# פרויקט סיום קורס DATA ANALYST

## מטרה הפרויקט

מטרת הפרויקט היא לעזור לעריית ניו יורק לנתח את פעילות אכיפת עבירות החניה בשטחה. לצורך זה עליכם להשתמש בכלים שלמדתם בקורס על מנת לתת תמונה כוללת של הפעילות.

בעיריית NY ניתנים מדי שנה יותר מ-10 מיליון (!) דוחות עבור עבירות החניה השונות. על מנת להקל על העבודה נבחרו באופן רנדומלי שלושה אחוז מהנתונים. הנתונים נלקחו מאתר NYC Open data.

ישנם דרכים שונות לבצע פרויקט מסוג זה. בחרנו בדרך הכי נכונה לבצע את המשימה, דרך המשלבת ההיבטים השונים של עבודת אנליסט בעולם ה-BI. הפרויקט מדמה עבודת אנליסט המתחיל בפרויקט חדש של ניתוח הנתונים ומעוניין לנתח את הנתונים. המשימות בפרויקט כוללות פיתוח שאילתות SQL לניתוח הנתונים, הכנת הנתונים לניתוח באמצעות ה-query editor של ה-Power BI ובניית ויזואליזציות ב-Power BI.

מילה על איכות הנתונים: בניגוד לבסיסי הנתונים ששימשו אתכם במהלך הקורס, הנתונים בפרויקט הנם נתונים אמיתיים על כל מה שמשתמע מכך: ערכים חסרים, קודים שאין עבורם תרגום, קודים שגויים. עליכם להתמודד עם מצב זה במהלך הפרויקט.

## המשימות בפרויקט

1. **מקורות הנתונים ומודל הנתונים של הפרויקט**

- מודל הנתונים של הפרויקט נבנה לפי העקרונות של המידול הממדי במבנה מסוג פתית שלג (ראו נספחי א' ו-ב')
- נתוני המקור לפרויקט הנם בסיס נתונים ב-SQL SERVER בשם DWH_DATA_ANALYST . בסיס הנתונים בפורמט BAK. יש לשחזר את בסיס הנתונים לשרת ה-SQL SERVER לפני תחילת העבודה (ראו הסבר בנספח ג.).

2. **פיתוח שליפות SQL לניתוח הנתונים**

באמצעות SQL Server Management Studio יש לפתח מספר שאילתות לצורך ניתוח נתוני דוחות החנייה בעיריית ניו יורק.

**שליפה 1**

יש לפתח שליפה שתציג את כמות דוחות החנייה לפי רובעים (Borough) בעיריית ניו יורק לשנים 2015 עד 2017.

- יש להפוך את תאריך הדוח בשליפה מסוג מחרוזת לסוג תאריך על מנת לאפשר הפעלת פונקציות של תאריכים במסנן.
- יש להציג את שם הרובע.
- יש למיין את התוצאות לפי סדר יורד של כמות דוחות החניה

בסיום פיתוח השאילתה יש להפוך אותה ל-Stored Procedure כשהפרוצדורה תופעל עם פרמטר של שם הרובע.

**שליפה 2**

יש להוסיף לשליפה הקודמת את היום בשבוע שבו ניתנו דוחות החנייה כך שהשליפה תציג את כמות דוחות החנייה לכל רובע ולכל יום בשבוע.

- יש להציג את שם היום בשבוע (לא את המספר)
- תוצאת השליפה תהיה ממוינת לפי רובע ויום בשבוע

בסיום פיתוח השאילתה יש להפוך אותה ל-Stored Procedure כשהפרוצדורה תופעל עם פרמטר של יום בשבוע.

**שליפה 3**

יש לפתח שליפה שתציג את חמשת סוגי עבירות החניה הכי נפוצות בעיריית ניו יורק בשנים 2015 עד 2017.

בסיום פיתוח השאילתה יש להפוך אותה ל-Stored Procedure כשהפרוצדורה תופעל עם פרמטר של מספר העבירות הכי נפוצות (Top N).

**שליפה 4**

יש להציג את שתי סוגי העבירות הכי נפוצות **לכל צבע רכב** בעיריית ניו יורק בשנים 2015 עד 2017.

- יש להימנע מלהציג צבע רכב לא ידוע.
- אם ישנו אותו מספר דוחות החניה לסוג חניה עבור אותו צבע רכב, יש להציג אחד מהקודים בלבד.

בסיום פיתוח השאילתה יש להפוך אותה ל-Stored Procedure כשהפרוצדורה תופעל עם פרמטר של מספר העבירות הכי נפוצות (Top N).

