תרגיל בית 1

חלק א' – רטוב (85%):

בחלק זה נתמקד בשלב הבסיס למיצוי ידע ראשוני מנתונים תוך שימוש בשפת התכנות python. לשם כך נשתמש במדדים סטטיסטיים בסיסיים.

סט הנתונים שנעבוד איתו בתרגול נתון במודל תחת השם "london.csv". הנתונים מכילים כ17,400 רשומות, כאשר כל רשומה מתעדת את: מספר האופניים החדשים שנשכרו בפרויקט בלונדון הדומה לתל אופן בתל אביב, פרמטרי מזג אוויר ותקופה בשבוע ובשנה.
כל רשומה חדשה מייצגת שעה עגולה, החל מה04/01/2015 בשעה 00:00 עד ל03/01/2017 בשעה 23:00.

אנחנו נתעניין רק בחמשת התכוניות:

משתנים רציפים	משתנים קטגוריאליים
hum (humidity) in %	**season** $\in \{0,1,2,3\}$
t1 (temperature) in $C°$	**is_holiday** $\in \{0,1\}$
cnt (no. of new rented bikes) $\in \{0,1,\dots\}$	

לתיאור מלא של הנתונים אנא בקרו ב:

https://www.kaggle.com/hmavrodiev/london-bike-sharing-dataset

תיאור המשימות

נגדיר:

- **סכום**:

$$X = \sum_{i=1}^{n} x_i$$

- **ממוצע**:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **חציון**:

חציון הוא הערך שמחצית מהתצפיות קטנות או שוות לו, ומחצית מהתצפיות גדולות או שוות ממנו. יהי n מספר התצפיות. אזי, אם n זוגי:

$$Med = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2}$$

אחרת:

$$Med = X_{\left(\left\lceil\frac{n}{2}\right\rceil\right)}$$

כאשר $X_{(i)}$ הינו האיבר הi ברשימת הערכים הממוינת בסדר לא יורד.

שאלה 1:

עליכם לחשב ערכי סכום, ממוצע וחציון לתכוניות hum, t1, cnt עבור: כלל האוכלוסיה, עונת הקיץ בלבד (season = 1), וימות החג בלבד (is_holiday = 1). על הפלט להיות מהצורה:

Question 1:

Summer:

hum: value of sum, value of mean, value of median

t1: value of sum, value of mean, value of median

cnt: value of sum, value of mean, value of median

Holiday:

hum: value of sum, value of mean, value of median

t1: value of sum, value of mean, value of median

cnt: value of sum, value of mean, value of median

All:

hum: value of sum, value of mean, value of median

t1: value of sum, value of mean, value of median

cnt: value of sum, value of mean, value of median

שאלה 2:

בשאלה זו נסתכל על הרשומות השייכות לימי החורף בלבד (season=winter). נחלק את רשומות החורף לשתי קבוצות – רשומות השייכות לימי החג (is_holiday=1), ולרשומות השייכות לימי החול (is_holiday=0).

נרצה לבדוק לכל קבוצה את היחס בין מספר האופניים שהושכרו (cnt), לטמפרטורה (t1). כלומר, לכל קבוצה של רשומות חורף (בזמן חג/בזמן חול) נתעניין בתכוניות cnt וt1 בלבד.

עליכם לחשב את הממוצע והחציון של מספר האופניים שהושכרו עבור שתי הקבוצות, כאשר נבצע שתי חלוקות:

$t_1 \leq 13.0$ ו - $t_1 > 13.0$. הפלט למשימה הזו יהיה מהצורה:

Question 2:

If t1<=13.0, then:

Winter holiday records:

cnt: value of mean, value of median

Winter weekday records:

cnt: value of mean, value of median

If t1>13.0, then:

Winter holiday records:

cnt: value of mean, value of median

Winter weekday records:

cnt: value of mean, value of median

דוגמה לפלט שמבוסס על מדגם קטן ושרירותי של תצפיות מתוך סט הנתונים:

```
Question 1:
Summer:
hum: 58197.0, 65.90826727066818, 67.0
t1: 16380.0, 18.55039637599094, 18.0
cnt: 1357301, 1537.1472253680633, 1265.0
Holiday:
hum: 6664.5, 76.60344827586206, 78.5
t1: 922.0, 10.597701149425287, 10.5
cnt: 58122, 668.0689655172414, 233.0
All:
hum: 251416.0, 72.18374964111398, 74.5
t1: 43552.5, 12.50430663221361, 12.5
cnt: 4019693, 1154.0892908412288, 834.0

Question 2:
If t1<=13.0, then:
Winter holiday records:
cnt: 393.76666666666665, 135.0
Winter weekday records:
cnt: 801.6467532467533, 589.0
If t1>13.0, then:
Winter holiday records:
cnt: 1080.3333333333333, 813.0
Winter weekday records:
cnt: 987.9324324324324, 1001.0
```

מבנה הפלט צריך להראות בדיוק על-פי הדוגמה המצורפת, מאחר ומתבצעת השוואת קבצים אוטומטית.

