

פרויקט סיום קורס Business Analyst

מטרת הפרויקט

מטרת הפרויקט היא לעזור לעיריית ניו יורק לנתח את פעילות אכיפת עבירות החנייה בשטחה. לצורך זה יש להשתמש בכלים שנלמדו בקורס על מנת לתת תמונה כוללת של הפעילות.

בעיריית NY ניתנים מדי שנה יותר מ-10 מיליון (!) דוחות עבור עבירות החנייה השונות. על מנת להקל על העבודה נבחרו באופן מקרי שלושה אחוזים מהנתונים. (הנתונים נלקחו מאתר [NYC Open data](https://data.cityofnewyork.us/))

ישנם דרכים שונות לבצע פרויקט מסוג זה. בחרנו דרך המשלבת בין ההיבטים השונים של עבודת אנליסט בעולם ה-BI. הפרויקט מדמה עבודת אנליסט המעוניין לנתח את הנתונים. המשימות בפרויקט כוללות פיתוח שאילתות SQL, הכנת הנתונים לניתוח באמצעות ה-Query Editor של ה-Power BI ובניית ויזואליזציות ב-Power BI.

תוצרי הפרויקט

1. פיתוח שליליות SQL לניתוח נתונים - קובץ SQL הכוללת את השאילתות שנדרשו.
2. פיתוח מודל ובניית דוחות ב-Power BI Desktop – קובץ PBIX על פי הדרישות.

פיתוח שליליות SQL לניתוח הנתונים

באמצעות SQL Server Management Studio יש לפתח מספר שאילתות לצורך ניתוח נתוני דוחות החנייה בעיריית ניו יורק. (אין צורך לבנות טבלאות חדשות בתוך בסיס הנתונים ולא לתקן טבלאות קיימות.)

מקורות הנתונים

נתוני המקור לפרויקט הינו בסיס נתונים ב-SQL SERVER בשם DWH_DATA_ANALYST. יש לשחזר את בסיס הנתונים לשרת ה-SQL SERVER לפני תחילת העבודה (ראו הסבר בנספח [ג](#)).

שליפה 1

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

- יש להציג את שם הרובע.
- יש למיין את התוצאות לפי סדר יורד של כמות דוחות החנייה
- בסיום פיתוח השאילתה יש להפוך אותה ל-Stored Procedure כך שהפרוצדורה 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 כך שהפרוצדורה תופעל עם פרמטר של יום בשבוע.

שליפה 3

יש לפתח שליפה שתציג את חמשת סוגי עבירות החנייה, על-פי קוד עבירה, ViolationCode, 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.

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.
- 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 דוחות חניה בין השנים 2015 ל 2017 לפי קבוצות של:

- 10 או יותר
- בין 5 ל-9
- מתחת ל-5

שליפה 6

יש להציג שליפה המציגה לכל מדינה שבה רשום הרכב את העמודות הבאות:

- כמות דוחות החנייה בשנת 2015
- כמות דוחות החנייה בשנת 2016
- כמות דוחות החנייה בשנת 2017
- אחוז השינוי של כמות דוחות החנייה בין שנת 2017 לבין שנת 2015 (יש להציג את המספר באחוזים)

פיתוח המודל ובניית דוחות ב- Power BI Desktop

רקע

מטרת הפרויקט היא לעזור לעיריית ניו יורק לנתח את פעילות אכיפת עבירות החניה בשטחה כדי להתאים ולמקד את ההסברה לתושבים ובכך לצמצמם את הצורך במתן דוחות בשטחה. העירייה מעוניינת לצאת בקמפיין הסברה שיכיל מספר פרסומים בערוצי שיווק שונים, כך שכל פרסום יותאם לסוגי העבירות, האזורים הגיאוגרפיים בעיר, שעות מתן הדוחות וסוגי הרכבים. בנוסף, העירייה מעוניינת לנתח את הכדאיות הכלכלית בהפעלת פקחים בעיר לפי הסוכנויות השונות, הרובעים וסוגי הקנסות השונים כדי לבחון היכן פעולות האכיפה הינן הרווחיות ביותר (על מנת לתת המלצה היכן כדאי להגביר את אכיפת הדוחות על פי סוגי קנסות, רובעים או סוגי רכבים)

שלב הכנת הנתונים (ETL)

מקורות הנתונים

בסיס הנתונים: DWH_DATA_ANALYST ([מקור](#))

