Elemzés repülőjáratok értékeléséről

Az elemzéshez egy repülőjáratok értékeléséről szóló adatforrást választottam, mely elérhető az alábbi linken:

https://www.kaggle.com/datasets/sujalsuthar/airlines-reviews/

Az adatforrás 2023. évi 10 legjobban értékelt légitársaságáról tartalmaz a közelmúltból véleményeket 17 oszlopban. Az oszlopok a következők:

- Title: Az értékelés címe. (Például: this was the worst experience)
- Name: Az értékelő neve. (Pl: Susanne Haas)
- Review Date: Az értékelés dátuma. (Például: 2018-03-10)
- Airline: A légitársaság neve. (10 különböző érték lehet, például: Air France)
- Verified: Ellenőrizve lett-e az értékelés. (True vagy False)
- Reviews: Az értékelés. (Például: Traveled on Air France many times, but this ...)
- Type of Traveller: Az utazó típusa: (4 különböző érték lehet, például: Couple Leisure)
- Month Flown: Repülés éve és hónapja. (Például: 2018-02)
- Route: Honnan történt az utazás hova. (Például: Geneva to Paris)
- Class: Melyik osztályon történt az utazás. (4 különböző érték lehet, például: Economy Class)
- Seat Comfort: Az ülés mennyire volt kényelmes. (1-5-ös skálán, például: 4)
- Staff Service: A Személyzeti szolgáltatás értékelés. (1-5-ös skálán, például: 4)
- Food & Beverages: Ételek és italok értékelés. (1-5-ös skálán, például: 4)
- Inflight Entertainment: A fedélzeti élmény/program értékelése. (1-5-ös skálán, például: 4)
- Value For Money: A pénzért kapott érték. (1-5-ös skálán, például: 4)
- Overall Rating: Összesített értékelés. (1-10-ös skálán, például: 9)
- Recommended: Ajánlaná-e az utazást: (yes or no)

Elsőként az adatok beolvasásával kezdtem, ahol dátummá alakítottam a 'Month Flown' és 'Review Date' oszlopokat. Egyszerű parancsokkal kiderítettem, hogy 8100 sorunk és 17 oszlopunk van a táblázatban, amely nem tartalmazz null értéket. Illetve a vélemények 2013 március és 2024 március közötti repülőutakról szólnak. Látható, hogy az értékelések dátuma viszont jóval, későbbi dátummal kezdődik, hiszen ez 2016. március 22.-től 2024. március 6.-ig tartalmaz adatokat.

Elemzés célja

A fő cél a számszerű értékelések légitársaságonkénti összehasonlítása volt: Mely légitársaságok teljesítettek a legjobban az egyes területeken, illetve miben kéne fejlődniük? Mely tényező befolyásolja legjobban az összesített értékelést?

Dátum szerint hogyan alakul az értékelések száma és a pontszámok átlaga, illetve függ-e az adott hónaptól, hogy mennyi az utazás és az értékelés megírása között eltelt idő?

Melyek a legnépszerűbb útvonalak, utazási célpontok, legforgalmasabb városok?

Mely szavak fordulnak elő a leggyakrabban a szöveges értékelésekben?

Van-e szignifikáns különbség a különböző osztályokon belüli és utazótípusokon belüli értékeléseknél?

Ezek után kezdődött az adatok tisztítása

A duplikált sorok száma 0 volt, így azzal nem kellett foglalkozni. Elsőként A 'Route' oszloppal kezdtem el dolgozni. Ez olyan adatokat tartalmaz, mint például „Budapest to Rome Fiumicino via Praga". Két információt szeretnék kinyerni ebből. Az indulás és a célpont városát. Ezt egyszerű „split" parancs segítségével megoldottam, azonban egy esetben nem eredményezett megfelelő szétválasztást, amikor az „and" szót használták. Viszont az „and" szót több város is tartalmazza a nevében, ezért nem írhattam be az „split" parancs argumentumába, így ezt az egy sort külön kezeltem és manuálisan állítottam be.

Ha megvizsgáljuk különböző városokat, akkor láthatjuk, hogy sokan nem a város nevét, hanem a repülőtér 3 betűs IATA kódját adták meg. Ezért az alábbi linkről letöltött táblázat segítségével oldottam meg ezt a problémát.

https://github.com/lxndrblz/Airports/blob/main/airports.csv

A táblázat tartalmazza az összes reptér kódját és adatokat az adott városról. Ebből készítettem egy szótárt, melynek segítségével tovább alakítottam az adatokat. A 'Departure' és 'Destination' esetében is több mint 150-150 IATA kódot alakítottunk át várossá. Azonban még mindig tartalmaz az oszlop hibás adatokat. Például különbözőnek veszi a „London Heathrow" és „London", illetve „New York JFK" és „New York" mezőket. Korábban beolvastam az összes város nevét, ahol reptér található. Ezt felhasználva végignéztem az 'Destination' és 'Departure' adatait, hogy megtalálható-e bennük valamely város neve, mert akkor inkább azt használom helyette. Ezáltal a 2 oszlop esetében összesen körülbelül 100 helyen végeztem javítást.

