מבוא למדע הנתונים

מספר קורס: 2686

מסלול כלכלה-יזמות

תרגיל מספר 1

חלק א פייתון: היכרות עם סביבת העבודה Anaconda, ספריות לניתוח מידע, קריאה כתיבה לקובץ

חלק ב אקסל: פקודות בסיסיות, Pivot, Solver

מרצה:

ד"ר גייל גלבוע פרידמן

מתרגלים:

מר עידן לופו

מר אפי פקאני

עבודה ביחידים

תאריך הגשה - 8 לאפריל 2021

IDC HERZLIYA | Adelson School of Entrepreneurship

בית ספר טיומקין לכלכלה | הבינתחומי הרצליה

מטרת התרגיל

היכרות עם פקודות בסיסיות וחבילות Python אשר תומכות בקריאה ניתוח וכתיבה של נתונים

התרגיל מתייחס לקבצי דאטה עם נתונים שנאספו ממערכת מובילאיי בת"א ובברצלונה:

Mobileye_risk_Barcelona.csv. Mobileye_risk_TelAviv.csv.

הגשה

- הגש תיקיית zip ששמה S1_YOURNAME_YOURID המכילה 3 קבצים (יודגש באדום בהנחיות): Word, xlsx, ipynb
- ההגשה דרך הMoodle, בתיקייה ייעודית לתרגיל זה.

קוד פייתון:

- לשם הגשת התרגילים בקורס עליך להתקין את סביבת העבודה Anaconda (התקנת אנקונדה)
- דרך אנקונדה עליך להיכנס לאפליקציה בשם Jupyter (אפליקציית web חינמית עבור עבודה בפייתון)
- ב- Jupyter תוכל לפתוח קובץ (מחברת) עריכה, המכונה Jupyter Notebook, הקובץ המתקבל בעל סיומת ipynb.

  נא קרא לו בשם : S1_YOURNAME_YOURID.ipynb למשל S1_DANILEVI_111
- במחברת תוכל לכתוב בתוך תאים:
  - תאים מסוג Code שבהם תוכל לרשום קוד ולהריץ אותו בגוף המחברת
  - תאים מסוג Markdown שבהם תוכל למלא את התשובות המילוליות
- אנא הקפידו להפריד בין חלקי התרגיל בעזרת תא Markdown בו יירשם מס׳ הפרק (Part A, Part B..).
- עבור כל שאלה רשמו את התשובה והציגו תצלום מסך של הפלט כסימוכין.

הגישו את המסמך הסופי בפורמט Word המכיל תשובות כתובות וצילומי מסך לדוגמא:

ניתן לראות שבקובץ שלנו ישנם 2823 שורות ו12 עמודות:

```
rows, cols = data.shape

print('There are', rows, 'rows.')
print('There are', cols, 'cols.')
```

```
There are 2823 rows.
There are 12 cols.
```

- אם התבקשתם להשתמש באקסל צרפו גם אותו, ותשובות שמסתמכות על תוצאות המתקבלות ממנו ניתן לצרף גם לגיליון התשובות הכתובות כתצלום מסך.

בהצלחה !

(15 נק') - חלק א': מתחילים...

יבוא ספריות (Import)

1. קוד (1 נק') ייבאו את 5 החבילות numpy, pandas, matplotlib, seaborn, xlwt
2. קוד (1 נק') ייבאו את 2 המודולים matplotlib.pyplot, xlwt.Workbook
3. קוד (1 נק') הדפיסו למסך את כל הפונקציות והמשתנים של כל אחד משני המודולים אותם ייבאתם.
4. קוד (6 נק') הסבירו בקצרה (מהאינטרנט) על השימוש בכל אחת מהחבילות אותן ייבאתם.
5. קוד (6 נק') הסבירו בקצרה מהו ההבדל בין למידת מכונה Supervised (מבוקרת) ללמידת מכונה Unsupervised (לא מבוקרת).

