

# פרויקט ניתוח סיכוני אשראי

הפרויקט מבוסס על [מערך נתוני סיכוני אשראי מאתר kaggle.com](https://www.kaggle.com) המציגים היסטוריית הלוואות של לקוחות. הנתונים מאפשרים לבצע ניתוח סיכוני אשראי ולבחון אילו לקוחות נוטים יותר להגיע לחדלות פירעון, כמו גם לחזות יכולת ההחזר של הלוואות בעתיד. לניתוח הנתונים והבנת סיכוני האשראי, ניסיתי לזהות דפוסים ומאפיינים המבדילים לווים מסוכנים מלוויים יציבים, ומכאן עלו שאלת המחקר ושאלות המשנה.

## שאלת המחקר: מה הפרופיל של לווה מסוכן?

שאלות שנעזרתי בהם כדי לענות על שאלה זו:

- האם לווים עם היסטוריית חדלות פירעון מצליחים להחזיר הלוואות חדשות?
- האם לווים צעירים מסוכנים יותר מלוויים מבוגרים?
- האם יש קשר בין החזר מההכנסה לוותק תעסוקתי?
- האם סכום גבוה מגדיל את סיכון ההלוואה?
- האם דירוג ההלוואה מנבא את סיכויי ההחזר?

כדי להבין את סיכוני האשראי בצורה מקיפה, התחלתי בניקוי ראשוני באקסל להסרת ערכים חריגים ולא הגיוניים. לאחר מכן, המשכתי בארגון הנתונים בפאנדס, עברתי ל-SQL לביצוע שאילתות מורכבות וניתוח דפוסים בין טבלאות, ולבסוף השתמשתי בטאבלו להצגת התובנות בצורה ויזואלית ונוחה להבנה.

## פאנדס:

הכנת נתונים:

טיפולתי בנתוני האשראי על ידי מילוי הערכים החסרים בריבית ההלוואה לפי הממוצע של כל דירוג הלוואה, ובערכים החסרים של ותק העבודה רשמתי 0. בנוסף, יצרתי עמודת יחס חוב להכנסה והפכתי קודים למילים ברורות ( N/Y הפכו ל-Yes/No ו-0/1 הפכו ל-Non-Default / Default).

```
import pandas as pd

# טעינת הנתונים
df = pd.read_csv('credit_risk_dataset.csv')

# מילוי ריבית חסרה לפי ממוצע של כל דירוג
df['loan_int_rate'] =
df.groupby('loan_grade')['loan_int_rate'].transform(lambda x:
x.fillna(x.mean()))

# מילוי ותק עבודה 0 בערכים חסרים
df['person_emp_length'] = df['person_emp_length'].fillna(0)

# החלפת 0 ו-1 למילים ברורות (Default / Non-Default)
df['loan_status'] = df['loan_status'].replace({0: 'Non-Default', 1:
'Default'})

# החלפת Yes ו-No ל-Y ו-N
df['cb_person_default_on_file'] =
df['cb_person_default_on_file'].replace({'N': 'No', 'Y': 'Yes'})

# יצירת עמודת יחס חוב להכנסה
df['debt_to_income'] = df['loan_amnt'] / df['person_income']

# שמירת הקובץ
df.to_csv('cleaned_credit_data.csv', index=False)
```

## פיצול הקובץ:

יצרתי מזהה ייחודי לכל לווה, בניתי מילון סוגי הלוואות שמאפשר שינוי קל של סוג ההלוואה מבלי לערוך ידנית, ופיצלתי את הנתונים לשלושה גיליונות: פרטי הלקוח, פרטי ההלוואה ומילון סוגי הלוואות.

```
import pandas as pd
import os

desktop_path = os.path.join(os.path.expanduser("~"), "Desktop",
                             "Credit_Project_Final_15.xlsx")

# טעינת הנתונים
df = pd.read_csv('credit_risk_data.csv')
df['person_id'] = range(1, len(df) + 1)

# יצירת המילון
loan_types = pd.DataFrame({
    'type_id': range(1, len(df['loan_intent'].unique()) + 1),
    'loan_intent': df['loan_intent'].unique()
})

df['type_id'] =
df['loan_intent'].map(dict(zip(loan_types.loan_intent,
                               loan_types.type_id)))

# פיצול לגיליונות:
persons = df[['person_id', 'person_age', 'person_income',
              'person_home_ownership',
              'person_emp_length', 'cb_person_cred_hist_length']]

loans = df[['person_id', 'type_id', 'loan_grade', 'loan_amnt',
            'loan_int_rate',
            'loan_status', 'loan_percent_income',
            'cb_person_default_on_file', 'debt_to_income']]

# שמירת הקובץ
```

```
with pd.ExcelWriter(desktop_path) as writer:
    persons.to_excel(writer, sheet_name='Persons', index=False)
    loans.to_excel(writer, sheet_name='Loans', index=False)
    loan_types.to_excel(writer, sheet_name='Loan_Types',
index=False)
```

