



Data science - תרגיל בית 1

מועד הגשה: 24.3 בשעה 23:55

הנחיות כלליות להגשת תרגילי הבית:

- הגשת התרגילים תתבצע באמצעות אתר הקורס בלבד.
- יש להגיש את התרגיל לתיבת ההגשה המתאימה, בהתאם לפורמט הבא: HW#_ID1_ID2, כאשר התו # מציין את מספר התרגיל, ID1 מציין את תעודת הזהות של הסטודנט הראשון, ID2 מציין את תעודת הזהות של הסטודנט השני.
- לדוגמה: HW1_000000000_111111111.py זהו קובץ בודד שפותר את תרגיל בית 1.
- בנוסף, יש לכתוב כהערה (בתוך הקוד) בראש כל קובץ את השמות המלאים (שם פרטי ושם משפחה) בלועזית ואת תעודת הזהות של כל אחד מהשותפים, כאשר הפרטים של כל אחד רשומים בשורה נפרדת.
- ההגשה מתבצעת באמצעות קבוצות ההגשה.
- תרגיל אשר יוגש באיחור לא ייבדק, למעט אישורים מיוחדים כגון: מילואים, אשפוז וכו' המתקבלים לפני מועד ההגשה הרשמי.
- יש לכתוב הסברים ומענה מילולי על השאלות בגוף הקובץ בעזרת הערות ובאנגלית בלבד.
- יש להפריד בין פתרונות של כל שאלה בצורה הבאה. לפני המענה לשאלה מספר 1 כתבו בהערה Q1 וכן הלאה. במידה והשאלה מכילה מספר סעיפים (נניח 2 סעיפים) יש לכתוב בהערה Q1a לפני פתרון סעיף a בשאלה 1, ו-Q1b לפני פתרון סעיף b בשאלה 1 וכן הלאה.

בהצלחה!

1) צרו שני וקטורים (בשורת קוד אחת):

- a) v1: from 1 to 100 (100 included) with spacing of 10
- b) v2: from 1 to 100 (100 included) with 8 elements evenly spaced
- c) Reshape v2 to matrix form of [2X4]

הקובץ customerData.csv מכיל אינפורמציה על לקוחות של חברת ביטוח בארה"ב. יש לטעון את הנתונים הקיימים בקובץ לתוכנית שלכם ולענות על השאלות הבאות.

2) תחילה הריצו פקודת describe וסקרו את הדאטה (בדקו טווחים של המשתנים השונים ונסו להבחין בין סוגי משתנים).

- a) הדפיסו בשורה אחת את ממדי הדאטה-פריים (שורות ועמודות).
- b) קבעו עבור כל עמודה האם היא קטגורית או נומרית.
- c) הסבירו מדוע המשתנה "custid" אינו רלוונטי, כלומר אינו יכול לעזור לנו בלמידת מידע חדש.

3) צרו דאטה-פריים חדש מ customerData, המכיל רק את העמודות הזוגיות ושורות שהן כפולות של 10.

4) בצעו את הפקודות הבאות.

- a) הדפיסו את תוצאת הפקודה df.shape, שימו לב שטיפוס התוצאה הינו tuple.
- b) הדפיסו את תוצאת הפקודה df.size.
- c) מה ההבדל בין 2 התוצאות של 2 הפקודות הנ"ל? האם ניתן לחשב את תוצאת df.size באמצעות תוצאת df.shape? במידה וכן הראו כיצד, אחרת הסבירו מדוע לא.

5) צרו דאטה-פריים חדש המכיל את כל האינפורמציה לגבי הלקוחות בתחום הגילאים בין 38 ל 50 (כולל).

- a) צרו דאטה-פריים חדש המכיל רק את העמודות הנומריות עבור הלקוחות המבוגרים מ 50.
- b) חזרו על אותה משימה רק בשימוש אינדקסים מספריים בלבד (ללא שימוש בשמות העמודות).

7) צרו דאטה-פריים חדש מתוך 6 המכיל אך ורק את עמודות הגילאים ואת 100 הלקוחות הראשונים בה. שימו לב מומלץ לוודא שהתוצאה חוזרת כדאטה פריים.

8) הדפיסו את מספרי הזיהוי של הלקוחות שהסטטוס המשפחתי שלהם הוא Married או Divorced/Separated וגילם קטן מ 18. יש לענות בשורת קוד אחת.

9) על הסעיפים הבאים יש לענות בשורת קוד אחת :

- a) הדפיסו את הגיל הממוצע של הלקוחות שהכנסתם גדולה מ 16,000 והם גרים במדינת וושינגטון.
- b) מהו הגיל המקסימלי בקבוצה זו? (הדפיסו את התשובה) נכון הוא קצת זקן..
- c) מהי ההכנסה המינימלית בקבוצה זו? (הדפיסו את התשובה)
- d) הדפיסו את מספר הלקוחות המקיימים את התנאי.

10) קבצו את הדאטה-פריים לפי המשתנה sex ו housing_type וענו :

- a) איזו צורת מגורים הכי פופולארית אצל הנשים?
- b) איזו צורת מגורים הכי פופולארית אצל הגברים?