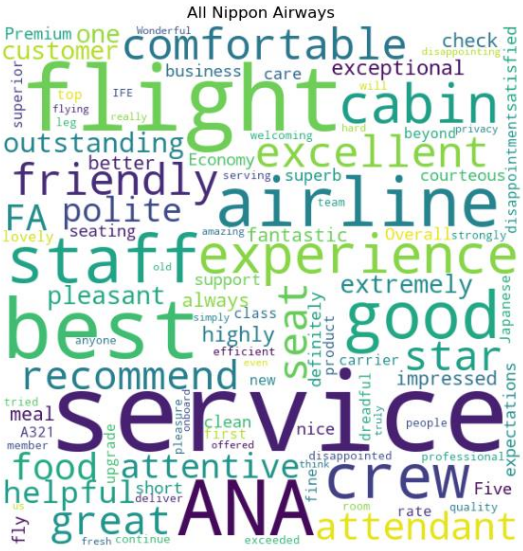
Ezután elkezdtem végignézni az összes oszlop egyedi értékeit. 3 esetben kellett javításokat végeznem. A 'Recommended' és 'Verified' értékeit boolean típusra állítottam, illetve az 'Inflight Entertainment' értékelésénél, egy 0 adatot találtam, pedig ez 1-5-ös skálán mozog. A 'Review'-t végig olvasva szinte biztosan állíthatom, hogy a legalacsonyabb pontszámot akarta adni, így ezt az értéket 1-re írtam át.

Az adatok elemzése

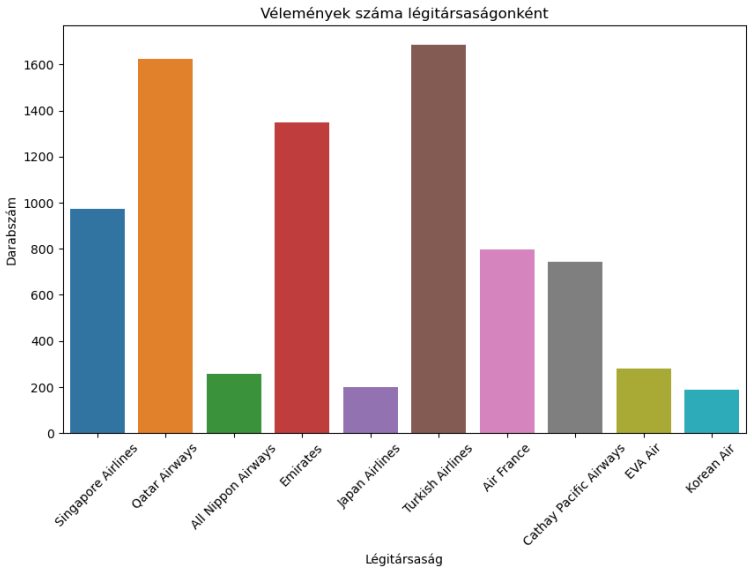
Szöveges értékelések vizsgálata

A szöveges értékelések vizsgálata nehéz feladat. Ezekből egy szófalhőt készítettem:

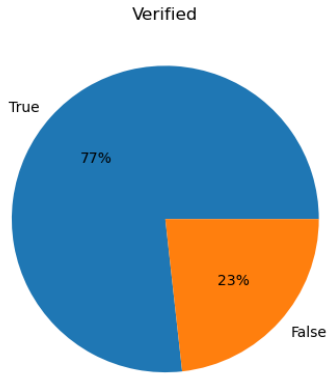
Természetesen pozitív és negatív jelentésű szavakat is találhatunk. Remek, hogy a „good" és „comfortable" szó ilyen nagyban szerepel és több kisebb hasonló jelentésűt is láthatunk. Feltehetőleg az embereknek egyik legfontosabb tényező, hogy időben megérkezhessenek a célpontjukra, hiszen a „time" és a „hour" is sok véleményben megtalálható volt. Ezek mellett a „food" és „meal" is jól kivehető, tehát erre is oda kell figyelni a légitársaságoknak, hogy milyen ételeket szolgálnak fel.



A legrosszabb összetett értékelése a Turkish Airlines-nak volt 3,67-el, a legjobb az All Nippon Airways-nak 7,95-el. Látható a két szófelhő közti különbség. A bal oldalon rengeteg olyan szót láthatunk, melyet egy légitársaság nem szívesen olvas a vélemények között. Ezek például a 'worst airline', 'avoid', 'rude', 'worst experience', 'poor', 'disappointed', 'horrible'. A jobb oldali szófelhőn alig találunk negatív szavakat. Ott például a 'best', 'excellent', 'comfortable', 'outstanding' és ehhez hasonló jelentésű szavak dominálnak.

