ניהול מידע מבוזר - פרויקט חלק ב'

מגישים:

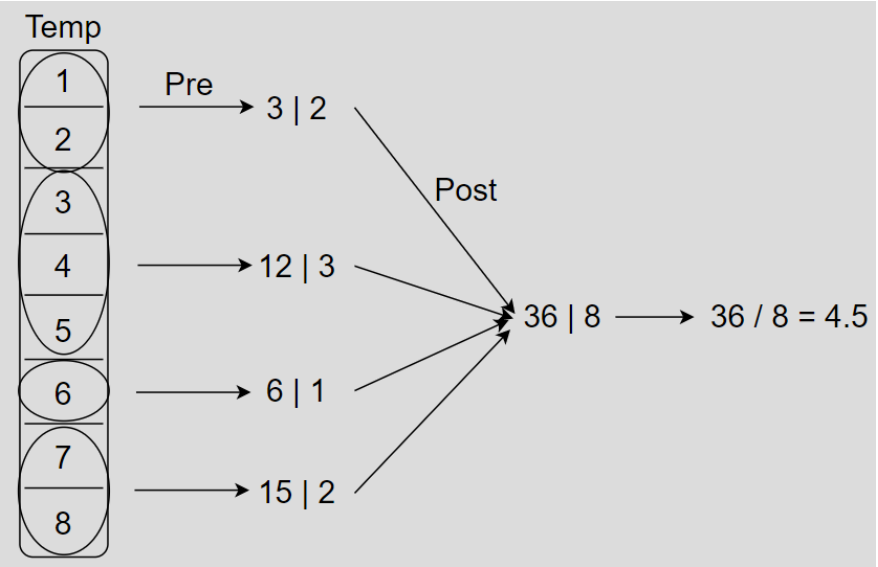
ירין בן שטרית 206230021

רועי פאפו 316327451

תומר פרץ 318295029

חלק 2 - Data Analysis:

העבודה על כל אחת מהאבחנות מתחלקת לשני שלבים - Pre & Post. דוגמה:

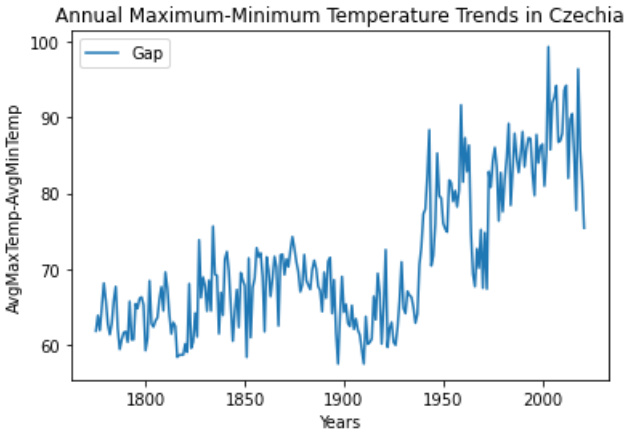


העמודה Temp מכילה 8 רשומות עם הערכים 1, 2, ..., 8. כל אליפסה מייצגת Batch המכיל את הרשומות שנמצאות בתוכו. המטרה - מציאת ממוצע הערכים של כל הרשומות. בשלב ה- Pre, עבור כל Batch, מחושבים שני ערכים: סכום ערכי הרשומות ומספר הרשומות ב- Batch. למשל עבור ה- Batch הראשון המכיל את הרשומות 1, 2, הערכים המתקבלים הם 3 (סכום הרשומות) ו- 2 (מספר הרשומות). בשלב ה- Post, מחברים את כל זוגות המספרים לקבלת סכום כל הרשומות ומספר כל הרשומות. מחלקים את המספר הראשון בשני כדי לקבל את הממוצע.

האבחנות הן:

1. **Temporal based insight:** בדקנו מהו ממוצע ההפרש בין הטמפרטורה המקסימלית הממוצעת (TMAX) שנמדדת לבין הטמפרטורה המינימלית הממוצעת (TMIN) שנמדדת בתחנות השונות בצ'כיה בכל שנה, כאשר המשתנה הבלתי תלוי הוא הזמן. השלבים:
   a. Pre: מתוך כל הדאטה, השתמשנו רק בעמודות Year, Variable, Value, כאשר העמודה Year מתקבלת מהפעלת הפונקציה year() של Spark על העמודה fullDate. מתוך רשומות אלו, נבחרו רק הרשומות עם המשתנה TMAX או TMIN. לבסוף בוצע groupBy לפי Year ובוצעו הפעולות של שלב Pre כפי שפורטו לעיל. התוצאה המתקבלת נכתבה לשרת בטבלה זמנית.
   b. Post: בוצעו הפעולות של שלב Post כפי שפורט לעיל, ונוצרה טבלה חדשה שבה לכל שנה מופיע ההפרש בין ממוצע הטמפרטורות המקסימליות שנמדדו לבין ממוצע הטמפרטורות המינימליות שנמדדו, נסמנה Gap.

