ניהול מידע מבוזר - פרויקט חלק ב'

מגישים:

ירין בן שטרית 206230021

רועי פאפו 316327451

תומר פרץ 318295029

חלק 1 - Extract, Transform, Load:

הדאטה שמגיע מהשרת נשמר עם הסכמה:

```
StructType([StructField('StationId', StringType(), False),
            StructField('Date', IntegerType(), False),
            StructField('Variable', StringType(), False),
            StructField('Value', IntegerType(), False),
            StructField('M_Flag', StringType(), True),
            StructField('Q_Flag', StringType(), True),
            StructField('S_Flag', StringType(), True),
            StructField('ObsTime', StringType(), True)])
```

בנוסף, על ה- DataFrame מופעלות הטרנספורמציות הבאות:

1. הסרת העמודות M_Flag, S_Flag, ObsTime: החלטנו שעמודות אלו אינן מוסיפות ערך מוסף לצורך הניתוח אותו אנו מבצעים בשלבים הבאים.
2. הסרת הרשומות בהן הערך בעמודה Q_Flag אינו Null: זאת על מנת להבטיח שכל הרשומות בהן נשתמש לצורך ניתוח ולמידה יכילו ערכים שהתקבלו ממדידות תקינות בלבד.
3. הסרת העמודה Q_Flag: לאחר הסרת הרשומות הבעייתיות בשלב הקודם, אין צורך בעמודה זו, שכן כל הרשומות שנותרו יכילו ערך Q_Flag = None.
4. הוספת עמודת fullDate שמתרגמת את עמודת Date הנתונה לפורמט תאריך בעזרת הפונקציה to_date() של Spark, זאת כדי לשלוף בצורה נוחה את התאריך של כל רשומה.

האבחנות עליהן עבדנו בחלק 2 דרשו לבצע ניתוח של הנתונים לפי מדינות ויבשות. לשם כך, יצרנו קובץ CSV בשם GeoData בעל 3 עמודות: עמודת FIPS המכילה את קוד ה- FIPS של המדינה, עמודת Continent המכילה את היבשת אליה שייכת המדינה ועמודת CountryName המכילה את שם המדינה. בחרנו להשתמש בקובץ זה כדי לאתר ולשלוף את הנתונים הרלוונטיים לאבחנות שלנו בקלות.

שתי פונקציות עזר בהן השתמשנו לצורך תקשורת עם השרת הם:

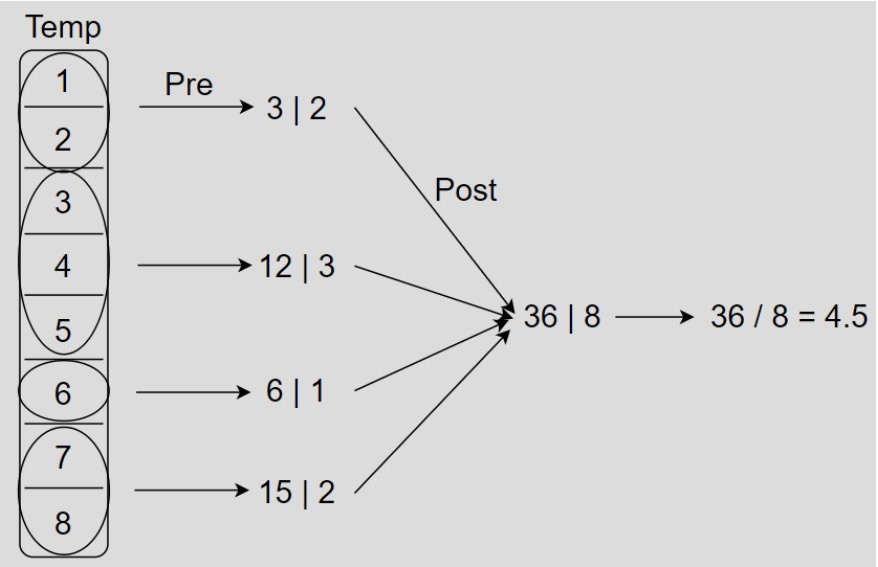
1. writeToServer(): כותבת spark.DataFrame לטבלה קיימת בשרת.
2. read_df_from_sql_server(): קורא טבלה קיימת בשרת ומחזיר spark.DataFrame עם המידע.

