COVID-19 adatok elemzése

MS Excel alkalmazással

Készítette: Szikora Viktor

Tartalom

[Bevezetés 4](#_Toc55156114)

[A letöltött táblázat elsődleges elemzése 5](#_Toc55156115)

[OLTP – OLAP különbségek 5](#_Toc55156116)

[1. ábra Eredeti táblázat fejléce 5](#_Toc55156180)

[2. ábra Módosított fejlécű táblázat 6](#_Toc55156181)

# Bevezetés

Témaválasztásomnál figyelembe vettem, hogy lehetőleg aktuális, a társadalom széles körét foglalkoztató kérdéskörben keressek olyan adatsort, amely OLAP[[1]](#footnote-1) elemzés elkészítésére alkalmas.

A 2019. év végétől a világ figyelme a Kínából indult COVID-19 járványra összpontosult. A hírek, az online és hagyományos média vezető hírei között a napi eset- és halálozási számok állnak. A vírus és a veszélyek értelmezése az emberek, de akár a családok között és családon belül is megosztott a lakosságot. Megjelentek amatőr, félhivatalos és hivatalos mérvadók, vírustagadók, járványtagadók, és „az itt a világvége” nézetet vallók egyaránt.

Sokan, sokféle képpen értelmezik a médiában megjelent híreket, adatokat, pedig a Wikipédia[[2]](#footnote-2) jól érthetően definiálja a járvánnyal kapcsolatos fogalmakat:

* **Endémia** egy adott betegség megszokott előfordulása egy adott népességben.
* **Epidémia** a neve, ha fenti egy adott mérőszámot meghalad - például influenza esetén 10%-os fertőzöttségi arány -. Az epidémia úgy is definiálható, mint a fertőzöttségi arány időszakos, a megszokottnál nagyobb mértékű megemelkedése egy adott területen.
* **Pandémia** akkor, ha a fertőzöttségi arány több kontinensen is nagy mértékben megemelkedik. Az egyes besorolások csak a fertőzöttségi aránytól függenek, és semmi közük sincs a betegség súlyosságához, lefolyásához, vagy a betegség miatti halálozás arányához. Hivatalosan pandémiát a WHO hirdethet.

A pandémiához három feltétel együttes teljesülése szükséges[[3]](#footnote-3):

* A népesség számára új betegség megjelenése.
* A kórokozó embereket fertőz meg, komoly betegséget okozva.
* A kórokozó könnyen és fenntarthatóan terjed az emberek között.

A **European Centre for Disease Prevention and Control** [honlapja](https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases) jó kiindulási pont a COVID-19 járvánnyal összefüggő ismeretanyag megszerzéséhez.

**A választott téma naponta frissülő táblázatát az alábbi címről töltöttem le:** <https://www.ecdc.europa.eu/sites/default/files/documents/COVID-19-geographic-disbtribution-worldwide.xlsx>

