ביג דאטה, תרגיל 2, אביב 2022

גרסא 1.00

הגשה ביחידים

תאריך הגשה: יום שני, 16/5/2022 עד שעה 11:59 בלילה

בתרגיל זה תנתחו מידע הנוגע לשינויי טמפרטורות ומספר ימי שלג בתחנה כלשהי עפנ״י כדור הארץ. שאלת המחקר עליה אנחנו מנסים לענות היא האם קיימים שינויי טמפרטורה מובהקים במקומות שונים בעולם לאורך תקופות זמן של 60 שנה או יותר.

data

באתר הבא ניתן למצוא מידע אקלימי עבור מקומות שונים בעולם:

https://www.ncdc.noaa.gov/cdo-web/search

המידע נאסף בתחנות, כשבכל תחנה נאסף מידע יומי של טמפרטורות מקסימום (TMAX) ומינימום (TMIN), כמויות שלג שירדו (SNOW) ועוד. למשל, התחנה הבאה נמצאת בניו יורק ועבורה יש מידע בין השנים 1894 ל-2014:

https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USC00369933/detail

בתרגיל הזה תנתחו מידע שנוגע לטמפרטורות המקסימום היומיות וכמות השלג שירד בתחנה כלשהי. כאן יש פרטים על תחנות שונות:

https://tinyurl.com/2p8erdcj

מצאו את השם שלכם בטבלה ונתחו את המידע עבור התחנה שבאותה שורה. הורידו את המידע עבור התחנה וודאו שהעמודות TMAX ו-SNOW קיימות בקובץ ה-csv שהורדתם.

משימות

1. הכנת המידע

- טענו את הקובץ ל-DataFrame
- שנו את האינדקס להיות התאריך (העמודה DATE)
- שנו את העמודה DATE מ-str ל-datetime
- צרו טבלה חדשה בשם tmax שתכיל רק את התאריך (DATE) ואת העמודה TMAX. נקו מהטבלה את כל השורות עבורן לא קיים ערך TMAX
- צרו טבלה בשם snow שתכיל רק את התאריך (DATE) ואת העמודה SNOW. נקו את כל השורות שעבורן SNOW לא קיים
- הורידו מכל טבלה (tmax, snow) את כל השורות עבור חודשים (חודש בשנה מסויימת) שבשבילם קיימים פחות מ-28 מדידות

2. Exploratory data analysis

- כתבו פונקציה בשם show_monthly_temp שמקבלת את tmax ויוצרת איור ובו 12 גרפים (scatter plots) שמציגים, עבור כל חודש, את הטמפרטורות הממוצעות לאותו חודש לאורך כל השנים עבורן יש נתונים. בנוסף, הוסיפו קו רגרסיה לינארית אותו תצבעו באדום במידה והשיפוע חיובי או בכחול אם השיפוע שלילי.
- כתבו פונקציה בשם show_snow_days שמקבלת את snow ומציגה boxplot של מספר ימי השלג עבור כל חודש (סה״כ 12 boxplots).

3. T-test

כתבו פונקציה בשם calc_diff_snow שתשווה, עבור החודשים דצמבר, ינואר ופברואר את התפלגות מספר ימי השלג ב-15 השנים המוקדמות ביותר שעבורן יש מספיק מידע ל-15 השנים המאוחרות ביותר שעבורן יש מספיק מידע (השנים יכולות להשתנות עבור כל חודש). מצאו את הפרש הממוצעים ובדקו האם ההפרש מובהק סטטיסטית. תקנו את ה-P-value בעזרת FDR כך שיתאים ל-12 חישובים.

הדפיסו טבלה שבה יופיע, לכל חודש, המידע הבא:

- החודש
- טווח השנים המוקדם (שנה ראשונה ואחרונה)
- טווח השנים המאוחר (שנה ראשונה ואחרונה)
- ממוצע ימי שלג עבור טווח השנים המוקדם
- ממוצע ימי שלג עבור טווח השנים המאוחר
- ההפרש במספר הימים
- ה-P-value לפני תיקון FDR
- ה-P-value אחרי תיקון
- סטטוס: המילה HIGHER אם מספר ימי השלג עלה או LOWER אם ירד
- סיגניפיקנטיות: סימון *** אם ה-P-value המתוקן קטן מ-0.001, * אם הוא קטן מ-0.01 ו-* אם הוא קטן מ-0.05.

שאלות

הגישו קובץ pdf אשר מכיל את האיורים שיצרתם (עבור הטמפ׳ ועבור מספר ימי השלג) ואת הטבלה. בנוסף, ענו על השאלות הבאות:

1. האם קיימים הבדלים משמעותיים בין שתי תקופות הזמן שהשוותם עבור חלק/כל החודשים?
2. האם t-test הוא מבחן סטטיסטי מתאים במקרה הזה?
3. לסיכום – האם, לדעתכם, הנתונים שאותה בדקתם מצביעים על מגמה כללית של התחממות?

הערות

- הקפידו על איכות הקוד.
- השתדלו ליצור איורים אסתטי וברורים ככל האפשר.
- תעדו את הקוד שלכם במידת הצורך.

הגשה

ארזו את הקבצים הבאים לקובץ tar או zip והגישו אותם דרך המודל:

- הקוד בו השתמשתם על מנת לייצר את קבצי המידע.
- הקובץ עם האיורים, הטבלה והתשובות לשאלות.

בהצלחה!