ÓBUDAI EGYETEM
ÓBUDA UNIVERSITY

NEUMANN JÁNOS
INFORMATIKAI KAR

SZAKDOLGOZAT

OE-NIK Hallgató neve: **Gál Aliz**
2023 Hallgató törzskönyvi száma: **T/007083/FI12904/N**

Óbudai Egyetem
Neumann János Informatikai Kar
Kiberfizikai Rendszerek Intézet

SZAKDOLGOZAT FELADATLAP

Hallgató neve: **Gál Aliz**
Törzskönyvi száma: T/007083/FI12904/N
Neptun kódja: [redacted]

A dolgozat címe:

Tőzsdei adatok előrejelző modellje
Forecasting model of Stock Exchange data

Intézményi konzulens: Sziklai Zsolt
Külső konzulens: Fésüs Péter

Beadási határidő: 2022. december 15.

A záróvizsga tárgyai: Számítógép architektúrák
Big Data és üzleti intelligencia specializáció

A feladat

Tervezzen és készítsen egy olyan modellt, amely a múlt idősoros adatai alapján előre jelzi a kiválasztott tőzsdei adatok alakulását.
Vegye figyelembe, hogy a részvény- és pénzügyi piacok kiszámíthatatlanok és gyakran nem logikusan működnek, egy-egy váratlan esemény miatt turbulens szerkezetűek, ami megnehezíti a releváns mintázatok beazonosítását, valamint áruk jövőbeli alakulását.
Ennek megfelelően hozzon létre egy olyan modellt, amely megadott, több éves idősoros adatok alapján képes az adott részvény árfolyamának hosszútávú becslésére, előre jelzésére.
A turbulens struktúrák modellezéséhez támaszkodjon az idősoros elemzések során használt algoritmusokra és a megfelelő statisztikai módszertanokra.
A megfelelő elemzési technikák alkalmazását követően értékelje az előrejelzés pontosságát (illeszkedését), már elérhető időintervallumú adatokkal összevetve, illetve készítsen jövőbeli projekciót.
A vizsgálatba bevont részvényeket szervezze kategóriákba, csoportokba.

A dolgozatnak tartalmaznia kell:

- a megoldandó feladat pontos leírását és részletes elemzését,
- a lehetséges megvalósítási módokat és megoldásokat,
- a részletes rendszertervet és annak indoklását,
- a megvalósítás során használt technológiák bemutatását,
- a megvalósítás menetét, a munkafázisok és a tapasztalatok leírását,
- a tesztelési módszereket, tesztelési terveket és a tesztelés eredményeit,
- az elkészült rendszer alkalmazási, és továbbfejlesztési lehetőségeit,
- online a kifejlesztett rendszer forráskódját, illetve a készített dokumentációkat.

ÓBUDAI EGYETEM * Neumann János Informatikai Kar * 1. Budapest
Ph.

.....
Dr. Fleiner Rita
intézetigazgató

A szakdolgozat elévülésének határideje: **2024. december 15.**
(OE TVSz 55.§ szerint)

A dolgozatot beadásra alkalmasnak tartom:

.....
külső konzulens

.....
intézményi konzulens

ÓBUDAI EGYETEM
ÓBUDA UNIVERSITY

Neumann János Informatikai Kar

HALLGATÓI NYILATKOZAT

Alulírott Gál Aliz kijelentem, hogy a szakdolgozat / diplomamunka saját munkám eredménye, a felhasznált szakirodalmat és eszközöket azonosíthatóan közöltem. Az elkészült szakdolgozatomban / diplomamunkámban található eredményeket az egyetem és a feladatot kiíró intézmény saját céljára térítés nélkül felhasználhatja.

Budapest, 2022. 12. 06.

.....
hallgató aláírása

ÓBUDAI EGYETEM
ÓBUDA UNIVERSITY

Neumann János Informatikai Kar

KONZULTÁCIÓS NAPLÓ

Hallgató neve: GÁL ALIZ

Neptun kód: [redacted]

Tagozat: ESTI

Telefon: [redacted]

Levelezési cím (pl: lakcím): [redacted]

Szakdolgozat / Diplomamunka[1] címe magyarul:

TŐZSDEI ADATOK ELŐREJELZŐ MODELLJE

Szakdolgozat / Diplomamunka[2] címe angolul:

FORECASTING MODEL OF STOCK EXCHANGE DATA

Intézményi konzulens: FÉSÜS PÉTER

Külső konzulens:

Kérjük, hogy az adatokat nyomtatott nagybetűkkel írja!

Alk.	Dátum	Tartalom	Aláírás
1.	2022.02.11.	Témaválasztás, feladatkiírás előkészítése	
2.	2022.03.08.	Feladatlap egyeztetése	
3.	2022.04.05.	Tartalomjegyzék és az eddig összeállított szakdolgozat hibáinak átbeszélése	
4.	2022.05.06.	Elkészített szakdolgozat 1. során felmerülő hibák átbeszélése	

A Konzultációs naplót összesen 4 alkalommal, az egyes konzultációk alkalmával kell láttamoztatni bármelyik konzulenssel.

A hallgató a Szakdolgozat I. / Szakdolgozat II. (BSc) vagy Diplomamunka 1 / Diplomamunka 2 / Diplomamunka 3 / Diplomamunka 4[3] tantárgy követelményét teljesítette, beszámolóra / védésre[4] bocsátható.

Budapest, 2022. május 08.

.....
intézményi konzulens

[1] Megfelelő aláhúzandó!
[2] Megfelelő aláhúzandó!
[3] Megfelelő aláhúzandó!
[4] Megfelelő aláhúzandó!

ÓBUDAI EGYETEM
ÓBUDA UNIVERSITY

Neumann János Informatikai Kar

KONZULTÁCIÓS NAPLÓ

Hallgató neve: GÁL ALIZ

Neptun kód: [redacted]

Tagozat: ESTI

Telefon: [redacted]

Levelezési cím (pl: lakcím): [redacted]

Szakdolgozat / Diplomamunka[1] címe magyarul:

TŐZSDEI ADATOK ELŐREJELZŐ MODELLJE

Szakdolgozat / Diplomamunka[2] címe angolul:

FORECASTING MODEL OF STOCK EXCHANGE DATA

Intézményi konzulens: SZIKLAI ZSOLT

Külső konzulens: FÉSÜS PÉTER

Kérjük, hogy az adatokat nyomtatott nagybetűkkel írja!

Alk.	Dátum	Tartalom	Aláírás
1.	2022.10.11.	Elemzés felépítésének átbeszélése	
2.	2022.11.03.	Trend és kiugró értékek elemzése	
3.	2022.11.10.	ARIMA modellről való egyeztetés	
4.	2022.11.17.	Egymásra hatások bemutatása	

A Konzultációs naplót összesen 4 alkalommal, az egyes konzultációk alkalmával kell láttamoztatni bármelyik konzulenssel.

A hallgató a Szakdolgozat I. / Szakdolgozat II. (BSc) vagy Diplomamunka 1 / Diplomamunka 2 / Diplomamunka 3 / Diplomamunka 4[3] tantárgy követelményét teljesítette, beszámolóra / védésre[4] bocsátható.

Budapest, 2022. november 30.

intézményi konzulens

[1] Megfelelő aláhúzandó!
[2] Megfelelő aláhúzandó!
[3] Megfelelő aláhúzandó!
[4] Megfelelő aláhúzandó!

TARTALOMJEGYZÉK

1. BEVEZETÉS

Ezen szakdolgozat célja egy olyan modell építése, amely alkalmas lehet a tőzsdei árfolyam trendek előrejelzésére, amely a döntések szakszerűbb meghozatalát segíti. Napjainkban az adatok kiváltképp nagy jelentőséggel és értékkel rendelkeznek, minél szélesebb körű tudás birtokában vagyunk, és minél nagyobb perspektívájú látókörrel bírunk a statisztikai és matematikai metodológiák terén, annál pontosabban fel tudjuk mérni, hogy az adott adathalmaz milyen többletinformációt biztosít számunkra. Képesek lehetünk megállapítani:

- a szabályszerűséget, változást az adatsorban, amely akár a szezonalitásból is fakadhat;
- a ciklikus mintákat, ismétlődéseket, amelyek segítenek megállapítani az ismétlődés gyakoriságát, fellendülések, avagy visszaesések mértékét;
- trendeket az adatokban;
- a trendek növekedési ütemét. [1]

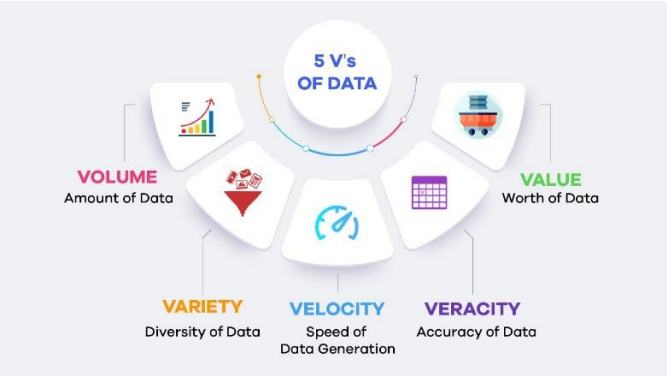
Mindezen megállapítások segítségével akár egy vállalat beszerzési stratégiáját is támogathatjuk vagy akár befektetések által hozamot érhetünk el.
A vállalati vezetők gyakorlatilag minden döntésük előtt figyelembe vesznek valamilyen előrejelzést, amely segíti őket a döntéshozásban az érintett területen. Erre van szükségünk a tőzsdei befektetések esetében is, ahol különösen fontos a turbulens szerkezet miatt a legmegfelelőbb előrejelző rendszer kiválasztása, az adatok tisztítása. Számos előrejelző rendszer ismert már manapság, azonban megfelelő döntést csak és kizárólag úgy tudunk hozni, ha az adott adathalmazhoz leginkább releváns technológiát használjuk. A módszer kiválasztása sok tényezőtől függ – előrejelzés kontextusa, pontosság foka, múltbeli adatok elérhetősége, relevanciája, az előrejelzésre rendelkezésre álló idő, a hozzá szükséges költség, illetve az általa nyert haszon.

Fontos mindezen ismérveket mérlegelni és az alapján felállítani a szempontrendszerünket. Egy adott szakterület nem tartja szükségesnek a 100% pontosságú becslést, hanem minél előbb elfogadható pontosságú kimutatást akar látni, hiszen egy 90%-os pontosságú modell, amennyiben előbb a rendelkezésünkre áll, akkor akár nagyobb profitot/hozamot is elérhetünk vele. Épp ezért a tervezés és a szempontok átgondolása egy előrejelzés során nagy súllyal bír.
A számunkra megfelelő előrejelző rendszer kiválasztását fontos, hogy az adattisztítás folyamata kövesse, hiszen, ahogy fent is említettem az adatok tartalmazhatnak szezonálisan kiugró értékeket, amelyek az előrejelzést nagyban befolyásolják és torzított becslést tárnak elénk.

1.1. A TŐZSDE ÉS A BIG DATA KAPCSOLATA

Az adatok gyarapodása és a növekvő technológiai összetettség átalakítja az iparágak működését, versenyét. A Big Data segít ezen strukturált és strukturálatlan adatok gyűjtésében, feldolgozásában és elemzésében is. Különösen a pénzügyi szolgáltatások tekintetében fontos a Big Data elemzés, annak szempontjából, hogy jobb pénzügyi, befektetési döntéseket tudjunk hozni. A portfólió hozamának maximalizálása miatt, illetve tekintettel a nagy adatmennyiségre, összetett matematikai modellek alkalmazására van szükség. Ahogy ebben a cikkben Trevir I Nath is említi, a New York-i tőzsde mindennap 1 terabájt adatot, információt állít elő, így ennek hatékony és gyors feldolgozása nagyon fontos, megkülönböztető hatása lehet a versenytársakkal szemben. [2]
A Big Data szempontjából alapvető fontosságú a 5 V, amelyet az 1.ábrán szemléltetek.[3]

Az 5 V



1. ábra: Az 5V

Bővebben kifejtve az 5 tényező jelentése:[4][5]

- **Volume:** az adatok méretét jelenti ez az ismérv, amelyek manapság már nagyobbak, mint terabájt és petabájt. A mai számítógépek pedig nem alkalmasak arra, hogy ilyen nagy mennyiségű adatot kezeljenek, főként úgy, hogy a feldolgozásuk is általában eltérő technológiát igényel, mint a hagyományos tárolási és feldolgozási lehetőségek. A tőzsde esetén perc szintű adatokról is beszélhetünk a különböző indexeket tekintve, amely egy óriási adatmennyiséget jelent.
- **Velocity:** az adatok sebességét jelöli, mint például a nagy sebességgel létrejövő Facebook bejegyzések. A több millió felhasználó ugyanabban az időpillanatban posztolhatja az életeseményeit.
- **Variety:** az adatforrások sokféleségének a kifejezése, hiszen az adatok lehetnek strukturáltak, félig strukturáltak és strukturálatlanok is. Ezek közül mindegyik különféle feldolgozást és algoritmust is igényelhet.

- **Veracity:** az adatok minőségére vonatkozó tulajdonság. Az adatok tartalmazhatnak fontos információkat is az elemzésekhez, azonban lehet, hogy az adott adathalmaz nagyobb részt csak értelmetlen adatokat hordoz, amelyeket ki kell szűrni.
- **Value:** az üzlet szempontjából az egyik legfontosabb V az érték. Ez arra utal, hogy a Big Data mit nyújthat, a szervezetek mit tehetnek az összegyűjtött adatokkal. A cél, hogy a Big Data-ból értéket tudjunk levonni, amely hatékonyabb működést, erősebb ügyfélkapcsolatot eredményez és számszerűsíthető üzleti előnyhöz juthatunk általa.

1.2. A VÁLASZTOTT PORTFÓLIÓ ÉS INDOKLÁSA

A fejezet célja annak a bemutatása, hogy miért az adott indexek letöltését választottam. A vizsgálatban olyan indexek lettek kiválasztva, amelyek egy-egy piac reprezentatív indexeként is funkcionálnak több tucat, olykor több száz tőzsdén kereskedett részvényt foglalnak magukban, és nem feltétlenül limitálódnak egyes piaci szektorokra. Mindegyikükre igaz továbbá, hogy a pénzügyi szektor legnagyobb figyelemmel övezett, legjelentősebb és legnagyobb tőkét csoportosító indexei.

Földrajzi értelemben egy igen széles skálát sikerült felvonultatni a kiválasztott indexek terén, a legnagyobb részvénypiacok, legnagyobb indexei kivétel nélkül a vizsgálat tárgyát képezik.

Az S&P 500 ez egyik legismertebb, a világon leginkább követett és az amerikai tőzsdét leginkább reprezentáló indexként van számontartva. Az S&P 500 Annual Survey of Asset publikációja alapján 13 500 milliárd dollár van indexelve vagy benchmarkolva az indexhez, amelyből az indexelt eszközök 5400 milliárd dollárt tettek ki 2020 végén. A bizottság által kiválasztott 500 tőzsdén jegyzett amerikai nagyvállalat az USA teljes tőzsdepiacának 80%-át teszi ki. [6]

Az amerikai NASDAQ composite index szintén egy jelentős, az S&P 500 után talán a legjelentősebb index az amerikai (és globális) tőzsdék világában. Ez az index szinte az összes Nasdaq-on (az USA második legnagyobb részvénytőzsdéje) kereskedett részvényt (~3800 cég papírjai) magában foglalja, megközelítőleg 19,4 ezer milliárd dollár értékben. [7]

A vizsgálatba bevont harmadik, szintén nagyon meghatározó jelentőségű indexe a DOW Jones 30 index (Dow Jones Ipari Átlag) amely egy klasszikus, nagyipari, telekommunikációs, gyártói vállalatokat csoportosító index, a benne foglalt cégek ipari kapitalizációja megközelíti a 50 milliárd dollárt. [8]

Amerikából kilépve, számos európai részvényindex a vizsgálat látókörébe került, az első ilyen az Egyesült Királyság tőzsdéjét reprezentáló FTSE 100, a London Stock Exchangen jegyzett 100 legnagyobb piaci kapitalizációjú Egyesült Királyságban jegyzett cége, összességében 1814 ezer milliárd Font értékben. [9]

Szintén jelentős index az Euronext 100, amely a 100 legnagyobb európai „blue chip” (legnagyobb kereskedett forgalmú, leglikvidebb és legnagyobb kapitalizációjú értékpapírok gyűjtőneve) cégek indexe, jelenleg 64 francia, 17 holland, 7 norvég, 6 belga, illetve 3-3 ír és portugál cég papírjait tartalmazza. [10]

Fontos megemlíteni persze, hogy bizonyos átfedések is akadnak az egyes indexek között, kiváltképp igaz ez a makróregionális vagy éppen a globális indexek esetében.