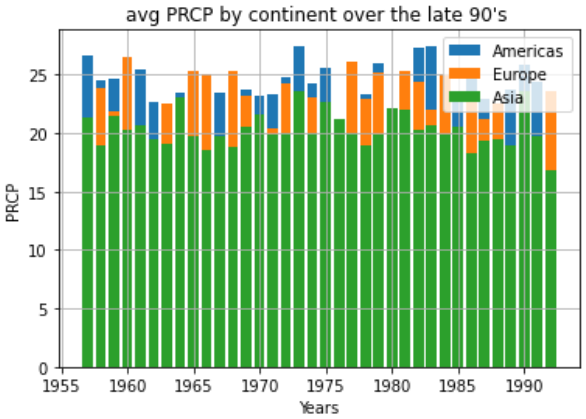
הגרף שהתקבל:



הגרף מתבסס על 100,000,000 רשומות שנאספו ממדידות בתחנות שונות בצ'כיה. מהגרף נוכל להסיק כי בעוד שמתחילת המדידות ועד לאמצע המאה ה- 20 אין שינוי במגמות הטמפרטורה, החל מאמצע ועד לסוף המאה ה- 20 ניתן לראות עלייה בהפרשים בין הטמפרטורה המקסימלית והמינימלית. נוכל להסיק מכך שהטמפרטורות הנמדדות נעשות קיצוניות יותר וגבוהות יותר, מה שעשוי להעיד על מגמת התחממות ביבשת אירופה בפרט ובסקאלה גלובלית בכלל.

2. **Spatio-temporal based insight:** בדקנו מהי כמות המשקעים הממוצעת בין השנים 1950-1990 בשלוש יבשות - אמריקה, אסיה ואירופה. המדינות הנציגות מכל יבשת הן סין, ארה"ב, צ'כיה, צ'ילה וצרפת. השלבים:
   a. Pre: מתוך כל הדאטה, בחרנו רק את הרשומות בהן המשתנה הנמדד הוא PRCP והשתמשנו בעמודות FIPS, Year, Variable, Value, כאשר העמודה FIPS מתקבלת ע"י חיתוך שני התווים הראשונים של העמודה StationID. ביצענו Join לפי FIPS עם הטבלה GeoData (הסבר בחלק 1) וביצענו את הפעולות של שלב Pre כפי שפורטו לעיל. התוצאות המתקבלות נכתבות לשרת בטבלאות זמניות, כל אחת מתאימה ליבשת נפרדת.
   b. Post: ביצענו את הפעולות של שלב Post כפי שפורט לעיל על כל טבלה, ומכל טבלה לקחנו רק את הרשומות עם ערך Year בטווח (1993, 1956).

הגרף שהתקבל:



הגרף מתבסס על 100,000,000 רשומות עם מדידות משקעים (PRCP) שנמדדו בסין, ארה"ב, צ'כיה, צ'ילה וצרפת בשנים 1956-1993. מהגרף נוכל להסיק כי ממוצע המשקעים באסיה נוטה להיות נמוך

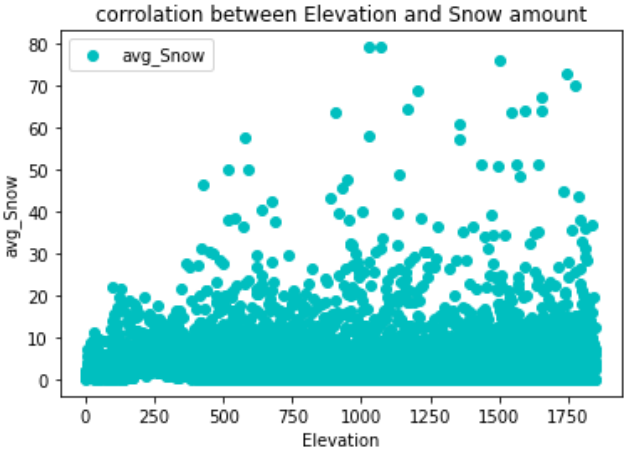
יותר מממוצע המשקעים באמריקה ואירופה. למרות זאת, ניתן לראות הפכפכות בין המדידות באירופה למדידות באמריקה בשנים שונות.

3. **Spatial based insight:** עבור תחנות בארה"ב, בדקנו האם יש קורלציה חזקה בין כמות השלג הנמדדת לבין גובה התחנה המודדת. השלבים:

a. Pre: מתוך כל הדאטה, בחרנו רק את הרשומות שבהן המשתנה הנמדד הוא SNOW והחודש הוא ינואר. בנוסף, מתוך הקובץ GHCND-Stations המצורף, יצרנו DataFrame נוסף המכיל את ה- ID של תחנת המדידה והגובה שלה. בין שני ה- DataFrames שהתקבלו ביצענו Join לפי StationID וביצענו את הפעולות של שלב Pre כפי שפורטו לעיל. התוצאה המתקבלת נכתבה לשרת בטבלה זמנית.

b. Post: ביצענו את הפעולות של שלב Post כפי שפורט לעיל, ולקחנו רק את הרשומות עם ערך Elevation בטווח [0 ,1850] וכמות שלג ממוצעת בטווח (0 ,82) וזאת כדי להסיר נתונים חריגים.

הגרף שהתקבל:



שאלת המחקר שעניינה אותנו לבדוק בחלק זה היא האם ישנה קורולציה בין גובה תחנת המדידה לבין כמות השלג היורדת בה. להפתעתנו מהחיזוי עולה כי אין קורולציה. הדבר הפתיע אותנו, והינו מאוד לא אינטואיטיבי שכן אנו מצפים שככל שהגובה עולה כך גם הסיכוי לשלג , תופעה זו אפילו יותר מוזרה היות והיא נצפתה על ידינו במגוון של מדינות בארצות הברית, ממדינות חמות יותר בדרום ועד מדינות קרות יותר בצפון. הגרף מתבסס על יותר מ- 100,000,000 רשומות שנאספו ממדידות בתחנות שונות בארה"ב. התוצאות שקיבלנו מראות מביאות אותנו להבנה כי תפיסת העולם לפיה מקומות מושלגים תמיד נוטים להיות גבוהים מוטעית, משום שבמדינות קרות כמו ארה"ב גם במקומות נמוכים יתכנו כמויות שלג לא מבוטלות.