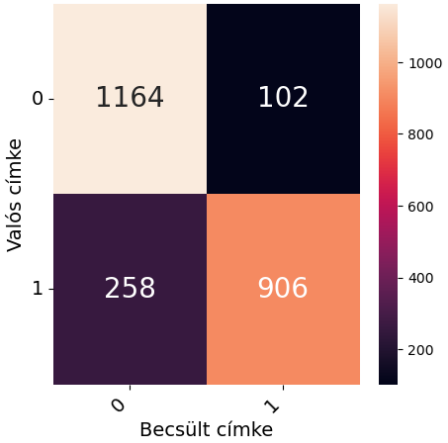


A 3 legtöbb értékelést kapó repülőgéptársaság a Turkish Airlines, Qatar Airways és az Emirates, tehát feltehetőleg ezeket vették legtöbbször igénybe.



Az összes értékelés 77%-át fogadták el.

Kategorizáltam az értékeléseket jó és rossz kategóriákba az átlagos értékelés alapján. Ennek a kategóriának a szöveges értékelés alapján történő megtippelésére betanítottam egy modellt. Ehhez először az úgynevezett „stopword”-öket töröltem az értékelésekből. (A „stopword”-ök olyan gyakori szavak, amelyek általában nem hordoznak sok jelentést az adott nyelv szempontjából, és gyakran előfordulnak a szövegekben. Például az angolban: is, at.) Majd az ugyanazon szavak más alakú előfordulását lecsökkentettem „PorterStemmer” segítségével (pl. „argue”, „arguing”, „argued”). Ezeket a feldolgozott szövegeket vektorizáltam TF-IDF módszerrel, és egy Naive Bayes osztályozót alkalmaztam a tanítóadatokra. A tesztadatokon a modell 85 százalékos pontossággal tippelt, ami jelen esetben megfelelő. A konfúziós mátrix látható a képen.



A modell a jövőben felhasználható az új értékelések esetében. Ha a modell a szöveges értékelést jónak ítéli, de a számszerű összesített értékelés mégsem ezt mutatja, akkor érdemes megvizsgálni, hogy mi okozhatta azt, hogy egy összeségében jó véleményt, valami mégis annyira lerontott, hogy a végén rossz ’Overall Rating’-et adott az illető. A modellt a jelenlegi adatainkon is megvizsgálhatjuk, azonban mivel annak nagyobbik része volt a tanító adathalmaz, ezért itt nem biztos, hogy releváns információkat kapunk.

	Reviews	Seat Comfort	Staff Service	Food & Beverages	Inflight Entertainment	Value For Money	Overall Rating
8041	Bali Denpasar to Seoul Incheon. Check-in was fast and easy by Korean's contracted ground handlers in Denpasar. The flight boarded and pushed back on-time. While the flight overall was fairly comfortable (good seat pitch, good seat-back inflight entertainment, pillow & blanket provided), Korean's cabin crew are very cold and without any personality. They are very robotic in everything they do. The cabin crew was very unfriendly and disingenuous while attending to the passengers - forced, half-smiles when serving the meal and drink service (the breakfast served - was soggy and virtually inedible). This is such a shame because they are dressed to impress and fly onboard such nice airplanes. If only they provided warm, caring hospitality, they would truly be an outstanding airline. Until then, I would not fly on them again by choice.	5	5	4	4	4	4

Ennél az értékelésnél olvasható, hogy az illető majdnem mindennel teljes mértékben meg volt elégedve, egyedül a barátságtalan személyzet, miatt adott csak 4 pontot a 10-ből az összesített értékelésnél. Ennek a légitársaságnak szükséges lenne kijavítani ezt a problémát.

