Learning:

בחרנו להשתמש במודל GLM – generalized linear model על מנת לחזות את המשתנה התלוי PRCP באמצעות סט פיצ׳רים. בחרנו במודל רגרסיה כיוון שהמשתנה אותו אנו רוצים לחזות הוא משתנה רציף ובנוסף ניסינו להשתמש גם במודל רגרסיה לינארית פשוטה אך קיבלנו תוצאות מעט יותר טובות עבור מודל ה-GLM, למרות שאין באפשרותנו להניח כי הנתונים מגיעים מהתפלגות גאוסיאנית כלשהי.

את סט הפיצ׳רים נבחר מבין כמה אפשרויות בכך שנבחר את הסט שממזער את הRMSE עבור מודל הGLM.

נריץ את המודל על טבלת station_year שהוצאנו בתהליך הETL, לאחר סינון רשומות שמכילות מתחת ל150 ימים או ששייכות לשנים הקודמות ל-1950. בחרנו בסינונים אלו על מנת לחזק את אמינות הנתונים, כך שתצפיות חריגות יטו אותו פחות. את הטבלה נפצל באופן אקראי לחלוטין לסט אימון וסט מבחן, ביחס 30-70. מודל הGLM משתמש ב-cross validation, בחלוקה ל-10 folds, על מנת לבחור את הפרמטר המיטבי. את המודל נאמן על סט האימון ונבחן מדד RMSE על פי סט המבחן. נבחן האם המודל חוזה יותר טוב עבור קווי אורך, קווי רוחב, קווי גובה ושנה.

בחרנו 6 אפשרויות שונות לסט הפיצ׳רים שעל פיו תיבנה הרגרסיה:

[שנה, אורך, רוחב, גובה]

[אורך, רוחב, גובה]

[אורך, רוחב]

[אורך]

[רוחב]

[שנה, אורך, רוחב]

השאלות שאנו רוצים לבחון הן האם ניתן לחזות ממוצע משקעים שנתי (ליום) על בסיס מיקום גיאוגרפי בלבד והאם הוספת שנה לחיזוי משפרת את הביצועים.

אלו תוצאות הRMSE שקיבלנו:

```
RMSE by feature list:
['lat', 'lon', 'ele', 'yea']:
13.229423412124387
['lat', 'lon', 'ele']:
12.859627933317993
['lat', 'lon']:
14.669280801714653
['lat']:
14.96886613914424
['lon']:
15.392656387375808
['lat', 'lon', 'yea']:
15.269297983894653
```

מהתוצאות שקיבלנו עולה כי החיזוי הטוב ביותר מתקבל עבור סט הפיצ׳רים [אורך, רוחב, גובה]

ללא שימוש בשנה. נשים לב כי אלמנט הגובה עוזר מאוד לחיזוי ולעומת זאת לאלמנט השנה קיימת השפעה שלילית על התוצאות שקיבלנו.

נוסף על כך, חיזוי המבוסס רק על פיצ׳ר בודד לא מניב תוצאות טובות וניתן להסיק כי חיזוי רק על סמך שתי המדינות שבחרנו הוא איננו מיטבי שכן ה-RMSE שקיבלנו בכל האפשרויות הוא יחסית גבוה.