ת.ז. _____

מבחן מועד א' במבוא לניתוח נתונים בפייתון (096202) סמסטר אביב 2019

מרצה: עפרה עמיר, מתרגלים: רפי שללה, שרון הירש, בודק: תום בר

1. לרשותכם **שעתיים וחצי** לפתור את הבחינה.
2. הבחינה היא עם **חומר סגור**. אסור להשתמש בכל חומר עזר.
3. הבחינה כוללת **12 עמודים** ובהם **11 שאלות**. יש לענות על כל השאלות (השאלה האחרונה היא שאלת בונוס). הניקוד של כל שאלה מופיע לצידה.
4. **את התשובות יש לכתוב רק על גבי טופס הבחינה.**
5. על טפסי שאלות ותשובות הבחינה יש לכתוב רק מספר תעודת זהות (ולא שם). חובה לכתוב מספר ת.ז. על כל דפי הטופס והמחברת.
6. אין להפריד את דפי טופס הבחינה.
7. חובה להחזיר בסיום הבחינה את טופס שאלות הבחינה עם כל המחברות בהן השתמשתם. שימו לב: אי החזרה של טופס שאלות הבחינה, או המחברות תגרור כישלון מידי בבחינה.

בהצלחה!

ת.ז. _____

שאלה 1 (12 נק')

על כל אחד מההיגדים הבאים ציינו אם הוא נכון או לא נכון והסבירו. **תשובה ללא הסבר לא תקבל ניקוד.**

1. נניח שהמדגם שלנו כולל את הדגימות הבאות {2, 5, 6, 6}. בדגימת בוטסטראפ ניתן לקבל את הדגימה {6, 6, 6, 2, 5}. נכון/לא נכון. הסבר:

_____

_____

2. הגדלת מספר הקלאסטרים ב k-means תמיד תוביל לירידה במדד של המרחק הממוצע בין דגימות למרכז הקלאסטר שלהן. נכון/לא נכון. הסבר:

_____

_____

3. ערך החציון של משתנה יותר רגיש לערכים קיצוניים מאשר הערך הממוצע של המשתנה.
נכון/לא נכון. הסבר:

_____

_____

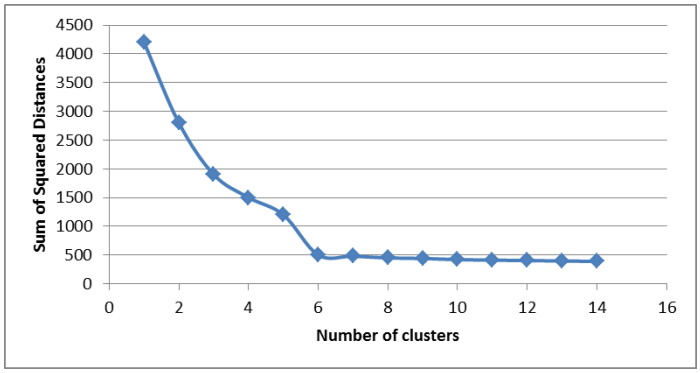
4. אם p-value=0.05, המשמעות היא שיש הסתברות של 5% שהשערת האפס נכונה.
נכון/לא נכון. הסבר:

_____

_____

שאלה 2 (5 נק')

הרצתם את אלגוריתם k-means עם ערכים שונים של k וקיבלתם את הגרף הבא:



באיזה ערך של K תבחרו? הסבירו.

_____

_____

_____

ת.ז. _____

שאלה 3 (18 נק')

סטודנטים רצו להבין את התפלגות מחירי הדירות להשכרה בAirBnB בניו יורק. לצורך כך הם כתבו קוד שהוריד נתונים על 1000 דירות מהאתר. טבלת הנתונים נראית כך (מודפסות 5 השורות הראשונות):

	id	name	host_id	area	room_type	price
0	2454	superCondo	2688	Manhattan	Entire home/apt	137
1	2539	Clean & quiet apt home by the park	2787	Brooklyn	Private room	149
2	2595	Skylit Midtown Castle	2845	Manhattan	Entire home/apt	225
3	3330	++ Brooklyn Penthouse Guestroom ++	4177	Brooklyn	Private room	70
4	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Manhattan	Private room	150