## טבלאות:

טבלת Persons – לקוחות	
שם עמודה	תיאור בעברית
person_id	מזהה לקוח
person_age	גיל
person_income	הכנסה שנתי
person_home_ownership	בעלות על דירה
person_emp_length	ותק תעסוקתי (בשנים)
cb_person_cred_hist_length	אורך היסטוריית אשראי (בשנים)

  

טבלת Loans – הלוואות	
שם עמודה	תיאור בעברית
person_id	מזהה לקוח
type_id	מזהה סוג הלוואה
loan_grade	דירוג הלוואה
loan_amnt	סכום הלוואה
loan_int_rate	ריבית הלוואה
loan_status	סטטוס הלוואה
loan_percent_income	אחוז החזר מההכנסה
cb_person_default_on_file	היסטוריית חדלות פירעון
debt_to_income	יחס חוב להכנסה (DTI)

  

טבלת Loan_Types – סוגי הלוואות	
שם עמודה	תיאור בעברית
type_id	מזהה סוג הלוואה
loan_intent	מטרת ההלוואה

## :SQL

- קודם כול בדקתי שכל השדות מלאים, כדי לוודא שאין ערכים חסרים שעלולים לפגוע  
בניתוח הנתונים.

```
select *  
from persons  
where person_id is null  
or person_age is null  
or person_income is null  
or person_home_ownership is null  
or person_emp_length is null  
or cb_person_cred_hist_length is null
```

```
select *  
from loans  
where person_id is null  
or type_id is null  
or loan_grade is null  
or loan_amnt is null  
or loan_int_rate is null  
or loan_status is null  
or loan_percent_income is null  
or cb_person_default_on_file is null  
or debt_to_income is null
```

```
select *  
from loan_types  
where type_id is null  
or loan_intent is null
```

- חילקתי את הלווים לקבוצות גיל לפי עשורים ובחנתי עבור כל קבוצה את כמות הלווים, ההכנסה החצינית, יחס חוב הכנסה, אחוז השוכרים ושיעור חדלות הפירעון, כדי לבדוק האם לוויים צעירים מסוכנים יותר ממבוגרים.

```
select
case
when p.person_age >= 60 then '60+'
else concat(floor(p.person_age / 10) * 10, '-',
floor(p.person_age / 10) * 10 + 9)
end as age_group_text,
count(*) as borrower_count,
percentile_cont(0.5) within group (order by p.person_income) as
median_income,
avg(l.debt_to_income) as avg_dti,
avg(case when p.person_home_ownership = 'RENT' then 1.0 else 0.0
end) as renter_percentage,
avg(case when l.loan_status = 'Default' then 1.0 else 0.0 end)
as default_rate
from persons p
join loans l on p.person_id = l.person_id
group by
case
when p.person_age >= 60 then '60+'
else concat(floor(p.person_age / 10) * 10, '-',
floor(p.person_age / 10) * 10 + 9)
end
order by
min(p.person_age)
```

- חילקתי הלוואות לפי מטרת ההלוואה ובחנתי עבור כל מטרה את סכום ההלוואה הממוצע, שיעור חדלות הפירעון ומספר ההלוואות, כדי לבדוק האם סכומי הלוואה גבוהים יותר מגדילים את סיכון ההלוואה.

```
select
t.loan_intent,
round(avg(l.loan_amnt), 0) as avg_loan_amount,
```

```

round(avg(case when l.loan_status = 'Default' then 1 else 0 end)
* 100, 2) || '%' as default_rate,
count(l.person_id) as total_loans
from loans l
join loan_types t on l.type_id = t.type_id
group by t.loan_intent
order by avg_loan_amount desc

```

- חילקתי את ההלוואות לפי דירוג ובחנתי עבור כל דירוג את מספר ההלוואות ושיעור חדלות הפירעון, כדי לבדוק האם דירוג ההלוואה מנבא את סיכויי ההחזר.

```

select
loan_grade,
count(person_id) as total_loans,
round(avg(case when loan_status = 'Default' then 1 else 0 end) *
100, 2) as default_rate
from loans
group by loan_grade
order by loan_grade asc

```

## טאבלו:

בטאבלו יצרתי גרפים שממחישים את שאלות המחקר ומסייעים להבין את הפרופיל של לווים הנוטים יותר לחדלות פירעון.

## מסקנות הניתוח:

מאפייני לווה בסיכון גבוה: אדם שהגיע בעבר לחדלות פירעון, בגילאי 20-29 או מעל גיל 60, שלוקח הלוואה לאיחוד חובות, בעל ותק עבודה נמוך מ-4 שנים ודירוג ההלוואה שלו D ומטה.