**שליפה 5**

יש לבנות שליפה המציגה כמות הרכבים שקיבלו יותר מ-10 דוחות חניה, כמה בין 5 ל-9 וכמה מתחת ל-5 דוחות בשנים 2015 עד 2017.

**שליפה 6**

יש להציג שליפה המציגה לכל מדינה שבה רשומה הרכב את עמודות הבאות:

שם המדינה

- כמות דוחות החנייה בשנת 2015
- כמות דוחות החנייה בשנת 2015
- כמות דוחות החנייה בשנת 2015
- אחוז השינוי של כמות דוחות החנייה בין שנת 2017 לבין שנת 2015 (יש להציג את המספר באחוזים)

**תוצר השלב:** תוכנית SQL הכולל את השאילתות שנידרשו.

### 3. הכנת הנתונים ובניית דוחות ב-Power BI for Desktop

#### a. שלב הכנת הנתונים

מקור הנתונים העיקרי לדוחות הנו בסיס הנתונים DWH_DATA_ANALYST. יש לטעון את הנתונים מבסיס הנתונים לתוך מודל הנתונים של ה-Power BI.

- על מנת להעשיר את הנתונים לדוחות יש לטעון ל-Power BI נתוני תיאור עבירת החנייה מקובץ ה-CSV (DOF_Parking_Violation_Codes.csv). לאחר טעינת הטבלה למודל הנתונים יש ליצור קשר גומלין עם הטבלה הרלוונטית.
- תאריך ושעת מתן הדוח (Issue Date ו-Issue Time) בטבלאות הנם מסוג טקסט. יש להפוך אותם לשדות מסוג תאריך ושעה בהתאם באמצעות ה-**Query Editor**.
  - ✓יש לשנות את השדה של שעת הדו"ח כך שהוא יהיה בפורמט HH:MM AM (או PM) . לאחר מכן יש להפוך את סוג השדה ל-TIME.

- ✓ לגבי שדה התאריך, יש להשתמש באפשרות של שינוי סוג השדה הלוקח בחשבון את הגדרת האזור ("Using Locale").
- ✓ לאחר מכן יש לסנן את הנתונים ב-**Query Editor** על מנת לשמור את הדוחות שניתנו בשנת 2015 עד 2017 בלבד.

- טיפול בטעויות בטעינה. על מנת להימנע מהשמטת הרשומות בהבאת הנתונים ל-PowerBI יש להפוך את הטעויות לערכים חסרים (**Null**) ב-**Query Editor** .
- בתוך ממשק קשרי הגומלין יש ליצור את הקשרים המתאימים בין כל הטבלאות במודל (ראה נספח ב').
- במודל שנבנה חסר ממד תאריכים. יש להגדיר באמצעות שפת DAX טבלה חדשה במודל עבור ממד זה באמצעות בניית טבלה מחושבת ושדות מחושבים. ראו דוגמה כאן:

http://www.mssqltips.com/sqlservertip/4857/creating-a-date-dimension-table-in-power-bi/

**b. שלב בניית הדוחות**

יש להשקיע מאמצים במראה המקצועי לדוחות באמצעות ויזואליזציות מתאימות וברורות, שימוש בכותרות, עיצוב השדות (כולל עיצובים מותנים). על מנת להקל על בניית הנוסחאות יש להפעיל את האופציה של ה-Quick Measures ב- Power BI.

**דוח 1 – ניתוח סוגי עבירות החנייה**

הדוח יציג ניתוח כולל של סוגי עבירות החנייה וייתן מענה לשאלות העסקיות הבאות:

- מה הם 5 סוגי עבירות החנייה הנפוצות ביותר לאורך השנים? ומה עם כל שנה בנפרד? האם חל שינוי משנה לשנה?
- באיזה יום בשבוע יש יותר עבירות חניה? האם יש הבדל בין הרובעים השונים?
- באילו שעות של היום (בפרקי זמן של שעתיים) יש יותר עבירות חניה? האם זה תלוי ברובע?
- הציגו בכל השנים השוואת מספר הדוחות מחודש לחודש וקצב הגידול החודשי.
- תבנו ויזואליזצה המציגה את מספר המצטבר של הדוחות בכל חודש לאורך השנה (YTD) לכל השנים.
- **שאלת בונוס:** הציגו ויזואליזציה המראה כמה רכבים ביצעו יותר מ-10 עבירות חניה, כמה בין 5 ל-9 וכמה מתחת ל-5 עבירות.

**דוח 2 - ניתוח סוגי הרכבים המעורבים בעבירות החנייה.**

דוח זה אמור לספק תובענות עסקיות לגבי סוגי הרכבים המעורבים בעבירות החנייה.

