Mérési adatok feldolgozása és vizuális elemzése

Rendszertervezés laboratórium 2

Mérési útmutató

Készítette: Gönczy László, Nádudvari György, Burján Dezső

Verzió: 2.0

2018.

Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs Rendszerek Tanszék

1 A mérésről

A mérés célja, hogy a hallgatók megismerkedjenek a vizuális adatelemzéssel és a Microsoft Power BI[1] eszköz használatával alapszinten. Az eszköz részletes bemutatása túlmutat jelen dokumentáció keretein, így ajánlott a megfelelő internetes oldalak felkeresése azoknak, akik mélyebben szeretnének megismerkedni ezzel a technológiával. A Power BI-hoz kapcsolódó információ források gyűjteménye az 13. ábra. fejezetben megtalálható.

Jelen dokumentáció nem választja ketté a méréshez kapcsolódó segédanyagokat a mérés során elvégzendő feladatoktól. A mérés során nem kell külön jegyzőkönyvet készíteni, minden dokumentálás a Power BI-ban készített jelentésben történjen (**rendszeresen mentse el munkáját**)!

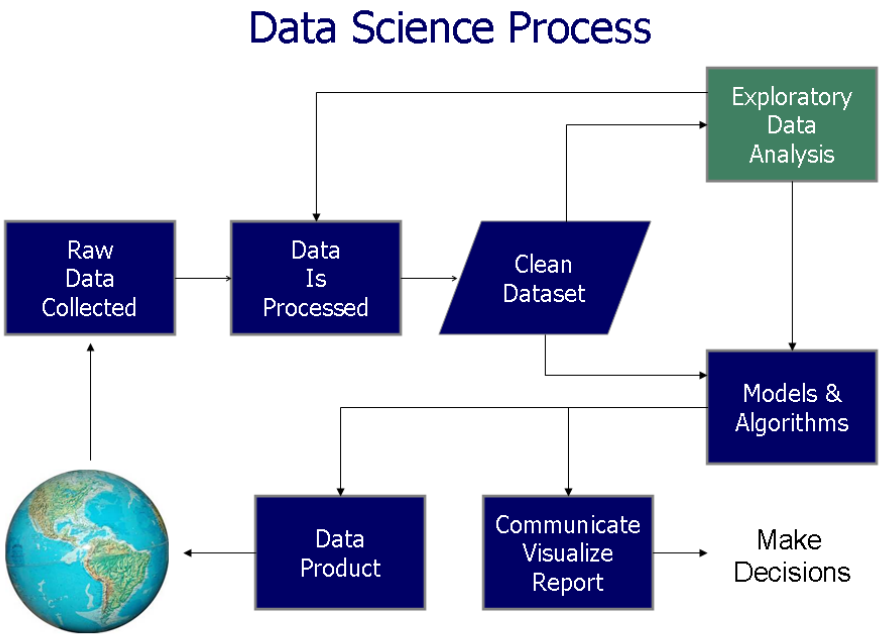
A mérés célja egy olyan módszer/eszköz bemutatása, mellyel a későbbiekben akár az önálló munka során (pl. szakdolgozatban, adott szoftver teljesítménytesztelésénél, stb.) áttekinthető, könnyen módosítható és értelmezhető vizuális megjelenítők készíthetők.

2 Felkészüléshez: Az adatelemzés alapjai és a vizsgált adatkészlet

Informatikai rendszerekben, összetettebb szoftverekben vagy komplex architektúrájú rendszerekben sok adat keletkezik, melyeket a ma elterjedt naplózási/monitorozó rendszerekkel tipikusan könnyű mérni, azonban a gyakorlatban sokszor okoz problémát ezen adatok megfelelő kiértékelése. A mérés célja egy olyan megközelítés bemutatása, mellyel interaktív jelentések készíthetőek jelentősebb programozói tudás nélkül. A mérés során használt eszköz támogatja bonyolultabb elemzések készítését is (pl. alapszintű előrejelzés készítését), erre példát mutatunk a mérésben.

2.1 Az adatelemzés főbb lépései

Az adatelemzéssel foglakozó módszertanok (pl. a CRIPS-DM [5]) tipikusan jól elkülöníthető lépésekre bontják az adatok feldolgozásának és megértésének folyamatát. Ennek a folyamatnak egy általánosnak mondható modellje látható az alábbi ábrán:



Az adatelemzés tipikus lépései [6]

[1] https://powerbi.microsoft.com/en-us/

A mérés során az adatgyűjtéssel nem foglalkozunk (a továbbiakban az adatkészlet ismertetésénél röviden leírjuk az adatok eredetét), a mérés során az adatfeldolgozás (data processing), adattisztítás (data cleaning), a feltáró adatelemzés (visual data analysis) és főleg a jelentéskészítés (reporting) feladatára koncentrálunk. A begyűjtött adatokra néhány egyszerűbb modellt illesztünk, mélyebb algoritmustervezéssel most nem foglalkozunk. A mérés offline elemzést feltételez, bár az adott eszköz képes eseményforrásokból beérkező adatot online is feldolgozni.

2.2 A felhasznált adatkészlet: a TPC-C benchmark eredményei

A mérés során az informatikai rendszerek teljesítményméréséből dolgozunk fel egy példát. A TPC Council TPC-C benchmarkja ([8], érdeklődőknek a teljes specifikáció itt [10]) alapvetően összetett adatbázis rendszerek összehasonlítható teljesítménykiértékelését támogatja. Hasonlóan a legtöbb elfogadott benchmarkhoz, a TPC-C is definiál egy mért mintarendszert (System Under Test): hardvert, operációs rendszert és adatbáziskezelőt mérünk együtt.

A mért rendszer lényegében egy raktárkészlet-kezelést megvalósító relációs adatbázis és az afölött értelmezett tranzakciók (pl. rendelés felvétele). A mérés során egy reprezentatív mintaterhelésnek vetik alá a rendszert (írás-olvasási műveleteket végrehajtó tranzakciók előre meghatározott mixe, adott méretű adatbázis fölött végrehajtva, adott számú emulált felhasználó által).

A benchmarkról publikusam alapvetően két eredmény érhető el: **tpmC** mért rendszer **áteresztőképessége** (egységnyi idő alatt végrehajtott tranzakciók száma), ahol nem egy-egy meghatározott egyedit tranzakció típus, hanem a teljes tranzakció mix végrehajtását értékeljük ki, valamit a **$/tpmC (a teljesítmény fajlagos költsége),** vagyis annak a mérőszáma, hogy egységnyi teljesítmény mennyibe kerül (a mért rendszer beszerzési árát véve alapul).

Bár a TPC-C nem tartozik a legújabb benchmarkok közé, a mai napig használják relációs adatbázis alapú rendszerek összehasonlítására, és kellően jól definiált ahhoz, hogy értelmezhető következtetéseket le lehessen vonni az eredményekből. A mérés során a TPC Council által hitelesített eredményeket használjuk, természetesen egy adott benchmarking mérés kiértékelése, azaz pl. az adott konfiguráció/SUT várható áteresztőképességének meghatározása egy adott mérési kampány adataiból szintén adatfeldolgozási/elemzési feladat lehetne.

2.3 A beugró során számonkért előismeretek

A beugró során nem az eszközismeretet kérjük számon (bár felhívjuk mindenkinek a figyelmét, hogy az eszköz ingyenesen letölthető előre), hanem az adatelemzéssel és a benchmarkokkal kapcsolatos alapismereteket. Ezeket a Rendszermodellezés tárgyban mindenki tanulta, ennek aktuális előadásait, valamint a hátteret bemutató jegyzetet kell a mérés előtt elolvasni:

- Vizuális adatelemzés előadás ([1]) és jegyzet ([3])
- Modellek paraméterezése és benchmarkok előadás ([2])

A beugró során néhány alapvető statisztikai fogalom ismeretére és értelmezésére (medián, átlag, percentilis) és az adatelemzés során alkalmazott alapvető fogalmakra (adatkeret, változók típusai, plotok/megjelenítők főbb típusai és értelmezésük – barchart, hisztrogram, scatter plot, boxplot, mosaic plot), valamint a lineáris regresszió céljára/alkalmazására kérdezünk rá.