הסטודנטים רצו לאפיין את חציון מחירי הדירות. **במדגם, ערכו של חציון מחיר הדירות הוא 108 דולר ללילה.** הסטודנטים כתבו את הקוד הבא כדי לחשב את רווח הסמך לערכו של החציון:

```
def bootstrap_median(original_sample, column_name, num_replications):
    original_sample_size = original_sample.shape[0]
    medians = np.zeros(num_replications)
    for i in range(num_replications):
        bootstrap_sample = original_sample.sample(original_sample_size, replace=True)
        resampled_median = bootstrap_sample[column_name].quantile(0.5, interpolation='higher')
        medians[i] = resampled_median

    return medians

medians_bootstrapped = bootstrap_median(airbnb_df, 'price', 5000)
ax = sns.distplot(medians_bootstrapped, kde=False);
```

(א) [4 נק'] מה הטעות בתהליך הדגימה בקוד שנכתב? ציינו את מספר השורה, את הבעיה ומדוע זו בעיה [מדובר בבעיה עקרונית בתהליך הדגימה, לא בשגיאת syntax שתגרום לקוד לא לרוץ].

_____

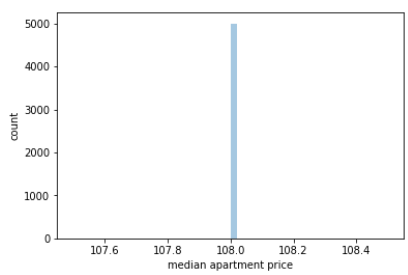
_____

_____

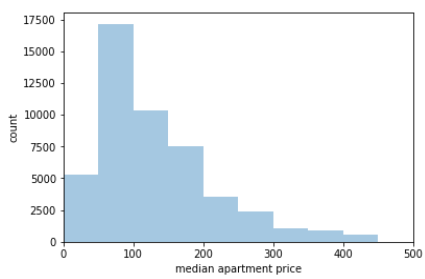
_____

(ב) [4 נק'] איזה מהגרפים הבאים קיבלו הסטודנטים כשהריצו את הקוד? הסבירו.

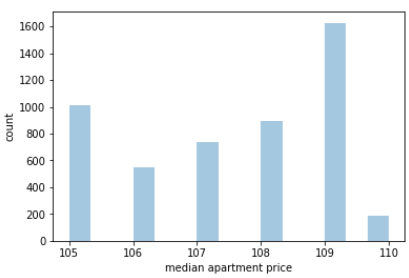
גרף א':



גרף ב':



גרף ג':



_____

_____

_____

_____

(ג) [5 נק'] אחרי שתיקנו את הטעות בקוד, הסטודנטים רצו לבדוק אם יש הבדל במחירים בין חציון המחירים של דירות במנהטן ודירות בברוקלין. נסחו את השערת האפס וההשערה החלופית של הסטודנטים, וציינו מה יהיה סטטיסטי המבחן.

(ד) [5 נק'] בנוסף, הסטודנטים רצו לבדוק אם משכירי דירות במנהטן ובברוקלין משתמשים במילים דומות או שונות בשם שהם נותנים לדירה (עמודת "name"). הציעו שני מאפיינים (פיצ'רים) בהם אפשר להשתמש כדי לענות על השאלה הזו.

שאלה 4 (16 נק')

אתם חושדים שספקית האינטרנט בה אתם משתמשים אינה עומדת בהתחייבות שלה לגבי מהירות הגלישה המובטחת למשתמשים, ומעוניינים לבדוק האם הטענה שלכם נכונה.

(א) [4 נק'] נסחו את השערת האפס וההשערה החלופית, ומה יהיה סטטיסטי המבחן.

(ב) [6 נק'] הציעו דרך לאסוף נתונים כדי לבחון את השערת המחקר. נסחו בבירור מה האוכלוסיה הכללית שמעניינת אתכם, אילו נתונים תכללו במדגם שלכם, ובאיזו דרך תאספו את הנתונים. אתם יכולים להניח שיש אפליקציה למדידת מהירות גלישה.