שימו לב: המדגם הנ"ל נתון לכם תחת הקובץ london_sample.csv.

דרישות למימוש:

המימוש **חייב** להכיל לפחות את המודולים והמתודות הבאות:

1. **main.py** – ממשק ראשי לריצת התוכנית, כפי שהיא מתוארת בפרק "תיאור המשימות". בין היתר, בקובץ זה יופיעו השורות:

```
def main(argv):
    pass


if __name__ == '__main__':
    main(sys.argv)
```

כאשר במקום "pass" יופיע קטע הקוד. עליכם לייבא את הספריה sys בראש הקובץ, ע"י import sys. ע"י המבנה הנ"ל תוכלו להריץ את התוכנית שלכם באמצעות:

python **your_path**/main.py arguments

בתרגיל בית זה:

python /home/student/**your_path**/main.py /home/student/**your_path**/london.csv "hum, t1, cnt, season, is_holiday"

שימו לב:

argv[0] = /home/student/**your_path**/main.py

argv[1] = /home/student/**your_path**/ london.csv

argv[2] = "hum, t1, cnt, season, is_holiday"

כאשר **your_path** הוא הנתיב בו שמורים: main.py, data.py, statistics.py and london.csv.

הסבר:

דרך שורת הפקודה הנ"ל, מערכת ההפעלה מעבירה למתודת המain את argv, שזהו list מסוג string המציין את הארגומנטים לתוכנית. argv[0] לא תופס תפקיד בתוכנית, אלא מציין את הנתיב המלא למודול שעל מערכת ההפעלה להפעיל (ובו נמצאת מתודת המain). שאר הפריטים בargv, כלומר argv[1] שמחזיק את הנתיב לקובץ הנתונים, וargv[2] שמחזיק את רשימת התכוניות שבהן נשתמש בתרגיל זה, יהיו הקלט למודול data.py בסעיף הבא.

(20%)

2. **data.py** – ממשק להכנת הנתונים וניהולם.

מתודות שעליכם לממש:

load_data(path, features)

פרמטרים:

- path הינו הנתיב המלא לקובץ (שבסופו שם הקובץ).
- features הינה רשימת התכוניות הרלוונטיות שבהן אנחנו מתעניינים.

המתודה קוראת את סט הנתונים מקובץ csv ומעלה את המידע לזיכרון הראשי. לצורך כך יש להשתמש בספרייה pandas, (על ידי Import pandas), ובקטע קוד הבא ב scope -של המתודה:

```
df = pandas.read_csv(path)

data = df.to_dict(orient="list")
```

data הוא מילון מהצורה הבאה:

{"cnt": [512, 358, 308, … , 40], "t1": [0, -5, 13 , …, 10], "is_holiday": [1, 0, 1, …, 0]}

אם כן, ערכי הרשומה ה-i (שורה בסט הנתונים) הם כל אותם ערכים שנמצאים במקום ה-i ברשימות של התכוניות במילון.

פלט המתודה הוא המילון data , כאשר הkeys של המילון יהיו התכוניות הרלוונטיות בלבד (5 מתוך 10 תכוניות סה"כ). (5%)

filter_by_feature(data, feature, values)

פרמטרים:

- data הינו מילון שהkeys שלו הם תכוניות מסט הנתונים, והvalues הם רשימות שמכילים את ערכי התכוניות.
- feature הינו שם של תכונת **קטגוריאלית**, כלומר, תכונותיה הן קטגוריות.
- values הינו set ערכים, כאשר התכונית בfeature יכולה לקבל את כל הערכים בvalues.