1.3. A FELADAT BEMUTATÁSA

A feladat során célom egy olyan előrejelző modell kiépítése, amely a múlt adataiból minél pontosabb becsléssel próbálja megállapítani, hogy az egyes indexek árfolyamai várhatóan, hogy fognak a jövőben alakulni. Erre még nem áll rendelkezésre tökéletes modell, azonban nem is célom egy 100%-os előrejelző modellt kiépíteni, a hibák lehetőségével a modell kiépítése során számolni kell, amelyet majd a későbbiekben lehet tökéletesíteni.

Tehát a szakdolgozatban a múlt nagyobb adathalmaza alapján építek ki egy a statisztikában közismert ARIMA modellt, amely megbecsüli a jövőbeli árfolyamok alakulását a kiválasztott indexeket tekintve.

Erre akkor van lehetőség, ha az idősoron a megfelelő transzformációkat elvégeztem, melynek segítségével az előrejelzés torzításai kizárhatóak. Ehhez hozzátartozik az adatsor trendjének vizsgálata, illetve a kiugró értékek elemzése. Ezen vizsgálódást követően pedig az adathalmaz stacionaritásának a vizsgálata szükséges a modell kiépíthetőségéhez.

A modell kiépítését követően pedig az eredeti adathalmaz adataival összevetem a becsült adatokat, amelynek segítségével megállapítható a modell pontossága.
A szakdolgozat végén pedig a kiválasztott indexek korrelációját vizsgálom, amelynek a fő indikátora az egymásra hatások vizsgálata.

1.4. A SZAKDOLGOZAT FELÉPÍTÉSE

A szakdolgozat felépítése a következő:

- Az első fejezetben bemutatom a tőzsde kapcsolatát a Big Data világával.

- Adathalmaz: következőkben az adathalmaz struktúráját, illetve tárolását fejtem ki.

- Irodalomkutatás:
  - az idősor elemzéshez rendelkezésre álló internetes irodalmak, könyvek, cikkek feldolgozása valósul meg, amelynek során bemutatok néhány rendelkezésre álló modellt;
  - az idősor elemzés során szükség van az adatok előfeldolgozására, annak érdekében, hogy megfelelő elemzéseket lehessen készíteni. Ennek a szakirodalmát mutatom be.

- Megvalósítás: az ezt követő részben a részvényárfolyamok tényleges elemzése, előrejelzése történik meg, illetve az egyes egymásra hatások vizsgálata valósul meg.

- Következtetések levonása és továbbfejlesztés: ezen fejezetben a szakdolgozat értékelése, a modell kiértékelése, illetve a fejlesztési lehetőségek kerülnek feltüntetésre.

- A szakdolgozat végén pedig az irodalomjegyzék tartalmazza a felhasznált szakirodalmat, illetve az ábrajegyzékben láthatóak a bemutatott ábrák listája.

2. ADATHALMAZ

Az adathalmaz a Yahoo Finance weboldal által áll rendelkezésemre, Python library segítségével letölthetőek a tőzsdei adatok.

Date: a kereskedés napja.

High: a maximum árra vonatkozik egy adott időszakon belül.

Low: a minimum árat jelenti egy kiválasztott időszakot vizsgálva.

Open: a nyitási ár az az ár, amellyel aznap a kereskedés nyitott.

Close: a zárási ár az az ár, amellyel az adott napon a tőzsde zárt.

Volume: a teljes összeg a kereskedési tevékenység során.

Adj Close: a korrigált értékek olyan vállalati tevékenységeket is figyelembe vesznek, mint például az osztalék, részvényfelosztás, illetve az új részvények kibocsátása.

Az adathalmazra minta a 2.ábrán látható Apple részvényének az alakulásáról:

Minta adathalmaz (Apple részvény)

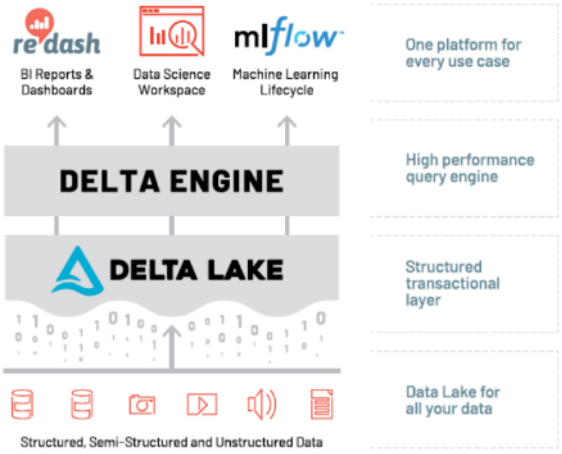
Date	Open	High	Low	Close	Adj Close	Volume
2018-12-31	39.632500	39.840000	39.119999	39.435001	38.233898	140014000
2019-01-02	38.722500	39.712502	38.557499	39.480000	38.277531	148158800
2019-01-03	35.994999	36.430000	35.500000	35.547501	34.464806	365248800
2019-01-04	36.132500	37.137501	35.950001	37.064999	35.936069	234428400
2019-01-07	37.174999	37.207500	36.474998	36.982498	35.856091	219111200

2.ábra: AAPL alakulása-minta adathalmaz

2.1. ADATHALMAZ TÁROLÁSA: DATABRICKS

A szakdolgozatom során a Databricks-et fogom használni az adatok tárolására és feldolgozására. Az irodalomkutatás alátámasztja, hogy ez az eszköz az egyik legalkalmasabb a Big Data feladatokhoz. Ezt egy összehasonlító oldallal is megerősítem, amelyen a Databricks magasabb értékelést kapott a felhasználók között, főként az egyszerűbb használata és könnyű beállítása miatt.[11] A Databricks célja a Data Lakehouse koncepció megvalósítása egy felhőalapú platformon, amely a 3.ábrán látható.[12]

Databricks adatplatform



3.ábra: Mit csinál a Databricks

A Databricks adatplatform-architektúra rétegeinek lebontása:

- A Delta Lake egy olyan tároló réteg, amely megbízhatóságot ad. A Delta Lake ACID tranzakciókat, skálázható metaadat-kezelést biztosít, és streaming és a batch adatfeldolgozást tesz lehetővé.
- A Delta Engine egy olyan optimalizált lekérdező motor, amely a Delta Lake-ben lévő adatok hatékony feldolgozását segíti.
- Ezenkívül számos beépített eszköz elérhető, amelyek a Data Science-hez és az üzleti intelligenciához nyújtanak támogatást. [12]

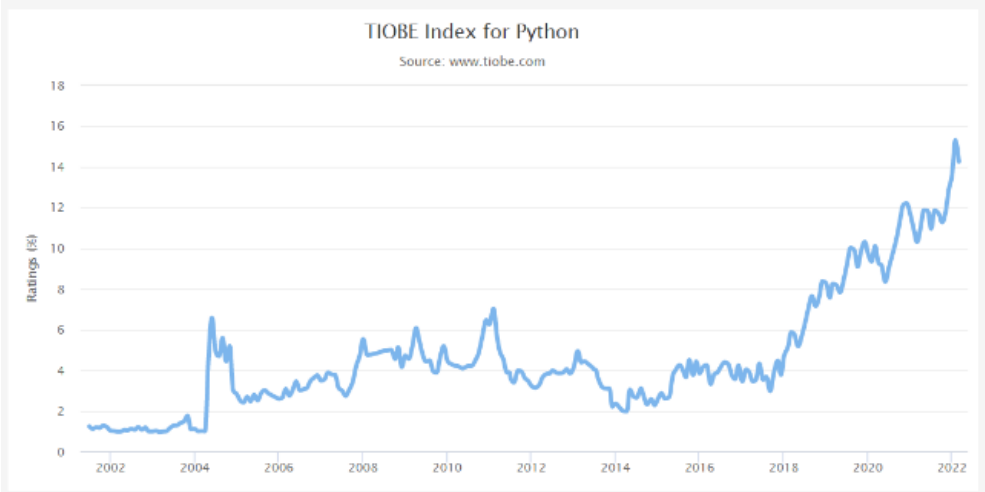
A Lakehouse az alábbi jellemzőkkel rendelkezik [13]:

- Tranzakciók támogatása: ACID tranzakciók segítségével biztosíthatóvá válik a konzisztencia. Ez kiváltképp fontos, hiszen több felhasználó hajthat végre író vagy olvasó műveleteket egy időben.
- BI támogatás: közvetlenül a forrásadatokon lehetséges a BI eszközöket használni, így az adatok frissítése adott a Lakehouse-ok által. Naprakész, friss adatok állnak rendelkezésre. Mindezzel a várakozási idő is csökkenthető.

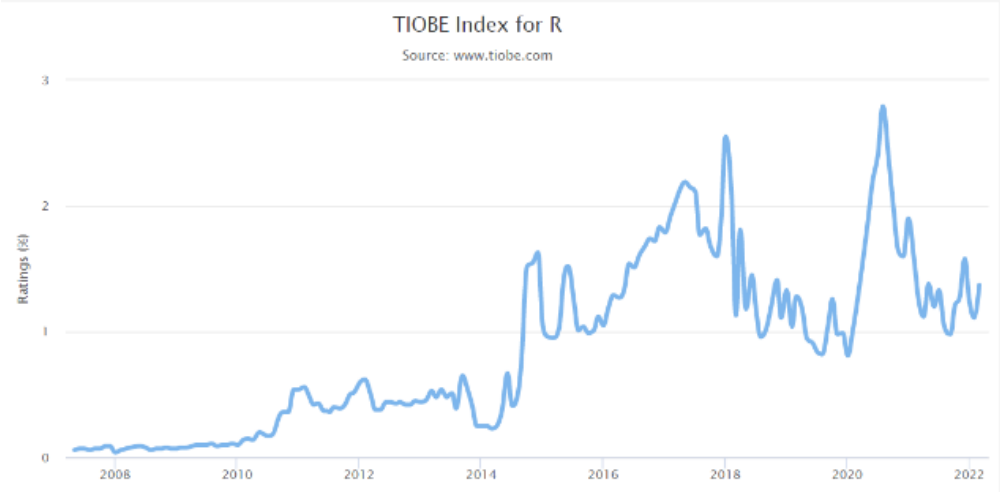
- A tárolás elkülönített a számítástól: azt jelenti, hogy a tárolás és a számítás nem ugyanazt a klasztert használja, ezáltal a rendszer több párhuzamos felhasználóra és nagyobb adatméretre skálázható.
- Nyitottság: Python/R könyvtárakat használ, amelyek segítségével biztosítva van az adatokhoz való hatékonyan és közvetlenül hozzáférés.
- Adattípusok támogatása: lehetőséget teremt a strukturálttól a strukturálatlan adattípusokig való használatra.
- End-to-end streaming: streaming támogatása.

3. TECHNOLÓGIA

Az idősoros elemzésre két fő programnyelv is jó megoldásnak tekinthető: az R és a Python. A két nyelv közötti választás alapja lehet a TIOBE index is, amely a Programozási nyelvek népszerűségének az indexe. [14] A Python 2022. márciusában az első helyezést érte el az R-rel szemben, az R pedig a 11. helyen szerepel, azonban a COVID óta az R egyre népszerűbb a programozók körében, hiszen egy év alatt 2 helyezést javított.
A 4. és 5. ábrán pedig a két programnyelv éves szintű értékelése látható.



4.ábra: Python- TIOBE index alakulása



5.ábra: R- TIOBE index alakulása

A döntést azonban nem csupán ez alapján lehet meghozni. Véleményem szerint azonban a feladatom megoldásához mindkét nyelv ideális lenne, rengeteg előnyük létezik, amelyet a Datacamp oldalán strukturáltan összegyűjtöttek, ahol a tanulók is sokszor felteszik ezt a kérdést. [15]

A cikkből az alábbiakat emelném ki:

Szempontok	Python	R
Használhatóság	Főként programozók használják.	Nem szükséges programozói tudás.
Előnyök	Kódolvashatósággal és a sebességével szerzett figyelmet. Megfelelően használható a matematikai számításokhoz és az algoritmusok működésének megértéséhez.	Sok funkcionalitás található benne az adatelemzéshez, az egyik legjobb eszköz a vizualizációhoz, főként statisztikai elemzésekhez tökéletes.
Hátrányok	Nem tartalmaz olyan nagyszámú könyvtárat a Data Science-hez, mint az R. A vizualizáció elkészítése összetettebb, mint a másik vizsgált programnyelvnél.	Főként statisztikusoknak lett tervezve, így nem a kódolás egyszerűségére törekedtek első sorban. Deep Learning-hez nem olyan elterjedt, mint a Python.

4. IDŐSORELEMZÉS

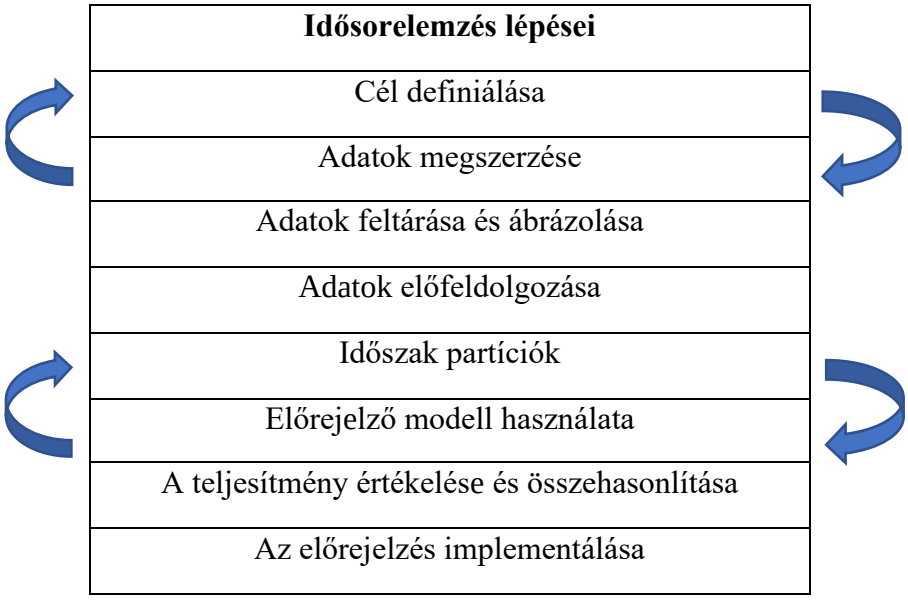
Az idősor definícióját Ben Auffarth [16] az alábbiak szerint fogalmazta meg: „Az idősorok olyan adatkészletek, amelyekben a megfigyelések időrendi sorrendben vannak elrendezve. Tehát egy idősorról $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})$ azt feltételezzük, hogy egyenlő időtávolságra lévő valós értékek sorozata, ahol $t = 1$-től $t = n$ időpontig történik a vizsgálat.”

