פרויקט סיום

מבוא למדע הנתונים

ד"ר אריה יעקבי

בפרויקט זה התלמידים יישמו באמצעות שפת התכנות פיתון את החומר שנלמד בקורס. הפרויקט יכלול קריאה של מקורות מידע מסוגים שונים (כגון CSV, XLSX ועוד), ניקוי הנתונים והכנתם לקראת ניתוחים שונים, וביצוע ניתוחים בסיסיים על נתונים אלה.

הנחיות לפרויקט

1. **בחירת מקורות המידע**:

1.1. יש לבחור קובץ נתונים בכל תחום שנראה לכם. קראו לו input1_df. הקובץ יכול להיות בפורמט CSV, XLSX, JSON, XML בקובץ חייבים להיות לפחות 1000 רשומות ו- 5 עמודות נתונים. ניתן להוריד קבצים כאלה ברשת הווב. ישנן דוגמאות רבות של אתרים בהם ניתן להוריד קבצי נתונים באתר: https://www.kdnuggets.com/datasets/index.html

1.2. מצא את אחוז הערכים החסרים. ניתן לבצע זאת על ידי מציאת מספר הערכים החסרים –

```
df.isnull().sum()
```

1.3. אם אחוז הערכים החסרים קטן מ- 10% אז קובץ נתונים זה לא מתאים לפרויקט זה ותחפשו קובץ אחר.

1.4. יש לשים לב שהקובץ אינו נקי מלכתחילה - לדוגמה בדקו האם ישנם ערכים חסרים, ערכים שלא רשומים בפורמט נכון בעמודה (לדוגמה מספר במקום שם). אם הקובץ כבר נקי ומסודר – פסלו אותו. לא אכפת לי שתיקחו קובץ שניקו אותו ופשוט תקלקלו אותו באופן משמעותי.

1.5. לאחר קריאת הקובץ יש להדפיס את 5 השורות הראשונות, האחרונות ובאיזה שהוא מקום באמצע שתבחרו.

1.6. לאחר מכן יש להדפיס סטטיסטיקה תיאורית של הנתונים ב- data-frame באמצעות הפונקציה describe(). חשוב להדפיס סטטיסטיקה תיאורית של משתנים רציפים ומשתנים קטגוריים כאחד.

1.7. יש לבחור קובץ נתונים נוסף בכל תחום שנראה לכם. קראו לו input2_df. הקובץ יכול להיות בפורמט CSV, XLSX, JSON, XML . גם בקובץ חייבים להיות לפחות 1000 רשומות ו- 5 עמודות נתונים. וגם עבורו בצעו את מה שכתוב בסעיפים 1.2 – 1.6 .

1.8. צרו data-frame ע"י פעולת Full Outer Join. קראו לו outer_join_df. אם אין להם מפתח משותף, תדאגו להוסיף עמודה מתאימה לאחד מה- data-frame כך שבשני הדטהפרמס תהיה עמודת מפתח עם ערכים דומים (כמו לדומה: מס מוצר, או מס זיהוי,...), כך שתוכלו להשתמש במתודה

```
df1.merge(df2, on='id', how='outer')
```

בדקו שוב האם ישנם ערכים חסרים בקובץ הממוזג.

1.9. צרו 2 data-frame חדשים ע"י בחירת עמודות מתוך ה- data-frame שנותר מפעולת ה- JOIN בסעיף הקודם. קראו להם df2, df1 בהתאמה. אנחנו נעבוד על df2, df1 בסעיפים הבאים.