קובץ: DOF_Parking_Violation_Codes.xlsx ([מקור](#))

Dim Tables

DimVehicle

DimLocation

DimIssuer

Dim_Parking_Violation_Codes

DimDate - יש ליצור טבלה מחושבת בתוך POWER BI. (ראו דוגמה [כאן](#)):

DimTime – יש ליצור במידת הצורך (אפשר לפתור עם או ללא מימד זה)

Fact Table

FactParkingViolation

- העמודות תאריך ושעת מתן הדוח, Issue Date ו-Issue Time, הנם מסוג טקסט. יש להמיר אותם לשדות מסוג תאריך ושעה בהתאם.
- יש להציג את השדה של שעת הדו"ח כך בפורמט HH:MM AM/PM. לאחר מכן יש להפוך את סוג השדה ל-TIME.
- יש לסנן את הנתונים ב- Query Editor על מנת להציג את הדוחות שנכתבו בין השנים 2015 ל 2017 בלבד.
- אופציונלי: מזג לתוך FactParkingViolation את שמות הקנסות וגובה הקנס לכל שורה.
- טיפול בשגיאות בטעינה: על מנת להימנע מהשמטת הרשומות בהבאת הנתונים ל- Power BI יש להפוך את הטעויות לערכי NULL

מידול

יש לתכנן את מודל הנתונים על פי עקרונות סכמת כוכב. צרו את קשרי הגומלין המתאימים בין כל הטבלאות במודל.
(ראו נספחים א' ו-ב').

בניית הדוחות

יש להשקיע בניראות כמו גם בפונקציונאליות של הדוחות באמצעות ויזואליזציות מתאימות וברורות.

דוח 1 – ניתוח סוגי עבירות החנייה

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

1. מה הם 5 סוגי עבירות החנייה הנפוצות ביותר לאורך השנים? האם חל שינוי משנה לשנה?
2. באיזה יום בשבוע ישנן יותר עבירות חניה? האם יש הבדל בין הרובעים השונים?
3. באילו שעות של היום, בפרקי זמן של שעות, יש יותר עבירות חניה? האם זה תלוי ברובע?
4. הציגו את מספר הדוחות בכל חודש לאורך השנים (2015-2017) ואת קצב הגידול החודשי.
(קצב הגידול החודשי ייחשב כהפרש בין מספר עבירות החניה בחודש מסוים לבין החודש הקודם חלקי מספר עבירות החניה בחודש הקודם).
5. בנו ויזואליזציה המציגה את מספר המצטבר של הדוחות בכל חודש לאורך השנה, YTD, לכל השנים.
6. 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 (Body Types) המעורבים בעבירות החנייה.

הדוח יספק תובנות עסקיות לגבי סוגי הרכבים המעורבים בעבירות החנייה.

1. מהו סוג הרכב, המקבל הכי הרבה דוחות חנייה ב-NY?
2. מהו גובה הקנס הממוצע לכל סוג רכב או לקבוצת סוגי רכב?
3. האם יש צבע רכב דומיננטי?
4. מאיזו מדינה, State, באים רוב הרכבים המעורבים בעבירות חניה?

דוח 3 – ניתוח גאוגרפי של עבירות החנייה של משאיות במנהטן

הערייה מעוניינת להבין היכן ומתי מתבצעות רוב עבירות החניה של משאיות, Delivery Trucks, ברובע מנהטן.

1. הציגו על מפה גיאוגרפית את עבירות החניה שבוצעו עם יכולת סינון של שעות היום והלילה.
(הציעו חלוקה הגיונית של השעות ליום ולילה) (השתמשו בשדה Address עבור זיהוי המקומות על גבי המפה)
האם אפשר לזהות בצורה ברורה אזורים ו/או חלקי יום בעייתיים?
2. מה הן עשרת הרחובות במנהטן שיש בהם הכי הרבה דוחות חניה?
3. מה גובה הקנס הממוצע שמקבלים בעלי רכב אלו? (לא רק העשרה העליונים אלא כלל המשאיות)

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:

1. מה ההכנסה, הרווח ואחוז הרווח (רווח חלקי הכנסות) של עיריית NY מאכיפת חוקי החנייה בכל שנה לכל סוכנות?
2. מה הרובע, borough, הכי רווחי בעיר NYC?
3. מה הם חמשת סוגי הקנס הרווחיים ביותר?