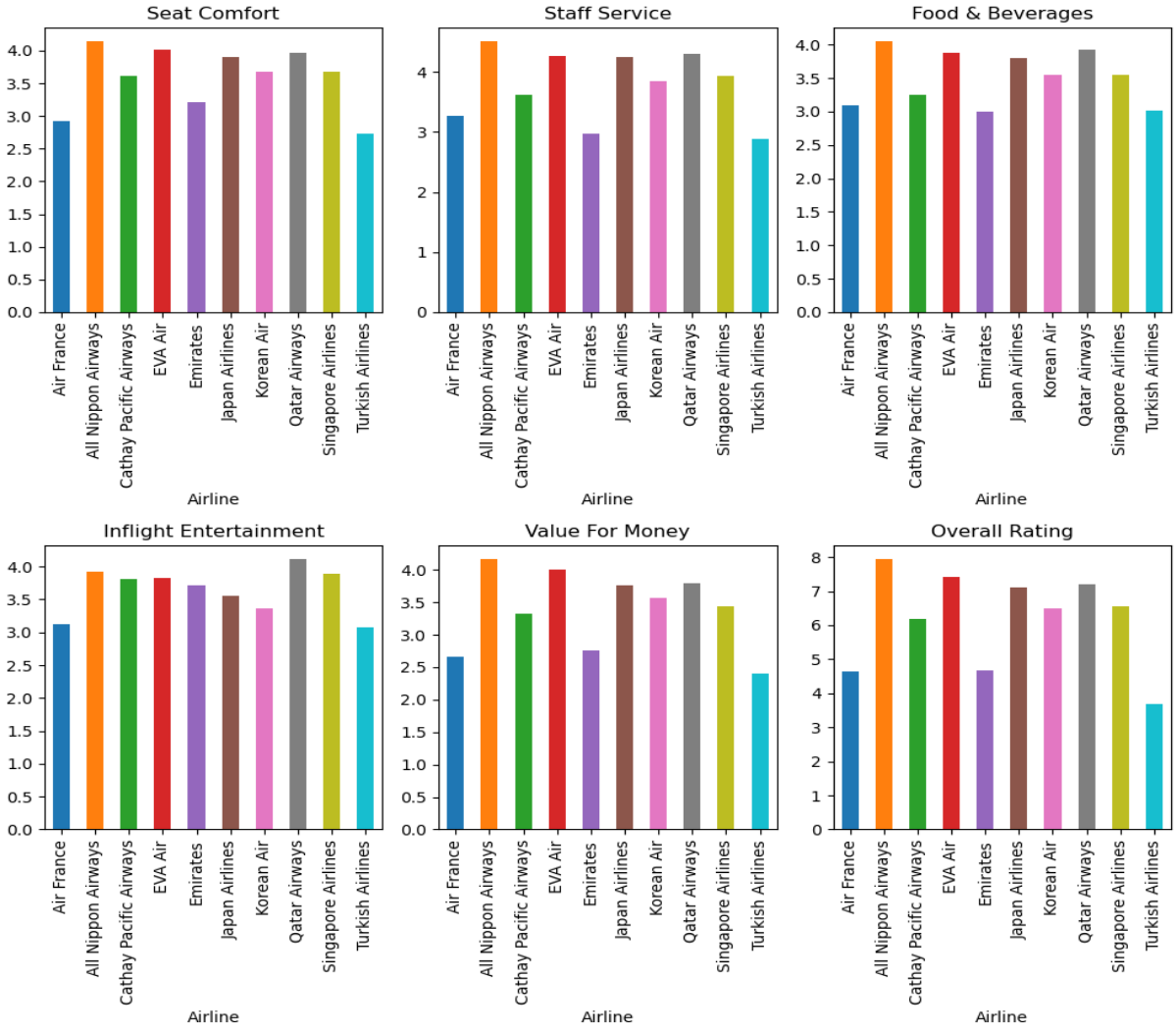
Utazás és a vélemény megírása között eltelt idő vizsgálata

Azt láthatjuk, hogy 2013-ban még nagy volt az utazás és a vélemény megírása közötti idő, azonban ez 2017-re megváltozott és azóta legtöbben még az adott vagy az azt követő hónapban megírják véleményüket. Ha ezt hónapokra vizsgáljuk, azt vesszük észre, hogy a nyári időszakban, az emberek sokkal hamarabb írják ezt meg, mint az októbertől májusig tartó időszakban a januárt leszámítva. A legtöbb utazás decemberben történik valószínűleg az ünnepek miatt, és ilyenkor a legnagyobb a várakozási idő az értékelésre. Ahhoz, hogy minél naprakészebb információink legyenek, azt javaslom, hogy téli időszakban több emailt írjunk a vevőknek, hogy értékeljék az utazásukat, és akár nyereményjátékokkal vagy bármilyen más módon ösztönözhetjük is őket erre.