2. **ניקוי וארגון קובץ הנתונים:**

עבור ה - df2, df1 data-frames בצעו את הניקיון של הנתונים והפורמט שלהם לפי הדרישות הבאות:

2.1. אם בעמודה מספרית ישנם ערכים לא מתאימים כמו לדוגמה ערכים בולאניים, או סטרינגים, יש למצא אותם ולהוריד את השורות עם ערכים אלה (אם במקור df2, df1 frames אינם מכילים ערכים "לא מתאימים" תדאגו להכניס אותם) . יש לוודא ששורות אלה נמחקו ולסדר את האינדקס של השורות באופן הנכון. כמו כן יש לדאוג שכל הערכים העמודה יהיו מסוג *float64 ויש לוודא זאת.*

2.2. אם בעמודה שיש בה סטרינגים (כמו לדוגמה עמודת שמות) ישנם ערכים לא מתאימים כמו לדוגמה ערכים בולאניים, או מספריים, יש למצא אותם ולהוריד את השורות עם ערכים אלה (אם במקור frames df2, df1 אינם מכילים ערכים "לא מתאימים" תדאגו להכניס אותם) . יש לוודא ששורות אלה נמחקו ולסדר את האינדקס של השורות באופן הנכון. כמו כן יש לדאוג שכל הערכים העמודה יהיו מסוג *string ויש לוודא זאת.*

2.3. לאחר מכן ניתן לוודא שאם ישנם ערכים חסרים (תאים ריקים אן ערכי null) יש להחליף ערכים אלה באמצעות הממוצע, השכיח, או החציון של הערכים בעמודה. ניתן להשתמש במתודה: `fillna` .
<u>בונוס</u>: מי שיחליף ערכים אלה באופן שונה על יד לדוגמה מציאת הממוצע בין המספר הקודם למספר הבא. לדוגמה, אם באמצע העמודה יש את המספרים:

100

N\A

200

אז אפשר להחליף את ערך ה-N\A במממוצע של (100+200) / 2 = 150.

שימו לב למה עושים במקרים חריגים כמו לדוגמה כיש רצף של ערכי null בעמודה אז אי אפשר לחשב את הממוצע של המספר הבא, או כשמדובר במספר האחרון בעמודה. במקרים אלה חשבו על פתרון יצירתי יותר.

2.4. <u>נרמול עמודת נתונים מספריים</u> – בחרו את אחת מהעמודות באחד מה- df2, df1 data-frames , בה יש מספרים. עברו בלולאה על העמודה, החליפו כל מספר בעמודה במספר (בין 1-0) שהוא תוצאת חלוקת המספר במספר המקסימלי בעמודת המספרים. לדוגמה, אם המספר הנכחי בעמודה הוא 750, והמספר המקסימלי בעמודה הוא 1200, אז נחליף את 750 במספר $750/1200 = 0.625$. אם המספר הוא 800-, אז נחליף אותו בערך המוחלט שלו (800) חלקי המספר המקסימלי בסדרה (1200) $|-800|/1200 = 800/1200 = 0.66$.

2.5. הדפיסו את כל השורות הכפולות (אם אין אז תכניסו שורות כאלה באמצעות קוד פייתון) והורידו אותן על ידי: `drop_duplicates()`.

3. הצגה ויזואלית של הנתונים:

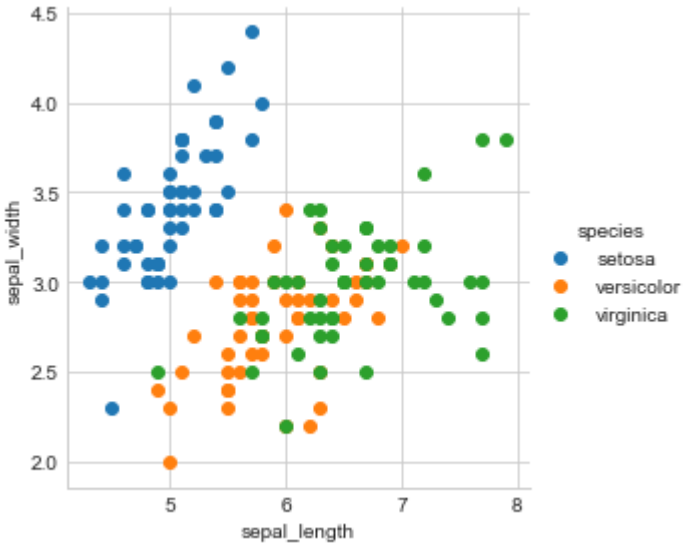
מספרים בפני עצמם אינם נוטים לעורר תגובה רגשית. אבל הדמיית נתונים יכולה לספר סיפור שנותן משמעות לנתונים. בעוד שהשוואת מספרים עשויה להרשים את הקוראים, חיזוק המספרים הללו באמצעות דיאגרמות עוזר להשפיע עוד יותר.

