**Iszlai Ferdinánd BI projekt**

A projektem témája egy erdélyi kisváros demográfiai tényezőinek megjelenítése. Pontosabban Nyárádszeredának, amely egy Maros megyei kisváros, Marosvásárhelytől 20 km-re található. Ebben a városban nőttem fel, ezért is választottam ezt a témát. Egy másik fontos ok, amiért ezt a témát választottam a 2020-as szeptemberi helyhatósági választások. A választások alatt lehetőségem nyílt a helyi RMDSZ csapatával együtt dolgozni és segédkezni, adatokat gyűjteni a lakósokról és elemezni azokat. Számos szempontot figyelembe véve gyűjtöttünk adatokat, különböző módszerek által próbáltunk mindenkiről a lehető legtöbb információt begyűjteni, ezzel előre jelezni a választási eredményeket, mivel, nagyon fontos volt aránylag pontosan megjósolni a választási eredményt, mert a tanácsosi helyek száma függött a pontos előre jelzéstől, aszerint állítottuk be a kampány menetét, mozgósítást stb.

Az adatokat egy korábbi (2019-es) választói névjegyzékből szereztük meg, majd ezeket kiegészítettük a számunkra értékes adatokkal, így nagyjából össze is állt a kép, hogy milyen eredményre lehet számítani. Mivel számos személyről nem tudtunk adatokat szerezni, ezért több mint 1000 személy a bizonytalan kategóriába került, melyet arányosan leosztottunk a pártok között. Végül egész jól sikerült meghatározni a választás eredményét, mindössze 3% os tévedési arányunk volt. Ezért is választottam ezt az adatsort, mivel nagyon sokat segítettem az adatok gyűjtésében és valós adatokról van szó. A választói névjegyzék mellet még a 2011es népszámlálási adatokat is figyelembe vettem, mivel a választói névjegyzékben nincsenek benne a 18 évnél fiatalabb személyek adatai.

Az adatokat nagyjából normalizálva gyűjtöttük, így nem volt sok dolgom az adatbázis normalizálásával, az órákon elhangzott módszerek szerint Google Colab-ot használva végeztem el még pár simítást az adataimon mielőtt dolgozni kezdtem volna velük, ilyen volt pl.: az index beállítása, a megfelelő adattípusok beállítása, CSV állomány létrehozása, hibás adatok kiszűrése.

Miután az adatok normalizálva voltak, újra néztem az előző előadásokat, megfigyelve, hogy mik a lehetséges hibák, amiket el lehet követni egy vizualizáció elkészítésekor. Ezután tanulmányozni kezdtem az általam választott ezközt a Flourish-t. Youtube oktatóvideókat nézve lassan megértettem, hogyan is kell használni a felületet, nagyon sokat segített a vizualizációk elkészítésében a tesztadatok megléte, amelyek szerint fel tudtam építeni a különböző adatsoraimat, így a megfelelően elkészített adatsort beillesztve már csak pár simítást kellet végezni egy adott vizualizáción: színek, méretek, cím, magyarázó cetlik beállítása stb.

Összesen 9 vizualizációt készítettem, próbáltam úgy felépíteni a projektet, hogy logikus sorrendbe kövessék egymást a vizualizációk:

* Első sorban az átlagos demográfiai adatokat próbáltam megjeleníteni:

1. Kor szerinti eloszlás: A fő adatsorból szűréseket végezve leválasztottam az általam fontosnak talált korcsoportok egyedeit és ezen egyeteket megszámolva készítettem egy táblázatot, melyben a következő korcsoportok szerepeltek: 0-18 azaz a kiskorúak, 18-30 általában a tanuló fázisban lévő emberek vagy friss munkakezdők, akiknek nincs még sok tapasztalatuk, 30-65 azon személyek, melyek már rendelkeznek munka tapasztalattal egészen a nyugdíjba vonuló személyekig, illetve a 65+ os korosztály, ezalatt az időseket értem. Egy Column Chart segítségével ábrázoltam az adatokat, mivel itt jól összehasonlítható az oszlopok mérete. Mint a vizualizációból is kiderül eléggé elöregedő kisvárosról beszélünk, a 30 év alatti korosztály elég kevés százalékát teszi ki a lakosságnak.
2. Nemek szerinti eloszlás: Szűrések elvégzése után, a nemek szerinti eloszlást egy Pie Chart-tal ábrázoltam, tudom külön megemlítettük az előadásban, hogy nem előnyös Pie Chatot használni, viszont számomra mégis ez tűnt a legjobban értelmezhetőnek a két lehetőség miatt. Ugyanakkor még egy szín béli kódolást is végrehajtottam, a nőket pirossal míg a férfiakat kékkel jelöltem, így könnyen megkülönböztethető a kés körcikk.
3. Kor és nemek szerinti eloszlás: a harmadik vizualizációmban az előző két vizualizációt egyesítettem valamilyen szinten, hogy meg tudjuk állapítani, hogy melyik korosztályban mennyi nő és mennyi férfi egyed szerepel, általában hasonló arányokat kapunk, mint a 2. vizualizációnál, viszont a 65+ os korosztályban a nők száma jelentősen nagyobb, mint a férfiaké, ez lehet a háborúk következtében, vagy csak egyszerűen a fizikai munka okozta kisebb átlagéletkornak köszönhetően.
4. Anyanyelv szerinti eloszlás: Ezen vizualizáció elkészítéséhez szintén egy Column Chart-ot használtam, mivel így látszik legjobban a mennyiségek közötti különbség. Jól látszik, hogy a lakosság nagyrésze magyar, de ugyanakkor román és roma családok is élnek a településen
5. Településenkénti nemzetiség szerinti eloszlás: Ebben a vizualizációban a Nyárádszereda és közigazgatásilag hozzátartozó falvak nemzetiség szerinti ábrázolása figyelhető meg. Grouped Column Chart segítségével ábrázoltam, ahol a csoportok a Települések és a különböző oszlopok pedig az egyes nemzetiségű emberek számát ábrázolják.
6. Vallási hovatartozás szerinti eloszlás: Mivel itt jóval több lehetőség van, ezért itt nem Column Chart-ot hanem Treemap típusú vizualizációt alkalmaztam, minden vallást egy-egy téglatest ábrázol, ezek nagysága határozza meg az adott vallású személyek számát. A szem sokkal jobban össze tudja hasonlítani egy téglatest területét, mint egy körét. Ha egymás mellet vannak e téglalapok, jól látszanak a méret béli különbségek, ezért akár a kurzor rávitele nélkül is könnyedén meg tudjuk mondani, hogy arányaiban melyik vallásnak hány követője van.

* Másod sorban a politikai tényezőket jelenítettem meg:

1. A 2020-as szavazószám eloszlás előrejelzés: Ez a vizualizáció az eredeti adatsorból készült, egy Bar Chart típusú vizualizáció, melyben a pártok és a bizonytalan illetve nem szavazó személyek száma jelenik meg. A sorrend sem véletlen, mivel az ember fentről lefele kezd általában olvasni, ezért felül tettem a legeredményesebb pártot, és tovább csökkenő sorrendbe, illetve az alsó két oszlop a bizonytalan illetve a nem szavazók számát ábrázolja.
2. Nem szavazás okai: Itt a fenti Chart-ban használt nem szavazó személyeket bontottam két kategóriába a nem szavazás okából kifolyólag: vallási okok miatt, illetve itthonlét miatt. Stacked Bar Chart-ot használtam, jól látszanak az arányok, hogy jóval több az itthonlét miatt nem szavazó ember, mint a vallási okok miatt nem szavazó.
3. Tanácsosi helyek eloszlása: Az utolsó vizualizáció, mely a választások utáni adatokra épül, megmutatja a 2016-ban illetve 2020-ban elért tanácsosi helyeket, színkódolást is alkalmaztam, ugyanakkor a diagram alatti táblázatból jól látszik az előző mandátumhoz képest a változás mértéke. 2016-ban az RMDSZ és az MPP egyesült, ezért nincs MPP-s tanácsos, viszont 2020-ban újra indult immár EMSZ néven (ezt nem vittem be a vizualizációba, az egyszerűség kedvéért) és három tanácsosi helyet is szereztek.

A vizualizációkban adott alakzatra rávíve a kurzort megjelenik pontosan az adott érték, illetve kiválasztva különböző kategóriákat, ezek külön megjeleníthetőek, így átláthatók és testreszabhatóak a vizualizációk. Színválasztáshoz a következő palettageneráló weboldalakat használatam: coolors.co, paletton.com Egyes vizualizációk helyettesíthetőek, Barchartok Columnchartokkal, illetve a településenkénti eloszlás megvalósítható valamiféle Map vizualizációval is(csomó próbálkozás után inkább hagytam így, sehogy sem jött össze az egyes falvak külön ábrázolása).

Számos dolgot tanultam a félév során, nagyon élveztem a tantárgyat. Sok olyan dologra világított rá, amire ez elött nem is gondoltam. Normalizálás és adattáblák kezelése Google Colab-ban Pythonnal, különböző dimenziók megjelenítésének leghatékonyabb módszerei, további adatvizualizációs felületekbe való betekintés, azok kipróbálása, adatok átlátása könnyebb értelmezése. Elégedett vagyok, fölülmúlta a tantárgy az elvárásaimat, remélem, hogy a következő félévi tantárgy is legalább ennyire izgalmas lesz.