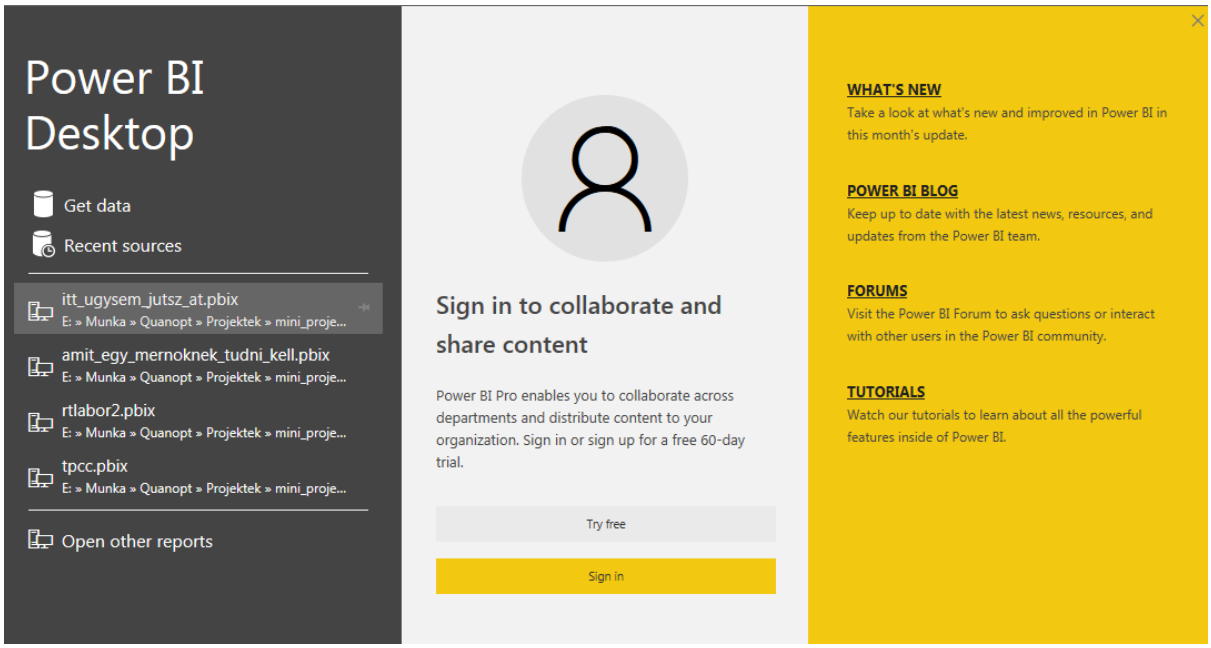
Az alábbiakban ismertetjük a mérés során végrehajtandó feladatokat.

3 Adatok betöltése és transzformálása

Az adatok elemzését azok betöltése és transzformálása előzi meg, amelynek során valamilyen technológia/eszköz segítségével az adatok beolvasásra kerülnek, majd az elemzésnek megfelelő átalakítások hajtódnak rajtuk végre.

3.1 Alkalmazás indítása

A Microsoft Power BI (PBI) Desktop indítása után az 1. ábran látható felület fogad minket. Itt lehetőségünk van a hivatalos termékoldal meglátogatására, vagy a Pro változat esetében bejelentkezni a kapcsolódó szolgáltatásba. Ha már használtuk az alkalmazást, akkor a legutóbbi dokumentumok listájából kiválasztva visszatérhetünk korábbi munkánkhoz, vagy a `Get data` gomb esetében új adat feldolgozásához, elemzéséhez láthatunk hozzá.

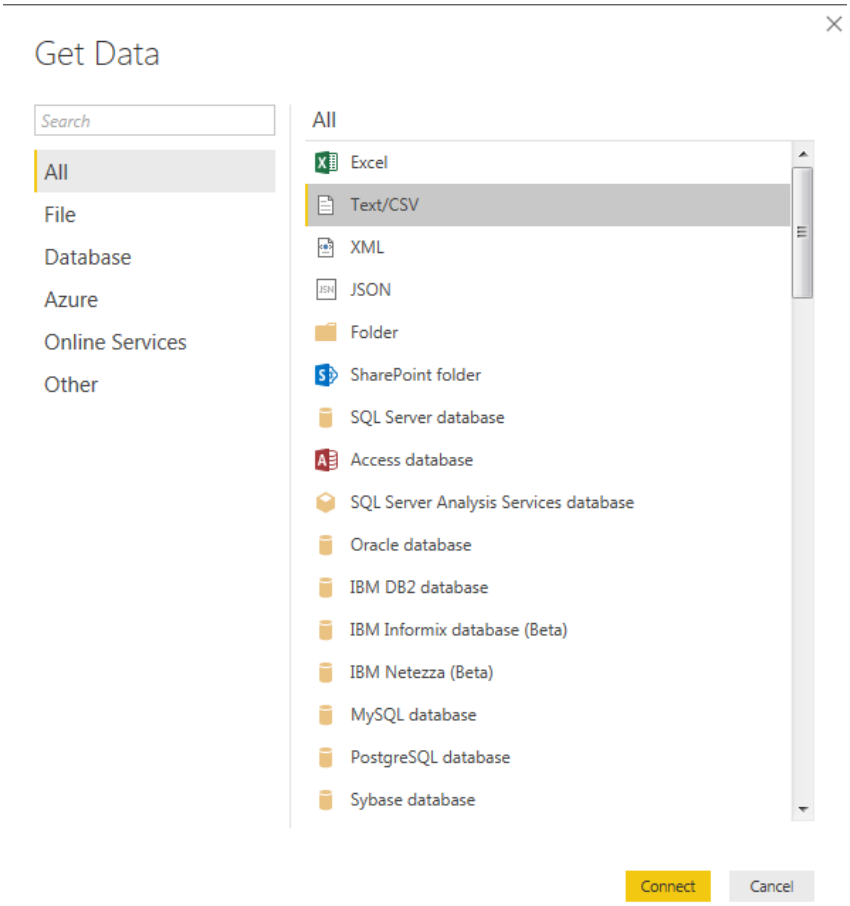


1. ábra A Power BI Desktop nyitóképernyője

3.2 Adatok betöltése szöveges állományból

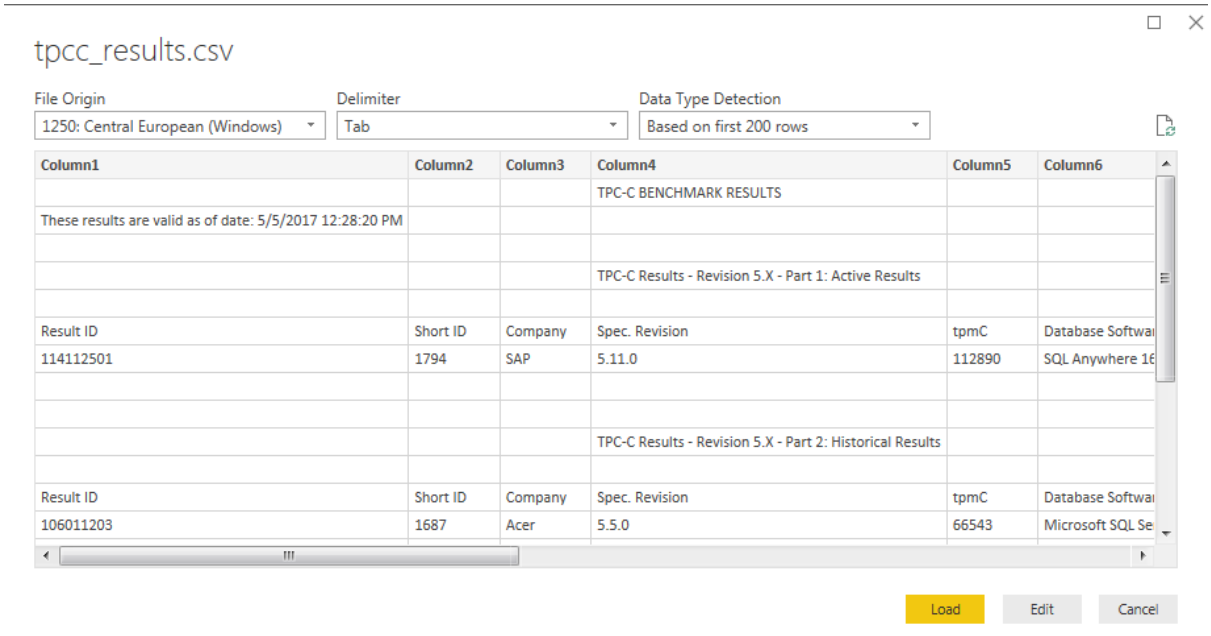
A PBI Desktop változata rengeteg adatforrást támogat. Az olyan alap bemeneteken kívül, mint az Excel, CSV, JSON fájlok olvasása mellett a fontosabb adatbázisokhoz (MySQL, MSSQL, PostgreSQL, stb.), felhős és webes szolgáltatásokhoz is csatlakozni tud. A mérés során először szöveges állományokból fogunk adatokat betölteni.