קישור: איך לספר סיפור עם נתונים (מדריך למתחילים):

https://venngage.com/blog/data-storytelling/

עכשיו כשיש לנו data-frames df1, df2 שעברו ניקיון של הנתונים, בחרו אחד מהם או את שניהם כרצונכם, והציגו 10 דיאגרמות שונות של הנתונים עשו זאת כפי שעשינו באפליקציית הדוגמה של המניות בקורס (שיעור 11). ציירו כמה שיותר גרפים מכמה שיותר סוגים והסיקו מסקנות על הנתונים בהתאם ורשמו אותן. גרפים מיוחדים ומסקנות יפות יזכו **בבונוס**.

גרפים פשוטים כגון:



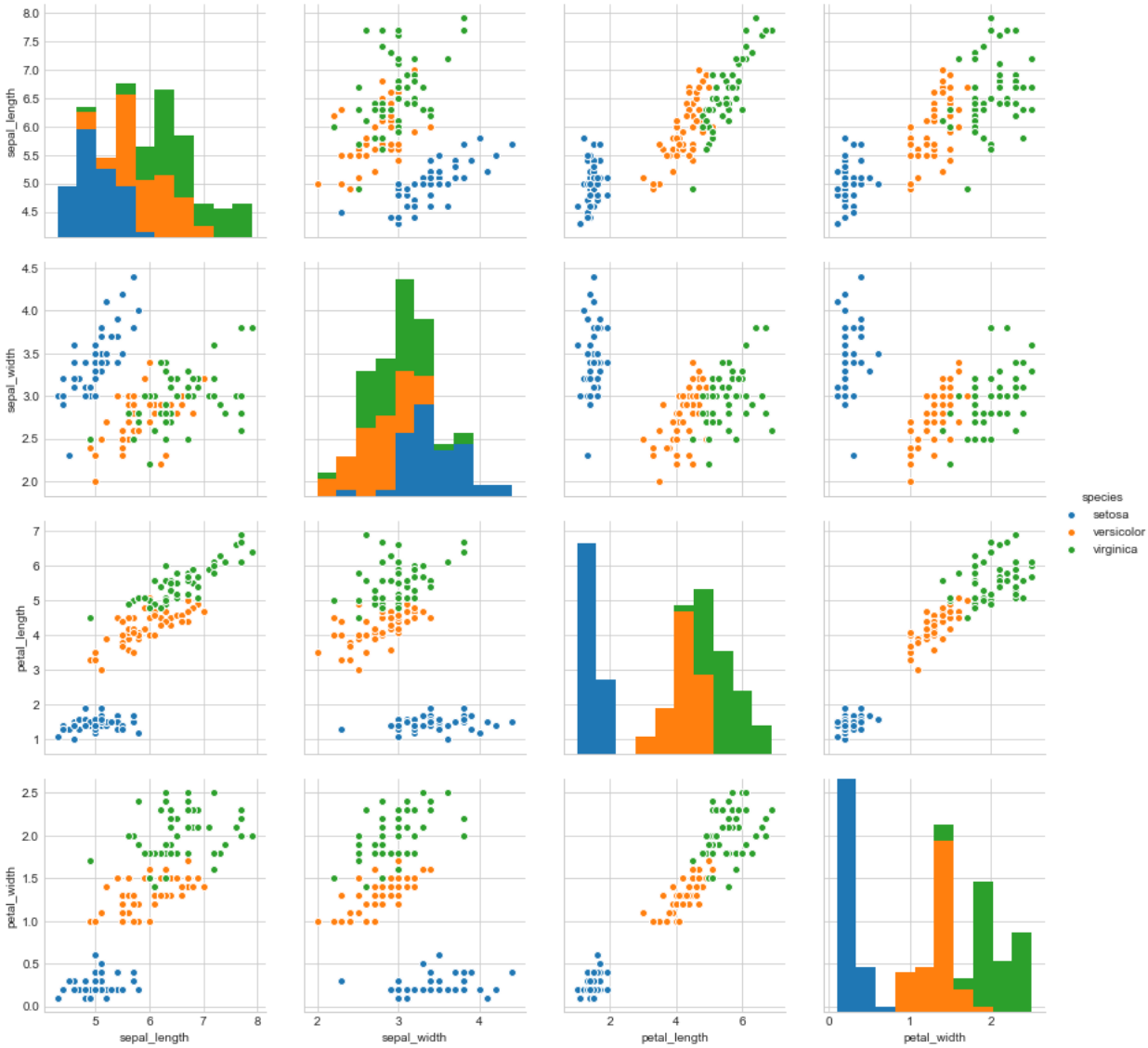
שהקוד הבא מציג:

```
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.datasets import load_iris

iris = load_iris()
print(iris.head())

plt.scatter(iris['sepal_length'],iris['sepal_width'],color=['r','b','g'])
plt.xlabel('Sepal length')
plt.ylabel('Sepal width')
plt.title('Scatter plot on Iris dataset')
```

