2022. évi magyar országgyűlési választási eredmények elEMzése python segítségével

Készítette: Szujó Róbert

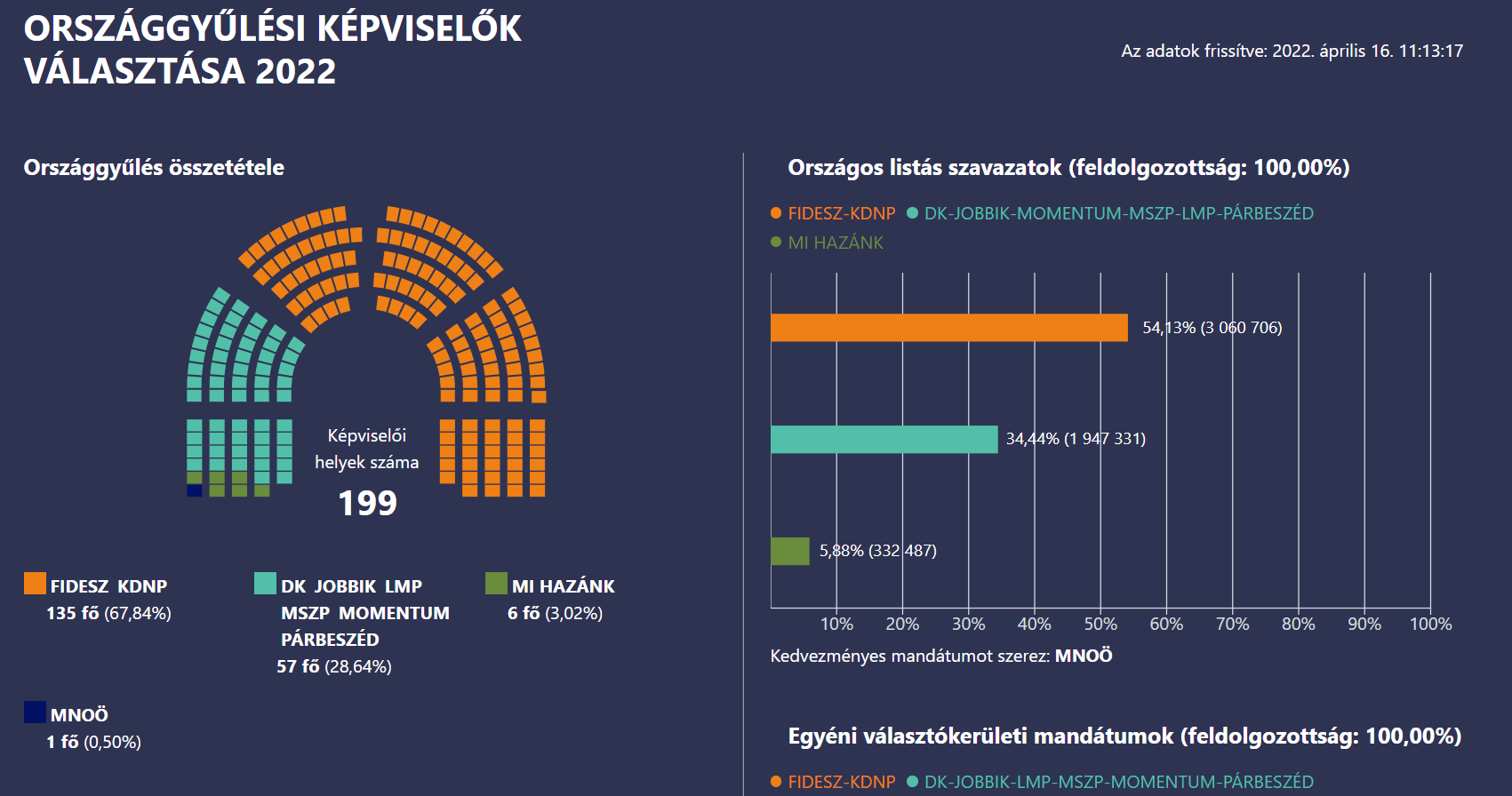
A képen szöveg látható

Automatikusan generált leírás

# Bevezető

2022. április 4-én nagy esemény történt a magyar közéletben, hiszen ismét megtartották az országgyűlési választásokat. A választást eredetileg nagy érdeklődés fogadta, hiszen a sajtó és a különböző közvélemény-kutatások szerint éles versenyre lehetett számítani, elsősorban a már 12 éve kormányt alakító FIDESZ-KDNP koalíció, illetve a hat legnagyobb ellenzéki pártokból frissen megalapult, kifejezetten a kormánypárt leváltására létrehozott pártszövetség között. Természetesen más pártok is részt vettek a választáson, azonban a közvélemény figyelme elsősorban az előbbi két tömörülésre szegeződött.

A választás gördülékenyen lezajlott, azonban az eredmény nemcsak az ellenzék győzelmében bízók, hanem még a kormány támogatói között is nagy meglepetést okozott, hiszen a FIDESZ-KDNP sorban 4. alkalommal alapíthat kormányt, ráadásul a közel „korlátlan hatalmat” biztosító, úgynevezett „2/3-os többséggel” teheti ezt meg.



1. ábra: A választás eredményeképp kialakult mandátumok („parlamenti patkó”)

Természetesen, a sajtóban szinte azonnal megjelentek a különböző elemzések, találgatások arról, hogy ki a felelős az újabb ellenzéki vereségért, illetve, hogy milyen tényezők vezettek el az újabb elsöprő győzelemhez a jelenlegi kormánypárt részéről.

Ennek az elemzésnek azonban nem célja, hogy a különböző felek kampányának stratégiáját, döntéseit és annak következményét vizsgálja. Célom volt azonban, hogy összefüggéseket keressek a választási eredményeken belül, illetve – saját meglátásaim alapján válogatott adatok alapján - különböző társadalmi statisztikai mutatók, és a választási eredmények között.

