עבודה עם Dataset, חילוץ נתונים, ושימוש בפונקציות בסיסיות של tidyverse

2022-03-22

בתרגיל זה תתנסו במספר פונקציות בסיסיות ב-R (ולמעשה פונקציות של tidyverse).

חלק ראשון

קראו את התיעוד בעמוד הבא:

https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-09-22/readme.md
((https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-09-22/readme.md

השתמשו בפונקציה `read_csv` מחבילת `readr` על מנת לקרוא את שלושת הקבצים הבאים. ניתן להיעזר בקוד הבא:

```
members <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytues
day/master/data/2020/2020-09-22/members.csv')
expeditions <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidy
tuesday/master/data/2020/2020-09-22/expeditions.csv')
peaks <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesda
y/master/data/2020/2020-09-22/peaks.csv')
```

על מנת לתאר את הדאטה, באופן כללי, השתמשו בפונקציות `distinct` , `arrange` , `filter` , `count` וענו על השאלות הבאות:

1. אילו שנים מתועדות בקבצים (טווח השנים של משלחות להימלאיה)
2. כמה אזרחויות שונות היו מעורבות במשלחות להימלאיה, עד לשנת 1950?
3. על בסיס הנתונים, מה לדעתכם העונה הטובה ביותר לטפס על פסגת האוורסט?
4. כמה משלחות מתועדות בקובץ, שניסו לטפס על האוורסט (ולא הצליחו להעפיל לפסגה) לפני שנת 1953?
5. כמה פסגות ברכס ההימלאיה מגיעות לרום של מעל 800 מטר?

חלק שני

בחלק זה תשתמשו בפונקציה `mutate` על מנת לבחור משתנים ולערוך טרנספורמציות.

היעזרו בקוד הבא, על מנת לייצר טבלה חדשה עם שני משתנים בוליאניים חדשים: `is_leader` , `is_doctor` .

הסבירו מה עושה הפונקציה `str_detect` , ומה המשמעות של כל שורה בקוד (מה עושה כל שורה).

```
leader_table <- members %>%
  mutate(is_leader = str_detect(expedition_role, "Leader")) %>%
  mutate(is_doctor = str_detect(expedition_role, "Doctor"))
```

השתמשו בפונקציה `count` על מנת לייצר טבלה שתראה את כל הצירופים האפשריים של is_leader, is_doctor, ומספר התצפיות. כמה מובילי משלחות יש שהם גם רופאים?

השתמשו ב- `mutate` ובפונקציה `cut` על מנת לבנות טבלה חדשה שבה יש משתנה שנקרא `decade` , המתאר את העשור שבו יצאה המשלחת. ניתן להיעזר בקוד הבא (השלימו את הקוד).

לאחר מכן, צרו תרשים שיציג את מספר המשלחות בכל עונה בכל עשור.

```
expeditions_new <- expeditions %>%
  mutate(decade = cut(year, breaks = c(___, ___, ___, ...))) %>%
  count(season, ___)

ggplot(___, aes(x = decade, y = ___, fill = season)) +
  geom_col(position = position_dodge())
```

מה ההבדל בין שימוש ב- `position_dodge` לבין `position_stack` ו- `position_fill` בקוד לעיל?

חלק שלישי

נניח שאתם מתכננים להעפיל לפסגה ברכס ההימלאיה, של מעל ל-8000 מטר.

1. צרו טבלה חדשה עם רשימה של פסגות אלו.
2. מיהן עשרת סוכנויות המסע (`trekking_agency`) שהוציאו הכי הרבה משלחות?
3. השתמשו ב- `group_by` וב- `summarize` על מנת לחשב את ממוצע התמותה לכל משלחת, באופן כללי, וגם עבור הפסגות מעל 8000 מטר בלבד (ניתן לחלק סעיף זה לשני חישובים נפרדים).
4. עם אילו חברות הייתם שוקלים לצאת, ועם אילו לא?