FAKULTA RIADENIA A INFORMATIKY

ŽILINSKÁ UNIVERZITA V ŽILINE

ANALÝZA DÁT DOPRAVNÝCH NEHÔD

SEMESTRÁLNA PRÁCA Z PREDMETU DATABÁZOVÁ ANALYTIKA A VÝKONNOSŤ

Autor: **Bc. Matej Poljak**

Učiteľ: **doc. Ing. Michal Kvet, PhD.**

Akademický rok: **2024/2025**

Obsah

[Cieľ semestrálnej práce 3](#_Toc191731537)

[Popis dát 3](#_Toc191731538)

[Transformácia dát do databázy 3](#_Toc191731539)

[Databázová analytika 5](#_Toc191731540)

[Prehľad o počtoch dát 5](#_Toc191731541)

[TOP mesiac/e pre každý kraj podľa počtu umrtí za rok 2024 6](#_Toc191731542)

[N-tý mesiac s najväščím počtom nehôd 9](#_Toc191731543)

[Alkohol a dopravné nehody 11](#_Toc191731544)

[Optimalizácia výkonnosti 16](#_Toc191731545)

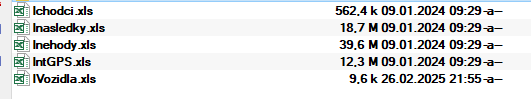
[Záver 16](#_Toc191731546)

# Cieľ semestrálnej práce

Cieľom práce bolo získať dáta, nad ktorými následne vykonáme analýzu. Touto analýzou zistíme zaujímavé informácie vyplývajúce z dát a tak preukážeme schopnosť analytického pohľadu na ne. Na to sme mali využiť analytické a agregačné funkcie, ktoré platforma Oracle poskytuje. Okrem toho sme sa mali pokúsiť optimalizovať rýchlosť dopytu dát s využitím indexov.

# Popis dát

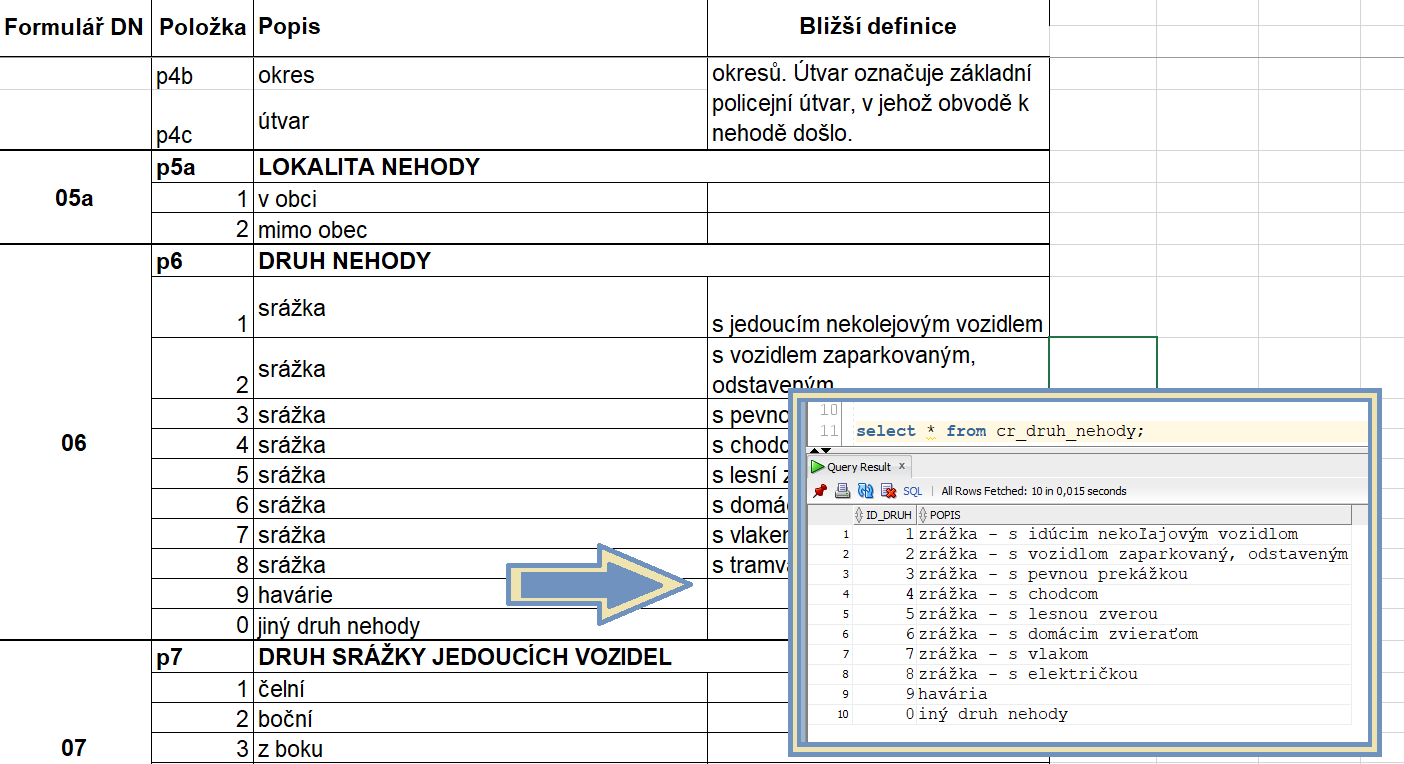
Na prácu sme si zvolili údaje o dopravných nehodách v Českej republike, ktoré poskytuje česká polícia na mesačnej báze. Dáta môžeme získať zo stránky [*https://policie.gov.cz/clanek/statistika-nehodovosti-900835.aspx*](https://policie.gov.cz/clanek/statistika-nehodovosti-900835.aspx), kde si zvolíme rok, z ktorého dáta požadujeme. Stiahneme RAR zložku, ktorá obsahuje súbory s príponou *xls*. Súbory sú rozdelené na záznamy o nehodách, o vozidlách spojených s nehodou, o chodcoch spojených s nehodou a záznamy o následkoch dopravných nehôd (obrázok 1). My si vyberieme pre ďalšie spracovanie dáta o nehodách všeobecne a dáta o vozidlách, čiže dva typy *xls* súborov, pričom použijeme dáta z 3 rôznych rokov: 2023, 2024, 2025.