בחרנו להשתמש בסטרימינג דינאמי המורכב מ- readStream ו- writeStream. המידע מגיע ב- Batch'ים בגודל של 5,000,000 רשומות, ועבור כל אחד מתבצעת הטרנספורמציה לעיל. לאחר בדיקה אמפירית גילינו שלאחר פרק זמן של שעתיים (7200 שניות) מתקבלות מעל 100,000,000 רשומות כנדרש.

ביצענו פרגמנטציה של שרת ה-Kafka כך שכל אבחנה תתבסס על חלק מעמודות הטבלה המקורית בפרגמנט יחיד (לפי המונחים שנלמדו בקורס נוכל לומר שביצענו חלוקות אופקית ואנכיות לסירוגין).

חלק 2 - Data Analysis:

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

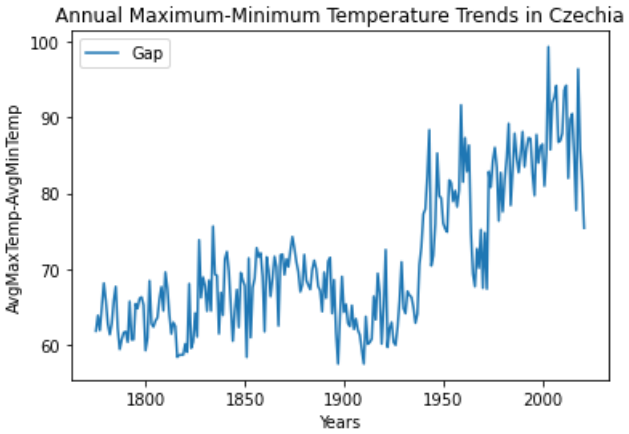


העמודה Temp מכילה 8 רשומות עם הערכים 1, 2, ..., 8. כל אליפסה מייצגת Batch המכיל את הרשומות שנמצאות בתוכו. המטרה - מציאת ממוצע הערכים של כל הרשומות. בשלב ה- Pre, עבור כל Batch, מחושבים שני ערכים: סכום ערכי הרשומות ומספר הרשומות ב- Batch. למשל עבור ה- Batch הראשון המכיל את הרשומות 1, 2, הערכים המתקבלים 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.

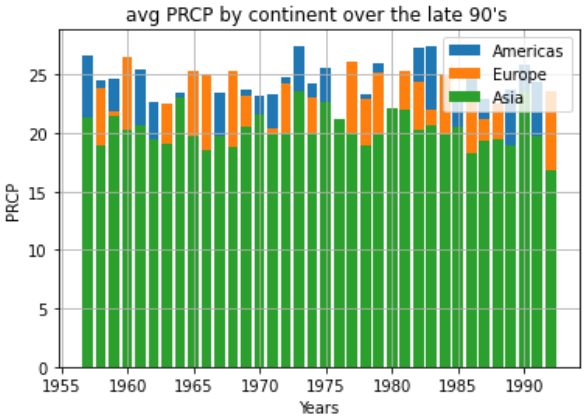
הגרף שהתקבל:



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

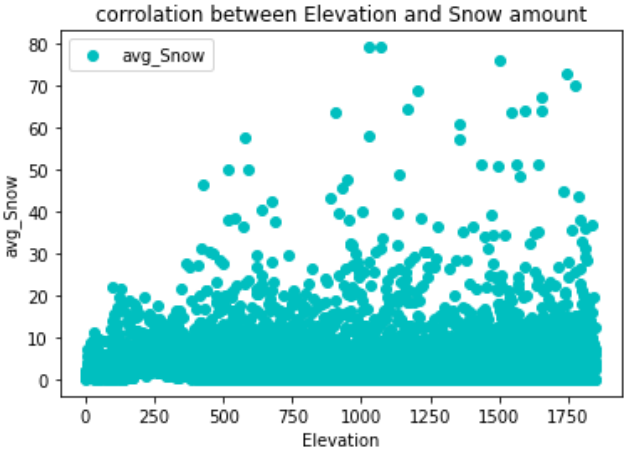
יותר מממוצע המשקעים באמריקה ואירופה. למרות זאת, ניתן לראות הפכפכות בין המדידות באירופה למדידות באמריקה בשנים שונות.