Erre példa:

- a szakdolgozatomban is elemzett tőzsdei záró árfolyamok értéke;
- a tömegközlekedési eszközzel utazók számának alakulása napi szinten;
- heti autóbalesetek száma;
- az időjárás alakulása;
- népesség növekedése évenként.

Ez csupán néhány példa adat arra vonatkozóan, hogy milyen adathalmaz során szoktak az adatok időbeli változásával foglalkozni.

Gailt Shmueli Kenneth C. Lichtendahl Jr. könyvében az idősor elemzés lépéseinek az alábbiakat tűnteti fel.[17]



6. ábra: Az idősor elemzés lépései

Az idősor elemzéssel kapcsolatban kétfajta megközelítés vált ismertté: a determinisztikus és a sztochasztikus idősorelemzés. A determinisztikus feltevés azon alapszik, hogy az idősort befolyásoló tényezők teljes mértékben számbavehetőek, tehát megfelelő pontossággal fel lehet írni az idősort és a véletlennek „gyakorlatilag” kisebb szerep jut, csak hibával tudunk mérni. A véletlenről itt csak úgy beszélhetünk, mint ami az adott

időszakbeli pontos értéket annak megfelelően alakítja ki, hogy valamilyen mértékben eltér a becslésünktől. Ennek azonban az ezt követő időpontokra semmilyen hatása nem lesz. Később is tapasztalni fogunk véletlent, azonban az előző véletlen hatása már nem fog szerepet játszani. Ezt alkalmazzák főként akkor, ha a hosszútávú előrejelzés a cél. Ezzel szemben a sztochasztikus idősorelemzés filozófiája, hogy a véletlen eltéréseknek a jövőben is hatása lesz, vagyis a véletlennek úgynevezett folyamatépítő szerepe van. Ezt a rövid távú idősoros modellezésnél szokták használni. [18]

4.1. TREND -, SZEZONÁLIS-, CIKLIKUS-, IRREGULÁRIS MOZGÁS

Az idősoros adatok sokféle mintát mutathatnak, ezért hasznos az idősorokat összetevőkre bontani. Ezt teszik a dekompozíciós modellek, melynek a képlete [19]:

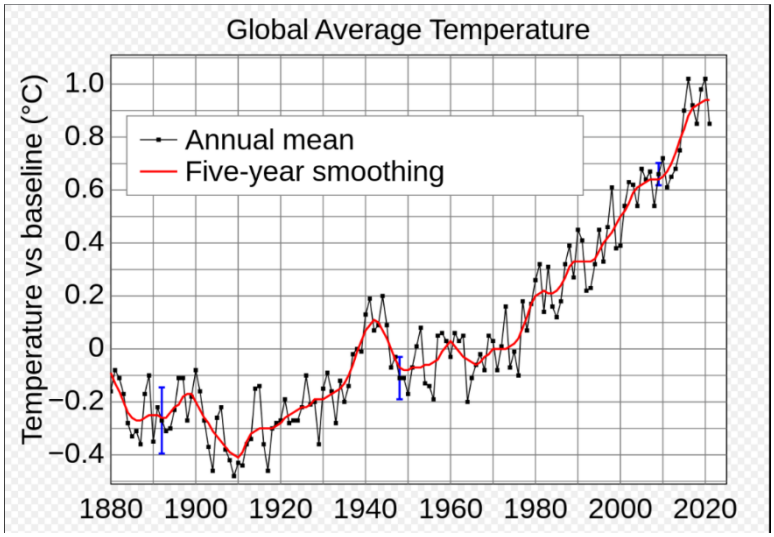
(1) $$X = T \times C \times S \times I$$

Az idősorok tehát az alábbiakra bonthatók:

- T, a trendkomponenst jelöli, amely a sorozat hosszú távú előrehaladását szemlélteti. Akkor beszélhetünk tendenciáról, ha az adatokban tartós növekedés, avagy csökkenés figyelhető meg. Ez a mozgás egy trendgörbével adható meg.

  Minden idősor elemzés esetén az első lépés, hogy a trendeket kiszűrjük, hiszen abban az esetben, ha az adatpontok átlagát vennénk, amely erőteljesen kiugró értékeket is tartalmaz és ezt az átlagpontot rávetítenénk a jövőre, tehát ezt a kivetítést illesztenénk egyenesként a jövőbeli adatokra, abban az esetben torz modellt kapnánk, hiszen a növekedést, illetve a trendet összemosná a modell és téves megállapításra jutna. Ezt a fajta hibát fontos elkerülnünk, ezért a trend meghatározására szokás használni a mozgóátlag technikát.

  A trendre szemléletes példa a 6.ábrán látható globális felmelegedés éves szintű alakulása [20]:



7.ábra: Globális felmelegedés alakulása

- C, a ciklikus mozgást jelenti. Az alábbi komponens a hosszú távú változásokat mutatja meg a trend körül, amely lehet periodikus és nem periodikus is. A lényege, hogy nem kell szükségszerűen ismétlődő mintát követnie, tehát nem ugyanolyan időközönként következnek be a változások.

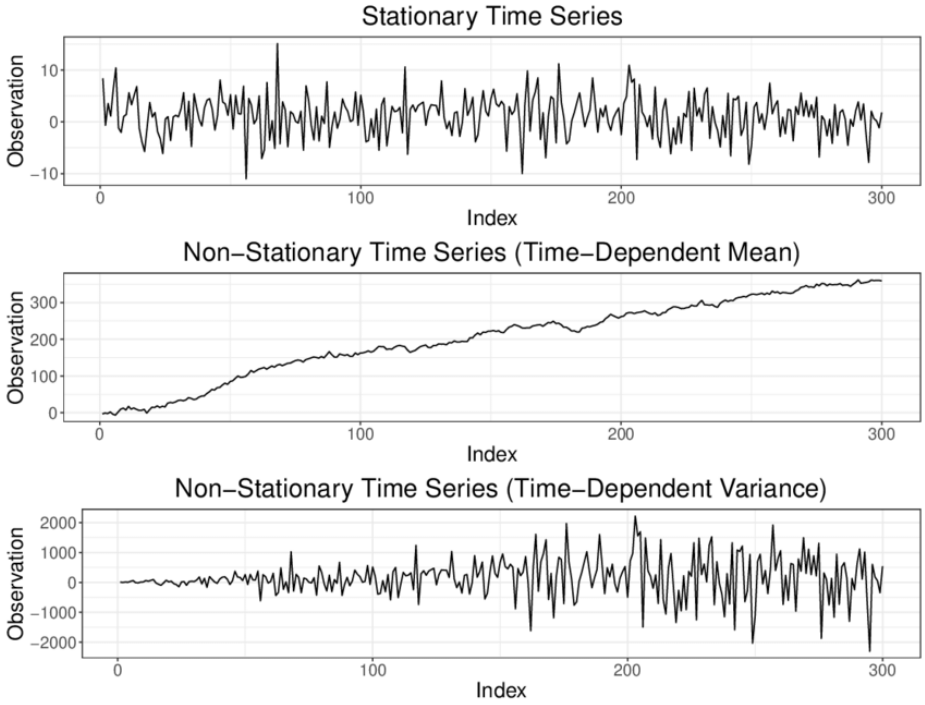
- S, azaz a szezonalitás komponens. Az alábbi változás adott időközönként (egy évnél rövidebb) jelentkezik. Például naponta, hetente, havonta vagy évente. Tehát a fő eltérés a ciklusokhoz képest, hogy egy évnél rövidebb a periódus ideje és mindig periodikus. Például az anyák napján megnövekedett virág eladások száma.
- I, azaz az irreguláris komponens. Az alábbi a véletlenszerű, szabálytalan hatásokat írja le. Erre példa: háborús események, Brexit, stb. [19]

A világ tőzsdéin az árfolyamok esetén nagy átlagban nagyon markáns szezonalitási trendek nem figyelhetőek meg egyértelműen, hiszen akkor a spekulánsok mindig az adott időpontban fektetnék be a pénzüket, ezzel nyereséget elérve és nem tekinthetnénk kockázatosnak a tőzsdei kereskedést. Azonban ebben az ágazatban sokkal inkább a gazdasági hangulat befolyásolja az árfolyamokat, amelyek kiszámíthatatlanok. Ezzel szemben a trend kiszűrésére figyelmet kell fordítani.

4.2. STACIONARITÁS

Az idősor elemzés során fontos, hogy stacioner legyen az adathalmaz. A stacionaritás azt jelenti, hogy egy idősort létrehozó folyamat statisztikai tulajdonságai nem módosulnak az idő előrehaladtával. Ebből nem az következik, hogy az adathalmazban nem áll be változás az idő múlásával, hanem azt jelenti, hogy a változás módja marad ugyanaz. Egy lineáris függvény lesz az algebrai kifejezése, vagyis egy lineáris függvény értéke változik, ahogy x nő, és maga a változás módja is állandó marad - konstans meredekséggel rendelkezik.
Erre azért van szükség, mert így jobban lehet elemezni, modellezni és az előrejelzés pontosságához is hozzájárul, enélkül téves becslést hozhatnánk létre. [21]

Az 7.ábra jól szemlélteti a stacionáris és nem stacionáris idősorokat [22]:



8.ábra: Példa a stacionáris és nem stacionáris idősorra

Megkülönböztethető erős és gyenge stacionaritás:

- Erős stacionaritás esetén Ferenci Tamás [23] oktatóanyagában lévő definíciót emelném ki: „Az (Y_t) idősor erős értelemben stacionárius, ha minden véges dimenziós vetületének együttes eloszlása eltolásinvariáns. Azaz $\forall k \geq 1$ esetén $\forall t_{1,\ldots,} t_k$ indexhalmazra $(Y_{t_1}, \ldots, Y_{t_k})$ és $(Y_{t_1+h}, \ldots, Y_{t_k+h})$ eloszlása megegyezik bármely $h \in \mathbb{R}$ esetén."
- Gyenge stacionaritásra pedig azért van szükség - ahogy a fent említett szerző is említi -, mert gyakorlatban nehezen ellenőrizhető. Tehát ebben az esetben már nem minden k-ra, hanem csupán $k = 1$ és $k = 2$ esetén vizsgálódunk. Csak azt

várjuk el, hogy a várható értékek megegyezzenek és a szórásnégyzetekben legyenek egyenlőek, azzal ellentétben, amelyet az erős stacionaritásnál elvártunk, tehát már nem a komplett eloszlásra követeljük meg. Így látszólag 4 megkötést kell kapni, amelyek az alábbiak:

- $k = 1$ esetén nem a teljes eloszlásnak, hanem csupán a várható értéknek kell egyezőnek lennie, így: $EY_{t_1} = EY_{t_1+h}$. Így a gyenge stacionaritás első megkötése, hogy legyen várható érték időben állandó:

(2) $$\boldsymbol{\mu_t = \mu}$$

- $k = 1$ esetén a szórásnégyzetnek kell egyeznie: $D^2Y_{t_1} = D^2Y_{t_1+h}$, a szórásnégyzet függvény időben állandó:

(3) $$\sigma_t = \sigma$$

- $k = 2$ esetén, ahol a várható érték megegyezik: nem található új megkötés, hiszen a $EY_{t_1} = EY_{t_1+h}$ és $EY_{t_2} = EY_{t_2+h}$ egyenlőségeket kell teljesíteni. A kikötés azonban az volt, hogy $t_1$és t_2 is tetszőleges, így ugyanúgy a $\mu_t = \mu$ egyenlőséget kell teljesíteni;
- $k = 2$ esetén, ahol a szórásnégyzetben egyezik esetben:

(4) $$D^2\begin{pmatrix} Y_{t_1} \\ Y_{t_2} \end{pmatrix} = D^2\begin{pmatrix} Y_{t_1+h} \\ Y_{t_2+h} \end{pmatrix}$$

Egy vektor D^2 -e pedig a kovariancia mátrixa, így az alábbi mátrix-szal írható fel:

(5) $$\begin{matrix} D^2Y_{t_1} & cov(Y_{t_1}, Y_{t_2}) \\ cov(Y_{t_2}, Y_{t_1}) & D^2Y_{t_2} \end{matrix} = \begin{matrix} D^2Y_{t_1+h} & cov(Y_{t_1+h}\, Y_{t_2+h}) \\ cov(Y_{t_2+h}, Y_{t_1+h}) & D^2Y_{t_2+h} \end{matrix}$$

Mivel a mátrixok akkor egyenlőek, amennyiben minden elemük egyenlő, így a $D^2Y_{t_1} = D^2Y_{t_1+h}$ *és* $D^2Y_{t_2} = D^2Y_{t_2+h}$ esetén szintén arra a megkötésre lehet jutni, hogy a szórásnégyzet függvény időben állandó. A harmadik egyenlőségből $-\, cov(Y_{t_1}, Y_{t_2}) = cov(Y_{t_1+h}\, Y_{t_2+h})\, -$ pedig az következik, hogy a kovariancia eltolás invariáns, tehát:

(6) $$\boldsymbol{\gamma_{t,s} = \gamma_{t+h,s+h}}$$

Ha megtartom azt a távolságot (késleltetés), amellyel eltolom, abban az esetben ugyanazt a kovarianciát kapom.

Tehát a gyenge stacionaritás esetén két megkötést kell teljesíteni, amelyet a (2) és a (6). egyenlet szemléltet. A (3) egyenlettel jelzett megkötés, amely kimondja, hogy a szórásnégyzet függvény időben állandó legyen, az tulajdonképpen redundáns egyenlet, hiszen, ha a hatodik egyenletben $s = t$ egyenlőséget használom, akkor a harmadik egyenletet kapom. [24]

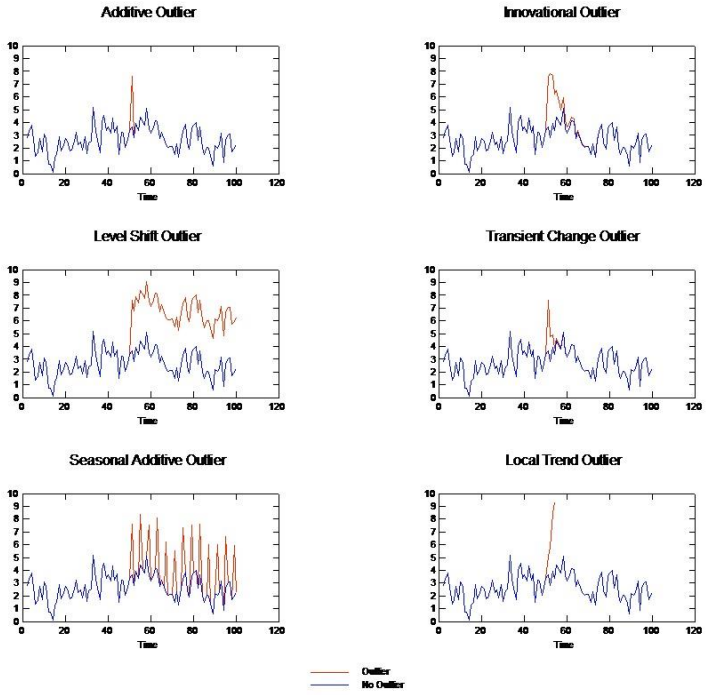
Néhány idősor alapvetően kielégíti a gyenge stacionaritás megkötéseit, azonban, ha ez nem valósul meg, akkor biztosítanunk kell az elemzéshez ezt a feltételt, amelyhez transzformációk használatára van szükség.