Tekintve, hogy ez az esemény a közelmúltban történt, és a teljes magyar közéletre (illetve valamilyen szinten mindannyiunk életére) hatással van, nem mellesleg a mai napig sok kérdést rejteget a kialakult eredmény, különösen érdeklődéssel fordultam az elemzési feladat felé. Továbbá, úgy éreztem, hogy ez a feladat hasznos lesz abból a szempontból is, hogy az eddigi adatelemzői tanulmányaim eredményeit bemutathassam, különös tekintettel az adatelőkészítés területére (melynek menetét később részletezem), illetve az előkészített adatok elemzésére, vizualizációjára is. Az elemzés végén, egyfajta bemutató jelleggel, egy egyszerű gépi tanulási algoritmust (döntési fát) is létrehoztam az adatokon. Az adatok elemzéséhez az ingyenes és szabadon hozzáférhető Python programozási nyelvet, illetve annak különböző kiegészítő csomagjait (scikit-learn, matplotlib, seaborn, pandas) használtam, a programkódot pedig Jupyter Notebook-ban írtam meg a könnyebb átláthatóság érdekében. A kód szerkesztéséhez a szintén ingyenesen igénybe vehető Visual Studio Code szoftvert használtam, a szükséges Python kiegészítőkkel.

# Az általam vizsgált kérdések

Mielőtt bármilyen előkészítő munkát végeznénk, érdemes feltenni magunknak azt a kérdést, hogy valójában mi a célja az elemzésnek, milyen eredményeket szeretnénk látni, mit szeretnénk elérni. Én is így tettem, és rövid gondolkodás után a következő kérdések fogalmazódtak meg bennem:

* Mekkora volt az országos részvételi arány? Hogyan alakult országos szinten a pártokra leadott szavazatok megoszlása?
* Mely pártoknak kedvezett, mely pártok eredményét hátráltatta a magas részvételi arány az egyes településeken?
* Mennyire segített volna az egyes településeken, ha a FIDESZ-KDNP-vel szemben valóban csak egy párt indul a választásokon?
* Hogyan alakult a pártokra leadott szavazatok megoszlása a különböző típusú településeken? Mely pártok értek el jobb eredményeket a nagyobb/kisebb településeken?
* Mely pártok szavazói között volt a legnagyobb ellentét, és mely pártok szavazatai mozogtak a leginkább együtt, vagy épp a többitől függetlenül?
* Van-e kapcsolat az egy-egy településre jellemző, különböző társadalmi mutatók, és a pártokra adott szavazatok megoszlása között? Ha igen, fel tudjuk-e használni ezeket a mutatókat, hogy a jövőben valamilyen módon előre jelezhessük a szavazatok alakulását?

# Az elemzés korlátai

Ez az elemzés elsősorban az általam tanult módszertanok és algoritmusok bemutatása céljából jött létre, nem tudományos publikáció. Igyekeztem az előkészítés és az elemzés során is tudományosan igazolt, szakmailag elfogadott módszereket alkalmazni, azonban az általam hozzáférhető adatok több tekintetben is hiányosak voltak, ezekkel szemben kénytelen voltam kompromisszumos megoldásokat alkalmazni.

Ezek a „kompromisszumok” a következők voltak a fő adathalmazra (a választási eredményekre vonatkozóan):

* Csak a pártlistára leadott szavazatokat vettem figyelembe!
* NEM vettem figyelembe a külföldön/levélben leadott szavazatokat!
* Az eredményeken településenként, összesített formában végeztem elemzést (az egyes fővárosi kerületek külön településnek számítottak).

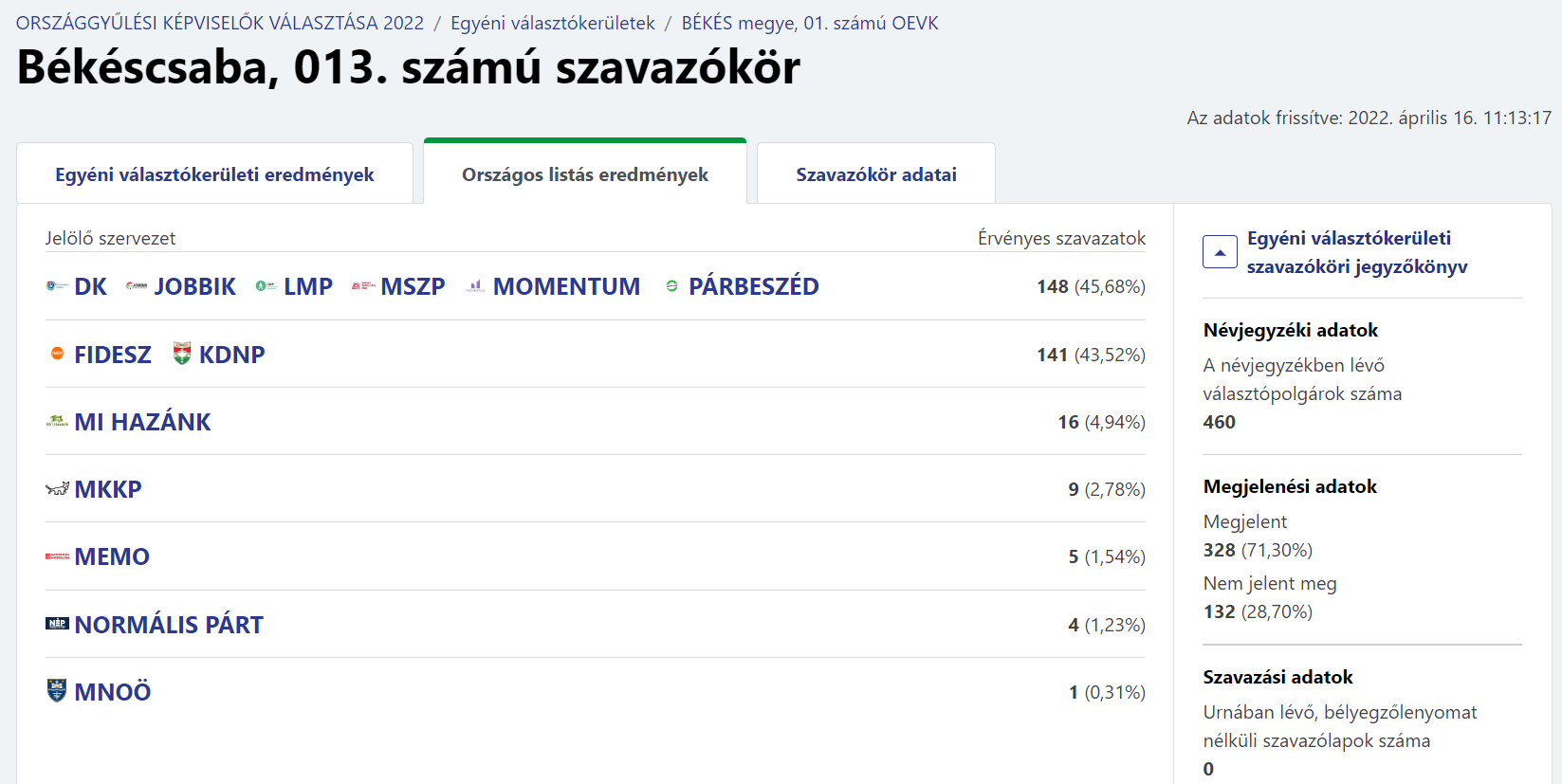
Az elemzés során használt, KSH-ról származó statisztikai mutatók csak megyékre és/vagy településtípusokra volt elérhetőek, ezek időben sem voltak feltétlenül azonosak, hiszen az adatokat részben a KSH legutóbbi STADAT tábláiból gyűjtöttem össze (melyek jellemzően 2020-2021 között frissültek utoljára), az adatok másik részét pedig a KSH Mikrocenzus 2016 letölthető állományaiból gyűjtöttem össze. A későbbiekben részletezem, milyen módszert alkalmaztam a KSH adatainak településekhez való illesztésére, melyek a későbbi elemzéshez voltak szükségesek.