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.

חלק 3 - Learning:

במשימת החיזוי, בחרנו בנתוני מדידות המגיעים מקנדה בלבד, וזאת משתי סיבות:

1. כמות התחנות בקנדה היא 8910, המדינה השנייה בגודלה מבחינת כמות תחנות מדידה.
2. זוהי מדינה הומוגנית מבחינת אופי האקלים. האקלים במדינה קר ומושלג יחסית, לכן ציפינו לקבל דגימות דומות מבחינת ההתפלגות שלהן, דבר שיאפשר לבצע חיזוי אמין עבור דגימות חדשות.

בנוסף, הפיצ'רים שבחרנו הם הטמפרטורה המקסימלית (TMAX), הטמפרטורה המינימלית (TMIN) וכמות השלג (SNOW). בחרנו למצע את הפרמטרים האלו על פני חודש עבור כל תחנה על מנת לייצר יציבות בדאטה ולמתן מקרי קצה. כך גם כמות הגשם שאנו מנסים לחזות היא כמות גשם ממוצעת על פני חודש עבור תחנה ספציפית.

משימת החיזוי התחלקה למספר שלבים:

1. **עיבוד המידע:** שלב זה מתחלק לשני חלקים - Preprocessing & Postprocessing, כפי שתוארו בחלק השני של הפרויקט. חלק Preprocessing נמצא ב- mlib_data וחלק Postprocessing נמצא ב- mlib_data_post, אלו שתי טבלאות שנמצאות על השרת שיצרנו כדי לאחסן את המידע הרלוונטי לחלק זה. בנוסף, בחלק ה- Preprocessing יצרנו טבלה בשם StationsForClustering המכילה רק תחנות מקנדה, בנוסף לנתוני גובה וקווי אורך ורוחב שלהן.
2. **K-Means:** בהמשך להנחתנו הראשונית של ההומוגניות של הדגימות מקנדה, החלטנו לנסות לעדן ולשפר את ההומוגניות של הדגימות ע"י בחירת תחנות שהמרחק האוקלידי ביניהן קטן מספיק. חישבנו ידנית את סכום המרחקים של כל הנקודות ממרכז הקלסטר שלהן. עבור ערכי K שונים, הרצנו את אלגוריתם K-Means כך שיתקיימו שני תנאים:
   a. אחד מהקלסטרים שנוצרו מכיל כמות של אלפי דגימות.
   b. סכום הריבועים הכולל של כל דגימה ממרכז הקלסטר שלה נמוך ביחס לסכום המרחקים הכולל בשאר הקלסטרים, עבור כל קלסטר.

   לאחר מכן, וידאנו את אמינות המתודה שפיתחנו ע"י אימון הדאטה על מודל של רגרסיה לינארית ובדיקת מדד ה- RMSE עבור סט מבחן שנבחר באקראי בגודל של 20% מתוך כל הדגימות של הקלסטר מהתנאי הקודם. אכן ראינו שקיימת קורלציה גבוהה בין קלסטר שסכום ריבועי המרחקים ממרכזו נמוך, כמות הדגימות שבו גבוהה מספיק (אלפי דגימות, כפי שנדרש בתנאי a) לבין הכללה טובה של מודל הרגרסיה הלינארית שאומן על 80% מהדאטה בקלסטר ונבחן על 20% הנותרים.
   הפיצול לסט האימון והמבחן נבחר בצורה אקראית.
   מתוך כל תתי-הקבוצות שמקיימות את התנאים לעיל, בחרנו את תת-הקבוצה שלדעתנו מקיימת את האיזון הטוב ביותר בין כמות הדגימות לבין מדד RMSE. הדאטה שנבחר אוחסן בטבלת SQL בשם model_data, זוהי הטבלה הסופית בה אחסנו את הנתונים עבור אימון ובחינת המודל.
   הסיבה שבחרנו להשתמש במודל זה היא שזהו מודל פשוט, אינטואיטיבי להבנה וחשנו כי הוא ממדל נכונה את המציאות, היות ואנו מעריכים שמרחק אוקלידי בין אזורים גיאוגרפיים מעיד על התנהגות אקלימית דומה.

