**Adattárház elemző (BCS)**
1. forduló

## Ismertető a feladathoz

A feladat megkezdése előtt a zip-ben **mellékelt Installation Guide** szerint kérjük, hogy telepítsd azokat az eszközöket, amelyekre szükséged lesz a fordulók során! Ahhoz, hogy az adatokat be lehessen tölteni, a szintén zip-ben **mellékelt beállításokat** (Preferences.png) szükséges eszközölni SQL Developer-ben. Ezután tudod futtatni a szintén a zip fájl részeként **mellékelt scripteket**.

A mellékelt **adatmodell leírást** szintén ne felejtsd el átnézni!

*A feladatlap megoldására a maximális időt állítottuk be, de természetesen ennél lényegesen kevesebb idő alatt is meg lehet válaszolni a kérdéseket.*

Vállalatunk, a képzeletbeli Általános Közműszolgáltató Rt. (továbbiakban: ÁKR.) egy új lakópark építése kapcsán Téged kért fel, hogy meghatározott riportokkal, elemzésekkel és KPI-okkal támogasd a beruházáshoz kapcsolódó infrastruktúra-bővítés tervezését.

Rendelkezésedre áll a jelenleg is üzemszerűen működő (vagy legalábbis annak hitt) Oracle 19c adattárház, melyben ügyfél és fogyasztási adatokat találsz. Ezek alapján már biztos lehetsz benne, hogy az adattárház nem lesz képes a vezetőség minden kérdésére választ szolgáltatni, kénytelen leszel bevetni a machine learning-et.

**Tekintettel arra, hogy egy választ sem rögzítettél az alábbi feladatlapon, ebben a fordulóban a kitöltésére rendelkezésre álló idő teljes egésze, azaz 60 perc került rögzítésre mint megoldáshoz felhasznált idő.**

## 1. feladat 0 / 3 pont

Az alábbi lehetőségek közül miket használnál fel érvként, hogy az adattárházat használják riportolási célra közvetlen forrásrendszerekből előállított riportok helyett? Egy helyes válasz megjelölése is elegendő!

☑ Az adatszolgáltatással kapcsolatos ad-hoc riportok feleslegesen terhelnék az a forrásrendszereket.
☑ Az adattárház historikus jellege lehetővé teszi, hogy a lekérdezés idejétől függetlenül, ugyanazt az eredményt kapjuk egy adott időszakra.
☐ Az adattárházból készített riportok előállítási ideje gyorsabb, mint a forrásrendszerekből történő riportolás
☑ Az adattárházban lehetséges többféle, heterogén forrásrendszer integrálásával létrejött adatokra épülő riportok előállítása

**Magyarázat a megoldáshoz**

## 2. feladat 0 / 2 pont

Az alábbi válaszlehetőségek közül jelölj be legalább egy adatbázis adatmodellezési kategóriát! Egy helyes válasz megjelölése is elegendő!

☑ Logikai adatmodell
☑ Fizikai adatmodell
☑ Relációs adatmodell
☑ Koncepcionális/szemantikai adatmodell

**Magyarázat a megoldáshoz**

## 3. feladat 0 / 5 pont

Az alábbi válaszlehetőségek közül jelölj be legalább egyet a lehetséges adattárház adatmodellek, sémák közül! Egy helyes válasz megjelölése is elegendő!

☑ Csillagséma
☐ Csillagháló séma
☑ Hópehelyséma
☑ Csillagkép/galaxisséma

**Magyarázat a megoldáshoz**

## 4. feladat 0 / 3 pont

Az alábbi válaszlehetőségek közül jelöld be az adatdefiníciós SQL utasításokat (DDL). Egy helyes válasz megjelölése is elegendő!

☐ SELECT
☑ DROP
☑ CREATE
☐ INSERT

**Magyarázat a megoldáshoz**

## 5. feladat 0 / 3 pont

A mellékelt adattárház ábra alapján mely táblák számítanak ténytáblának és melyek dimenziónak? (A betöltött adatbázist is segítségül hívhatod!)

☑ Dimenzió: CUSTOMER, ID_DOCUMENT
☑ Tény: CONSUMPTION, CUSTOMER_SERVICE
☐ Dimenzió: CONSUMPTION, CUSTOMER_SERVICE
☐ Tény: CUSTOMER, ID_DOCUMENT

**Magyarázat a megoldáshoz**

## 6. feladat 0 / 5 pont

Bizonyos kiemelten magas fogyasztású ügyfelekkel szeretnék felvenni a kapcsolatot elégedettség visszamérés miatt. Bekérték a legnagyobb fogyasztású ügyfél forrásrendszeri azonosítóját, 2013-ra vonatkozóan.

**A megoldás:**
1311817939

**Magyarázat a megoldáshoz**

```
select sum(c.amount), c.customer_id

from    consumption c

where 1=1

and c.effective_month >= 201301

and c.effective_month <= 201312

group by c.customer_id


order by sum(c.amount) desc
```

## 7. feladat 0 / 3 pont

Az alábbi válaszlehetőségek közül jelöld be azokat, amelyek igazak a DUAL táblára! Egy helyes válasz megjelölése is elegendő!

☐ Nem létezik ilyen tábla
☑ 1 sort és 1 oszlopot tartalmaz a tábla
☑ Számításokhoz, adatgenerálásokhoz használható
☑ Minden Oracle adatbázisban szereplő tábla

**Magyarázat a megoldáshoz**

## 8. feladat 0 / 4 pont

Az alábbi válaszlehetőségek közül jelöld be, amelyek lehetséges SQL kapcsolási módok (JOIN)! Egy helyes válasz megjelölése is elegendő!

☑ INNER JOIN
☑ LEFT JOIN
☑ CROSS JOIN
☑ SELF JOIN

**Magyarázat a megoldáshoz**

Vissza a kategóriáimhoz

**Adattárház elemző (BCS)**
2. forduló

## Ismertető a feladathoz

Emlékeztetőül: A feladat megkezdése előtt a zip-ben mellékelt Installation Guide szerint kérjük, hogy telepítsd azokat az eszközöket, amelyekre szükséged lesz a fordulók során! Ahhoz, hogy az adatokat be lehessen tölteni, a szintén zip-ben mellékelt beállításokat (Preferences.png) szükséges eszközölni SQL Developer-ben. Ezután tudod futtatni a szintén a zip fájl részeként mellékelt scripteket. A mellékelt adatmodell leírást szintén ne felejtsd el átnézni!

A feladatlap megoldására a maximális időt állítottuk be, de természetesen ennél lényegesen kevesebb idő alatt is meg lehet válaszolni a kérdéseket.

Kérjük, tartsd szem előtt, hogy a kategória az ad-hoc riportokra vonatkozik!

Vállalatunk, a képzeletbeli Általános Közműszolgáltató Rt. (továbbiakban: ÁKR.) egy új lakópark építése kapcsán Téged kért fel, hogy meghatározott riportokkal, elemzésekkel és KPI-okkal támogasd a beruházáshoz kapcsolódó infrastruktúra-bővítés tervezését.

Rendelkezésedre áll a jelenleg is üzemszerűen működő (vagy legalábbis annak hitt) Oracle 19c adattárház, melyben ügyfél és fogyasztási adatokat találsz. Ezek alapján már biztos lehetsz benne, hogy az adattárház nem lesz képes a vezetőség minden kérdésére választ szolgáltatni, kénytelen leszel bevetni a machine learning-et.

**Tekintettel arra, hogy egy választ sem rögzítettél az alábbi feladatlapon, ebben a fordulóban a kitöltésére rendelkezésre álló idő teljes egésze, azaz 60 perc került rögzítésre mint megoldáshoz felhasznált idő.**

## 1. feladat 0 / 5 pont

Az igazolványadatokat tartalmazó táblában van egy ügyfél, akinek 2013.06.30-án látszólag két aktív jogosítványa is van eltérő számmal. Add meg a hibás rekord egyedi mesterséges kulcsának értékét!

**A megoldás:**
5691

### Magyarázat a megoldáshoz

Az alábbi scriptek alapján jól látszik, hogy az adattisztításból visszamaradt rekord:

```
select count(*), customer_id, doc_type

from co_id_document t

where date'2013-06-30' between valid_from and valid_to

and t.active_flag = 'A'

group by t.customer_id, t.doc_type

having count(*) > 1

;

select t.*,to_char(valid_to,'YYYYMMDD')

from co_id_document t

where t.customer_id = '9901140720'

and   t.doc_type = 'L'

;

select t.*,to_char(valid_to,'YYYYMMDD')

from co_id_document t

where t.unid = '5691'


;
```

## 2. feladat 0 / 5 pont

Az ügyféladatokat tartalmazó táblában van egy ügyfél, aki látszólag abban az időszakban nem születhetett még meg, de az adott időszakra volt aktív fogyasztása. A hiba egy 2013.08.30-ra vonatkozó riportban került elő. Add meg a hibás rekord egyedi mesterséges kulcsának értékét (ügyfélrekordra vonatkozóan)!