Kattintson rá a `Get data` gombra, majd a 2. ábraán is látható ablakban válassza ki a *Text/CSV* bemenet típust!



2. ábra A Power BI által támogatott bemeneti adatforrások

A `Connect` gombra való kattintás után keresse ki a tpcc_results.csv[2] állományt, nyissa meg! Ezután a 3. ábra látható ablakot láthatja, ahol a CSV állomány beolvasásával kapcsolatos paramétereket lehet módosítani. Itt az alapértelmezett értékek most megfelelőek, ezért kattintson az `Edit` gombra!



3. ábra CSV állomány beolvasásával kapcsolatos paraméterek beállítása

[2] retelab2_mit2\data\tpcc_results.csv

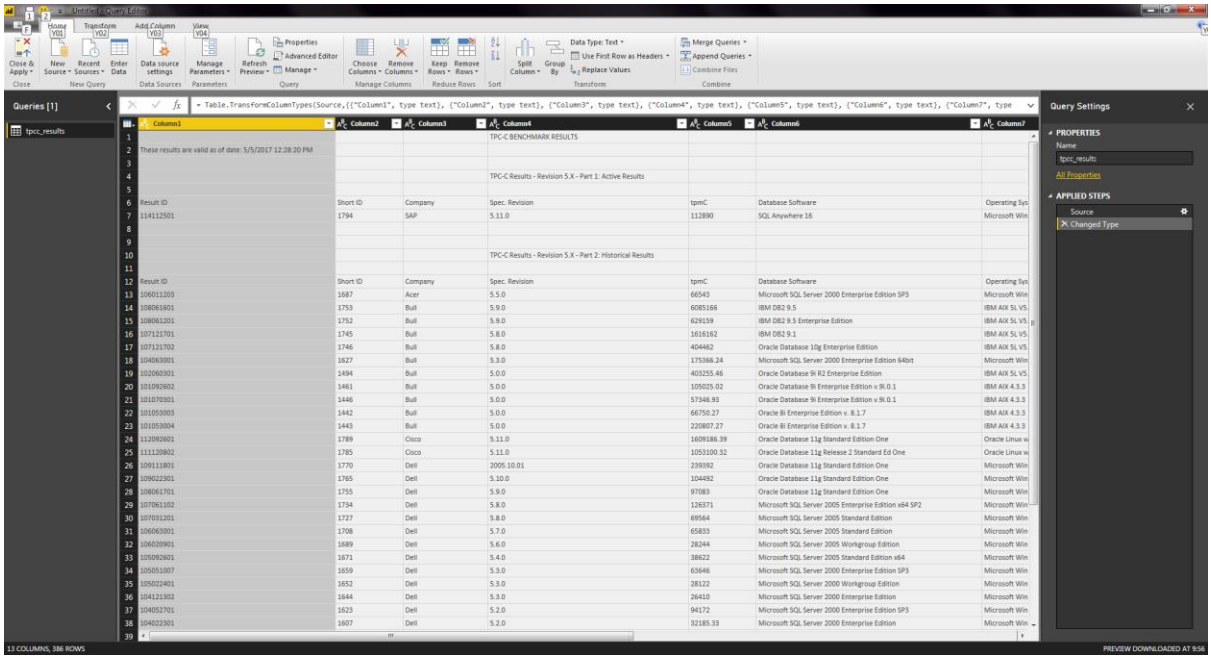
3.3 Adatok áttekintése

Az adatok transzformálásának megkezdése előtt érdemes lehet azok értelmezése. Fontos, hogy az adatelemzőnek legyen valami fogalma, tudása arról, hogy mit jelentenek azok az adatok, amelyekkel dolgozni akar.

Az Edit gombra való kattintás után a 4. ábralátható *Query Editor* felület jelenik meg, ahol a következő lépésekben majd megfelelő formátumúra hozhatjuk adatainkat.

Kattintson az Enter Data gombra, majd hozza létre az 5. ábra látható táblázatot, kiegészítve az oszlopok egy-egy rövid leírásával! Ehhez segítségül hívhatja a http://www.tpc.org/tpcc/default.asp oldalt.

Az elkészült táblázatot nevezze el *col_desc* néven!



4. ábra A Power BI Query Editor felülete

5. ábra Adatoszlopok

3.4 Adatforrás átalakítása – szűrések

A Power BI segítségével egyszerűbb, néhány kattintással megoldható és bonyolultabb (akár a Power Query[3]-vel végrehajtható) átalakításokat lehet elvégezni. Az egyszerűbb műveletek közé tartozik az adatszerkezetből bizonyos oszlopok eltávolítása (pl. adatot nem tartalmazó, vagy szükségtelen adatokat tartalmazó oszlopok), sorok eltávolítása (a Power BI-ben csak az első vagy utolsó n db sort lehet így eltávolítani) vagy szűrése (pl. bizonyos értékek kiszűrése), első sor fejléccé alakítása.

Térjen vissza a `tpcc_results` query-hez! Alakítsa át a táblázatot úgy, hogy csak a szükséges adatokat és fejlécet tartalmazzon! Ehhez használja a `Keep/Remove rows`, `Use First Row as Headers`, és a szűrés funkciókat! Távolítsa el az adatokat nem tartalmazó oszlopokat is (`Remove Columns`)!

Az elkészült query-t duplikálja (jobb klikk a `tpcc_results` query-n majd Duplicate) `tpcc_results_2_feladat` néven, hogy a mérésvezető később ellenőrizhesse az eredményt. Térjen vissza a `tpcc_results` query-re (kattintson a query-re)!

Figyeljen oda, hogy a továbbiakban az eredeti `tpcc_results` query-n dolgozzon tovább!

3.5 Adatforrás átalakítása – adattípusok

Az adatelemzések során fontos pontosan ismerni az egyes adatok típusát, hiszen eltérő ábrák eltérő adattípusokat igényelhetnek. A Power BI az adatforrások beolvasása után megpróbálja automatikusan meghatározni az egyes adatoszlopokban található adatok típusát, ám ezek nem feltétlenül helyesek. Gyakori hiba, hogy a tizedes jegyet elválasztó karakterek (tizedesvessző és pont) keverednek, amely okból kifolyólag a számokat szöveges mezőként próbálja meg az eszköz értelmezni. Ezeket a problémákat a felhasználónak kell megoldania a Query Editor felületen.

Módosítsa az egyes oszlopok adattípusait a tartalmuknak megfelelően! Figyeljen oda, hogy a területi beállítások miatt eltérő lehet a tizedes jegyek elválasztó karaktere! Szükség esetén cserélje ki azokat úgy, hogy a Power BI szám típusra állítás esetén ne dobjon hibát (`Replace Values`)! Figyeljen oda, hogy a dátum/idő típusok esetében csak az érdemi tartalmat visszaadó típust válassza ki (pl. nem feltétlenül szükséges a pontos idő ismerete, elégséges lehet az év, hónap és nap)! A „Result ID" oszlop szöveges típusú legyen! Duplikálja az elkészült táblázatot `tpcc_results_3_feladat` néven!

Figyelje oda, hogy a továbbiakban az eredeti `tpcc_results` query-n dolgozzon tovább!

A `Close & Apply` gombra kattintva alkalmazza a változtatásokat!

4 Elemzések

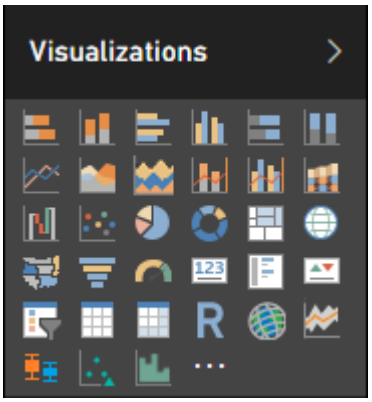
Az adattok transzformációja után kezdődhetnek azok elemzései. A mérés során a vizuális adatelemzést fogjuk alkalmazni, amely során az adatokat különböző módszerekkel ábrázoljuk és az elkészült diagram segítségével vonunk le azokból következtetéseket.

