תרגיל מסכם – ניתוח ועיבוד מידע.

לצורך התרגיל ניתן לעבוד גם בסביבת און ליין ללא צורך בהתקנות:
https://jupyter.org/try

בתרגיל זה נתרגל טעינת קובץ מידע, ניקוי הקובץ והפעלת מודל רגרסיה ליניארית.

1. הכרת המידע

הקובץ restaurant_data.csv מכיל מידע על מסעדות שונות.
פתחו את הקובץ והתבוננו במידע.
הקובץ מכיל את העמודות הבאות:
url – כתובת אתר המסעדה
address – כתובת המסעדה
name – שם המסעדה
online_order – האם המסעדה מאפשרת הזמנות און ליין.
book_table – האם המסעדה מאפשרת הזמנת שולחן.
rate – דירוג המסעדה (דירוג משוקלל של כל הלקוחות שדירגו).
votes – מספר הלקוחות שדירגו את המסעדה.
phone – טלפון המסעדה.
rest_type – סוג המסעדה.
approx_cost(for two people) – מחיר משוער לסעודה של שני אנשים במסעדה.
reviews_list – רשימה שמכילה את הביקורות המילוליות שסועדים נתנו למסעדה.
listed_in(type) – לאיזו קטגוריה שייכת המסעדה.
listed_in(city) – באיזו עיר רשומה המסעדה.

2. ניקוי המידע

טענו את הקובץ לתוך dataframe ובצעו את משימות הניקוי הבאות על הקובץ:

2.1 מחקו את העמודות url, address, name, phone מהקובץ.
2.2 מחקו שורות שמכילות תאים ריקים.
2.3 מחקו שורות כפולות.
2.4 המירו את הערכים בעמודות online_order ו-book_table מ-"yes" ו-"no" ל-True ו-False.
2.5 עמודת הדירוג מציגה את הדירוג בפורמט "4.5/5" כלומר, הממוצע מתוך 5. שנו את הערכים כך שיכילו רק את הממוצע כמספר (כלומר במקום 4.5/5 יופיע 4.5).
2.6 הפכו את העמודות listed_in(type) ו-listed_in(city) לעמודות קטגוריאליות, כלומר במקום מחרוזות יופיעו מספרים שמייצגים את הערכים השונים.
2.7 מחקו את העמודה reviews_list ובמקומה צרו עמודה שתכיל את מספר הביקורות שהיו למסעדה (כלומר אורך רשימת הביקורות שמופיעה בעמודה המקורית).
2.8 שנו את שמות העמודות listed_in(type) ו-listed_in(city) ל- listed_type ו-listed_city.

3. ויזואליזציה של המידע

צרו את הדיאגרמות הבאות:

3.1 צרו דיאגרמת פיזור (scatter plot) שתציג את הקשר בין המחיר הממוצע לסעודה לבין דירוג המסעדה.

3.2 צרו דיאגרמת מקלות (histograma) שתראה כמה מסעדות יש מכל סוג (listed_type).

3.3 צרו תרשים פאי שיראה כמה מסעדות יש מכל עיר (listed_city).

4. מודל רגרסיה ליניארית

נבנה מודל שינסה לחזות את דירוג המסעדה לפי שאר העמודות.

צרו מודל רגרסיה ליניארית והריצו אותו על המידע.

הריצו הערכה למודל ומצאו מה אחוז ההצלחה של המודל.

הנחיות

- הריצו כל סעיף בתא נפרד ב-notebook, או שכתבו פונקציה נפרדת לכל סעיף אם אתם עובדים ב-pycharm. לפני כל תא/פונקציה רשמו את מספר הסעיף ותיאור קצר (למשל – "בסעיף זה נבצע מחיקה של שורות כפולות").
- כתבו בראש הקובץ את שמכם המלא ומספר ת.ז. שלכם.
- הגישו את הקובץ למייל שלי shay.tavor@gmail.com

בהצלחה!