הכנת נתונים עם Knime, חלק א'

נלמד טכניקות הכנת נתונים לאנליזה דרך למידה מסרטונים תוך כדי ביצוע תרגילים קצרים. נשתמש בכלי חינמי מצוין בשם [Knime](https://www.knime.com).

מומלץ להצטייד באזניות על מנת לשמוע את הסרטונים המלווים את העבודה.

# התקנה (אופציונלי).

הכלי מותקן כבר במחשבים במעבדה. הרוצה לתקין את הכלי במחשב הפרטי, ישתמש ב- [סרטון](https://youtu.be/8ISIeFKkoOE?list=PLz3mQ6OlTI0Zoz5pcha7tiPP9vGU8UDza).

# יסודות עבודה עם Knime

## יצירת פרויקט

תבצעו בעזרת ה- [סרטון](https://www.youtube.com/watch?v=-JtO7DW9Jr0) (3:44):

1. תריצו את Knime.
2. תיצרו Workflow Group. כל הפרויקטים (workflows) שלך יהיו בקבוצה זאת.
3. תיצרו Workflow (פרויקט).

## עבודה עם קדקודים

קדקוד מייצג פעולה כלשהי של הכנת נתונים (כגון קריאת קובץ, סינון נתונים וכו'). אחרי שנגדיר מספר קדקודים וקשרים ביניהם, נוכל לחזור על הפעולות בנקל.

תבצעו בעזרת ה- [סרטון](https://youtu.be/fMM_w4v5zZc) (4:02):

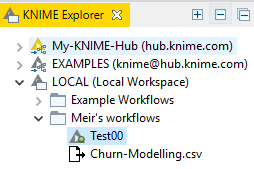
1. תמצאו קדקוד File Reader (בקבוצת קדקודים IO->Read. ניתן גם להשתמש בשורת חיפוש).



1. תגדירו את הקדקוד כך שהוא יקרא את קובץ ChurningModel.csv הנמצא ב- מודל.
2. תבצעו את הקדקוד, תפתחו את הנתונים לציפייה.
3. תוסיפו קדקוד Row Filter ותחברו אותו לקדקוד הקודם, כך שהסינון יתבצע אחרי קריאת הקובץ.
4. תגדירו את הקדקוד כך שרק שורות עבור בני 50 (כל מגדר) ומטה יישארו. הערה: הסרטון לא מראה את זה. עליכם לחשוב לבד איך למלא את טופס הגדרות הקדקוד.
5. תבטלו ביצוע הקדקוד **של קריאת הקובץ**.
6. תבצעו את קדקוד הסינון (שימו לב שקריאת הקובץ התבצע גם כן).
7. תפתחו את הנתונים שאחרי הסינון לציפייה.
8. תמצאו דרך למיין את הנתונים לפי סדר יורד של גיל. האם הסינון עבד?
9. תפתחו את הנתונים לציפייה בעזרת הקדקוד הקודם (של קריאת הקובץ). מה אתם רואים?
   1. זה אומר ש**אפשר לראות את הנתונים שהיו אחרי כל שלב.**
   2. זה כמו פיתוח תכנית באמצע ניפוי שגיאות!
10. תבטלו ביצוע קדקוד של קריאת הקובץ. מה קרה עם קדקוד סינון?

## דרכי עבודה עם קבצים

* 1. Explorer של Knime יכול להציג קבצים:



* 1. הקבצים צריכים להיות בתיקייה של ה Workflow **group**.
  2. כדי להציג קובץ שהוכנס לתיקייה הנ"ל באמצע עבודה ב Knime, יש להשתמש בכפתור Refresh של Explorer 
     1. Refresh לא עבד אצלי והייתי צריך לסגור את Explorer ולפתוח אותו מחדש מתפריט View.
  3. אפשר גם למשוך קובץ ל Explorer.
  4. תכניסו את Churn-Modelling.csv ל Explorer.

## יצירת תיאור קדקודים

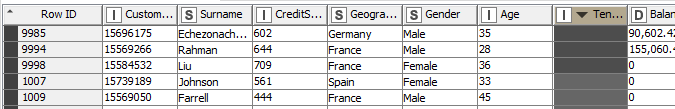
תבצעו בעזרת ה- [סרטון](https://youtu.be/AHURYB_O8sA) (3:02):

1. תנו תיאורים לקדקודים שיצרתם.
2. תיצרו מסגרת צבעונית מסביב הקדקודים עם תיאור של פעולת שניהם ביחד.