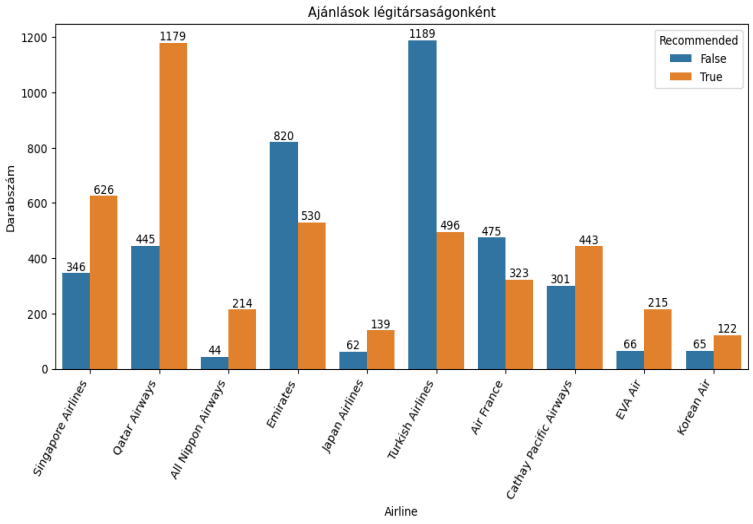
A legrosszabb vélemények decemberben születtek, ezt követik a nyári hónapok. Ezek a legforgalmasabb időszakok a repülésben. Ennek javítására szükséges lenne ilyenkor időszakos munkára alkalmazni több embert, melyek elősegíthetnék az emberek elégedettségének növekedését.

Értékelések légitársaságonként

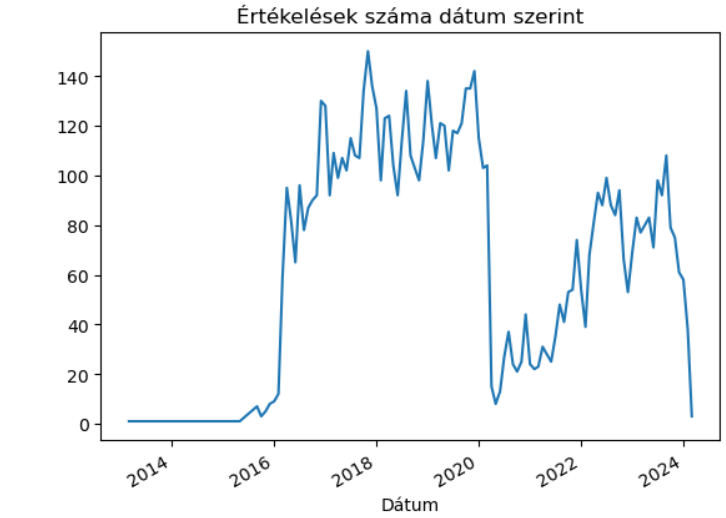
Airline	Seat Comfort	Staff Service	Food & Beverages	Inflight Entertainment	Value For Money	Overall Rating
Air France	2.922306	3.274436	3.098997	3.131579	2.669173	4.637845
All Nippon Airways	4.139535	4.507752	4.046512	3.930233	4.170543	7.949612
Cathay Pacific Airways	3.618280	3.615591	3.244624	3.815860	3.331989	6.169355
EVA Air	4.017794	4.263345	3.875445	3.832740	3.996441	7.419929
Emirates	3.208148	2.973333	3.000000	3.722963	2.752593	4.674074
Japan Airlines	3.895522	4.243781	3.791045	3.562189	3.766169	7.099502
Korean Air	3.684492	3.850267	3.545455	3.358289	3.561497	6.491979
Qatar Airways	3.969828	4.294335	3.927956	4.119458	3.798645	7.195813
Singapore Airlines	3.683128	3.940329	3.549383	3.886831	3.445473	6.542181
Turkish Airlines	2.735312	2.884866	3.018991	3.081306	2.397033	3.679525



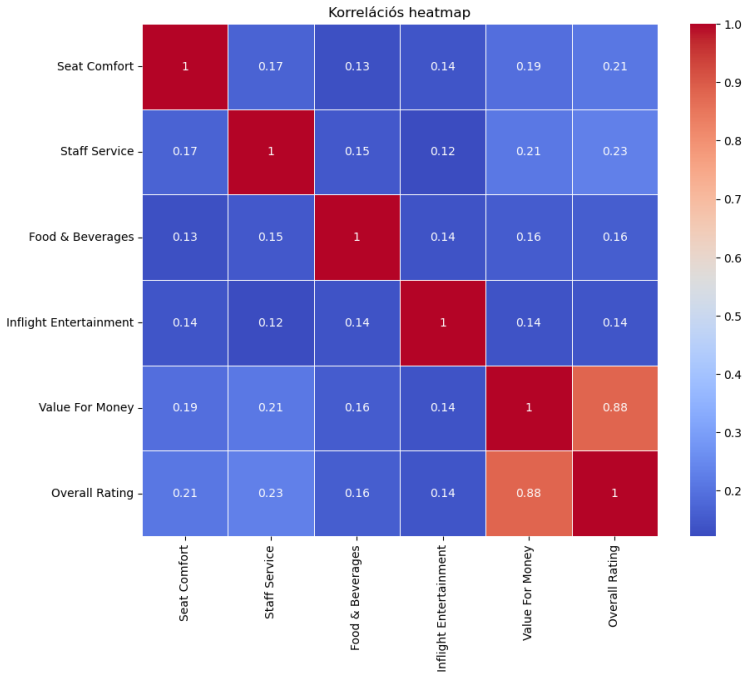
Ezeket végignézve légitársaságonként sok információt lehet megállapítani. Majdnem minden kategóriában a 4 legjobb légitársaság az All Nippon Airways, EVA Air, Japan Airlines és a Qatar Airways. Egyedül a fedélzeti élmény értékelésénél lépett a Qatar Airways és az All Nippon Airways mögé, a harmadik helyre egy másik légitársaság a Singapore Airlines. Így ezen kéne javítania a másik kettő cégnek. A legtöbb véleményt kapó Turkish Airlines a pénzért kapott érték és az ülés kényelmességének értékelésénél kapta a legrosszabb pontszámokat.



