**Adatelemzés összefoglalás**

Amikor megkaptam az adatfájlokat, hirtelen azt sem tudtam, hogyan kezdjek neki. Miután elolvastam a feladatot, logikus döntésként megnyitottam a .csv fájlokat, és megnéztem a tartalmukat, hogy tisztában legyek azzal, mivel állok szemben. Első gondolatom az volt, hogy egyik fájlban sincs túl sok oszlop. Ezután feltűnt, hogy mindkét adatbázis tartalmazza a cégek azonosítóját, így olyan eszközt kellett választanom, amellyel a táblákat össze tudom kapcsolni.

Elsőként a Python + Jupyter Notebook páros jutott eszembe, de később felmerült bennem az SPSS, a SAS, vagy akár egy relációs adatbázis-kezelő (pl. Oracle vagy Access) használata is. Végül azonban az első megérzésemre hallgattam, mivel úgy gondoltam, ezzel a kettőssel minden szükséges lépést el tudok végezni.

Első lépésként beolvastam az adatokat két külön pandas DataFrame-be, átneveztem az oszlopokat úgy, hogy mindenki számára beszédesek legyenek, majd egy **inner join** segítségével egyesítettem őket, így egy egységes adathalmazt kaptam. Az egyesítés után kiírtam az első öt sort, hogy ellenőrizzem, működik-e minden a tervek szerint.

Ezután megnéztem az alapvető statisztikai jellemzőket. Sajnos az adathalmazban mindössze egyetlen folytonos attribútum szerepelt, mégpedig a **bruttó ár**. Emiatt ez az attribútum központi szerepet kapott a későbbi elemzési lépések során. Az alacsony számú folytonos változó miatt sajnos a gépi tanulási algoritmusok nagy része kiesett, mivel a szöveges adatokkal ezek nehezebben, vagy egyáltalán nem működnek. Ettől függetlenül akadtak érdekes szöveges attribútumok is, például a vállalkozások mérete, illetve a TEÁOR-kategóriák. Csoportosítás szempontjából a cégazonosító volt a legalkalmasabb.

A képen szöveg, képernyőkép, diagram, Diagram látható

Előfordulhat, hogy a mesterséges intelligencia által létrehozott tartalom helytelen.Megvizsgáltam az adathalmaz alapinformációit is, például az attribútumok típusait, valamint azt, hogy találhatóak-e „null” értékek. Szerencsére nem volt szükség sem típuskonverzióra, sem hiányzó értékek kezelésére.

A tényleges elemzést a vállalkozások méretével kezdtem: megvizsgáltam, milyen méretkategóriák léteznek, és ezek milyen gyakorisággal fordulnak elő. Azt tapasztaltam, hogy ahogy nő a vállalkozás mérete, úgy csökken a számuk – bár ez nem meglepő. Ezt a későbbiekben oszlopdiagramon is vizualizáltam, illetve készítettem egy hőtérképet is.

Ezután a TEÁOR-kategóriák eloszlását vizsgáltam. A legtöbb vállalkozás a kereskedelem, szállítás és egyéb szolgáltatások kategóriáiba tartozik – ez a három terület az adatbázis 67%-át fedi le. A legkevésbé képviselt szektor az energiaszektor volt.

A számlák kapcsán azt tapasztaltam, hogy átlagosan a középvállalkozások költöttek a legtöbbet, ami meglepő eredmény, mert általában ezt a szerepet a nagyvállalatokhoz társítanánk. Ezt követően részletesen elemeztem, hogy a különböző méretű vállalkozásoknak milyen típusú számlákból mennyit kellett befizetniük. Az adatbázisból kimagaslott a könyveléssel kapcsolatos kategória, amely közel tízszer annyi előfordulást mutatott, mint a második leggyakoribb „n\_szoft” kategória. Ezt az eredményt táblázatban is bemutattam, a vállalkozásméret szerinti bontásban. A képen szöveg, képernyőkép, Betűtípus, szám látható

Előfordulhat, hogy a mesterséges intelligencia által létrehozott tartalom helytelen.Elemzésem végén pedig megnéztem azt a tíz céget, amely a legtöbbet költötte számlákra.

Összességében úgy érzem, hogy érdekes adatbázissal dolgozhattam. Bár az elején nem volt egyértelmű, mit lehet kihozni belőle, végül úgy gondolom, sikerült a legtöbbet kihoznom a lehetőségekből. Bízom benne, hogy sikerült érthetően bemutatnom a gondolkodásmódomat és az elemzés menetét.