**Összességében, szeretném felhívni a figyelmet arra, hogy az elemzés kizárólag tanulási célból jött létre, a fenti korlátokból eredő torzító hatások miatt nem feltétlenül tükrözi a valós körülményeket!**

# Az adatok beszerzése, előkészítése

A Nemzeti Választási Iroda weboldalán bármikor megtekinthetjük az országgyűlési választások eredményét, mind az egyéni jelöltekre, mind a pártlistára vonatkozóan. Az oldal lehetőséget biztosít arra, hogy megnézzük az országos összegzést (amely tartalmazza a külföldön és levélben leadott szavazatokat is), illetve a választókerületenként összesített adatokat is. Kicsit rejtve ugyan (több linken keresztüljutva), de meg tudjuk nézni az egyes szavazókörökben született eredményeket is, ezek a legrészletesebb adatok, melyeket elérhetünk (természetesen a választás titkossága miatt egyéni adatokhoz nem juthatunk hozzá).

Az elemzéshez a pártlistára leadott szavazatokat fogom felhasználni, mivel az egyéni jelöltek minden választókerületben különbözőek, és bár az indító pártokban rengeteg átfedés van, de ezeknek az összegyűjtése és kezelése külön nehézségeket okozna, melyekre ez az elemzés nem tér ki. Természetesen, ezt is figyelembe kell venni, mint torzító hatást, amikor a végleges eredménnyel, azaz a „parlamenti patkóval” szeretnénk összehasonlítani a mi elemzésünk eredményét (nem beszélve a különböző törvényi befolyásoló tényezőktől, mint például a parlamenti küszöb szabálya, és ehhez kapcsolódóan a szavazat-kompenzáló rendszer működése).



2. ábra: Országos listára leadott szavazatok egy adott szavazókörben

Külön nehezíti a feladatunkat, hogy az NVI weboldaláról semmilyen formában nem tölthetjük le a választási eredményeket, különösen nem olyan formátumban, amely a gépi adatelemzéshez megfelelő lenne. Azonban, a böngészőnk fejlesztői eszköztárát megnyitva megfigyelhetjük, hogy az egyes szavazókörök adatai nincsenek „beleégetve” a weboldalba, hanem egy REST-alapú háttérrendszer szolgáltatja a front-end felé. Ez a háttérrendszer „.json” formátumú fájlokban juttatja el a szavazókörök adatait, mely tartalmazza az egyes pártokat, és a rájuk leadott szavazatok számát, illetve egyéb statisztikákat is.

A képen szöveg látható

Automatikusan generált leírás

3. ábra: Egy szavazókör adatait tartalmazó .json fájl

Ezeket a fájlokat azonban össze kell gyűjteni, és mivel települések szintjén szeretnénk elvégezni az adatelemzést, ezért összesíteni is kell őket, amennyiben több szavazókör is található egy településen. Maguknak a fájloknak a tartalmát is értelmeznünk kell valamilyen módon, hiszen nincs minden adatra szükségünk, és amelyekre szükségünk van, azok sem feltétlenül egy helyen szerepelnek a fájlon belül. Erre szolgál az első, általam megírt Python kód, mely összegyűjti ezeknek a szavazóköröknek az adatait, településenként összegzi őket, majd kimenetként előállít egy .csv fájlt, amely már gond nélkül használható további műveletekhez.

A képen szöveg látható

Automatikusan generált leírás

4. ábra: A szavazókörök adatait összegyűjtő programkód egy részlete

# További előkészítő műveletek, KSH adatok illesztése a településekhez

Az első programkód által előállított fájl a következő adatokat tartalmazza:

* + Az összes települést, a hozzá tartozó megyével
  + Az egyes pártokra leadott szavazatok darabszámát (az Egyéb kategória a különböző nemzetiségi szavazatokat jelenti)
  + A választásra jogosult személyek számát
  + A ténylegesen leadott szavazatok számát (= az érvényes szavazatot leadott személyek száma)

A képen asztal látható

Automatikusan generált leírás

5. ábra: Az első programkód futtatása után létrejött táblázat (Excel-ben megnyitva)

Az egyes településeken leadott szavazatok azonban még nem hasonlíthatóak össze, hiszen a leadott szavazatok darabszáma erősen függ attól, hogy az adott településen mennyien szavazhatnak. Az összehasonlíthatóság érdekében ezeket az adatokat százalékos értékre alakítjuk át, eredményképp megkapjuk, hogy az egyes pártok az összes leadott szavazatok hány %-át kapták meg. A későbbi elemzéshez fontos lesz még a részvételi arány megállapítása, mely a leadott szavazatok számának és a választásra jogosult személyek számának hányadosa. A választásra jogosult személyeknek a számát azonban önállóan is meghagyjuk a táblázatban, mert ez még hasznos lesz a későbbiekben. Továbbá, minden település esetén megállapítjuk, mely párt kapta a legtöbb szavazatot.