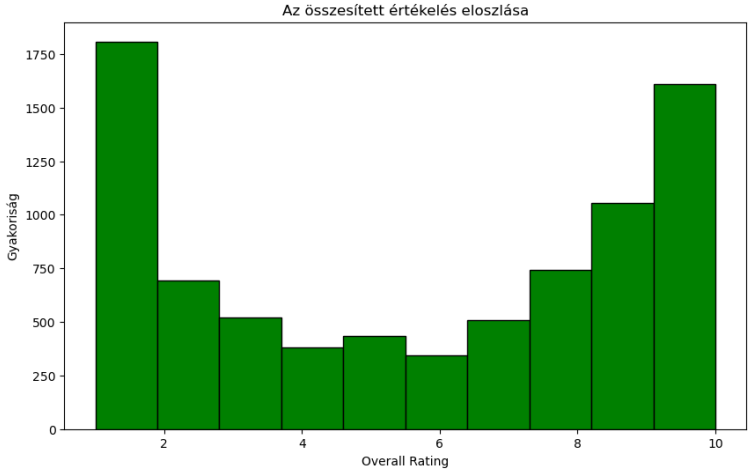
Az ajánlások vizsgálatánál nagy eltéréseket láthatunk. A Turkish Airlines-t ugyan rengetegen használják, azonban több, mint kétszer annyi a negatív vélemény róla, mint a pozitív. Valószínűleg ennek ellenére azért választják sokan, mert szinte a világ összes pontjára indítanak járatokat. Az Emirates és az Air France légitársaságok is szörnyen teljesítenek a vélemények alapján, így változtatásokra van szükség, hogy megtartsák ügyfeleiket.



Az értékelések száma, nagyjából megmutatja az utazások számát. A diagrammon látható, hogy nagy visszaesés történt 2020-21 környékén, ugyanis ekkor robbant ki a covid járvány. Annak ellenére, hogy 2022-től megindult az utazások számának növekedése, még nem sikerült elérni a korábban produkált számokat.

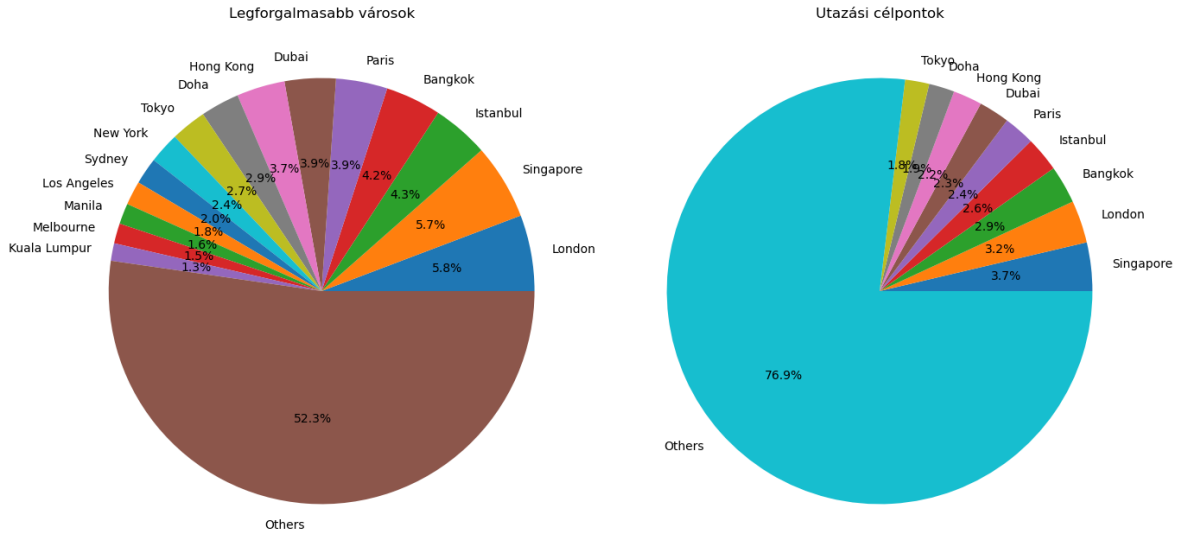


A korrelációs hőtérképből leolvashatjuk, hogy torony magasan a pénzért kapott érték határozza meg a végső értékelést. Legkevésbé a fedélzeti élmény és az ételek, italok számítanak.