דף המלצות

1. האם העירייה צריכה לשפר את ההסבר בגין עבירות החנייה למי שאינו תושב מדינת NY או לתושבי מדינת NY עצמה?
2. לאיזה קבוצות מיקוד היית ממליץ לבצע פעולות הסברה ובאיזה אזורים?
3. איזה פעילות היא הרווחית ביותר ואיפה היית ממליץ לעירייה להגביר את האכיפה?
4. האם לאורך השנים היו שינויים מהותיים ברווח מדוחות חנייה?

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)	

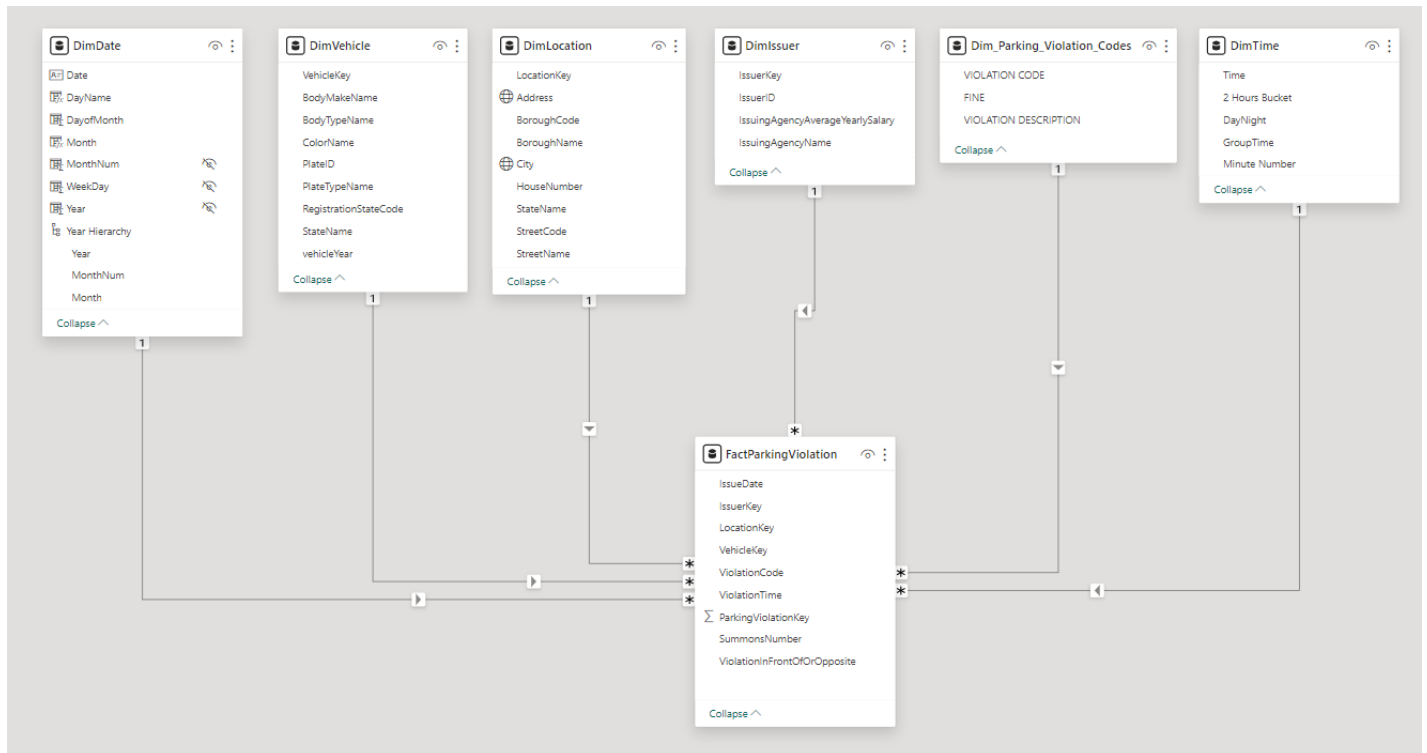
DimIssuer		
Column Name	Condensed Type	
IssuerKey	int	
IssuerID	varchar(20)	
IssuingAgencyCode	varchar(5)	

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)	

