מבוא למדע הנתונים

תרגיל הגשה 1

<u>הנחיות להגשת התרגיל:</u>

(1) יש לממש את המודלים בתוכנת R/Python

(2) התרגיל הוא אישי.

(3) את הפתרונות הסופיים יש להגיש דרך המודל בשתי תיבות הגשה נפרדות:

<u>תרגיל בית 1 – שאלות</u>

יש להגיש דרך המודל, שימו לב שהבדיקה של חלק זה היא ממוחשבת.

<u>תרגיל בית 1- קוד R/Python</u>

בנוסף למענה על השאלות, יש להגיש בנפרד קובץ R/Py מתועד היטב, הקובץ צריך להחיל את הניתוחים שהרצתם לכלל חלקי העבודה. יש לוודא שהקובץ לא מכיל חלקים לא רלוונטים (מודלים/ניתוחים שלא בוצע בהם שימוש) בחלקי העבודה השונים.

(4) אין להתייעץ עם סטודנטים אחרים (עבודות דומות תיפסלנה)!

<u>הבעיה:</u>

בפני קבלן עומדות 2 אופציות לקניית שטח במרכז הארץ לצורך בניית מתחם מגורים חדש. מדובר במתחם גדול שיכול להכיל בתוכו גם שירותים שונים כגון מכולות/מסעדות/גינות וכו' (אותם הקבלן יכול להקים כחלק מהפרויקט). הקבלן מתייעץ אתכם, **<u>מומחי נדל"ן בעלי ניסיון עסקי רלוונטי</u>**, על מנת לקבל המלצות שיעזרו לו להחליט איזה שטח כדאי לו לקנות, וכיצד כדאי לו לתכנן את הדירות במתחם, ואת תכולת המתחם. ברשותכם מסד נתוני דירות שנמכרו במרכז הארץ. מטרת הקבלן, ומטרתכם כיועציו, לתכנן את המתחם **<u>כדי למקסם את הרווח</u>** מעסקת הנדל"ן.

מתחם א – קרוב מאוד לים, רחוק ממרכז העיר (אין חנויות מזון וגינות כלבים בקרבת המתחם).
מתחם ב – רחוק מן הים, קרוב למרכז העיר ולאפשרויות שהיא מציעה (בקרבת המתחם יהיו חנויות מזון וגינות כלבים).

מסד הנתונים שעומד לרשותכם כולל רשומות של 20,932 דירות שנמכרו במרכז הארץ, כאשר:

Price	המחיר בו נמכרה הדירה [שקלים חדשים]
MtrsToBeach	מרחק הדירה מהים [מטרים]
SqMtrs	גודל הדירה [מטרים רבועים]
Age	גיל הדירה [שנים]
NumStores	מספר חנויות המזון ברדיוס של 100 מטר מהדירה
DogParkInd	האם יש גינת כלבים ברדיוס של 100 מטר מהדירה {1 = כן, 0 = לא}
SchoolScores	ממוצע ציוני הבגרות באזור הרישום של הדירה [0-100]

שלב א – מטרת המחקר, הבנת הבעיה והכרת המסד:

1. האם מטרת המחקר היא מפוקחת או לא מפוקחת?
2. האם תריצו מודל הסברתי או מודל חיזוי? מדוע? **שימו לב**, כמומחי נדל"ן מנוסים אתם רשאים להניח (בהסתמך על ההיכרות שלכם עם השוק) כי קשרים בין המשתנים המסבירים למשתנה התלוי מייצגים סיבתיות.
3. בחנו את המשתנים השונים במסד הנתונים:
   א. אילו משתנים אינם רלוונטים לקבלן? מדוע?
   ב. אילו משתנים רלוונטים לבחירת המתחם (א או ב)?
   ג. לאחר שמיקום המתחם כבר נקבע (א או ב), על אילו משתנים יכול הקבלן להשפיע? הסבירו.
   ד. אילו משתנים רלוונטים לתכנון הדירות עצמן?
   ה. הציעו שני משתנים נוספים שלא נמצאים במסד שהיו עוזרים לקבלן לתכנן את הדירות במתחם באופן מיטבי. הסבירו.
   ו. הציעו שני משתנים נוספים שלא נמצאים במסד שהיו עוזרים לקבלן להחליט על מיקום המתחם (א או ב). הסבירו.

שלב ב – ויזואליזציה

1. צרו היסטוגרמה של משתנה המחיר. תארו את מאפייני ההתפלגות (התייחסו לחציון, ממוצע, מדד פיזור, וצורתה הכללית של ההתפלגות, האם דומה או שונה מהתפלגות נורמלית).
2. צרו גרף המתאר את הקשר בין NumStores ל- DogParkInd. האם יש קשר בין משתנים אלו?

שלב ג – אמידת מודלים

1. בנו שני מודלי רגרסיה לינארית לפי ההנחיות הבאות:
   א. מודל האומד את הקשר בין מחיר הדירה (Price) ל**כל** המשתנים המסבירים שבמסד.
   למודל זה נתייחס כ-m1
   ב. מודל האומד את הקשר בין מחיר הדירה (Price) למשתנים המסבירים MtrsToBeach, SqMtrs, Age.
   למודל זה נתייחס כ-m2

2. מלאו את הטבלה הבאה לפי פלטי הרגרסיות (השאירו דיוק של 3 ספרות אחרי הנקודה העשרונית):

	m1	m2
אחוז שונות מוסברת רגילה (R^2)		
משתנים מובהקים ברמת מובהקות של 0.05		
האם המודל כולו מובהק?	א. מובהק ב5% ב. לא מובהק ב5%	א. מובהק ב5% ב. לא מובהק ב5%
לפי המודל, ההבדל בין מחירן של שתי דירות, שהפרש הגדלים ביניהן הוא 18 מ״ר (וכל יתר המאפיינים זהים) הינו	א. בין 0-20 אלף שח. ב. בין 20-100 אלף שח ג. בין 100-200 אלף שח ד. בין 200-500 אלף שח ה. למעלה מ500 אלף שח	א. בין 0-20 אלף שח. ב. בין 20-100 אלף שח ג. בין 100-200 אלף שח ד. בין 200-500 אלף שח ה. למעלה מ500 אלף שח

3. אם היינו מורידים ממודל m1 את משתנה NumStores, מה היה קורה למקדם של DogParkInd? קטן/ גדל/ לא משתנה (ודאו כי אתם מבינים מדוע!)