A Power BI működésében két fő módot különböztetünk meg. Az első a már fentebb használt Query szerkesztő mód, amelyben az adatok transzformálást, átalakítását tudjuk elvégezni, a másik pedig a következőekben bemutatásra kerülő jelentés szerkesztő mód, ahol különböző ábrákat, szövegdobozokat, képeket tudunk elhelyezni. A két mód között pl. az **`Edit Queries`** (jelentés → query szerkesztés) és a **`Close & Apply`** (query → jelentés) gombok között tudunk váltani.

[3] https://support.office.com/en-us/article/Getting-Started-with-Power-Query-7104fbee-9e62-4cb9-a02e-5bfb1a6c536a

4.1 Adatkészlet áttekintő vizsgálata - Stacked bar/column chart, treemap

A Power BI segítségével jelentések készíthetők, amely jelentésekre diagramok, képek, szövegek kerülhetnek. Az egyes diagramtípusokat a *Visualizations* panelen találhatjuk (lásd 6. ábra). A beépített diagramok használata mellett lehetőség van ún. custom visualok importálására is, amelyek a mérés során kipróbálásra fognak kerülni.



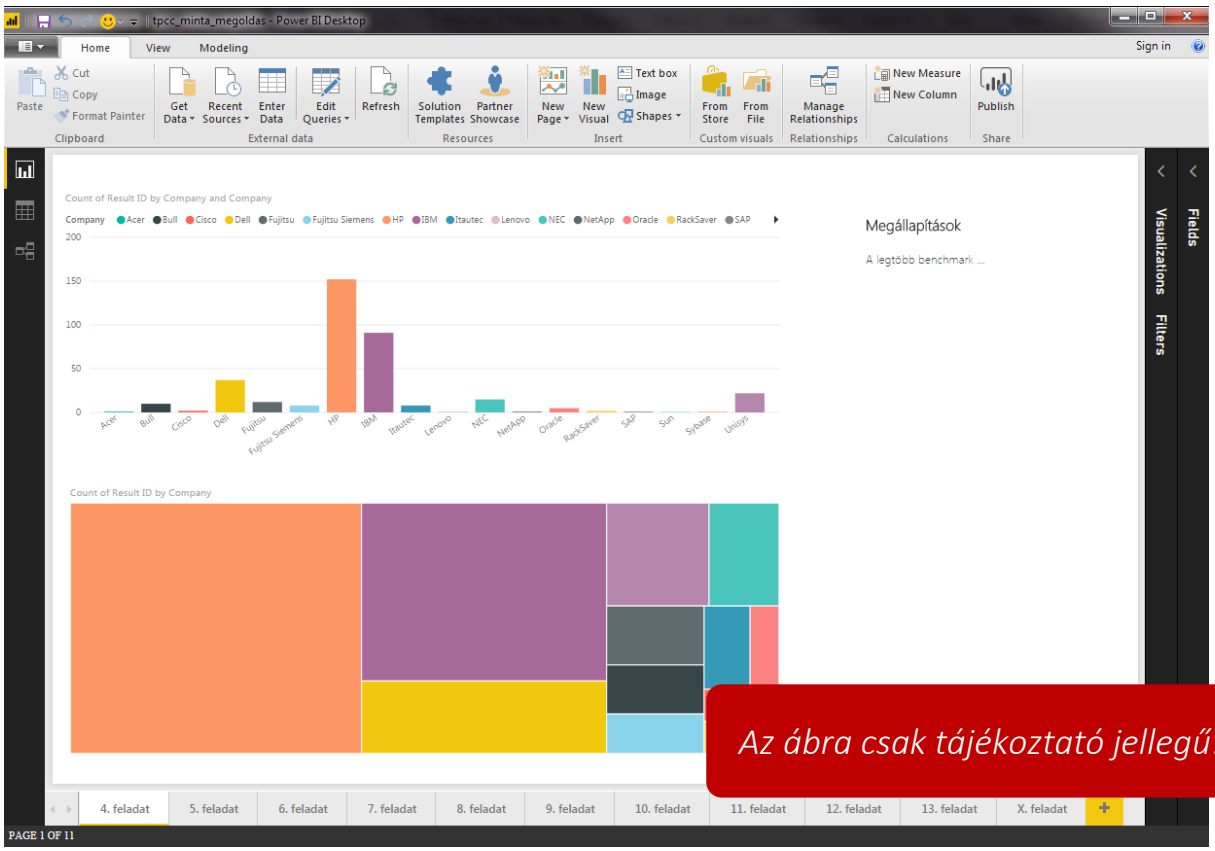
6. ábra A Visualizations panel

A következők az oszlopdiagramok (Stacked bar chart, Stacked column chart és százalékos változatainak) használatát mutatja be. Az oszlopdiagramok segítségével diszkrét változó egyes értékeinek gyakoriságát tudjuk megjeleníteni, ahol egy oszlop magassága az adott érték abszolút gyakoriságát jelenti. A Power BI Stacked bar chartja ettől annyiban tér el, hogy a változó számossága mellett egy plusz kategorizálását is meg tudja jeleníteni egy oszlopon (pl. diszkrét változó az év, az oszlop magassága az adott évből származó benchmark eredmények száma, kategóriák az egyes cégek).

A Treemap hasonló elemzések megjelenítésére alkalmas, de itt az egyes téglalap területek méreteiből lehet következtetni az értékek gyakoriságára.

Stacked bar/column chart **segítségével készítsen ábrát, amelyből megállapítható, hogy mely beszállítók/cégek szerepelnek a legtöbb benchmark eredménnyel az adatkészletben! Ugyanezt mutassa meg treemap-pel is! Az oldal neve legyen "*4. feladat*", és szövegdobozban írjon néhány összefoglaló gondolatot a megállapításokról!**

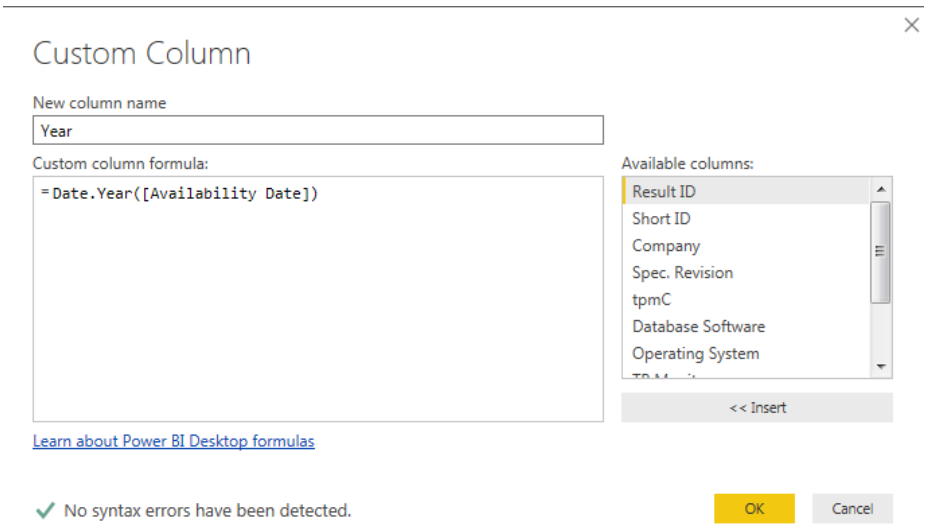
Megjegyzés: A Power BI egyes diagramjainak paraméterezése nem feltétlenül magától értetődő, ilyenkor érdemes lehet az adott visualra rákeresni az interneten, ahol rengeteg blog és videó található a témában. A mostani feladat esetében azt szeretnénk, ha annyi oszlop jelenne meg az ábrán, amennyi céget (Company) tartalmaz az adatsor, ezért ennek az oszlopnak kell szerepelnie az Axis paraméternél. Az oszlopok magasságának az adott cég által beküldött benchmark eredmények számát kell reprezentálnia. Az eredmények egyértelműen megkülönböztethetők a Result ID alapján, amit ha a diagram Values paraméterére húzunk, automatikusan kiegészül a Count of előtaggal, vagyis a Power BI összeadja azoknak az egyedi ID-knak a számát, amik az adott céghez tartoznak. Ha a Legend paraméternek megadjuk ugyanazt az oszlopot, mint amit az Axisnak is, akkor minden oszlopunk egyedi színt kap. Próbálja ki, hogy mi történik, ha ide más (pl. Database software) oszlopot adunk meg!



7. ábra Adatkészlet áttekintő vizsgálata - Stacked column chart, treemap

4.2 Elemzések - Mely években megjelent konfigurációkat tartalmazza a benchmark? Mennyire használható/releváns napjainkban?