(20 נק') -חלק ב': קריאת נתונים

6. הכנת הנתונים:
   a. קוד (2 נק') – בידיכם 2 קבצים עם נתונים על מקטעי נסיעה מהערים תל אביב וברצלונה, קלטו את קבצי הדאטה לתוך שני משתנים שונים, השמיטו את העמודה הראשונה (במיקום 0) מכל אחת מהטבלאות.
   b. קוד (2 נק') הוסיפו לכל טבלה עמודה נוספת בשם city, בה רשמו Tel-aviv או Barcelona עבור כל שורה בהתאמה לטבלה.
   c. קוד (2 נק') – מזגו בעזרת pandas את שני משתני הדאטה שיצרתם ל-Dataframe יחיד, הדפיסו את מס' המימדים שלו <u>בעזרת פונקציה של Pandas</u> (רשמו את כמות השורות והעמודות).
   d. קוד (2 נק') – שנו את הערכים בעמודה city מערכי text לערכים נומריים, עבור תצפית מהעיר תל אביב רשמו 1 ועבור תצפית מברצלונה 0.
   e. קוד (2 נק') – הדפיסו 10 תצפיות לדוגמא מה-dataframe שלכם על מנת לוודא שהכנתם את הנתונים כראוי.
7. קוד (2 נק') – הדפיסו רשימה של שמות העמודות של הDataframe.

8. קוד (4 נק') – הדפיסו בטבלה את הנתונים עבור מקטעי הנסיעה בהם שיעור ההתקרבות של כלי-רכב הוא גדול מ-0.2 (**near_miss_vehicle_ratio**), כמה תצפיות קיבלתם עבור כל אחת מהערים?
מה אפשר להגיד על מקטעי נסיעה אלו?
9. קוד (4 נק') – הדפיסו בטבלה את הנתונים עבור 10 מקטעי הנסיעה בהן שיעור הסטיות מהנתיב (**harsh_cornering_ratio**) הוא הגבוה ביותר ב-Dataset, מהי התצפית בעלת השיעור הגבוה ביותר (מה מס' האינדקס שלה), מה המשמעות של שיעור כל כך גבוה על אותו מקטע?

חלק ג': סטטיסטיקה תיאורית (15 נק')

10. קוד (4 נק') – הצג תיאור סטטיסטי של כל אחת מהעמודות בדאטה סט בעזרת פונקצית describe של pandas, רשמו מה שיעור התצפיות מהעיר תל-אביב ומה שיעור התצפיות מהעיר ברצלונה?
11. קוד (4 נק') – באיזה מבין הערים מתבצעות יותר נסיעות בממוצע עבור מקטע דרך (detection_drives_count)? הסבירו כיצד קבעתם זאת.
12. (7 נק') – היעזרו בפונקצית corr על מנת לבנות מטריצת קוריאלציות בין כלל המשתנים ב-Data Set שלנו, כתבו 2 תובנות הגיוניות שעולות לכם מטבלת הקורלציה (עבור קורלאציות הגבוהות מ-0.3).

(15 נק')

חלק ד': ויזואליזציה של פונקציות

13. קוד (2 נק') - הציגו את תוצאות מטריצת הקוריאלציות משאלה 12 כ"מפת חום" (seaborn - heatmap).
14. קוד (3 נק') – הציגו את מס' התצפיות מכל עיר בגרף ברים בעזרת הפונקציה value_counts כך שעמודה אחת תייצג את ברצלונה (0) ואת מס' התצפיות מעיר זו והעמודה השנייה תייצג את תל-אביב (1) ומס' התצפיות מעיר זאת.
15. קוד (5 נק') – הציגו עבור כל עיר בדאטה בדיאגרמת scatter את הקשר בין מספר הנסיעות במקטע דרך לכמות הסטיות מהנתיב (פיזור נקודות, כשלכל נקודה: מספר הנסיעות במקטע הינו ציר x ומספר סטיות מהנתיב הינו ציר y).
16. קוד (5 נק׳) הציגו Boxplot עבור אחת העמודות בדאטה שלנו (לבחירתכם), הסבירו אילו נתונים ניתן להסיק מהתוצאה המתקבלת.