(ג) [6 נק'] ציינו הטיה אחת מסוג הטיית בחירה (selection bias) והטיה אחת מסוג הטיית מדידה (measurement bias) שעלולות להיות בנתונים שתאספרו וכיצד הן ישפיעו על הניתוח שתבצעו.

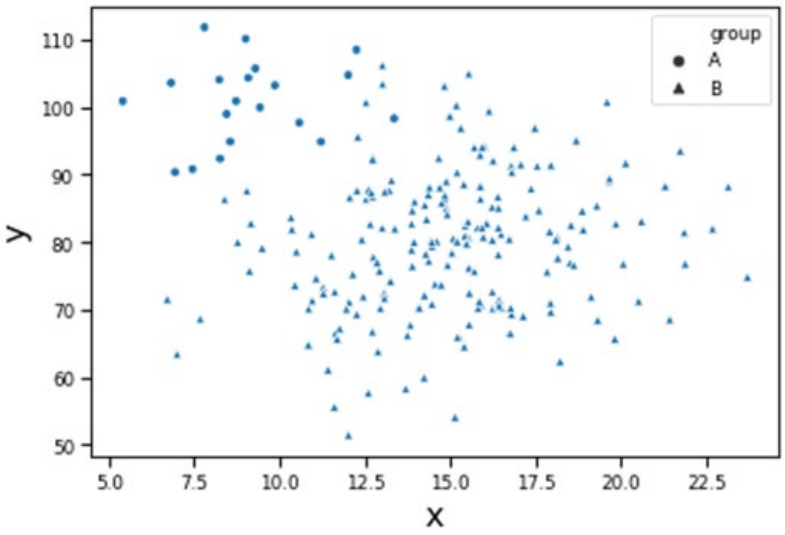
ת.ז. _____

שאלה 5 (12 נק')

בידיכם טבלת נתונים, כאשר לכל רשומה בנתונים יש שלושה מאפיינים: X (מספר רציף), Y (מספר רציף) וgroup (קבוצה, או A או B). כך נראות 5 השורות הראשונות מהטבלה:

	x	y	group
0	5.407682	100.880892	A
1	6.822332	103.594943	A
2	8.237156	104.056405	A
3	12.255190	108.485141	A
4	13.356959	98.316662	A

ברצונכם לאמן מסווג שיחליט אם דוגמה חדשה שייכת לקבוצה A או לקבוצה B על פי ערכי X וY. בתור שלב ראשון, ציירתם גרף אשר מראה את ערכי X וY של כל הנקודות מהטבלה (קבוצה A בעיגולים, קבוצה B במשולשים):



(א) [4 נק'] איזו בעיה קיימת בנתונים אשר עשויה להקשות על הסיווג?

_____

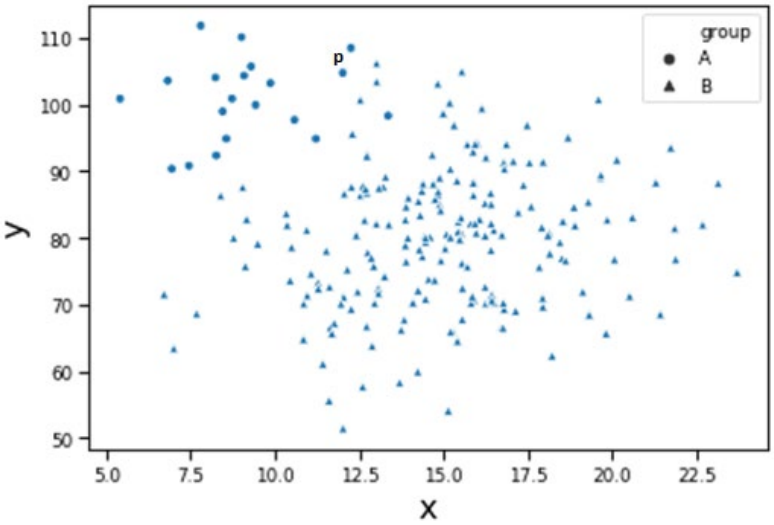
_____

_____

_____

_____

(ב) [4 נק'] נניח שאתם מנסים לסווג נקודה חדשה, המסומנת בP על הגרף (y=108, x=12):



לאיזו קבוצה תסווג הנקודה אם נשתמש בk=2?

