מטלת מנחה (ממ"ן) 15 - חובה

הקורס: תכנות וניתוח נתונים בשפת פייתון (20606)

חומר הלימוד למטלה : יחידות 14-13 **נושאי המטלה:** קבצים וניתוח נתונים

מספר השאלות: 1 **משקל המטלה:** 10 נקודות

סמסטר: 2025א **מועד אחרון להגשה:** 22.2.2025

(ת)

שימו לב:

- **יש לתעד את התכניות בתיעוד פנימי באנגלית בלבד (בתחילת התכנית התיעוד מסביר מה מבצעת התכנית באופן כללי ובמהלך התכניות התיעוד מסביר את הקוד) על פי תקן PEP 8.**
- **ניתן להוסיף פונקציות עזר מעבר לאלה הנדרשות באופן מפורש במטלה.**
- **יש לשים דגש על מענה פונקציונאלי של המשימה. ניתן ואף רצוי להציג את התוצרים הגרפיים (תרשימים) באופן אסתטי ככל הניתן אך הבדיקה תתמקד בצד הפונקציונאלי ובלבד שהתוצרים יענו על דרישות המשימה.**
- **אין להשתמש בחומר מתקדם או שלא נלמד בקורס.**
- **יש להשתמש בקבועים היכן שאפשר.**
- **יש להקפיד על הזחה (אינדנטציה - עימוד) נכונה, ועל שמות משתנים בעלי משמעות (באנגלית) ולפי המוסכמות בקורס.**
- **יש להקפיד על פורמט הפלט בדיוק כפי שמצוין בשאלה: איות נכון, אותיות גדולות וקטנות, רווחים, וכו.**
- **הגשת המטלה נעשית אך ורק בעזרת מערכת המטלות המקוונת שבאתר הקורס.**
- **אל תשכחו לשמור את מספר האסמכתא שתקבלו מהמערכת לאחר ההגשה.**

שאלה 1 (100 נקודות)

בדף המטלה תמצאו קובץ nasa.csv המכיל מידע על אסטרואידים בחלל הקרובים למערכת השמש של כדור הארץ. הקובץ הינו בפורמט csv.

הסבר על קבלת נתונים מקובץ בפורמט csv מפורט ביחידה 13 פרקים 13.1 ו- 13.4 באתר הקורס.

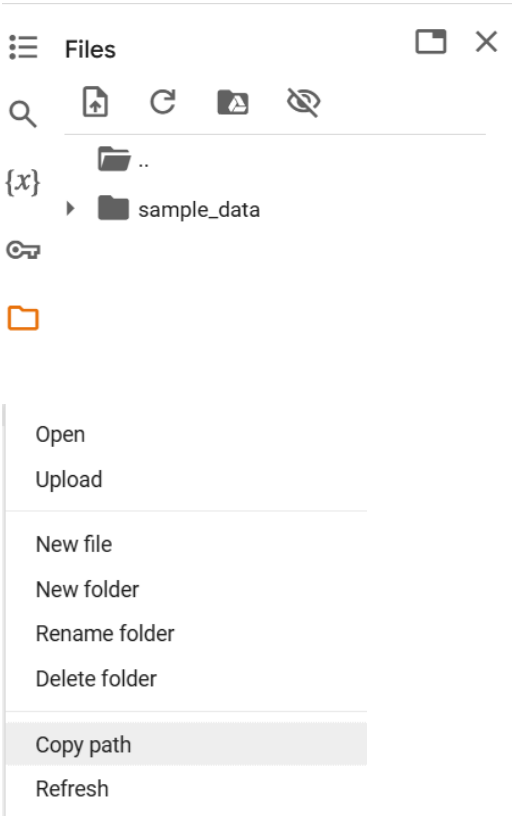
במסגרת מטלה זו, עליכם להכין ולטייב את הנתונים הגולמיים בקובץ ה- csv, לנתחם אותם ולהציג מממצאים עיקריים.

שלב טיוב והכנת נתונים

א. כתבו פונקציה בשם load_data המקבלת שם קובץ מסוג csv ומחזירה Data Frame של Pandas (יסומן כ df בהמשך מסמך זה). יש לטפל במקרי חריגות רלוונטיות (שם קובץ לא תקין, קובץ לא זמין, כשלון בטעינת הקובץ מסיבה אחרת וכיו"ב) ולהציג הודעה מתאימה בפלט הסטנדרטי.

<u>הנחה</u>: הקובץ nasa.csv נמצא בתיקיית הפרויקט ביחד עם קובץ הפיתון של פרויקט זה.

שימו לב! סטודנטים העובדים בסביבת Google Colab נדרשים להעלות את הקובץ למחברת בתיקיית sample_data. השתמש בשלוש נקודות מצד ימין של ה sample_data כדי להעלות את הקובץ. כדי לטעון אותו לתוכנית שלכם השתמשו בקליק ימני (על הקובץ) -> בחירה ב- copy path:



לפעמים יש ניתוק/הסרה של הקובץ ונדרש להעלותו מחדש למחברת.

ב. כתבו פונקציה בשם mask_data המקבלת df. הפונקציה תעדכן ותחזיר df מעודכן שבו כל האסטרואידים שתאריך הקרבה שלהם (Close Approach Date) לכדור הארץ הוא משנת 2000 ואילך.

ג. כתבו פונקציה בשם data_details המקבלת df. הפונקציה תנקה מה df את העמודות הבאות: Neo Reference ID, Orbiting Body ו- Equinox ותחזיר רשימה סטטית (tuple) הכוללת שלושה איברים: מספר שורות, מספר עמודות ורשימה הכוללת את כותרות הטבלה.