(12 נק') - חלק ה': שמירת נתונים לקובץ

בשאלה זו עלייך לכתוב קוד שמייצר קובץ אקסל שנקרא S1_YOURNAME_YOURID.xslx בקובץ שתייצרו:

1. עמודה A תהיה עמודת Index שבה כל תצפית תקבל מס' סידורי.
2. עמודות B-F יהיו עמודות: **detection_drives_count**, **avg_speed, braking_count**, **cornering_count** אשר ילקחו מקובץ הנתונים של **<u>העיר ברצלונה</u>** אותו יצרתם בחלק ב'.

רמז: ההתחלה של הקובץ תראה כך

	A	B	C	D	E
1	Index	detection_drives_count	avg_speed	braking_count	cornering_count
2	1	53	16.67762442	2	30
3	2	78	20.44686955	41	51
4	3	108	19.86951061	25	19
5	4	266	29.25569345	16	185
6	5	86	51.4346366	14	56
7	6	126	27.61654117	10	95
8	7	52	19.13000817	15	20
9	8	107	28.55917352	1	92
10	9	106	29.1092541	3	92
11	10	75	15.24046277	0	0
12	11	146	27.36014973	28	81
13	12	62	42.08545263	20	47
14	13	109	29.53364857	7	39
15	14	220	28.39391732	23	63
16	15	133	29.33302629	10	54
17	16	96	19.7695308	30	36
18	17	108	23.09465676	7	33
19	18	612	28.95406005	39	116

21. קוד (2 נק') ייצרו dataframe מצומצם שמכיל רק את הדאטה מהעמודות המוזכרות לעיל.
22. קוד (2 נק') פתחו workbook והוסיפו בו גליון ששמו זהה לשמכם (בהתאם לשם באדום לעיל).
23. קוד (2 נק') כתבו 5 כותרות לשורה הראשונה בגליון, בסדר הבא:

    Index, detection_drives_count, avg_speed, braking_count , cornering_count
24. קוד (5 נק') כתבו לולאה שעוברת על התאים ב-dataframe שייצרתם ובעזרת פונקציית ()iterrows עדכנו את הגליון:
    a. קוד (2 נק') כך שבעמודה A שכותרתה Index יופיע מספר של התצפיות <u>החל מהמס' 1</u>.
    b. קוד (3 נק') עמודות B-E יהיו העתק של העמודות הרלוונטיות מהדאטה המקורי.
25. קוד (1 נק') ייצאו את ה-dataframe שייצרתם לקובץ CSV בעזרת פונקציה של pandas (על מנת להוסיף עמודת אינדקס שמתחילה ב-0 השתמשו ב- index=True).

(23 נק') - חלק ו': Excel-Pivot & Solver

עבור החלק הבא השתמשו בקובץ האקסל שנקרא "Part F - pivot & solver" בגליון שנקרא "risk_Barcelona_filtered":

26. קוד (2 נק') הכניסו עמודה חדשה לאחר העמודה הנקראת "detection_drive_count" וקראו לה בשם "detection_drive_count_groups". בעמודה זו ספירת הנסיעות תהיה מחולקת לקבוצות של 10. (כלומר: 0<=x<10, 10<=x<20).
27. קוד (2 נק') הכניסו עמודה חדשה לאחר העמודה הנקראת "avg_speed" וקראו לה בשם"avg_speed_groups". בעמודה זו חלקו את המהירות הממוצעת לקבוצות של 10. (כלומר: 0<=x<10, 10<=x<20).

צרו את טבלאות הפיבוט המתאימות עבור השאלות הבאות:

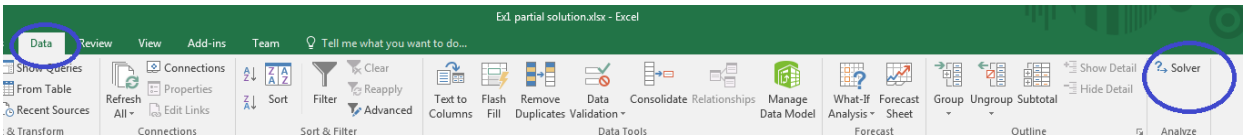
28. קוד (3 נק') מהי קבוצת ה-"avg_speed_group" השכיחה ביותר במונחים של מיקום גלובלי (geometry), בכמה אזורים קיימת קבוצת מהירות זו?
29. קוד (4 נק') מהי קבוצת ה-"avg_speed_group" הכי פחות שכיחה במונחים של "detection_drive_count" ? כמה נסיעות קיימות בקבוצת מהירות זו?
30. קוד (4 נק') באיזו קבוצת "avg_speed_group" יש את ההפרש הגדול ביותר בין סך כל העצירות (braking_count) לבין סך החריגות מהנתיב (cornering_count)? [זכרו להשתמש בהפרש אבסולוטי] מהו סך העצירות וסך החריגות מהנתיב עבור קבוצת מהירות זו?

IDC HERZLIYA | Adelson School of Entrepreneurship

בית ספר טיומקין לכלכלה | הבינתחומי הרצליה

In the data tab, you can find the solver button:

In case you can't find this button:



Load the Solver Add-in

Applies To: Excel 2016, Excel 2013, Excel 2010, Excel 2007

In Excel 2010, many improvements have been made to the Solver add-in.

Read a blog post or try Office 365!

The Solver Add-in is a Microsoft Office Excel add-in program that is available when you install Microsoft Office or Excel.

To use the Solver Add-in, however, you first need to load it in Excel.

1. In Excel 2010 and later goto **File** > **Options**

   NOTE: For Excel 2007 click the **Microsoft Office Button** , and then click **Excel Options**.

2. Click **Add-Ins**, and then in the **Manage** box, select **Excel Add-ins**.

3. Click **Go**.

4. In the **Add-Ins available** box, select the **Solver Add-in** check box, and then click **OK**.

   a. **Tip** If the **Solver Add-in** is not listed in the **Add-Ins available** box, click **Browse** to locate the add-in.

   b. If you get prompted that the Solver Add-in is not currently installed on your computer, click **Yes** to install it.

5. After you load the Solver Add-in, the **Solver** command is available in the **Analysis** group on the **Data** tab.

Solver

Solver is a super useful tool in excel which helps us solve complicated optimization problems of many variables. We will illustrate the tool by the following problem:

The Electronic Products Factory Problem

You are a manager of a factory that produce electronic products. Your factory can produce four different types of products: Speakers, Amplifiers, Receivers, & Screens. Your aim is to maximize the factory profit by selecting which products to produce, and considering the profit from each type of product, the production capabilities of your factory and the component details of each product.

- **Production capabilities:** Each product is produced by using 5 types of components: Motherboard, Working Hours, Cables, Resistors, and several meters of Wires. Your inventory estimation is:
  - Motherboards: 350 units
  - Working Hours: 80 hours
  - Cables: 1250 units
  - Resistors: 3500 units
  - Meters of Wires: 145 meters
- **Required Components**

	Motherboard	Working Hours	Cables	Resistors	Meters of Wires
Speaker	1	0.25	3	7	6
Amplifier	1	0.3	28	44	3
Receiver	1	0.35	28	52	3
Screen	1	0.15	10	6	1

- **Monetary Data**

	Unit Price ($)	Unit Cost ($)
Speaker	120	13
Amplifier	390	58
Receiver	560	32
Screen	210	23

מה הוא שילוב המוצרים האופטימלי שהמפעל צריך למכור ללקוחותיו בשביל למקסם את רווחיו?

עבור החלק הבא השתמשו בקובץ האקסל שנקרא "Part F - pivot & solver":

1. קוד (2 נק') הכניסו לגליון Solver את הדאטה מהעמ' לעיל בתאים המתאימים.
2. קוד (2 נק') פתור את הבעיה למעלה בעזרת שימוש ב-"Solver".
3. קוד (2 נק') זהה אילו מהתאים השתנו. מי מבין התאים הללו מייצג את תאי ההחלטה עבור הבעיה?
4. קוד (2 נק') הסבר את הפתרון שנמצע ע"י ה-Solver