**A megoldás:**
1434

### Magyarázat a megoldáshoz

```
select *

from co_customer t

inner join co_consumption c

on t.customer_id = c.customer_id

and c.effective_month = '201308'

where date'2013-08-30' between t.valid_from and t.valid_to

and t.active_flag = 'A'

and t.birth_date>date'2013-08-30'


;
```

## 3. feladat 0 / 5 pont

Az ügyfelek igazolványaira vonatkozó táblában van két ügyfél, akiknek látszólag azonos a személyigazolvány számuk. A hiba egy 2013.06.30-ra vonatkozó riportban került elő. Add meg a kérdéses igazolványszámot!

**A megoldás:**
777579QQ

### Magyarázat a megoldáshoz

```
select count(*), t.doc_number

from co_id_document t

where 1=1

and date'2013-06-30' between valid_from and valid_to

and t.active_flag = 'A'

and t.doc_type = 'I'

group by t.doc_number

having count(*) > 1


;
```

## 4. feladat 0 / 5 pont

Az ingatlanokat tartalmazó táblában van egy rekord, amit látszólag semmilyen ügyfélhez sem lehet kötni, időszaktól függetlenül. Add meg a hibás rekord egyedi mesterséges kulcsának értékét!

**A megoldás:**
1290

### Magyarázat a megoldáshoz

```
select *

from co_property p

left join co_customer t

on t.customer_id = p.customer_id

where 1=1

and t.customer_id is null


;
```

Vissza a kategóriáimhoz

**Adattárház elemző (BCS)**
3. forduló

## Ismertető a feladathoz

Emlékeztetőül: A feladat megkezdése előtt a zip-ben mellékelt Installation Guide szerint kérjük, hogy telepítsd azokat az eszközöket, amelyekre szükséged lesz a fordulók során! **Ahhoz, hogy az adatokat be lehessen tölteni, a szintén zip-ben mellékelt beállításokat (Preferences.png) szükséges eszközölni SQL Developer-ben**. Ezután tudod futtatni a szintén a zip fájl részeként mellékelt scripteket. A mellékelt adatmodell leírást szintén ne felejtsd el átnézni!

A feladatlap megoldására a maximális időt állítottuk be, de természetesen ennél lényegesen kevesebb idő alatt is meg lehet válaszolni a kérdéseket.

Kérjük, tartsd szem előtt, hogy a kategória az ad-hoc riportokra vonatkozik!

Vállalatunk, a képzeletbeli Általános Közműszolgáltató Rt. (továbbiakban: ÁKR.) egy új lakópark építése kapcsán Téged kért fel, hogy meghatározott riportokkal, elemzésekkel és KPI-okkal támogasd a beruházáshoz kapcsolódó infrastruktúra-bővítés tervezését.

Rendelkezésedre áll a jelenleg is üzemszerűen működő (vagy legalábbis annak hitt) Oracle 19c adattárház, melyben ügyfél és fogyasztási adatokat találsz. Ezek alapján már biztos lehetsz benne, hogy az adattárház nem lesz képes a vezetőség minden kérdésére választ szolgáltatni, kénytelen leszel bevetni a machine learning-et.

**Tekintettel arra, hogy egy választ sem rögzítettél az alábbi feladatlapon, ebben a fordulóban a kitöltésére rendelkezésre álló idő teljes egésze, azaz 60 perc került rögzítésre mint megoldáshoz felhasznált idő.**

## 1. feladat 0 / 5 pont

Tesztelni szeretnék az adattárház teljességét. Bekérték a 2013-as első két negyedévének összfogyasztását az aktív ügyfelek körében ESOMAR státusz bontásban egész számra kerekítve. A megoldást az alábbi formában várják a legmagasabb fogyasztási értéktől a legalacsonyabbig: A:2800,B:1200,C1:1000,C2:900

**A megoldások:**
C1:2917,C2:2856,D:2317,B:1994,A:1338,E:1194
C2:32619,C1:27144,B:20778,D:19653,E:13465,A:12569
C2:32619,C1:27144,B:20778, D:19653,E:13465,A:12569

### Magyarázat a megoldáshoz

```
"select round(sum(c.amount),0) as fogyasztas

     ,a.esomar_status

from consumption c

inner join co_customer a

on a.customer_id = c.customer_id

and c.effective_month between to_char(a.valid_from,'YYYYMM') and
to_char(a.valid_to,'YYYYMM')

and a.active_flag = 'A'

where 1=1

and c.effective_month >= 201301

and c.effective_month <= 201306

group by a.esomar_status

order by round(sum(c.amount),0) desc

;"
```

## 2. feladat 0 / 5 pont

2013. június hó végére vonatkozóan tettek fel kérdést az aktív ügyfelek köréről, mégpedig azt szeretnék látni, hogy milyen arányban vannak azok a ügyfelek, akiknek az ingatlanjukhoz tartozik napelem az összes ügyfélhez képest. Nem releváns, hogy volt-e az adott ügyfeleknek az adott időszakban mért fogyasztása. Hisztorikus adat sajnos nem áll rendelkezésre a lakhellyel kapcsolatos leíró információkról, de állítólag időszaktól függetlenül helyes adatokat tartalmaz a CO_PROPERTY tábla. A riport alapját nem képzik azok az ügyfelek, ahol az ingatlanról nem áll rendelkezésre adat.

A választ százalékos formában várják 2 tizedesjegyre kerekítve, az alábbi szerint: 11,11

**A megoldások:**
27,92
28,33

### Magyarázat a megoldáshoz

```
select round(sum(p.solar_panels)/count(*)*100,2) arany --27,92
     ,sum(p.solar_panels) as napelemmel_birok --86
     ,count(*) as osszes_fogyaszto --308
from co_customer a
inner join co_property p
on p.CUSTOMER_ID = a.customer_id
where 1=1
and date'2013-06-30' between a.valid_from and a.valid_to --érvényes a rekord
and a.active_flag = 'A' --aktív
;"
```

## 3. feladat 0 / 5 pont

Kértek egy riportot 2013 első két negyedévére vonatkozóan. Az aktív, fogyasztási méréssel rendelkező ügyfelek körében szeretnék vizsgálni, hogy mekkora az egy négyzetméterre jutó fogyasztás, ezen belül is megbontva az alapján, hogy az ingatlan rendelkezik-e napelemmel. A két esetből képzett négyzetméterre vonatkozó szám különbségére van szükség.

A megoldást az alábbi formában várják abszolút értékben, két tizedesjegyre kerekítve: 11,11

**A megoldások:**
0,01
00,01
0,11

### Magyarázat a megoldáshoz

```
"select round(sum(c.amount)/sum(p.squaremeter),4) as fogy_per_nm
,p.solar_panels
from co_consumption c
inner join co_customer a
on a.customer_id = c.customer_id
and c.effective_month between to_char(a.valid_from,'YYYYMM') and
to_char(a.valid_to,'YYYYMM')
and a.active_flag = 'A'
inner join co_property p
on p.CUSTOMER_ID = c.customer_id
where 1=1
and c.effective_month >= 201301
and c.effective_month <= 201306
group by p.solar_panels
;
/*
fogy/nm napelem
0,4143    1
0,4259    0
*/
select abs(round(0.4143-0.4259,2)) from dual
;
"
```

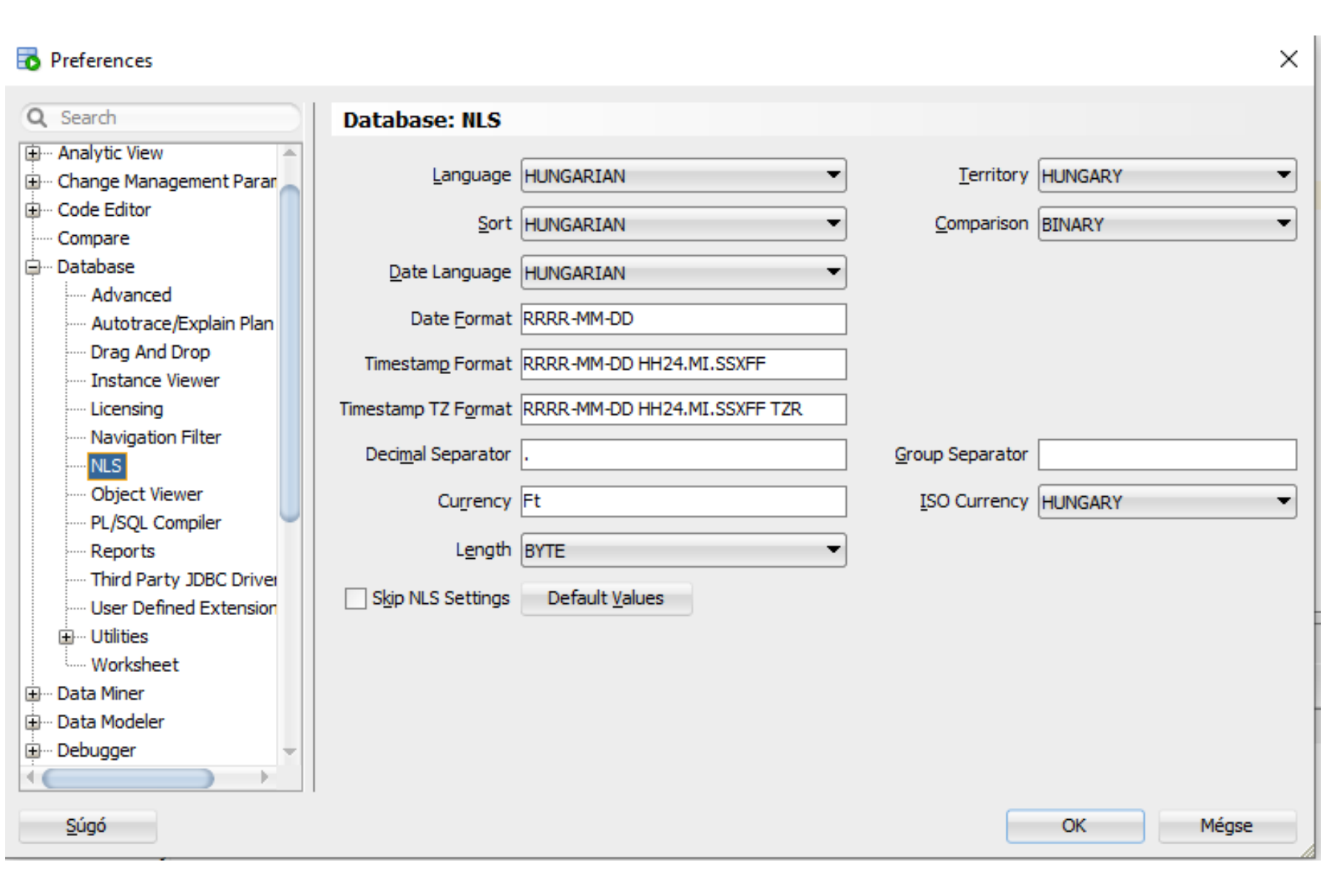
Vissza a kategóriáimhoz

**Adattárház elemző (BCS)**
4. forduló

## Ismertető a feladathoz

A korábbi hibakereséseidnek köszönhetően létrehozott a BI osztály egy tisztított részletet az adattárházból. A részlet csak olyan rekordokat tartalmaz, ami a projekt céljainak eléréséhez releváns. Ezentúl a CO előtagú táblák helyett a REPCO előtagú táblákat kell majd használnod, hogy a riportigényeket ki tudd elégíteni.

**Az új táblákat insert script formájában kapod (ld. Mellékletek), amihez az alábbi beállításokat kell eszközölnöd:**



Database: NLS

Language: HUNGARIAN
Territory: HUNGARY
Sort: HUNGARIAN
Comparison: BINARY
Date Language: HUNGARIAN
Date Format: RRRR-MM-DD
Timestamp Format: RRRR-MM-DD HH24.MI.SSXFF
Timestamp TZ Format: RRRR-MM-DD HH24.MI.SSXFF TZR
Decimal Separator: .
Group Separator:
Currency: Ft
ISO Currency: HUNGARY
Length: BYTE
☐ Skip NLS Settings

**Tekintettel arra, hogy egy választ sem rögzítettél az alábbi feladatlapon, ebben a fordulóban a kitöltésére rendelkezésre álló idő teljes egésze, azaz 60 perc került rögzítésre mint megoldáshoz felhasznált idő.**

## 1. feladat 0 / 5 pont

Összesen mennyi ügyfélszolgálati megkeresés történt átlagosan havonta a 2012-es év kezdetétől a 2018-as év végéig? Csak azok a megkeresések képezik a riport alapját, ahol létezik az időpontra vonatkozóan érvényes ügyféladat!

**A számot két tizedesjegyre kerekítve kérik, az alábbi formában: 11,11**

**A megoldások:**
530,69
134,88

### Magyarázat a megoldáshoz

```
with darabszam as --530,69

(

select COUNT(*) as kontaktus_szam

from repco_customer_service t

inner join repco_customer c

on t.customer_id = c.customer_id

and t.effective_time between c.valid_from and valid_to

where to_char(t.effective_time,'YYYYMM') between '201201' and '201812'

),

honapok_szama as

(

select count(distinct to_char(t.effective_time,'YYYYMM')) honap_db

from repco_customer_service t

where to_char(t.effective_time,'YYYYMM') between '201201' and '201812'

)

select round(kontaktus_szam / honap_db,2) as atlag

from darabszam

full join honapok_szama

on 1=1

;
```

## 2. feladat 0 / 5 pont

Az ügyfélszolgálat szeretné megtudni, hogy melyik kontaktus típus volt a legnépszerűbb a 2016-os év során nem szerinti bontásban. Csak azok a megkeresések képezik a riport alapját, ahol létezik az időpontra vonatkozóan érvényes ügyféladat!

**A megoldást az alábbi formában várják F:X,M:X, ahol az F a nőket, az M a férfiakat jelenti, az X a kontaktus típus kódját jelöli.**

**A megoldások:**
F:F,M:F
F:F ,M:F
M:F,F:F

### Magyarázat a megoldáshoz

```
select COUNT(*) as kontaktus_szam --F:F,M:F

, c.sex

, t.type_of_contact

from repco_customer_service t

inner join repco_customer c

on t.customer_id = c.customer_id

and t.effective_time between c.valid_from and valid_to

where to_char(t.effective_time,'YYYYMM') between '201601' and '201612'

group by c.sex

, t.type_of_contact

order by sex, COUNT(*) desc
```

## 3. feladat 0 / 5 pont

"Az ügyfélszolgálat szeretné visszamérni, hogy nagyságrendileg mennyi munkaóra ráfordítás lehetett az ügyfelek kiszolgálása a 2014-es év kezdetétől a 2016-as év végéig. Ehhez megadták, hogy kontaktus típusonként mi a várt átlagos ideje egy ügyfél kiszolgálásnak, ami a következő:

M - Levél - 30 perc

E - Email - 12 perc

C - Chatablak - 8 perc

P - Telefonos hívás - 15 perc

F - Facebook - 9 perc

Csak azok a megkeresések képezik a riport alapját, ahol létezik az időpontra vonatkozóan érvényes ügyféladat!

**A megoldást az alábbi formában várják két tizedjegyre kerekítve: 11,11**

**A megoldások:**
5163,65
1189,23

### Magyarázat a megoldáshoz

```
select
round(sum(decode(t.type_of_contact,'M',30,'E',12,'C',8,'P',15,'F',9,0))/60,2) -
-5163,65

from repco_customer_service t

inner join repco_customer c

on t.customer_id = c.customer_id

and t.effective_time between c.valid_from and valid_to

where to_char(t.effective_time,'YYYYMM') between '201401' and '201612'

;
```

## 4. feladat 0 / 5 pont

Hiba történt a 2016-os elszámolásban, egy meghirdetett kedvezményt nem vittek fel a rendszerb, és ezt most utólag szeretnék megtéríteni az ügyfeleknek.

A jogosultak ismérvei a következők:

- 2 vagy több gyermekük van, de nem feltétlenül aktív ügyfelek,
- mért fogyasztással rendelkeznek 2016-ban,
- nem rendelkeznek napelemmel (vagy nincs róla információnk).

Az alábbi kérdésekre várják a választ rendre a lenti formában:

Hány ügyfélt érint a visszatérítés?

Mekkora összeget kell nekik kiutalni, ha fogyasztási egységenként 12 pénz jár vissza?

**A választ az alábbi formában várják, a pénzösszeget két tizedesre kerekítve: CUST:11,MON:11,11**

**A megoldások:**
CUST:302,MON:2086806,32
CUST:302 , MON:2086806,32
CUST:302,MON:2066892

### Magyarázat a megoldáshoz

```
select count(distinct c.customer_id) as CUST --CUST:302,MON:4173612,64

,round(sum(c.amount*12),2) as MON

from repco_consumption c

inner join repco_customer a

on a.customer_id = c.customer_id

and c.effective_month between to_char(a.valid_from,'YYYYMM') and
to_char(a.valid_to,'YYYYMM')

and a.num_of_children >=2

left join repco_property p

on p.customer_id = a.customer_id

where substr(c.effective_month,1,4) = '2016'

and nvl(p.solar_panels,0)<> 1
```

## 5. feladat 0 / 5 pont

Felhív egy barátod az ügyfélszolgálatról, hogy kellene egy kis segítség. Sajnos kitörölt egy e-mailt, és nem emlékszik pontosan, hogy ki ez az illető, vagy hogy mi az e-mail címe.

Az alábbiakra emlékszik:

- Az email címben a kukac és a pont között vagy c vagy j betű volt, vagy talán mind a kettő.
- Az email címben a kukac előtt volt c és s betű is, egymás után, de nem biztos hogy követték egymást.
- Az emailben említette a pasas, hogy van autója, mert azzal viszi a gyerekeit suliba.
- Tippre a stílus alapján középiskolai végzettségűnek ítélné az illetőt.

Azt ígérte, hogy kapsz egy csokit, ha megírod az e-mail címet még ma, mert holnap már tárgytalan lesz, hiszen reggelre bekerül a customer_service táblába.

