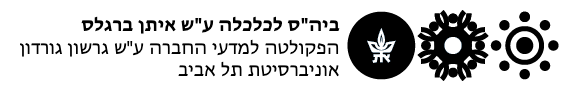
**מטלה ראשונה - לספר סיפור בעזרת נתונים**

**כלכלה בעולם ה Big Data**

**ד״ר רועי ששון; אופיר בצר**

להגיש לכל המאוחר עד:

**23:59 2022-05-22**

בשל טווח הזמן הארוך לא יתקבלו הגשות באיחור.

**המטרות המרכזיות של המטלה הראשונה בקורס הן:**

1. להתנסות בתכנות וניתוח-נתונים בשפת R.
2. לתרגל את תהליך ה EDA עם נתונים מורכבים.
3. לתרגל את תקשור הממצאים מהאנליזה שלכם.

פרטים חשובים על המטלה:

1. יש לבצע את המטלה **בזוגות**.
   1. רק אחד מחברי כל קבוצה צריך להגיש את המטלה במודל.
   2. נא לציין באופן ברור מי הם חברי הקבוצה בתוך הדוח ובשם של הקובץ של הדוח באופן הבא:

EinBDW22\_A1\_{name #1}\_{name #2}.pdf

לדוגמא:

EinBDW22\_A1\_roy\_sasson\_ophir\_betser.pdf

1. משקל המטלה הוא **20%** מציון הקורס.
2. את כל הניתוחים יש לבצע ב **R בלבד**.

יש להגיש:

1. דו״ח מפורט ובו התשובות שלכם על השאלות, בצירוף גרפים וטבלאות רלוונטים. על הדו״ח להיות ממוקד - אין להגיש יותר מ 6 עמודים, וינתן משקל על אסתטיות של הגרפים ו/או הטבלאות המצורפים בו.
   1. יש לצרף לכל אחד מהגרפים כותרות המסבירות את תוכנו.
   2. יש לוודא שהגרפים שאתם מצרפים קריאים, וברורים, מובנים ואסתטיים.
   3. אין צורך לכתוב הרבה מלל, אלא לענות באופן קצר וממוקד.
   4. פורמט הקובץ צריך להיות **pdf**.
   5. שפת הדוח צריכה להיות **עברית**, גופן דוד 12, רווח בין השורות 1.5.
2. וידאו קצר, בן לא יותר מ 5 דקות שיכיל סיכום של המסקנות העיקריות שלכם מהניתוח, בדגש על ההמלצות העסקיות (שאלה #4) שעלו מתהליך המחקר.
   1. עדיף לצרף את הוידאו בקישור לגוגל דרייב (לוודא שיש אליו גישה) מאשר להעלות למודל, מכיוון שלעיתים יש בעיות טכניות עם קבצים גדולים במודל.
3. **אין צורך** לצרף את הסקריפט שלכם (אבל אפשר, והוא לא יהיה חלק מהציון).

**על סט הנתונים:**

במטלה זו תבצעו ניתוח של נתוני [Airbnb](https://en.wikipedia.org/wiki/Airbnb) בבוסטון.

הנתונים נמצאים בשתי טבלאות:

1. הטבלה הראשונה, [listings.csv](https://github.com/ophirbetser/EinBDW_github/blob/master/datasets/listings_clean.csv), מכילה מידע כללי על כל נכס (בערך 3k שורות, כל שורה מייצגת נכס ייחודי).
2. הטבלה השניה, [calendar.csv](https://github.com/ophirbetser/EinBDW_github/blob/master/datasets/calendar_clean.csv), מכילה מידע על מה שקרה לכל נכס בפועל ברמה היומית בתקופה של בערך שנה (בערך 1m שורות, כל שורה מייצגת נכס בתאריך מסויים).
3. פירוט נוסף על הנתונים ניתן למצוא פה: [Airbnb\_code\_book](https://docs.google.com/spreadsheets/d/1jEaGRw9v9e3_kTBJcoExh06aNXq385mfMyJ8ztw_czk/edit?usp=sharing)

**שאלות מחקר: (חלקים 1 עד 4)**

חלק 1 - סטטיסטיקה תיאורית 35%)

סטטיסטיקה תיאורית היא חלק חשוב מאוד בכל עבודת ניתוח נתונים. הסטטיסטיקה התיאורית מאפשרת לנו להבין לעומק את הנתונים שלנו, לזהות קשרים חשובים בין משתנים, לאתר בעיות בנתונים ועוד. סטטיסטיקה תיאורית איכותית היא הבסיס לבניית מודלים טובים בהמשך, והיא קריטית כאשר אנחנו רוצים לתקשר את העבודה שלנו לאחרים.

