**AirBnb Data Analysis**

בתרגיל זה תבצעו עיבוד של הנתונים לפי רגרסייה לינארית ל**חיזוי המחיר** כולל שינוי קנה מידה וone hot encoding

תוך כדי עיבוד הנתונים אתם צפויים להתקל בעמודה עם מידע חסר.

בתרגיל תנסו 3 שיטות להתמודד עם המידע החסר

לפני שתתחילו נסו לחשוב איזו שיטה לדעתכם תניב תוצאות מיטביות

לבסוף תצרכו לנתח איזו שיטה הייתה טובה ביותר ולמה לדעתכם זו היתה התוצאה

שלוש השיטות יהיו:

1. החלפת המידע החסר

2. מחיקת העמודה שיש בה מידע חסר

3. מחיקת השרות שיש בהן מידע חסר בעמודה הזו

**התרגיל:**

**א. הכרת המידע והכנתו**

1. תטענו את הנתונים בקובץ csv המצורף למשתנה df
2. סקרו את הנתונים שקיבלתם
3. בדקו איזה עמודות הן מועמדת טובות לטיפוס קטגוריה ושנו אותן בהתאם
   1. רמז( העמודות שאתם צריכים לבחור בהן : ['neighbourhood\_group','room\_type','bedrooms','baths'])
4. בעזרת one hot encoder הפכו את העמודות הללו לעמודות מקודדות
   1. הקפידו על drop first
   2. שרשרו את העמודות בעזרת concate לdf המקורי
   3. לא לשכוח למחוק את העמודות המקוריות שעשיתם איתן encode
5. מחקו המ df את העמודות שאינן מספריות וגם לא מתאימות להיות קטגוריות (ולכן לא יוכלו לעזור לנו ברגרסיה לינארית):
6. ['id', 'name', 'host\_id', 'host\_name', 'neighbourhood', 'last\_review', 'license']
7. שימו לב שעמודת rating אינה מספרית. שנו אותה לעמודה מספרית
8. ברוכים הבאים לבעיה- המון שורות הן למעשה עם raiting חסר. בינתיים **המירו**  אותם ל np.na וחשבו **איזה אחוז** מהנתונים חסר.

**ב. גישה ראשונה החלפת הנתונים החסרים ב חציון**

1. צרו העתק של df וקרא לו df\_median
2. **החליפו את הנתונים החסרים בעמודת הרייטינג לחציון של הרייטינג**
3. חלקו את df\_median לנתוני אימון ומבחן. שימו לב שעמודת המחיר היא העמודה שאנחנו מעוניינים לחזות
4. בעזרת MinMaxScaler הפכו את כל הסקאלות **של הפיצ'רים** לאחידות.
5. אמנו מודל רגרסיה לינארית בעזרת נתוני האימון
6. בצעו תחזית בעזרת נתוני המבחן
7. חשבו את RSME ושמרו את העלות של המודל במשתנה rsme\_median\_rating

ג. **גישה שניה: מחיקת עמודת הרייטינג**

1. צרו העתק של df וקרא לו df\_no\_raiting
2. **מחקו מ df\_no\_rating את עמודת הרייטינג**
3. חלקו את df\_no\_rating לנתוני אימון ומבחן. שימו לב שעמודת המחיר היא העמודה שאנחנו מעוניינים לחזות
4. בעזרת MinMaxScaler הפכו את כל הסקאלות של הפיצ'רים לאחידות.
5. אמנו מודל רגרסיה לינארית בעזרת נתוני האימון
6. בצעו תחזית בעזרת נתוני המבחן
7. חשבו את RSME ושמרו את העלות של המודל במשתנה rsme\_no\_rating\_column

**ד. גישה שלישית: מחיקת עמודת הרייטינג**

1. צרו העתק של df וקרא לו df\_no\_rows

2. **מחקו את השורות בהן נתוני הרייטינג חסרים**

3. חלקו את df\_no\_rowsלנתוני אימון ומבחן. שימו לב שעמודת המחיר היא העמודה שאנחנו מעוניינים לחזות

4. בעזרת MinMaxScaler הפכו את כל הסקאלות של הפיצ'רים לאחידות.

5. אמנו מודל רגרסיה לינארית בעזרת נתוני האימון

6. בצעו תחזית בעזרת נתוני המבחן

7. חשבו את RSME ושמרו את העלות של המודל במשתנה rsme\_no\_rows

ה. **רגע האמת**

תשוו את התוצאות.

איזה מודל היה הכי יעיל?

האם זה מה שציפיתם שיקרה?

מה אפשר ללמוד מזה?