. Finally, some inconsistencies were spotted in the data: in some rows the birth dates featured future dates (i.e. 2035-2050 years). Such rows were dropped. Additionally, one more attribute was calculated as described in Analysis of drivers' age at licence acquisition date. After these clean-up stages the loss was negligible (around 100 data points).

2.2 Analysis of category and category to gender distributions

In this section we will take a look on a comparative plot of the number of drivers that hold permits of a certain type. The treemap plot type was chosen for this task. While it would seem that a pie chart is an obvious choice for showcasing categorical data, in this particular situation it was not suitable because according to [1]p. 64, a piechart should be used for displaying *shares* of a small number of categories. Firstly, it is really important to understand that total sum of permits on our chart would not add up to the dataset size. Thus, our plot does not show shares. That happens because the same person is usually permitted to operate several different means of transport (e.g. mopeds and cars - categories $A + B$) and the licence is added to the category X if it permits operating vehicles of type X. Secondly, number of categories is way too big for a piechart to allow labels, which are essential.

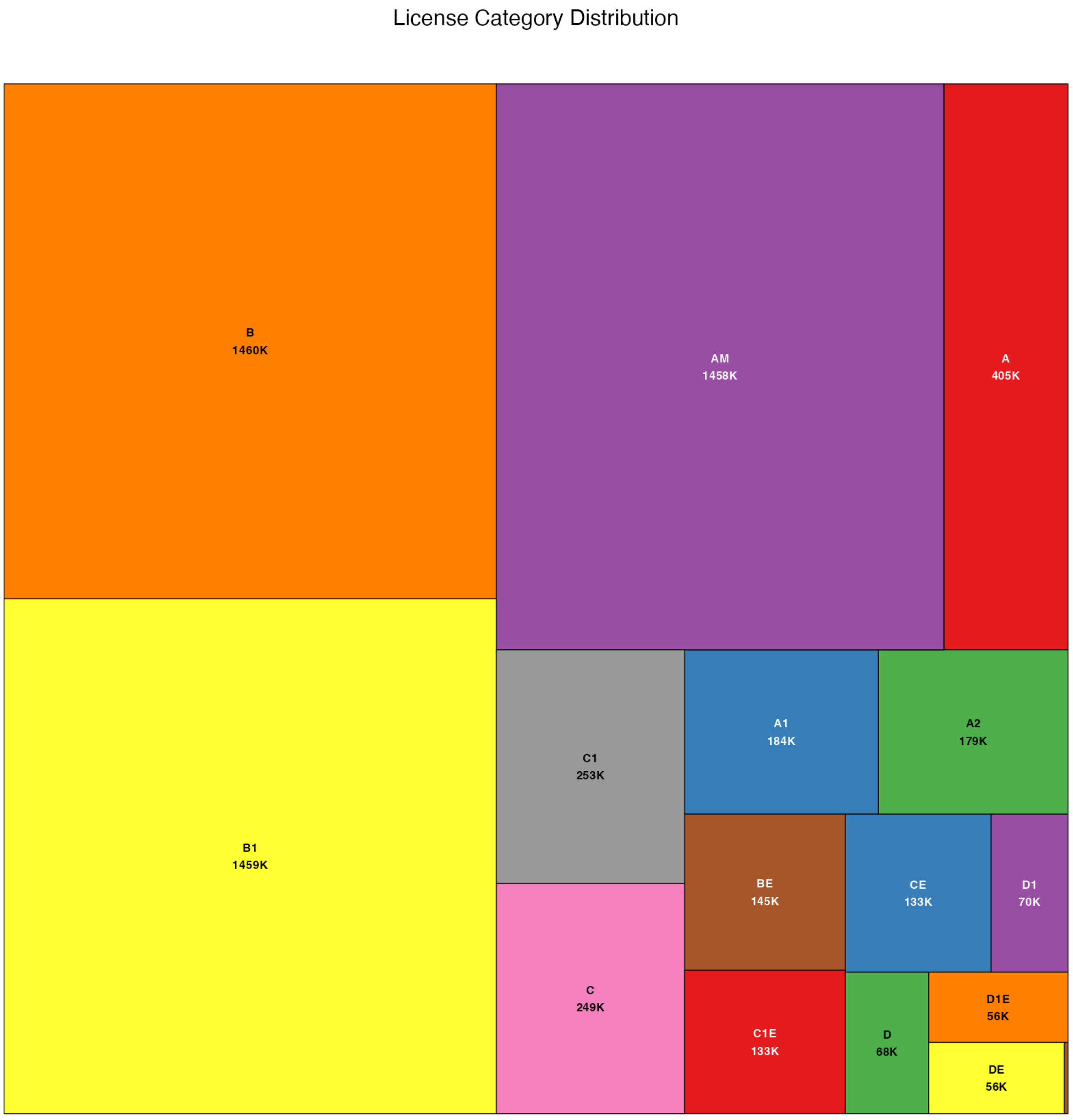


Figure 1. Representation of driving licence category frequencies, thousands

The data in Fig. 1 shows clear prevalence of categories for personal casual use (A-B) over those that are likely to be used in commercial or communal activities (C-D, T). Just under 100% (1.46 mln) of licence holders may operate regular personal vehicles (B), light vehicles (B1), and mopeds (AM). Additionally, high percentages (up to 17%) of C-categories permits point to the fact that trucking is popular in Lithuania. It may suggest that logistics companies often use Lithuania as a host country for their non-EU workers, but it is no more than a calculated guess and requires more data to know for sure.

One more interesting fact the plot showcases - the Lithuania-specific trolleybus (T) category. Its tile is barely visible in the right bottom corner because of the small number of licence holders (appr. 1500), which is exactly 0.1% of all drivers. Considering that only 2 Lithuanian cities (Vilnius and Kaunas) boast a trolleybus system [3] - the data seems coherent.

Now, we shall move on to analysing gender-related patters within the categories. For that, an interactive plot system was built using the **Shiny** package. The interface allows selecting a category in a drop-down and seeing both static treemap (Fig. 1) on the left and interactive category-specific bar chart (Fig. 2) on the right.