צרו סטטיסטיקה תיאורית עבור הנתונים שנמצאים בטבלה listings בלבד: (בחרו מה להציג בהקשר של חלק 4 - ההמלצה העסקית)

1. צרו 3 גרפים שונים שמזהים תכונה חשובה או חשודה על 3 משתנים (סטטיסטיקה עבור משתנה יחיד). כתבו פרשנות שלכם למה שרואים בגרף, ומה החשיבות של זה.
2. צרו 3 גרפים שונים שמזהים קשרים מעניינים וחשובים בין המשתנים (סטטיסטיקה עבור 2 משתנים או יותר). כתבו פרשנות שלכם למה שרואים בגרף, ומה החשיבות של זה.

חלק 2 - מטריקות (30%)

מטריקות עסקיות הן כלי מדידה אובייקטיבים חשובים על מנת לבחון אספקטים שונים בביצועים ובמצב העסק. המטריקות מאפשרות לנו להבין מה מצב העסק בהווה, ולראות את השינוי בהן לאורך זמן.

דוגמא למטריקות: פייסבוק - מספר משתמשים חדשים בכל שבוע; משרד האוצר - הכנסות ממס בחודש לעומת אותו החודש בשנה שעברה; וויז - אחוז הפער בין זמן הגעה שהוצהר בתחילת הנסיעה לעומת זמן הגעה בפועל.

עבור המימדים הבאים של AirBNB (כלומר, צרו בסך הכל 3 ניתוחים)

1. צמיחה
2. צמיחה לפי שכונה
3. שינוי במחיר

בצעו את השלבים הבאים:

1. הגדירו בבירור מטריקה עסקית (אתם חופשים להגדיר מה שנראה לכן מעניין, כל עוד הוא מתקשר לנושא)
2. כתבו מה לדעתכם המטריקה שהגדרתם תופסת ומה היא מפספסת (אף מטריקה בודדה לא תופסת את התמונה המלאה, וזה בסדר. לכן משתמשים במספר רב של מטריקות)
3. אמדו את המטריקה שהגדרתם בעזרת הנתונים
4. חוו דעה על מצב העסק ביחס לתוצאות שקיבלתם, וחוו דעה על הגדרת המטריקה שלכם ביחס לאיכות הנתונים שיש בידכם (כלומר האם בעזרת נתונים עשירים יותר הייתם יכולים לאמוד את המדד באופן מדויק יותר?)

חלק 3 - זיהוי ערכים חריגים (20%)

הנתונים שאנו עובדים איתם אף פעם לא מושלמים. התמודדות עם ערכים חסרים וחריגים היא חלק חשוב בעבודה עם נתונים. ערכים חריגים הנובעים בשל חריגות בהזנת הנתונים יעוותו את האנליזה שלנו. ערכים חריגים הנובעים מהתנהגות חריגה אמיתית במציאות יכולים להיות חשובים מאוד לאנליזה שלנו, ונלמד מהם רבות. בכל מקרה, את הערכים החריגים אנחנו יכולים למתן, לשפץ או למחוק לחלוטין אם נראה לנו שזה כדאי.

1. זהו **טווח זמנים** שיש בו התנהגות חריגה של הנתונים (ערך מספרי, או מספר ערכים מספריים שמתנהגים מוזר על פני תקופה של מספר ימים רצופים ביחס לתקופה הקודמת או העוקבת).
2. צרפו גרף שמבסס את טענתכם.
3. כתבו מה הסיבה שלדעתכם גרמה לקבלת הערכים החריגים בתקופה ההיא? (טעות בהזנת הנתונים, אירוע חריג במציאות וכו')
4. מזווית של ניתוח נתונים, האם ואיך כדאי לטפל בערכים אלו?

בנוסף בצעו את אותו התהליך על 2 העמודות הבאות: (ללא התחשבות במימד הזמן)

1. מספר מיטות.
2. מחיר (ניתן לבחור עמודה מאחת משתי הטבלאות).

חלק 4 - המלצה עסקית (15%)

צוות מנהלי השיווק של Airbnb בבוסטון מעוניין להגדיל את ההכנסות ב 100% בשנתיים הבאות.

מה הייתם מייעצים להם לעשות? בהתבסס על ניתוח הנתונים שביצעתם, וההבנה והאינטואיציה הכלכלית שלכם או כל דבר אחר. (אין ״נכון״ או ״לא נכון״ מוחלט בחלק הזה…)

בחלק זה תוכלו לצרף ניתוחים נוספים שנראים לכם רלוונטים וגיליתם במהלך חקירת וניתוח הנתונים. אתם יכולים לצרף גרפים, או להגדיר מטריקות כדי לחזק את הטיעונים וההמלצות שלכם.

יש לדון בתשובה הזאת בוידאו המצורף.

על הוידאו להיות מין סיכום ממוקד של הניתוח שלכם בו אתם מציגים את הגילויים המרכזיים שלכם, ואת ההמלצות העסקיות הנגזרות מהם.