# 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.
   * + 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.

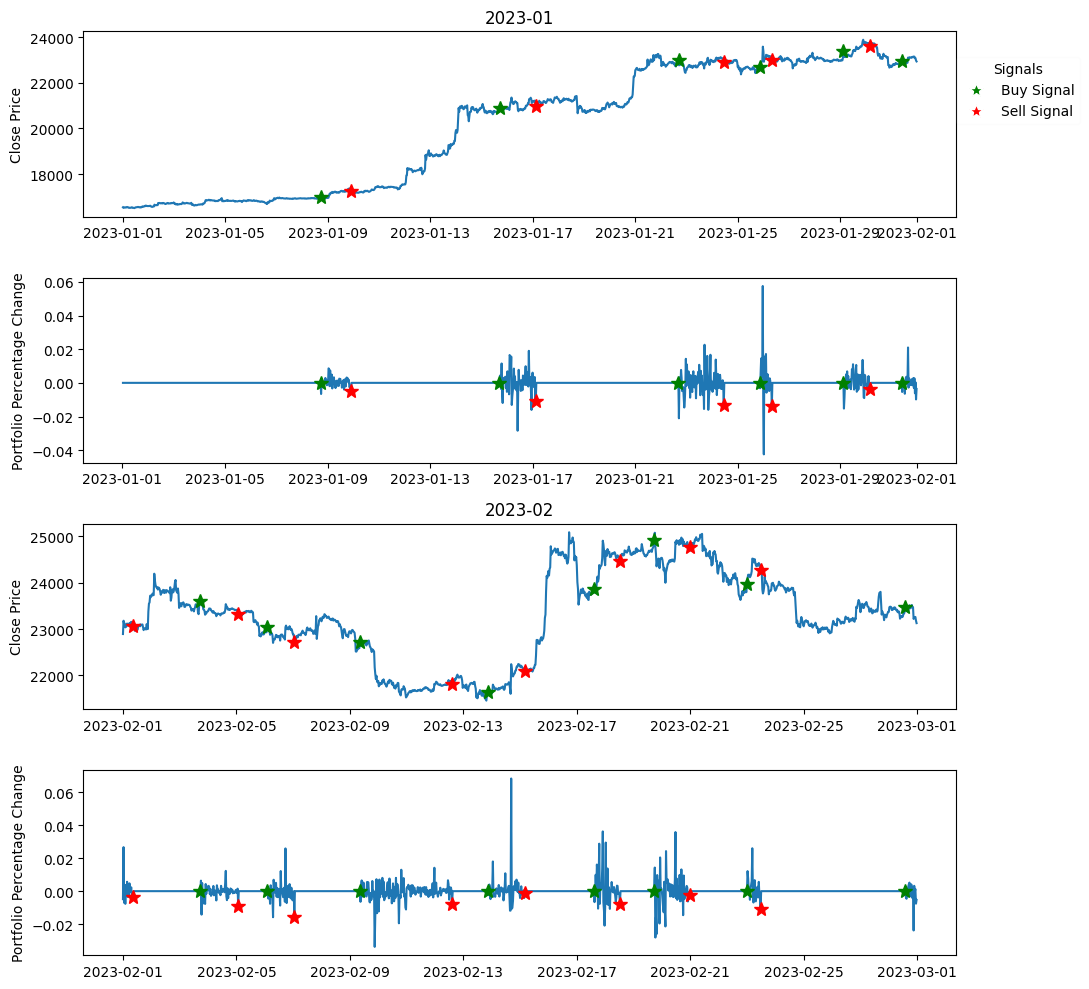
* + - את הערכים הריקים שנשארו בנתונים (1 שנוצר מתוך התהליך של הערכים הלא הגיוניים ו1 שהיה מלכתחילה) החלטנו להשלים כהחציון של הערכים בעמודה בה יש ערך ריק בשורות שקדמו לשורה בה יש ערך ריק, בנוסף לכך אותם ערכים מהם מחושב החציון באים משורות אשר הערכים בשאר העמודות דומים לערכים בשורה בה יש ערך ריק, אם לא 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)

