# מעבדה לניתוח נתונים פיננסיים

מגישים: טום גבע (318334901), גיא גלאט (206745440), הלל שרביט (318757572)

## חלק 1 – ניקוי ועיבוד נתונים ב-WRDS

1. פתחו מחברת עבודה ב-WRDS.
   * בוצע בWRDS
2. העלו את הקובץ final\_task\_bad\_data.pickle לסביבת העבודה של המחברת.
   * בוצע בWRDS
3. טענו את הקובץ ל-pandas DataFrame.
   * בוצע במחברת קוד
4. תארו מה הן הבעיות שיש בנתונים בפירוט לפי עמודות.
   * מצאנו בנתונים את הבעיות הבאות:
     + קיימות 6 שורות בהם הערכים לא הגיוניים.
       - תאריכי השורות בהם הClose לא הגיוני הם:
         * 2022-09-23 (9999 אשר גבוה מHigh ששווה ל100.851316)
         * 2022-09-26 (0 אשר נמוך מLow ששווה ל97.741452)
         * 2022-09-27 (9999 אשר גבוה מHigh ששווה ל99.316317)
         * 2023-05-01 (20- אשר נמוך מLow ששווה ל101.279918, בנוסף הClose לא יכול להיות שלילי)
       - תאריכי השורות בהם הOpen נמוך מהLow הם:
         * 2023-02-16 (Open שווה ל97.200627 וLow שווה ל105.356631)
         * 2023-04-25 (Open שווה ל89.057157 וLow שווה ל97.691612)
     + קיימות 77 שורות בנתונים אשר אחד או יותר מערכיהם מהווים outlier (זו כמות גדולה במיוחד של נתונים מכדי להיפטר מהם מתוך ה502).
     + אותם 77 השורות הללו נמצאות בתאריכים שמאוד קרובים אחד לשני (כאשר עושים ממוצע של מרחק בימים של התאריך של הoutlier הנוכחי לoutlier הבא מקבלים ממוצע של 9.46 ימים, אבל כאשר מורידים את הoutliers מתוך המרחקים הללו הממוצע הוא 2.12 ימים, מרחק בין תאריכים שהוא קיצוני מידי בעינינו כדי להיפטר מכמות כזו של נתונים).
     + קיימת שורה אחת בה בעמודת volume הערך ריק.
5. כיצד אתם מציעים להשלים ולתקן את הנתונים, בהנחה שמשימתנו הצפויה היא ניתוח סטטיסטי היסטורי של המניה? פרטו למה בחרתם בשיטה זו לכל אחת מהעמודות בנפרד.
   * בעקבות הסוגים השונים של
   * את הערכים הלא הגיוניים נתקן בדרך הבאה:
     + קיימים 2 סוגים של ערכים לא הגיוניים בנתונים
       - הסוג הראשון בו הClose לא הגיוני טופל בצורה הבאה:  
         אם השורה הנוכחית (הלא הגיונית מהסוג הזה) לא השורה העוקבת של השורה הקודמת (הלא הגיונית מהסוג הזה) בנתונים שטופלה, אז הClose בשורה הזו מוגדר להיות ערך ריק והוא יטופל בהמשך.  
         אם השורה הנוכחית היא כן השורה העוקבת, היא מסומנת להסרה מאוחר יותר.
       - הסוג השני בו הOpen נמוך מהLow הוגדר מחדש ערכו של הLow להיות הקטן מבין הClose והOpen.

לבסוף כל השורות שסומנו להסרה מוסרות מהנתונים, בפועל הוסרו 2 שורות בלבד.  
בחרנו לטפל כך בערכים הלא הגיוניים מכיוון שאנחנו רוצים לשמור כמה שיותר נתונים לעבוד איתם בעת הניתוח הסטטיסטי היסטורי, אבל לא רוצים לייצר מעין הטייה אשר תשפיע על הניתוח בהמשך, לכן חלק מהשורות בהם הערכים היו לא הגיוניים "הוצלו" והם יהוו הטייה מועטה יותר מאשר היוו לפני כן.

* + - את הערכים הריקים שנשארו בנתונים (1 שנוצר מתוך התהליך של הערכים הלא הגיוניים ו1 שהיה מלכתחילה) החלטנו להשלים כהחציון של הערכים בעמודה בה יש ערך ריק בשורות שקדמו לשורה בה יש ערך ריק, בנוסף לכך אותם ערכים מהם חושב החציון באים משורות אשר הערכים בשאר העמודות דומים לערכים בשורה בה יש ערך ריק, אם לא קיימות מספיק שורות שעונות לקריטריונים הללו החציון מחושב מתוך כל השורות שקדמו לשורה הנוכחית בלי להתחשב בדימיון של השורות הללו לשורה הנוכחית.  
      גם כאן בחרנו לטפל בצורה הזו כדי לצמצם את ההטייה כמה שניתן, החציון מהווה חוסר הטייה, וחישובו של החציון מתוך שורות קודמות שדומות מספיק לשורה הנוכחית בשאר הערכים גם זה ווידוא נוסף המונע הטייה של הערך שמחליף את הערך הריק להיות דומה יותר לערכים שאינם דומים לו.
    - את כל השורות בהם אחד או יותר מהערכים מהווים outlier תיקנו באמצעות ביצוע winsorize על 5% מהערכים בכל צד של ההתפלגות של אותם ערכים(על 5% מהערכים אשר האחוזון שלהם היה מעל 95 ועל 5% מהערכים אשר האחוזון שלהם היה מתחת ל5)

1. כתבו קוד לתיקון הנתונים בדרכים שהצעתם ושמרו את הנתונים המתוקנים בקובץ pickle בשםfixed\_data.pickle.
   * בוצע במחברת קוד.
2. האם הפתרון שהצעתם חסין בפניdata leakageכפי שהגדרנו אותו בכיתה? אם כן הסבירו למה כן, ואם לא הסבירו למה לא ואיך הייתם פותרים זאת,והציעו פתרון אחד (ללא צורך לממש בקוד).
   * הפיתרון שלנו לא חסין מפני data leakage בעקבות הטיפול בכל השורות שהחילו outlier, אשר בוצע בעצם winsorize אשר מתייחס לכל השורות לפי ערכיהם ולא לפי סדר כרונולוגי.  
     כדי לפתור את זה בעינינו צריך להשתמש בסט נתונים יותר ספציפי במהלך היום, במקום שכל שורה תבטא יום שלם, כל שורה תבטא רבע מהיום או שעה ביום, ועל בסיס כך יהיה אפשר לסכם את היום לשורות תקינות יותר בהם מה שהביא לקיומם של הערכים הoutlier יקבל התייחסות קרובה יותר ובהתאם מניעה שלהם, כך אין צורך בכלל לבצע winsorize.
3. כעת טענו מ-CRSP את הנתונים היומיים (ohlcv, dividends, splits\_adjustment\_factor) של מניית Abercrombie & Fitch Co מהראשון לינואר 2018 עד הראשון לאפריל 2020.
   * בוצע במחברת קוד.
4. חשבו את ה-adjusted close של המניה בהתאם לספליטים והדיבידנדים שלה.
   * בוצע במחברת קוד.
5. בדקו מול ה-historical data ב- yahoo finance אם החישוב שלכם נכון (מומלץ לבדוק מול תאריכים ישנים שחולקו אחריהם לפחות שני דיבידנדים).
   * קיים הפרש מינורי אשר קטן עם הזמן ונובע מעיגול הערכים, זאת משום שהערכים שמחושבים ע"י yahoo finance מחושבים על בסיס הרבה יותר נתונים מאשר נגישים לשימושינו.
6. חשבו את ה- adjusted open, adjusted high ו- adjusted lowבהתאם ל-adjusted close.
   * בוצע במחברת קוד.