שלב ניתוח נתונים (יבוצע לאחר טיוב הנתונים שבוצע בשלב הקודם)

ד. כתבו פונקציה בשם max_absolute_magnitude המקבלת df ומחזירה tuple שבו האיבר הראשון מכיל את שם האסטרואיד (Name) בעל גודל הקרבה המקסימלי (Absolute Magnitude) ביחס למרחק לכדור הארץ והאיבר השני, מכיל את ערך הגודל הקרבה המקסימלי (Absolute Magnitude).

ה. כתבו פונקציה בשם closest_to_earth המקבלת df ומחזירה את שם האסטרואיד (Name) הקרוב ביותר לכדור הארץ על פי מרחקו מכדור הארץ בק"מ (Miss Dist.(kilometers)).

ו. כתבו פונקציה בשם common_orbit המקבלת df ומחזירה מילון (dict) שבו ערך המפתח הוא מזהה המסלול (Orbit ID) והערך הוא כמות האסטרואידים בכל מסלול.

ז. כתבו פונקציה בשם min_max_diameter המקבלת df ומחזירה את כמות האסטרואידים שהקוטר המקסימלי (Est Dia in KM(max)) שלהם הוא מעל לממוצע הקוטר המקסימלי (Est Dia in KM(max)) בקרב כלל האסטרואידים ב- df.

שלב הצגת נתונים (שלב זה יבוצע לאחר ביצוע שלבי טיוב הנתונים וניתוח הנתונים)

<u>הנחיות כלליות</u>

- יש לעשות שימוש ב- df שעודכן בשלב טיוב והכנת הנתונים.
- השתמש בחבילת matplotlib להצגת הנתונים.
- יש להקפיד להציג בכל תרשים: כותרת, מקרא, הסבר על כל ציר והצגת ערכי הצירים.
- יש להשתמש, ככל הניתן, בפונקציות שהוגדרו בשלב ניתוח הנתונים. במקרה הצורך, ניתן להשתמש בפונקציות עזר נוספות, כרצונכם.
- ניתן, ואף רצוי, להציג את התוצרים הגרפיים (תרשימים) באופן אסתטי ככל הניתן אך בדיקת שלב זה תתמקד בצד הפונקציונאלי תוך שימת דגש שהתוצרים עונים על דרישות המשימה.

ח. כתבו פונקציה בשם plt_hist_diameter המקבלת df ומציגה בגרף היסטוגרמה את **כמות** האסטרואידים בהתאם לקוטר **הממוצע** שלהם בק״מ. יש לכלול בגרף 100 טווחים רציפים.

ממוצע הקוטר של כל אסטרואיד הוא הממוצע בין ערך הקוטר הממוצע המינימלי לק״מ (Est Dia in KM(min)) לבין ערך הקוטר הממוצע המקסימלי לק״מ (Est Dia in KM(max)).

ט. כתבו פונקציה בשם plt_hist_common_orbit המקבלת df ומציגה בגרף היסטוגרמה את **כמות** האסטרואידים בהתאם למסלולם (Minimum Orbit Intersection). יש לכלול בגרף 10 טווחי מסלולים רציפים, החל מערך המסלול המינימלי לערך המסלול המקסימלי.

י. כתבו פונקציה בשם plt_pie_hazard המקבלת df ומציגה בגרף עוגה את **אחוז** האסטרואידים המסוכנים והלא מסוכנים על פי הסיווג (Hazardous) ב- df.

יא. כתבו פונקציה בשם plt_linear_motion_magnitude המקבלת df ובודקת האם יש קשר לינארי בין גודל הקרבה המקסימלי (Miss Dist.(kilometers)) לכדור הארץ של כל אסטרואיד לבין למהירות התנועה שלו בשעה (Miles per hour). לצורך כך, הציגו גרף רגרסיה לינארית פשוטה. בנוסף, הסבירו במילים שלכם, כהערה בתיאור הפונקציה, האם קיים מתאם בין שני משתנים אלו.

1.9. הקפידו לתעד כל פונקציה באמצעות docstring. ובהתאם למוסכמות התיעוד שהוצגו ביחידה

הגשה

1. הגשת הממ״ן נעשית בצורה אלקטרונית בלבד, דרך מערכת שליחת המטלות.
2. יש לכלול את הקובץ nasa_asteroid_ds.py בלבד (<u>ללא</u> קובץ nasa.csv). לחילופין, ניתן להגיש באמצעות מחברת Google Colab. במידה והחלטתם להגיש מחברת של Google Colab יש להוריד את המחברת בפורמת py ולהגישה. במקרה זה, יש לבדוק שקובץ ה- py רץ באופן תקין.
3. ארזו את קובץ הפתרון בקובץ zip **(ולא rar)** יחיד ושלחו אותו בלבד.
4. **אל תשכחו לשמור את מספר האסמכתא שקיבלתם מהמערכת לאחר ההגשה. אם לא קיבלתם מספר אסמכתא, סימן שההגשה לא התקבלה.**
5. שימו לב, אתם יכולים לשלוח שוב ושוב את המטלה במערכת, אם אתם רוצים לתקן משהו בה. כל הגשה דורסת את ההגשה הקודמת. **אבל עשו זאת אך ורק עד לתאריך ההגשה**. אחרי התאריך, ייחשב לכם כאילו הגשתם באיחור, גם אם ההגשה הראשונה הייתה בזמן! כמו כן, אם המנחה הוריד כבר את המטלה שלכם מהמערכת, לא תוכלו לשלוח עותק מעודכן יותר.

בהצלחה