4.3. KIUGRÓ ÉRTÉKEK DETEKTÁLÁSA

Az idősor elemzés során feladat a kiugró értékek detektálása, az alábbi fejezetben csak a definícióját és típusait sorolom fel és az elemzés fejezetben fogok kitérni a kezelésére.
D.M. Hawkins [25] az alábbiak szerint definiálja: „a kiugró érték egy olyan adatpont, amely olyan jelentősen eltér a többi megfigyeléstől, hogy azt más mechanizmus is létrehozhatta volna.” Azonban számos definíció elterjedt a szakirodalomban, amelyek közül Hawkinst emelném ki.
A kiugró értékeknek több típusa ismert [26]:

- Additív kiugró érték: egy megfigyelés esetén van hatása, tehát csak az adott időpontot befolyásolta. Ezután pedig folytatódik az idősorozat, úgy mintha az adott esemény be sem következett volna.[27]
- Az innovatív kilógó érték ezzel ellentétben már a bekövetkezés időpontjától kezdve, minden megfigyelésnél érzékelhető lesz.[27]
- A szint eltolódás (level shift outlier): a kiugró érték után megjelenő megfigyelések új szintre kerülnek, ez sok megfigyelést érint és állandó hatással bír.
- Az átmeneti változás (transient change outlier) esetén a kiugró érték hatása a következő megfigyeléseket tekintve exponenciálisan csökken.
- Szezonális additív kiugró érték: az adott megfigyelés értéke kiváltképp nagy vagy kicsi, amely rendszeresen időközönként ismétlődő értékként tapasztalható.
- Lokális trend kiugró érték (local trend outlier): ez a kiugró érték is az idősor eltolódását okozza, amely után további kiugró értékek is felbukkannak.

Az említett típusokat pedig a 9. ábrán szemléltetem:



9.ábra: Kiugró értékek típusai

5. BOX JENKINS MÓDSZER

Az idősorelemzésben a Box-Jenkins-módszer [28] vált ismertté, amely az elnevezését George Box és Gwilym Jenkins statisztikusokról kapta, autoregresszív mozgóátlag vagy autoregresszív integrált mozgóátlag modelleket alkalmaz, amiatt, hogy az idősor múltbeli értékeihez megtalálja egy idősor modelljének legjobb illeszkedését.

Az alábbi módszer 3 lépésből áll:

- modell azonosítása;
- paraméterek becslése;
- modell ellenőrzése.

A modell azonosítása esetén az első lépés, hogy a változók stacionáriusak legyenek. Ezenfelül fontos, hogy az idősor szezonalitása azonosítva legyen és a szezonalitás el legyen távolítva a modellből, amelyet szezonális kiigazításnak neveznek.
A stacionaritás és a szezonalitás az autokorrelációs diagram (ACF) segítségével megállapítható. Ez megmutatja, hogy az időben adott távolságokra (lag) lévő értékek mutatnak-e korrelációt, tehát az idősor elemei pozitívan, negatívan függenek egymástól, avagy függetlenek. Az autokorrelációs diagram az autokorrelációs függvény értékét mutatja a függőleges tengelyen, melynek értékkészlete: [-1, +1]. [29] A nem stacionaritást gyakran a nagyon lassan csökkenő autokorrelációs diagram jelzi.
Box és Jenkins a differenciálási megközelítést javasolja a stacionaritás elérésére, azonban megemlíti, hogy egy görbe illesztése és az illesztett értékek kivonása az eredeti adatokból is egy megfelelő megoldást jelenthet.
Az ezt követő lépés pedig az autoregresszív és mozgóátlag sorrendjének (p és q) meghatározása. Brockwell és Davis [30] az Akaike információs kritérumot tekintik meghatározónak, azonban más szerzők a fent említett autokorrelációs és részleges autokorrelációs (PACF) diagramot használják. Tehát a modell első lépése az ARIMA p, d, q értékét adja meg.

A második lépés a paraméterek becslése, amely nemlineáris egyenletek megoldásainak numerikus közelítését fejezi ki. Erre a gyakorlatban rendelkezésre állnak statisztikai csomagok, így erre a lépésre nem térek ki bővebben.

Az utolsó lépés pedig a modell által előre jelzett adatok összevetését jelenti az eredeti adatokkal.

5.1. ARIMA MODELL

A szakdolgozatom során az ARIMA modellt választottam, amelynek oka, hogy az irodalomkutatásom során kiváltképp nagyszámú hivatkozásként bukkant fel az ARIMA modell, amely ezen felül széles körben ismert, elemzett és bizonyított módszertanról tett tanúbizonyságot.

Az ARIMA az Autoregressive Integrated Moving Average angol szó rövidítése, amely az ARMA modell általánosítása. [31] Az imént említett modell a stacionárius sztochasztikus folyamatok leírására alkalmas, amely két komponensből épül fel [32][33]:

- AR(p), autoregresszív modell rész: előrejelzést végez a vizsgált változó múltbeli értékeinek lineáris kombinációjával. Tehát azt jelenti, hogy a változó értéke függ ugyanezen változó korábbi értékeitől.

Képlete:

(7) $$y_t = c + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \varepsilon_t$$

- MA(q), mozgóátlag modell rész: az autoregresszív modellel ellentétben a mozgóátlag modell a múltbeli előrejelzési hibákat használja egy regressziószerű modellben. Fontos megemlíteni, hogy a mozgóátlag modellek nem azonosak a mozgóátlag-simítással. A mozgóátlag modell a jövőbeli értékek előrejelzésére, míg a mozgóátlag-simítás a múltbeli értékek trendciklusának a becslésére szolgál.

Képlete:

(8) $$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

Abban az esetben pedig, ha a differenciálást az autoregresszióval és mozgóátlag-modellel kombinálom, akkor megkapom az ARIMA-modellt, amely a következőképp írható fel:

(9) $$y'_t = c + \emptyset_1 y'_{t-1} + \cdots + \emptyset_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Tehát $ARIMA(p, d, q)$-nak használja a szakirodalom, ahol:

- p: az autoregresszív rész paramétere;
- d: a differenciálás mértéke/foka;
- q: a mozgóátlag rész paramétere.

6. ELEMZÉS

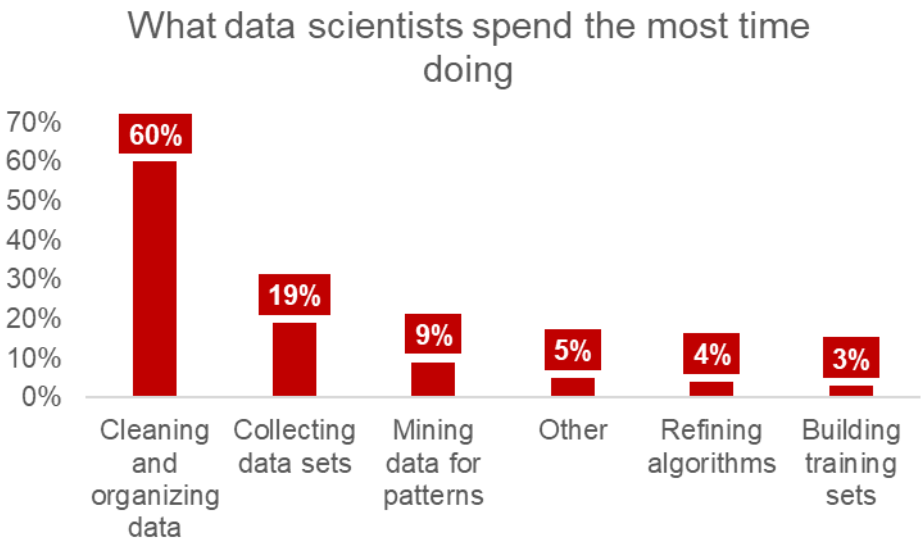
A szakdolgozat elején említésre került, hogy mely részvény indexek esetén jelzem előre az árfolyamokat. Azonban a szakdolgozat keretein belül csak egy index kerül szemléltetésre, hiszen az összes index elemzése túlmutatna a megszabott oldalszámon, illetve mindnél ugyanazokat a vizsgálatokat hajtom végre, így egy index esetén szemléletesen bemutathatóak az elemzési lépések.
Következésképpen az indexek statisztikai mutatószámai, illetve az alkalmazott modell paraméterei eltérőek, azonban a módszertan az összes esetén ugyanaz az ARIMA modell kiépítése lesz.
Az általam kiválasztott index az S&P500, hiszen az egyik legismertebb, a világon leginkább követett és a New York-i tőzsdét leginkább reprezentáló indexnek tekinthető és az USA teljes tőzsdepiacának 80%-át teszi ki.[6]

Az elemzés során az Adjust Close Price-t választottam az időtől függő változónak, a fentebb említett idősor adatok pedig nem kerültek a vizsgálat tárgyába, hiszen az Adjust Close Price nem csak a záró árat, hanem már egy korrigált értéket vesz figyelembe, amely a legjobban mutatja az index alakulását.

Fontos megemlíteni, hogy minden egyes elemzés előtt az alábbi lépések hajtandók végre:

- Adatgyűjtés: a releváns adatok adattárházból, egyéb adatforrásból való összegyűjtése. Az én elemzésem során ezek az adatok a Yahoo Finance weboldala segítségével állnak rendelkezésemre, ahol az indexek jelölését ismerve automatikusan a Python programnyelv segítségével le lehet hívni.
- Adatfeltárás: meg kell érteni, hogy az adathalmaz milyen mintázatot tartalmaz, így az elemzés előtt kapunk egy képet az adatokról. Itt főként az inkonzisztencia vizsgálatra, a hiányzó értékek eltávolítására és kezelésére lehet gondolni. Az általam vizsgált idősor esetén is fontos a hiányzó értékek tanulmányozása, hiszen az torzítaná az elemzést.
- Adattisztítás: a hibás adatok kijavítása, pótlása (imputáció) érthető ez alatt.
- Adatok átalakítása: az adatok aggregálása, új oszlopok létrehozása, adatok előszűrése is egy lényeges lépés az elemzés előtt. Mindez megkönnyíti az elemző munkáját, hiszen már egy olyan adathalmazon tud dolgozni, amelynél az adatok vizualizációját követően összefüggéseket tud keresni.

Mindezen felvetésemre egy a Forbes által publikált felmérés is alátámasztást nyújt. A kutatás azt bizonyította, hogy az adatkutatók a munkájuk 80%-át az adatelőkészítéssel töltik. [34] Ezen kutatás végeredményét pedig a 10. ábra szemlélteti:



10. ábra: Adatkutatók által végzett munkák megoszlása

6.1. ELŐELEMZÉS

Az alábbi fejezetben az S&P500 idősorát szemléltetem és vizsgálom az egyes statisztikai mutatóit.
Az előkészület során fontos leellenőrizni, hogy az idősor ne tartalmazzon hiányzó értéket, hanem egy összefüggő, ugyanannyi adatpontot tartalmazó adathalmaz kerüljön elemzésre az összes indexet vizsgálva. Ezt tanulmányozva pedig a kiválasztott index adataiban nem található N/A adat.

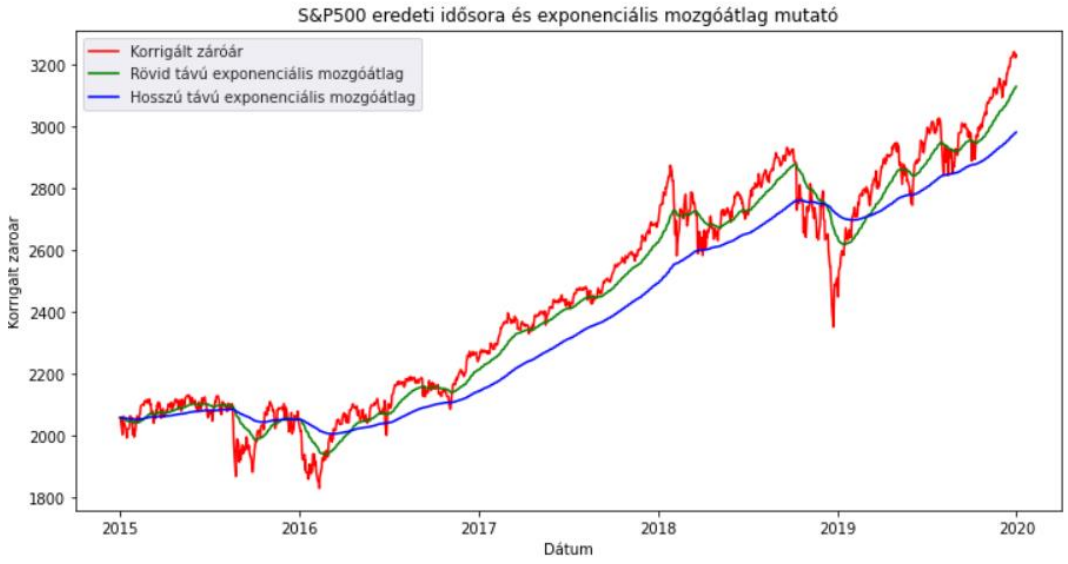
Az elemzett Adjust Close Price statisztikai mutatóit pedig a 11. ábra szemlélteti. Jól látható, hogy a részvényárfolyam széles intervallumot mutat, hiszen a vizsgált időszakban az index árfolyama 1800 és 3200 pont között mozgott, sőt a medián 2434-es értékkel viszonylag magasnak is mondható. A képen látható kvartilis értékek pedig a későbbi fejezetekben még előtérbe kerülnek, hiszen ezek szükségesek egy dobozdiagram ábrázolásához is.

S&P500 statisztikai mutatói

count	1258.000000
mean	2452.643027
std	357.451845
min	1829.079956
25%	2102.082520
50%	2434.145020
75%	2771.179993
max	3240.020020

11.ábra: S&P500 statisztikai mutatói

Az elemzéshez ezen kívül még fontos az adatok idősoros ábrázolása is, az adatok előtanulmányozása, amelyet a 12. ábra szemléltet.



12.ábra: S&P500 idősora

A fenti diagramot tekintve az eredeti adatsor, illetve egy rövid és egy hosszú perióduson vizsgált exponenciális mozgó átlag (EMA) görbe látható. Mindebből már lehetne egy olyan stratégiát felállítani, hogy a rövid és hosszú távú EMA mutatót megvizsgálva, ahol a rövid távú mutató a hosszú távú alatt van, akkor medve piac tapasztalható, és ilyenkor érdemes a befektetőknek eladni a megvásárolt részvényindexeket. Azonban ez a taktika nem hozott kiemelkedő eredményt, így ezt nem vettem bele a vizsgálatba.