A képen asztal látható

Automatikusan generált leírás

6. ábra: Az előkészített választási eredmények

Az előkészítés utolsó lépése, hogy a KSH adataiból kiválogatott statisztikai mutatókat összeillesszük a választási eredményben szereplő településekkel. Ezeken belül a következő adatokat használtam fel:

* + Az iskolázottság mértékét (érettségivel rendelkező 18 év felettiek %-a)
  + A munkanélküliségi rátát (%-ban)
  + Az átlagéletkort (évben)
  + Az interneteléréssel rendelkező háztartások arányát (%-ban)
  + A családos (gyermekkel rendelkező) háztartások arányát (%-ban)
  + A súlyos anyagi deprivációban (~”szegénységben”) élők arányát (%-ban)

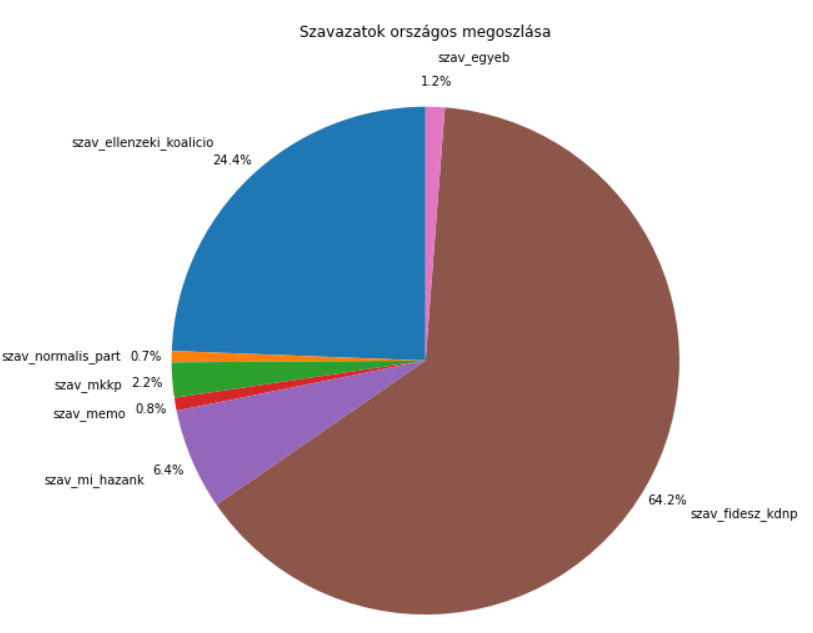
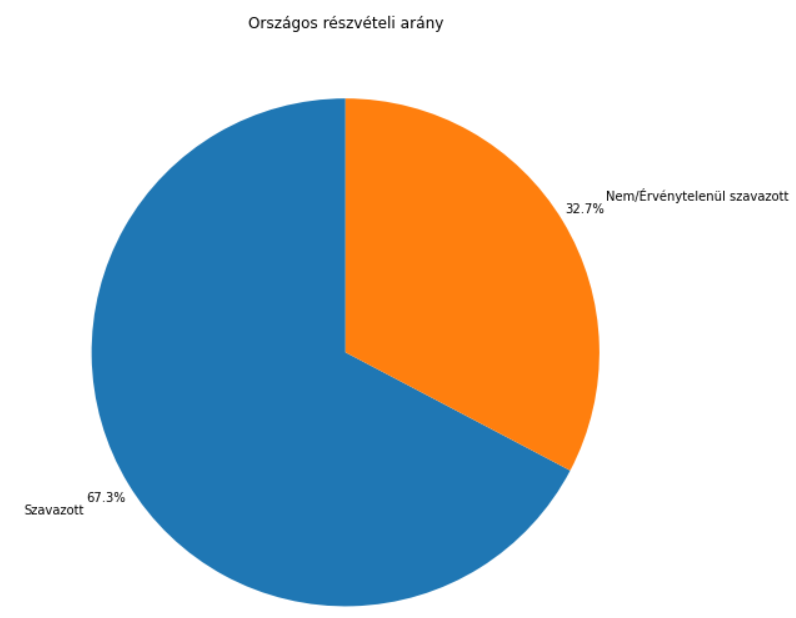
Az adatok megyénként, illetve településtípusonként állnak rendelkezésre, ezek alapján kötöttem őket az egyes településekhez. Ha egy adat megyére és településtípusra is elérhető volt, a két adat átlagát vettem a település besorolásának megfelelően. Ezek az adatok szabadon letölthetőek a KSH weboldaláról .xls formátumban, azonban önmagában nem használhatóak gépi adatelemzésre, mivel számos formázási opciót tartalmaznak. A gépi adatkezelésre alkalmas formátumba alakítást kézzel végeztem el, mivel véleményem szerint feleslegesen sok időt és energiát vett volna igénybe ennek az automatizálása.

Az adatok összeillesztését a pandas „merge” funkciója segítségével végeztem el, különös tekintettel arra, hogy minden településhez tartozzon adat. A statisztikai mutatók felvétele mellett, felvettem az egyes településekhez tartozó településtípust is, melyeket szintén felhasználók az elemzésben.

A végeredményként kapott, elemzésre kész táblázat 3177 sorból (ennyi település van), és 20 oszlopból áll. Hiányzó adatokra nem végeztem vizsgálatot, mivel a korábbi műveleteket úgy végeztem el, hogy azok kizárják a felmerülés lehetőségét.