**Mi ez az e-mail cím?**

**A megoldások:**
francis9@cde.com
francis9@

### Magyarázat a megoldáshoz

```
select * --francis9@cde.com

from repco_customer c

where 1= 1



and (


 lower(substr(c.email_address,instr(c.email_address,'@')+1,instr(substr(c.email_address,instr(c.email_address,'@')+2),'.')))
like '%c%'

     or


 lower(substr(c.email_address,instr(c.email_address,'@')+1,instr(substr(c.email_address,instr(c.email_address,'@')+2),'.')))
like '%j%'

     )

and substr(c.email_address,1,instr(c.email_address,'@')-1) like '%c%s%'



and num_of_cars >0

and education like 'High%'

and num_of_children > 1

and sex = 'M'"
```

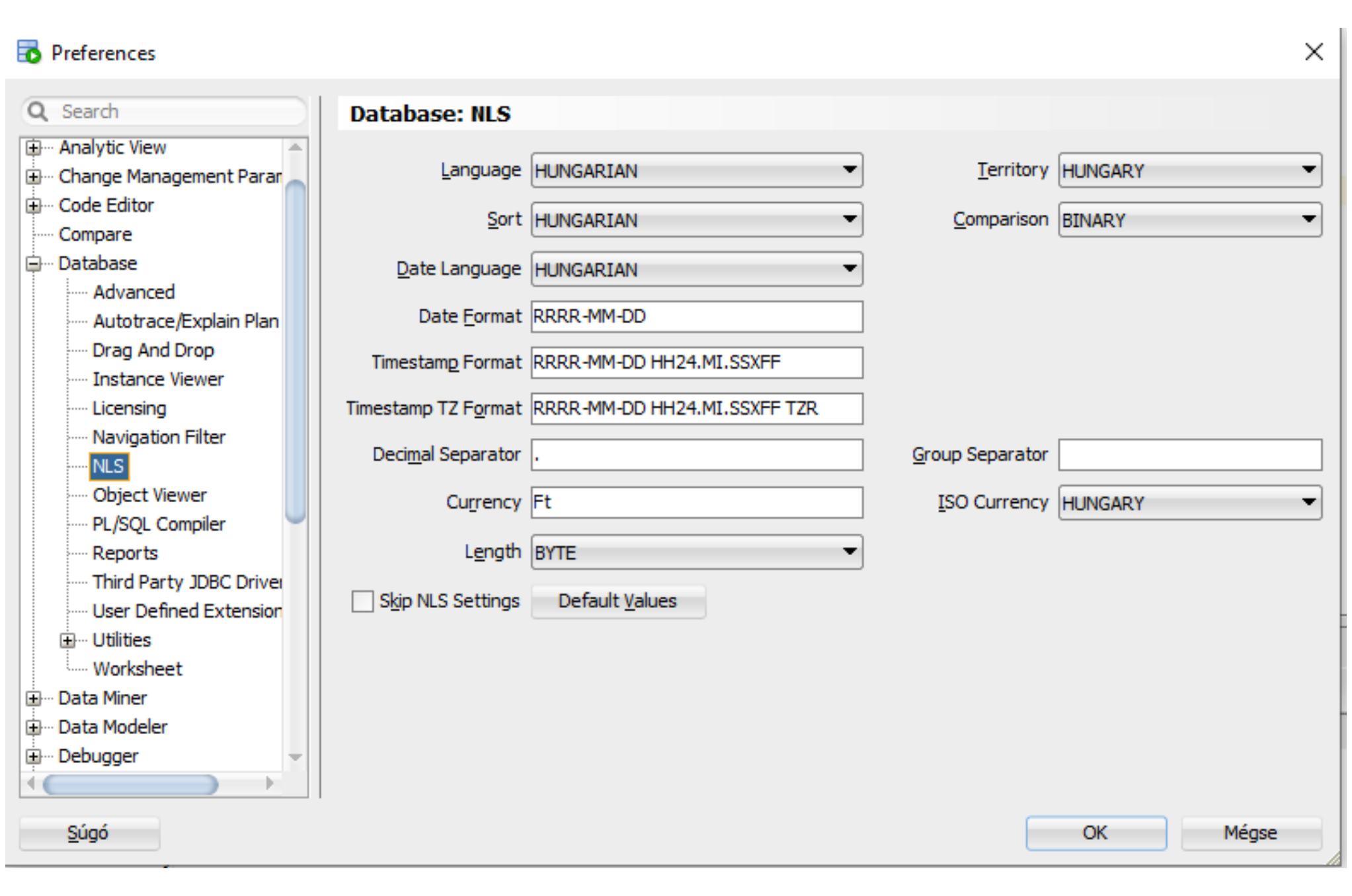
Vissza a kategóriáimhoz

**Adattárház elemző (BCS)**
5. forduló

## Ismertető a feladathoz

Az új lakópark elektromos hálózatának tervezése során hasznos lehet, ha előre tudjuk jelezni, hogy az ingatlanokba beköltöző új lakók mennyi **elektromos áramot** fognak fogyasztani.
Ennek prediktálására machine learning-et, lineáris regressziót fogunk alkalmazni. Ehhez viszont először adatelőkészítő lépésekre és adatfeltáró elemzésekre lesz szükség. A következő kérdések az ilyen irányú ismeretekre támaszkodnak.

A forduló 6. feladatának megoldásához szükség lesz a 4. fordulós insert scriptekre (ld. Mellékletek).



Database: NLS

Language: HUNGARIAN
Sort: HUNGARIAN
Date Language: HUNGARIAN
Date Format: RRRR-MM-DD
Timestamp Format: RRRR-MM-DD HH24.MI.SSXFF
Timestamp TZ Format: RRRR-MM-DD HH24.MI.SSXFF TZR
Decimal Separator: .
Currency: Ft
Length: BYTE
Territory: HUNGARY
Comparison: BINARY
Group Separator:
ISO Currency: HUNGARY
☐ Skip NLS Settings

**Tekintettel arra, hogy egy választ sem rögzítettél az alábbi feladatlapon, ebben a fordulóban a kitöltésére rendelkezésre álló idő teljes egésze, azaz 60 perc került rögzítésre mint megoldáshoz felhasznált idő.**

## 1. feladat 0 / 5 pont

Az alábbi válaszlehetőségek közül jelöld be azokat, amelyek egyváltozós elemzések!

☐ Korrelációs vizsgálat (Pearson-féle korrelációs együttható)

☐ One-way ANOVA

☑ Minimum

☑ Szórás (Standard Deviation)

☐ Kereszttáblás elemzés (Cross-Table Analysis)

### Magyarázat a megoldáshoz

A Pearson-féle korrelációs együttható olyan mérőszám, amely az erősségét és az irányát mutatja meg egy lineáris kapcsolat két változója között.
(https://hu.wikipedia.org/wiki/Korrel%C3%A1ci%C3%B3s_egy%C3%BCtthat%C3%B3#Pearson)
Nominális vagy ordinális változókat intervallum szintűekkel vet össze.
(https://hu.wikipedia.org/wiki/Varianciaanal%C3%ADzis)
Két vagy több paraméter közti összefüggést (befolyásoltságot) vizsgál.
(http://eit.bme.hu/sites/default/files/Oktatas/2014-2015/Berke_David_-_D_Eloadas_Terinfo.pdf)

## 2. feladat 0 / 5 pont

Az alábbi válaszlehetőségek közül jelöld be azokat, amelyek igazak a korrelációs vizsgálatra (Pearson-féle korrelációs együtthatóról van szó)!

☐ Egy változós elemzés.

☑ A lineáris kapcsolatot jelzi.

☑ A korrelációs együtthatónak van előjele.

☑ A korrelációs együttható abszolút értékben egy 0-100 közötti szám.

☐ Csak kategorikus és numerikus változók között tudjuk kiszámolni.

### Magyarázat a megoldáshoz

A Pearson-féle korrelációs együttható olyan mérőszám, amely az erősségét és az irányát mutatja meg egy lineáris kapcsolat két változója között.
(https://hu.wikipedia.org/wiki/Korrel%C3%A1ci%C3%B3s_egy%C3%BCtthat%C3%B3#Pearson)
A Pearson-féle korrelációs együttható olyan mérőszám, amely az erősségét és az irányát mutatja meg egy lineáris kapcsolat két változója között.
(https://hu.wikipedia.org/wiki/Korrel%C3%A1ci%C3%B3s_egy%C3%BCtthat%C3%B3#Pearson)
A Pearson-féle korrelációs együttható olyan mérőszám, amely az erősségét és az irányát mutatja meg egy lineáris kapcsolat két változója között.
(https://hu.wikipedia.org/wiki/Korrel%C3%A1ci%C3%B3s_egy%C3%BCtthat%C3%B3#Pearson)
Különböző típusú korrelációs együtthatók léteznek. Mindegyik −1 és +1 közötti értéket vehet fel, ahol ± 1 a lehető legerősebb egyezést és 0 a lehető legnagyobb eltérést jelzi.
(https://hu.wikipedia.org/wiki/Korrel%C3%A1ci%C3%B3s_egy%C3%BCtthat%C3%B3#Pearson)
Gyakran elõfordul, hogy két változó mennyiség közötti kapcsolatot vizsgálunk. A kapcsolat szorosságát célszerû egy mérõszámmal jellemezni. Nagyon sok ilyen mérõszám létezik, ezek közül a legelterjedtebb az ún. korrelációs együttható, vagy Pearson-féle korrelációs együttható.
(http://rs1.szif.hu/~szorenyi/elm/bioselm7.htm)

## 3. feladat 0 / 5 pont

Az alábbi válaszlehetőségek közül jelöld be azokat, amelyek igazak a kvantilisekre!