Ezt követően az egyik legfontosabb lépés a trend, szezonalitás elemzése a vizsgált idősorban. Az alábbiakban használni lehetne a Hodrick Prescott szűrőt, azonban ezt inkább makro, aggregált adatokra fejlesztették ki. A szakdolgozat során azonban napi adatok álltak rendelkezésemre, amelyet természetesen át lehetne alakítani a Python programnyelv által biztosított resample függvénnyel, de a feladatom során napi adatokat vizsgálatát találtam célszerűnek. Így egy másik csomaggal dolgoztam a trend felismerése során. [35]

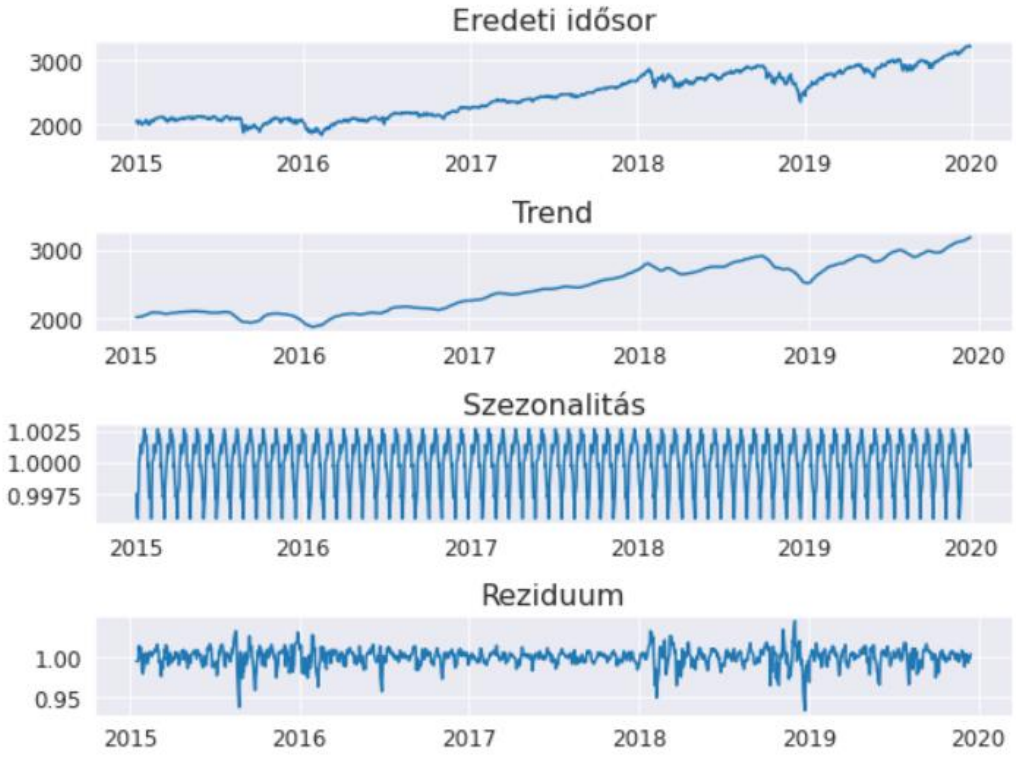
Ezenfelül ismert még a Holt's Winter metódus, amelyet alkalmazni szoktak a szezonalitás kiszűrésére. [33] Azonban a gyakorlatban az egyik legelterjedtebb a Python programnyelvben a seasonal_decompse használata, amelynek segítségével megállapítható, hogy az idősor bizonyos periódusonként tartalmaz -e ciklust, trendet, illetve egyéb hibatagot.

A modell additív vagy multiplikatív értékre is állítható. [36] A megfelelő modell kiválasztásának a szabálya, hogy a diagramunkban meg kell nézni, hogy a trend és a szezonális változás időben viszonylag állandó, tehát lineáris-e. Ha ez a feltétel teljesül, akkor az additív modellt kell választani. Ellenkező esetben, ha a trend és a szezonális

változás idővel nő vagy csökken, akkor a multiplikatív modell használata a javasolt. Ebben az esetben a multiplikatív került kiválasztásra, hiszen nem teljesen lineáris trend fedezhető fel az adatsorban. A periódus esetén pedig 20 napot vizsgáltam.

Ezenfelül a Python által biztosított Seasonal-Trend decomposition using LOESS (STL) segítségével is megvizsgáltam a trendet, ciklust és szezonalitást. Mindkettő természetesen ugyanazt eredményezte, csupán azt szerettem volna érzékeltetni, hogy számtalan lehetőség ismert a programnyelvben, amelynek segítségével megállapíthatóak az egyes torzító hatások.
Ennek segítségével kirajzoltattam az eredeti idősort, a trendet, a szezonalitást és az úgynevezett maradékokat, amelyet a 13. ábra mutat.

Látható, hogy az idősor nem stabil és tartalmaz trend komponenst, emiatt is szükséges az adatok további előfeldolgozása, elemzése.

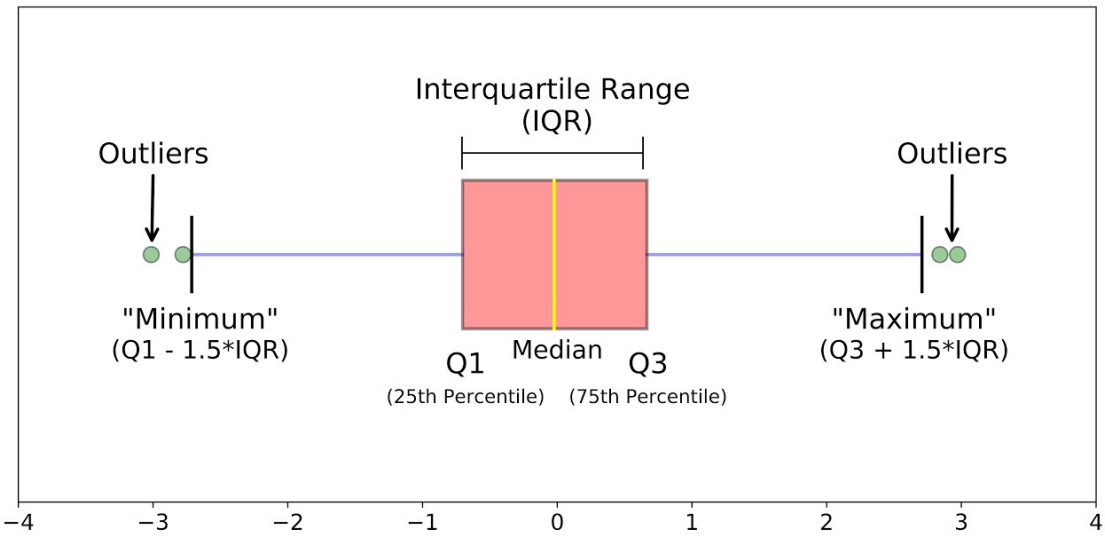


13.ábra: S&P500 idősor adatainak szétbontása

Az ábrán is látható trend eltávolítására a szakértők sűrűn használják a differenciálást, amelyet a stacionaritás fejezetben fogok bővebben bemutatni. Az elsődleges célom annak a demonstrálása volt, hogy az idősor trendet tartalmaz, amely miatt nem jelezhető még előre a jövőbeli várható árfolyam.

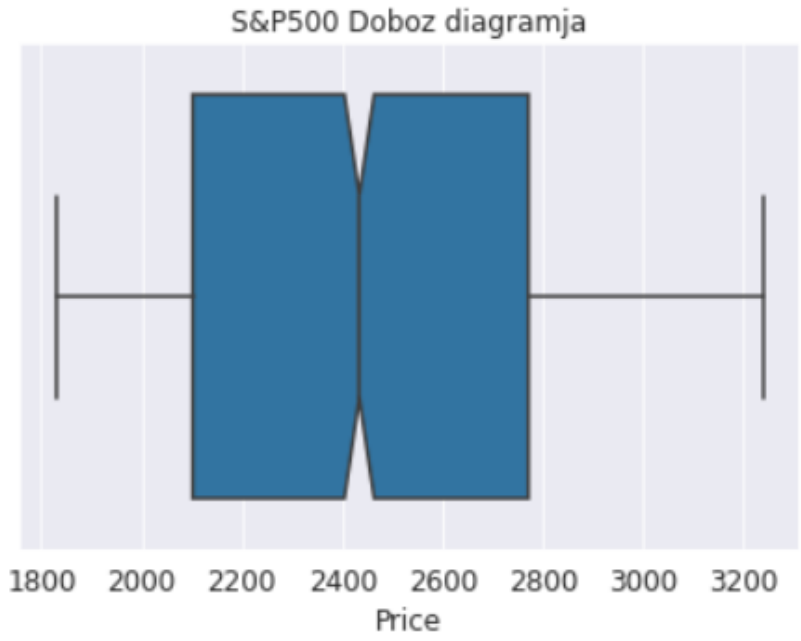
6.2. KIUGRÓ ÉRTÉKEK ELEMZÉSE

Az adatok tartalmazhatnak hibás vagy szélsőséges kiugró értékeket, amelyeket azonosítani és kezelni szükséges.
Ahogy fentebb bemutattam az általam vizsgált adatsor sem stacionárius, látható, hogy tartalmaz szélsőséges értékeket, illetve a trend komponens is azonosíthatóvá vált. Így először megvizsgáltam, hogy az adatsor dobozdiagramja mit mutat meg számomra. A diagram megértéséhez a 14. ábra szemléletes. [37] Eszerint az az adatpont tekintendő kilógó értéknek, amely esetén az értékek az ábrán feltüntetett minimum és maximum értéken kívül esnek. Szükséges tehát az egyes kvartiliseket megvizsgálni, és ez alapján megállapítani, hogy az adott adatpont anomáliának tekinthető-e.



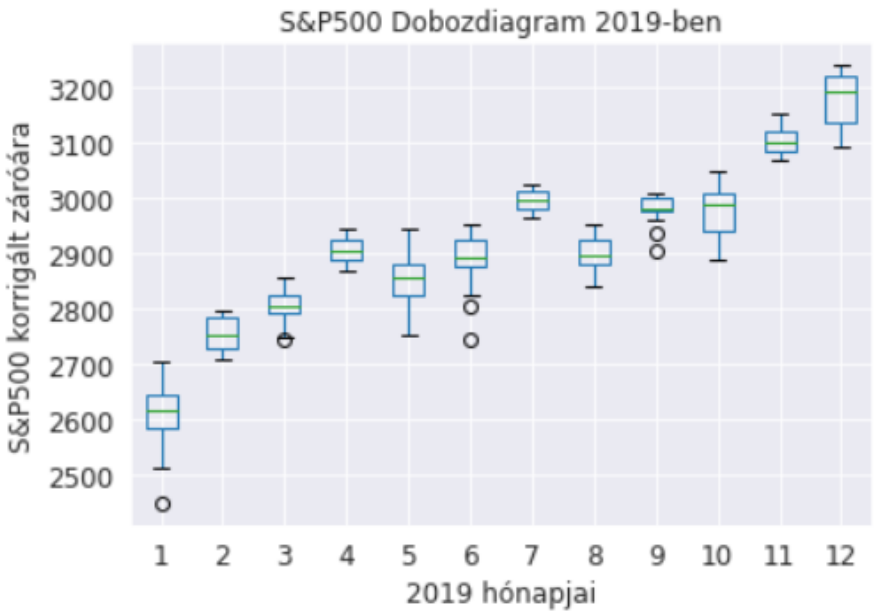
14.ábra: A doboz diagram számítása

Az S&P 500 adatsorát tekintve, azonban nem találtam az alábbi számolás szerinti kilógó értékét, amelyet a 15. ábra is mutat:



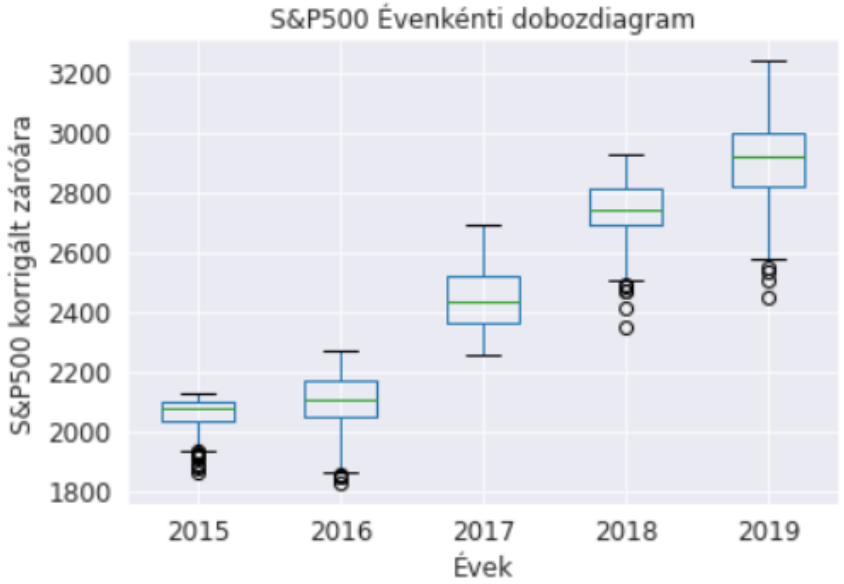
15. ábra: S&P500 doboz diagramja

Habár az adatsornál szemmel látható, hogy vannak kilógó értékek, azonban abban az esetben, ha egy idősor trendet tartalmaz, akkor a kilógó értékek azonosítása doboz diagrammal kétséges, hogy helyesen kivitelezhető. Befolyásoló tényező, hogy az évek között jelentősen eltérő nagyságú index értékek találhatóak. Abban az esetben, ha megvizsgálom azt, hogy a 2019-es év esetén, hogy nézett ki az adatsor havi doboz diagramja, akkor már látható, hogy vannak kiugró értékek hónapon belül. Tehát egy éven belül az egyes hónapokat vizsgálva már a trend nem torzította olyan mértékben az adatokat, hogy ne lennének azonosíthatóak a kiugró értékek. Mindezt a 16. ábra mutatja be, amely azt szemlélteti, hogy 2019-ben az egyes hónapokat tekintve milyen medián részvényárfolyam mutatkozott, illetve melyik hónapban találhatóak kiugró értékek. A diagramon a kiugró értékeket a körrel jelölt egyes adatpontok jelentik.



16. ábra: S&P500 doboz diagramja 2019-ben

Mindez megvizsgálható éves bontásban is, kilistázva, hogy az adott vizsgált évben mennyi kiugró érték látható. Ezt mutatja a 17. ábra, ahol a 2017-es év esetén kiugró érték nem fedezhető fel. Az már az idősorról is megállapítható volt, hogy nagy mértékű eltérés az előző adatpontokhoz képest nem figyelhető meg, sokkal inkább a trend jelensége tapasztalható, hiszen egy folyamatosan közel azonos meredekséggel emelkedő egyenes lenne ráilleszthető a 2017-es évet vizsgálva.



17. ábra: S&P500 doboz diagramja éves bontásban

A kiugró értékek észlelésére ajánlott megoldás még a z pontszám alkalmazása is, amely az alábbi képlettel írható le:

(10) $$(X_i - átlag)/szórás$$

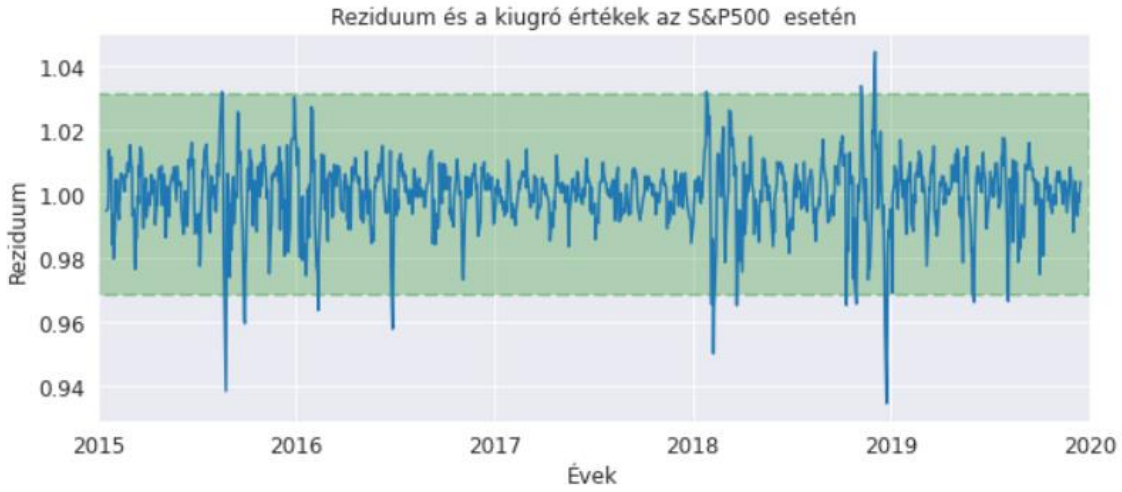
Abban az esetben, ha ezen érték bármelyik adatpontot vizsgálva nagyobb, mint a küszöbérték, azaz 3, akkor kiugró értéknek tekinthető. [38] Azonban itt ugyanaz a megállapítás tehető, mint a teljes idősor doboz diagramjának a vizsgálatakor. A trend nagyban befolyásolja az adatsort, így ezen módszerrel nem volt megállapítható kiugró érték.