# A választási eredmények elemzése

## Országos szintű eredmények

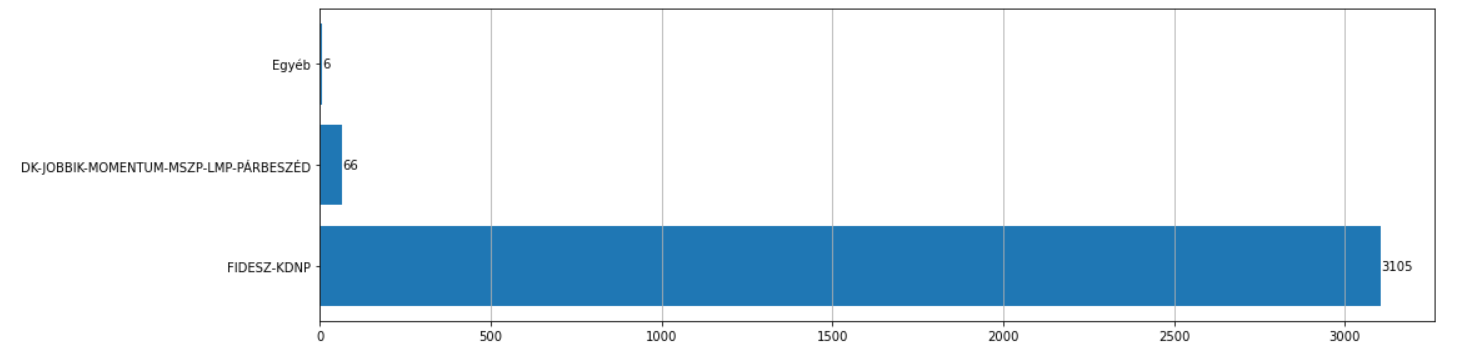


7. ábra: Országos részvételi arány és a pártokra adott szavazatok megoszlása

A két kördiagramm alapján a következő megállapításokat tehetjük:

* A szavazásra jogosultak 67,3%-a elment szavazni, ÉS érvényes szavazatot adott le.
* A legtöbb szavazatot a FIDESZ-KDNP koalíciója kapta, összesen 64,2%-ot.
* A második legtöbb szavazatot az ellenzéki koalíció kapta, 24,4%-ot.
* A többi párt 11,4%-nyi szavazaton osztozik (a nemzetiségi szavazatokat is beleértve)

## A legtöbb szavazatot kapott pártok az egyes településeken

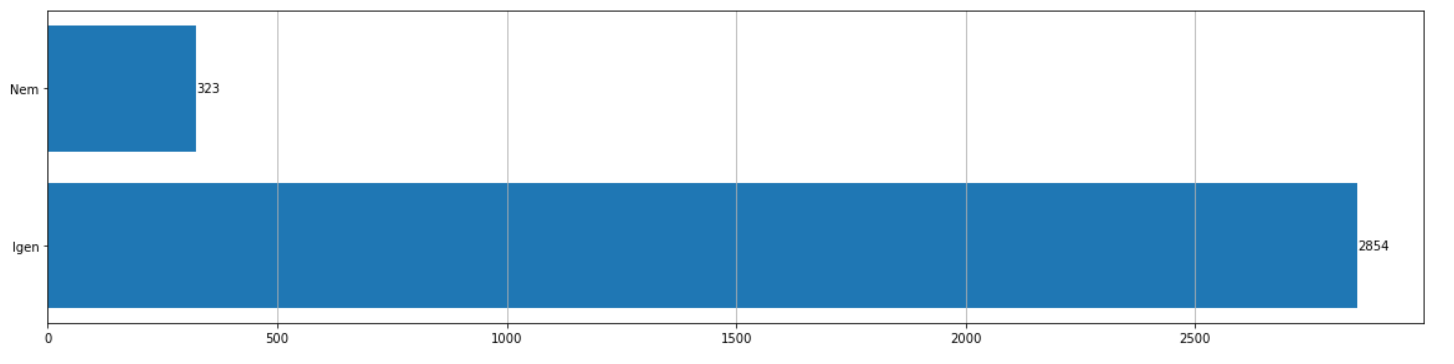


8. ábra: A legtöbb szavazatot kapott pártok az egyes településeken

A települések ~97,7%-án (3105 db) a FIDESZ-KDNP szerezte a legtöbb szavazatot. Az ellenzéki koalíció mindössze a települések ~2,1%-án (66 db) tudott „többségbe kerülni”. Ezek között szerepel a legtöbb (21 db) fővárosi kerület. További 6 településen (~0,02%) valamely nemzetiség képviselője kapta a szavazatok többségét.

Érdemes megjegyezni, hogy itt relatív többségről van szó, tehát arról, hogy mely párt kapta a legtöbb szavazatot a többihez képest. Azonban ez nem jelenti azt, hogy a szavazók több, mint fele az adott pártot választotta.

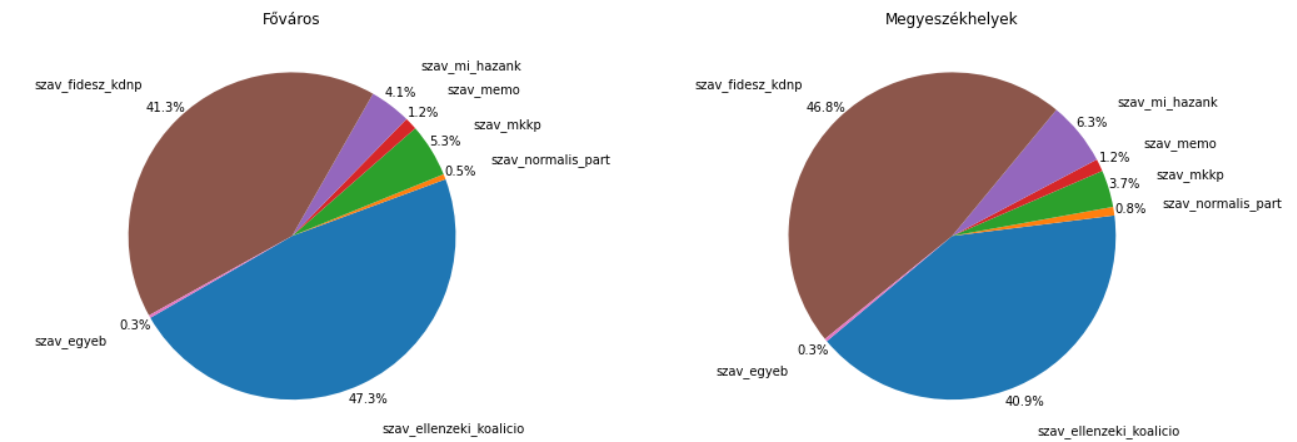
Vizsgáljuk meg, hogy a legtöbb „legtöbb szavazattal” rendelkező FIDESZ-KDNP koalíció mennyi településen kapta meg a szavazatok legalább 50%-át:

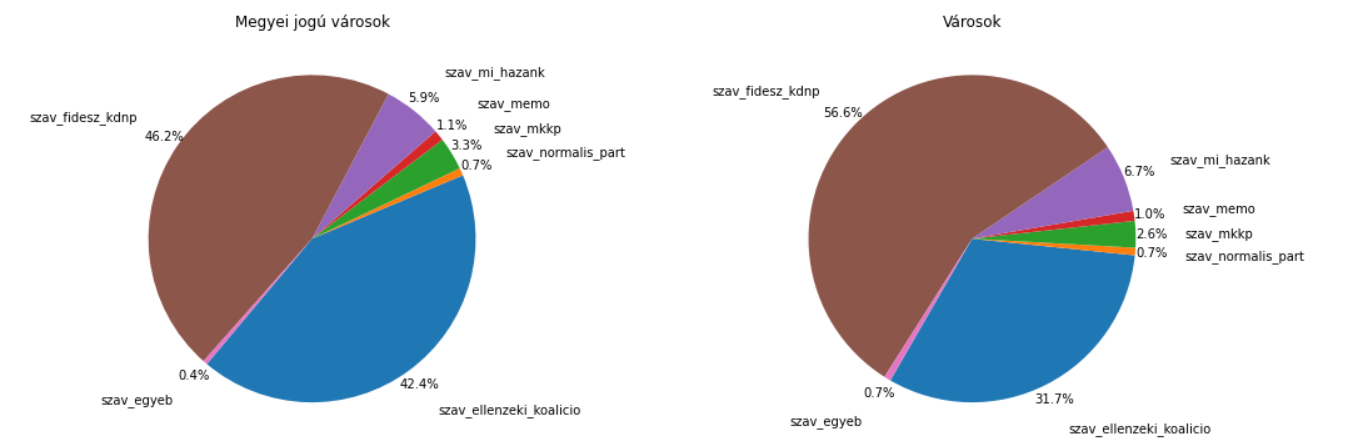
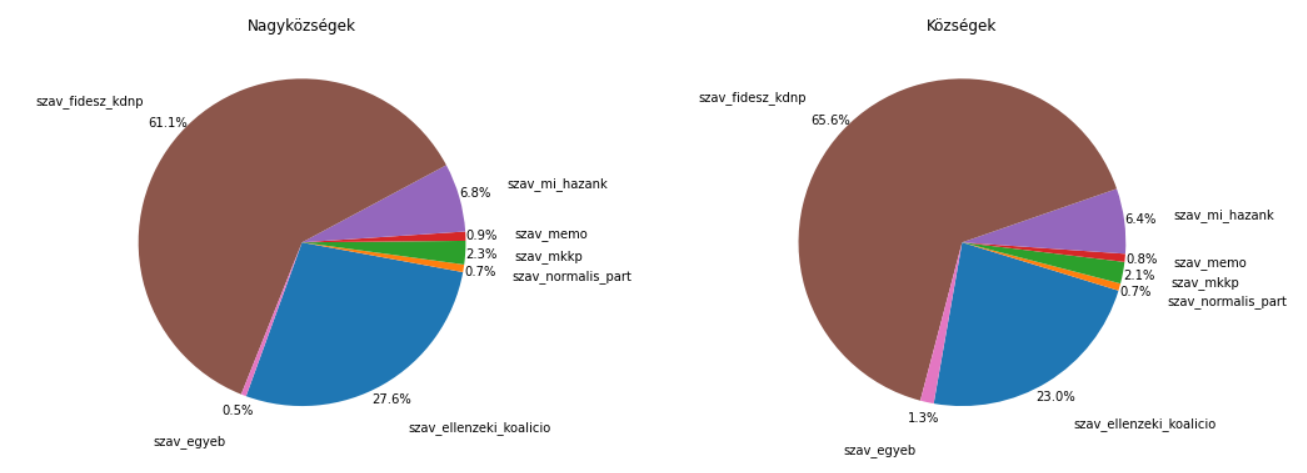


9. ábra: Települések száma, ahol a FIDESZ-KDNP kapta a szavazatok legalább 50%-át

Ez körülbelül a települések 89,8%-a, ami így 7,9%-kal kevesebb azokhoz a településekhez képest, ahol a párt relatív többséggel nyert. Ez azt jelenti, hogy a települések 7,9%-án elméletileg nyerhetett volna az ellenzék, amennyiben nem több párt, hanem valóban egyetlen pártként indul a FIDESZ-KDNP-vel szemben. Természetesen vegyük figyelembe, hogy a valós életben ez az arány jóval kevesebb lenne (illetve maga a szituáció sem történne meg nagy valószínűséggel), hiszen nem minden ellenzéki szavazó lenne hajlandó arra, hogy egy teljeskörű ellenzéki összefogásra is szavazzon.