- מה סוג הרכב (Body Type) המקבל הכי הרבה דוחות חנייה ב-NY? תחשבו אם לא כדאי לקבץ את סוגי הרכב לקטגוריות. החלוקה לקטגוריות תיעשה לפי ראות עינכם עם דגש על קיבוץ ערכים לא נפוצים ביחד.
- מה הוא גובה הקנס הממוצע לכל סוג רכב (או להקבצת סוגי הרכב)?
- האם יש צבע רכב דומיננטי?
- מאיפה באים רוב הרכבים המעורבים בעבירות חניה? האם העירייה צריכה לשפר את ההסבר למי שאינו תושב מדינת NY (או המדינות הסמוכות לה) ?

**דוח 3 - ניתוח גאוגרפי של עבירות החנייה של משאיות במנהטן**

העירייה מעוניינת להבין איפה ומתי רוב עבירות החניה של משאיות (Delivery Trucks) מתבצעות ברובע מנהטן.

- הציגו על מפה גיאוגרפית את עבירות החניה שביצעו משאיות ברובע מנהטן עם חלוקה של שעות היום והלילה (תחשבו על חלוקה הגיונית של השעות). יש לחלק את הניתוח לשדרות (Avenues) מצד אחד ולרחובות (Streets) מצד שני. האם אפשר לבודד בצורה ברורה אזורים ו/או חלקי יום בעייתיים?
- מה הן עשרת הרחובות במנהטן שיש בהם הכי הרבה דוחות חניה עבור משאיות המספקות סחורה? אולי העירייה צריכה לחשוב על פתרונות חניה באזורים אלו?
- מה גובה הקנס הממוצע שמקבלים בעלי רכב אלו?

**הערה:** הדרך הכי מדויקת להציג נתונים על מפות היא להשתמש בקואורדינטות. היות שאין בקבצים נתונים מסוג זה, יש להסתמך על היכולת של ה-POWER BI "לנחש" את המיקום על פי נתוני המיקום שברשותכם.

**דוח 4 - ניתוח הכדאיות הכלכלית של פעילות האכיפה**

מעבר לצורך לאכוף את עבירות החניה בעיר, עיריית NY מעוניינת לבחון את הכדאיות הכלכלית של פעילות אכיפת עבירות החניה בעיר לפי הסוכנויות השונות. ההכנסות מהפעילות יחושבו באמצעות הנתונים במודל על גובה הקנסות. עלות מתן דוח לעירייה תחושב לפי 5$ לדוח חניה.

הרווח יהיה מחושב כהפרש בין הכנסות לעלויות. הדוח ייתן מענה לשאלות העסקיות הבאות:

- מה הרווח ואחוז הרווח (הכנסות פחות עלויות חלקי הכנסות) של עיריית NY מאכיפת חוקי החנייה בכל שנה לכל סוכנות?
- מה הרובע (borough) הכי רווחי בעיר NYC?
- האם היו שינויים מהותיים לאורך השנים ברווח מדוחות חנייה?
- מה הם עשרת סוגי הקנס הרווחיים ביותר בשנת 2016?

## בהצלחה!

## נספח א' שמות שדות וטבלאות בבסיס הנתונים

**FactParkingViolation**

| Column Name | Condensed Type |
|---|---|
| ParkingViolationKey | int |
| [Summons Number] | varchar(50) |
| [Issue Date] | varchar(50) |
| [Violation Code] | varchar(50) |
| [Violation Time] | varchar(50) |
| [Violation In Front Of Or Opposite] | varchar(50) |
| VehicleKey | int |
| IssuerKey | int |
| LocationKey | int |

**DimPlateType**

| Column Name | Condensed Type |
|---|---|
| PlateTypeCode | varchar(3) |
| PlateTypeNa... | varchar(40) |

**DimState**

| Column Name | Condensed Type |
|---|---|
| StateCode | varchar(2) |
| StateName | varchar(40) |

**DimBodyType**

| Column Name | Condensed Type |
|---|---|
| BodyTypeCode | varchar(5) |
| BodyTypeName | varchar(40) |

**DimLocation ***

| Column Name | Condensed Type |
|---|---|
| LocationKey | int |
| BoroughCode | varchar(2) |
| StreetCode | varchar(10) |
| StreetName | varchar(50) |
| HouseNumber | varchar(20) |
| City | varchar(20) |
| StateCode | varchar(2) |