A PBI-ben nem csak eltávolítani, hanem létrehozni is lehet új oszlopok. A *Query Editor* `Add Column` fülén található `Custom Column` gombja a 8. ábra látható ablakot nyitja meg, ahol már meglévő oszlopok és Power Query formulák segítségével lehet létrehozni új oszlopot.

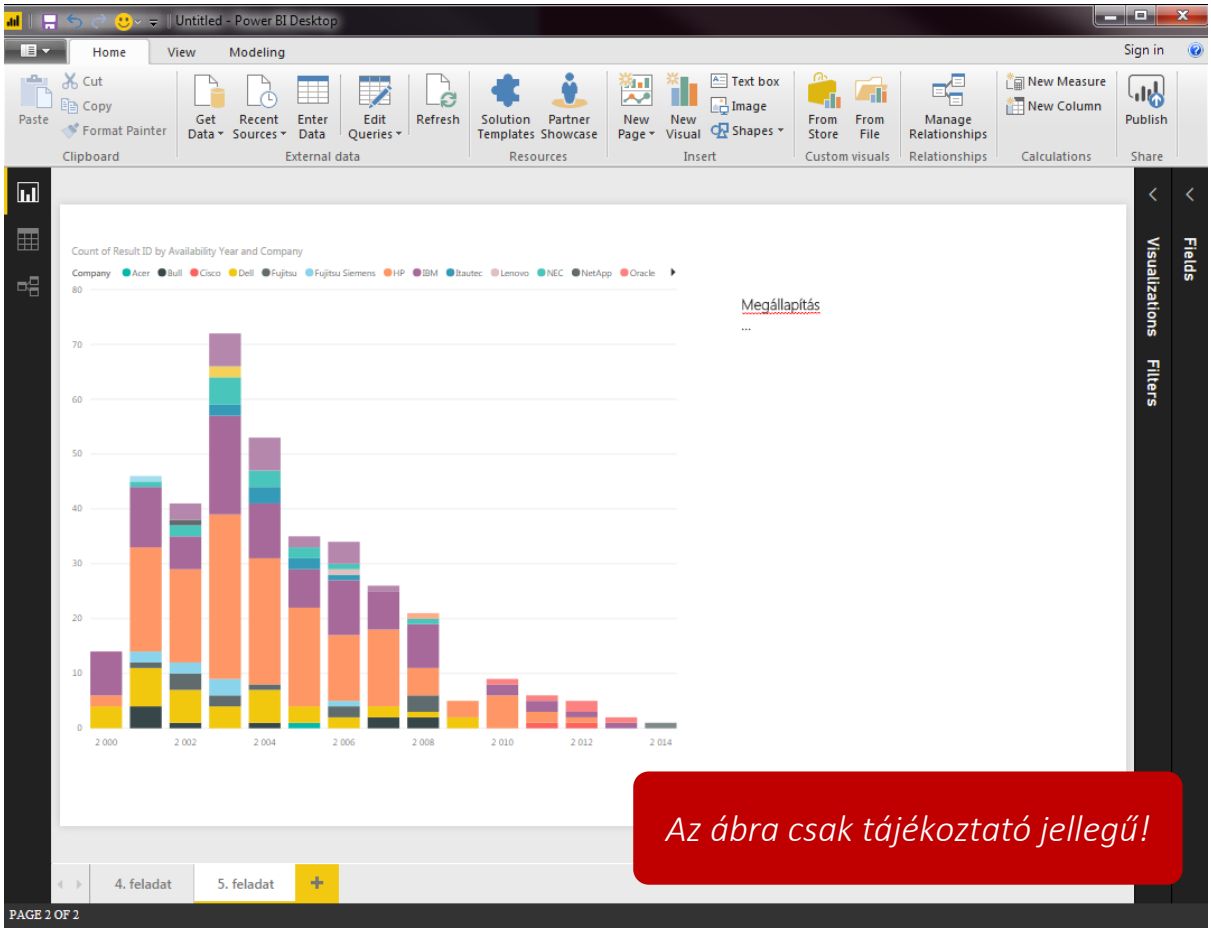


8. ábra Új oszlop létrehozása

***Stacked bar/column chart* segítségével készítsen egy ábrát, amelyről kiderül, hogy mely évekből vannak benchmark eredmények! Ehhez használja az `Availability Date` oszlopban található adatokat!**

Segítség: készítsen új oszlopot az `Availability Date`-ből, amely csak az évet tartalmazza! Ehhez kattintson az `Edit Queries` gombra, majd az `Add Column` fülön a `Custom Column` gombra!

A kész ábra külön oldalra kerüljön, amelynek a neve "5. feladat" legyen! Ne felejtse el megállapításait is rögzíteni az oldalra!



9. ábra Elemzések - Mely években megjelent konfigurációkat tartalmazza a benchmark? Mennyire használható/releváns napjainkban?

4.3 Elemzések - Az egyes beszállítók mely években voltak aktívak? Mely beszállító a legjelentősebb piaci szereplő?

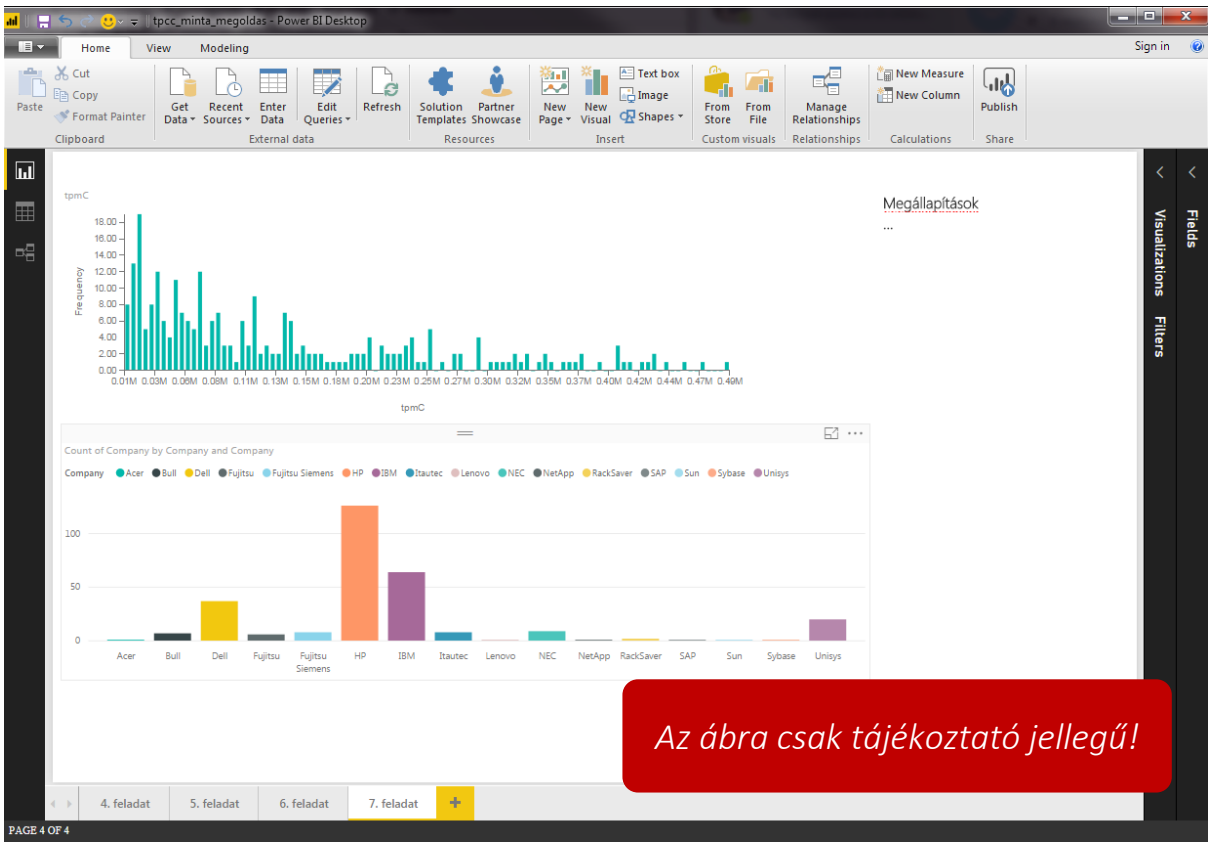
A Power BI egyik kiemelkedő funkciója a crossfiltering/selection. Ez a funkció lehetőséget biztosít arra, hogy az egy oldalon szereplő diagramok és egyéb elemek között szűrési és kiválasztási kapcsolatok aktiválódjanak, vagyis egy elemen történi kiválasztás vagy szűrés az oldal többi elemén is végrehajtódik. Ezen funkció segítségével könnyebbé válik az egyes adatok értelmezése.

