**עיבוד הנתונים**

על מנת לדמות נתונים על קרנות פנסיה, השתמשנו במאגר מידע על קרנות נאמנות בשוק האמריקאי. מבחינת ההשקעה בשוק ההון, ניתן לראות בקרן פנסיה כסוג של קרן נאמנות. ישנם הבדלים בין המוצרים – כדוגמת אופי דמי הניהול (בקרן פנסיה קיימים גם דמי ניהול מהפקדה) או אגרות חוב מיועדות מהמדינה – אך בהיעדר נתונים מדויקים ומגוונים על קרנות פנסיה, קרנות נאמנות מהוות חלופה טובה.

מאגר הנתונים שהשתמשנו בו מבוסס על הנתונים שנמצאים ב-yahoo finance. את המאגר עצמו ניתן למצוא ב-(קישור), כאשר בקישור הנ"ל השתמשנו במאגר בשם Mutual Funds.

המאגר כולל מידע על קרוב ל-25,000 קרנות נאמנות שונות בשוק האמריקאי. הנתונים שבמאגר מעודכן לנובמבר 2020, כאשר על כל קרן (המיוצגת ע"י רשומה) יש אינפורמציה רבה מעודכנת לתאריך הנ"ל (מיוצג ע"י עמודות).

בשלב הראשון בעיבוד הנתונים, עברנו כל העמודות השונות – האינפורמציה שיש לכל קרן – ובחרנו איזה מידע רלוונטי עבור הסימולציה שלנו.

האינפורמציה שסיננו החוצה כולל בעיקר:

* מידע שרלוונטי בעיקר לקרנות של אגרות חוב, כמו למשל דרגות אגרות, מח"מ וכו'. הבחירה לסנן החוצה את המידע הזה נובעת מההיכרות שלנו עם שוק ההון. מטרת הפרויקט היא מציאת אסטרטגיה שמניבה תשואה טובה. קרנות אג"ח הן יניבו בממוצע תשואה נמוכה בהשוואה לקרנות מנייתיות (על אחת כמה וכמה בעשור 2010-2020 שבו יש לנו את הנתונים, בעקבות הריבית הנמוכה) ולכן אלו נתונים מיותרים מבחינתנו. לאור הצורך להישאר עם כמות סבירה של אינפורמציה על כל קרן, זאת הייתה בחירה סבירה.
* פרמטרים שונים מהשנים האחרונות: בשוק ההון ישנם דרכים שונים לאמוד סיכון ותשואה, כמו למשל אלפא, בטא ועוד. במאגר קיים מידע המידע שקיים במאגר הוא רק עבור 2020 (כלומר קיים מידע של "חמש שנים אחרונות" וכאלה) שהוא לא רלוונטי, ולכן החלטנו לסנן החוצה גם מידע שהוא נכון ל-2020 ומאוד תלוי זמן.

את המידע שנשאר לנו עבור כל קרן:

* שם הקרן (מזהה ייחודי) והחברה המנהלת
* תאריך הקמה
* קטגוריה (מעיד על טיב הקרן)
* דירוג הקרן, ההחזר שלה והסיכון שלה – לא ברור לנו לחלוטין על מה נתונים אלה מתבססים, אך החלנו להשאיר אותם בכל מקרה כדי להתייחס אליהם כמעין "המלצת הסוכן"
* דמי ניהול של הקרן ודמי ניהול ממוצעים בקטגוריה
* תשואת בחלוקה לרבעונים החל מ-2010 ועד ל-2020 (למעט הרבעון האחרון של 2020), תשואה של הקטגוריה ברבעון השלישי של 2020
* חלוקת ההשקעה של הקרן לפי נכסים: מניות, אג"ח, מזומן וכו'
* חלוקת ההשקעה של הקרן לפי סקטורים: פיננסים, אנרגיה, תעשייה וכו'
* היקף ההשקעות בקרן (שווי שוק)
* מכפילים שונים כמו מכפילים רווח

**התמודדות עם ערכים חסרים**

כאמור, מספר הקרנות במאגר המקורי הוא קרוב ל-25,000. לאחר שסיננו ובחרנו איזו אינפורמציה אנו רוצים על כל קרן, בדקנו האם יש לנו אכן אינפורמציה מלאה על כל הקרנות. גילינו שישנן לא מעט קרנות בעלות אינפורמציה חסרה ולכן היה עלינו לטפל בכך בדרך כלשהי.

מתוך כל הקרנות הנ"ל, על כ-8011 קרנות יש אינפורמציה מלאה. נזכיר שלכל הרצה של סימולציה אנחנו בוחרים אקראית מתוך המאגר מספר מצומצם של קרנות שישמש אותנו. אם נבחר לדוגמה בכל הרצה רק שלוש קרנות מתוך ה-8000, כמות הסימולציות השונות שנוכל לייצר היא למעלה מ-85 מיליארד (8000 בחר 3). בפועל נבחר כמות גדולה יותר של קרנות לכל סימולציה, ולכן כמות הסימולציות השונות שלנו היא גדולה מאוד.

על כן יכולנו לוותר על קרנות עם נתונים חסרים ולהישאר רק עם קרנות בעלות נתונים מלאים – וזאת על אף שמדובר בקיצוץ של שני שליש ממאגר הקרנות.

עמודה נוספת שנדרשנו אליה היא category\_return. ישנה עמודה בשם category עבור כל קרן, שמקטלגת את סוג הקרן. במאגר המקורי ישנן עמודות category\_return עבור כל שנה – מה הייתה התשואה עבור קרנות דומות באותה קטגוריה. המידע הזה זמין מ-2010 עד 2019, כלומר אין נתונים לשנת 2020. על מנת להיות עקביים עם שאר הנתונים החלטנו להשלים ידנית את המידע הנ"ל עבור 2020.