**DimVehicle ***

| Column Name | Condensed Type |
|---|---|
| VehicleKey | int |
| PlateID | varchar(10) |
| RegistrationStateCo... | varchar(2) |
| PlateTypeCode | varchar(3) |
| BodyTypeCode | varchar(5) |
| BodyMakeName | varchar(5) |
| VehicleColorCode | varchar(10) |
| vehicleYear | varchar(4) |

**DimIssuer**

| Column Name | Condensed Type |
|---|---|
| IssuerKey | int |
| IssuerID | varchar(20) |
| IssuingAgencyCode | varchar(5) |

**DimColor**

| Column Name | Condensed Type |
|---|---|
| ColorCode | varchar(10) |
| ColorName | varchar(40) |

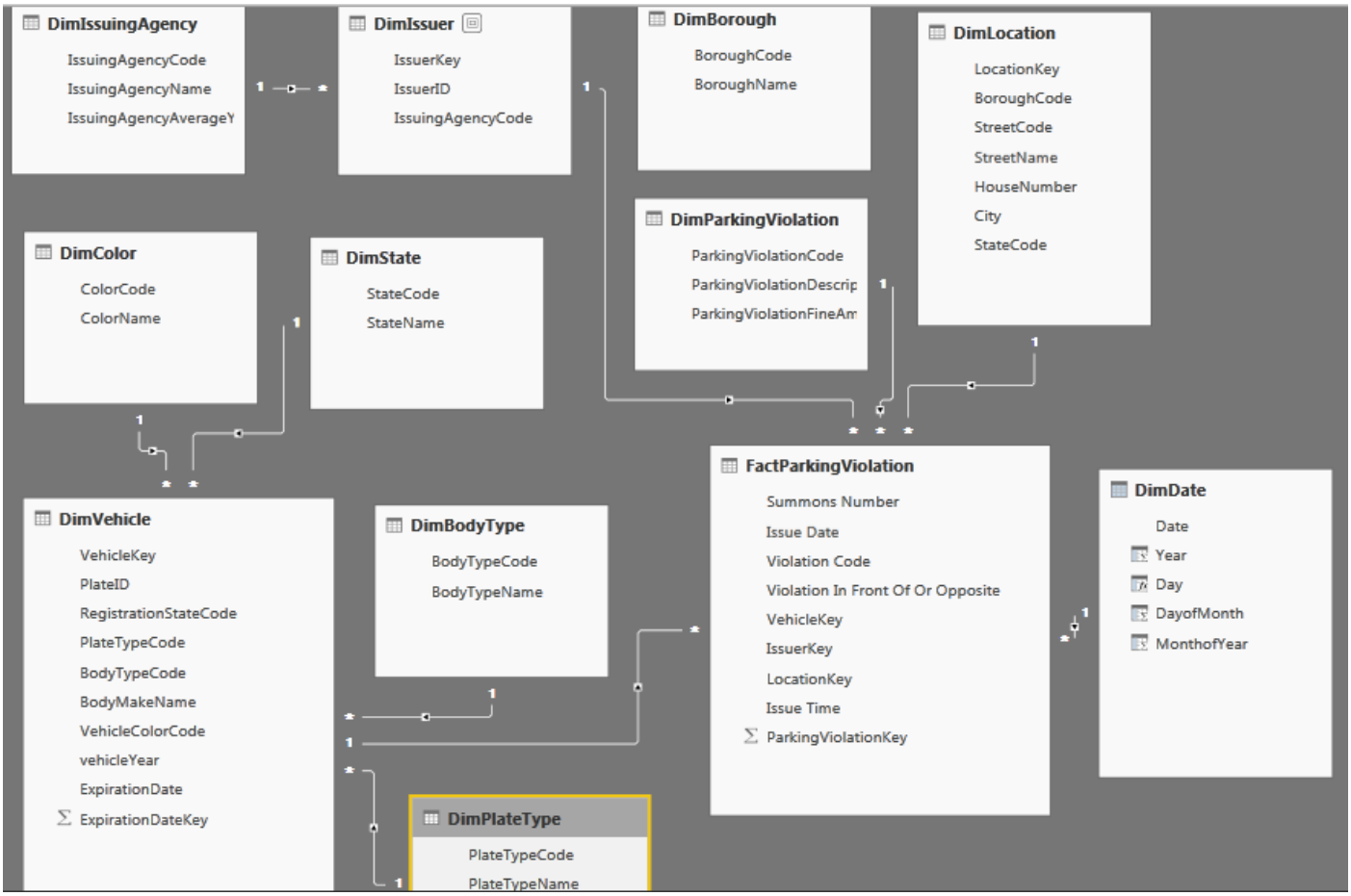
**DimIssuingAgency**

| Column Name | Condensed Type |
|---|---|
| IssuingAgencyCode | varchar(5) |
| IssuingAgencyName | varchar(100) |
| IssuingAgencyAverageYearlySala... | money |