☐ A medián az a kvantilis, ami 3 egyenlő részre osztja a rendezett adat sokaságot.

☑ Az 50-edik percentilis 2 egyenlő részre osztja a rendezett adat sokaságot.

☑ Használhatóak a numerikus változók egyenlő elemszámú binekre bontására.

☐ A következőek mind kvantilisek: tercilisek, kvartilisek, kvintilisek, setilisek, decilisek, percentilisek.

☐ Kategorikus és numerikus változóknak is léteznek kvantilisei.

### Magyarázat a megoldáshoz

A medián az az érték, amely a sorba rendezett adatokat két egyenlő részre osztja. (https://hu.wikipedia.org/wiki/Medi%C3%A1n)
Percentilis: n-edik percentilis a változó azon kategóriája, amely az összes érték éppen n százalékánál nagyobb. Például a medián az 50. percentilis.
(https://hu.wikipedia.org/wiki/Le%C3%ADr%C3%B3_statisztika)
(https://hu.wikipedia.org/wiki/Kvantilisek)

## 4. feladat 0 / 5 pont

Az alábbi válaszlehetőségek közül jelöld be azokat, amelyek igazak a hisztogramokra!

☑ Egy változó értékek szerinti eloszlását mutatja.

☐ Változók közötti kapcsolatot mutatja.

☐ Csak kategórikus változókra működik.

☑ Az adat feltáró elemzések részét képezheti.

### Magyarázat a megoldáshoz

Histograms are used to show distributions of variables.
(https://www.forbes.com/sites/naomirobbins/2012/01/04/a-histogram-is-not-a-bar-chart/?sh=412832446d77)
Histograms are used to show distributions of variables while bar charts are used to compare variables.
(https://www.forbes.com/sites/naomirobbins/2012/01/04/a-histogram-is-not-a-bar-chart/?sh=412832446d77)
Histograms plot quantitative data with ranges of the data grouped into bins or intervals while bar charts plot categorical data.
(https://www.forbes.com/sites/naomirobbins/2012/01/04/a-histogram-is-not-a-bar-chart/?sh=412832446d77)

## 5. feladat 0 / 5 pont

Az alábbi válaszlehetőségek közül jelöld be azokat, amelyek igazak a z-score-ra!

☑ Standard Normális eloszlású.

☐ Várható értéke 1.

☑ Szórása 1.

☑ Standard-score néven is használják.

☑ Értéke lehet akár -3 is.

### Magyarázat a megoldáshoz

Contrary to what many people believe, z-scores are not necessarily normally distributed. (https://www.spss-tutorials.com/z-scores-what-and-why/)
The mean of the z-scores is always 0.
(https://www.uth.tmc.edu/uth_orgs/educ_dev/oser/L1_6.HTM)
The standard deviation of the z-scores is always 1.
(https://www.uth.tmc.edu/uth_orgs/educ_dev/oser/L1_6.HTM)
Simply put, a z-score (also called a standard score) gives you an idea of how far from the mean a data point is. (https://www.statisticshowto.com/probability-and-statistics/z-score/)
Z-scores range from -3 standard deviations (which would fall to the far left of the normal distribution curve) up to +3 standard deviations (which would fall to the far right of the normal distribution curve).
(https://www.statisticshowto.com/probability-and-statistics/z-score/)

## 6. feladat 0 / 10 pont

Az új lakópark építése előtt szeretnénk megvizsgálni, hogy:

**az ingatlanok melyik paramétere (SQUAREMETER, PERFORMANCE_OF_SOLAR_PANELS, THERMAL_INSULATION_THICKNESS) van a leginkább hatással az elektromos áram átlagfogyasztására?**

Ehhez korrelációs vizsgálatot (PEARSON) alkalmazunk.

**Add meg az elektromos áram átlagfogyasztására leginkább hatással lévő paramétert annak értékével együtt a következő alakban: PARAM:+/-0,1111** (szóközök nélkül)

**A megoldások:**
SQUAREMETER:+0,6752
SQUAREMETER:+0,7176
SQUAREMETER:+0,7177
SQUAREMETER:0,7176
SQUAREMETER:0,6752

### Magyarázat a megoldáshoz

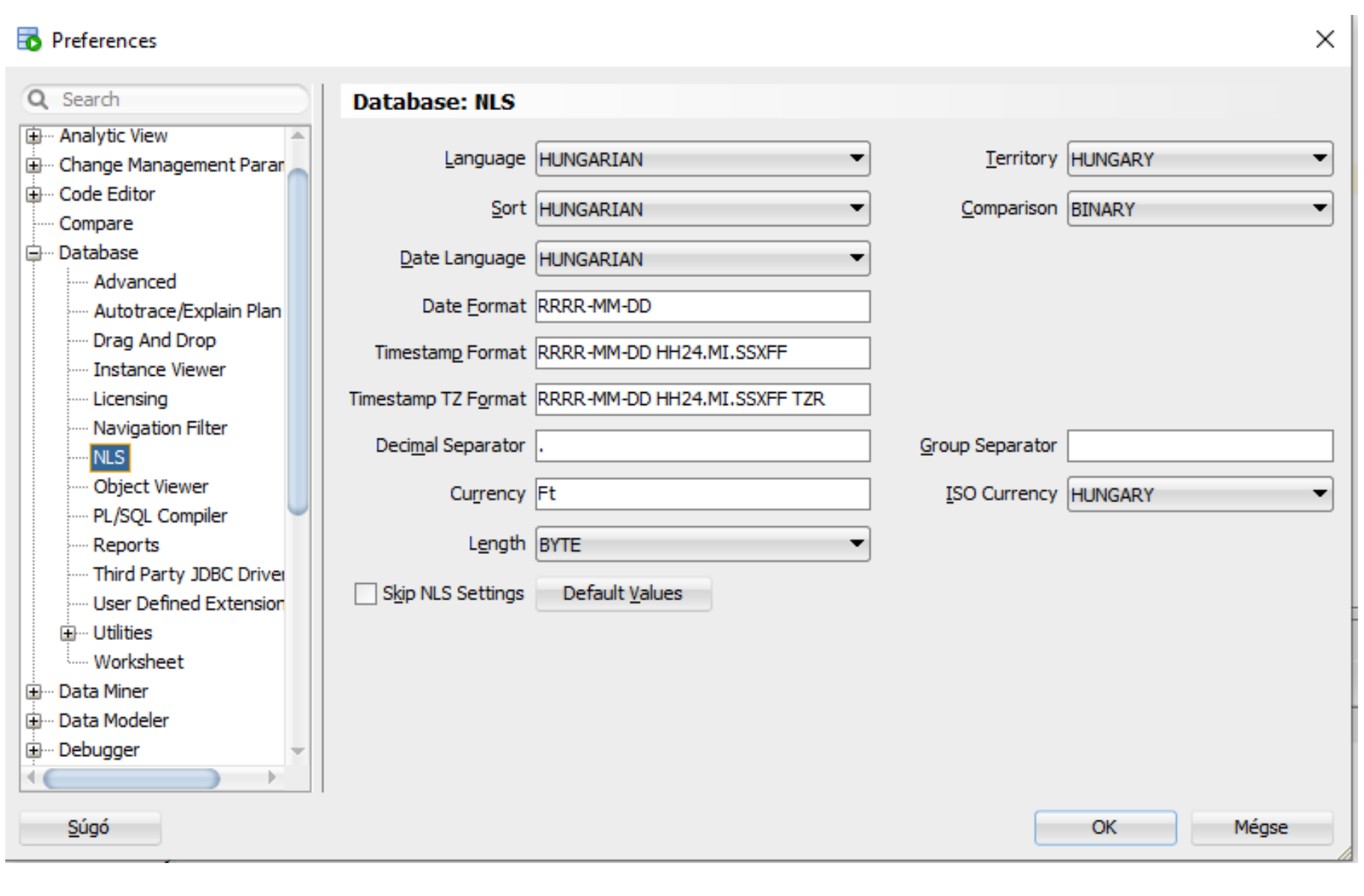
```
select
corr(d.atlag,e.squaremeter),corr(d.atlag,e.performance_of_solar_panels),corr(d.atlag,e.thermal_insulation_thickness)
from
(select avg(c.amount) atlag, c.customer_id
from repco_consumption c
inner join repco_customer cu
on c.customer_id = cu.customer_id
and c.effective_month between to_char(cu.valid_from,'YYYYMM') and
to_char(cu.valid_to,'YYYYMM')
and cu.active_flag= 'A'
group by c.customer_id) d
join repco_property e
on d.customer_id = e.customer_id;
```

Vissza a kategóriáimhoz

**Adattárház elemző (BCS)**
6. forduló

## Ismertető a feladathoz

Az előző fordulóban említettek szerint később regressziós algoritmus segítségével fogjuk előre jelezni a fogyasztási adatokat. A következő kérdések a regressziós modellek ismeretére támaszkodnak.