Obrázok 1 – obsah stiahnutej RAR zložky

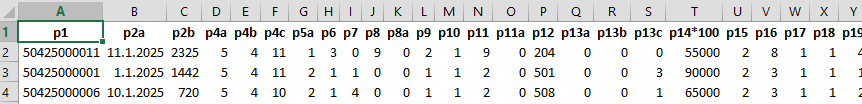
# Transformácia dát do databázy

Na stránke, uvedenej v predchádzajúcej kapitole, sme si stiahli aj dokument *Položky\_formuláře\_WEB.xlsx* (príklad zobrazený na obrázku 2), ktorý obsahuje význam a hodnoty pre jednotlivé stĺpce, ktorými sú naplnené záznamy o nehodách a vozidlách zúčastnených pri nehode. Na základe týchto popisov si pre každý referencujúci stĺpec vytvoríme databázovú tabuľku, ktorá obsahuje významy jednotlivých ID hodnôt, na ktoré sa záznamy o nehodách a vozidlách odkazujú.



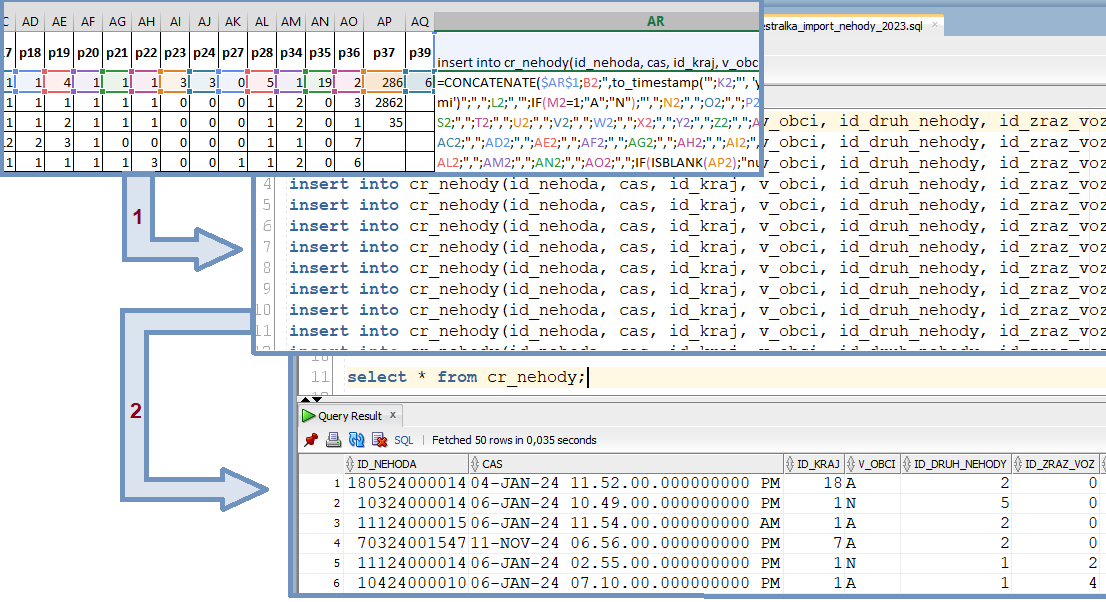
Obrázok 2 – popis hodnôt stĺpcov, na ktoré sa záznamy referencujú

Potom som si vytvoril databázovú tabuľky *cr\_nehody* a *cr\_vozidla* podľa štruktúry jedného riadku súboru *lnehody.xls*, resp. *lvozidla.xls*. Takýto záznam je ukázaný na obrázku 3.



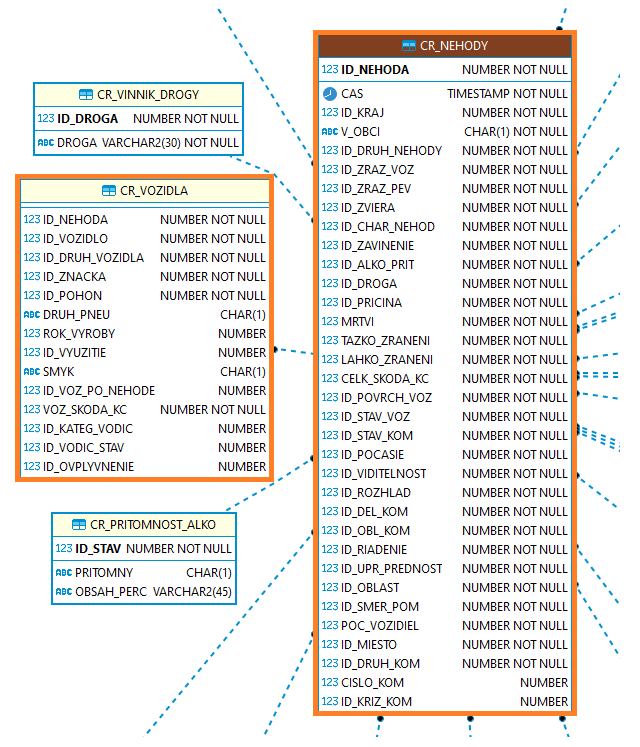
Obrázok 3 – ukážka záznamov o nehodách v xls súbore

Následne som mohol naplniť tabuľky o nehodách a vozidlách hodnotami zo súborov. Najprv som si napísal príkaz, ktorý mi v *xls* súbore vygeneroval sql príkaz *insert*. Všetky vytvorené sql príkazy som si uložil do súboru a potom som spustil vykonanie skriptu v nástroji *SqlDeveloper*. Tento postup je znázornený na obrázku 4.



Obrázok 4 – transformácia záznamov do databázy

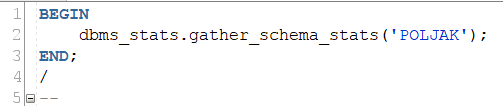
Pre prácu teda budeme využívať (okrem všetkých tabuliek s popismi) databázovú reprezentáciu záznamov zodpovedajúcu obrázku 5.



Obrázok 5 – dátový model najpodstatnejších tabuliek zvýraznených oranžovým rámikom

# Databázová analytika

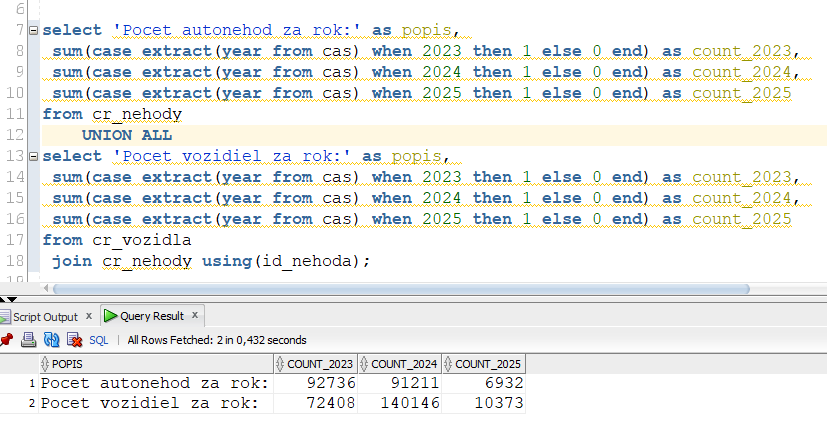
V tejto časti si ukážeme zopár zaujímavých informácií, ktoré sme získali databázovým príkazom *select* spolu s využitím analytických a agregačných funkcií. Pred analýzou výkonnosti si najprv nad všetkými tabuľkami, ktoré prislúchajú k schéme „*POLJAK*“ pregenerujeme štatistiky, pretože sme vložili veľké množstvo nových dát (obrázok 6).