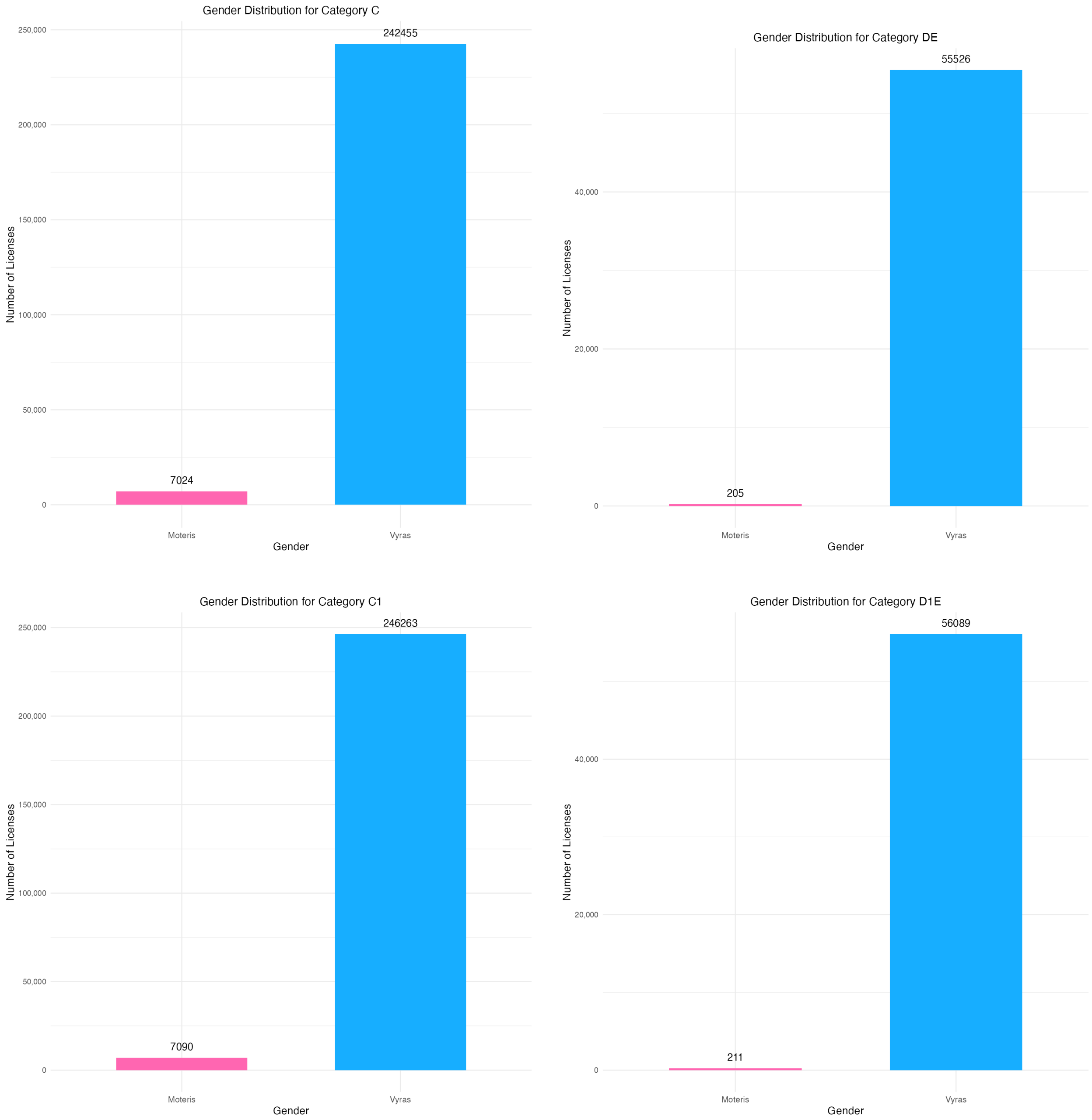