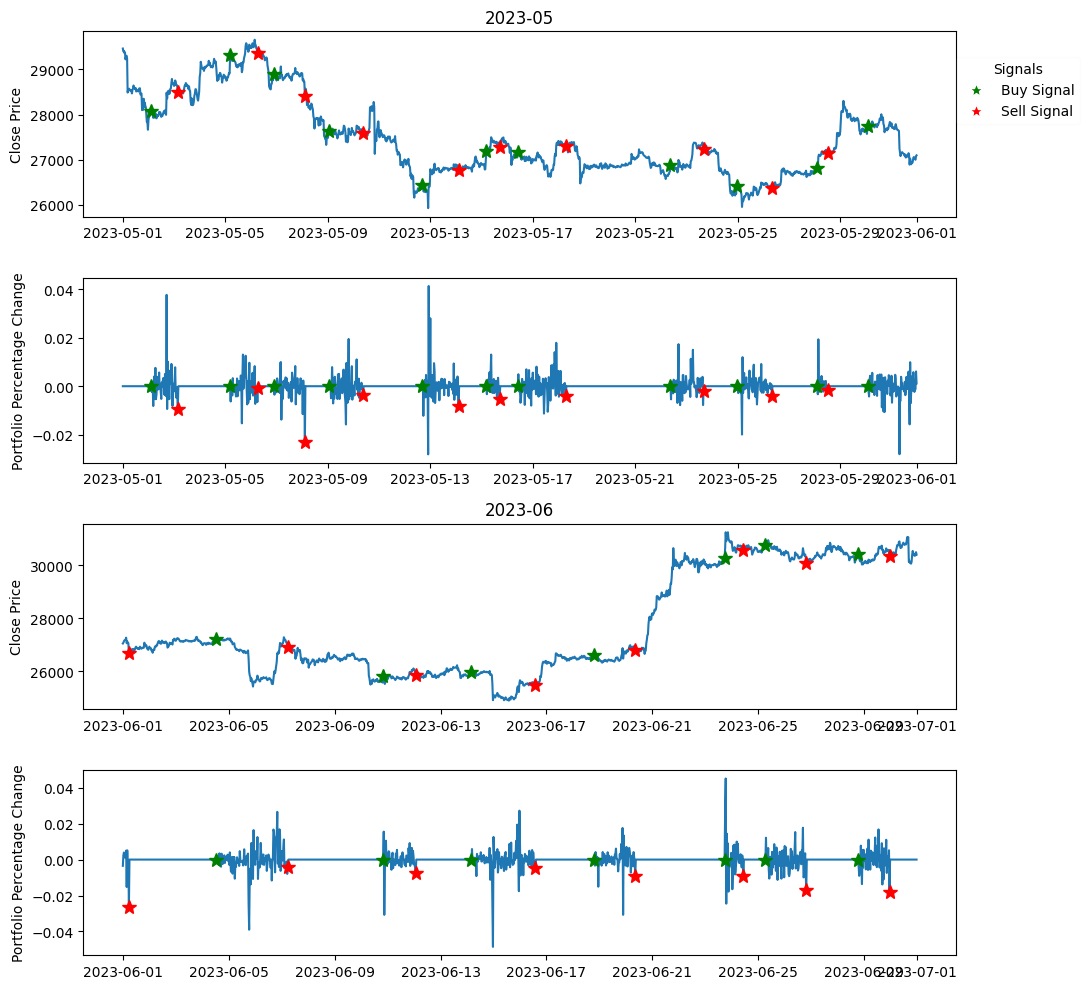
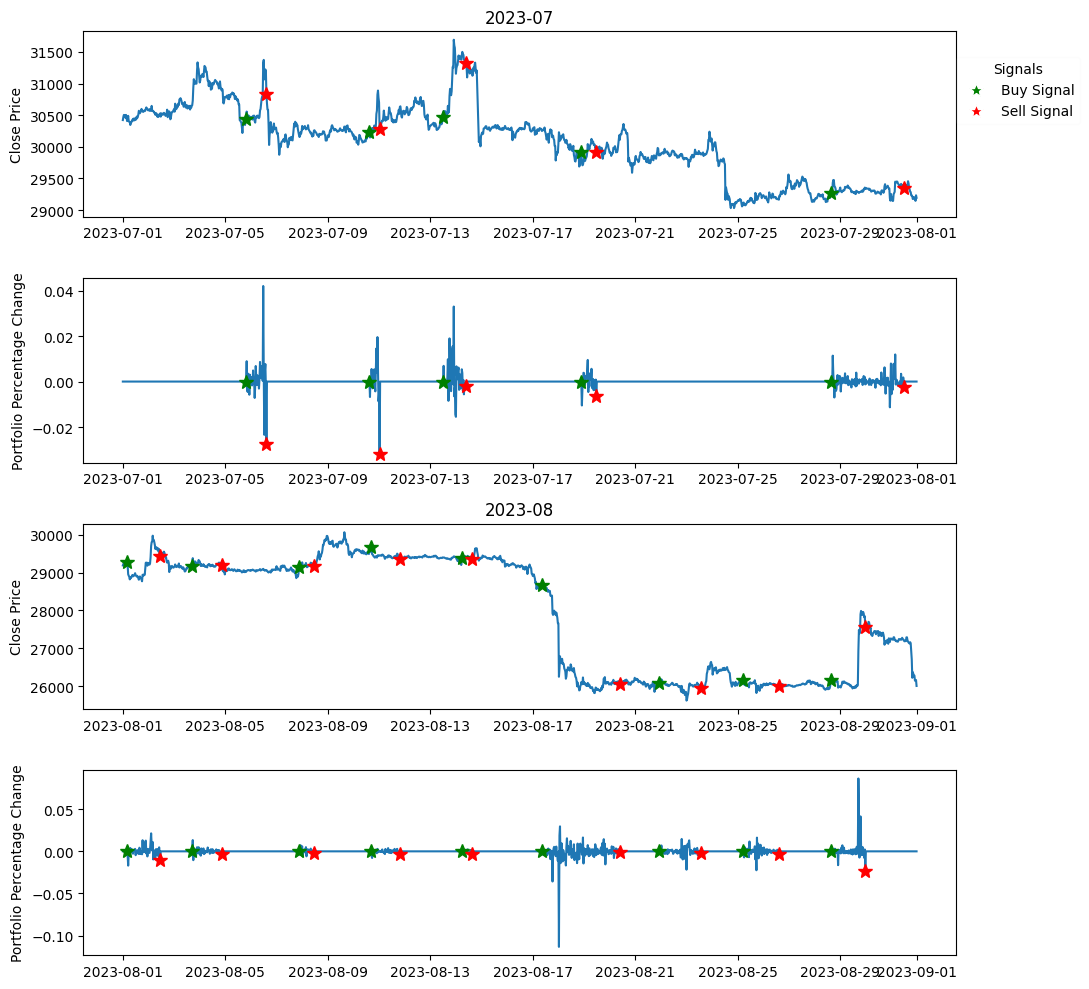
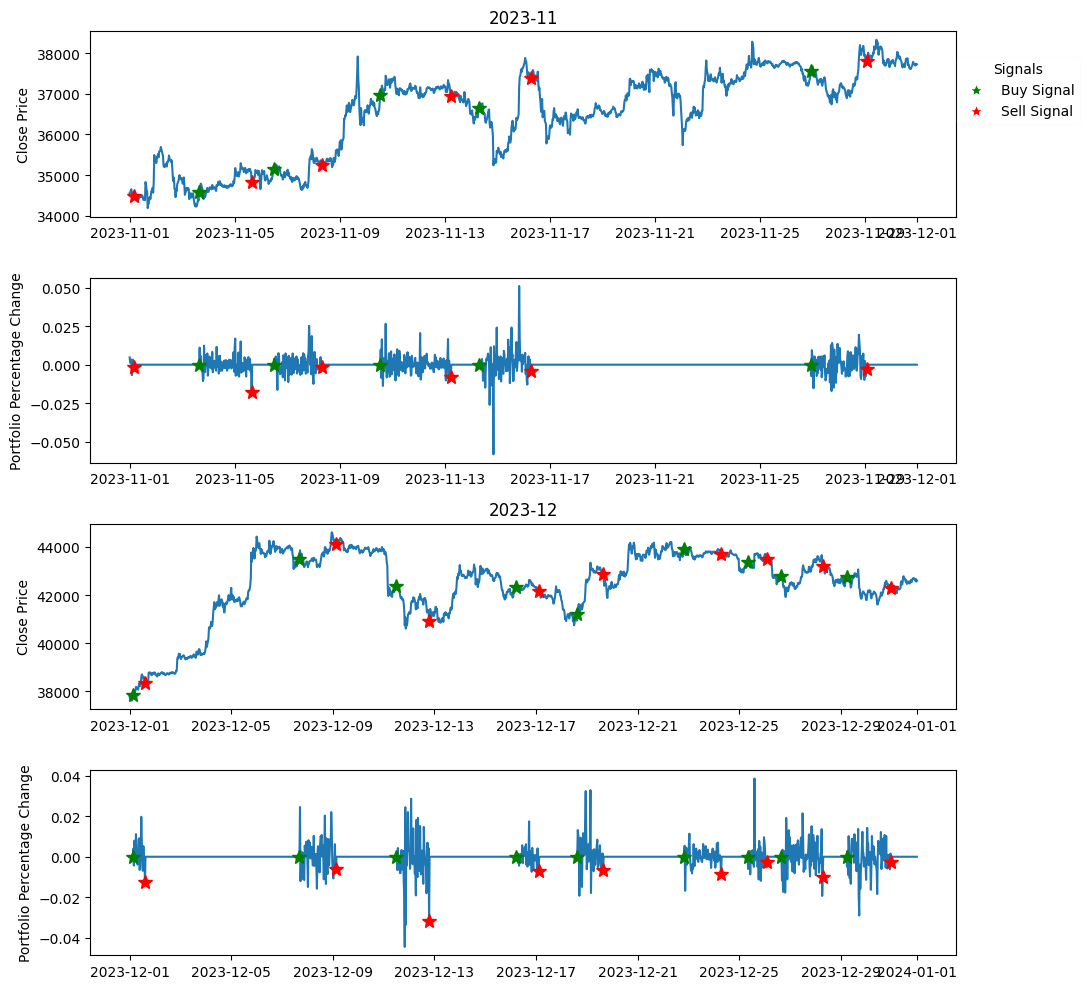
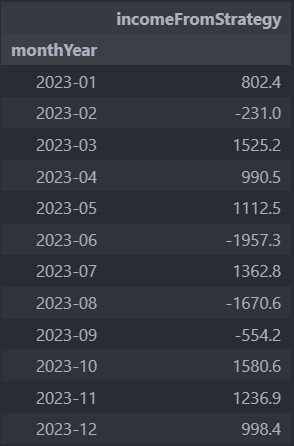
1. כתבו קוד לתיקון הנתונים בדרכים שהצעתם ושמרו את הנתונים המתוקנים בקובץ pickle בשםfixed\_data.pickle.
   * בוצע במחברת קוד.
2. האם הפתרון שהצעתם חסין בפני data leakage כפי שהגדרנו אותו בכיתה? אם כן הסבירו למה כן, ואם לא הסבירו למה לא ואיך הייתם פותרים זאת,והציעו פתרון אחד (ללא צורך לממש בקוד).
   * 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.
