ביג דאטה, תרגיל 1, אביב 2024

גרסא 1.00
הגשה ביחידים
תאריך הגשה: יום שני, 17/6/2024 עד שעה 11:59 בלילה

בתרגיל זה ננסה לענות על שתים מהשאלות החשובות ביחסים בינלאומיים: האם האירויזיון הוא תחרות פוליטית ומי המדינות שעוינות את מדינת ישראל בזירת השירה הבינלאומית?

data

נשתמש בשני קבצי נתונים:

1. contestants.csv – קובץ עם נתונים על השירים שייצגו כל מדינה לאורך ההיסטוריה של התחרות
2. votes.csv – קובץ עם נתוני ההצבעה של כל המדינות לאורך ההיסטוריה של התחרות

משימות

1. נתונים כלליים על התחרות

השתמשו בקובץ contestants.csv בשביל להפיק את המידע הבא:

א. טבלה שמכילה את המידע הבא עבור כל מדינה: שם המדינה, מספר התחרויות בהן היא השתתפה, החציון של המקומות בהן היא סיימה את כל התחרויות, מספר הפעמים בהן היא ניצחה בתחרות, ואחוז הפעמים בהם היא ניצחה מתוך סך כל הפעמים בהן השתתפה. אינדקס השורות צריך להיות קוד המדינות. מיינו את הטבלה לפי העמודה האחרונה. מה מצבה של ישראל בטבלת הזכיות?
הדרכה: השתמשו ב-group_by על העמודה to_country_id והריצו את כל החישובים על העמודה place_final. התוצאה תהיה טבלה עם העמודות המבוקשות וה-to_country_id כ-index.

ב. הדפיסו את כל האמנים (performers) שהשתתפו יותר מ-3 פעמים בתחרות.

ג. הדפיסו את כל המלחינים (composers) שהלחינו יותר מ-3 שירים לתחרות.

ד. הדפיסו את כמות הפעמים שבהם ישראל סיימה בכל מקום בתחרות לאורך כל שנותיה. הדפיסו בסדר יורד מהמקום שבו סיימה ישראל הכי הרבה עד למקום שבו סיימה ישראל הכי מעט.

2. יצירת "מילון" להמרת שמות מקוצרים של מדינות לשמות ארוכים

הנתונים בקובץ votes.csv שמורים לפי קודים של מדינות (למשל il בשביל ישראל). כדי לתרגם את הקודים לשמות מדינות ניצור Series מהעמודות to_country_id ו-to_country בקובץ contestants.csv. השתמשו במתודה drop_duplicates על חלק מהטבלה שכולל את שתי העמדות בשביל לשמור עותק אחד של כל זוג קוד-מדינה, הפכו את התוצאה ל-pd.Series שבו העמודה to_country_id היא האינדקס ו-country_name היא הערך. שמרו את התוצאה במשתנה code2country.
ישנם שני מקרים של קוד מדינה עם יותר מערך אחד לקוד. בשביל להסיר את המקרים האלה ניתן להשתמש בקוד:

```
code2country = code2country[~code2country.index.duplicated(keep='first')]
```

3. ניתוח הפוליטיזציה של האירויזיון

באירוויזיון ניתן ניקוד על ידי צוות שופטים של כל מדינה ובשנים האחרונות גם על ידי הקהל. בניתוח הזה נתמקד בניקוד שניתן על ידי השופטים (jury) מכיוון שזו שיטת הניקוד המסורתית וקיים עבורה הרבה יותר מידע מאשר הצבעת הקהל. כמו כן, נתמקד בניקוד בגמר ונתעלם מחצי הגמר. אפשר לקרוא על שיטות הניקוד השונות באירוויזיון בלינק הזה.

שיטת הניקוד ומספר המדינות השתנו לאורך השנים, ולכן אציע לבדוק האם הניקוד שנתנה מדינה למדינה אחרת בתחרות מסויימת גבוה או נמוך מהממוצע של המדינה המקבלת באותה תחרות. למשל: אם באירויזיון כלשהו השתתפו 20 מדינות, ישראל קיבלה את הניקוד הכולל 160 (ממוצע של 8 נק׳ מכל מדינה) והולנד נתנה לישראל 9 נק׳, אז באירויזיון הזה הולנד נתנה לישראל ניקוד שגבוה מהממוצע. בשביל לבדוק האם הולנד אוהדת את ישראל נבדוק באיזה אחוז מהתחרויות נתנה הולנד לישראל ציון מעל לממוצע.
מדינות שנותנות באופן יחסית עקבי ניקוד גבוה או נמוך יותר למדינות אחרות עשויות להחשב כאוהדות אותן, ולהיפך.