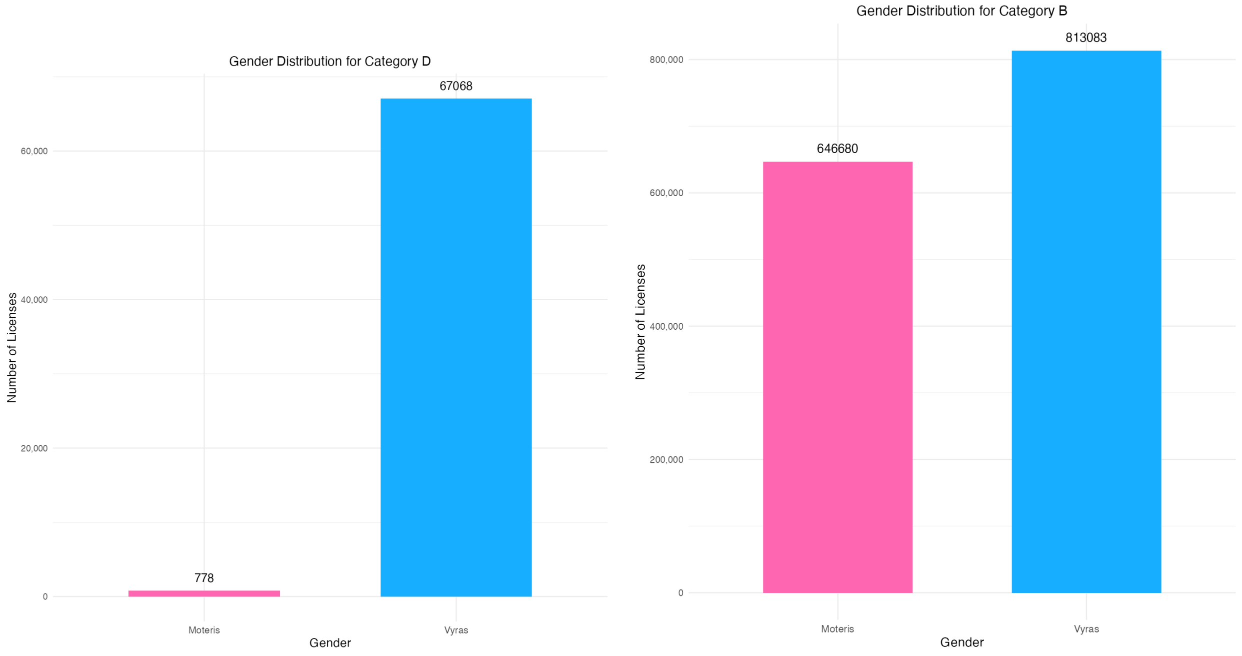