Obrázok 6 – pregenerovanie štatistík pre aktualizáciu kvôli novým tabuľkám

## Prehľad o počtoch dát

Na úvod si pre lepšiu predstavu zistíme najzákladnejšiu štatistiku a to počty záznamov o nehodách a vozidlách za jednotlivé roky. Túto informáciu vieme pekne zobraziť pomocou pivotovej transformácie dosiahnutej použitím agregačnej funkcie *sum*, ktorá pripočíta do stĺpca záznam prave vtedy, keď spĺňa kritérium príslušnosti roku, pre ktorý je daný stĺpec určený. Stĺpec *cas* je *NOT NULL*, preto nemusíme ošetrovať podmienku s chýbajúcou hodnotou. Takéto dopyty sme vytvorili nad oboma tabuľkami nehôd aj vozidiel a následne sme ich zjednotili pre ucelenú informáciu v jednom výpise. Skript aj výpis je zobrazený na obrázku 7.



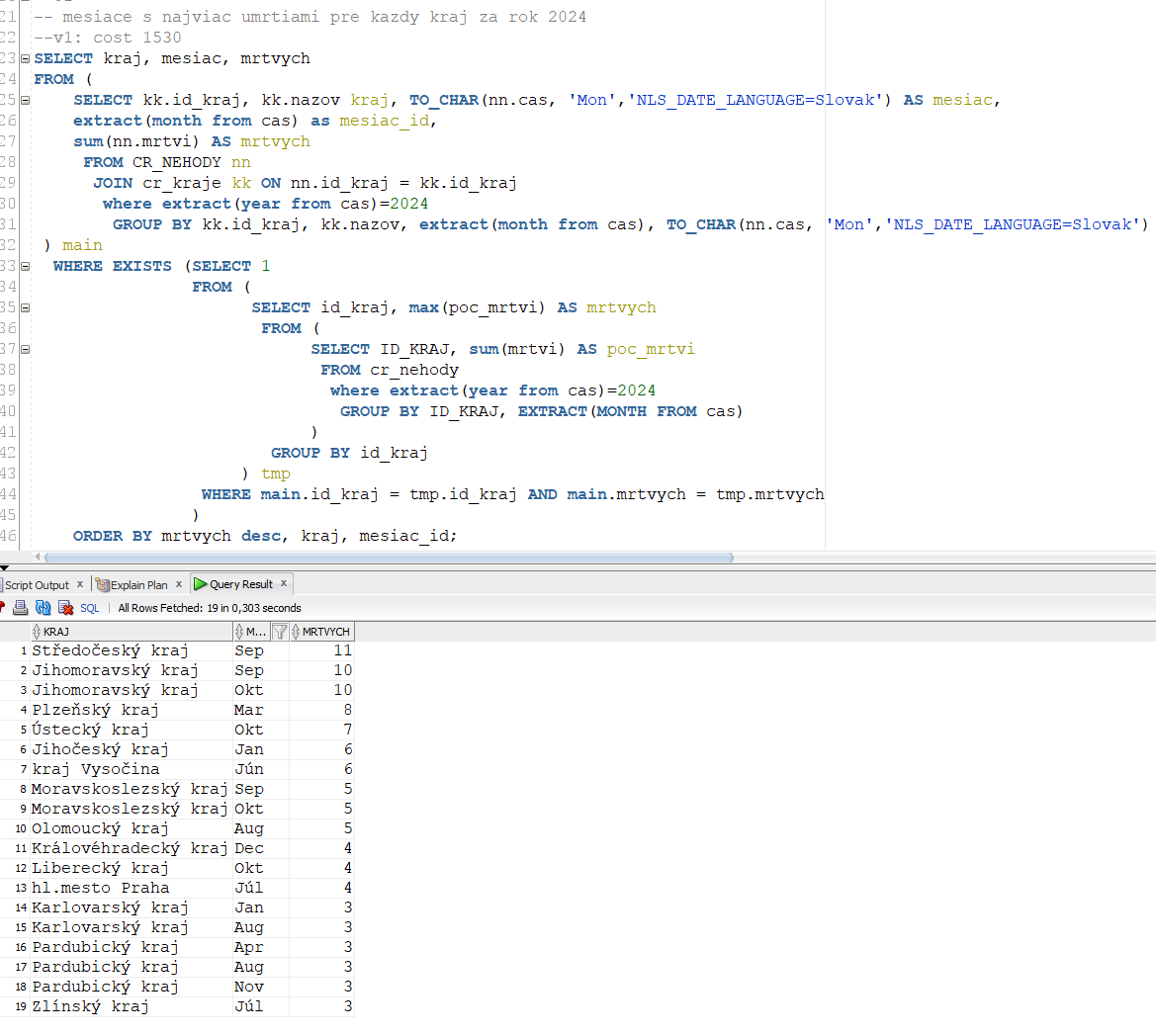
Obrázok 7 – počty nehôd a vozidel za jednotlivé evidované roky

Rok 2025 obsahuje pomerovo iba zlomok dát, pretože sme doteraz mali prístup len k záznamom za mesiac Január. Za rok 2023 je počet záznamov o vozidlách tiež zredukovaný – vybrali sme len necelú polovicu záznamov pre ukážku dát, ktoré sú vzorkou z tohto roku.

