ויזואליזציה של נתונים | הנחיות לעבודות הגמר

לעבודת הגמר קיימות שתי מטרות

1. ביצוע חקירה של dataset ב-Jupyter Notebook
בשפה המקצועית חקירה כזו נקראת Exploratory Data Analysis, ובקצרה EDA
2. הצגת הסיפור שעומד מאחורי סט הנתונים

בחירת סט הנתונים

- אתרו dataset שהוא אינו dataset טריוויאלי, שניתן להורדה באופן ציבורי.
  - מה הכוונה ב-dataset לא טריוויאלי?
בהקשר של הקורס שלנו, יספיקו התנאים הבאים כדי לקבוע ש-dataset הוא אינו טריוויאלי
    - מספר לא קטן של עמודות ומספר לא קטן של שורות
    - עדיף שהעמודות יהיו ממגוון של dtypes, למשל dataset שמכיל גם דאטה קטגורי וגם דאטה נומרי, או אפילו אחד שבנוסף מכיל גם עמודה/ות מסוג זמן (נניח שמכיל שנה, או חודש או תאריך מלא)
    - הערה: זו ממש אינה חובה שה-dataset יכיל את כל סוגי ה-dtypes, העיקר שתמצאו אחד שיש בו כמה סוגים.
    - ניתן למצוא סטים של נתונים באתרים הבאים:
    - Hugging Face, Data World, Kaggle וגם מאגר המידע הממשלתי
  - אם במקרה תמצאו שמבחינת הפרויקט שלכם זה נכון ואפשרי לחבר מספר datasets לכדי אחד מורכב יותר, אתם יותר מאשר מוזמנים לעשות את זה.

תיאור סט הנתונים

- מה מקורו, משמע מהיכן הורדתם אותו? רשמו את שם האתר וצרפו קישור ישיר אל ה-dataset או את ה-dataset עצמו (אם הוא גדול מדי, הסתפקו ב-1000 שורות)
- מהן השדות שהוא מכיל, מה ה-data type שלהם, כמה שורות הוא מכיל?
- פרטו את פעולות הנקיון / השלמה / נרמול שנדרשתם לבצע על סט הנתונים?
- האם נדרשתם לאחד מספר מקורות נתונים? אם כן, מהם? ובקצרה, כיצד עשיתם.ן זאת?
- האם חישבתם מטריקות? אם כן, כתבו במסודר את הגדרת המטריקה ואת הדרך בעזרתה חישבתם אותה.
  - בכיתה, למשל, חישבנו GDP per Capita וחישבנו גם Nobel Pieces לפי גודל האוכלוסיה

הגשת העבודה

הגשת העבודה באופן המתואר היא חלק מהציון הסופי. כל סטודנט מגיש למרצה שלו.

הגישו את העבודה בקובץ מכווץ ב-zip (שיכלול את הדברים הבאים:

1. **מסמך** מסכם המכיל את הדברים הבאים (**ולא עולה על 25 עמודים**)
   - תארו את סט הנתונים וענו על השאלות בפיסקה שלמעלה בשם "תיאור סט הנתונים"

- ניסוח הבעיה, שאלת/ות המחקר, השערות והציפיות
- בצעו EDA (דוגמא ל-EDA מצומצם, ניתן למצוא בהקלטות השיעור האחרון)
- ריכוז התובנות ***המרכזיות*** (אם ישנן כאלה שהן מאוד שוליות אל תכללו אותן) ביחד עם הויזואליזציה שהביא אתכם אל התובנה הזו.
  - אם קיימות ויזואליזציות שגרמו לכם לשנות ולדייק את שאלת המחקר / הצפיות / ההשערות שלכם, אנא הכלילו אותן במסמך.
- סיכום הנרטיב שעומד מאחורי סט הנתונים
- (רשות) האם עולות לכם שאלות המשך שהייתם ממליצים להמשיך לחקור? מהן? אילו נתונים נדרש לאסוף כדי לענות עליהן?

2. **מחברת** פייתון
3. **קישור** לדאטה סט או **הדאטה סט עצמו** (אם גדול מדי, ניתן לכלול 1000 שורות רנדומליות)

בנוסף שימו לב לדגשים הבאים:

- **תאריכים**, עבור שדות תאריכים אנו ממליצים להיעזר במחברת האחרונה שהועלתה המכילה דוגמה
- לניתוח של מכירות לאורך זמן, כמו גם בכמה מן הדוגמאות עליהן עברנו במהלך הסמסטר, למשל ניתוח של מניות.
- **ערכים חסרים או חריגים** נדרש מכם להפעיל שיקול דעת ולפרט/לנמק. כפי שראינו, ככל שעמודה יותר חשובה לנו כך לא נרצה שייגרעו ממנה רשומות (למשל, האם תבע או לא תבע בהקשר של חברת ביטוח). מצד שני, אם אין לנו חיווי טוב למה היה הערך האמיתי, נצטרך למחוק אותה. למשל, האם כל בעלי ההכנסה הגבוהה הגישו תביעה והעמודה חסרה, נוכל להשלים אותה ע״י כך. אך ייתכן שלא יהיה לנו שום רמז, ונצטרך למחוק את הרשומות. אם למשל חסרה לי עמודה שהיא (אולי) שולית יותר, למשל כמו שם פרטי אזי לא נצטרך להשלים ולא נצטרך למחוק. שימו לב: אתם לא בהכרח נדרשים שמסד הנתונים שלכם יכיל ערכים חסרים, אך זה יכול לתת נפח לתהליך ה-eda. אנו מודעים לכך שזה לא היה הפוקוס בסמסטר, אך זה יכול לתת בונוס של ניקוד על עבודה מעמיקה בנושא זה. אותו הדבר אמור לגבי ערכים חריגים.

קריטריונים להערכת הפרויקט וקביעת הציון

הציון הסופי בפרויקט ייקבע על בסיס הערכת המרכיבים הבאים:

1. בחירה, תיעוד וניקוי סט הנתונים

- מורכבות ורלוונטיות: הערכה תתייחס לבחירת סט נתונים שאינו טריוויאלי ("toy dataset").
- תיעוד הנתונים: ייבדק אם סופק תיאור מקיף של סט הנתונים, הכולל את מקורו, הגדרת המשתנים וסוגיהם (נומרי, קטגורי, תאריכי וכו').
- אפיון ראשוני: ביצוע אפיון ראשוני של הנתונים (למשל, באמצעות `info()` ו-`describe()`) וכן הגדרה והסבר של מדדים מחושבים.
- זיהוי וטיפול בבעיות: תינתן הערכה לזיהוי והתייחסות לבעיות פוטנציאליות בנתונים (ערכים חסרים, חריגות, סוגי נתונים שגויים).

2. ניתוח נתונים אקספלורטורי (EDA)

- ניתוח חד-משתני (Univariate Analysis):
  - ביצוע ניתוח מעמיק לכל משתנה משמעותי ותיאור ההתפלגות שלו.
  - בחירה מושכלת של ויזואליזציה המתאימה לסוג המשתנה
- ניתוח דו/רב-משתני (Bivariate/Multivariate Analysis):
  - הפגנת הבנה של מערכות היחסים בין המשתנים המרכזיים בסט הנתונים.
  - בחירת ויזואליזציה מתאימה לניתוח קשרים בין זוגות משתנים

3. איכות הויזואליזציה

- בהירות ואפקטיביות: כלל הגרפים והתרשימים צריכים להיות ברורים, קריאים ומציגים את המידע באופן אפקטיבי.
- סטנדרטים טכניים: הקפדה על מרכיבים חיוניים כגון כותרת ראשית, כותרות לצירים, ושימוש בסקאלה נכונה וברורה.
- שימוש מתקדם: שימוש יעיל במאפיינים ויזואליים (צבע, גודל, צורה) להדגשת תכונות ודפוסים משמעותיים בנתונים.

4. הפקת תובנות ופרשנות

- דיוק ועומק: מתן פרשנות מדויקת ובהירה לכל ויזואליזציה, החורגת מתיאור שטחי של הנתונים.
- זיהוי דפוסים: הדגשת התובנות המרכזיות והדפוסים המשמעותיים שעלו מניתוח הנתונים.
- בחינת השערות: התייחסות להשערות והציפיות שהועלו בתחילת התהליך ובחינתן לאור ממצאי ה-EDA.

5. בניית הנרטיב (Data Storytelling)

- בהירות המסר: הצגת סיפור נתונים בעל מסר מרכזי ברור וקל להבנה.
- עקיבות: הקפדה על קשר לוגי ועקיבות בין התובנות שהופקו בשלב ה-EDA ובין הנרטיב המוצג.
- רלוונטיות ואסתטיקה: בחירת ויזואליזציות רלוונטיות, ברורות, אסתטיות ואפקטיביות להעברת הסיפור.

כל אלה תוך יישום עקרונות תיאורטיים שנלמדו בכיתה, כמו שיפור גרפים והוכחת ידע והבנה במשימות תפיסתיות ועקרונות עיצוב ונראות.

שימו לב: הדגש הוא על איכות ה-EDA בהתייחס לויזואליזציות ולתובנות . אין צורך לבנות מודל חיזוי או משימה שלא במסגרת הקורס