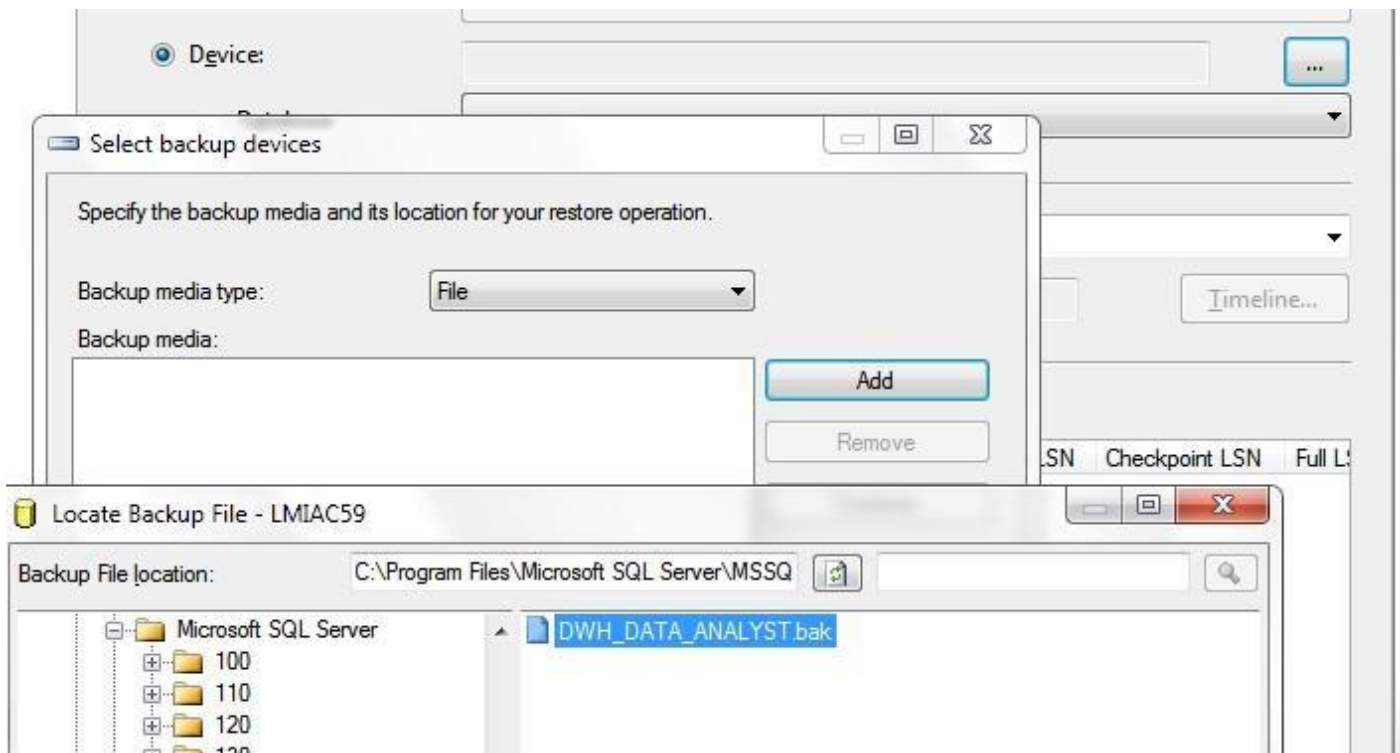
נספח ב' מודל הנתונים ב-Power BI



נספח ג' שחזור בסיס הנתונים לפרויקט

התקנה מקומית

1. לפתוח את SQL Server Management Studio
2. אם בסיס הנתונים DWH_DATA_ANALYST כבר קיים יש למחוק אותו (כפתור ימני על שם בסיס הנתונים ו-delete)
3. ללחוץ על הכפתור הימני של העבר על Databases ולבחור Restore Database...
4. בחלון שנפתח יש לבחור Device וללחוץ על הכפתור 
5. יש ללחוץ על ADD ולבחור את קובץ ה-BAK כפי שמופיע כאן:



6. בסיום יש ללחוץ על OK עד שחוזרים לחלון המקורי ושם ללחוץ שוב OK להפעלת תהליך השחזור.

חיבור לבסיס הנתונים בשרת של נאיה

לפתוח את SQL Server Management Studio ולהתחבר לבסיס הנתונים

http://Hotbox1.ddns.net	שרת
sql	יוזר
Password21345	ססמא
DWH_DATA_ANALYST	בסיס נתונים