סטטיסטיקה למדעי המחשב – תרגיל בית שבוע 3

שאלה 1 (20 נקודות, 5 נקודות לכל סעיף)

נסתכל על סטיית התקן המדגמית, S:

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

א. הוכיחו או הפריכו: סטיית התקן המדגמית היא תמצית לפיזור.

ב. מהי נקודת השבירה האסימפטוטית של S?

הוכיחו או הפריכו:

ג. ממוצע של שתי תמציות לנטייה הוא תמצית לנטייה.

ד. השברון ה-0.9 הינו תמצית חסינה.

נסתכל על מקדם המתאם המדגמי $\hat{r}$:

$$\hat{r} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}} = \frac{\widehat{Cov}(X,Y)}{\widehat{sd}(X)\widehat{sd}(Y)}$$

ה. (בונוס – 5 נקודות) מהו שיעור התצפיות שנדרש לשנות, ובאיזה אופן, על מנת ש $\hat{r}$ ישתנה מערך חיובי ממש לערך ששואף ל-0. האם זו תמצית חסינה?

שאלה 2 (25 נקודות)

שאלה זאת משתמשת בקובץ הנתונים heights.csv.

א. ציירו תרשים פיזור של גובה (X) ומשקל (Y) מקובץ הנתונים.

ב. הוסיפו את הקו החסין לתרשים בצבע אדום ואת קו הריבועים הפחותים בצבע כחול (מומלץ לממש באופן עצמאי את הקו החסין).

ג. מה ניתן להסיק מהסתכלות על תרשים לגבי הרגישות של כל אחד מהקווים להימצאות תצפיות חריגות?

ד. מצאו את הערכים הבאים: שיפוע קו הריבועים הפחותים, מקדם המתאם בין X ל – Y, אחוז השונות המוסברת. האם לפי נתונים אלו הייתם מסיקים שקיים קשר לינארי בין גובה ומשקל?

ה. חזרו על סעיפים ב' ו-ד' לאחר הסרת תצפית חריגה אחת.

ו. מה ניתן להסיק על ההשפעה של תצפיות חריגות על אחוז השונות המוסברת?

שאלה 3 (30 נקודות)

שאלה זאת משתמשת בקובץ הנתונים flatprices.csv.

א. ציירו תרשים פיזור של גודל הדירות (X) ו מחירי הדירות (Y) מקובץ הנתונים.

ב. הוסיפו את קו הריבועים הפחותים.

ג. מצאו את הערכים הבאים: שיפוע קו הריבועים הפחותים, מקדם המתאם בין X ל – Y, אחוז השונות המוסברת, SSE, SST.

ד. מה צפי המחיר לדירה בגודל 125 מ"ר לפי קו הריבועים הפחותים?

ה. בחנו את האופן שבו התצפיות מפוזרות מסביב לקו הריבועים הפחותים. האם אתם חושבים שכדאי לבצע טרנספורמציה לא לינארית לאחד המשתנים? הסבירו.

ו. בחרו 2 אפשריות לביצוע טרנספורמציה לא לינארית למשתנה אחד או לשני המשתנים וחזרו על סעיפים א'-ד' עבור כל אחת מהאפשרויות. באיזה מבין שלושת המודלים הייתם משתמשים על מנת להכריע מה המחיר של דירה בגודל 125מ"ר?

שאלה 4 (25 נקודות)

א. הגרילו באקראי וקטור X באורך 30, מתוך התפלגות נורמאלית עם תוחלת 5 וסטיית תקן 1, בעזרת הפונקציה ()rvs של scipy.stats.norm (הפרמטר loc של ()rvs מגדיר את התוחלת, scale את סטיית התקן ו-size את גודל הדגימה)

ב. צרו את וקטור Y על ידי הכפלת X ב-5 והוספת 2 (Y=5X+2).

ג. מה צפוי להיות המתאם בין X ו-Y , $\hat{r}$? הראו שצדקתם בעזרת חישוב בפייתון.

ד. מה צפוי להיות השיפוע של קו הריבועים הפחותים, $\hat{b}$? הראו שצדקתם.

ה. הוסיפו רעש ל-Y מתוך התפלגות נורמאלית עם ממוצע 0 וסטיית תקן 1 באופן הבא:

```
noise= norm.rvs(loc=0, scale=1, size=30)

Y=Y+noise
```

מהם ערכי $\hat{r}$ ו- $\hat{b}$ כעת?

ו. חשבו את ערכי $\hat{r}$ ו- $\hat{b}$ עבור ערכים שונים של סטיית התקן של הרעש (בין 0.5 ל 10).

ז. צרו גרף פיזור אחד המציג את ערכי $\hat{r}$ כפונקציה של סטיית התקן, וגרף פיזור נוסף המציג את ערכי $\hat{b}$ כפונקציה של סטיית התקן.

ח. מה ניתן להסיק מגרף זה על הקשר שבין הרעש של הנתונים למקדם המתאם, אחוז שונות מוסברת, ולאמינותו של קו הריבועים הפחותים.

שאלה 5 (20 נקודות בונוס)

כנסו לאתר https://www.gapminder.org/tools/

א. סיירו באפליקציה ומצאו 2 משתנים שאתם חושבים שיכולים לספר "סיפור" מעניין אחד על השני.

ב. הסבירו מה כל אחד מהמשתנים שבחרתם מתאר.

1. אם אתם לא בטוחים, חפשו בגוגל.
2. אם אתם עדיין לא בטוחים, ציינו זאת (זה לגיטימי).

ג. ציינו האם נדרשת טרנספורמציה לאחד או לשני הצירים.

ד. האם יש תצפיות חריגות בגרף? ציינו האם הן נובעות מנתונים בעייתיים או לא אמינים.

ה. האם לדעתכם יש סיבתיות בין גורם אחד לאחר?

1. אם כן, מי לדעתם גורם למי ובאיזה אופן.
2. נמקו את תשובתכם, רצוי בציון מקורות אם אתם מוצאים כאלה.

ו. בחרו מדינה אחת בגרף וספרו את הסיפור שלה:

1. צרפו גרף אחד המתאר את הסיפור שברצונכם לספר על מדינה זו.
2. בחרו בקפידה מה להדגיש ומה לא להדגיש בגרף, כך שיתאים לסיפור. חוסר במועד ועודף במידע לא רלוונטי לא רצויים באותה המידה.
3. ספרו בקצרה (פיסקה אחת) את הסיפור שהגרף מספר.

הערות:

א. סיפור יכול להיות "מעניין" כתוצאה מטיב הקשר בין המשתנים, כתוצאה מחוסר הקשר ביניהם, מתצפיות חריגות, מהתפתחות הקשר והמשתנים בזמן, כתוצאה ממצב המדינה שבחרתם לעומת מדינות אחרות ועוד.

ב. גם ויקיפדיה היא מקור טוב, כל עוד אתם מציינים אותו. זה תרגיל בית ולא עבודת מחקר.

ג. הסבר פשוט על שימוש באפליקציה מופיע בכפתור How to use בראש האתר.

ד. נא לא לבחור את שילוב ברירת המחדל בין Life expectancy ל- Income !