1. היי לכולם, אני יוליה והפרויקט שלי **עוסק בחיזוי רמת הפופולריות** של שירים בספוטיפיי.

2. שאלת המחקר שלי היתה: **מה גורם לשיר להיות פופולרי- הסאונד או החשיפה שלו? המטרה** היתה **לחזות את מדד הפופלריות** שהוא משתנה רציף ונמדד בסקאלה בין 0-100 **באמצעות מאפייני שמע ומידע על פלייליסטים, אלבומים ואמנים.** עבדתי על דאטה של כ-33 אלף שירים, מתוך קגל, שכלל 23 פיצ'רים ראשוניים.

3. תהליך העבודה כלל הכנת הנתונים לFlat file כאשר הורדתי כל מיני אימוג'ים כמו להבה, ניתוח EDA, ניקוי חריגים והשלמת מיסינג, יצירת פיצ'רים חדשים, בחירת פיצ'רים חשובים ולבסוף בניית מודלים וfine tuning. כל השלבים נעשו בפייתון דרך גוגל קולאב.

4. **בשלב ה-EDA ראיתי שרוב השירים יצאו ביום שישי ובחודש ינואר. רוב השירים נמצאים בטווח** ציוני פופולריות של 20-60. **הפיקים** ב2008 ו2014 מצביעים על % השירים עם פופולריות 0 כאשר שנת 2008 היא שנת השקת הספוטיפיי. חלק מהמשתנים כמו **ספיצ'נס, אקוסטיות, אינסטרומנטליות** הראו זנב ימני, כלומר **רוב השירים לא אקוסטיים או דיבוריים, אבל יש קבוצה קטנה שכן.**

5. **כאן בדקתי את הקשרים בין כל הפיצ'רים המספריים.** רואים למשל **קשר חיובי חזק בין אנרגיה ורעש** - לאודנס. לעומת זאת **קשר שלילי בין אקוסטיות לאנרגיה- כלומר שירים יותר אקוסטיים הם פחות אנרגטיים. קשר חיובי מתון בין עד כמה השיר נשמע שמח או אופטימי לבין עד כמה השיר מתאים לריקוד.** הקשרים של המשתנים עם הפופולריות עצמה חלשים. קורלציה 1 זה משך השיר – שזה אותו פיצ'ר ובסוף השארתי 1.

6. ניקוי החריגים במשתנים רציפים נעשה בשיטה של IQR כולל קורלציה והתפלגות. מה ששינה התפלגות אך לא קורלציה הוסר והושלם באמצעות מודל MICE.

7. כאן יצרתי משתנים חדשים. התווספו עוד מהשלב הקודם משתנים טמפורליים כמו שנה, חודש ויום בשבוע.
אחר כך משתני חשיפה- כמה שירים יש לאמן, כמה באלבום, וכמה פלייליסטים כוללים את השיר.
בנוסף יצרתי משתנים מבוססי טקסט מהמילים הנפוצות בשם השיר. **למשל המילה "original" היא דווקא קשורה לשירים פחות פופולריים מהממוצע.**
בסוף שלב Feature engineering, לאחר סינון משתנים חופפים, נותרו 57 פיצ'רים.

8. ניסיתי שלוש גישות לבחירת פיצ'רים:
1. – **One-Hot Encoding** אחרי שזרקתי את עמודות שמות האמן והפלייליסט.
2. – **Target Encoding** הפעלתי על **שם האמן ושם הפלייליסט.**
3. **גישה משולבת** Target Encoding –יחד עם הפיצ'רים החזקים מהגישה הראשונה.

בכל אחת מהגישות הראשונות, הפיצ'רים נבחרו לפי הסכמה של לפחות שני מודלים מבין Lasso, Ridge, Random Forest וGradient Boosting-.
בסוף נבחרה הגישה של **Target Encoding,** כי היא גם הסבירה הכי הרבה מהשונות בפופולריות — בערך **55%**.
זה היה אחרי Fine-Tuning ל XGBoost-עם **216שילובים ב.GridSearchCV-**
אפשר לראות בתרשים שהגורמים המשפיעים ביותר הם **החשיפה בפלייליסטים וביצועי האמן ,הרבה יותר ממאפייני הסאונד.**

9. ראיתי **שהפקטור הכי משמעותי לפופולריות** של שיר הוא **החשיפה שלו** —**דרך האמן או הפלייליסטים שהוא מופיע בהם.**
המאפיינים של הסאונד עצמו, כמו **קצב או אנרגיה ,כן משפיעים, אבל הרבה פחות.**
המודל מצליח להסביר בערך **55% מהשונות בפופולריות** ,מה שאומר שיש עוד גורמים שלא נכללו – כמו **שיווק, פרסום או טרנדים חברתיים.**

יש גם מגבלות שצריך לזכור:
מדד הפופולריות עצמו נקבע על ידי ספוטיפיי והוא **לא מדד אובייקטיבי.**
שני שירים עם אותה כמות השמעות יכולים לקבל ציוני פופולריות שונים, לפי פרמטרים כמו **מתי האזינו לאחרונה, כמה זמן שמעו, וכמה שיתפו.**
בנוסף, פופולריות היא משהו **דינמי** —מה שפופולרי היום לא בהכרח יהיה פופולרי מחר.
לסיכום אפשר לראות שפופולריות היא **לא רק עניין של סאונד, אלא הרבה יותר עניין של חשיפה.**