_____

_____

לאיזו קבוצה תסווג הנקודה אם נשתמש בk=5?

_____

_____

(ג) [4 נק'] באופן כללי (לא רק כלפי הנקודה P), האם הייתם ממליצים להשתמש בערכי K נמוכים או גבוהים במקרה זה? הסבירו, וציינו באיזה מדדי ביצועים הייתם משתמשים כדי לבחון אם התשובה שלכם נכונה.

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

ת.ז. _____

שאלה 6 (6 נק')

באוקטובר 2016 התפרסמה כתבה בעיתון המקומי של סיאטל שטענה שהעירייה בזבזה כסף מכיוון שהושקעו 74 מיליון דולר בהרחבת אחד הכבישים הראשיים, אבל החיסכון בזמן הנסיעה לאחר הרחבת הכביש הוא רק 2 שניות בממוצע. ציינו בעיה שקיימת בטיעון של הכתבה. הסבירו את תשובתכם.


SOUTH LAKE UNION | NEWS | SEATTLE CONSTRUCTION NEWS

$74 million later, Mercer Mess has improved by two seconds

Seattle's biggest traffic jam remains unchanged despite "improvements"

By Sean Keeley | @SeanKeeleyMe | Updated Nov 29, 2016, 11:56am PST

_____

_____

_____

_____

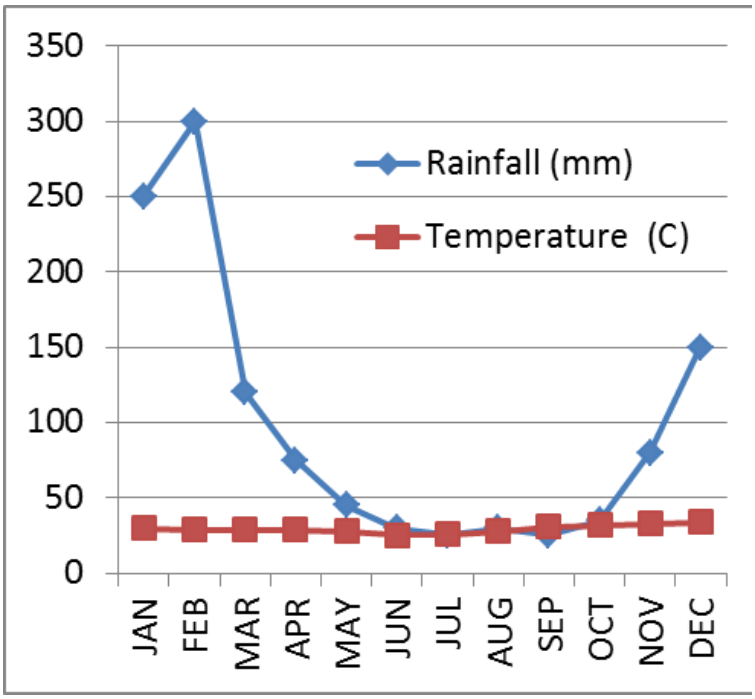
_____

_____

שאלה 7 (8 נק')

הגרף הבא מראה את כמות הגשם (קו עליון) והטמפרטורה (קו תחתון) בעיר Townsville באוסטרליה.

ציר ה X מציין את החודש בשנה.



ת.ז. _____

(א) [4 נק'] ציינו בעיה הקיימת בגרף ומדוע היא מקשה על הסקת מסקנות.

_____

_____

_____

_____

(ב) [4 נק'] הציעו דרך אחרת להציג את הנתונים (טמפרטורה וגשם). הדגימו באיור פשוט (לא חייבים שהמספרים יתאימו, רק שנבין את הרעיון).