Sok esetben előfordul, hogy az adathalmaz csak egy részét szeretnénk adott pillanatban megjeleníteni. Ebben lehet segítségünkre a Slicer megjelenítő, amely egy adott változó értékeit checkboxok listájaként jeleníti meg, így a jelentésben a szűrés eredménye látható.

Készítsen "6. feladat" néven egy új oldalt (akár az előző duplikálásával), amelyen egyértelműen megmutatható, hogy mely beszállító, mely évben volt aktív és melyik beszállító a legaktívabb a teljes időszakra nézve. Használja a *Slicer megjelenítőt* a könnyebb áttekinthetőség érdekében! Rögzítse szövegdoboz segítségével a fontosabb megállapításokat!

4.4 Elemzések - Ha cégünk igényeit egy alacsonyabb teljesítményű konfigurációval is ki tudjuk elégíteni, akkor mely beszállítók közül válasszunk?

A feladat megoldásához importálja a Histogram custom visualt[4], amelyen ábrázolja a `tpmC` értékeket! Az oldal szintű szűrő segítségével szűkítse a konfigurációs benchmark eredményeket úgy, hogy cégünk adatbázisának naponta csak átlagosan 25,92 – 31,68 millió közötti tranzakciót kell kiszolgálnia! Column chart segítségével ábrázolja, hogy melyik cég benchmark eredményei szerepelnek a szűrt adatkészletben! Mentse az oldalt "7. feladat" néven! Ne feledkezzen meg a megállapítások rögzítéséről!



10. ábra Elemzések - Ha cégünk igényeit egy alacsonyabb teljesítményű konfigurációval is ki tudjuk elégíteni, akkor mely beszállítók közül válasszunk?

4.5 Elemzések - Mit lehet megállapítani a teljesítmény változásáról?

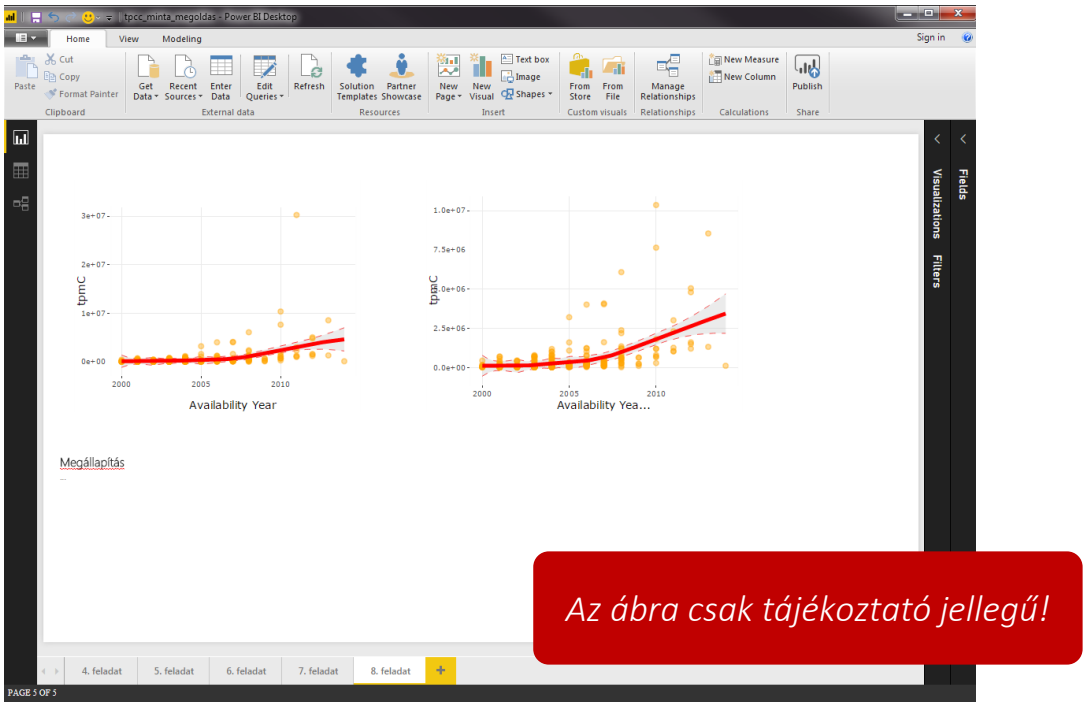
A Power BI lehetőséget biztosít arra, hogy azt összekapcsoljuk az R statisztikai programkörnyezettel. Az integráció megléte esetében mind az adattranszformáció, mind a megjelenítés során felhasználhatók az R nyelv által nyújtott képességek. Több olyan úgynevezett „Custom Visual" is található az interneten, amely valamilyen R csomag alkalmazásával jelenít meg valamilyen diagramot az adatokból.

Megjegyzés: természetesen az R nyelv és bizonyos csomagok megléte szükséges az adott rendszerkörnyezetben. A laboron használt környezetben a szükséges programok telepítésre kerültek.

[4] `retelab2_mit2\custom_visuals\Histogram.1.0.2.0.pbiviz`

Hozzon létre egy új oldalt "8. feladat" néven! Importálja a PowerBI-visuals-spline custom visualt[5] és ábrázolja vele a teljesítmények időbeli alakulását!

Készítsen még egy ábrát ugyanezzel a Spline megjelenítővel, de szűrő segítségével távolítsa el a kiugró értéket (outlier-t)! Szövegdobozban foglalja össze a teljesítmény változás alakulását, külön kitérve a két ábra közötti különbségre!

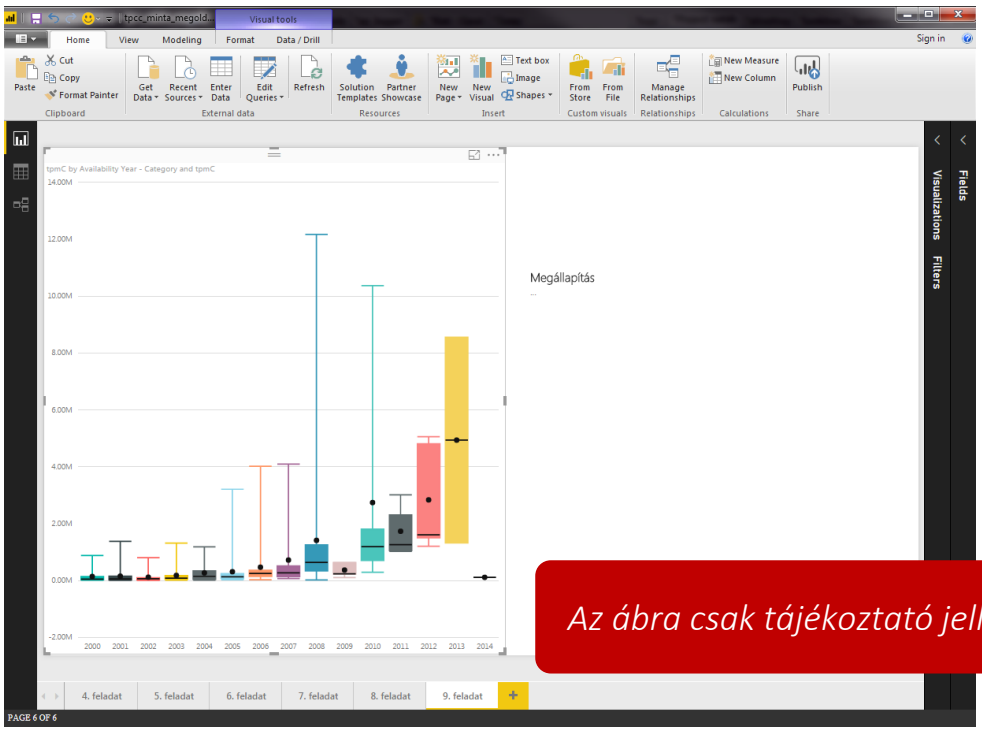


11. ábra Elemzések - Mit lehet megállapítani a teljesítmény változásáról?