Ennél azonban célszerűbb a reziduumokat, trendet és szezonalitást vizsgálni. Elsőként azt tanulmányoztam, hogy az eredeti idősorra, hogy illeszkedne a trend és a szezonalitás összege.
Tehát a reziduumok nélkül az adatsor, hogy nézne ki, hogy simulna az eredeti idősorra, amelyet a 18. ábra szemléltet:

18. ábra: S&P500 eredeti és hibatagok nélküli idősora

Az alábbi diagramról már leolvasható, hogy melyik nap látható nagyobb mértékű hibatag, de ezt a 19. ábra még szemléletesebben bemutatja.

A szélső értékek megállapítása azon alapszik, hogy alsó értékként definiálom azt, ahol a zaj átlagából kivonom a reziduum szórásának háromszorosát. A felső érték esetén pedig az átlaghoz hozzáadtam a szórás értékének a háromszorosát.
Elsődlegesen statisztikai próbálkozásaimmal választottam ki a háromszoros értéket, másrészt módszertani bevett gyakorlat az 1,5-3 -szoros szórás érték használata kiugró érték azonosításakor. [39] Esetemben a 3 tűnt a leginkább megfelelőnek.
Ezt mutatja a 19. diagram, ahol a két érték közötti sáv nem számít kiugró értéknek.



19.ábra: S&P500 hibatagjainak alakulása

Amennyiben a hibatag magasabb, mint a korábban definiált magasabb érték, vagy ha a hibatag alacsonyabb, mint a korábban említett alacsonyabb érték, abban az esetben anomáliának tekinthető, tehát ezzel megtaláltam a vizsgált időszakban lévő kiugró értékeket.

Mindezt az eredeti idősor ábrázolását követően piros pontokkal jelölve szemléltettem a 20. diagramon:

20.ábra: A kiugró értékek definiálása

Az alábbi fejezet célja szintén csak a kiugró értékek bemutatása volt, ezek simítása a következőkben kerül részletezésre.

6.3. IDŐSOR ÁTALAKÍTÁSA, STACIONARITÁS BIZTOSÍTÁSA

Ahogy fentebb is látható ez az idősor még nem jelezhető előre, hiszen trendet és kiugró értékeket is tartalmaz. Azonban a stacionaritás megállapítására a szakirodalom az Augmented Dickey Fuller tesztet használja.[40] Mindez azt teszteli, hogy van-e egységgyök folyamat egy adott idősorban.

Az alábbi hipotézis írható fel ehhez:

- H_0 = Egységgyök folyamat van az idősorban. Az idősor nem stacionárius, van időfüggő szerkezete, nincs állandó eltérése az időben.
- H_a = Nincs egységgyök folyamat az adott idősorban. Az idősor stacionárius.

A képlete pedig a következő:

(11) $$\Delta y_t = \alpha + \beta_t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t$$

Mindezt az S&P500 idősoránál is szükséges volt lefuttatni, amelynek a végeredményét a 21. ábra szemlélteti:

```
ADF Statistic: 0.036464
p-value: 0.961446
Critical Values:
        1%: -3.436
        5%: -2.864
        10%: -2.568
```

21. ábra: Augmented Dickey Fuller teszt eredménye

A teszt p értéke 0,96, amely nagyobb, mint a 0,05, így a null hipotézis elfogadható. Mindez azt jelenti, hogy az idősor jelenleg még nem stacionárius. Így az adatsor várható értéke nem állandó, nem egy vízszintes egyenes körül ingadozik.

Az adathalmaz stacionáriussá való alakítása során az első lépés, hogy a kiugró értékeket kezelem, amelynek több megoldása is ismert. Egyik megoldás, hogy az adott idősorból kitörlöm az adott napi kiugró értéket, azonban ezt az idősor folytonosságának a megszakítása miatt nem tekintem megfelelőnek, illetve az „n/a” értékek sem megfelelőek egy ilyen elemzés esetén.
Lehetséges akár a kiugró értékek kezelése az átlag és a medián segítségével, azonban az átlag, mint mérőszám sok esetben szintén nagyon torzító hatású lehet a szélsőséges értékek miatt. Az S&P500 idősorában ez nem lenne olyan mértékben befolyásoló, hiszen az átlag és a medián érték is közel megegyezik, azonban, ahogy fentebb is látható volt, az

idősor elején is vannak kiugró értékek, ahol még az árfolyam jóval alacsonyabb, emiatt egy magasabb medián értékkel való kicserélés nem lenne célravezető.
Mindezen vizsgálat után az exponenciális mozgó átlag technikát alkalmaztam, amelyet a korábbi fejezetben is már bemutattam, csak itt az alfa paramétert a com és nem a span paraméter segítségével állítottam be, amely a 22. ábrán látható.

22.ábra: Exponenciális mozgó átlag (com paraméterrel)

Az ábráról is megfigyelhető, hogy ezen átalakítással nem történt nagy mértékű változás a kiugró adatpontok tekintetében, illetve az adatok szórását sem befolyásolta az elvárt mértékben. Habár látszik az ábrán, hogy a 2019-es évben lévő kiugró érték már jobban közelít az előző adatponthoz, azonban ezt nem tekintem elfogadhatónak. A megoldásnál megfelelő lenne a korábban bemutatott rövid periódusú exponenciális mozgó átlag a kiugró értékek csillapítására, ahol a simítás jobban befolyásolta az index értékek alakulását. Azonban szeretném bemutatni az egyszerű mozgó átlag technikát is, amelynek segítségével hasonlóan jó megoldás érhető el a kilógó értékek simítását tekintve.

A következő lépésben így az egyszerű mozgó átlag technikát alkalmaztam, ahol 45 napos periódust vettem. A kiugró értékek kezelését egy 45 napos mozgó átlag technikával valósítottam meg, azonban fontos kiemelni, hogy csak a kiugró értékek felülírása, kezelése lett végrehajtva, a nem kiugró értékek esetében az adatsor megváltoztatására nincsen szükség. Ezen eljárás esetén azonban fel kell arra is hívni a figyelmet, hogy egyes adatpontok (idősor kezdeti és záró értékei) „n/a” értéket vesznek fel, hiszen 45 nap átlagát számítja ki a modell, amely a 40. napon még nem áll rendelkezésre. Azonban ez most nem befolyásoló hatású, hiszen csak adott adatpontok cseréje fog bekövetkezni. Az eredmény pedig a 23.ábrán látható.

23. ábra: S&P500 eredeti és 45 napos mozgó átlag értékei

Így a kiugró értékek simításával a 24. ábrát kaptam eredményül. Fontos megjegyezni, hogy a metódus sok kiugró értéket nem detektált, így főként a 2019-es év esetén szembetűnő az adatsor változása a kiugró értékek simításával.

24.ábra: Idősor simítása (kiugró értékek kezelése)

Ezzel azonban még csak a kiugró értékek kezelése történt meg, a következő lépés esetén az idősort fehérzaj folyamattá kell alakítani, hogy stacionáriusnak tekinthessem. Ehhez az első lépés, hogy a trendet eltávolítottam a differenciálás módszerével, ahol az adott adatpontból kivontam az előtte lévő értéket, így a 25. ábrán lévő adathalmazt kapom:

25.ábra: S&P500 idősorának differenciálása

Ezzel a módszerrel a 26. ábrán látható eredményt értem el az Augmented Dickey Fuller teszt statisztikája alapján:

```
ADF Statistic: -13.415763
p-value: 0.000000
Critical Values:
        1%: -3.436
        5%: -2.864
        10%: -2.568
```

26. ábra: Augmented Dickey Fuller teszt eredménye (első differencia képzés)

Jól látható, hogy a p érték nagyon kicsi, így hat tizedesjegyig nem is megjeleníthetőek az értékei, így már az idősor stacionáriusnak tekinthető.
Tulajdonképpen ez a differenciálás lesz látható a következő fejezetben is, amikor a megfelelő mutató alapján kiválasztásra kerül az ARIMA modell, ahol szintén 1 lesz az integrált tag értéke. Tehát a korábban említett d paraméter értéke lesz egyenlő eggyel.

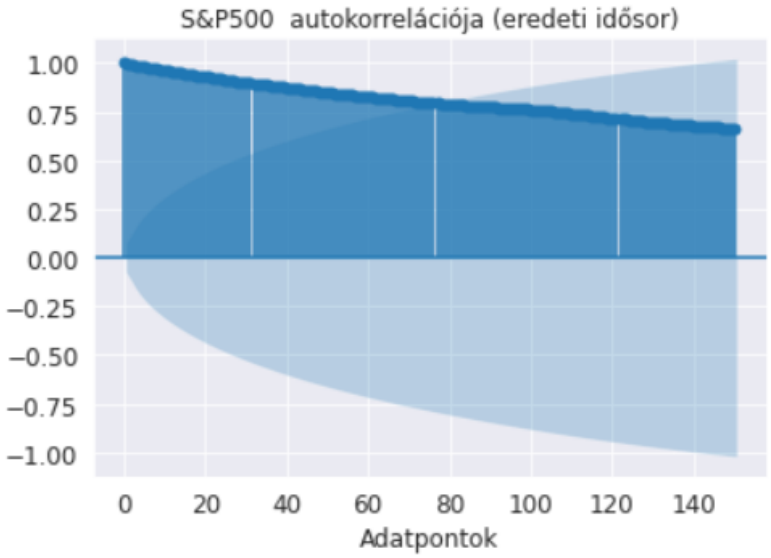
6.4. ARIMA MODELL KIÉPÍTÉSE

A következő fejezetben kerülnek kiválasztásra a p, d, q értékek.

Ahogy korábban bemutattam az adatsorhoz elegendő volt egy differenciálás az Augmented Dickey Fuller teszt alapján, amelynek köszönhetően az idősor már stacionárius, tehát az ARIMA modell segítségével előrejelezhető.

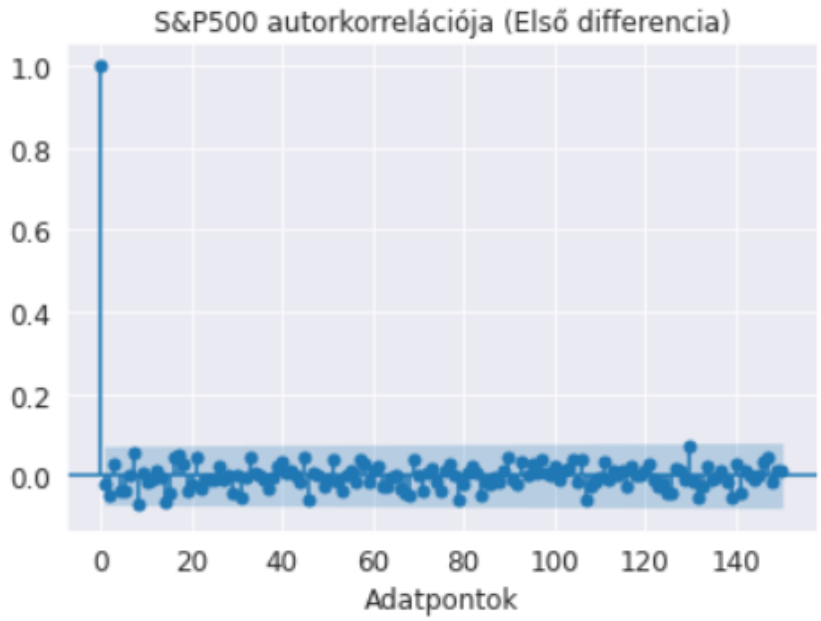
Ehhez egy másik hasznos technika az autokorreláció vizsgálata, amelyet a Python a plot_acf-el biztosít.
Az autokorreláció azt jelenti, hogy önmagával van korrelációban az adott adatpont, tehát a két érték esetén megállapítható, hogy milyen mértékű összefüggés található. Amennyiben ez az érték magas, akkor közel megegyező értékkel rendelkeznek és korrelálnak egymással. [41]

Első lépésben így megvizsgáltam, hogy az eredeti adathalmazra milyen az autokorreláció, abban az esetben, ha lassan cseng le, tehát sok adatpontra magas az autokorreláció, akkor differenciálás szükséges, hiszen a stacionaritást biztosítani kell. Ezt a differenciálást már az Augmented Dickey Fuller teszt alapján korábban megtettem, azonban fontosnak tartottam ennek a technikának a bemutatását, amely hasznos eszköz volt a vizsgálódás során.



27.ábra: S&P 500 autokorrelációja

A 27. ábrán jól látható, hogy az eredeti adathalmaz, ahogy azt korábban is bemutattam, differenciálást igényel. Így az autokorrelációs ábrázolást a differenciált idősorra is elvégeztem, amely a 28. ábrán figyelhető meg.



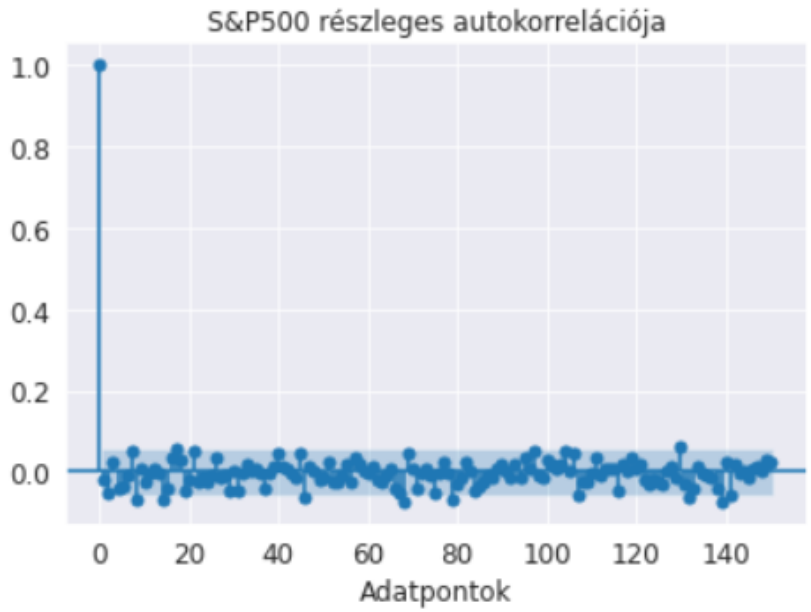
28.ábra: S&P 500 első differenciáltjának az autokorrelációja

Az ábrán megtekinthető, hogy az első adatpont után lecseng az autokorreláció és a megadott szignifikancia sávban mozog az összes érték, illetve korábban a teszt segítségével is látható volt, hogy stacioner az idősor. Az első érték esetén minden esetben 1 lesz látható, hiszen az a nulladik adatpont, tehát önmaga.

Így összegezve a d értékének az 1-et fogom választani, hiszen ezzel elértem, hogy az ARIMA modell előfeltétele, a stacionaritás biztosítva legyen.

A bemutatott ACF diagram használható az MA q paraméterének a megállapítására is, látható, hogy a késések mind a szignifikancia vonalon belül vannak, így 1-et választottam a paraméter értékének. Amennyiben lenne megadott tartományon kívüli érték, akkor érdemes lenne magasabb értéket hozzárendelni.

Az AR p paraméter értékét a PACF diagram segítségével teszteltem és választottam ki, amely a 29. ábrán látható. A részleges autokorreláció megmutatja az adott idősor és annak a késleltetettjei közötti korrelációt. [42]



29.ábra: S&P 500 részleges autokorrelációja (1.differenciált értéken)

A részleges autokorrelációs diagramon is látható, hogy már az első késleltetésnél alacsony lesz az érték, így a túlillesztés miatt 1-nél magasabb paramétert nem tartottam szükségesnek kiválasztani.