## Szavazatok megoszlása az egyes településtípusokon

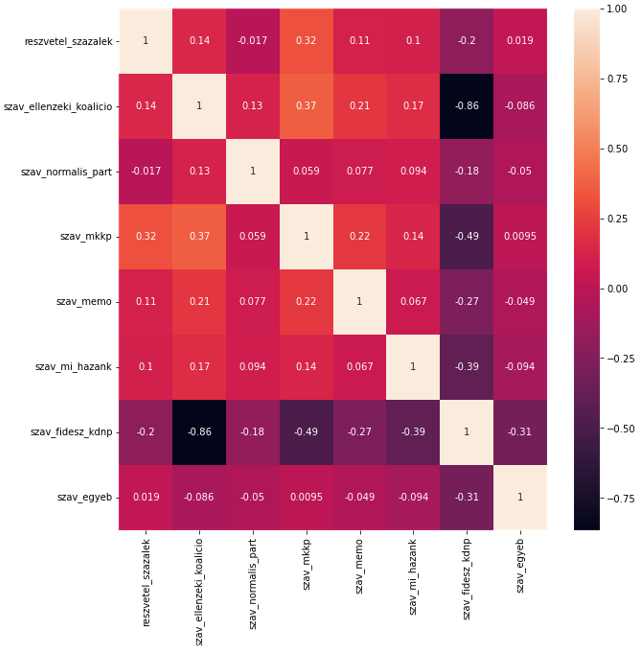


10. ábra: Szavazatok megoszlása a különböző településtípusok szerint összegezve

Általánosságban elmondható, hogy minél kisebb a település, annál nagyobb a FIDESZ-KDNP, és annál kisebb az ellenzéki koalíció (és az MKKP) támogatottsága. A Mi Hazánk elsősorban a vidéki szavazók körében volt „népszerű”, de ott településmérettől függetlenül. A kisebb pártok (MEMO, NP) teljesítménye azonban „független” volt a település jellegétől.

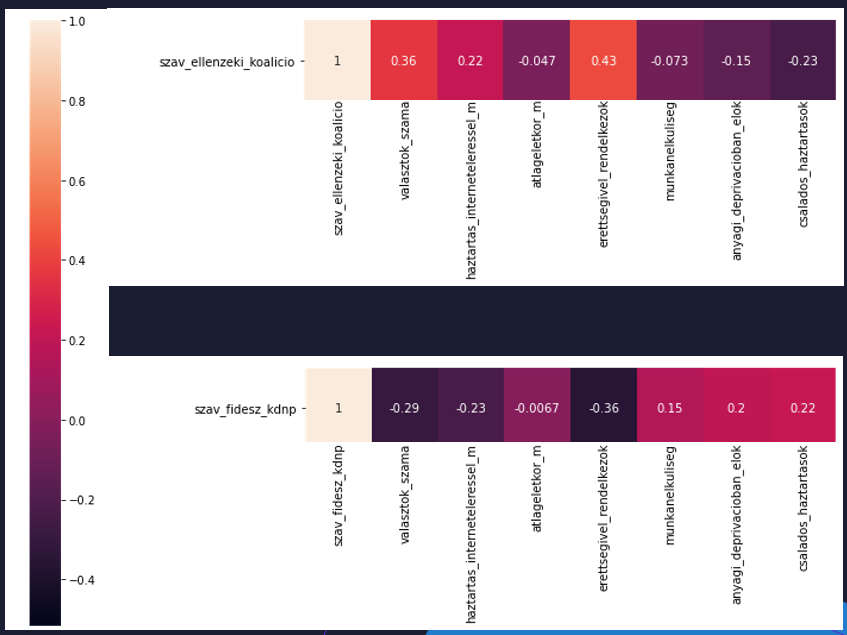
## Kapcsolat a pártokra adott szavazatok, illetve a részvételi arány között



11. ábra: Pártokra adott szavazatok, és a részvételi arány korrelációja Heatmap-en ábrázolva

A Seaborn-modul által nyújtott heatmap egy rendkívül robosztus és látványos eszköz arra, hogy egyszerre sok változó között megvizsgáljuk a korreláció nagyságát és irányát. Megfigyelhetjük például, hogy a részvételi arány egyik pártnak sem okozott komoly előnyt/hátrányt, habár az MKKP esetében már-már mondható, hogy a rájuk adott szavazatokra volt némi pozitív hatása. Természetesen, a legnagyobb ellentét (érthető módon) a FIDESZ-KDNP és az ellenzéki koalícióra szavazók között volt. Az „ellenzéki régióban” az MKKP is jobban teljesített, a FIDESZ-KDNP tőlük is számos szavazatot tudott elvonni. Érdekes módon azonban, a Mi Hazánk is hátrányt szenvedett a „fideszes” körzetekben. Végül még elmondható, hogy a kisebb pártok teljesítménye kevésbé függött a nagyobb pártoktól.

## Kapcsolat a települések társadalmi mutatói, és a leadott szavazatok között



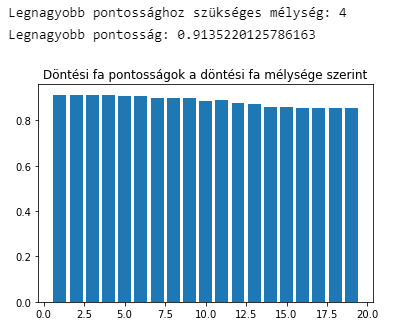
12. ábra: Kapcsolat a települések társadalmi mutatói, és a leadott szavazatok között

Talán, ami legelőször feltűnhet a két heatmap-et megfigyelve, hogy azokon a településeken, ahol a legalább érettségivel rendelkezők száma nagyobb volt, ott nagyobb volt az ellenzéki koalícióra szavazók száma is. Ugyanez fordítva is igaz a Fidesz-re nézve (bár valamennyivel gyengébb mértkében), a rájuk adott szavazatok aránya alacsonyabb a legalább érettségivel rendelkezők körében. Elmondható továbbá, hogy a nagyobb lélekszámú településeken általánosan nagyobb volt az ellenzéki koalíció támogatottsága. A szélesebb körű internet-hozzáférés szintén segítette az ellenzéket, és körülbelül ugyanolyan mértékben gyengítette a FIDESZ-KDNP koalíciót. A családosok magasabb aránya azonban inkább a FIDESZ-KDNP-nek kedvezett, amely valószínűleg nem meglepő a pártszövetség családpolitikája, és a számukra adott, különböző kedvezmények és támogatások miatt.

