**Homework 1 – Business Intelligence**

**40205**

**מרצה: אור יצחק פרץ**

מגישים:

נדב שפינדל 205612468

דור רון 311129936

ענבר מוצרי 315073353

רועי הניגסברג 205387848

**חלק 1 – אוסף נתונים**

סט הנתונים שבחרנו נקרא “Crimes In Boston”, הוא מתייחס לכלל התלונות שהוגשה למשטרה בבוסטון בין השנים 2015 ל2018. העמודות מכילות נתונים אודות התלונה: נתוני תיאור (שם, קטגוריה וכו'), מיקום (קורדינטה, רחוב ומחוז), תאריך (שנה, חודש, יום בשבוע ושעה). בנוסף יש עמודה בוליאנית של האם בוצע ירי וכן עמודה המציינת את חומרת הפשיעה(לפי חלוקה שהFBI קבע למיון חומרה של פשיעות)

**חלק 2 – שאלות מחקר / שאלות עסקיות**

Supervised - בהינתן מיקום, שעה ותאריך האם ניתן לחזות תלונה על פשע הכוללת ירי?

מדדים:

• מהו אחוז הפשעים הכוללים ירי?

• מה אחוז הפשעים הכוללים ירי בכל שנה?

• מה אחוז הפשעים הכוללים ירי לכל מחוז?

• האם ישנה קורלציה בין סוג הפשיעה לבין שימוש בירי

Unsupervised – האם בימי שבת וראשון (סוף שבוע בבוסטון) מתקבלות יותר תלונות חמורות (part one) מאשר בשאר הימים בשבוע ואם כן, מהן תלונות אלו?

מדדים:

• מה סכום התלונות החמורות לכל יום

• מה אחוז התלונות החמורות לכל יום

• מה סוג הפשיעה הנפוצה מתקבלת לכל יום

**חלק 3 – הבנת הנתונים**

1. מדדי פיזור של התכונות ( ממוצע, סטיית תקן, רבעונים):

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | OFFENSE\_CODE | YEAR | MONTH | HOUR | Lat | Long |
| mean | 2317.547 | 2016.560586 | 6.609719 | 13.118205 | 42.214381 | -70.908272 |
| std | 1185.286 | 0.996344 | 3.273691 | 6.294205 | 2.159766 | 3.493618 |
| min | 111 | 2015 | 1 | 0 | -1 | -71.178674 |
| q1 | 1001 | 2016 | 4 | 9 | 42.297442 | -71.097135 |
| median | 2907 | 2017 | 7 | 14 | 42.325583 | -71.077524 |
| q3 | 3201 | 2017 | 9 | 18 | 42.348624 | -71.062467 |
| max | 3831 | 2018 | 12 | 23 | 42.395042 | -1 |

1. תלויות וקשרים (קורלציה):

Calendar

Description automatically generated

1. מדדים סטטיסטיים נוספים: **ניתן לראות בטבלה בסעיף 1**
2. אנטרופיה עבור שאלת הSupervised:

* מיקום (מורכב ממדדי Lat ו Long(קורדינאטה)): nan
* שעה: 12.522096221949665
* תאריך (שנה, חודש, יום בשבוע ושעה) :
  + שנה : 12.673175073653946
  + חודש: 12.534841424683208
  + יום בשבוע: 12.54444965157038

1. מדדי Gini-index וInformation-gain עבור 2 התכונות בעלות האנטרופיה הנמוכה ביותר:

* Gini – Index:
  + שעה: 0.9516347019008494
  + חודש: 0.9145879202677099
* Information- gain: 0.0008875063369635505

1. מסקנות:

* לאור העובדה שהנתונים בסט הנתונים שבחרנו אינם רציפים, אלא קטגוריאליים- לא ניתן להסיק מסקנות מהימנות באמצעות ניתוח סטטיסטי פשוט.