## TOP mesiac/e pre každý kraj podľa počtu umrtí za rok 2024

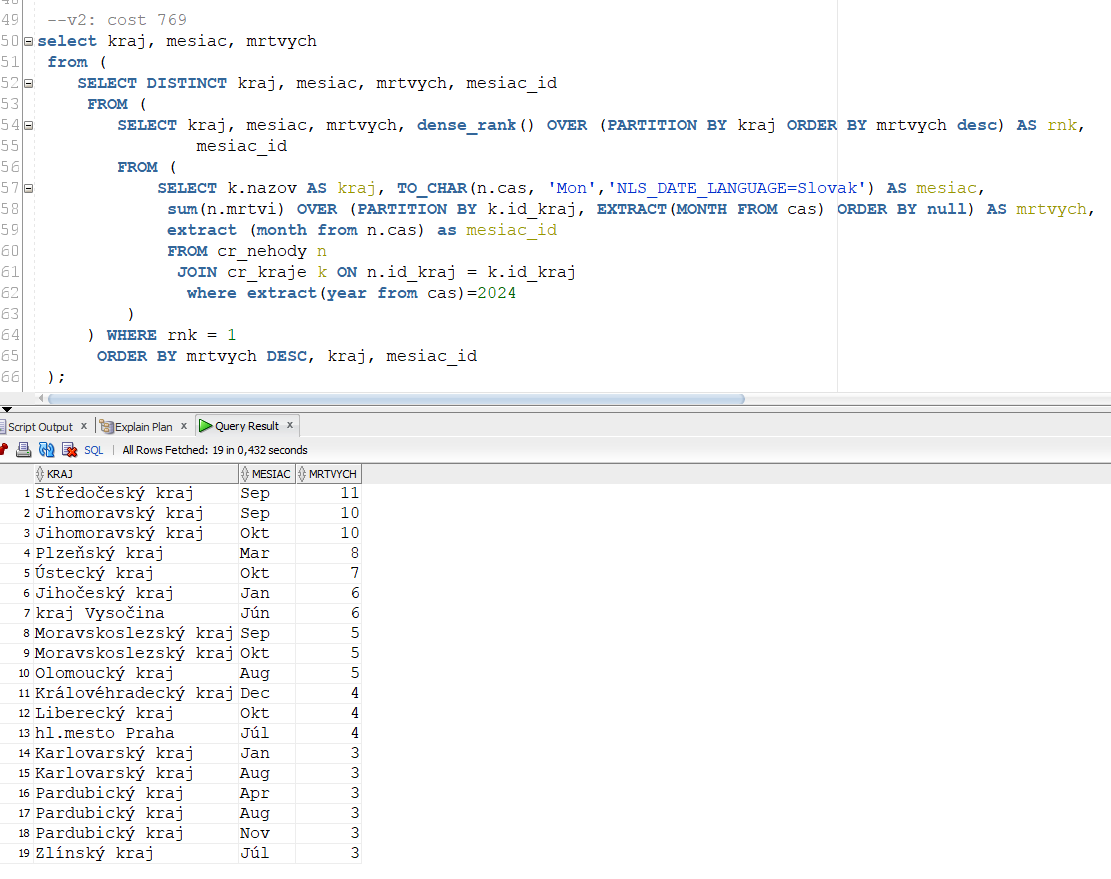
Prvou informáciou, ktorá nás môže zaujímať je koľko a v akých mesiacoch pre konkrétny kraj Českej republiky (ďalej ČR) bol zaznamenaný najvyšší počet úmrtí spôsobených dopravnou nehodou za vymedzené obdobie - rok 2024. Ukážeme si tri varianty ako napísať takýto dopyt na dáta a porovnáme ich náklady.

**Variant 1 – využitie vnoreného príkazu *select* (obrázok 8)**



Obrázok 8 – najviac mŕtvych pre kraj a mesiac za rok 2024 (variant 1)

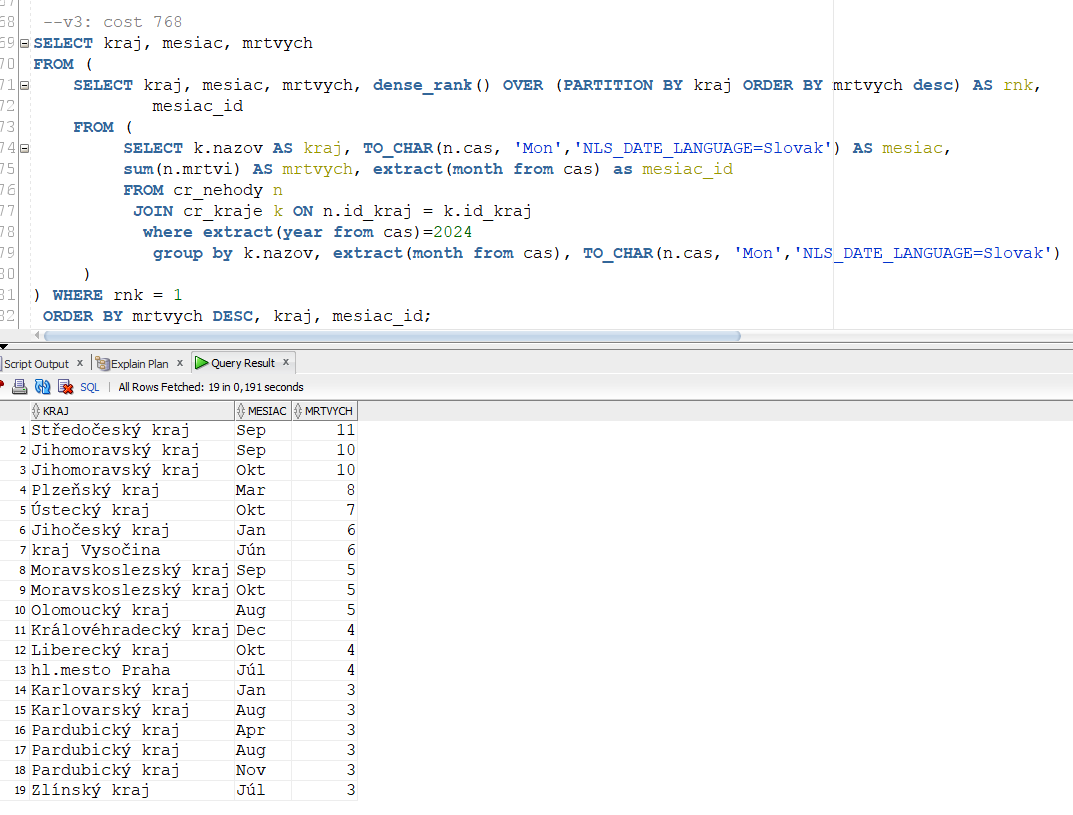
Vnorený *select*, ktorého výsledky označíme aliasom *tmp* vytvorí pre záznamy z tabuľky *cr\_nehody* skupiny podľa mesiaca a kraja, ku ktorým záznamy prislúchajú. Samozrejme, pred samotným zatriedením odfiltrujeme záznamy, ktoré sú mimo roku 2024. Pre skupiny potom získame počty záznamov, ku ktorým si zapamätáme identifikátor ich skupiny – stačí kraj. Následne pre každý kraj vyberieme iba záznam s maximálnym počtom nehôd. Výsledky tohto dopytu budeme vyhľadávať v druhom selecte označenom aliasom *main*, kde si budeme pamätať kraj, jeho názov, mesiac a počet mŕtvych dôsledkom dopravnej nehody. Ak sa takýto záznam zhoduje so záznamom z vnoreného príkazu *select tmp*, ktorý obsahuje maximálne počty, potom tento záznam budeme vypisovať ako výsledok nášho hľadania. Nevýhodou tohto riešenia je, že vnorený príkaz *select tmp* sa musel vykonať nad každým záznamom pre *select* *main*.

