עידו שקורי 206557332
יהונתן לייטנר 312474646
אסף אבירן 204620512

שיעורי בית 1 – הנדסת נתוני עתק

קבוצה - 19

בשלב הראשון נרצה לזהות שאלות עליהם נבסס את ניתוח הנתונים בהמשך , וזאת כדי לאפשר ניתוח מזווית חדשה אודות תאונות הדרכים בברצלונה וקבלת החלטה לצמצום מספר תאונות הדרכים בעיר.

שאלות החקר :

1) האם בשכונות בהם מספר המובטלים גדול מהממוצע , מתרחשות יותר תאונות דרכים במהלך יימות השבוע ?

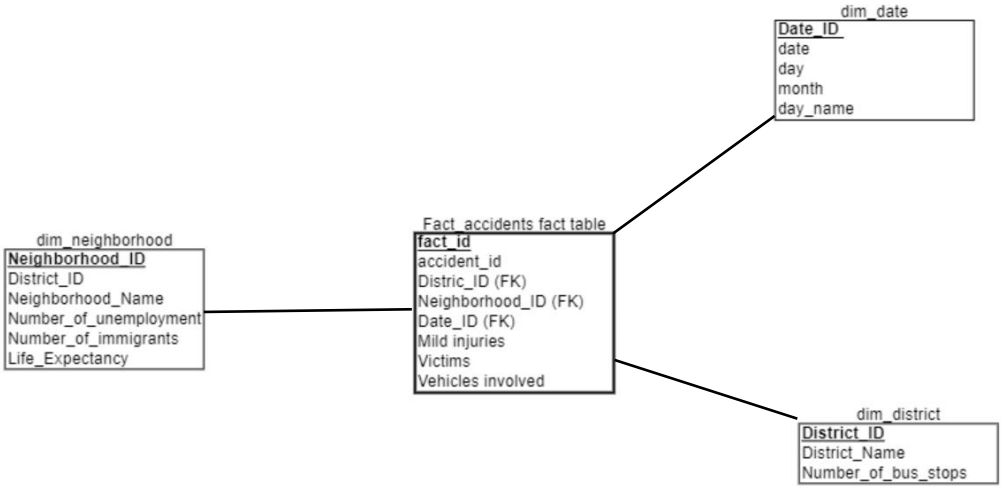
2) האם יש קשר בין כמות תאונות הדרכים בסוף השבוע במחוז לבין מספר תחנות האוטובוס הקיימות ?

3)האם יש קשר בין תוחלת החיים בשכונות לבין סך כמות הפציעות הקלות והחמורות המדווחות במהלך שנת 2017 ?

בשלב הבא נרצה לאפיין את מחסן הנתונים עבור ניתוח המידע.

1. **זיהוי נושא המחסן :** מידע אודות תאונות דרכים בעיר ברצלונה והקשר בינם לבין המיקום בו התקיימו וזמן התאונה.
2. **זיהוי העובדות:**
   - מידת הפציעות ומספר רכבים מעורבים
   - השכונה בה התרחשה התאונה
   - המחוז בו התרחשה התאונה
   - התאריך והשעה בה התרחשה התאונה
3. **הגדרת הגרעין** : כל תאונה מיוצגת כרשומה אחת בטבלת העובדות
4. **הגדרת המימדים :**
   - NEIGHBORHOOD – מימד השכונה
   - DISTRICT – מימד המחוז
   - DATE – מימד התאריך