Figure 2. Comparative gender distribution plots for categories C, C1, DE and D1E

The plots reveal some expected and some surprising trends. Firstly, the overall male prevalence was expected. In every single category, male drivers hold more licences than females do. The difference, as expected, exacerbates in truck categories (C) and trailer categories (CE, C1E, DE, D1E), where in some cases females own less than 1% of the licences issued (Fig. 2). On the other hand, bus drivers' licences were a surprise, showing the exact same male-leaning pattern. Judging from the author's subjective experience, the occurrences of female bus drivers are a lot more frequent in Vilnius public transport than the figure 3a suggests. Of course, this fact is easily explained by the existence of other, non-public transport bus companies.

Contrary to the commercial and service-leaning categories, personal vehicles show almost equal distribution where 44.3% of licences within the category belong to females (Fig. 3b).



(a) Bus licences (D) distribution

(b) Passenger vehicle licenses (B) distribution

2.3 Analysis of drivers' age at licence acquisition date

One more interesting characteristic that could be obtained from the dataset is driver's age at licence issuance date. It was calculated as $age = licence\_issuance\_date - driver\_birth\_date$ and the result was cast to years to obtain float value. Afterwards, a histogram and a density curve were built to analyse this aspect. The obvious pitfall of this method is that there is no way to know if the driver got their licence for the first time or if it is a renewed one although the dataset only contains valid licences.

The data of Fig. 4 indicates several interesting points:

1. Around 4.5% of currently valid licences were obtained at the age of 21 years old (mode).
2. Second mode - 31 years old. This makes sense because licence validity interval is exactly 10 years, so drivers who obtained their first licence in 21 are likely to renew it at 31 y.o.
3. The third mode happens almost exactly 10 years after the second - once again a multiple of 10-year expiration date.
4. Some categories (AM) allow obtaining a licence starting from 15 years old, and we can clearly see that roughly 0.5% of drivers obtain their first licence while being minors.

Apart from the facts mentioned above, Chen et al have an important, to my mind, subsection regarding choosing scales (in particularity histogram bin widths) in their Handbook of Data Visualization [1] (p. 65-66). That served as an inspiration to change bin width to 3 and produce figure 5. It reveals one more fact: almost the same percentage of drivers get their licence when