Database: NLS

Language: HUNGARIAN
Territory: HUNGARY
Sort: HUNGARIAN
Comparison: BINARY
Date Language: HUNGARIAN
Date Format: RRRR-MM-DD
Timestamp Format: RRRR-MM-DD HH24.MI.SSXFF
Timestamp TZ Format: RRRR-MM-DD HH24.MI.SSXFF TZR
Decimal Separator: .
Group Separator:
Currency: Ft
ISO Currency: HUNGARY
Length: BYTE
☐ Skip NLS Settings
Default Values

**Tekintettel arra, hogy egy választ sem rögzítettél az alábbi feladatlapon, ebben a fordulóban a kitöltésére rendelkezésre álló idő teljes egésze, azaz 60 perc került rögzítésre mint megoldáshoz felhasznált idő.**

## 1. feladat 0 / 10 pont

**Szeretnénk kideríteni, hogy mi okozhatta a legszélsőségesebb elektromos áram fogyasztási értékeket.**

Ehhez megnézzük azt, hogy a legmagasabb, illetve a legalacsonyabb átlag fogyasztási értéket produkáló ügyfél (csak az aktív ügyfelek számítanak) tulajdonában lévő ingatlanok méretei (SQUAREMETER) kimagaslóan nagyok vagy kicsik voltak-e (azaz a méretük kisebb-e a SQUAREMETER oszlop 10. percentilisénél vagy nagyobb-e 90. percentilisénél).

**Írjuk ki** (az alábbi formában) **a legmagasabb és legalacsonyabb átlagfogyasztású ügyfél közül annak vagy azoknak a CUSTOMER_ID-ját és a tulajdonában lévő ingatlan méretét (SQUAREMETER), amelyre vagy amelyekre igaz az előbbi feltevés!**

CUSTOMER_ID:1111111111,SQUAREMETER:111

**A megoldások:**
CUSTOMER_ID:9431752379,SQUAREMETER:190
CUSTOMER_ID:8551090261,SQUAREMETER:63
CUSTOMER_ID:9431752379,SQUAREMETER:190,CUSTOMER_ID:8551090261,SQUAREMETER:63
CUSTOMER_ID:9431752379,SQUAREMETER:190,8551090261:63
CUSTOMER_ID:9431752379,8551090261,SQUAREMETER:190,63
CUSTOMER_ID:9431752379,8551090261,SQUAREMETER:190,63
CUSTOMER_ID:9431752379,SQUAREMETER:190
CUSTOMER_ID:8551090261,SQUAREMETER:63
CUSTOMER_ID:9431752379,SQUAREMETER:190
CUSTOMER_ID:8551090261,SQUAREMETER:63
CUSTOMER_ID:9431752379,SQUAREMETER:190;CUSTOMER_ID:8551090261,SQUAREMETER:63
CUSTOMER_ID:8551090261,SQUAREMETER:63,CUSTOMER_ID:9431752379,SQUAREMETER:190
CUSTOMER_ID:8551090261,SQUAREMETER:63
CUSTOMER_ID:9431752379,SQUAREMETER:190
CUSTOMER_ID:8551090261,SQUAREMETER:063
CUSTOMER_ID:9431752379,SQUAREMETER:190
CUSTOMER_ID:9431752379,SQUAREMETER:190 ;
CUSTOMER_ID:8551090261,SQUAREMETER:63

### Magyarázat a megoldáshoz

```
select a.customer_id,b.squaremeter
from (
with atlag as
(
select avg(c.amount) atlag, c.customer_id
from repco_consumption c
inner join repco_customer cu
on c.customer_id = cu.customer_id
and c.effective_month between to_char(cu.valid_from,'YYYYMM') and
to_char(cu.valid_to,'YYYYMM')
and cu.active_flag= 'A'
group by c.customer_id
)
select * from atlag
where atlag in (select max(atlag) from atlag)
union all
select * from atlag
where atlag in (select min(atlag) from atlag)) a
join repco_property b
on a.customer_id = b.customer_id
where b.squaremeter <= (SELECT PERCENTILE_CONT(0.1) WITHIN GROUP
(ORDER BY squaremeter  ASC)
FROM repco_property) or
   b.squaremeter >= (SELECT PERCENTILE_CONT(0.9) WITHIN GROUP (ORDER
BY squaremeter  ASC)
FROM repco_property)
;
```

## 2. feladat 0 / 5 pont

Az alábbi válaszlehetőségek közül jelöld be azokat, amelyek igazak az outlier adatpontokra!

☐ Reprezentatív adatpontok, az egész adatra nézve.
☑ Okozhatja mérési vagy adminisztrációs hiba is.
☑ Lehet valós adatpont is.
☑ Az adat természetes varianciájából is fakadhat.
☐ Az átlaghoz közeli adatpontok.

### Magyarázat a megoldáshoz

Reprezentatív adatpontok, az egész adatra nézve. Helytelen https://hu.wikipedia.org/wiki/Kiugr%C3%B3_%C3%A9rt%C3%A9k
Okozhatja mérési vagy adminisztrációs hiba is. Helyes https://hu.wikipedia.org/wiki/Kiugr%C3%B3_%C3%A9rt%C3%A9k
Lehet valós adatpont is. Helyes https://hu.wikipedia.org/wiki/Kiugr%C3%B3_%C3%A9rt%C3%A9k
Az adat természetes varianciájából is fakadhat. Helyes https://hu.wikipedia.org/wiki/Kiugr%C3%B3_%C3%A9rt%C3%A9k
Az átlaghoz közeli adatpontok. Helytelen https://hu.wikipedia.org/wiki/Kiugr%C3%B3_%C3%A9rt%C3%A9k

## 3. feladat 0 / 5 pont

Az alábbi válaszlehetőségek közül jelöld be azokat, amelyek a regressziós modellekre igazak!

☐ Unsupervised learning
☑ Supervised learning
☑ Célváltozó diszkrét
☑ Prediktív analízis
☑ Célváltozó bináris

### Magyarázat a megoldáshoz

Unsupervised learning Helytelen https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/
Supervised learning Helyes https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/
Célváltozó diszkrét Helytelen A regression problem is when the output variable is a real or continuous value, such as “salary” or “weight”. (https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/)
Prediktív analízis Helyes Regression models are the mainstay of predictive analytics. (https://en.wikipedia.org/wiki/Predictive_analytics)
Célváltozó bináris Helytelen A regression problem is when the output variable is a real or continuous value, such as “salary” or “weight”. (https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/)

## 4. feladat 0 / 5 pont

Az alábbi válaszlehetőségek közül jelöld be azokat az algoritmusokat, amelyek használhatók regressziós probléma megoldására!

☑ Linear regression
☑ Support Vector Machine (SVM)
☑ Decision Tree
☑ Logistic regression
☐ K-means

### Magyarázat a megoldáshoz

Linear regression Helyes (https://analyticsindiamag.com/top-6-regression-algorithms-used-data-mining-applications-industry/)
Support Vector Machine (SVM) Helyes Support Vector Machine (SVM) can be leveraged both for classification or regression challenges. (https://analyticsindiamag.com/top-6-regression-algorithms-used-data-mining-applications-industry/)
Decision Tree Helyes Decision trees in classification and regression are very similar, in that both work by constructing trees of yes/no nodes. However, while classification end nodes result in a single class value , regression trees end with a continuous value. (https://medium.com/analytics-vidhya/5-regression-algorithms-you-need-to-know-theory-implementation-37993382122d)
Logistic regression Helyes Logistic regression can be used both in classification and regression problems. (https://www.analyticssteps.com/blogs/how-does-linear-and-logistic-regression-work-machine-learning)
K-means Helytelen K-means clustering as the name itself suggests, is a clustering algorithm, with no pre determined labels defined ,like we had for Linear Regression model, thus called as an Unsupervised Learning algorithm. (https://towardsdatascience.com/k-means-clustering-implementation-2018-ac5cd1e51d0a)

## 5. feladat 0 / 5 pont

Az alábbi válaszlehetőségek közül jelöld be azokat, amelyek a lineáris regressziós modellekre igazak!

☑ Kategorikus változókon is tanítható.
☐ Többváltozós lineáris regresszió minden rekordhoz (egy rekord n db magyarázó változóból áll) egy n-1 dimenziós hipersík legközelebbi pontját rendeli.
☑ Supervised learning
☑ Machine learning kategóriába tartozik.
☐ Deep learning kategóriába tartozik.

### Magyarázat a megoldáshoz

Kategorikus változókon is tanítható. Helyes Of course you can include categorical variables in a linear regression model, just codify first those variables as dummy variables. (https://www.researchgate.net/post/In_a_linear_regression_model_can_i_use_few_categorical_variables_as_independent_variables)
Többváltozós lineáris regresszió minden rekordhoz (egy rekord n db magyarázó változóból áll) egy n-1 dimenziós hipersík legközelebbi pontját rendeli. Helytelen https://datacadamia.com/data_mining/multiple_regression
Supervised learning Helyes Linear Regression is a supervised machine learning algorithm. (https://ml-cheatsheet.readthedocs.io/en/latest/linear_regression.html)
Machine learning Helyes Linear Regression is a supervised machine learning algorithm. (https://ml-cheatsheet.readthedocs.io/en/latest/linear_regression.html)
Deep learning Helytelen Deep learning (also known as deep structured learning) is part of a broader family of machine learning methods based on artificial neural networks with representation learning. (https://en.wikipedia.org/wiki/Deep_learning)

## 6. feladat 0 / 5 pont

Az alábbi válaszlehetőségek közül jelöld be azokat, amelyek egy regressziós modell teljesítményének kiértékelésére szolgáló KPI-ok vagy performancia mutatók!