תרשים מחסן הנתונים



עידו שקורי 206557332
יהונתן לייטנר 312474646
אסף אבירן 204620512

פירוט המס"ד

FACT_ACCIDENT

שדה	טיפוס נתונים	מקור נתונים	הסבר
Fact_id	Int(11) not null auto increment	מפתח רץ	מפתח ראשי רץ לטבלת העובדות
Accident_id	varchar(45) NOT NULL	טבלת ACCIDENT ממס"ד BARCELONA	מזהה ייחודי שקיים בטבלת תאונות לכל אירוע
District_ID	Int(10) not null	טבלת UNIMPLOYMENT ממס"ד ברצלונה	מפתח זר ייחודי לכל מחוז
Neighborhood_ID	Int(10) not null	טבלת UNIMPLOYMENT ממס"ד ברצלונה	מפתח זר ייחודי לכל שכונה
Date_ID	Int(11) not null auto increment	טבלת DIM_DATE	מפתח רץ זר לכל תאריך
Serious injuries	(11)int	טבלת ACCIDENT ממס"ד BARCELONA	מספר פציעות קשות בתאונה
Mild Injuries	(11)int	טבלת ACCIDENT ממס"ד BARCELONA	מספר פציעות קלות בתאונה
Victems	(11)int	טבלת ACCIDENT ממס"ד BARCELONA	מספר קורבנות בתאונה
Vehicle involved	(11)int	טבלת ACCIDENT ממס"ד BARCELONA	מספר הרכבים המעורבים בתאונה

DIM_NEIGHBORHOOD

שדה	טיפוס נתונים	מקור נתונים	הסבר
Neighborhood_ID	Int(10) not null	טבלת UNIMPLOYMENT ממס"ד ברצלונה	מפתח זר ייחודי לכל שכונה
District_ID	Int(10) not null	טבלת UNIMPLOYMENT ממס"ד ברצלונה	מפתח זר ייחודי לכל מחוז
Neighborhood_Name	(45)varchar	טבלת UNIMPLOYMENT ממס"ד ברצלונה	שם ייחודי לכל שכונה
Number_of_unempoyment	(11)int	טבלת UNIMPLOYMENT	לקיחת הנתונים לכל שכונה ויישום פונקצית SUM לכל דורשי העבודה , נשים וגברים באחד
Number_of_immigrants	(11)int	טבלת immigrants	לקיחת הנתונים לכל שכונה ויישום

עידו שקורי 206557332
יהונתן לייטנר 312474646
אסף אבירן 204620512

			פונקציית SUM לכל הלאומים
Life_Expectancy	(11)int	טבלת Neighborhood	לקיחת הנתונים וביצוע פונקצית AVG לכל השנים

DIM_DISTRICT

שדה	טיפוס נתונים	מקור נתונים	הסבר
District_ID	Int(10) not null	טבלת STOPS.BUS	מפתח ייחודי לכל מחוז
District _Name	(45)varchar	טבלת UNIMPLOYMENT ממס"ד ברצלונה	שם ייחודי לכל מחוז
Number_of_bus_stops	(11)int	טבלת STOPS.BUS	ביצוע פונקצית COUNT לכל תחנות האוטובוס במחוז

DIM_DATE

שדה	טיפוס נתונים	מקור נתונים	הסבר
Date_ID	Int(11) not null auto increment	מפתח רץ	מזהה ייחודי לכל תאריך
date	Date	טבלת ACCIDENT	תאריך התאונה
day	(11)int	טבלת ACCIDENT	יום התאונה (מספרי)
month	(45)varchar	טבלת ACCIDENT	חודש התאונה
Day_name	(45)varchar	טבלת ACCIDENT	שם יום התאונה

עידו שקורי 206557332
יהונתן לייטנר 312474646
אסף אבירן 204620512

תוצאות השאילתות וקוד SQL

1) האם בשכונות בהם מספר המובטלים גדול מהממוצע , מתרחשות יותר תאונות דרכים במהלך יימות השבוע?

number of accidents	Number_of_unemployment	Neighborhood_Name	Neighborhood_ID
83	2011	el Raval	**1**
108	1797	Sant Andreu	**60**
258	1398	la Nova Esquerra de l'Eixample	**9**
122	1312	la Vila de Gràcia	**31**
235	1299	la Sagrada Família	**6**
113	1178	el Camp de l'Arpa del Clot	**64**
211	1118	les Corts	**19**
122	1096	Sants	**18**
135	1015	Sant Antoni	**10**

תוצאות השאילתא מראות שממוצע התאונות בשכונות בהם רמת האבטלה גבוהה מהממוצע , כמות התאונות גם כן עומדת מעל הממוצע העומד על 139 תאונות לשכונה כלומר , ישנו קשר ישיר בין כמות התאונות במהלך שבוע העבודה לבין כמות האבטלה בשכונה.