4.6 Elemzések - Mit lehet megállapítani a teljesítmény változásáról? – Boxplot

Importálja a Power BI-ba a BoxWhisker visualt[6] és ábrázolja rajta a teljesítmények alakulását az évek során. A boxplot kategória paraméterének érdemes lehet létrehozni az "Availability Year" oszlopból egy kategorikus változót (oszlop duplikálása és típus módosítása szöveges típusra). Fontos a BoxWhisker visual alapértelmezetten nem a Rendszermodellezés tárgyból megtanultan paraméterezi fel a boxplot diagramot, állítsa be a paramétereket úgy, hogy az ott megtanultak szerinti legyen (lsd. segédlet[1]). Távolítsa el itt is az outliert. Megállapításait szövegdobozba rögzítse! Mentse az oldalt "9. feladat" néven!

[5] `retelab2_mit2\custom_visuals\PowerBI-visuals-spline.1.0.3.0.pbiviz`
[6] `retelab2_mit2\custom_visuals\BoxWhisker(JanPieter).2.0.0.0.pbiviz`



12. ábra Elemzések - Mit lehet megállapítani a teljesítmény változásáról? – Boxplot

4.7 Elemzés - Hogyan alakulnak a tranzakciós és összköltségek?

A Power BI támogatja több adatforrás használatát, amelyek között kapcsolatokat is létre lehet hozni, mint ahogy azt a nagyobb adatbázis sémák kezelése esetében is. Az így létrehozott relációk segítségével egy diagramnak több adatforrásból származó paramétert is megadhatunk.

A jelenleg használt adatkészletben nem szerepelnek költségekkel kapcsolatos értékek, ezeket egy másik CSV állományból kell beolvasni. Ehhez a *Query Editorban* kattintson a `New Source` gombra és a korábban már megismert módon importálja a `tpcc_costs.csv`[7] fájlt! Ellenőrizze, hogy az egyes oszlopok megfelelő típusúak-e! A szükséges módosítások után lépjen ki a *Query Editorból* (ne felejtse el alkalmazni a változtatásokat)!

A baloldalon található `Relationships` gombra kattintva megnyílik egy felület, amelyen az eddigi query-k eredménytáblái láthatóak. Fogja meg a `tpcc_results` tábla „Result ID” mezőjét, és húzza a `tpcc_costs` táblára! Ennek eredményeként a két tábla között 1 - 1 kapcsolatnak kell megjelennie. Ez után a Reportra visszatérve már a `tpcc_costs` és `tpcc_results` táblából származó paramétereket is megadhat egy visual típusnak.

Készítsen diagramot, amelyen jól látható, hogy az évek során csökkent a teljesítményre vett költség a teljes piacra nézve, majd egy másikat, amelyiken látható, hogy ugyanez nem áll fent a teljes költség esetében! Szövegdobozba írja össze a megállapításokat! Az oldalt mentse “10. feladat” néven!

4.8 Elemzés - Milyen adatbázis-kezelő szoftvert válasszunk, ha még mindig az olcsóbb megoldást szeretnénk használni?

Amikor egy diagram paraméterének megadunk egy változót (oszlopot), akkor a Power BI sok esetben az egyes paramétereket megpróbálja aggregálni azok típusának megfelelően. Ez akár nem várt eredményeket okozhat.

[7] `retelab2_mit2\data\tpcc_costs.csv`

Hibás aggregációt eredményezhet, amikor egy általunk szám típusnak gondolt változó típusa nem szám, ilyenkor ugyanis kategorikus változóként szerepel az adatkészletben, amire nem értelmezhetők az olyan aggregáló függvények, mint a minimum, átlag vagy medián. Ilyenkor az adott változó típusának ellenőrzése segíthet a probléma megoldásában.

Bar/Column chart segítségével ábrázolja az egyes adatbázis-kezelő szoftvereket és a futtató konfigurációk teljes költségét! Milyen aggregációs műveletet választott ki, hogy megfelelő eredményt kapjon az egyes oszlopok magasságára? Milyen problémával szembesült?

Hozzon létre új oszlopot (`Conditional Column`), amely új kategóriákba rendezi a DB szoftvereket (pl. MSSQL, IBM DB2, stb.)! Ábrázolja az új oszlop segítségével is az előző ábrát! Szövegdobozban gyűjtse össze az észrevételeket, megállapításokat! Melyik kategóriát választaná, ha egy olcsóbb megoldást keres? Az oldalt "11. feladat" néven mentse!

4.9 Párhuzamos koordináták – Változók közötti kapcsolat keresése

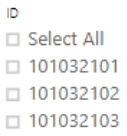
A feladat megoldásához ismételje át a párhuzamos koordináták módszeréről tanultakat.

A korábban már megismert módon olvassa be a `tpcc_longformat_data.csv`[8] állományt. Alakítsa át az ID oszlopot Text típusúvá. Ábrázolja Párhuzamos Koordinátáták Grafikonon a változókat.

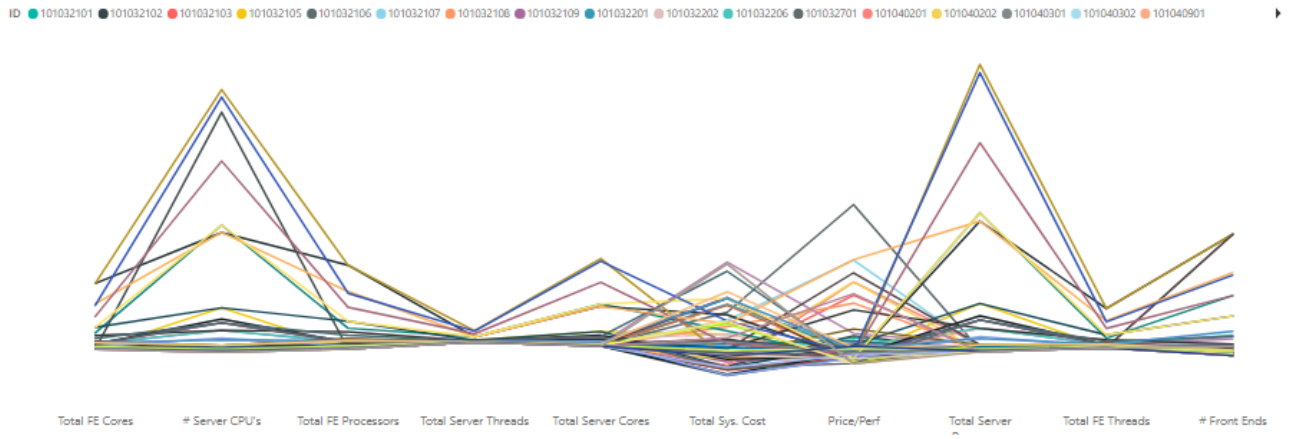
A Power BI programhoz jelenleg nem található külön megjelenítő, amely ilyen grafikont tud ábrázolni, éppen ezért használja a beépített Linechart diagramtípust ügyelve arra, hogy mi kerül az Axis, a Legend, a Values és a Tooltip mezőbe.

Ugyanerre az oldalra tegyen a Slicer visual segítségével egy olyan szűrőt, amellyel akár több ID-t is ki lehet jelölni egyszerre.

Milyen összefüggéseket tud leolvasni az ábráról? Gondolja végig, mi a különbség az eddigiekben használt adatformátum, és a mostani közt! Melyiket milyen célra/körülmények közt lehet érdemes használni?



Az ábra csak tájékoztató jellegű!