## ציפייה ב- נתונים

תבצעו בעזרת ה- [סרטון](https://youtu.be/5kj0ESmaL4k) (4:28):

1. תלמדו איך נתונים שמורות בטבלאות ב Knime. האם כל הנאמר חוזר על מה שלמדתם בקורס בסיסי נתונים או יש משהו ספציפי ל Knime?
2. תציגו את הנתונים שקראתם מהקובץ כך שחלק מהתצוגה יראה באופן הבא:



# הכנת נתונים לאנליזה

קיימות מספר פעולות קלסיות עבור הכנת נתונים לאנליזה. נקדיש סעיף לכל פעולה.

## סינון (Filtering)

המטרה של סינון היא להוריד מהנתונים שורות ועמודות המפריעות לאנליזה איכותית, כמו שהוסבר בסרטונים [1](https://youtu.be/NJwWwpIEBpg) (0:59) ו- [2](https://youtu.be/wQE_cXwDH-I) (0:43). אפשר לחלק פעולות סינון לקטגוריות הבאות:

### סינון שורות בסיסי [3](https://youtu.be/j3YhdEgu0Z0) (5:52), [4](https://youtu.be/rBmGjMu9EG4) (3:15), [5](https://youtu.be/nomaYlGJwmA) (2:52).

אחרי ציפייה בסרטונים הנ"ל, תיצרו קבוצת קדקודים הדרושים כדי לקבל טבלת לקוחות העונים לתנאים הבאים עם תיאורים של קדקודים ומסגרת (ראה צילום למטה):

1. יש ערך בעמודת Exited (בדיקה: 10,000 שורות אחרי הסינון עד כה).
2. אינם גרים בגרמניה (בדיקה: 7,491 שורות אחרי הסינון עד כה).
3. בגיל מ 40 עד 49, כולל הקצוות (בדיקה: 1,878 שורות אחרי הסינון עד כה).
4. שם משפחה מתחיל באות B והאות השנייה מ a עד e כולל (בדיקה: 44 שורות אחרי הסינון). (רמז – ביטויים רגולריים).

### סינון שורות מתקדם [6](https://youtu.be/WcpEIzzZ-yc) (6:40).

אחרי ציפייה בסרטון הנ"ל, תיצרו קבוצת קדקודים הדרושים כדי לקבל טבלת לקוחות העונים לתנאים הבאים עם תיאורים של קדקודים ומסגרת (ראה צילום למטה):

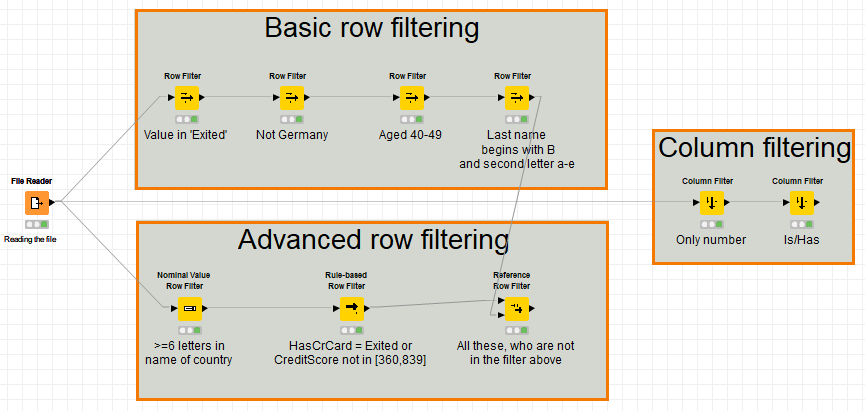
1. גרים במדינה שיש לפחות 6 אותיות בשמה. (בדיקה: 7,523 שורות אחרי הסינון עד כה). חובה להשתמש ב- Nominal Value Row Filter ו- ביטויים רגולריים (לא ניתן לבחור מדינות באופן ידני).
2. הערך בעמודה HasCrCard שווה לערך בעמודה Exited או הערך בעמודה CreditScore אינו בין 360 ל 839 כולל (בדיקה: 3,005 שורות אחרי הסינון עד כה).
3. אינם בין הלקוחות שקיבלנו בעזרת סינון שורות בסיסי בסעיף 3.1.1 (בדיקה: 2,998 שורות אחרי הסינון).

### סינון עמודות [7](https://youtu.be/b9k39b-6oxo) (4:49).

אחרי ציפייה בסרטון הנ"ל, תיצרו קבוצת קדקודים הדרושים כדי לקבל טבלה עם עמודות העונים לתנאים הבאים עם תיאורים של קדקודים ומסגרת (ראה צילום למטה):

1. טיפוס נומרי כלשהו.
2. שם העמודה מתחיל ב Is או Has (לא ניתן לבחור עמודות באופן ידני).

ל- יתר בהירות: הפתרון שלי לשאלות על סינון נראה כך:



### להרחבת ידע: [8](https://youtu.be/miUZNhePBLg) (1:40).