☑ Confusion matrix
☑ RMSE
☑ MAE
☐ Accuracy
☐ F1-score

### Magyarázat a megoldáshoz

Confusion matrix Helytelen https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b
RMSE Helyes https://becominghuman.ai/understand-regression-performance-metrics-bdb0e7fcc1b3
MAE Helyes https://becominghuman.ai/understand-regression-performance-metrics-bdb0e7fcc1b3
Accuracy Helytelen https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b
F1-score Helytelen https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b

Vissza a kategóriáimhoz

**Adattárház elemző (BCS)**
7. forduló

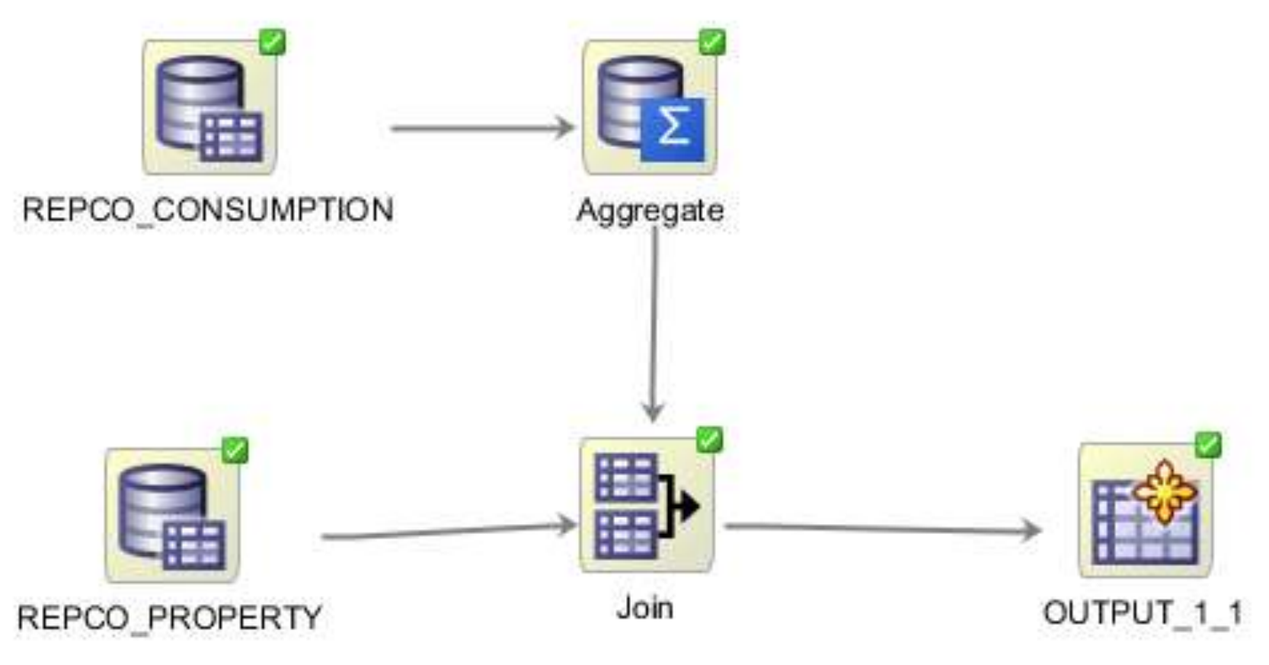
## Ismertető a feladathoz

Regressziós algoritmusok segítségével szeretnénk előrejelezni az ügyfelek áram fogyasztási adatait. Ehhez az **Oracle Data Miner**-t fogjuk használni, ami az Oracle SQL Developer egy bővítménye. Ez a korábban letöltött SQL Developer-ben megtalálható és a következő videó alapján konfigurálható: https://www.youtube.com/watch?v=tZCaQJsNVSs.

A Data Miner használatához további segítséget nyújt a következő lejátszási lista: https://www.youtube.com/playlist?list=PL99-DcFspRUq8VbbgXe2lQ559VDr7BSCr

A célváltozónk a **REPCO_CONSUMPTION** tábla AMOUNT változójából képzett ügyfélszintű átlagfogyasztás lesz, a magyarázó változóink pedig a **REPCO_PROPERTY** tábla változói lesznek (SQUAREMETER, PERFORMANCE_OF_SOLAR_PANELS, THERMAL_INSULATION_THICKNESS, SOLAR_PANELS, THERMAL_INSULATION). Ehhez először a REPCO_CONSUMPTION táblát kell aggregálni, majd a REPCO_PROPERTY táblával kell összekötni azt.

Ezen lépések végeztével egy az alábbi képen látható Workflow diagramhoz hasonlót kell kapni.



Ez után futtass regressziós algoritmusokat az adaton, ügyelj arra, hogy a **megfelelő cél- és magyarázó változókat használd** (a fent említett módon, ettől több változót ne használj a regresszió során). (Nem kell train és test setre bontani az adatot, ez a Regression Node-on belül megtörténik.)

**A megoldások beírásának megkönnyítésére a letölthető Word dokumentumban megtalálod a megoldások formátumát, be kell helyettesítened az értékeket, és utána be tudod másolni a válaszokat.**

**Tekintettel arra, hogy egy választ sem rögzítettél az alábbi feladatlapon, ebben a fordulóban a kitöltésére rendelkezésre álló idő teljes egésze, azaz 60 perc került rögzítésre mint megoldáshoz felhasznált idő.**

### 1. feladat 0 / 4 pont

Állítsd a test set nagyságát 30%-ra, és futtasd a regressziót! (Tipp: Case ID a CUSTOMER_ID lesz.)
Milyen algoritmusokat használ a Regression Node? Add meg az összes rendelkezésre álló algoritmus nevét a következő formátumban: Abc Xy, Abc Xy

**A megoldások:**
Generalized Linear Model, Support Vector Machine
REGR_GLM_1_1 Generalized Linear Model, REGR_SVM_1_1 Support Vector Machine
Generalized Linear Model,Support Vector Machine
Generalized Linear Models GLM, Support Vector Machines SVM
Generalized Linear Models, Support Vector Machine

**Magyarázat a megoldáshoz**

### 2. feladat 0 / 9 pont

Add meg az egyes algoritmusok MAE és RMSE értékeit, 4 tizedes jegyig, a következő formában:

Algorithm: Abc Xy, MAE: 11,1111, RMSE: 11,1111, Algorithm: Abc Xy, MAE: 11,1111, RMSE: 11,1111

**A megoldások:**
Algorithm: Generalized Linear Model, MAE: 3,5069, RMSE: 4,3386, Algorithm: Support Vector Machine, MAE: 3,5048, RMSE: 4,3472
Algorithm: Generalized Linear Model, MAE: 3,5069, RMSE: 4,3386, Algorithm: Support Vector Machine, MAE: 3,5052, RMSE: 4,3475

**Magyarázat a megoldáshoz**

### 3. feladat 0 / 9 pont

Állítsd a test set nagyságát 90%-ra, és nézd meg mi történik a Predictive Confidence-el! Hány százalékot változott a Predictive Confidence algoritmusonként külön-külön az előző 30%-os méretű test seten történt visszaméréshez képest?

(Tipp: ne százalékos változást számolj, a Perdiction Confidence értéke van %-ban)

Az értékeket a következő formában add meg 4 tizedesjegyig:

Algorithm: Abc Xy, Difference: 11,1111%, Algorithm: Abc Xy: Difference: 11,1111%

**A megoldások:**
Algorithm: Generalized Linear Model, Difference: -4,3123%, Algorithm: Support Vector Machine: Difference: -6,0489%
Algorithm: Generalized Linear Model, Difference: -4,3123%, Algorithm: Support Vector Machine: Difference: -8,2361%

**Magyarázat a megoldáshoz**

### 4. feladat 0 / 9 pont

Állítsd vissza a test set nagyságát 30%-ra, és vedd ki (ignore) a bináris magyarázó változókat!
Add meg a MAE és az RMSE értékeit az egyes algoritmusoknak, 4 tizedes jegyig, a következő formában:

Algorithm: Abc Xy, MAE: 11,1111, RMSE: 11,1111, Algorithm: Abc Xy, MAE: 11,1111, RMSE: 11,1111

**A megoldás:**
Algorithm: Generalized Linear Model, MAE: 3,4977, RMSE: 4,3304, Algorithm: Support Vector Machine, MAE: 3,4999, RMSE: 4,3324

**Magyarázat a megoldáshoz**

### 5. feladat 0 / 9 pont

Csak a SQUAREMETER magyarázó változót használva milyen performancia értékek jönnek ki a 30%-os méretű test seten visszamérve a modellt?
Add meg az egyes algoritmusok MAE és a RMSE értékeit, 4 tizedes jegyig, a következő formában:

Algorithm: Abc Xy, MAE: 11,1111, RMSE: 11,1111, Algorithm: Abc Xy, MAE: 11,1111, RMSE: 11,1111

**A megoldás:**
Algorithm: Generalized Linear Model, MAE: 4,0039, RMSE: 5,0110, Algorithm: Support Vector Machine, MAE: 4,0041, RMSE: 5,0262

**Magyarázat a megoldáshoz**

Vissza a kategóriáimhoz

**Adattárház elemző (BCS)**
7. forduló

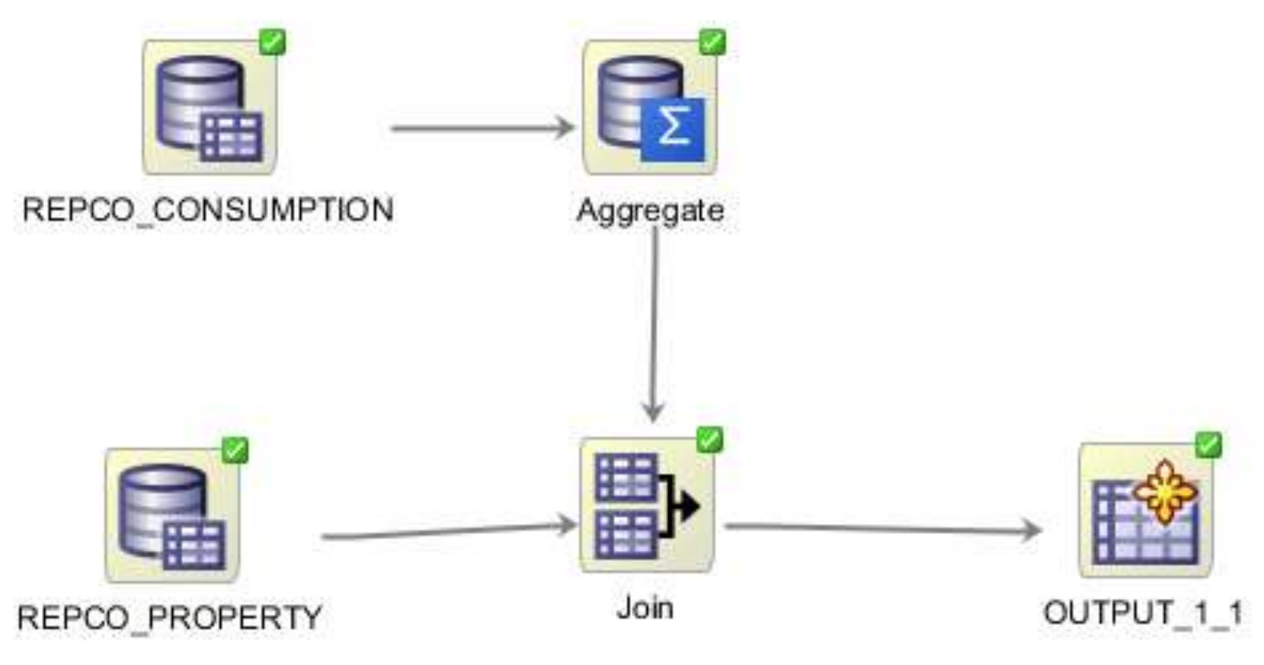
## Ismertető a feladathoz

Regressziós algoritmusok segítségével szeretnénk előrejelezni az ügyfelek áram fogyasztási adatait. Ehhez az **Oracle Data Miner-**t fogjuk használni, ami az Oracle SQL Developer egy bővítménye. Ez a korábban letöltött SQL Developer-ben megtalálható és a következő videó alapján konfigurálható: https://www.youtube.com/watch?v=tZCaQJsNVSs.

A Data Miner használatához további segítséget nyújt a következő lejátszási lista: https://www.youtube.com/playlist?list=PL99-DcFspRUq8VbbgXe2lQ559VDr7BSCr

A célváltozónk a **REPCO_CONSUMPTION** tábla AMOUNT változójából képzett ügyfélszintű átlagfogyasztás lesz, a magyarázó változóink pedig a **REPCO_PROPERTY** tábla változói lesznek (SQUAREMETER, PERFORMANCE_OF_SOLAR_PANELS, THERMAL_INSULATION_THICKNESS, SOLAR_PANELS, THERMAL_INSULATION). Ehhez először a REPCO_CONSUMPTION táblát kell aggregálni, majd a REPCO_PROPERTY táblával kell összekötni azt.

Ezen lépések végeztével egy az alábbi képen látható Workflow diagramhoz hasonlót kell kapni.



Ez után futtass regressziós algoritmusokat az adaton, ügyelj arra, hogy a **megfelelő cél- és magyarázó változókat használd** (a fent említett módon, ettől több változót ne használj a regresszió során). (Nem kell train és test setre bontani az adatot, ez a Regression Node-on belül megtörténik.)

**A megoldások beírásának megkönnyítésére a letölthető Word dokumentumban megtalálod a megoldások formátumát, be kell helyettesítened az értékeket, és utána be tudod másolni a válaszokat.**

**Tekintettel arra, hogy egy választ sem rögzítettél az alábbi feladatlapon, ebben a fordulóban a kitöltésére rendelkezésre álló idő teljes egésze, azaz 60 perc került rögzítésre mint megoldáshoz felhasznált idő.**

### 1. feladat 0 / 4 pont

Állítsd a test set nagyságát 30%-ra, és futtasd a regressziót! (Tipp: Case ID a CUSTOMER_ID lesz.)
Milyen algoritmusokat használ a Regression Node? Add meg az összes rendelkezésre álló algoritmus nevét a következő formátumban: Abc Xy, Abc Xy

**A megoldások:**
Generalized Linear Model, Support Vector Machine
REGR_GLM_1_1 Generalized Linear Model, REGR_SVM_1_1 Support Vector Machine
Generalized Linear Model,Support Vector Machine
Generalized Linear Models GLM, Support Vector Machines SVM
Generalized Linear Models, Support Vector Machine

**Magyarázat a megoldáshoz**

### 2. feladat 0 / 9 pont

Add meg az egyes algoritmusok MAE és RMSE értékeit, 4 tizedes jegyig, a következő formában:

Algorithm: Abc Xy, MAE: 11,1111, RMSE: 11,1111, Algorithm: Abc Xy, MAE: 11,1111, RMSE: 11,1111

**A megoldások:**
Algorithm: Generalized Linear Model, MAE: 3,5069, RMSE: 4,3386, Algorithm: Support Vector Machine, MAE: 3,5048, RMSE: 4,3472
Algorithm: Generalized Linear Model, MAE: 3,5069, RMSE: 4,3386, Algorithm: Support Vector Machine, MAE: 3,5052, RMSE: 4,3475

**Magyarázat a megoldáshoz**

### 3. feladat 0 / 9 pont

Állítsd a test set nagyságát 90%-ra, és nézd meg mi történik a Predictive Confidence-el! Hány százalékot változott a Predictive Confidence algoritmusonként külön-külön az előző 30%-os méretű test seten történt visszaméréshez képest?

(Tipp: ne százalékos változást számolj, a Perdiction Confidence értéke van %-ban)

Az értékeket a következő formában add meg 4 tizedesjegyig:

Algorithm: Abc Xy, Difference: 11,1111%, Algorithm: Abc Xy: Difference: 11,1111%

**A megoldások:**
Algorithm: Generalized Linear Model, Difference: -4,3123%, Algorithm: Support Vector Machine: Difference: -6,0489%
Algorithm: Generalized Linear Model, Difference: -4,3123%, Algorithm: Support Vector Machine: Difference: -8,2361%

**Magyarázat a megoldáshoz**

### 4. feladat 0 / 9 pont

Állítsd vissza a test set nagyságát 30%-ra, és vedd ki (ignore) a bináris magyarázó változókat!
Add meg a MAE és az RMSE értékeit az egyes algoritmusoknak, 4 tizedes jegyig, a következő formában:

Algorithm: Abc Xy, MAE: 11,1111, RMSE: 11,1111, Algorithm: Abc Xy, MAE: 11,1111, RMSE: 11,1111

**A megoldás:**
Algorithm: Generalized Linear Model, MAE: 3,4977, RMSE: 4,3304, Algorithm: Support Vector Machine, MAE: 3,4999, RMSE: 4,3324

**Magyarázat a megoldáshoz**

### 5. feladat 0 / 9 pont

Csak a SQUAREMETER magyarázó változót használva milyen performancia értékek jönnek ki a 30%-os méretű test seten visszamérve a modellt?
Add meg az egyes algoritmusok MAE és a RMSE értékeit, 4 tizedes jegyig, a következő formában:

Algorithm: Abc Xy, MAE: 11,1111, RMSE: 11,1111, Algorithm: Abc Xy, MAE: 11,1111, RMSE: 11,1111

**A megoldás:**
Algorithm: Generalized Linear Model, MAE: 4,0039, RMSE: 5,0110, Algorithm: Support Vector Machine, MAE: 4,0041, RMSE: 5,0262

**Magyarázat a megoldáshoz**

Vissza a kategóriáimhoz