ועד דוגמאות של גרפים מעניינים:



שמציגים 4 עמודות, אלה עם אלה. הקוד שמציג גרף זה:

```
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.datasets import load_iris

iris = load_iris()

sns.set_style("whitegrid");
sns.pairplot(iris, hue="species", size=3);
plt.show()
```

ממליץ מאוד על הספרי: https://plotly.com/python/plotly-express/ שבה יש דוגמאות רבות וקוד מוכן.

במסמך: https://towardsdatascience.com/data-visualization-for-machine-learning-and-data-science-a45178970be7
יש דוגמאות שונות ויפות. דוגמאות מעניינות נוספות יש ב:

https://plotly.com/python/plotly-express/
https://matplotlib.org/stable/gallery/index
https://seaborn.pydata.org/generated/seaborn.jointplot.html
https://seaborn.pydata.org/

4. ניתוח הנתונים:

4.1. על אתם DF שבחרתם להשתמש בסעיף הקודם, בצעו חלוקה לאשכולות של הנתונים לפי חתכים שונים והציגו בגרף את התוצאות. כמובן שחשוב שתשתמשו במבחן המרפק (כולל גרף) על מנת לקבוע את מספר האשכולות הנכון. הסיקו מסקנות מתאימות (לפחות 5 מסקנות). **בונוס** יינתן ליותר מסקנות ולמסקנות מעניינות יותר.
4.2. יש לבצע רגרסיה ליניארית (זה מופיע בשקפים) ולצייר את זה. הציגו והסבירו את מסקנות הרגרסיה.
4.3. כל ניתוח נוסף מסוג אחר ומעניין יזכה **בבונוס**.