**Variant 2 – využitie analytických funkcií (obrázok 9)**

Obrázok 9 - najviac mŕtvych pre kraj a mesiac za rok 2024 (variant 2)

V druhom variante využijeme namiesto vnoreného príkazu *select* analytickú funkciu *sum*, v ktorej si spočítame všetkých mŕtvych pre daný kraj a mesiac, čo definujeme v klauzule *PARTITION BY*. Keďže ide o analytickú funkciu, počet záznamov sa neredukuje, iba každému záznamu podľa jeho skupiny priradí vypočítané číslo. Každý záznam v skupine dostane rovnakú hodnotu sumy, pretože nerobíme tzv. rolling sum, lebo podľa ničoho netriedime (*order by null*). Takto získané hodnoty ohodnotíme ďalšou analytickou funkciou *dense\_rank,* ktorá priradí každému záznamu v rámci skupiny (definovanej krajom) ohodnotenie podľa počtu mŕtvych s TOP ohodnotením pre záznamy s najvyšším počtom mŕtvych. Následne len vyberiem tie záznamy, ktoré dostali ohodnotenie 1 – „najlepšie“, to znamená, najvyšší počet mŕtvych pre každý kraj. Náklady v porovnaní s variantom 1 sú polovičné, pretože sme nemuseli zakaždým vyhodnocovať vnorený príkaz *select*, ale len raz sme si ku každému záznamu priradili jeho ohodnotenie v rámci skupiny na základe počtu.

**Variant 3 – využitie kombinácie agregačných a analytických funkcií (obrázok 10)**



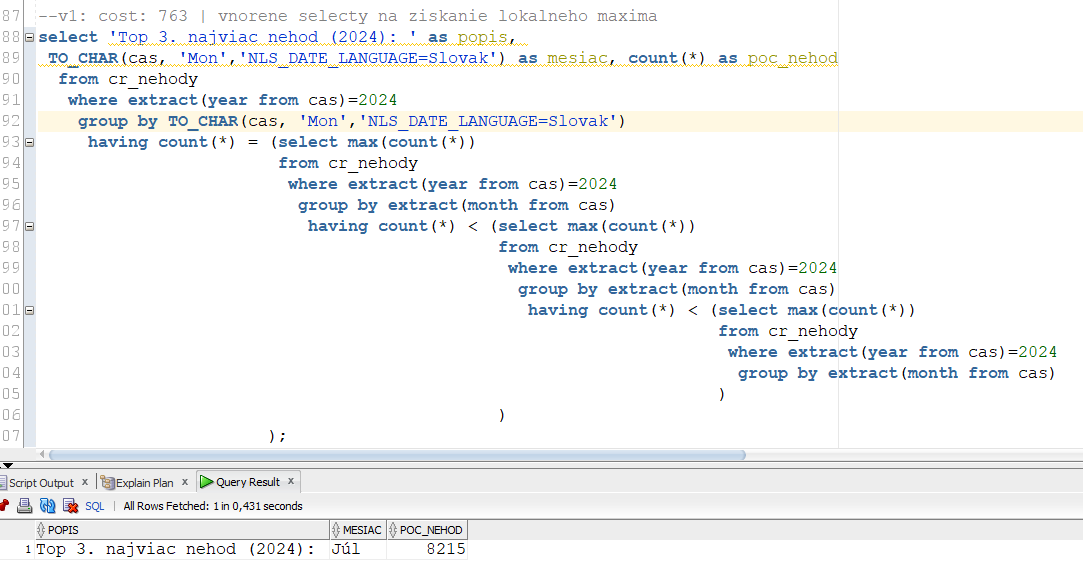
Obrázok 10 - najviac mŕtvych pre kraj a mesiac za rok 2024 (variant 3)

Tretí variant sa od druhého líši iba v spôsobe spočítania počtu mŕtvych pre skupiny definované krajom a mesiacom. Na to sme použili namiesto analytickej funkcie agregačnú funkciu *sum* tak, že sme použili klauzulu *GROUP BY*. Vďaka tomu sme spočítali sumy pre skupiny a tak sme aj, na rozdiel od analytickej funkcie, zredukovali počet záznamov tak, že sme dostali práve jeden záznam pre každú skupinu. Následne sme, podobne ako vo variante 2, iba ohodnotili počty a vybrali pre každý kraj mesiace s najvyšším počtom úmrtí. Porovnanie nákladov na vykonanie variantu 2 a variantu 3 sa javia byť skoro rovnaké (769 verzus 768, t.j. pri variante 3 sme ušetrili 1 jednotku nákladov).

## N-tý mesiac s najväščím počtom nehôd

Ďalej by nás mohlo zaujímať, ktoré 3 mesiace bolo zaznamenaných najviac nehôd a koľko ich presne bolo. Pre prvých *n* mesiacov by bol obtiažnejší výpis bez použitia analytickej funkcie. Síce si taký príklad neskôr ukážeme, ale zamerajme sa teraz len na výpis tretieho mesiaca v poradí s počtom najviac nehôd za rok 2024.

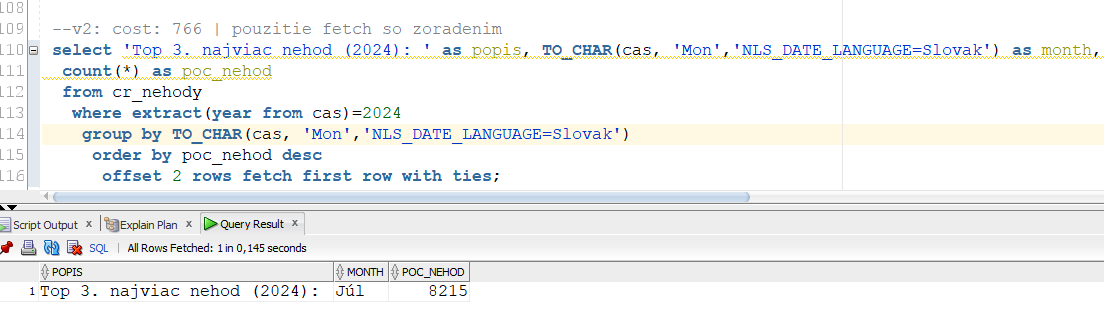
**Variant 1 – využitie vnoreného príkazu *select* (obrázok 11)**



Obrázok 11 - TOP 3. mesiac s najviac nehodami v 2024 (variant 1)

Prvý variant je postavený na vnorených príkazoch *select*, ktoré získajú maximum pri splnení podmienky roku 2024 a zároveň maximum nesmie prekročiť hranicu definovanú vnorenejším príkazom *select*. Takto si vytvoríme toľko vnorení, ktorú n-tú najväčšiu hodnotu v poradí potrebujeme, pričom najvnorenejší *select* predstavuje globálne maximum. Najmenej vnorený *select* obsahuje našu hľadanú n-tú hodnotu, na ktorú sa pozeráme z vonkajšieho príkazu *select*. Toto porovnávanie počtov na najvrchnejšej úrovni nerobíme kvôli získaniu daného počtu, pretože ten sme získali už prvým vnoreným príkazom *select*. Robíme to, aby sme zistili, pre ktorý mesiac je tento počet priradený a tak ho mohli spolu s počtom vypísať.