Így összességében az $ARIMA(1,1,1)$ modellt alkalmaztam, amelyet a 30. ábra mutat. Mindezt a Python által biztosított auto_arima segítségével is leteszteltem és a legjobb modellnek a program is ezt tartotta.

30. ábra: ARIMA (1,1,1) modell (S&P500)

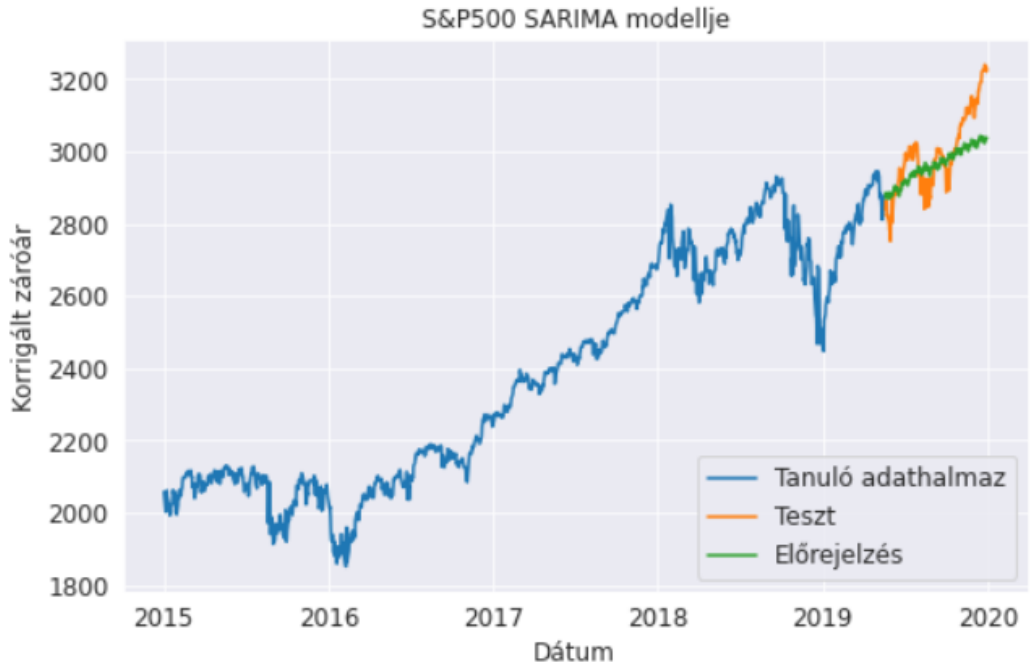
Azonban látható, hogy a modell pontosságán lehet javítani, illetve a trend elemzésnél is egyértelmű volt, hogy az idősor szezonalitást is tartalmaz, így az ARIMA modell szezonális változatát fogom használni a következőkben.
A modell kiválasztása esetén a tanuló adatsorhoz közel négy és fél évnyi értéket rendeltem hozzá, amely az adathalmaz 80%-át jelenti, az ezt követő értékek pedig a teszt adathalmazba kerültek. A függvény esetén beállítottam, hogy vizsgálja a modell a szezonalitást is. Ezen felül a korábban említett differenciálási értéket eggyel tettem egyenlővé, amely a stacionaritást biztosította. Így a modell, ahogy a 31. ábrán is látható az $ARIMA(3,1,0)(5,1,0)$, amely azt jelenti, hogy az AR értékének 3-at választott, a szezonális komponens esetén pedig 5-öt rendelt hozzá.
Az alábbi előrejelző modell a legjobb illeszkedést az Akaike információs kritérium alapján választotta ki, ahogy a 31.ábrán is megfigyelhető. Ez a mutatószám a modell minőségét ellenőrzi és hasonlítja össze a többi modellel való illeszkedést vizsgálva, ez alapján pedig a legalacsonyabbat választja ki. Így azt a modellt rendeli hozzá az előrejelzéshez, ahol a legkevesebb információ elvesztése valósult meg. [43]

Az előrejelző modell kiválasztása

```
ARIMA(4,1,1)(5,1,0)[12]             : AIC=10157.044, Time=87.20 sec
ARIMA(3,1,0)(5,1,0)[12] intercept   : AIC=10155.175, Time=139.63 sec

Best model:  ARIMA(3,1,0)(5,1,0)[12]
```

31. ábra SARIMA modell kiválasztása



32. ábra: SARIMA által becsült S&P500 árfolyam alakulása

A 32. ábrán pedig megfigyelhető, hogy a teszt adathalmazhoz képest a kiválasztott modell, milyen becslést adott az S&P500 árfolyam értékeinek az alakulására.
Mindezt pedig a jövőre is előrevetítettem, amely a 33. ábrán kerül bemutatásra, illetve a diagramon látható egy hibahatár is, amelyen belül vizsgálódik, tehát a szürkével jelölt területig bezárólag alakulhat a jövőbeli árfolyam a vizsgált modell szerint.

33. ábra: Jövőre előrevetített SARIMA modell (S&P500)

6.5. MODELL PONTOSSÁGA

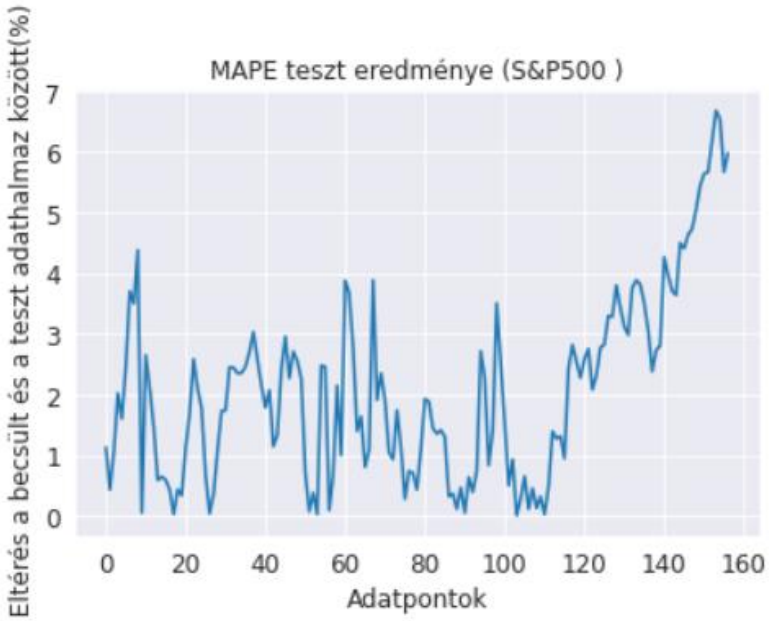
Az előző fejezetben bemutatott SARIMA modell pontosságát szemléltetem az alábbi fejezetben, ahol a már betöltött teszt adatokat vetem össze a becsült értékekkel.

Ennek a mérésére egyik alkalmas módszer a MAPE mutató használata. [44] Ezen formulát az alábbiak szerint definiálták és a feltüntetett 12. egyenlet mutat be:

(12) $$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

Az alábbi összefüggés azt fejezi ki, hogy az eredeti értékhez képest a becsült érték átlagosan hány százalékban tér el. Ezen képlet pedig a számítás során abszolút értéket használ annak elkerülésére, hogy a pozitív, illetve negatív értékek kioltsák egymást.

A modell pontosságának MAPE tesztjét elvégezve a 34. ábrán látható százalékos eltérések tapasztalhatóak a tesztelt és a becsült adathalmaz között. A diagramon megfigyelhető, hogy az utolsó adatpontok esetén tudta a modell a legkevésbé megállapítani az S&P500 index árfolyamának az alakulását. A nem várt piaci környezet, felpattanás, gyors korrekció olyan külső hatások, amelyet a modell természetesen nem tudott előre jósolni, ezért is tapasztalható lényeges eltérés a predikciós és valós értékek között a modell vizsgált időszakának az utolsó szakaszában.
Átlagosan azonban a mutató értéke 2,08% lett, amely azt jelenti, hogy a teszt és az előrejelzett érték átlagosan 2,08%-ban tér csupán el, amely egy közel 98%-os modellnek tekinthető, így elfogadhatónak ítéltem a modell becslését.



34. ábra: MAPE mutató a becsült és teszt adathalmaz között

Ezen mérőszámon kívül még az RMSE és MSE mutató szeretném kiemelni, amelyet Barnston [45] is kifejt a publikációjában. A képletet a 13. számú egyenlet írja le:

(13) $$r_{f\,o} = \sqrt{\sum_{i=1}^{N}(z_{f_i}\, z_{o_i})/N}$$

Az alábbi összefüggés a maradékok eloszlását mutatja meg, vagyis azt méri, hogy a hibatagok milyen messze vannak a regressziós egyenestől. Tehát máshogy megfogalmazva azt jelenti, hogy mennyire illeszkednek az adott regressziós egyenesre. Abban az esetben pedig, ha nem veszem az adott érték gyökét, akkor az MSE mutatóját kapom meg.

A korábban bemutatott modell esetén pedig a 35. ábrán látható RMSE és MSE értékek állapíthatóak meg:

```
S&P500 pontosságának a mérőszáma
Az S&P500 MSE értéke:6319.6487
Az S&P500 RMSE értéke:79.4962
```

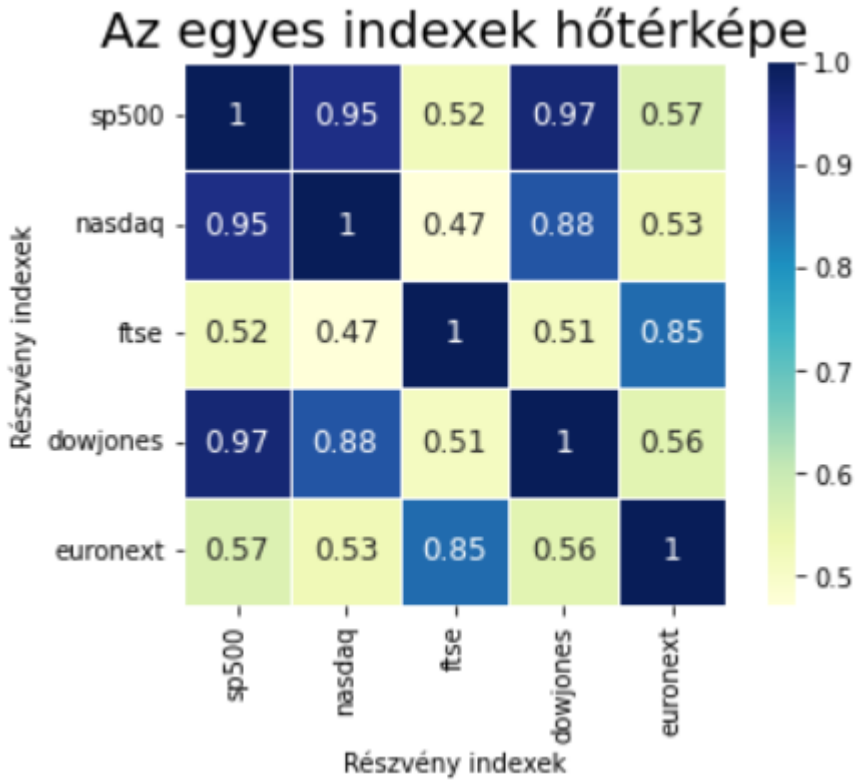
35. ábra: S&P500 RMSE és MSE mutatója

Így elmondható, hogy a modell adatpontjainak a regressziós egyenestől való átlagos eltérése csupán 79 pont, ami a vizsgált idősor értékeinek nagyságát tekintve nem nevezhető magasnak.

6.6. EGYMÁSRA HATÁSOK ELEMZÉSE

A feladat megoldása során célom volt az egyes indexek csoportokba rendezése, amelyeknél a korrelációs vizsgálódást szeretném kiemelni. Mindez tükrözi, hogy az egyes vizsgált részvény indexek milyen mértékben korrelálnak egymással.
Ezt a vizsgálatot azért is tartom fontosnak, mert egy megfelelően diverzifikált portfólió kiépítéséhez fontos tanulmányozni, hogy mely idősorok, hogy hatnak egymásra, hiszen az egyik idősornál esetlegesen elszenvedett veszteséghez jóval kisebb lesz a kitettségünk. Azonban a korreláció megállapításánál mélyebben a szakdolgozat keretein belül nem vizsgálódtam, mert egy diverzifikált portfólió kiépítésének a stratégiája, szempontjai túlmutatnak az általam a szakdolgozat elején felvázolt célrendszeren.

Erre a hőtérkép megoldást választottam, amely bemutatja színek és értékek segítségével is, hogy az egyes indexek esetén milyen mértékű összefüggés látható az idősorok között. Természetesen ez csupán egy felső és egy alsó háromszögben is ábrázolható lett volna, hiszen a két érték megegyezik. Azonban ezt a fajta négyzetes ábrázolást szemléletesebbnek találtam.



36. ábra: Részvényindexek korrelációs hőtérképe

A 36. ábrán feltüntetett hőtérkép esetén a világosabb színek jelölik az alacsonyabb korrelációs összefüggést, míg a magasabbak értékek kék színnel láthatóak.

Így tulajdonképpen 3 csoportra osztható az 5 index:

- 1.csoport: a szorosan összefüggő részvény indexek: S&P500-val összefüggő Nasdaq és DowJones,
- 2. csoport: DowJones és Nasdaq 88%-ban korrelál egymással, illetve az FTSE és az EuroNext mutat hasonló mértékű összefüggést.
- A 3. csoportot pedig a maradék kevésbé korreláló indexek alkotják.

Jól látható, hogy a legnagyobb mértékű összefüggés az S&P500 és a DowJones, illetve az S&P500 és a Nasdaq között látható. Ezek közül ki is emelném az S&P500 és a Nasdaq idősorát, ahol a két idősor szinte ugyanolyan növekedési trendet mutat, amelyben viszont eltérnek, hogy más kezdő értéke van az adathalmazoknak. Ezenfelül a kiugró értékek esetén is tapasztalható eltérés. A két idősor a 37. és a 38. ábrán tekinthető meg.

37. ábra: Nasdaq idősora kiugró értékekkel

38.ábra: S&P500 idősora kiugró értékkel

7. KÖVETKEZTETÉSEK LEVONÁSA

Az alábbi fejezet tartalmazza a szakdolgozat összefoglalóját, illetve itt mutatom be az elemzés közben tett megállapítások konklúzióját.

A feladat megoldása során egy több részvényt magában foglaló indexet, az S&P500 -ot vizsgáltam, amelynek során az egyik legfontosabb szempontnak az előelemzést tekintettem. Megemlítettem egy kutatást is, amely alátámasztja ezen kijelentésemet. Ez azt mutatta ki, hogy az elemzések 80%-át elővizsgálódás folyamata teszi ki.
A szakdolgozat felépítése is ezen módszerek, statisztikai szempontok kutatására és bemutatására irányult.

Első lépésként az idősor statisztikai mutatóit – szórás, átlag, maximum, minimum értékek– vizsgáltam meg, továbbá megállapítottam, hogy az adathalmaz nem tartalmaz hiányzó értéket, így az imputációra nem volt szükség a folyamat során.
Kiváltképp fontos az idősor elemzés alatt a trend vizsgálata is az adatsorban, hiszen az nagyban befolyásolhatja a becslésünk pontosságát, illetve nem stacionárius adathalmazra nem is illeszthető a kiépített ARIMA modell.

A trend elemzést első lépésben a kiugró értékek detektálására használtam, ahol a reziduumok eltérésével megállapítottam, hogy az adott adatpont kiugró értéknek tekinthető-e. Ez alkalmas módszernek bizonyult, hiszen az 5 éves adatsor doboz diagramon való ábrázolása nem hozta meg a várt végeredményt a nagy eltérések miatt.
A kiugró értékek felismerését követően az idősor stacionaritásának a biztosítására volt szükség, hiszen az Augmented Dickey Fuller teszt egyértelműen 5%-nál magasabb p értéket mutatott, így megállapításra került, hogy az idősor nem stacioner.

