**פרויקט בינה עסקית – הלוואות**

**קורס:**

בינה עסקית – מס' קורס 234020500

שנת ה'תשפ"ג – סמסטר ב'

**שמות הסטודנטים בפרויקט:**

עידן לוי – 208995191  
ליאור ברשדסקיי – 319000337

**מרצה מוביל הפרויקט:**אור יצחק פרץ

**קישור לפרויקט בגיטהאב:**<https://github.com/liorber13/BI-Project-3rd-year-2nd-semester>

**קישור לפרויקט בקולאב:**<https://colab.research.google.com/drive/1qcQNTntspgw1SrpQWwyTFfF6QeAHh-nS#scrollTo=HyJDFjuK4v0z>

**חלק 1: הבנה עסקית**

אוסף הנתונים נלקח מאתר Kaggle.

הנתונים עוסקים באישור למתן הלוואה על בסיס הכנסה שנתית של בני זוג.

<https://www.kaggle.com/datasets/mirzahasnine/loan-data-set>

**חלק 2: הבנת הנתונים**

בהתבסס על נתונים אלו, הועלו 3 שאלות עסקיות אותן נרצה להעלות מניתוח הנתונים:

1. מהו סוג האזור שבו סך ההלוואות שנלקחו הוא הגבוה ביותר וכמה כמות זו מהווה באחוזים מסך ההלוואות שנלקחו?
2. האם קיים קשר ישר/קשר הפוך/אין קשר בין מספר ילדים לגובה ההלוואה ומה עוצמתו של קשר זה חזק/בינוני/חלש?
3. מה הרבעון שבו ממוצע ההלוואות הוא הגבוה ביותר אצל נשואים?

לטובת המענה על שאלות אלו, נבחן מספר KPI's שיהוו את המדדים עבור שאלות אלו:

1. מהו הסכום הכולל של ההלוואות שנלקחו?
2. האם ניתן להגיש שבאזורים מסוימות נלקחות הלוואות יותר גדולות?
3. מה כמות הילדים הממוצעת של לוקחי ההלוואות?
4. מה סכום ההלוואות הממוצעת בחלוקה להשכלת הלווים?
5. מה ממוצע ההלוואות בחלוקה לרבעונים?

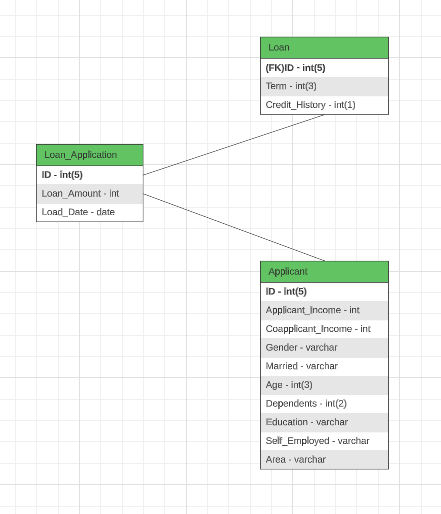
**חלק 3: הכנת הנתונים – ניקוי וסידור הנתונים**

**חלק Extract:**קראנו את הנתונים מקובץ CSV לקובץ פייתון באמצעות csv.DictReader כך שכל שורה מקובץ ה-CSV צורפה לרשימת הנתונים.

