תרגיל כתה: קריאת נתונים והתייחסות לערכים חסרים

בתרגיל זה נבדוק מגמות בכמויות המשקעים בסן פרנסיסקו. לצורך כך נשתמש בקובץ נתוני אקלים יומיים מהעיר סן פרנסיסקו עם נתונים (לא מלאים) עבור השנים 1949-2016. את הנתונים הורדתי מהאתר https://www.noaa.gov.

הורידו מה-Moodle את הקובץ SFWeather.csv. הקובץ מכיל מספר עמודות, מתוכן אלו שמעניינות אותנו הן:

- DATE – תאריך המדידה בפורמט YYYYMMDD (למשל 19840526 עבור התאריך 26/5/1984)
- PRCP – כמות המשקעים שנמדדה בעשיריות מילימטר.

חלק ראשון: הכנת הטבלה

1. קראו את קובץ הנתונים ל-DataFrame בשם sfrain. הסירו מהטבלה את כל העמודות חוץ משתי העמודות הרלוונטיות. הדפיסו את הטיפוס של הערכים שנשמרים בכל אחת מהעמודות.
2. המירו את העמודה DATE לטיפוס datetime. הטיפוס הזה מאפשר לבצע פעולות של תאריכים כמו שליפת השנה (בעזרת התכונה year) והחודש (בעזרת התכונה month).
הדרכה: ניתן להפעיל פונקצית lambda בעזרת apply על העמודה DATE ובה להמיר את התאריך בעזרת הפונקציה datetime.strptime לאחר שממירים את התאריך ל-str.
3. מיינו את הטבלה לפי התאריך.
4. הפכו את העמודה DATE להיות האינדקס של הטבלה בעזרת set_index. אל תשכחו להשתמש בפרמטר inplace כדי שהשינוי יתבצע על הטבלה. התוצאה היא טבלה עם עמודה אחת (PRCP).
5. הדפיסו את מספר המדידות (שורות) שמופיעות בטבלה (בעזרת shape)
6. בידקו את הערכים בעמודה PRCP ומיצאו את הערך שמסמן מדידה חסרה.
7. הורידו את כל השורות שבהן יש ערך חסר בעמודה PRCP. הדפיסו שוב את מספר השורות כדי לראות כמה ערכים ירדו.

חלק שני: הכנת טבלת סיכום חודשית

נרצה ליצור טבלה חדשה שבה השורות מייצגות נתונים עבור חודש+שנה (למשל 1/1949, 2/1949 וכו) ובה שלוש עמודות: מספר ימי מדידה לחודש, מספר ימי גשם בחודש, וסך כל הגשם בחודש. לאחר מכן נוסיף עמודה רביעית שתכיל את ממוצע הגשם היומי.

לצורך בניית העמודות נשתמש ב-groupby על sfrain. בשביל ליצור קבוצות של חודש+שנה אפשר להעביר ל-groupby את הביטוי הבא:

```
sfrain.index.map(lambda x: datetime(x.year, x.month, 1))
```

הביטוי הזה יוצר מהאינדקס של sfrain מערך תאריכים שבו היום הוא תמיד 1. למשל: כל הימים של ינואר 1949 (1/1/1949, 2/1/1949 וכו) יכללו בקבוצה של 1/1/1949.

8. צרו Series בשם ndays שמסכם את מספר המדידות עבור כל חודש בכל שנה.
9. צרו Series בשם nraindays שמסכם את מספר ימי הגשם בכל חודש/שנה. יום גשם הוא יום שבו PRCP שונה מ-0.
10. צרו Series בשם prcp שמסכם את כמות המשקעים עבור כל חודש בכל שנה.

11. צרו DataFrame בשם sfrain_summary משלוש המערכים שיצרתם. שמות העמודות תהיינה שמות המערכים.
12. הסירו את כל השורות של חודשים שעבורם יש מדידות עבור פחות מ-90% מהימים בחודש.
אפשר ליצור רשימה עם מספר הימים המינימלי בכל חודש עפ״י הקריטריון הזה בעזרת הקוד הבא:

```
from calendar import monthrange

min_days = [monthrange(i.year, i.month)[1]*0.9 for i in sfrain_summary.index]
```

השתמשו בתוצאה בשביל לסנן את כל החודשים עם פחות מ-90% מהימים.
13. הוסיפו עמודה בשם daily_prcp שמכילה את כמות המשקעים הממוצעת ליום בכל חודש.

חלק שלישי: קצת גרפים

14. צרו barplot שמציג את כמות הגשם היומית הממוצעת בכל חודש. הוסיפו labels לצירי ה-x וה-y וכן כותרת לאיור. מהם החודשים הגשומים ביותר בסן פרנסיסקו?
15. צרו 12 scatterplots (כ- 3 שורות על 4 עמודות של subplots של אותו איור) עבור 12 החודשים עם כמות הגשם היומית הממוצעת בכל חודש בכל שנה (daily_prcp). כלומר: עבור ינואר תהיינה נקודות עבור כל השנים שעבורן יש מדידות לינואר בין 1949 ל-2016, וכו. הגבילו את ציר ה-y בכל האיורים לאותו טווח (0 עד לקצת מעל הערך המקסימלי בעמודה daily_prcp), תנו לכל איור כותרת שתכיל את החודש.
האם אפשר להבחין במגמות כלשהן לאורך השנים?
16. בונוס: צרו איור scatterplot שמציג את כמות המשקעים היומית הממוצעת לאורך כל השנה. האם ניתן להבחין במגמה מפתיעה כלשהי?

הגישו את הקוד שיצרתם.