Érdekes eredményt kapunk az összesített értékelés eloszlásának vizsgálatakor. Azt látjuk, hogy az emberek vagy nagyon jó, vagy nagyon pocsék pontszámot adnak. Ritka a közepes értékek használata.

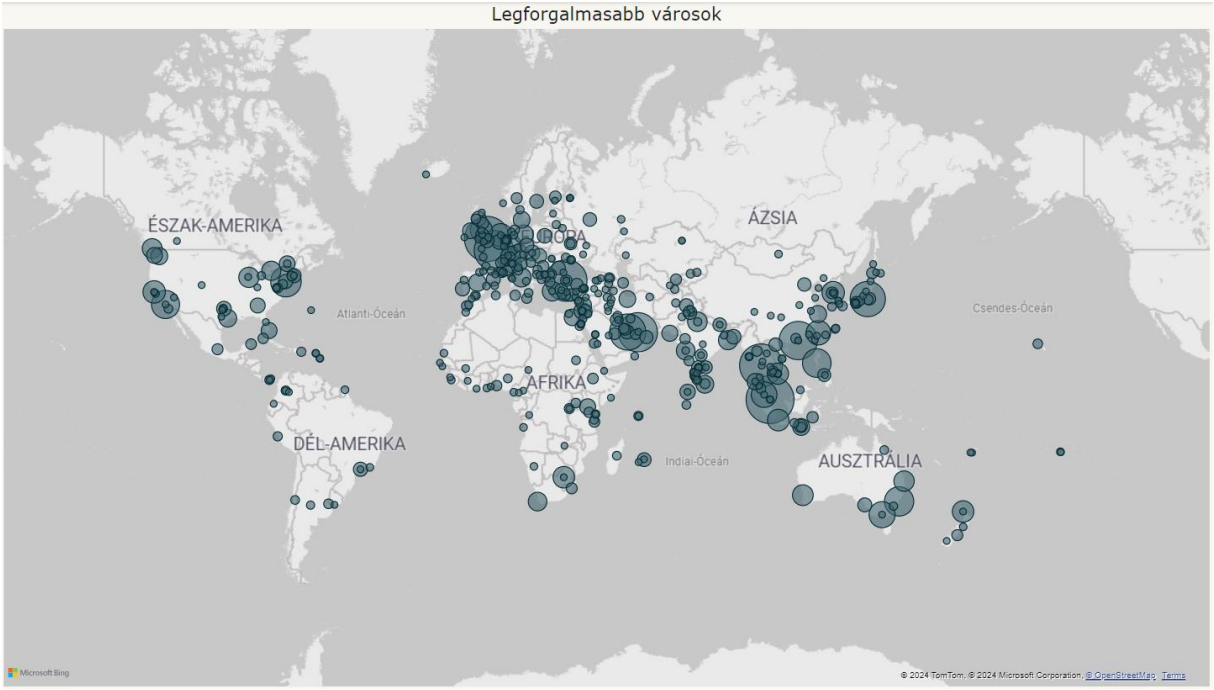
Városok forgalma



A 15 leggyakoribb reptérrel rendelkező város bonyolítja le a forgalom majdnem felét, ha csak az elemzésben szereplő 613 repteret nézzük. Csak az utazási célpontokat nézve sorrendben a 3 leggyakoribb Szingapúr, London és Bangkok.

```
London, Bangkok: 93 utazás
London, Singapore: 73 utazás
Hong Kong, London: 67 utazás
Dubai, London: 56 utazás
Hong Kong, Bangkok: 50 utazás
```

Ezek a leggyakoribb útvonalak, ha azt nem nézzük, hogy melyik a kiindulási és az érkezi város. London a 4 legyakoribb útvonalban szerepel, ezért a város nagy forgalmat bonyolít le.