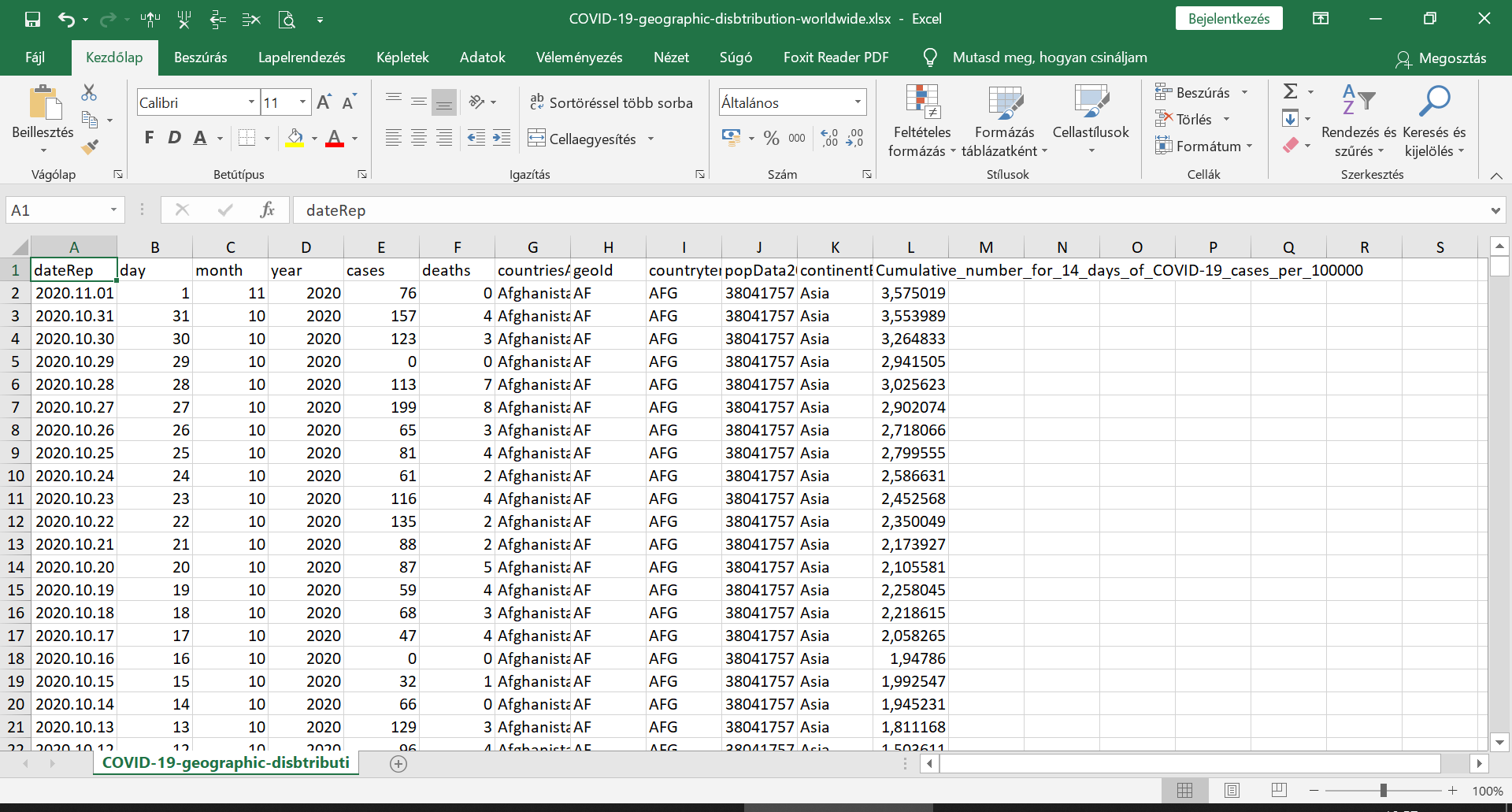
# **A letöltött táblázat elsődleges elemzése**

