BME AUT
Automatizálási és
Alkalmazott
Informatikai Tanszék

Opcionális NagyHF dokumentáció

Üzleti Intelligencia

2024 ősz

„A pénz nem boldogít” – Vagy mégis?!

Hajnal Árpád Erik - (IKW7OU)

hajnalarpaderik@gmail.com

MŰEGYETEM 1782

ETL folyamatok

Az ETL folyamataim 3 adatforrásból merítkeztek. Ezek pedig a következőek:

- Globális fizetés: https://www.kaggle.com/datasets/zedataweaver/global-salary-data
- Globális boldogsági-ráta: https://www.kaggle.com/datasets/PromptCloudHQ/world-happiness-report-2019
- Globális népesség: https://www.kaggle.com/datasets/tanishqdublish/world-data-population

A specifikációmmal szemben, végül nem használtam a globális indikátorok adathalmazát, mivel a megalósítás során láttam, hogy a fenti 3 adathalmaz is bőségesen elég a logika megvalósításához.

Az adathalmazokat .csv formátumban töltöttem le, majd Microsoft SQL Server Managment Studio segítségével lokális környezetben létrehoztam egy új adatbázist, illetve a megfelelő adattáblákat a megfelelő típusú oszlopokkal. Ezen létrehozási SQL utasításokat a beadott zip fájlba is beletettem.

Az adatbázis és a táblák létrehozása után a Visual Studioban hoztam létre egy új projektet. Ezen belül pedig 3 Data Flowt, egyet-egyet az adatforrásoknak. Minden Data Flow indítása előtt TRUNCATE paranccsal kitakaríttatom a megfelelő táblát.

Magában a Data Flowkban a .csv fájlból való beolvasást, különböző transzformációkat és a legvégén az adatbázisba való mentést hajtottam végre.

A legkönnyebb dolgom a „boldogság” adatforrással volt. Itt csak kiválasztottam a számomra lényeges oszlopokat, a biztonság kedvéért szűrtem a sorokat, hogy ne legyen duplikáció és rendeztem a sorokat boldogsági-ráta szerint. Utána pedig mentettem természetesen.

A „fizetés” adatforrás Data Flowjának alapjai teljesen megegyeznek a „boldogság”éval, annyival kiegészülve, hogy a fizetések statisztikai számait mind két tizedesjegyre kerekítettem egységesen.

A „népesség” adatforrás Data Flowja a legösszetettebb. Itt a már fentikben felsorolt műveleteken kívül, két oszlop transzformációját kellett elvégeznem. Ezek string típusúak voltak, százalékot jelöltek, minden érték végén ott is szerepelt a % jel. Ezeket a jeleket levágtam, majd kasztoltam a típusukat és végül ahogy a „fizetés” statisztikáinál is eljártam, itt is két tizedesjegyre kerekítettem.

A feladat elvégzése során ütemezést nem sikerült megvalósítanom, ami a Microsoft SQL Server Managment Studio esetében a Server Agent eszközzel kellett volna megtörténjen. Az eszköz számomra elérhetetlen volt, több órányi próbálkozás során se sikerült előcsalogatnom. Pentaho renszerben tudtam volna megvalósítani, de egyrészt a specifikációmba Microsoft eszközöket határoztam meg, másrészt a határidő közeledtével nem szerettem volna ha a többi funkció és elvégzett munka esetleges minőségromlását jelentette volna az átalakítás. Bízom benne, hogy az ezen kívüli elvégzett munkám kompenzálja a az ütemezés hiányát.

Reportok

A reportok kapcsán próbáltam minél széleskörűbb aspektusait bemutatni a domainnek. Összesen 9 reportot készítettem, ezek közül mindegyik dinamikus, vagy szűrhető, vagy rendezhető. A legtöbb report esetében, tehát ahol csak tehettem, megvalósítottam több lefúrást is, a komplexebb kép alkotásához. Táblákat, bar- és pie-chartokat. területdiagramot és fatáblát használtam mindezekhez.

Az alábbiakban röviden bemutatom az egyes reportokat és azok lényegét.

1. Országok medián keresete és boldogsági indexük

Egy táblázatot használtam az országok medián keresetének és boldogság indexének összehasonlítására. A táblázat minden oszlopa rendezhető. A táblázatban két lefúrás is van, az egyik az ország alapján a 2. reportra fúr, a kontinens szerint pedig a 9. reportra fúrhatunk le.

2. Adott ország népességének növekedése

Bar chart segítségével egy adott ország népesség növekedését ábrázoltam. A report lefúrással használható, úgy a lefúrt ország 1970-2020 időszak népesség növekedését láthatjuk 10 éves periódusokkal.

3. Országok területeinek eloszlása

A report egy pie-chartot tár elénk, ami az országok területeinek eloszlását mutatja be szerte a világban. Az országok szerint lefúrhatunk a 2. reportra.

4. Országok medián keresete és szabadság-függésük

Táblázatos nézetbe jelenítem meg az országok medián kereseteit és szabadság függésüket. A szabadság mutató, azt mutatja, hogy a világ országai között az adott ország boldogsága mennyire függ a szabadságuktól. Minden oszlop rendezhető, az országok lefúrhatóak a 2. reportra.

5. Top 10 legmagasabb fizetésű ország és azok különbsége a legalacsonyabb fizetéssel

Területdiagram segítségével ábrázoltam a világ 10 legmagasabb fizetéssel rendelkező országát és ezeknek a legkisebb fizetését. A reportból jól kiolvasható a legmagasabb és legalacsonyabb fizetések közötti differenciál.

6. Kontinensenkénti össz-medián kereset

A reportban lévő pie-chart a kontinensek szerinti összesített medián fizetéseket mutatja meg. A kontinensek szerint lefúrhatunk a 9. reportra.

7. Kontinensek és azokat befolyásoló szabadság index mértéke

Ez a pie-chart a kontinensek szabadság-függőségének eloszlását mutatja. A legnagyobb „szelettel” rendelkező kontinens boldogságát befolyásolja legjobban a szabadság. Kontinens szerint lefúrhatunk a 9. reportra.

8. Kontinensek boldogság inverze

A fatérkép segítségével a boldogság inverzét szeretném bemutatni kontinensenként, azaz azt, hogy hogy oszlik meg a boldogság kontinensenként. A legnagyobb területtel rendelkező kontinens tekinthető a legboldogtalanabbnak, míg a legkisebb területtel rendelkező a legboldogabbnak. Kontinens szerint lefúrhatunk a 9. reportra.

9. Adott kontinensen belüli országok medián béreinek eloszlása

Ez a report a lefúrásban játszik szerepet. Segítségével egy adott kontinensen belül mutatja meg a medián bérek eloszlását országok szerint.

Adatelemzés

Az adatelemzés feladatrész elvégzéséhez a népesség adatforrást használtam fel, ezt egy Facebook (Meta) Prophet modelljének alapjául átadva predikciót hajtottam végre a jövőbeni népesség alakulásának kiderítéséből. A Google Colabban elvégzett műveleteket végig pontosan dokumentáltam, így az egyszerűség kedvéért ide nem illeszteném be, a noteboookban teljesen részletezve olvasható a folyamat.

Google Colab notebook link:

https://colab.research.google.com/drive/1BI1LBKlA8uy75q2gLkli3RpOUqrIXkou?usp=sharing

Github repó link:

https://github.com/hajnalarpaderik/BI_HF

-