שאילתת SQL

```
select F.Neighborhood_ID , N.Neighborhood_Name ,N.Number_of_unemployment,
'count(accident_ID) as 'number of accidents

from accidents_barcelona_dw.fact_accidents as F , accidents_barcelona_dw.dim_neighborhood as N ,
accidents_barcelona_dw.dim_date as D

where F.Neighborhood_ID = N.Neighborhood_ID and F.Date_ID = D.Date_ID and (D.day_name =
'Monday' or D.day_name = 'Tuesday' or D.day_name = 'Wednesday' or D.day_name = 'Thursday')

group by F.Neighborhood_ID , N.Neighborhood_Name , N.Number_of_unemployment

having N.Number_of_unemployment > (select avg(Number_of_unemployment) from

(accidents_barcelona_dw.dim_neighborhood
```

206557332 עידו שקורי
312474646 יהונתן לייטנר
204620512 אסף אבירן

2) האם יש קשר בין כמות תאונות הדרכים בסוף השבוע במחוז לבין מספר תחנות האוטובוס הקיימות?

number of accidents	Number_of_Bus.Stop	District_Name	District_ID
255	167	Ciutat Vella	**1**
222	228	Sant Andreu	**9**
257	207	Les Corts	**4**
176	210	Gràcia	**6**

בשאילתא זאת דווקא קיבלנו את התוצאות ההפוכות ממה שחשבנו תחילה. מספר תאונות הדרכים בסוף השבוע במחוזות בהם מספר תחנות האוטובוס קטן מהממוצע קטן גם הוא ממוצע שאר המחוזות. ייתכן שהדבר נובע מכך שבקרבת תחנות אוטובוס ישנה תנועה גדולה של הולכי רגל , עובדה היכולה להסביר את הנתונים וריבוי התאונות. סיבה נוספת שיכולה להשפיע היא גודל המחוז , אך מידע זה לא כלול במס"ד שנבנה.

<u>שאילתת SQL</u>

```
select N.District_ID , N.District_Name ,N.`Number_of_Bus_Stops` ,count(F.fact_ID)

from accidents_barcelona_dw.dim_district as N , accidents_barcelona_dw.fact_accidents as F , accidents_barcelona_dw.dim_date as D

where N.District_ID = F.District_ID and F.Date_ID = D.Date_ID and (D.day_name = 'Friday' or D.day_name = 'Saturday' or D.day_name = 'Sunday')

`group by N.District_ID , N.District_Name , N.`Number_of_Bus_Stops

having N.`Number_of_Bus_Stops`<(select avg(`Number_of_Bus_Stops`) from

(accidents_barcelona_dw.dim_district
```

עידו שקורי 206557332
יהונתן לייטנר 312474646
אסף אבירן 204620512

3) האם יש קשר בין תוחלת החיים בשכונות לבין סך כמות הפציעות המדווחות החל משנת 2017, וזאת כדי לבדוק את השפעת פציעות התאונה על ממוצע תוחלת החיים.

sum(A.`Serious injuries`)	sum(A.`Mild injuries`)	Life_Expectancy	Neighborhood_Name
1	6	61.58	**Torre Baró**
0	54	63.73	**Baró de Viver**
2	9	78.62	**Can Peguera**
0	52	80.37	**Montbau**
5	142	80.87	**el Raval**
0	31	80.93	**Vallvidrera, el Tibidabo i les Planes**
3	48	80.99	**les Roquetes**
2	150	81.25	**la Barceloneta**
0	111	81.27	**la Trinitat Nova**
1	64	82	**la Salut**

בשאילתא זו לא הצלחנו למצוא קשר בין כמות הנפגעים בתאונות הדרכים לבין תוחלת החיים בשכונה. ייתכן והקשר אינו קיים , אך עם זאת ייתכן שחוסר המידע בקשר למספר התושבים בשכונה וגודל המקום יכול היה לשפוך אור נוסף על הקשר.