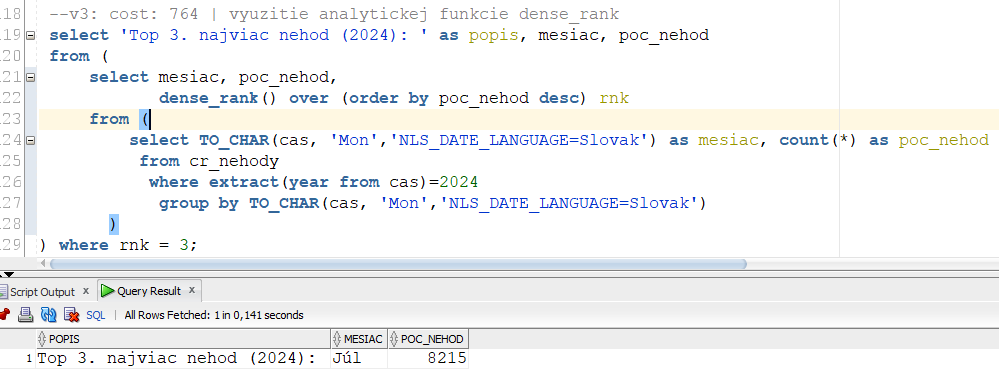
**Variant 2 – využitie klauzuly *fetch* (obrázok 12)**



Obrázok 12 - TOP 3. mesiac s najviac nehodami v 2024 (variant 2)

V druhom variante si pre každý mesiac roku 2024 zistíme počet nehôd a následne je dôležité zoradiť výsledky príkazu *select* vzostupne podľa počtu nehôd. Potom využijeme klauzulu *offset*, pomocou ktorej preskočíme prvé dva najlepšie mesiace a zoberiem s využitím *fetch first row with ties* iba riadky, ktoré podľa počtu predstavujú náš hľadaný top 3. mesiac. Toto riešenie ale nemusí byť správne, pokiaľ napríklad na druhom mieste sme mali dva mesiace. Vybrali by sme teda jeden záznam z dvoch, ktoré prislúchajú k druhému mesiacu a nedozvedeli by sme sa, že sme sa dopustili chyby. Náklady sa zhoršili voči prvému variantu iba o 3 jednotky, takže variant 1 a 2 sú podobne výkonné.

**Variant 3 – využitie analytickej funkcie *dense\_rank* (obrázok 13)**



Obrázok 13 - TOP 3. mesiac s najviac nehodami v 2024 (variant 3)

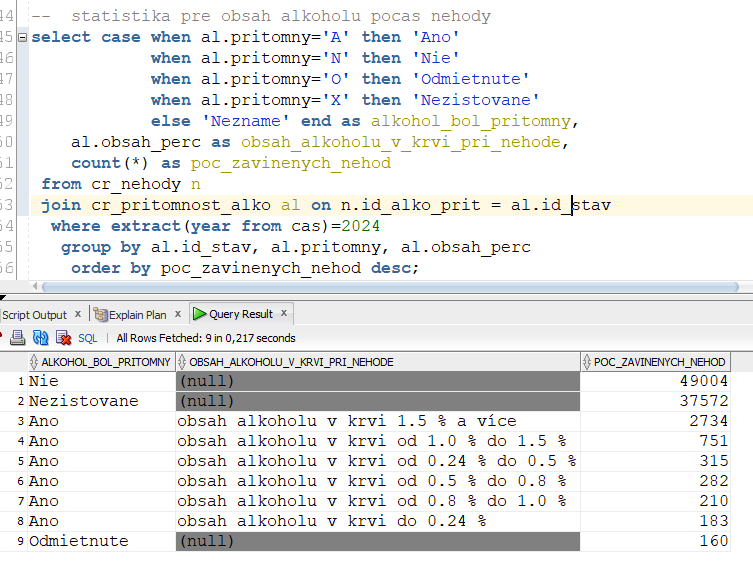
V poslednom variante si pomocou agregačnej funkcie *count* pre každý mesiac spočítame počet záznamov a následne tieto počty zostupne ohodnotíme pomocou analytickej funkcie *dense\_rank* a vyberieme záznamy s ohodnotením *3*. Výhodou tohoto variantu oproti prvým dvom je, že malou zmenou môžeme získať ľubovoľné *n*-té umiestnenie, resp. ľubovoľný interval za sebou idúcich umiestnení. Zmenili by sme iba podmienku, ktoré ohodnotenia chceme získať (t.j. aké čísla môže obsahovať stĺpec *rnk*). Čo sa týka nákladov, tie sa nezmenili (náklady sa zhoršili len o 1 jednotku oproti variantu 1), ale silno sa nám zvýšila miera možnej modifikácie výberov na základe umiestnenia, ktoré je výstupom analytickej funkcie.

## Alkohol a dopravné nehody

Niektoré dopravné nehody bývajú zbytočne zapríčinené požitím alkoholu pred jazdou, v dôsledku čoho sa schopnosti účastníkov cestnej komunikácie byť pozorný výrazne znižujú.