**DimBorough**

| Column Name | Condensed Type |
|---|---|
| BoroughCode | varchar(2) |
| BoroughName | varchar(20) |

**נספח ב' מודל הנתונים ב-Power BI**



## נספח ג' שחזור בסיס הנתונים לפרויקט

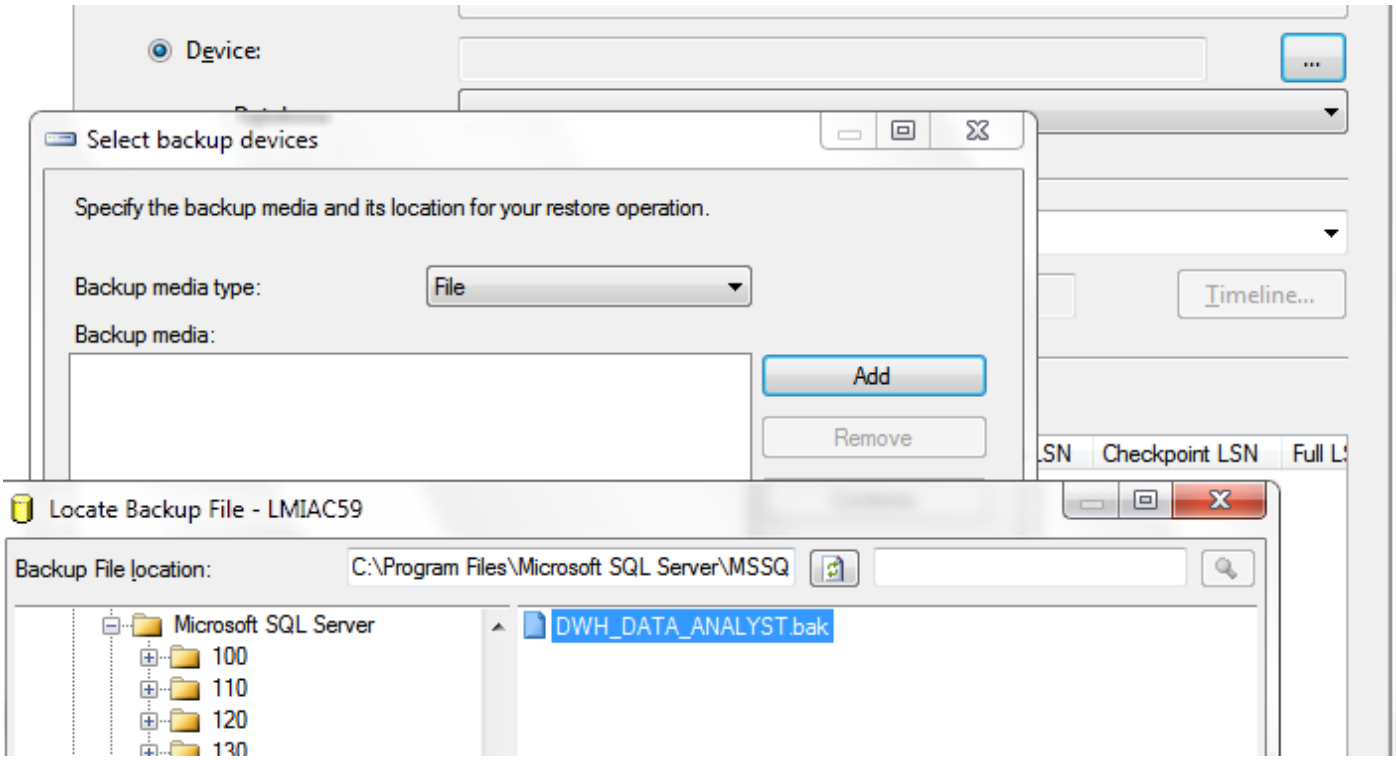
1. לפתוח את SQL Server Management Studio.

אם בסיס הנתונים DWH_DATA_ANALYST כבר קיים יש למחוק אותו (כפתור ימני של העכבר על שם בסיס הנתונים ו-delete)

2. ללחוץ על הכפתור הימני של העכבר על Databases ולבחור Restore Database...

3. בחלון שנפתח יש לבחור Device וללחוץ על הכפתור ... .

4. יש ללחוץ על ADD ולבחור את קובץ ה-BAK כפי שמופיע כאן:




5. בסיום יש ללחוץ על OK עד שחוזרים לחלון המקורי ושם ללחוץ שוב OK להפעיל את תהליך השחזור.