<u>שאילתת SQL</u>

```
select N.Neighborhood_Name , N.Life_Expectancy , sum(F.`Mild injuries`) , sum(F.`Serious injuries`)

from accidents_barcelona_dw.fact_accidents as F
, ,accidents_barcelona_dw.dim_neighborhood as N

accidents_barcelona_dw.dim_date as D

where F.Neighborhood_ID =  N.Neighborhood_ID and F.Date_ID = D.Date_ID and (D.date >2017/01/01)# and (D.date<2017/12/31)

group by N.Neighborhood_Name , N.Life_Expectancy

limit 10
```

206557332 עידו שקורי
312474646 יהונתן לייטנר
204620512 אסף אבירן

קוד SQL ליצירת המס"ד :

טבלת העובדות

```
) `CREATE TABLE `accidents_barcelona_dw`.`fact_accidents
,NULL AUTO_INCREMENT      NOT          fact_ID` int(11)`
,accident_ID` varchar(45) NOT NULL`
,District_ID` int(11) NOT NULL`
,Neighborhood_ID` int(11) NOT NULL`
,Date_ID` int(11) NOT NULL NOT NULL`
,Mild injuries` int(11) DEFAULT NULL`
,Serious injuries` int(11) DEFAULT NULL`
,Victims` int(11) DEFAULT NULL`
,Vehicles involved` int(11) DEFAULT NULL`
,PRIMARY KEY (`fact_ID`)
,KEY `idx_fk_District_ID` (`District_ID`)
,KEY `idx_fk_Neighborhood_ID` (`Neighborhood_ID`)
KEY `idx_fk_Date_ID` (`Date_ID`)
ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci;CREATE  (
) `TABLE `accidents_dw`.`dim_district
,District_ID` int(11) NOT NULL`
,District_Name` varchar(45) DEFAULT NULL`
,Number_of_Bus.Stop` int(11) DEFAULT NULL`
PRIMARY KEY (`District_ID`)
;ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci (
```

206557332 עידו שקורי
312474646 יהונתן לייטנר
204620512 אסף אבירן

מימד השכונה

```
) `CREATE TABLE `accidents_barcelona_dw`.`dim_neighborhood
,District_ID` int(11) NOT NULL`
,Neighborhood_ID` int(11) NOT NULL`
,Neighborhood_Name` varchar(45) DEFAULT NULL`
,Number_of_unemployment` int(11) DEFAULT NULL`
,Number_of_immigrants` int(11) DEFAULT NULL`
,Life_Expectancy` double DEFAULT NULL`
PRIMARY KEY (`Neighborhood_ID`)
;ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci (
```

מימד המחוז

```
) `CREATE TABLE `accidents_barcelona_dw`.`dim_district
,District_ID` int(11) NOT NULL`
,District_Name` varchar(45) DEFAULT NULL`
,Number_of_Bus_Stops` int(11) DEFAULT NULL`
PRIMARY KEY (`District_ID`)
;ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci (
```

מימד התאריך

```
) `CREATE TABLE `accidents_barcelona_dw`.`dim_date
,NULL AUTO_INCREMENT      NOT          Date_ID` int(11)`
,date` date DEFAULT NULL`
,day` int(11) DEFAULT NULL`
,month` varchar(45) DEFAULT NULL`
,year` int(11) DEFAULT NULL`
,day_name` varchar(45) DEFAULT NULL`
PRIMARY KEY (`Date_ID`)
;ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci (
```

206557332 עידו שקורי
312474646 יהונתן לייטנר
204620512 אסף אבירן

פקודות INCSERT :

הכנסה למימד המחוז

```
insert into accidents_barcelona_dw.dim_district(District_Name , District_ID ,
`Number_of_Bus_Stops`)

select table1.District_Name , table2.District_Code , table1.number_of_bus_stops

)from

'select B.District_Name , count(B.`Bus.Stop`) as 'number_of_bus_stops

from barcelona.bus_stops as B

group by B.District_Name

table1 (

JOIN

)

select B.District_Name , D.District_Code

from barcelona.bus_stops as B , barcelona.deaths as D

where B.District_Name = D.District_Name

group by B.District_Name , D.District_Code

table2 (

on table1.District_Name = table2.District_Name

where table1.District_Name is not null

##############################################################################
```