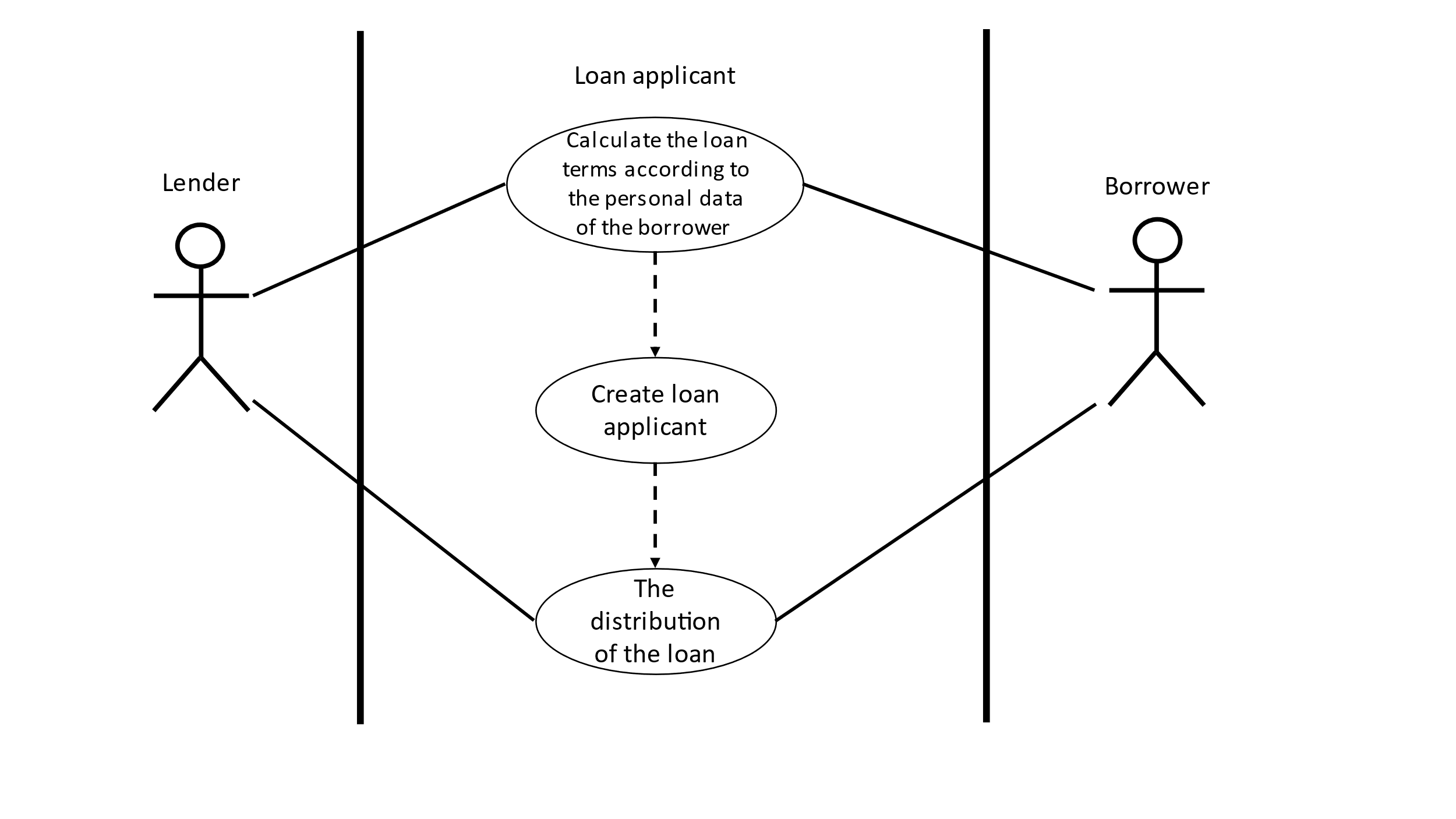
**חלק Transform:**הגדרנו מספר שורות כדי לאפיין את העמודות שאנו רוצים להציג לפי Loan\_fact – שמות השדות שקיימים בקובץ, Loan\_numerics – עבור ערכים מספריים שנרצה לטפל בהם, Loan\_non\_numerics – עבור ערכים לא מספריים שנרצה לטפל בהם. טבלת Masks מהווה כלי עזר שמכיל בתוכו 2 פונקציות לטיפול בשדות. טבלת ה-Fact מגדירה לנו את לב הנתונים, הטבלה שמקשרת בין כל מספרי הזיהוי של שאר הטבלאות ובעזרתה מבוצעים הניתוחים שמשלבים נתונים ממספר טבלאות על מנת לקבל משמעויות עמוקות ולא שטחיות.  
מחלקות ה-Database הוגדרו על מנת לעזור לנו לאפיין את המקור, את היעד, את פעולת הטרנספורמציה, ואת המיפוי בתהליך ה-ETL. בנוסף, השתמשנו גם במתודיות להוספת מקורות, יעדים, טרנספורמציות ומיפוי שהיוו את התהליך בפועל. המחלקות בבסיס הנתונים יורשות ממסד הנתונים ומספקות שיטות למשיכת המידע על בסיס מספרי זיהוי או שמות משתנים.

**חלק Load:**מחלקת ה-STTM הוגדרה על מנת לבצע את תהליך הטרנספורמציה לנתונים. בתוך מחלקה זו השתמשנו במתודה עבור מיפוי הנתונים וסידורם על פי ההגדרות שבחרנו. המתודה מחלצת את נתוני המקור ומחילה את "מסכות" הטרנספורמציה שציינו בטבלת Masks.  
הנתונים לבסוף מאוחסנים במילון.

**תרשים ERD**



**תרשים use-case**



לדעתנו מסד הנתונים יחסית עשיר ומאפשר תמיכה בתהליך קבלת ההחלטה על גובה ההלוואה הצפוי בהתבסס על הנתונים שברשותנו. התהליך דורש קבלת מידע ראשוני מהלווה, המידע מנותח על פי הנתונים שברשותנו ולבסוף מתקבל סכום הלוואה שתואם לנתוני הלווה. התהליך לא אמור לקחת יותר ממספר דקות בודדות על מנת לקבל הצעה לקבלת הלוואה.