5. מודולים או אובייקטים שחובה להשתמש בהם בפרויקט הם:

5.1. Xmltodict(רשות)
5.2. האובייקט json
5.3. Pandas
5.4. plotly.express
5.5. Matplotlib
5.6. Seaborn

6. מועד ואופן הגשה:

6.1. מועד הגשת הפרויקט יירשם באתר העבודות של הקורס.
6.2. כל דמיון חלקי או אחר בין התרגילים של תלמידים שונים יגרור פסילה של שתי העבודות ומכאן פסילת זכאות לקורס של צוותי התלמידים.
6.3. הפרויקט יוגש באתר. בדקו במועד שהתרגיל הוגש כראוי. (בדקו שהפרויקט מופיע באתר לאחר העלאתו).
6.4. יש להגיש את הפרויקט במועדו . לא יתקבלו טענות מכל סוג כולל "חשבנו שהגשנו, טעינו בכתובת, יש לחץ או בחינות בסוף הסמסטר וכיו"ב". הערכו מראש ועמדו בקצב הזמנים המוצג.

7. כיצד להגיש את התרגיל המסכם?
7.1. יש לצרף לעבודה קובץ WORD שיכלול:
- כל הערה שברצונכם שאני אקרא טרם בדיקת התרגיל.

- שם הפרויקט, שמות המגישים ומספרי זיהוי, משפט המתאר את הפרויקט שלכם, הסברים וההערות לפי צורך.
- ציינו ב WORD את שם קובץ הנתונים .
- כל גרף ומסקנה שבקשתי יכנסו בקובץ הוורד.

7.2. יש להכניס את קובץ הוורד כל קבצי הפיתון והנתונים לתיקיה ששמה מורכב משמות התלמידים ובתוכה יהיו כל הקבצים המוזכרים לעיל.

7.3. יש לדחוס (לקווץ) את התיקייה על ידי תכנת WinRAR או ZIP. למי שאין שיתקין זה חינם.

7.4. יש להעלות את התיקיה המכווצת לאתר כפתרון לעבודה.

8. כללי:

1.1. צוותי הפרויקט ימנו עד ארבעה סטודנטים.

1.2. אי הגשת פרויקט תגרור כישלון בקורס.

1.3. התלמידים יצטרכו להגן בעל פה על הפרויקט. הגנה זו משמעותה הוא 40% מהציון הסופי של הפרויקט. בהגנה זו אנו יכולים לשאול כל שאלה הקשורה בפרויקט או בחומר (ההגנה היא עם חומר ומחשב פתוח וניתן להיעזר בהם ותתקיים בזום). הערות מצד סטודנטים כגון, "אני לא עשיתי את החלק הזה" יגררו הורדת נקודות. זמן ההגנה יכול לנוע מזמן קצר מאוד (דקות בודדות) עד הגנה ארוכה יותר (כרבע שעה).

<u>האחריות על למידת כל החומר והביצוע היא על כל הצוות.</u> לא ניתן לבצע משימה או ללמוד חומר מתקדם מבלי ללמוד היטב את החומר הקודם לכך.

הפרויקט יוגש באתר העבודות של הקורס. בדקו במועד שהפרויקט הוגש כראוי (בדקו שהתרגיל מופיע באתר לאחר העלאתו). <u>לא נקבל טעויות בהגשה – בדקו בבקשה 5 פעמים לפני שאתם מגישים.</u>
יש לשמור את העבודה אצלכם לגיבוי לפחות עד קבלת הציון.

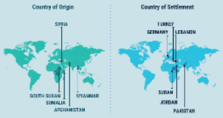
<u>בהצלחה!</u>

The Coronavirus Pandemic's
Impact on the Refugee Crisis

One group of people impacted the most by the Coronavirus Pandemic are the world's population of refugees and displaced people. There are approximately 25 million refugees across the globe, facing unique struggles in access to healthcare, government assistance, and sanitation.

These stats help paint a picture of how the COVID-19 outbreak is impacting the Refugee Crisis.

The Scale of the Refugee Crisis



2/3 of refugees come from just five countries: **Syria, Afghanistan, South Sudan, Myanmar** and **Somalia**

WHO ARE REFUGEES?

80% of the world's refugees live in the neighboring countries of their country of origin

2.3 million refugees are from South Sudan

80% are women and children

50,000 children are orphaned

Children of refugees are **x5** more likely to be out of school than their non-refugee peers

6 people a day lost their lives attempting to reach Europe last year

x2 people displaced in the world over the last 20 years

6.7 million refugees are hosted by the poorest countries

2.6 million refugees live in camps, which are crowded and without proper sanitation

World Health Organization's Advice On Stopping The Spread Of COVID-19

Lack of access to sanitation products and a safe space to isolate make it hard for many displaced people to follow these guidelines.

- Regularly wash your hands with soap and clean water
- Avoid crowded places and social distance if you must
- Stay home and isolate if you have symptoms

COVID-19 And The Refugee Crisis Timeline

March 11, 2020
World Health Organization declares **a global pandemic**

March 18, 2020
All new refugee admissions to the U.S. **are suspended**

April 4, 2020
1 million confirmed cases of COVID-19 worldwide

April 10, 2020
Refugee settlements begin building COVID-19 **isolation centers**

May 14, 2020
COVID-19 case found in Bangladesh Rohingya refugee camps

May 25, 2020
400 people attempt to flee Libya; 317 are caught and returned

June 10, 2020
Asylum applications reach lowest level in Europe

How Are Countries Handling The Refugee Crisis During The Coronavirus Pandemic?

America has the **19th highest** rate of refugees in the world, but the **highest** rate of death worldwide

10.7 million undocumented immigrants are not eligible for healthcare or government assistance

313,241 refugees in America

75,000 people outside of America are awaiting immigration trials

67% of the refugees in U.S. participate in the labour force

58% of the refugees do not speak English upon arrival in the U.S.

America's response to the Coronavirus Pandemic:

March 21, 2020
All non-essential cross border travel halted

43/50
states issue stay at home orders

7/50
states issue masks as a requirement

114,647*
COVID-19 deaths
*accurate as of June 12, 2020

Turkey has the **highest** rate of refugees in the world, but the **16th** highest rate of COVID-19 deaths worldwide

5 million refugees, irregular migrants, and asylum seekers

1/5 refugees lack access to clean drinking water

1/3 lack access to hygiene items

45% of Syrian refugees live in poverty

58% of refugees do not have access to healthcare

Refugees do not receive ~~FREE~~ accessible basic needs

Turkey's response to the COVID-19 Pandemic:

April 3, 2020
Masks become mandatory

March 13, 2020
Travel bans started

May 14, 2020
Schools closed

June 10, 2020
COVID-19 deaths: 4,729*
*numbers subject to change

167 countries so far have fully or partially closed their borders to contain the spread of the virus. At least 57 states are making no exception for people seeking asylum.

Issues Refugees Are Facing During The Coronavirus Pandemic

- Suspended visa applications
- No access to public funds
- Lack of sanitation in settlement camps
- No access to internet in camps
- Unfamiliarity with language and healthcare systems
- Fear of deportation if accessing services
- Lone children refugees without guardians
- Loss of jobs and lack of alternatives
- Inability to socially distance
- Host countries have poor health infrastructure

The struggles refugees face in accessing healthcare, government assistance, and safety will continue long beyond COVID-19. There are more displaced people in the world than ever before with no signs of flattening that curve.

Learn more about the lives of refugees and what you can do to help at:
https://donate.unhcr.org/gb/general/~my-donation

Join the conversation on social media:
#WorldRefugeeDay

VENNGAGE *Read the full blogpost at:* https://venngage.com/blog/refugee-infographic

Sources

- The World Bank
- [illegible]
- [illegible]
- World Health Organization
- The UN Refugee Agency
- [illegible]
- Wikipedia
- Refugees International
- [illegible]
- Center for American Progress
- Ballotpedia
- CNN
- [illegible]
- Migration Policy Institute
- The UN Refugee Agency
- Amnesty International
- [illegible]
- Wikipedia
- The UN Refugee Agency
- World Health Organization
- The UN Refugee Agency Southern Europe
- World Health Organization