_____

_____

[מקום לאיור]

שאלה 8 (5 נק')

נתונה התדירות שבה מופיעה כל מילה במספר טקסטים:

	חתול	כלב	כביש	בית	מכונית
טקסט 1	1	0	0	50	0
טקסט 2	0	1	1	2	1
טקסט 3	0	0	0	10	0

כעת אתם רוצים לייצג את אותם טקסטים בשיטת **TF-IDF**. בייצוג של **טקסט 2**, איזו מילה תקבל ערך גבוה יותר, "כביש" או "בית"? הסבירו. שימו לב, אין צורך לחשב את הערכים המדויקים, רק להסביר את ההיגיון.

_____

_____

_____

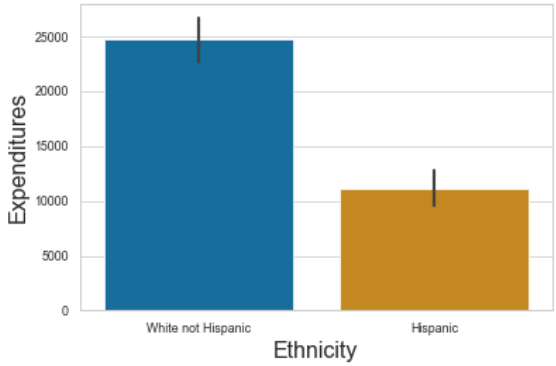
_____

_____

ת.ז. _____

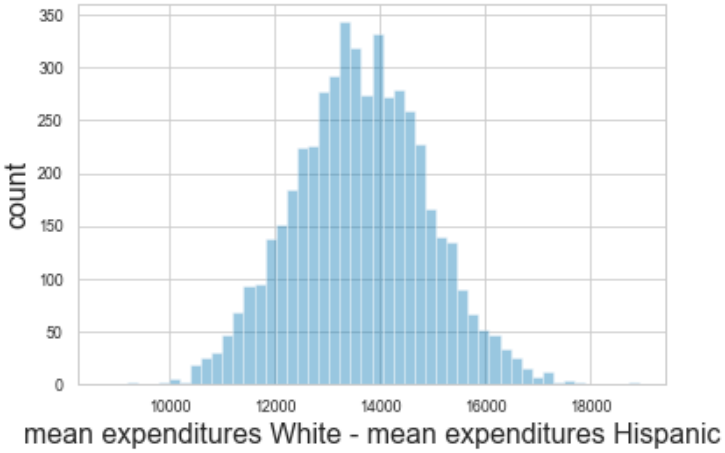
שאלה 9 (12 נק')

מדינת קליפורניה משלמת קצבאות נכות למשפחות הזקוקות לכך. לפני מספר שנים הועלתה טענה על אפלייה כנגד תושבים ממוצא היספני (מהגרים ממדינות מרכז אמריקה והסביבה). נטען שתושבים לבנים מקבלים קצבאות נכות גבוהות יותר. על מנת לבחון את הטענה, חוקרים בחנו את הנתונים על קצבאות הנכות הניתנות במדינה. הגרף הבא מציג את ממוצע קצבאות הנכות בדולרים אשר ניתנו לתושבים לבנים (לא היספניים) וממוצע הקצבאות לתושבים היספנים:



מקרא לגרף:
Expenditures: ממוצע הוצאות על קצבאות נכות
Hispanic: היספנים
White not Hispanic: תושבים לבנים

(א) [4 נק'] החוקרים השתמשו בשיטת בוטסטראפ כדי לבחון האם ההבדל בהוצאות על קצבאות הנכות בין לבנים להיספניים מובהק סטטיסטית. לצורך כך הם דגמו דגימות בוטסטראפ ועבור כל דגימה חישבו את ההפרש בין ממוצע ההוצאות על קצבאות לאוכלוסיית הלבנים פחות ממוצע ההוצאות על קצבאות לאוכלוסיית ההיספנים. להלן הגרף שהתקבל:



האם ניתן להסיק מהגרף שיש הבדל בממוצע קצבאות הנכות בין האוכלוסיה הלבנה לאוכלוסיית ההיספנים ברמת מובהקות של 0.05? הסבירו.