**חלק 4: מודל**

הסכמה שנבחרה היא סכמת Star. סכמה זו מתאימה למסד הנתונים שלנו מכיוון שסכמת הנתונים בנויה באופן של טבלה מרכזת אחת שממנה יוצאות מספר טבלאות משנה שאינן קשורות באופן היררכי כלשהו. למשל עבור מספר מזהה של כל הלוואה שניתנה, ניתן לראות את הגיל של לוקח ההלוואה, את ההכנסה של בעל ההלוואה, את החודשים להלוואה וכו'.  
בשלב זה כתבנו מספר שאילתות חוצות טבלאות על מנת לנתח משמעויות מתוך הנתונים שהתקבלו. שאילתות אלו יוצגו בנספח א'.

**חלק 5: מסקנות**על ידי בחינת תוצאות השאילתות, נוכל לזהות את האזורים עם סכומי ההלוואה הגבוהים ביותר, תוך מתן תובנות לגבי התפלגות ההלוואות באזורים שונים.  
בנוסף, יתאפשר לנו לצפות במגמות או דפוסים כלשהם בסכומי הלוואות לאורך רבעונים שונים, מה שיכול להצביע על שינויים עונתיים או שינוי בהתנהגות הלווים.  
ניתן יהיה להבין את התפלגות הגילאים של מבקשי הלוואות, ולזהות אילו קבוצות גיל נוטות יותר להגיש בקשה להלוואות. נדע מה סכום ההלוואה המקסימלי ומי משויך אליה.  
ניתוח התוצאות יעזור לנו להבין את הקשר בין מספר הילדים שיש למבקש ההלוואה לבין הכנסתו הממוצעת. בחינת התוצאות יכולה לספק תובנות לגבי אילו רמות השכלה קשורות לסכומי הלוואה גבוהים יותר.  
כל אלו יוכלו לעזור לנו לגבש תמונת מצב על אופי לוקחי ההלוואות ולהביא משמעויות קונקרטיות לבנקים ולמוסדות שמאפשרות לקיחת הלוואות מכל סוג.

**נספח א' – שאילתות SQL**

1. SELECT A.Area, SUM(LA.Loan\_Amount) AS Total\_Loan\_Amount  
   FROM Loan\_Applicant LA  
   JOIN Applicant A USING (ID)  
   GROUP BY A.Area;

שאילתה שתציג לכל אזור את סך כל ההלוואות שנלקחו.

1. SELECT QUARTER(L.Loan\_Date) AS quarter, AVG(LA.Loan\_Amount) AS average\_amount  
   FROM Loan\_Applicant LA  
   JOIN Loan L USING (ID)  
   GROUP BY QUARTER(L.Load\_Date)  
   ORDER BY quarter;

בשאילתה זו יהיה ניתן לראות את ממוצע ההלוואות בחלוקה לרבעונים.

1. SELECT \*  
   FROM Loan\_Application  
   Order by CASE

WHEN Age <= 25 THEN '25 and under'  
WHEN Age > 25 AND AGE <= 35 THEN '25-35'  
WHEN Age > 35 AND Age <= 45 THEN '35-45'  
ELSE '45 and over'  
END;

שאילתה זו מציגה את כל פרטי ההלוואות בחלוקה לקבוצות גילים שהגדרנו.

1. SELECT Education  
   FROM Applicant  
   GROUP BY Education HAVING AVG(Age) >=   
   ALL (SELECT AVG(Age)   
   FROM Applicant  
   GROUP BY Education);

השאילתה תחזיר את רמת ההשכלה שבה כל האנשים מבוגרים יותר מהגיל הממוצע של כל האנשים.

1. SELECT MAX(Loan\_Amount)  
   FROM Loan\_Application as LA  
   WHERE EXISTS (SELECT \* FROM Applicant as A WHERE A.ID = LA.ID);

השאילתה תעבור על כל ההלוואות של אותו האדם ותציג את הגבוהה ביותר.

1. SELECT Dependents, AVG(Applicant\_Income)  
   OVER(PARTITION BY Dependents)  
   FROM Applicant;

לכל כמות צאצאים תוצג ההכנסה הממוצעת של מגישי הבקשה להלוואה.

1. SELECT A.Education, SUM(LA.Loan\_Amount) AS Total\_Loan\_Amount  
   FROM Loan\_Applicant LA  
   JOIN Applicant A USING (ID)  
   GROUP BY A.Education;

שאילתה לחישוב סכום ההלוואה הכולל לפי רמת השכלה.

1. SELECT A.Education, AVG(A.Age) AS Average\_Age  
   FROM Applicant A  
   GROUP BY A.Education;

שאילתה למציאת הגיל הממוצע של מועמדים עם רמות השכלה שונות.

1. SELECT A.ID, A.Dependent AS Number\_of\_Dependents  
   FROM Applicant A  
   WHERE Area = Urban  
   GROUP BY A.ID;

שאילתה לקביעת מספר הילדים הממוצע עבור מבקשי הלוואות מאזורים עירוניים.

1. SELECT Loan\_Amount  
   FROM Loan\_Application  
   ORDER BY Loan\_Amount DESC  
   LIMIT 5;

שאילתה למציאת 5 סכומי ההלוואה המובילים.