# Döntési fa készítése arra vonatkozóan, hogy mely településeken várható 50% feletti FIDESZ-KDNP győzelem

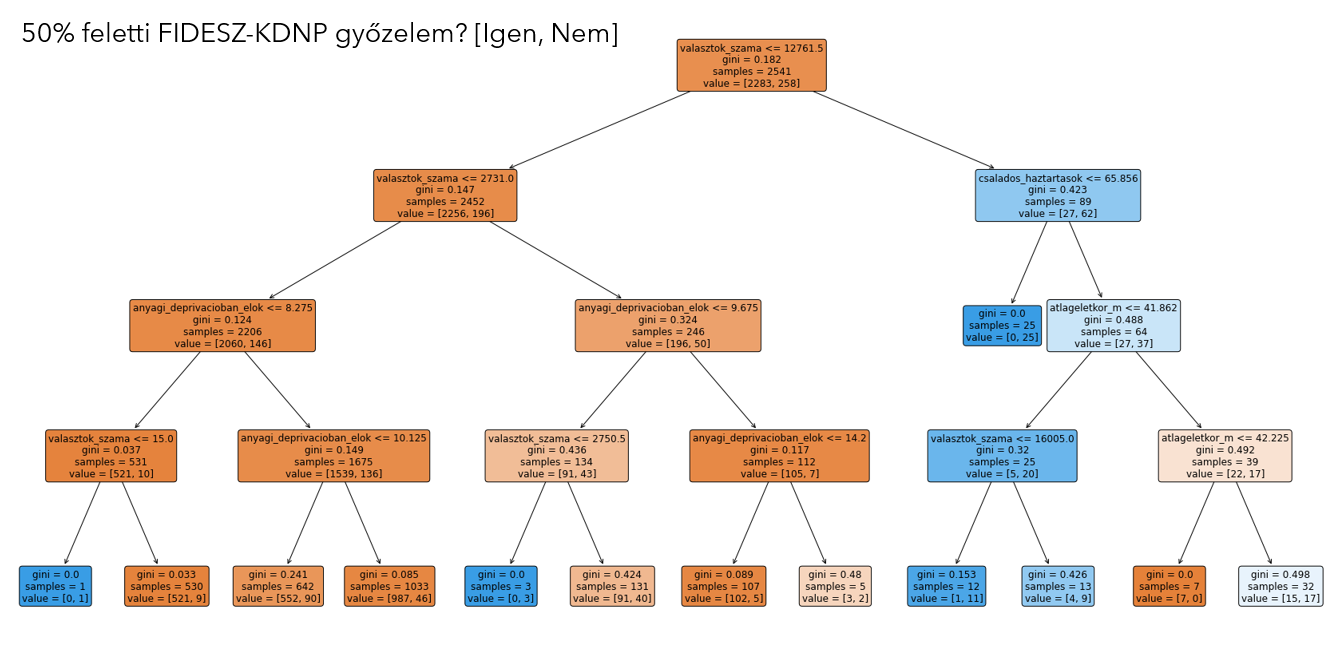
A Scikit-Learn modul segítségével viszonylag egyszerűen létrehozhatunk döntési fát, amely talán a legsokrétűbben alkalmazható osztályozó algoritmus. Előnye, hogy könnyen vizualizálható és értelmezhető, nem igényli az adatok előzetes sztenderdizációját (bár a nominális változóknál mindenképp javasolt a One-Hot kódolás használata). Ezek alapján, próbáljuk meg az előbb bemutatott társadalmi mutatókat felhasználni arra, hogy predikciót készítsünk vele egy-egy településre vonatkozóan.

A döntési fa optimális mélységét egy iterációs algoritmus segítségével valósítom meg, mely 1-20 közötti mélységnél megvizsgálja a pontosságot, és visszaadja (illetve vizualizálja), hogy mely mélységnél volt a legnagyobb a pontosság. Ez tulajdonképpen egy „Early-stopping” algoritmus, melynek segítségével elkerülhető a túltanulás, és a fa komplexitásának felesleges növelése.



13. ábra: Mélységekhez tartozó pontosságok

Látható, hogy 4-es mélységgel kaphatjuk meg a lehető legpontosabb predikciót készítő döntési fát. Készítsük ezt el, és próbáljuk meg ábrázolni.



14. ábra: „Optimális” döntési fa

Az adatokat még a fa betanítása előtt szétválasztottuk tanuló és validációs adatokra. Használjuk a validációs adatokat ahhoz, hogy megvizsgáljuk a döntési fa valódi pontosságát.

A képen szöveg látható

Automatikusan generált leírás

15. ábra: Döntési fa pontossága

Az ábrán a zöld pontok jelentik azokat az értékeket, melyek valójában az adatokhoz tartoznak. Amennyiben a predikció pontos volt az adott adatra (településre) vonatkozóan, a zöld pont eltakarja a pirosat (ami pedig a predikált értéket jelenti), azaz minél kevesebb piros pontot látunk, annál pontosabb az előrejelzés. Látható, hogy viszonylag pontos eredményünk született, ezt alátámasztja a konkrét százalékos érték is. Természetesen érdemes megjegyezni, hogy a korábban már bemutatott torzító hatások alapján aligha biztos, hogy ugyanez a döntési fa valóban alkalmazható lenne a választás eredményének „megjósolására”, azonban kiindulási alapnak mindenképp használható.

# Összegzés

Véleményem szerint, a korlátozottan elérhető adatok ellenére is sikerült egy olyan elemzést készíteni, amely rávilágít a választási eredmények bizonyos összefüggéseire, és közelebb hozza számunkra a véglegesen kialakult eredmény okozó tényezőit.

Az elemzéshez használt valamennyi adatfájl és programkód elérhető a projekt mappájában, remélem, más is hasznát veszi az általam elvégzett munkának, és további hasznos összefüggésekre tud rávilágítani azok felhasználásával. Külön köszönet a KSH-nak és az NVI weboldalának, melyek az elemzéshez szükséges adatok biztosították, és az összes általam használt Python modul fejlesztőjének, akik önkéntes munkájukkal hozzájárultak ahhoz, hogy ezeket az elemzéseket bárki, ingyenesen elvégezhesse.