עבור כל קטגוריה של קרן, לקחנו את ממוצע התשואה של כל קרנות באותה קטגוריה עבור כל רבעון ב-2020. הנתון הנ"ל אפשר לנו להשלים את המידע החסר. נציין שלקחנו את המידע הנ"ל רק מתוך 8011 הקרנות ש-"שרדו" את הסינון הראשוני (כלומר שיש לנו מידע מלא עליהן) מתוך נוחות. כמובן שמדובר בקירוב עם כן שלא כולל את כל הקרנות שנמצאות בשוק, אך לדעתנו זה היה מספיק טוב כדי להשלים נתונים חסרים – מדובר על השלמה של שלושה רבעונים מתוך סה"כ 43 רבעונים שקיימים במאגר (2010-2020 ללא הרבעון האחרון של 2020), ולכן ההשפעה שלהם היא ככה"נ זניחה גם בתוצאות הסופיות שנקבל.

**עדכניות הנתונים שבמאגר והוספת רעש אקראי**

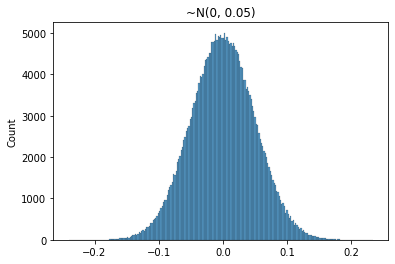
נשים לב שלמעט התשואה שנתונה לנו בחלוקה לרבעונים, כל הפרמטרים הנ"ל הם נתון בודד שנכון לשנת 2020. בסימולטור שלנו אנו מעוניינים בנתונים שונים עבור כל שנה, שידמו במידה מסוימת קרנות פנסיה. בעת בחירת מאגר הנתונים ידענו שנצטרך להתמודד עם בעיה זו, ומראש הגדרנו שהדרך להתמודד עם הנ"ל תהיה ע"י הוספת רעש אקראי.

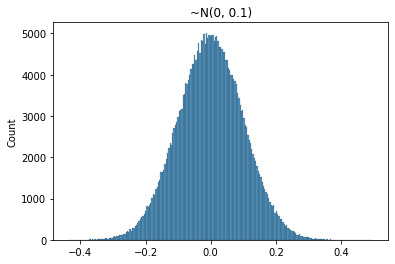
נדגים באמצעות הפרמטר דמי ניהול: יש לנו את דמי הניהול שנכונים לסוף שנת 2020. כדי לפרוס דמי ניהול שונים על פני כל התקופות (דמי ניהול שונים לכל שנה), העתקנו את דמי הניהול של סוף שנת 2020 לכל השנים ואז הוספנו רעש שמגיע מהתפלגות גיאוסינית. למעשה אנחנו מניחים שדמי הניהול מתפלגים נורמלית עם תוחלת מסביב למס' המקורי שנתון לנו. זאת הנחה מובלעת שאינה בהכרח נכונה במציאות (הרי שיכול להיות שבעקבות תחרות דמי ניהול יורדים למשך עשור מסוים – לא מתפלגים בצורה נורמלית).

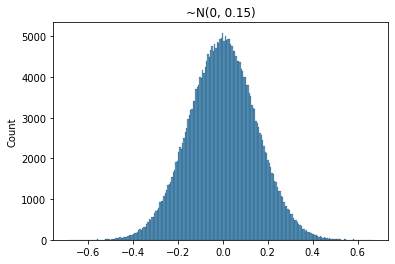
לכל פרמטר הוספנו רעש אקראי עם תוחלת אפס ועם שונות משתנה. הרעש נובע מהתפלגות בעלת רעש (מס' דוגמאות מסוים) כך שהוא נובע שינוי באחוזים: לדוגמה, עבור פרמטר של דמי ניהול, הגדרנו אחוזים יחסית קטנים (הרוב באזור ה-10%) ושינינו בהתאם, ואילו בפרמטרים אחרים הגדלנו את הרעש עוד יותר שיהווה עד ל-30%. להלן הדוגמאות הנפוצות שבהן השתמשנו.

(הסבר הסבר)

גרפים חשובים לרעש:







**נרמול הנתונים על אופי ההשקעה**

חלק מהנתונים שקיימים על כל קרן הן אופי ההשקעה, בחלוקה לשתי קבוצות.

הקבוצה הראשונה כוללת מידע על אופי הנכסים שהקרן מחזיקה: כמה מהנכסים הם מניות, אג"ח, מזומן וכו'.

הקבוצה השנייה כוללת מידע על אופי הסקטורים שבהם הקרן משקיעה: טכנולוגיה, פיננסים, תעשייה, אנרגיה, שירותים וכו'.

המידע הנ"ל מספק מידע חשוב על אופי הקרן, אך הוא נתון בערכים שונים. על מנת להפוך אותו לאחיד לכל קרן, נרמלנו אותו לאחוזים (מס' בין 0 ל-1) ואף הוספנו עמודת others במידת הצורך (אם הנתונים לא הסתכמו ל-1). זה נעשה כמובן ע"י החישוב של החלק היחסי של כל תת-תחום לפי הסכום בקבוצה שלנו.

לאחר שלב זה, יכולנו לקבל במפורש כמה אחוזים הקרן משקיעה בכל סקטור או כמה אחוזים מכל הנכסים של הקרן מושקעים בכל סוג נכס.

**הוספת עמודות נוספות**

פירוט על זה שהוספנו עוד עמודות של "דמי ניהול בשלוש השנים האחרונות" וכו'.