7. שמרו את המחברת בשם part1.ipynb והורידו אותה למחשב שלכם.

## חלק 2 – ניתוח טכני

1. צפו בסרטון על איך מתעשרים מניתוח טכני בקישור: התעשרות בקליק.
2. הורידו את הנתונים של BTCUSDT מבורסת Binance באמצעות הקוד שמימשנו בכיתה (שימו לב שאלו הנתונים של BTCUSDT.P ב-trading view) מהראשון לינואר 2023 עד הראשון לינואר 2024 באינטרוולים של חצי שעה.
   * בוצע במחברת קוד.
3. חשבו את כל האינדיקטורים הטכניים בהם המציג משתמש בסרטון וכתבו את הנוסחה המתמטית של אחד מהם. הסבירו בקצרה מה ההיגיון שעומד מאחוריו (למשל, ההיגיון של BB הוא להראות איפה המחיר נמצא ביחס לסטיות התקן שלו ובכך כמה המחיר הנוכחי חריג סטטיסטית).
   * **אינדיקטור UTBot**
   * הנוסחאות המתמטיות עליהן מתבסס האינדיקטור הם:
   * ההיגיון של UTBot הוא להראות שינויים במחיר שהם דרסטיים (לפי מקדם רגישות Key Value כלשהו המוגדר ע"י המשתמש) ביחס לתקופה האחרונה (לפי תקופה באורך כלשהו המוגדר ע"י המשתמש), כאשר שינוי דרסטי במחיר נקבע להיות לכל הפחות ערכו של מקדם הרגישות (Key Value) כפול ממוצע הגדלים של תחומי ההגדרה של מחיר בתקופה האחרונה.