הכנסה למימד השכונה

```
insert into accidents_barcelona_dw.dim_neighborhood
(District_ID,Neighborhood_ID,Neighborhood_Name,Number_of_unemployment,Life_Expect
ancy,Number_of_immigrants)

select table1.District_Code , table1.Neighborhood_Code , table1.Neighborhood_Name ,
(table1.male_unemployment + table2.female_unemployment) , (table3.male_life +
table4.female_life)/2 , table5.immegrants

)from

select U.District_Code , U.Neighborhood_Code , U.Neighborhood_Name , avg(U.Number) as
''male_unemployment

from barcelona.unemployment as U

where U.Demand_occupation = 'Registered unemployed' and
'U.Gender = 'Male

group by U.District_Code , U.Neighborhood_Code , U.Neighborhood_Name

table1(
```

206557332 עידו שקורי
312474646 יהונתן לייטנר
204620512 אסף אבירן

```
JOIN
)
select U.District_Code , U.Neighborhood_Code , U.Neighborhood_Name , avg(U.Number) as
''female_unemployment
from barcelona.unemployment as U
'where U.Demand_occupation = 'Registered unemployed' and U.Gender = 'Female
group by U.District_Code , U.Neighborhood_Code , U.Neighborhood_Name
(
table2
JOIN
)
select L.Neighborhood , (L.`2006-2010` + L.`2007-2011` + L.`2008-2012` + L.`2009-2013` +
'L.`2010-2014`)/5 as 'male_life
from barcelona.life_expectancy as L
'where L.Gender = 'Male
group by L.Neighborhood
(
table3
JOIN
)
select L.Neighborhood , (L.`2006-2010` + L.`2007-2011` + L.`2008-2012` + L.`2009-2013` +
'L.`2010-2014`)/5 as 'female_life
from barcelona.life_expectancy as L
'where L.Gender = 'Female
group by L.Neighborhood
(
table4
JOIN
)
'select I.Neighborhood_Name , sum(I.Number) as 'immegrants
from barcelona.immigrants as I
group by I.Neighborhood_Name
```

206557332 עידו שקורי
312474646 יהונתן לייטנר
204620512 אסף אבירן

```
(

table5

on table1.District_Code = table2.District_Code and table2.Neighborhood_Name =
table3.Neighborhood and table3.Neighborhood = table4.Neighborhood and
table4.Neighborhood=table5 .Neighborhood_Name

group by table1.Neighborhood_Code , table1.Neighborhood_Name

##############################################################################

;"2017-01-01" =:  SET @currdate

;"2020-12-31" =:  SET @enddate

$$ delimiter

$$DROP PROCEDURE IF EXISTS BuildDate

()CREATE PROCEDURE BuildDate

BEGIN

enddate DO@ >          WHILE @currdate

(date,day,month,year,day_name)  INSERT INTO dim_date


) VALUES


        DAY(@currdate),                  currdate,@
    ,MONTH(@currdate)


DAYNAME(@currdate)         YEAR(@currdate),


;(


;DAY)          1       DATE_ADD(@currdate, INTERVAL =:  SET @currdate

;END WHILE

$$END

;()CALL BuildDate
```

הכנסה למימד התאריך

```
insert into accidents_barcelona_dw.fact_accidents (accident_ID, District_ID,
Neighborhood_ID, Date_ID,`Mild injuries`,`Serious injuries`,Victims,`Vehicles involved`)

select A.Id ,N.District_ID , N.Neighborhood_ID ,D.Date_ID, A.`Mild injuries` , A.`Serious
`injuries` , A.Victims , A.`Vehicles involved
```

206557332 עידו שקורי
312474646 יהונתן לייטנר
204620512 אסף אבירן

from barcelona.accidents as A , accidents_barcelona_dw.dim_neighborhood as N , accidents_barcelona_dw.dim_date as D

where A.accident_date = D.date and A.Neighborhood_Name = N.Neighborhood_Name

הכנסה לטבלת העובדות