**שיעורי בית 1 – בינה עסקית:**

עדן בן דוד 206957797

אביב רביד 316018175

נועה דלאל 316045335

**חלק 1 – אוסף נתונים:**

בחרנו אוסף נתונים של מחירי דירות, המכיל 3 טבלאות.

**חלק 2 – שאלות מחקר:**

**שאלת מחקר 1: Supervised**

האם דירות בעלות 3 חדרים, אשר שטח המגורים שלהן הינו מעל הממוצע של שטחי המגורים, יעלו מעל הממוצע של מחירי הדירות?

**KPI:**

המדדים שנבדוק בפרויקט הינם:

* זמן מוגדר לביצוע הפרויקט: נרצה לענות על שאלת המחקר בטווח זמן של עד 3 חודשים.
* מדד ממוצע מחירי דירות כללי.
* מדד ממוצע מחירי דירות בעלות 3 חדרים.
* מדד ממוצע שטח מגורים כללי.
* מדד ממוצע שטח מגורים לדירות בעלות 3 חדרים.
* מדד כמות בתים בכל עיר.
* ממוצע מחירים, מחיר מינימלי, מחיר מקסימלי.

**Smart:**

S ספציפי: השאלה ספציפית כיוון שהיא מתמקדת במספר חדרים מסוים, ושטח מגורים אשר עונה על תנאי מסוים.

M מדיד: קיים לנו מידע בבסיס הנתונים על הפרמטרים אותם אנו רוצות לבדוק.

A בר השגה: ניתן לבדוק מדדי חוסר ודאות על כמות החדרים והאם שטח המגורים מעל או מתחת לממוצע, וכך ניתן לחקור את שאלת המחקר.

R רלוונטי: המדדים שבחרנו רלוונטיים לשאלת המחקר כי הם מציגים את הממוצעים איתם נרצה לעבוד ומהם להסיק לגבי שאלת המחקר.

T זמן: במסגרת הזמן של הקורס שהינו 3 חודשים, אנו מעריכות שנצליח לתת חיזוי מתאים ולבצע ניתוח נתונים נכון.

**שאלת מחקר 2: Unsupervised**

מה מאפיין דירות בעלות 3 חדרים ובעלות שטח מגורים מעל הממוצע?

**KPI:**

המדדים שנבדוק בפרויקט הינם:

* + - * מדד הצלחה: במהלך בחינת האלגוריתם נרצה להגיע לכ-80% הצלחה בניבוי המחיר.
* ממוצע החדרים בכל קבוצת סיווג.
* ממוצע מחירי הדירות בכל קבוצת סיווג, מעל או מתחת לממוצע.
* ממוצע גודל שטח המגורים בכל קבוצת סיווג, מעל או מתחת לממוצע.

**Smart:**

S ספציפי: בחרנו להתמקד בחיפוש אחר מאפיינים למחירי בתים שהינם מתחת לממוצע, לכן השאלה מתמקדת רק בטווח מחירים ספציפי.

M מדיד: קיים לנו מידע בבסיס הנתונים בעזרתו נוכל לבצע סיווג מדויק ולענות על שאלת המחקר.

A בר השגה: ניתן לבדוק מדדי חוסר ודאות על הפרמטרים בבסיס הנתונים וכך ניתן לחקור את שאלת המחקר.

R רלוונטי: המדדים שבחרנו רלוונטיים לשאלת המחקר כי הם מציגים את הממוצעים איתם נרצה לעבוד.

T זמן: במסגרת הזמן של הקורס שהינו 3 חודשים, אנו מעריכות שנצליח לתת חיזוי מתאים ולבצע ניתוח נתונים נכון על מנת להבין מי הם המאפיינים של בתים שעלותם מתחת לממוצע.

**חלק 3 – הבנת הנתונים:**

לאחר הניתוחים הסטטיסטים אשר ביצענו הבחנו בכמה דברים:

* ניתן לראות כי רוב העמודות הינן נומריות, ובעלות מגוון רחב של ערכים.
* אין ערכים חסרים.
* הדירה הנמצאת בחציון נחשבת כדירה בעלת 3 חדרי שינה ו 2 חדרי רחצה עם מחיר של כ – 460,943 דולר.
* גודל שטח המגורים הממוצע הוא כ ft² 4600 שהם כ427 m² במונחים שלנו.
* כאשר חיפשנו מהו המחיר המינימלי לדירה קיבלנו מחיר 0 – כמובן שזהו דבר בלתי הגיוני וניתן להבין כי נצטרך לבצע ניקוי לדאטה.
* ניתן לראות ממבחן הקורלציה שיש התאמה יחסית גבוהה בין העמודה sqft\_living לעמודה .sqft\_above קיימת קורלציה חיובית בין מחיר לכמות החדרים, בין מספר חדרי השינה למספר חדרי האמבטיה, קורלציה חיובית בין מספר חדרי השינה לשטח בנוי. בנוסף לכך קיימת קורלציה חיובית בין השנה בה נבנה הבית לבין מספר הקומות שקיימות בו.
* כאשר נסתכל על האנטרופיות נבחין שעבור מחיר וגודל הבית קיבלנו אנטרופיות גבוהות מאוד (9.83, 8.433). מכיוון ששני אלו פרמטרים רציפים זהו לא דבר שכיח למצוא שני בתים בעלי אותו מחיר או אותו גודל. לעומת זאת עבור מספרי חדרי השינה קיבלנו אנטרופיה נמוכה (1.882) מכיוון שהם בדידים ונמצאים בטווח מסוים כאשר המינימום הוא 1 והמקסימום 9. בהנחה והייתה התפלגות שווה בין אפשרויות מספר החדרים היינו מקבלים אנטרופיה מעל 3, ולכן ניתן להבין כי ההסתברויות של האופציות השונות אינן שוות.

לפי הסטטיסטיקות שבדקנו ראינו כי הנתונים יכולים לעזור לנו במענה על שאלות המחקר.