*Showcase of database analytics on accidents dataset*

Matej Poljak  
*Faculty of Management and Informatics*  
*University of Žilina*Žilina, Slovakia  
poljak@stud.uniza.sk

*Abstract*— The number of cars has been rapidly growing in recent years and so has the number of related accidents. Given the availability of a public resource providing this data, it would be interesting to analyze it from various aspects. During this work we try to quantify the impact of alcohol usage on car accidents, considering factors such as accident frequency, time of day, and associated damage. The results may be distorted because gathered data does not seem to correspond to reality always. The main goal of this paper involves suggesting database analytic tools (specifically aggregate and analytic functions) available on Oracle platform which provide us with an effective and easy extraction of information from vast amounts of data. Eventually, this study demonstrates the power of these tools, as well as the importance of proper data gathering. The results can motivate national institutions to focus more on precise data gathering for matters such as car accidents, as significant information can be gained from such data to implement smarter measures that contribute to a better world.

Keywords— analytics, alcohol usage, car accidents, database

# Introduction

This work is trying to emphasize the importance of gathering and processing data while we live in era of data explosion. Spread of the Internet, social networks, Internet of Things, cloud-computing, etc. This all comes with the vast amount of data generated. But there are also areas that are not still monitored well or at all, but they should. The reason we should care about it is that we can extract from them some interesting and important information which can help us to understand some phenomenon, process and (ideally) based on that create better solutions if possible. As a good example of consistent data gathering is Czech police, which reveal car accidents data every month on their web [1] and are free for everyone. These data are going to be used as an example of data processing using Oracle database analytic tools. Specifically, this work focuses on alcohol impact of car accidents from various points of view. It is possible to observe this while information about alcohol usage is provided within stored data.

# Dataset overview

## Preparing dataset

First, we need to download data from [1] in zip format. We will focus on the year 2024 only – that means we must download twelve structured files. These files contain IDs from many dials. Dials are specified in file ‘Položky formuláře - data’ also available on [1]. Then we can create database tables. Final structure is shown in Fig. 1. There are many references to database tables that serve as dials. For our work, we will use just these tables:

* Table CR\_NEHODY – this is the main table that stores all information about car accidents and IDs to other dial tables. We will refer to this table as ‘acc’ (shortly for accident).
* Table CR\_KRAJE – is a dial for regions in Czech Republic. We will refer to this table as ‘regions’.
* Table CR\_PRITOMNOST\_ALKO – is a dial table of observations of alcohol level in the blood. We will refer to this table as ‘alco’.

Downloaded zip file from [1] contains xls file with records. Records can be imported to the database by creating formulas that construct insert statements.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 1 - Structure of database tables

## Understanding possible cases of alcohol usage

We already know that acc table has a relation to alco table. Now, we could get insight into what alcohol cases were observed when an accident occurred. Results of select query are shown in Fig. 2. We see that recordings can be negative on alcohol presence or positive with different amounts of per mills. There can also be records with refusion of alcohol observation for an unknown reason. There are also cases when it was not observed at all.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 2 – presence of alcohol during an accident

# Basic statistics

With knowledge from the previous paragraph, we can dive into analysis of our dataset.

## All accidents and alco-related accidents

A screenshot of a computer

AI-generated content may be incorrect.

Before we start doing more in-depth analysis, we should get a general overview of accidents in 2024, as well as the number of accidents which were somehow related to alcohol usage. Script with results is shown in Fig. 3. In the first select of UNION statement we query for all accidents in 2024 with any per mill amount of alcohol found in the blood. We sum all cars that were related to those accidents also. In the second select of UNION statement we query for all accidents in 2024. Results of UNION statement are then grouped to one row using analytic function lag (assigns each row its predecessor) and filtering proper row. We see that only 4475 out of 91211 accidents had relation to alcohol usage, what is approximately 5%. 6808 crashed cars were found in accidents with alcohol presence, what is also 5% of all crashed cars (149 208). We might have expected a higher number of accidents related to alcohol usage.

Figure 3 – accidents for each possible state from alco table

## Observations for all possible states of alco table

Let us observe the number of accidents for each case of alco table to get a better idea of what is going on. Results from Fig. 4 reveal that 37 572 accidents, what is 41% of all accidents, were without any observation of alcohol presence what is interesting, and we do not know the real reason for that. In most cases, which is 53% of accidents, alcohol presence was not confirmed as shown in Chart 1. In the query from Fig. 4 we compute number of cases for each alcohol presence in the blood or absence of all three types {N,X,O}. Relative proportions of each state are also computed to understand better dominant cases.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 4 - ratio of alco-accidents to all accidents

Chart 1 - states of alcohol observation in the blood during an accident

Another interesting fact is that if there was an accident with confirmed alcohol, then the probability of measuring at least 1.5 per mills in the blood is greater than 60%. (Chart 2) This assumption was taken by extracting all result rows with positive presence of alcohol in the blood.

Chart 2 - amount of per mills of alcohol in the blood

# Based on time analysis

Now that we understand work with these database tables, we can conduct wider analysis. We will start with time analysis of car accidents for the year 2024.

## Monthly alcohol in-accident presence

In the previous chapter, we have found out that majority of accidents end up with any alcohol presence or it was not observed. We should investigate whether this behavior was the same throughout the entire year or if it has been changing. We are going to focus just on four alco-type states: alco present, not present, not observed, refused. We start by grouping records by extracting the month of accident occurrence. While we want to pivot counts of wanted alco-types states for each month, we can create sum expressions

with CASE clauses to count only cases when an accident contains desired alco-type state. Besides that, we compute overall counts of cases for each month to compute relative proportions of alco-types later. The whole script is shown in Fig. 5. Results are visualized from two perspectives in Chart 3 and 4. From Chart 3 we can see that proportions are similar throughout the entire year. Chart 4 suggests that most alco-present accidents were recorded in the summer months July and August. To confirm this observation and see more detailed numbers, we can look at Chart 5, which visualizes numbers of column alco\_yes from script in Fig.5. We can assume that this could be related to the time of vacation when people are more prone to drink alcohol and forget to be responsible.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 5 - monthly ratios of alcohol presence in accidents

Chart 3 - Monthly ratio of states of alcohol observations during accidents (percentual)

Chart 4 - Monthly ratio of states of alcohol observations during accidents (amounts)

Chart 7 - Number of accidents with found alcohol during months of 2024

## Amount of accidents based on the hour of day

Next, we can take a closer look at which hours of the day are most significant regarding accident frequency. We must group records by hour of the day. This can be done using EXTRACT function where we define that we want to get hour of day (script is shown on Fig. 5. We can see that most frequent hours are from 18 to 20 PM when people tend to meet in their free time.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 7- Amount of accidents for each hour of day

Chart 5 - Amount of accidents for each hour of day in 2024

## Amount of accidents for each day in the week

Another important observation could be to see in which days of the week most of accidents occur.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 6 -

# Optimization of queries

While we work whole time with joining acc table with alco table, it could be desirable to create index on foreign key, which by default does not exist. It could be a bitmapping index because the number of states is finite, more specifically, there are only 8 states.

Another optimization could be creating b-tree index on timestamp column in acc table because we use this column in where clause all the time.

##### References

1. Policie České republiky, "Statistika nehodovosti," \*Policie České republiky\*, 2024. [Online]. Available: https://policie.gov.cz/clanek/statistika-nehodovosti-900835.aspx. [Accessed: July 5, 2025].