פלט המתודה יהיה שני מילונים, כאשר איחודם הוא המילון data (return data1, data2). המילון הראשון מכיל את כל הרשומות בהן התכונית feature קיבלה ערך כלשהו המופיע בvalues, והמילון השני יהיה מילון המכיל את כל הרשומות שבהן התכונית feature קיבלה ערכים שאינם מופיעים בvalues, ורק אותן. (10%)

print_details(data, features, statistic_functions)

פרמטרים:

- data הינו מילון שהkeys שלו הם תכוניות מסט הנתונים, והvalues הם רשימות שמכילים את ערכי התכוניות.
- features הינה רשימה של תכוניות מתוך סט הנתונים.
- statistic_functions – רשימה המכילה מתודות סטטיסטיות שמצאות במודול statistics.py.

המתודה תדפיס מדדים סטטיסטיים על data אך ורק לפי התכוניות בfeatures, תוך שימוש המתודות בstatistic_functions. (15%).

3. **statistics.py** – מודול לחישוב מדדים סטטיסטיים.

מתודות שעליכם לממש:

sum(values)

mean(values)

median(values)

פרמטרים:

- values – רשימה של ערכים מספריים.

פלט המתודות הנ"ל הוא כפי שהוגדר בפרק "תיאור המשימות" (כאשר פלט המתודות sum, mean and median הינם הסכום, הממוצע והחציון בהתאמה). ניתן להשתמש בפו' ceil של הספריה math לטובת חישוב ערך עליון במתודה median. **חל איסור להשתמש בספריה math להפעלת מתודה אחרת, חל איסור להשתמש בספריות חיצוניות למימוש שאר המתודות במודול הנוכחי. כמו כן, אין להשתמש בפו' המובנית sum.**

(5% למתודה)

population_statistics(feature_description, data, treatment, target, threshold, is_above, statistic_functions)

פרמטרים:

- feature_descrription הינה מחרוזת המתארת שם הקבוצה (למשל "Winter holiday records").
- data הינו מילון שהkeys שלו הם תכוניות מסט הנתונים, והvalues הם רשימות שמכילים את ערכי התכוניות.
- treatment הינו שם של תכונית מסט הנתונים (למשל "t1").
- target הינו שם של תכונית מסט הנתונים (למשל, "Cnt").
- threshold הינו ערך סף לתכונית treatment.
- is_above הינו אינדיקטור המקבל את הערך True או False.
- statistic_functions – רשימה המכילה מתודות סטטיסטיות שמצאות במודול statistics.py.

המתודה מדפיסה מדדים סטטיסטים על data שמייצג את האוכלוסיה population ורק אותה בלבד, תוך שימוש במתודות בstatstic_functions. המדדים הסטטיסטיים נמדדים על התכונית target, לאחר איסוף הרשומות המתאימות לפי התנאים הבאים:

אם is_above קיבל את הערך True, יאספו הרשומות בהן התכונית treatment מקבלת ערכים גדולים ממש מthreshold. אחרת, יאספו הרשומות בהן התכונית treatment מקבלת ערכים קטנים או שווים לthreshold.

(20%)

בונוס: מימוש המתודה במספר מינימלי של שורות קוד יזכה ב- 5 נקודות בונוס לציון של תרגיל זה. המספר המינימלי יוגדר לפי התוכניות שיוגשו.

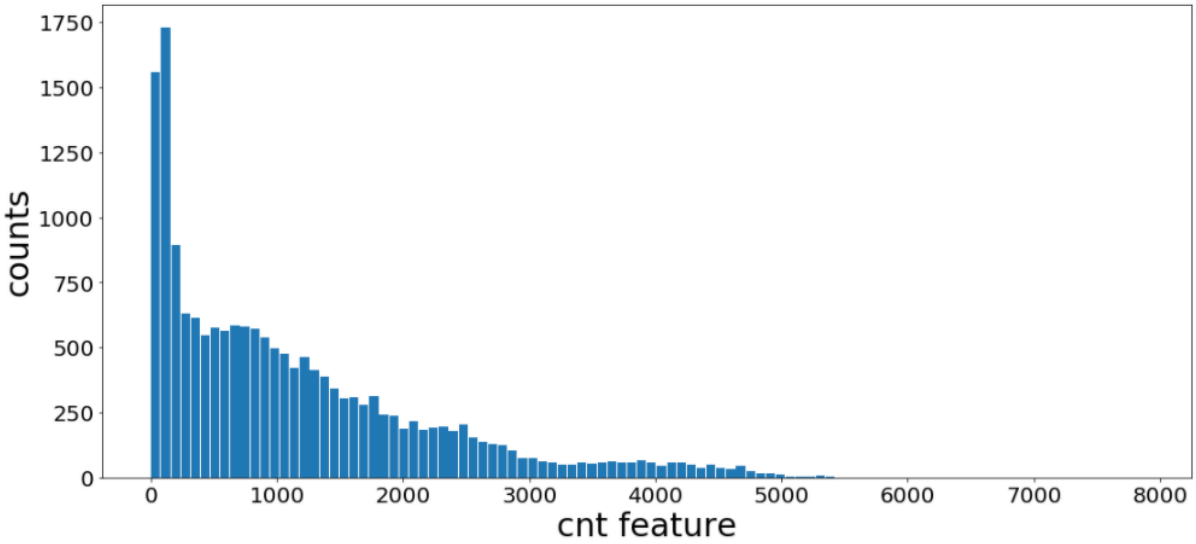
דגשים נוספים:

1. עליכם לכתוב את הקוד בהתאם לדגשים והסטנדרטים לפי pep8. לשימושכם המסמך Code Quality " Requirements " באתר ה moodle- של הקורס. קוד אשר לא יעמוד בסטנדרטים הנדרשים, יקבל ניקוד מופחת.
2. ניתן להוסיף מתודות נוספות, במידה ותמצאו לנכון. יש להימנע מכפילויות קוד.
3. ניתן להשתמש במתודות שהן in-built בשפה. קרי, מתודות אשר **לא** דורשות ייבוא של ספריות.
4. יש לתת שמות בעלי משמעות לכל משתנה.

5. חובה לתעד את הקוד באנגלית. בפרט עליכם לכתוב עבור כל מתודה docstring

חלק ב' – יבש (15%):

1. להלן היסטוגרמת נתונים המציגה את ערכי התכונית cnt של כלל האוכלוסיה (כל הרשומות). בהינתן ההיסטוגרמה, במידה והייתם מתבקשים לחשב עבור התכונית cnt רק אחד המדדים, ממוצע או חציון, באיזה מדד סטטיסטי הייתם בוחרים? נתון: הערך המקסימלי המתקבל הינו 7860. נמקו את בחירתכם. (10%)



2. בהינתן הפלט שהתקבל בשאלה 2 של החלק הרטוב, ובהסתמך על מספר ניסויים כרצונכם, האם לדעתכם קיים קשר בין הטמפרטורה למספר האופניים שהושכרו בקרב ימי חג חורפיים? האם קיים קשר בין הטמפרטורה למספר האופניים שהושכרו בקרב ימי חול חורפיים? (5%).

למרות שאין חובה להציג ניסויים בסעיף זה, **בונוס** של 3 נקודות לציון תרגיל זה ינתן למספר מצומצם של הגשות שיציגו את הניסויים היצירתיים ביותר על מנת לענות על השאלה, לפי שיקול צוות הקורס.

הוראות הגשה:

- התרגיל להגשה בזוגות בלבד.
- לפני ההגשה, חובה לוודא שהתוכנית עובדת במעבדת ההוראה ולא בסביבה אחרת.

- ההגשה חייבת להכיל קובץ אחד (קובץ zip) :
  - שם הקובץ חייב להיות hw1 _xxxxxxxx_yyyyyyyy.zip כאשר xxxxxxxx ו-yyyyyyyy הם מספרי תעודות הזהות של המגישים, כולל ספרת ביקורת.
  - הקובץ מכיל את כל קבצי הקוד וקובץ דו"ח שלכם עם תשובות לשאלות. אין להכיל תיקייה ובתוכה קבצי הקוד, אלא את קבצי הקוד עצמם.
  - **הערה :** עליכם לוודא שהתוכנית מתחילה לפעול מקובץ "main.py" בלבד.
  - תשובות לחלקים יבשים יש להקליד במעבד תמלילים. אין להגיש תשובות בכתב יד.

- ההגשה היא אלקטרונית בלבד, דרך אתר ה-moodle של הקורס. תרגילים שיוגשו בכל דרך אחרת לא ייבדקו.
- אין להגיש את אותו הקובץ פעמיים. התרגיל יוגש ע"י אחד מבני הזוג.
- שימו לב שההגשה תיחסם בדיוק בשעה 23:55 ביום ההגשה. מומלץ להגיש לפחות שעה לפני המועד האחרון.
- ניתן להגיש כמה פעמים. רק ההגשה האחרונה תישמר.
- תרגיל בית שלא יוגש לפי הוראות ההגשה – לא ייבדק (כלומר יקבל ציון 0).
- לצורך תרגיל הבית ייפתח פורום. ניהול שאלות ומתן תשובות בנושא התרגיל יתבצע דרך הפורום בלבד.

בהצלחה!