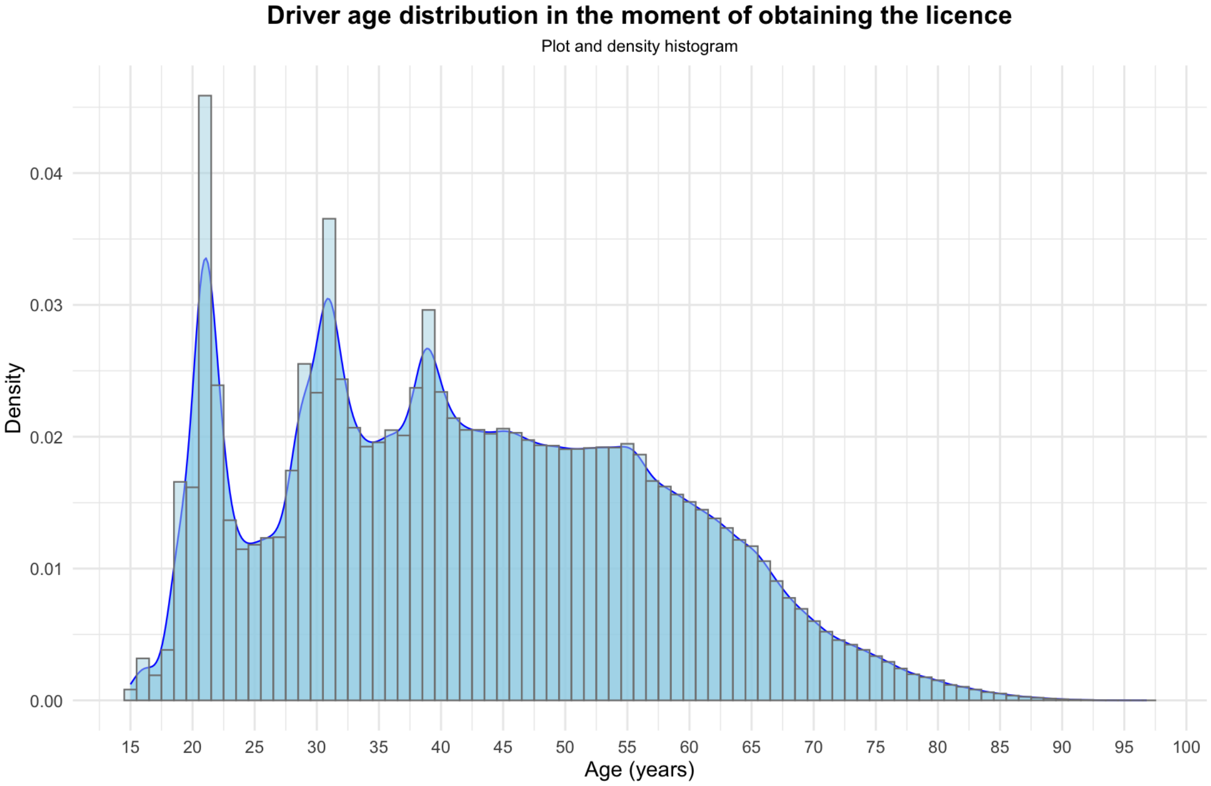


Figure 4. Driver’s age at licence issuance date density, bin = 1

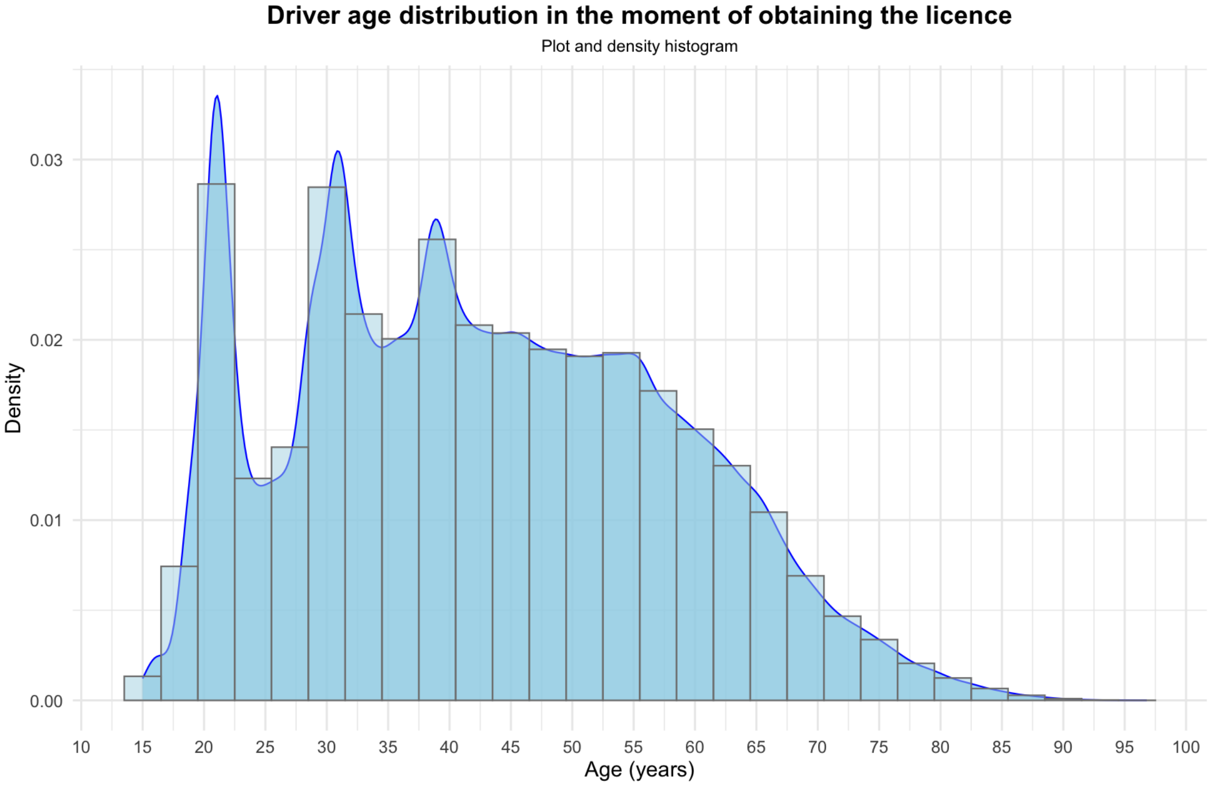


Figure 5. Driver’s age at licence issuance date density, bin = 3

they are 19-21 and 28-31. While on the first plot we could clearly see 31 y.o. as the second mode, aggregating drivers into the 3 year-wide bins helps us see that in fact, those intervals are almost equivalent and almost all of the drivers renew their licence in their late 20-s or early 30-s.

Conclusions

To conclude with this work, let me list results of the data analysis:

1. Approximately 50% of Lithuanian population has a valid driving licence;
2. Almost 100% of those who own a licence may operate AM, B, and B1 category vehicles;
3. Licences with categories A, B, and their subtypes dominate over C, D, and T;
4. One six drivers (17%) is allowed to operate a heavy truck ;
5. Male drivers prevail in all categories, particularly in C and D, where females make up from 0.03% to 2.7% of licences.
6. Female drivers show strong competitiveness in regular vehicles (B) (44.3% of licences);
7. Just under 1500 trolleybus drivers exist in Lithuania with 34.85% of them being females;
8. Mode for obtaining a driver's licence in Lithuania is 21 years old;
9. Almost the same amount of drivers get their licence in the interval of 19-21 and 28-31 years (likely, they are renewing expired 10-year licences);
10. 0.5% of drivers obtain their first licence being minors (e.g. for AM category);

References

[1] Chun-houh Chen, Wolfgang Hrdle, Antony Unwin, Chun-houh Chen, Wolfgang Hrdle, and Antony Unwin. *Handbook of Data Visualization (Springer Handbooks of Computational Statistics)*. Springer-Verlag TELOS, Santa Clara, CA, USA, 1 edition, 2008.

[2] Valstybės duomenų agentūra. Data on valid driving licences issued. https://data.gov.lt/datasets/2934/resource/16492/Pazymejimas/csv, 2025. Accessed: 2025-04-10.

[3] Wikipedia page of trolleybus systems. https://en.wikipedia.org/wiki/List_of_trolleybus_systems, 2025. Accessed: 2025-05-31.

[4] Lithuania wikipedia page. https://en.wikipedia.org/wiki/Lithuania, 2025. Accessed: 2025-05-30.