## OLTP – OLAP különbségek

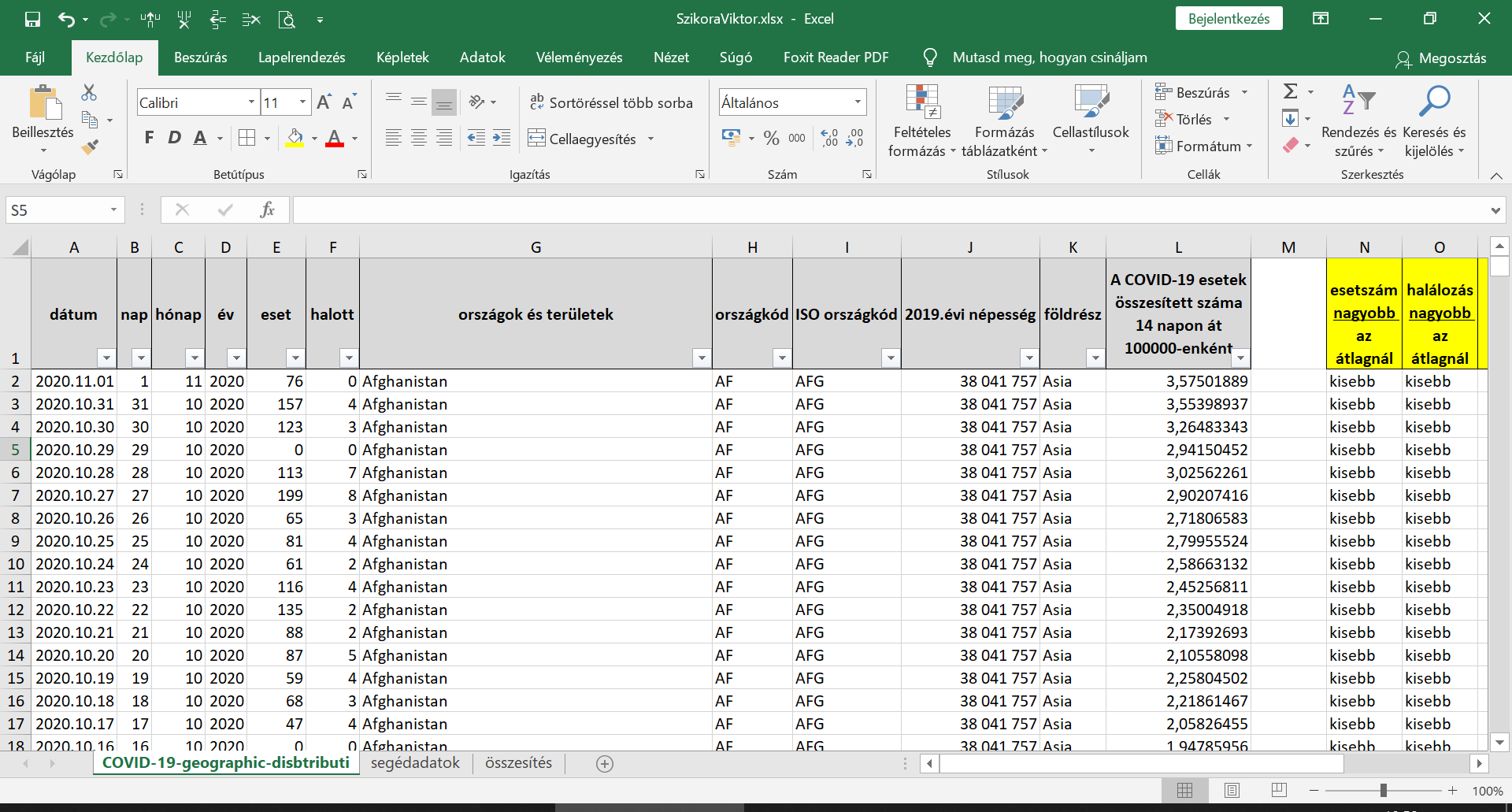
Az OLTP rendszerek a tranzakciók kezelésére összpontosítanak. Az ilyen rendszerek mögött normalizált adatbázisok találhatók, melyek biztosítják, hogy az adatműveletek végrehajtása a lehető leggyorsabban megtörténhessen. Például egy raktári rendszerben, ahol több tucat raktáros a vonalkódos rendszer segítségével másodpercenként rögzít adatokat, nem engedhető meg a felesleges adatok továbbítása és kezelése miatti hosszú tranzakciós várakozási idő. Arról nem is beszélve, hogy egy-egy tranzakció befejezéséig számos adattábla kerül zárolásra a többi felhasználó elől, így ők a megelőző tranzakció befejezéséig várakozni kényszerülnek.