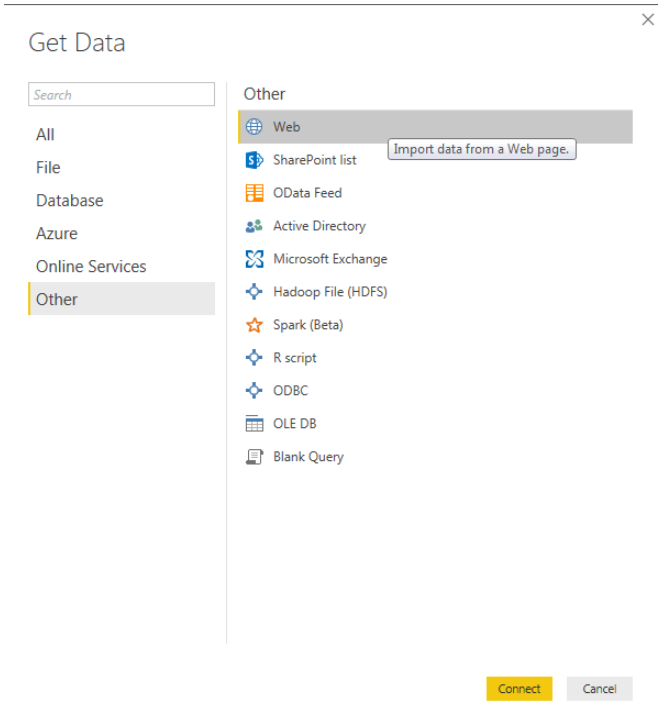


[8] `retelab2_mit2\data\tpcc_longformat_data.csv`

4.10 Extra feladat (IMSc) - Valuta árfolyamok problémája

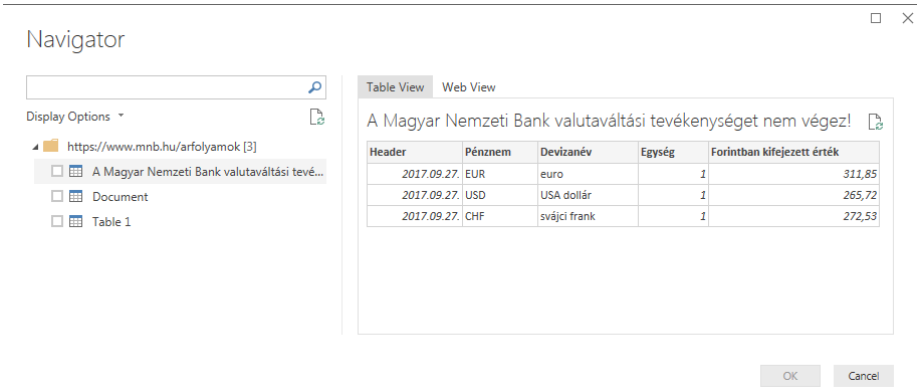
A 4.1 – 4.8 feladatok megoldása során elkövettünk egy hibát. A költségekkel kapcsolatos vizsgálatok során nem lett figyelembe véve, hogy azok eltérő valutákban voltak megadva. Váltsa át a valutákat Forintban kifejezett értékekre!

Adjon hozzá új adatforrást a meglévőekhez! Válassza ki az *Other* - *Web*-et, majd kattintson a `Connect` gombra!



Adja meg a következő URL-t: http://www.mnb.hu/arfolyamok

Nézze meg a talált elemeket (a baloldali fastruktúrában ha rákattint egy elemre, akkor a jobb oldalon megjelenik egy előnézeti kép)! Vizsgálja meg, hogy a tpcc_costs táblában milyen valutákat talál, és ennek alapján válassza ki azokat az oldalelemeket, amelyek szükségesek ahhoz, hogy minden valuta esetében el lehessen végezni az átváltást! A kiválasztás után kattintson az `OK` gombra!



Ellenőrizze a Query Editorban az újonnan importált táblázat(ok) oszlopainak típusait, szükség esetén módosítsa azokat! Távolítsa el azokat az oszlopokat, amelyekre nincs szükség!

Ha több táblázatot is használnia kell, akkor jelölje ki az egyik query-t, majd kattintson az `Append Queries` gombra, ezután pedig válassza ki azt a táblázatot (query-t), amelyiket hozzá szeretné fűzni! A már teljes táblázatot nevezze el "mnb_arfolyam" néven!

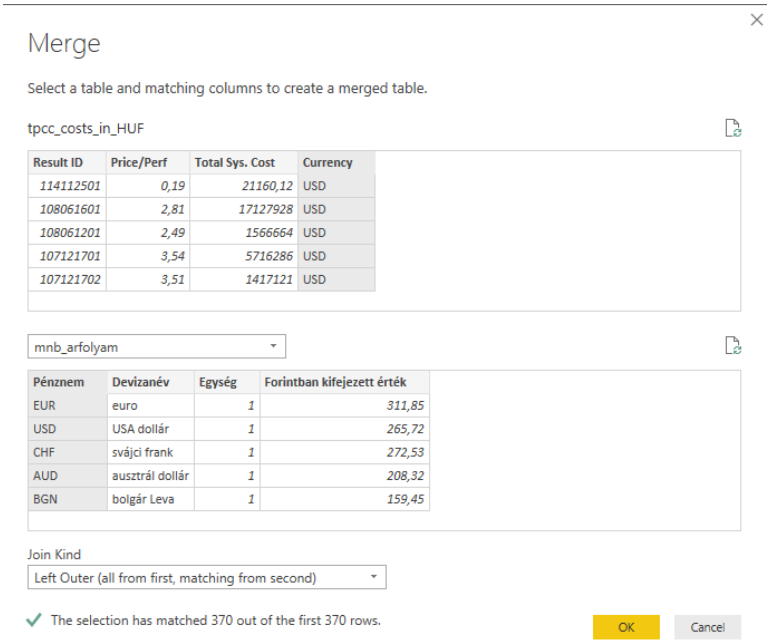
Hozzon létre egy referenciát a `tpcc_costs` query-ből (jobb kattintás a query-n majd, Reference menüpont) `tpcc_costs_in_HUF` néven! Kattintson rá, majd a `Merge Queries` gombra! A megjelenő ablakban kattintson a *Currency* oszlopra, a legördülű menüből válassza ki az `mnb_arfolyam` táblázatot, majd itt is kattintson a *Pénznem* oszlopra, ezután pedig az `OK` gombra (lásd 13. ábra)!

A megjelent mnb_arfolyam oszlopnál kattintsunk a két irányba mutató nyilas gombra, majd az OK-ra! Adjunk hozzá új oszlopot a query-hez „TSC in HUF" néven a

```
= [Total Sys. Cost]/[mnb_arfolyam.Egység]*[mnb_arfolyam.Forintban
                    kifejezett érték]
```

formulával! Hasonló módon hozzunk létre egy másik oszlopot a „Price/Perf" oszlopból „in HUF" utótaggal! `Close & Apply`-ra kattintással mentse a munkát!

Külön bar/column chartokon mutassa meg a teljes költséget, és az ár/teljesítmény mutatót Forintban adatbázis-kezelőkre és operációs rendszerekre nézve! Szövegdobozban foglalja össze a megállapításait!



13. ábra Merge ablak

4.11 Extra feladat (IMSc) – Található-e összefüggés az egyes konfigurációk CPU jellemzői és áteresztőképessége között?

A korábban már megismert módon olvassa be és kapcsolja össze a már meglévő query-kkel a `tpcc_hardware.csv`[9] állományt, majd készítsen egy új oldalt, amelyen összefüggéseket mutat a konfigurációk CPU jellemzői (CPU szám, szálak száma, stb.) és áteresztőképessége között! Megállapításait szövegdobozba gyűjtse!

[9] `retelab2_mit2\data\tpcc_hardware.csv`

5 Ajánlott irodalom a felkészüléshez

[1] Rendszermodellezés előadás diasor (2017) - Vizuális adatelemzés https://inf.mit.bme.hu/edu/courses/materials/rendszermodellez%C3%A9s/2017-tavasz/vizu%C3%A1lis-adatelemz%C3%A9s

[2] Rendszermodellezés előadás diasor (2017) – Modellek paraméterezése, benchmarkok https://inf.mit.bme.hu/edu/courses/materials/rendszermodellez%C3%A9s/2017-tavasz/modellek-param%C3%A9terez%C3%A9se-benchmarkok

[3] Antal Péter – Antos András – Horváth Gábor – Hullám Gábor – Kocsis Imre – Marx Péter – Millinghoffer András – Pataricza András – Salánki Ágnes: Intelligens adatelemzés (Typotex, 2014), 5. fejezet, Vizuális Analízis (A teljes könyv elérhető a http://www.interkonyv.hu/konyvek/antal_peter_intelligens_adatelemzes címen)

[4] Microsoft Power BI – Guided Learning https://powerbi.microsoft.com/en-us/guided-learning/

További háttéranyagok

[5] Shearer C., The CRISP-DM model: the new blueprint for data mining, J Data Warehousing (2000); 5:13—22.

[6] Schutt, Rachel, and Cathy O'Neil. Doing data science: Straight talk from the frontline. " O'Reilly Media, Inc.", 2013.

[7] Kocsis Imre, Salánki Ágnes: Vizuális adatanalízis - „Big Data” elemzési módszerek https://inf.mit.bme.hu/sites/default/files/materials/taxonomy/term/446/14/20141015_BigData_5-6_ea_EDA.pdf

[8] Francois Raab, Walt Kohler, Amitabh Shah. Overview of the TPC-C Benchmark: The Order-Entry Benchmark. http://www.tpc.org/tpcc/detail.asp

[9] TPC-C http://www.tpc.org/tpcc/default.asp

[10] TPC-C benchmark Standard Specification. 2010, TPC Council. http://www.tpc.org/tpc_documents_current_versions/pdf/tpc-c_v5.11.0.pdf