**Ročné početnosti podľa percentuálnej prítomnosti alkoholu v krvi pre rok 2024**

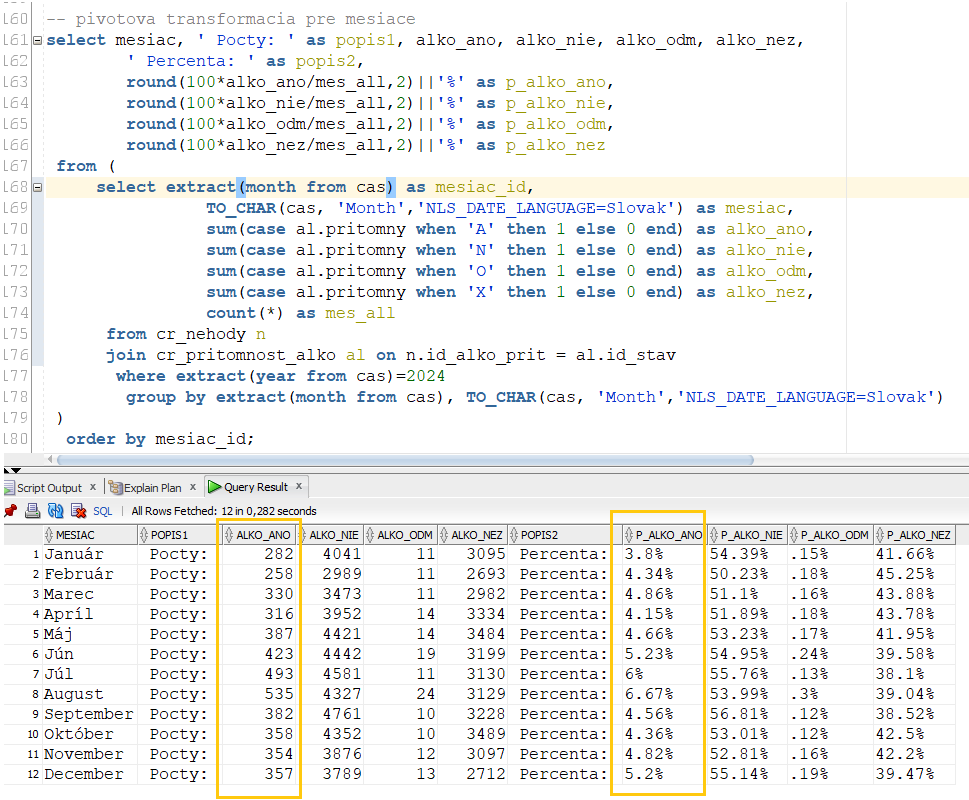
Ako prvé by nás mohlo zaujímať pri koľkých nehodách bola zistená prítomnosť alkoholu a v akom veľkom množstve. Pre zistenie sme vypočítali agregačnou funkciou *count* počty podľa druhov prítomnosti alkoholu v krvi pri nehodách, ako je to zobrazené na obrázku 14 spolu s výsledkami. Z výsledkov môžeme vyčítať, že okolo 50% prípadov alkohol nebol prítomný pri nehode. Zaujímavé je, že pri vyše 40% nehôd sa nezisťovala prítomnosť alkoholu v krvi. V približne 5% nehôd bola potvrdená prítomnosť alkoholu, pričom ak bola prítomnosť potvrdená, potom je vyše 60% pravdepodobnosť, že obsah alkoholu v krvi je viac ako 1.5%.



Obrázok 14 – ročné počty nehôd podľa prítomnosti alkoholu v krvi

**Mesačné početnosti podľa percentuálnej prítomnosti alkoholu v krvi pre rok 2024**

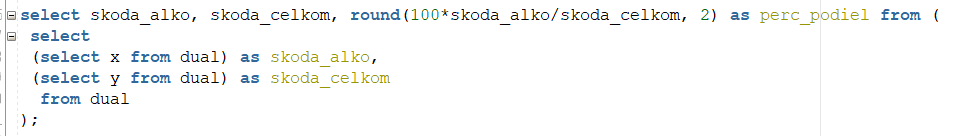
Okrem početností percentuálnej prítomnosti alkoholu v krvi ročnej nás môžu zaujímať aj mesačné početnosti alkoholu v krvi, ako sa celkovo počas roka vyvíjala. Pre lepšie pochopenie uvedieme aj relatívne percentuálne početnosti podielu zisteného alkoholu v krvi pre nehody v rámci mesiaca. Teraz už budeme abstrahovať od miery alkoholu v krvi a bude nás len zaujímať či alkohol bol prítomný, alebo nie. Výsledky spolu so skriptom sú uvedené na obrázku 15. Je vidno, že najviac nehôd so zisteným alkoholom v krvi sa zaznamenal v letných mesiacoch. Môže to byť spojené s tým, že v lete chodia ľudia väčšinou na dovolenky a viacej ľudí berie zodpovednosť vtedy menej vážne. Na konci roka v decembri môžeme tiež vidieť nárast prípadov, čo môže byť opäť istým spôsobom spojené s časom sviatkov.



Obrázok 15 – mesačné počty nehôd podľa prítomnosti alkoholu v krvi

**Percentuálny podiel škôd nehôd s prítomnosťou alkoholu v TOP 1000 najvyššie ocenených škodách**

Môžeme predpokladať, že nehody s prítomnosťou alkoholu môžu mať veľmi katastrofálne následky, preto nás bude zaujímať všeobecné ohodnotenie TOP 1000 najnákladnejších nehôd, čo je približne 1% všetkých nehôd za rok 2024. Z nich identifikujeme všetky, ktoré boli s prítomnosťou alkoholu a ich ohodnotenie dáme do pomeru s celkovými škodami TOP 1000 nehôd, ako to vysvetľuje šablóna príkazu *select* na obrázku 16.

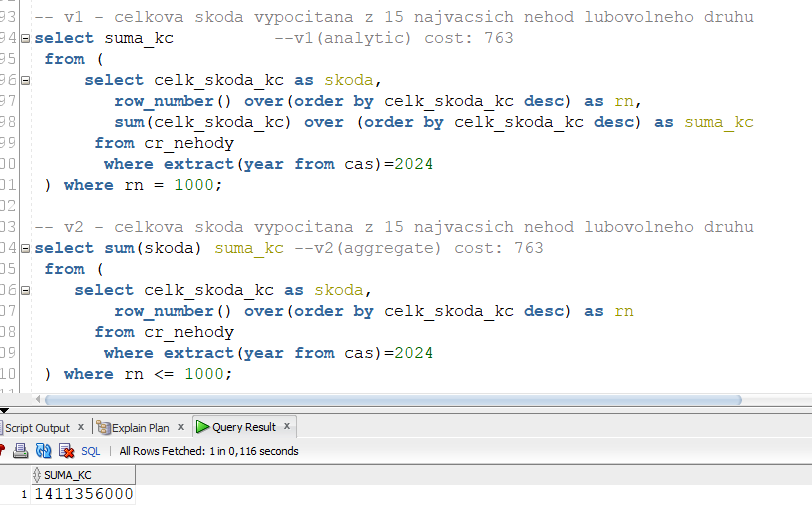


Obrázok 16 – abstrakcia výpočtu percentuálneho podielu pre hľadanú skutočnosť v jazyku SQL

Rozoberme si postupne príkaz *select* pre čitateľ a menovateľ zvlášť.

1. Menovateľ (obrázok 17) – celková škoda TOP 1000 nehôd s najväčšími vyčíslenými škodam v peňažnej mene Kč. Potrebujeme si zoradiť nehody za rok 2024 zostupne podľa veľkosti škody v Kč a následne vybrať prvých 1000 záznamov. Keďže nezohľadňujeme koľko záznamov môže mať teoreticky rovnakú hodnotu škody (nehodnotíme princípom hodnotenia olympijských hier), ale stačí nám jednoducho 1000 ľubovoľných najnákladnejších nehôd, môžeme použiť analytickú funkciu *row\_number* na získanie ohodnotenia pre každý záznam podľa jeho vyčíslenej škody. Tu máme 2 spôsoby sčítania prvých 1000 záznamov s najlepším ohodnotením.
   1. Prvý spôsob je pomocou analytickej funkcie *sum* s využitím tzv. rolling sum, to znamená postupné nasčítanie hodnôt ako suma hodnôt stĺpca predchádzajúcich riadkov s vyšším ohodnotením (príp. rovnakým pri variantoch funkcie *rank*) a aktuálnym riadkom. Aby nasčítavanie fungovalo správne, musíme uviesť, že sa majú hodnoty nasčítavať zoradené podľa hodnoty škody (*order by celk\_skoda\_kc desc*). Potom stačí, keď si vyberieme riadok s ohodnotnením *rn=1000*, kde je nasčítaných TOP 1000 škôd.
   2. Druhý variant je po vytvorení ohodnotení škôd využitie agregačnej funkcie *sum*, ktorou zrátame po odfiltrovaní nepotrebných záznamov (cez *where* podmienku) sumu všetkých zostávajúcich hodnôt stĺpca so s hodnotou škody.

Náklady oboch variantov sú rovnaké a ich skript je zobrazený na obrázku 17.

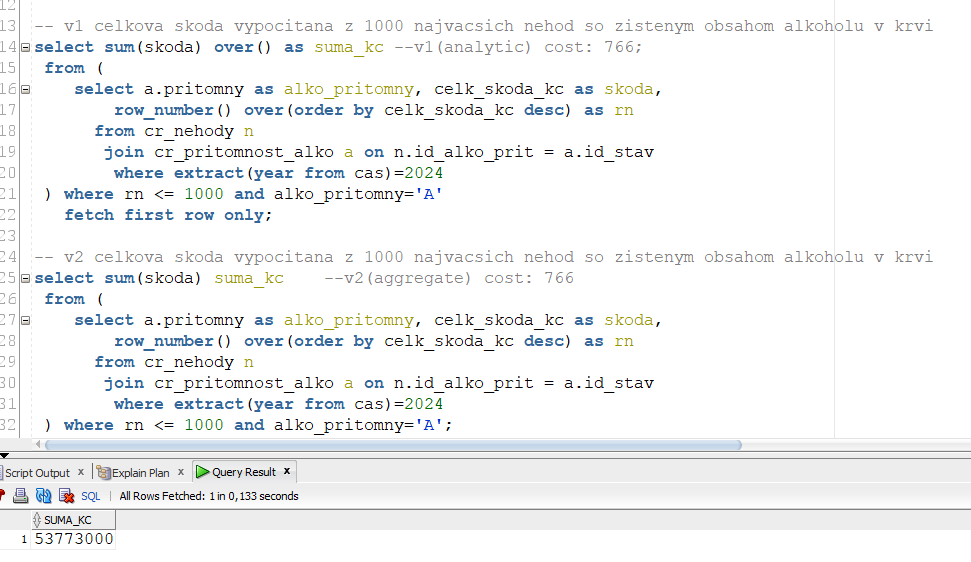


Obrázok 17 – dve možnosti spočítania sumy TOP 1000 najvyšších škôd roku 2024

1. Čitateľ (obrázok 18) – celková škoda s prítomnosťou alkoholu spomedzi TOP 1000 nehôd

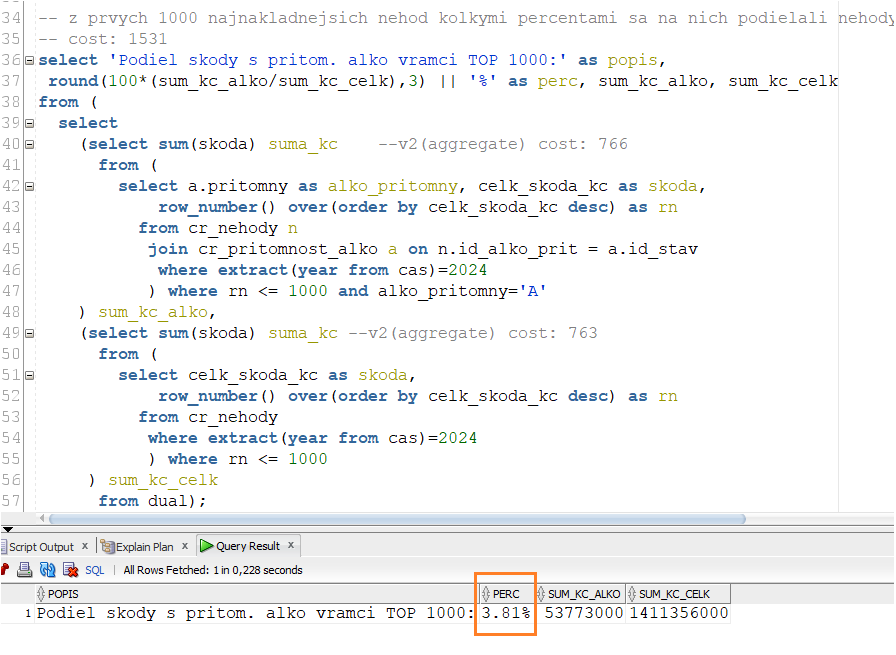
Najprv si zoradíme podľa veľkosti škody v Kč všetky nehody a priradíme im ohodnotenie pomocou analytickej funkcie *row\_number*. Okrem toho si musíme zapamätať či bol alkohol prítomný v krvi počas nehody, alebo nie. Následne, keď máme záznamy zoradené podľa ohodnotenia, vyberieme v podmienke *where* všetky záznamy, ktoré sú medzi prvými TOP 1000 a zároveň bol pri nehode alkohol prítomný v krvi. Potom už stačí len spočítať celkovú sumu škôd z vybraných záznamov a to môžeme dosiahnúť dvoma spôsobmi:

* 1. Využitie analytickej funkcie sum, v ktorej neuvedieme podmienku triedenia, čím zabránime využitiu rolling sum, ale namiesto toho sa sčíta suma zo všetkých záznamov a priradí sa ku každému záznamu. Potom nám stačí pomocou klauzuly *fetch* zobrať ľubovoľný, napr. prvý riadok, v ktorom získame hľadanú sumu.
  2. Druhý(jednoduchší) spôsob je využitie agregačnej funkcie *sum* bez definovania skupiny pre všetky vybrané záznamy.



Obrázok 18 - dve možnosti spočítania sumy škôd s obsahom alkoholu spomedzi TOP 1000 roku 2024

Po vytvorení príkazov *select* pre čitateľ aj menovateľ stačí napísať *select* podľa vzoru na obrázku 16. Príklad výsledného skriptu je na obrázku 19. Môžeme vidieť, že na TOP 1000 nehodách podľa veľkosti vyčíslených škôd mali škody s prítomnosťou alkoholu v krvi 3,81% podiel na sumárnej škode.



Obrázok 19 – výsledný skript pre získanie perc. podielu škôd s prít. alkoholu na TOP 1000 škodách

# Optimalizácia výkonnosti

# Záver