3. **אימון ושערוך המודל:** עד כה, בחרנו רק את הדאטה אשר נשתמש בה כדי לבצע אימון וחיזוי של המודל. המודל שבחרנו לבצע את החיזוי באמצעותו הוא מודל הרגרסיה הלינארית. הסיבה שבחרנו במודל זה היא הפשטות שלו והאפקטיביות שלו עבור הדאטה שבחרנו. בחרנו מסט הדגימות הכולל 20% בצורה רנדומלית שישמשו כסט המבחן ו- 80% כסט האימון. הסיבה לבחירת פיצול זה היא ש- 80% מהווה כמות מספקת לאימון המודל, ומצד שני, 20% מהווה כמות גדולה מספיק לבדיקת אמינות המודל על מספר רב של דגימות ותרחישים שונים. ה- RMSE שהתקבל עבור סט האימון (שנבחר רנדומלית בכל פעם) הוא 2.1.

המסקנות שקיבלנו מראות על קיום קשר לינארי בין הפיצ'רים שנבחרו.

<u>בונוס:</u>

בחלק זה, בחרנו לנסות לחזות את כמות השלג שיורדת בחודש בממוצע. באופן דומה, מאותן סיבות שפורטו בחלק 3, גם כאן בחרנו להתמקד רק במדידות שהגיעו מקנדה.

סט הפיצ'רים בו בחרנו להשתמש הוא טמפרטורה מינימלית (TMIN) וטמפרטורה מקסימלית (TMAX). בחרנו בסט פיצ'רים זה כי הנחנו שקיימת קורלציה גבוהה בין טמפרטורות מינימליות ומקסימליות לבין כמות השלג המצטברת על הקרקע.

מכיוון שיש כמות קטנה של רשומות המכילות מידע על שלג (SNOW) נאלצנו להשתמש בפחות דאטה כדי לבצע את החיזוי.

משימת הבונוס התחלקה למספר שלבים:

1. **עיבוד המידע + K-Means:** מתבצעים באופן זהה לחלק 3 של הפרויקט. שלב ה- Preprocessing מאוחסן ב- bonus_data ושלב ה- Postprocessing מאוחסן ב- bonus_post.
2. **אימון ושערוך המודל:** עד כה, בחרנו רק את הדאטה אשר נשתמש בה כדי לבצע אימון וחיזוי של המודל. המודל שבחרנו לבצע את החיזוי באמצעותו הוא Gradient Boosted Trees. הסיבה שבחרנו במודל זה היא האפקטיביות שלו עבור הדאטה שבחרנו, שהנחנו וראינו שאינו מתפלג בצורה ליינארית פשוטה כמו קודם, אלא בצורה מורכבת יותר, שדרשה מודל מורכב יותר על מנת לתפוס את הקשרים המורכבים יותר הקיימים בו. בחרנו מסט הדגימות הכולל 20% בצורה רנדומלית שישמשו כסט המבחן ו- 80% כסט האימון. הסיבה לבחירת פיצול זה היא ש- 80% מהווה כמות מספקת לאימון המודל, ומצד שני, 20% מהווה כמות גדולה מספיק לבדיקת אמינות המודל על מספר רב של דגימות ותרחישים שונים. ה- RMSE שהתקבל עבור סט האימון (שנבחר רנדומלית בכל פעם) הוא 5.5.

ניתן לראות כי יש קשר חזק מאוד בין טמפרטורות נמוכות לבין ירידת שלג. הקשר הוא חזק למרות שטמפ' נמוכה לא גוררת ירידת שלג באופן ישיר.