A stacionaritás biztosítása a differenciálás módszerével valósult meg, amelynek köszönhetően az idősor már előrejelezhetővé vált az ARIMA modell segítségével.
Az autokorreláció és a részleges autokorrelációs módszer segítségével pedig megállapítható vált, hogy milyen paramétereket szükséges a modell kiépítése során használni. Mindezt a Python által biztosított auto_arima függvény segítségével is teszteltem, amely ugyanazt az eredményt adta, így a későbbiekben már ezt vettem alapul.

Ahogy a szakdolgozatban is említettem az idősor szezonalitást is tartalmazott, így a kiválasztott ARIMA előrejelző modell szezonális változatára volt szükségem a további vizsgálódás során, amely kiválasztotta a legalacsonyabb Akaike információs kritérium alapján az $ARIMA(3,1,0)(5,1,0)$ modellt.
Ennek a modellnek a tesztelését is elvégeztem a feladat végrehajtása során, amelynek a pontossági eredmény közel 98%-os lett a MAPE mutatója alapján.

8. TOVÁBBFEJLESZTÉSI LEHETŐSÉGEK

A szakdolgozat során egy a szakértők között is leggyakrabban használt modellt mutattam be, azonban számos más modell is ismert, amelyeket a későbbiekben érdemes lenne megvizsgálni, hogy általuk milyen pontosságú becslés hozható létre, ilyen például [16][46]:

- ARCH,
- GARCH,
- Vector Autoregresszív modell (VAR),
- Deep Learning,
- Prophet,
- XGBoost

Egy másik fontos továbbfejlesztési lehetőségnek tekintem a tőzsdei hangulat elemzését, amely megmutatkozik a részvény árfolyamában, azonban számokban nehezen kifejezhető. Fontos tehát azt a tényezőt is figyelembe venni, hogy akár a közösségi platformokon tett kijelentések (például egyes neves cégtulajdonosok, közszereplők által) is komolyan befolyásolhatják az adott részvény pillanatnyi árfolyamát, rövid távú kilengéseket okozva. Szintén jó példa egy esetleges nagyobb részvényfelvásárlás árfelhajtó hatása, vagy akár csak egy leendő felvásárlás híre, amely természetesen jelentős kihatással lesz a befektetők várakozásaira, akik eszerint invesztálnak az adott részvénybe, avagy adják el őket.
Lehetséges tehát, hogy egy tőzsdén jelenlévő vállalat, habár a számok alapján nyereséges, azonban a befektetők (beszélve itt a nagy befektetőkről főként) aszerint kereskednek, hogy a vállalat jövőjét mennyire pozitívnak ítélik meg. Egy adott bejelentés ezt a jövőképet hirtelen gyorsasággal befolyásolhatja. [47]
Mindehhez hasznos lehet a piaci hangulat felmérésén túl, akár az egyes tweetek elemzése is, hogy az adott híreknek milyen a befektetői visszajelzése, adott céghez mennyire társítható pozitív érzés.
A Journal of Computational Science folyóirat 2011-ben ezt vizsgálta, hogy a Twitter-hangulatból mennyire előre jelezhető a tőzsde piaca, tehát a közhangulat korrelál-e a gazdasági mutatószámokkal. A cikk a DJIA (30 legfontosabb amerikai részvény) záróértékeinek változás-előrejelzésére fókuszált, ahol két hangulatkövető eszközt használtak: OpinionFinder-t és a GPOMS-t (Google-Profile of Mood States). Ez alapján megállapították, hogy 87,6%-os pontossággal lehetséges a DJIA napi záróértékeinek a felfelé és lefelé történő változásait előre jelezni. [48] Így a tőzsdei adatok vizsgálatát, predikcióját tekintve ezt tartom egy további nagyon fontos elemzési technikának.

9. KIVONAT

A tőzsdei kereskedés a Robin Hood és egyéb befektetési platformok berobbanásával, láthatóan utat tört a kis és egyéni befektetők körébe, a korábbiaknál jóval nagyobb teret nyert az átlagemberek között is, hiszen a modern gazdaságban az egyik legnagyobb információs központnak tekinthetjük.
Az árfolyamok percenként változnak, amelyeknek a megértéséhez komoly számítógépes erőforrásra, algoritmusok ismeretére és statisztikai gondolkozásra van szükség. Ez az adathalmaz pedig minden nap egyre nő, amely mögött a mintákat szeretné azonosítani az éppen aktuális befektető. Ehhez áll rendelkezésre a Big Data tudománya.

A feladat megoldása során az imént említett mintákat, összefüggéseket, egymásra hatásokat kutatom. Célom, hogy minél pontosabb, szakszerűbb előrejelző modellt lehessen kiépíteni a kiválasztott indexeket érintően a múlt adataiból tanulva.

A vizsgálódás során a legfontosabb az adatok előábrázolása és egyben előelemzése, amelynek hozzáadott értéke, hogy az idősor statisztikai mutatói mentén már el lehet indulni az előrejelző modell kiépítése felé.

A szakdolgozat készítése mentén látható volt, hogy a trendet, szezonalitást, hibatagokat megfelelően be lehetett azonosítani, azonban a kiugró értékek detektálására és kezelésére már számos módszert, technikát figyelembe kellett vennem. Ennek során a dobozdiagram nem bizonyult a legjobb megoldásnak a nagy intervallumú értékek miatt, azonban a hibatagokon keresztüli vizsgálódás segítségével az anomáliák azonosíthatóvá váltak.

A megfelelően átalakított adathalmaz tanulmányozásával pedig előrejelezhetővé vált az index jövőbeli alakulása a teszt adathalmazon és a későbbi időszakra is.

Következésképp az S&P500 indexét vizsgálva arra a megállapításra jutottam, hogy tökéletes modellt kiépíteni nem lehet, azonban a medve és bika piac felismerhető és beazonosítható. A kiugró értékek detektálásával, a trendek azonosításával, a megfelelő statisztikai vizsgálatok végrehajtásával egy 98%-os MAPE mutatójú modell létrehozható volt. Természetesen itt a becsült és a teszt adathalmaz közötti szórás nagyobb mértékűnek mondható.

10. ABSTRACT

Trading of shares and stock market securities, due to the revolution of Robin Hood and similar trading platforms, have evidently spread widely among small, individual investors. This is also thanks to the fact that never before we had so vast, wide, quick and truly massive amount of information regarding the modern economical and financial situation at the fingertip of the average people.
As the stock prices change by the minute, to understand the process, it requires significant computing resource, the knowledge of algorithms and statistical thinking. This huge volume of data, naturally grows by every day, to which the investor wants to identify patterns that affects the pricing. This is where the science of Big Data comes to the picture. During my thesis, I am investigating these patterns, correlations and causalities. My goal is to create an accurate, professional forecasting model based on historical data of selected indexes.

During my investigation, the most important thing was the visualisation and the pre analysis of data. This had the added benefit to provide the foundation for the building of the forecasting model.
During the thesis, it became visible that it is possible to identify trends, seasonality error terms, however the detecting and managing of outliers requires several other methods and techniques. Due to the big interval values, the box diagram didn't prove to be the best solution, however thanks to investigating the outlier values the anomalies were later identified.

Further investigating the transformed dataset made possible to create a forecast of the index's price on the test dataset and also on future timeframe.
As a result, analysing the S&P 500 index, I have concluded that building a perfect model is not feasible, however it is possible to properly identify a bear or bull market phase. By detecting outliers, identifying trends and executing the proper statistical analysis, the creation of a 98% accurate MAPE model was possible. Naturally, the deviation between the estimated and test datasets were more significant.

11. IRODALOMJEGYZÉK

[1] John C. Chambers, Satinder K. Mullick, and Donald D. Smith: How to Choose the Right Forecasting Technique. *Harvard Business Review*, Vol. 49., 1971, pp. 45-71.

[2] Investopedia: How Big Data has changed finance, (https://www.investopedia.com/articles/active-trading/040915/how-big-data-has-changed-finance.asp), utoljára megtekintve: 2022.04.30.

[3] Az 5V, (https://medium.com/@get_excelsior/big-data-explained-the-5v-s-of-data-ae80cbe8ded1), utoljára megtekintve: 2022.05.07.

[4] The 4 Characteristics of Big Data, (https://www.bigdataframework.org/the-four-vs-of-big-data/), utoljára megtekintve: 2022.04.30.

[5] What are the 5 V's of Big Data?, (https://www.teradata.com/Glossary/What-are-the-5-V-s-of-Big-Data), utoljára megtekintve: 2022.05.07.

[6] S&P 500, (https://www.spglobal.com/spdji/en/indices/equity/sp-500/#overview), utoljára megtekintve: 2022.04.30.

[7] Nasdaq Fact Sheet, (https://www.nasdaq.com/), utoljára megtekintve: 2022.04.30.

[8] Dow Jones Industrial Average Fact Sheet, (https://www.spglobal.com/spdji/en/indices/equity/dow-jones-industrial-average/#overview), utoljára megtekintve: 2022.04.30.

[9] FTSE 100 Index Factsheet, (https://www.ftserussell.com/products/indices/ukx), utoljára megtekintve: 2022.12.04.

[10] Investopedia, (https://www.investopedia.com/terms/e/euronext.asp), utoljára megtekintve: 2022.04.30.

[11] Compare Databricks Lakehouse Platform and Hadoop HDFS, (https://www.g2.com/compare/databricks-lakehouse-platform-vs-hadoop-hdfs), utoljára megtekintve: 2022.04.30.

[12] What does Databricks do, (https://towardsdatascience.com/what-does-databricks-do-8a6c4ef9071b), utoljára megtekintve: 2022.04.30.

[13] What is a data lakehouse,
(https://databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html),
utoljára megtekintve: 2022.04.30.
[14] TIOBE indexek,
(https://www.tiobe.com/tiobe-index),
utoljára megtekintve: 2022.04.30.
[15] Choosing Python or R for Data Analysis?,
(https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis),
utoljára megtekintve: 2022.04.30.
[16] Ben Auffarth: Machine Learning for Time-Series with Python: Forecast, predict, and detect anomalies with state-of-the-art machine learning methods. *Packt*, 2021, ISBN 978-1-80181-962-6
[17] Gailt Shmueli Kenneth C. Lichtendahl Jr.: Practical Time Series Forecasting with R, *Axelrod Schnall Publishers*, 2018, ISBN: 978-0-9978479-1-8
[18] *Ferenci Tamás*: A sztochasztikus idősorelemzés alapjai,
(https://adoc.pub/a-sztochasztikus-idsorelemzes-alapjai.html),
utoljára megtekintve: 2022.04.30.
[19] Jiawei Han, Micheline Kamber: Data Mining: Concepts and Techniques, Second Edition, *Morgan Kaufmann Publishers*, 2006, ISBN: 978-1-55860-901-3
[20] Wikipedia: A globális felmelegedés,
(https://hu.wikipedia.org/wiki/Glob%C3%A1lis_felmeleged%C3%A9s#/media/F%C3%A1jl:Global_Temperature_Anomaly.svg),
utoljára megtekintve: 2022.04.30.
[21] Stationarity in time series analysis,
(https://towardsdatascience.com/stationarity-in-time-series-analysis-90c94f27322),
utoljára megtekintve: 2022.04.09
[22] Példák a stacionárius és nem stacionárius idősorra,
(https://www.researchgate.net/figure/Examples-for-stationary-and-non-stationary-time-series_fig3_348592737),
utoljára megtekintve: 2022.04.30.
[23] Ferenci Tamás: Idősorelemzés,
(https://math.bme.hu/~ftamas/szrmea/szrmea789.pdf),
utoljára megtekintve: 2022.04.30.
[24] Ferenci Tamás: A gyenge stacionaritás fogalma,
(https://www.youtube.com/watch?v=lkhqtSD66J8),
utoljára megtekintve: 2022.04.28.
[25] D.M. Hawkins: Identification of Outliers. *Chapman and Hall*, 1980, ISBN: 978-0412219009

[26] IBM: Outliers,
(https://www.ibm.com/docs/en/spss-modeler/18.1.1?topic=series-outliers),
utoljára megtekintve: 2022.04.30.
[27] Fox, A.J.: Outliers in Time Series. *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol.34, 1972, pp.350-363.
[28] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, Greta M. Ljung: Time Series Analysis: Forecasting and Control.*Wiley*, 2015, ISBN: 978-1-118-67502-1
[29] Alan Anderson, David Semmelroth: Statistics for Big Data For Dummies. *For Dummies*, 2015, ISBN: 978-1-118-94001-3
[30] Peter J. Brockwell, Richard A. Davis: Time Series: Theory and Methods. *Springer*, 2009, ISBN: 978-1441903198
[31] Wikipedia: Autoregressive integrated moving average,
(https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average),
utoljára megtekintve: 2022.04.30.
[32] Wikipedia: Autoregressive–moving-average model,
(https://en.wikipedia.org/wiki/Autoregressive%E2%80%93moving-average_model),
utoljára megtekintve: 2022.04.30.
[33] Rob J Hyndman and George Athanasopoulos: Forecasting: Principles and Practice, *OTexts*, 2018, ISBN-13: 978-0987507112
[34] Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says,
(https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=730c65506f63),
utoljára megtekintve: 2022.10.23
[35] Wikipedia: Hodrick–Prescott filter,
(https://en.wikipedia.org/wiki/Hodrick%E2%80%93Prescott_filter),
utoljára megtekintve: 2022.10.29.
[36] Time Series From Scratch — Decomposing Time Series Data,
(https://towardsdatascience.com/time-series-from-scratch-decomposing-time-series-data-7b7ad0c30fe7),
utoljára megtekintve: 2022.10.29.
[37] Understanding Boxplots,
(https://builtin.com/data-science/boxplot),
utoljára megtekintve: 2022.10.27.
[38] Detecting and Treating Outliers | Treating the odd one out!,
(https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/),
utoljára megtekintve: 2022.10.30.
[39] Identifying outliers,

(https://help.highbond.com/helpdocs/analytics/15/en-us/Content/analytics/analyzing_data/identifying_outliers.htm), utoljára megtekintve: 2022.10.31.

[40] Wikipedia, (https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test), utoljára megtekintve: 2022.10.28.

[41] Charles Therrien, Murali Tummala: Probability and Random Processes for Electrical and Computer Engineers, CRC Press, 2011, ISBN: 978-1439826980

[42] Wikipedia, (https://en.wikipedia.org/wiki/Partial_autocorrelation_function), utoljára megtekintve: 2022.11.12.

[43] Wikipedia, (https://en.wikipedia.org/wiki/Akaike_information_criterion), utoljára megtekintve: 2022.11.14.

[44] Oracle: Working with planning, (https://docs.oracle.com/en/cloud/saas/planning-budgeting-cloud/pfusu/EPM-INFORMATION-DEVELOPMENT-TEAM-E94218-6693400D.pdf) utoljára megtekintve: 2022.11.13.

[45] Anthony G. Barnston: Correspondence among the Correlation, RMSE and Heidke Verification Measures; Refinement of the Heidke Score. *Weather and Forecasting*,Volume 7, Issue 4, 1992, pp. 699–709.

[46] Advancing Analytics: 10 Incredibly Useful Time Series Forecasting Algorithms, (https://www.advancinganalytics.co.uk/blog/2021/06/22/10-incredibly-useful-time-series-forecasting-algorithms) utoljára megtekintve: 2022.11.20.

[47] Roger Lowenstein: Origins of the Crash: The Great Bubble and Its Undoing. *Penguin Books*, 2004, ISBN 978-1-59420-003-8

[48] Johan Bollen, Huina Mao, Xiao-Jun Zeng: Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 2011, pp. 1-8.

12. ÁBRAJEGYZÉK