Az OLAP rendszerek ezzel szemben az adatok elemzésével foglalkoznak. OLAP szempontból az OLTP szerkezet hátrányos. A különböző, de egyébként összefüggő adatokat külön adattáblákban tároló OLAP rendszerek a bonyolult lekérdezések miatt hosszú várakozási idő után adhatnak eredményt. Az OLAP szerkezetben megengedettek a redundanciák, azaz egy-egy sorban (rekordban) megismételni az ismereteket. Az OLAP rendszer a sebesség rovására a fájlméret növekedésével biztosítja a gyorsaságot.

**A táblázat letöltése után annak fejlécét angol elnevezésről magyarra írtam át.**



**1**. ábra Eredeti táblázat fejléce

**A módosítás után az alábbi fejléccel ellátott táblázattal folytattam a munkát:**

**2**. ábra Módosított fejlécű táblázat

**Az OLAP szemlélet megjelenése a táblázatban:**

* **minden rekordban megjelenik a dátum megbontása nap, hónap, év részekre,**
* **minden rekordban az ország neve mellett megjelenik annak kétféle kódja, a földrész megnevezése és a népességi adat**

**OLTP szemlélet szerint a dátum nem lenne megbontva, valamint az országok adatai külön táblában jelennének meg. Mivel minden ország neve egyedi, ezért az ország neve meghatározza a kétféle országkódot, a népesség számát, valamint a földrészt, ahol az adott ország található. (Az eredeti adatsor nem foglalkozik a több földrészen fekvő országokkal, mint például Törökország, ami a táblázat szerint Ázsiában található.)**

**A táblázat sorainak száma meghaladja az 52 700 sort. A táblázat készítői a dátum oszlop mellé annak nap-hónap-év megbontását előre elkészíttették, így azzal külön nem kellett foglalkoznom. Ha nem bontották volna fel a dátumot, akkor az Excel függvényeivel készíttetem volna el:**

* **nap meghatározása a *NAP(időérték)*,**
* **hónap meghatározása a *HÓNAP(időérték)*,**
* **év meghatározása az *ÉV(időérték)* függvénnyel történt volna.**

**A következő lépésben egy új munkalapon néhány elemi függvénnyel az eset, és halott oszlopokból a *MIN*, *MAX*, és *ÁTLAG* értékeket állítottam elő. Meglepő módon az eset és halott oszlop minimum adatai között negatív(!) értékek is szerepelnek.**

**A negatív értékekre a honlapon található megjegyzés ad magyarázatot, miszerint egyes jelentő országok a korábbi eredményeiket korrigálni kényszerültek:**

<https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases>

***NOTE:***

\*United Kingdom: from 3 October 2020, due to a technical issue, COVID-19 cases from 24 September to 1 October will be reported by British authorities in the coming days.

On 2 October 2020, authorities in Ireland retrocorrected the total number of COVID-19 deaths leading to a negative value for the deaths reported on 3 October 2020.

On 10 September 2020, Jersey reclassified nine cases as old infections resulting in negative cases reported on 11 September 2020.

As of 7 September 2020, there is a negative number of cumulative cases in Ecuador due to the removal of cases detected from rapid tests. In addition, the total number of reported COVID-19 deaths has shifted to include both probable and confirmed deaths, which lead to a steep increase on the 7 September.

From end of August 2020, Swedish authorities are performing daily data consolidation leading to data retro-corrections. From week 38, the Swedish Public Health Agency will update COVID-19 daily data four times per week on Tuesday–Friday. Hence, the cumulative figures and related outputs include cases and deaths from the previous 14 days with available data at the time of data collection.

**https://www.ecdc.europa.eu/en/interpretation-covid-19-data**

**https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide**

1. OLAP: On-line Analitical Process – Online analitikai feldolgozás, [↑](#footnote-ref-1)
2. https://hu.wikipedia.org/wiki/J%C3%A1rv%C3%A1nytan [↑](#footnote-ref-2)
3. Egy betegség nem feltétlenül pandemikus akkor, ha elterjedt és sok ember veszíti az életét, fertőzőnek is kell lennie. Például a rák sok ember halálát okozza, de nem pandemikus, mert nem fertőző. [↑](#footnote-ref-3)