_____

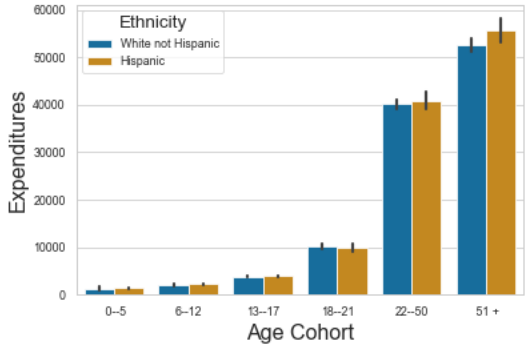
_____

_____

_____

ת.ז. _____

(ב) [5 נק'] בהמשך, החוקרים רצו לבדוק את ההבדלים בממוצאי קצבאות הנכות עבור תושבים בקבוצות הגיל השונות. הם הסתכלו על הגרף הבא, שמראה עבור כל קבוצת גיל (ציר X) את ממוצע הקצבאות לתושבים לבנים ולתושבים היספנים (ציר Y):



על פי כל הנתונים שבידיכם, האם לדעתכם ניתן להסיק כי יש אפלייה כלפי התושבים ההיספניים?

_____

_____

_____

_____

_____

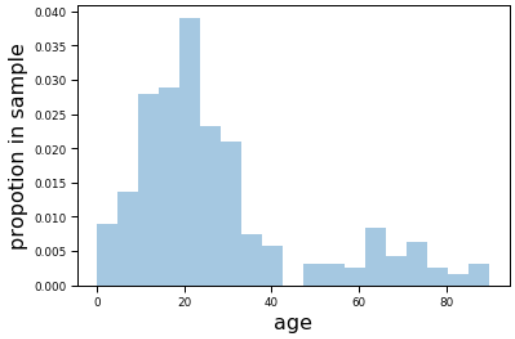
_____

_____

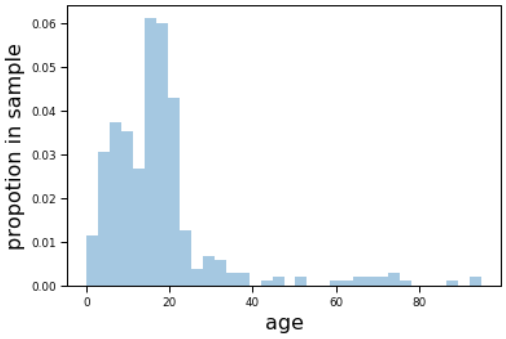
_____

(ג) [3 נק'] אחד מהגרפים הבאים מראה את התפלגות הגילאים של התושבים הלבנים ואחד מראה את התפלגות הגילאים של התושבים ההיספנים. איזה גרף מתאים לאיזו אוכלוסייה? הסבירו את תשובתכם.

גרף א':



גרף ב':



_____

_____

_____

_____

_____

_____

ת.ז. _____

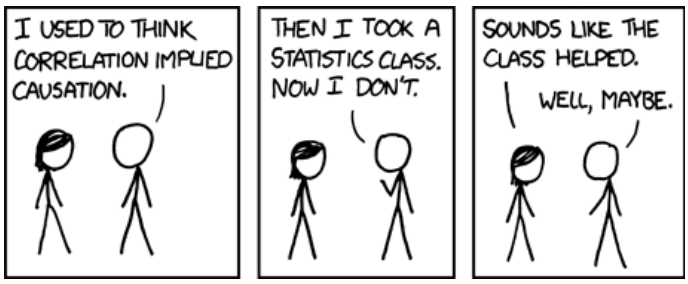
שאלה 10 (6 נק')

הסבירו את המושג "הזכות להישכח" (right to be forgotten) ותנו דוגמה למצב שבו יכול לעלות טיעון לזכות זו.

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

שאלת בונוס (2 נק'):

הסבירו מדוע הקומיקס הבא מצחיק:



_____
_____
_____
_____