הכנת הטבלה

א. טענו את הקובץ votes.csv למשתנה בשם votes.
ב. הסירו את כל השורות שבהן העמודה round שונה מ-final.
ג. השתמשו ב-code2country בשביל לשנות את הקודים בעמודות from_country_id ו-to_country_id לשמות המדינות.
ד. הסירו מהטבלה את כל המדינות שהשתתפו בתחרות פחות מ-30 פעמים. אנחנו צריכים מידע מספיק עבור המדינות שאותן אנחנו בודקים.
ה. עבור כל השורות שמתייחסות לתחרויות עד 1996 העמודה total_points מכילה את הניקוד שאליו נתייחס.
ו. אחרי 1996, נתייחס לעמודה jury_points. הסירו את כל השורות שמתייחסות לתחרויות אחרי 1996 שבהן jury_points היא ללא ערך. עבור שאר השורות של אחרי 1996 העתיקו את הערך של jury_points ל-total_points.
ז. שנו את שם העמודה from_country_id ל-from, את to_country_id ל-to ואת total_points ל-points
ח. השאירו בטבלה את העמודות הבאות בלבד: year, from, tom points

חישוב הניקוד הממוצע לכל מדינה בכל תחרות

צרו DataFrame שבו 3 עמודות: שנה, מדינה (העמודה to) וציון ממוצע. אפשר ליצור את הטבלה על ידי שימוש ב-groupby על העמודות year ו-to מתוך votes ואז הרצה של mean על העמודה points. כדי לקבל טבלה כמבוקש ניתן להשתמש ב-reset_index() על התוצאה.

שנו את האינדקס של הטבלה שהתקבלה לשנה.מדינה. למשל, 2023.Israel בשביל השורה שמייצגת את הניקוד הממוצע של ישראל ב-2023.

"נרמול" הניקוד בכל שורה לפי הממוצע

הוסיפו ל-votes עמודה בשם adjusted.points שמכילה, בכל שורה, את points פחות הציון הממוצע למדינה ב-to בשנה year.

הוסיפו עמודה בשם above.average שמכילה False אם adjusted.points נמוך מ-0 או False אחרת. בכל המקומות שמתייחסים לניקוד של מדינה לעצמה הכניסו True.

יצירת מטריצת מדינות נותנות ניקוד (שורות) למדינות מקבלות ניקוד (עמודות)

צרו DataFrame בשם from_to_above_average שבו כל מקום x, y מציין את אחוז הפעמים שבהם מדינה x נתנה למדינה y ניקוד גבוה מהממוצע. השתמשו ב-groupby על העמודות from, to ב-votes, חשבו את הממוצע על העמודה above_average. בשביל לקבל טבלה בגודל 30 על 30 עם שמות המדינות המנקדות כאינדקס ושמות המדינות המקבלות כשמות העמודות הפעילו את unstack() על התוצאה.

4. ניתוח התוצאות

א. הציגו heatmap שבו השורות והעמודות ימויינו על ידי hierarchical clustering (נלמד על זה). אפשר להשתמש ב-clustermap של seaborn למטרה זאת.

האם אתם יכולים לאתר קבוצות של מדינות שנוטות להצביע אחת לשניה? מה לגבי מדינות שבד"כ לא מצביעות זו לזו?

ב. עבור העמודה של ישראל (אחוז הפעמים שכל מדינה נותנת לישראל ניקוד גבוה מהממוצע), מיינו את העמודה בסדר יורד, הסירו את התא שמכיל את ה"ניקוד" שישראל נותנת לעצמה וציירו barplot שיציג את המידע. הוסיפו קו אופקי עבור הממוצע של כל הערכים בשביל ישראל.

ג. חזרו על סעיף ב עבור השורה של ישראל (מתייחסת לניקוד שישראל נותנת למדינות אחרות).

לסיכום, ענו על השאלות הבאות:

1. האם מהתוצאות עולה לדעתכם שהאירוויזיון הוא תחרות פוליטית?
2. מי המדינות שנותנות את הניקוד הכי נמוך וגבוה לישראל? האם לדעתכם יש סיבה לחשוב שהסיבות הן פוליטיות? האם המדינה שמעניקה לישראל את הניקוד הכי נמוך מקבלת מישראל ניקוד נמוך בחזרה? מה לגבי המדינה שמעניקה לישראל את הניקוד הכי גבוה?
3. לאיזה מדינות כדאי שנשפר את הניקוד בשביל לשמור על יחסים טובים?

הגשה

הגישו את הקוד (קובץ פייתון) וקובץ סיכום עם האיורים המבוקשים והתשובות לשאלות. אפשר להגיש jupyter notebook.

בהצלחה!