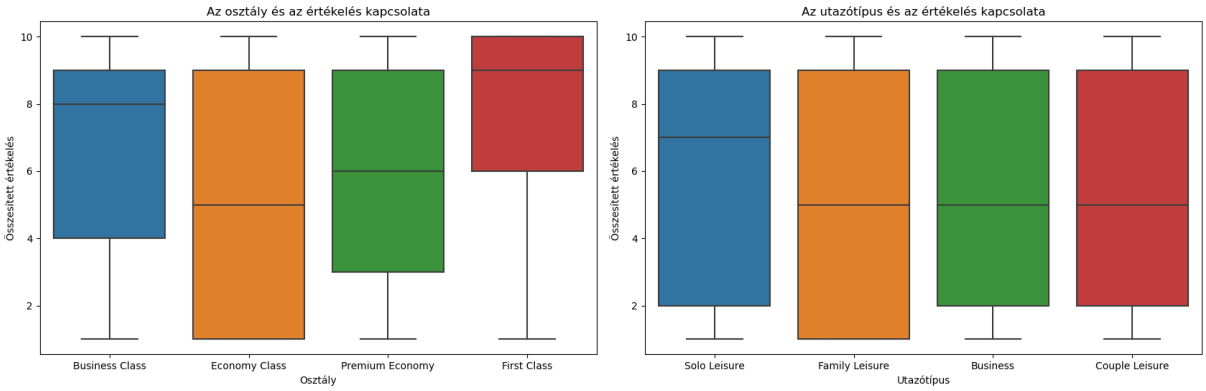


A legforgalmasabb városokat tartalmazó térkép Power BI segítségével készült. Európa számtalan repterén nagy forgalmat bonyolít le.

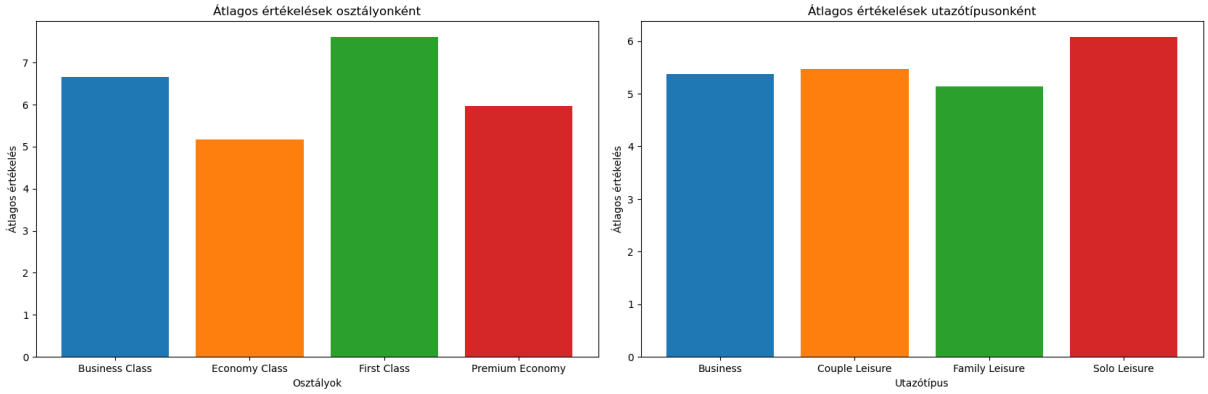
Utazási típusok és az osztályok vizsgálata

Type of Traveller Class	Business	Couple Leisure	Family Leisure	Solo Leisure
Business Class	628	502	216	758
Economy Class	702	1258	1273	2271
First Class	23	16	11	71
Premium Economy	60	123	51	137

Keresztábla vizsgálata esetén, azt látjuk, hogy egyedül az céges utazások esetében választják, sokan a Business Class-t, azonban ennél is, ahogy az összes kategóriában a legolcsóbb Economy Class dominál. Ez talán a nagy árkülönbségek miatt lehet. Talán ha nem kellene sokat ráfizetni, egy magasabb osztályért akkor többen választanák azt.



ANOVA táblával megvizsgáljuk, hogy van-e szignifikáns különbség az osztály, illetve az utazótípus változó különböző értékei között az összesített értékelés változó tekintetében. A p érték mindkét esetben, egy 0-hoz nagyon közeli szám lett, mely jóval kisebb, mint 0,05. Ez az eredmény azt mutatja, hogy a kategorikus változó különböző értékei között van szignifikáns különbség a számszerű változó tekintetében, tehát ez azt jelenti, hogy ez a különbség nagyon valószínűtlen, hogy véletlen lenne.



Semmi meglepőt nem tapasztaltam, amikor azt láttam, hogy az minél drágább osztályon utazunk, annál jobb az összesített értékelés.

Az egyedül utazó emberek átlagosabban jobb értékeléseket adnak. Családos értékelések a legrosszabbak. Esetleg a családok élményét lehetne javítani, akár kedvezőbb helyfoglalással számukra, hogy a repülőn ne keljen plusz összeget fizetniük, ahhoz, hogy egy helyre ülhessenek.

Type of Traveller	Airline
Business	Turkish Airlines
Couple Leisure	Qatar Airways
Family Leisure	Turkish Airlines
Solo Leisure	Qatar Airways

Az egyes utazótípusok által legtöbbször választott légitársaságok láthatóak a képen.

Készítette: Duli Bálint Adrián (MDI509)