   * **אינדיקטור STC**
   * הנוסחאות המתמטיות עליהן מתבסס האינדיקטור הם:

* + ההיגיון של STC הוא להראות את עקביות כיוון השינוי במחיר ביחס לתקופה האחרונה (לפי תקופה באורך כלשהו המוגדר ע"י המשתמש) דרך השוואה של ממוצע נע אקספוננציאלי של המחיר בטווח קצר (לפי תקופה באורך קצר כלשהו המוגדר ע"י המשתמש) למול בטווח ארוך (לפי תקופה באורך קצר כלשהו המוגדר ע"י המשתמש).

1. בדקו מול האינדיקטורים ב-trading view שהחישובים שלכם מדויקים.
   * נבדק ידנית מול אתר trading view, הסיגנלים זהים.
2. תארו את אסטרטגיית המסחר מבחינה לוגית ואת ההיגיון שעומד מאחוריה. כתבו האם אתם מאמינים בהגיון זה או לא ולמה, וכתוצאה מכך העריכו האם האסטרטגיה תרוויח או תפסיד.
   * האסטרטגיה משתמשת בשתי האינדיקטורים בצורה הבאה:
     + באינדיקטור UTBot האסטרטגיה משתמשת פעמיים, פעם אחת כדי להביא אינדקציה לקנייה ופעם שנייה כדי להביא אינדיקציה למכירה, לכל פעם מוגדרת גודל תקופה שונה אך בשתיהם מוגדרות רמת רגישות זהה של 2, כלומר, באסטרטגיה שינוי דרסטי במחיר יהיה לכל הפחות כפול מממוצעי הגדלים של תחומי ההגדרה בתקופות שהוגדרו:
       - האינדיקטור הקובע קנייה לוקח בחשבון תקופה יחסית ארוכה (300 אינטרבלים), בהתאם לכך ההחלטה של האינדיקטור זהירה במיוחד.
       - האינדיקטור הקובע מכירה לוקח בחשבון תקופה קצרה במיוחד (אינטרבל 1), בהתאם לכך ההחלטה של האינדיקטור פזיזה במיוחד.
     + באינדיקטור STC האסטרטגיה משתמשת פעם אחת, זאת כדי לבחון האם ההחלטות של שתי הUTBotים לא רק נכונות מזוית הראייה של האינדיקטורים האלו, אלה גם מבחינת עקביות כיוון השינוי במחיר, אם המחיר צונח באופן דרסטי מספיק כך שהUTBot של המכירה יתריע אבל הצניחה הדרסטית הזו היא מקרה חד פעמי ויוצא דופן שחורג מהמחיר הצפוי ע"י הטווח הקצר ו/או הארוך מתבצעת מכירה שתמנע רווחים פוטנציאליים במשך תקופה יחסית ארוכה אשר תיקח לUTBot של הקנייה להתריע כדי למנוע את ההפסד.
   * בעינינו לאסטרטגיה הזו יש פוטנציאל להצליח אבל מכילה בה גם סיכון גבוה התלוי בתקופה בה מתקיים המסחר, זאת משום שהאסטרטגיה נבחנת על תקופת זמן אחת בלבד, שגם היא עלולה לתת מענה מוטה לשאלת הרווחיות של האסטרטגיה, לכן אנחנו לא מצליחים להשתכנע שהאסטרטגיה היא רווחית.
3. ממשו את האסטרטגיה של המציג ומדדו את הביצועים שלה באמצעות הכלים שלמדנו. פרטו את התוצאות שקיבלתם.
   * לצורך החישוב ולצורך פרמטרים נוספים הצבנו ערך בסיס שרירותי בתחילת האסטרטגיה של 10,000.
   * ביצועי האסטרטגיה הם להלן:
     + הממוצע של הכנסות התיק באחוזים הוא – 0.003%
     + סטיית התקן של הכנסות התיק באחוזים היא – 0.467%
     + הכנסה מצברת נטו של התיק באחוזים היא –51.962%
   * בנוסף במהלך השנה האסטרטגיה ביצעה פעולות קניה ומכירה 88 פעמים, 49 מהם היוו רווח לעומת 39 מהם אשר היוו הפסד, (55.682% מפעולות הקניה ומכירה של האסטרטגיה היו חיוביות).
4. המציג המליץ להשתמש בפרמטרים שונים לחישוב האסטרטגיה. בחרו פרמטר אחד, הסבירו על מה הוא משפיע באסטרטגיה וכיצד השינוי שלו ישפיע על התנהגות האינדיקטור. הציעו עבורו ערך חלופי וחזרו על סעיף 5 עם השינוי של פרמטר זה.
   * בחרנו לשנות את הפרמטר STC Length של אינדיקטור STC מ80 ל90, במצב הזה אינדיקטורי הUTBot לא משתנים כלל, ואינדקטור הSTC לוקח בחשבון תקופה טיפה יותר ארוכה, מתוך כך הערך של הSTC מייצג קצת יותר על העקביות של כיוון השינוי ומונע יותר פעולות מסחר שיותר מסוכנות.
   * ביצועי האסטרטגיה הם להלן:
     + הממוצע של הכנסות התיק באחוזים הוא – 0.004%
     + סטיית התקן של הכנסות התיק באחוזים היא – 0.442%
     + הכנסה מצטברת נטו של התיק באחוזים היא – 59.127%
   * בנוסף במהלך השנה האסטרטגיה ביצעה פעולות קניה ומכירה 83 פעמים, 44 מהם היוו רווח לעומת 39 מהם אשר היוו הפסד, (53.012% מפעולות הקניה ומכירה של האסטרטגיה היו חיוביות).
5. עבור סעיפים 5 ו-6, הציגו בגרף את פעולות האסטרטגיה לאורך התקופה, ובגרף נוסף הציגו ניתוח סטטיסטי של הרווחים שלה (אחוזי השינוי ב-portfolio value).
   * 
   * A screenshot of a graph

     Description automatically generated
   * 
   * 
   * 
   * 
6. השוו את ההערכה שלכם לתוצאות שקיבלתם, האם הן תואמות? הסבירו.
   * כדי להשוות אם ההערכה שלנו מלפניי התוצאות תואמת לתוצאות שקיבלנו צריך למצוא דרך לקיים בחינה של האסטרטגיה על פני כמה תקופות, לכן חילקנו את רווחיות האסטרטגיה לפי חודשים וביצענו השוואה ביניהם כדי להשוות בין ההערכה המקורית שלנו לתוצאות עצמן.  
     כפי שניתן לראות בטבלה משמאל, 4 מתוך 12 חודשים בשנה מהווים הפסד מצטבר, במילים אחרות, שליש מהתקופות שנבדק השימוש באסטרטגיה בהן מניבות הפסד שהוא יחסית מהותי, ומכך ההערכה שלנו אכן תואמת לתוצאות היות ובעינינו הצלחה בסיכוי של 2 מתוך 3 זה סיכון גבוה מידי.
7. רשמו את התוצאות והגרפים שלכם בוורד ושמרו את הקוד שכתבתם בשם part2.ipynb

## חלק 3 – מציאת ארביטראז'

1. טענו את הנתונים של האופציות על חוזה QQQ בשנת 2023.
   * בוצע במחברת קוד.
2. נקו שורות בהן נפח המסחר קטן מ-100.
   * בוצע במחברת קוד.
3. מצאו את העסקה הרווחית ביותר בכל יום המבוססת על ארביטראז' PCP בהנחה שניתן לבצע עסקה אחת ביום הכוללת 100 יחידות מנכס הבסיס וקניה או מכירה של אופציה אחת מכל סוג.
   * בוצע במחברת קוד.
4. כעת חשבו מה הוא הרווח השנתי אותו הייתם מרוויחים מביצוע עסקאות אלה ורשמו את הפעולות שעליכם היה לבצע בחמשת הימים הכי רווחיים על מנת לעשות זאת.
   * בוצע במחברת קוד.
5. רשמו את התוצאות והפעולות שלכם בוורד ושמרו את הקוד שכתבתם בשם part3.ipynb.
   * קיבלנו כי 5 הימים בהם רווחיי הארביטראז' הם הגדולים ביותר וערכיהם הם: