פרויקט גמר תכנות מדעי בשפת פייתון
ניתוח Dataset מוזיקה בין השנים 1950-2019

עומר סלע 314977570
ראובן ונטורה 328759675

1. **תיאור הפרויקט:**
ניתוח של מידע מ-dataset בעל מעל ל-28,000 שורות, המכיל מידע על שירים שיצאו בין השנים 1950-2019. בפרויקט נרצה להוכיח או להפריך הנחות שהנחנו טרם השגת המידע ולקבל אינפורמציה חיונית לגבי האמנים והז׳אנרים המתוארים בו.

2. **עיבוד המידע:**
העמודות, לפני תהליך עיבוד המידע, הן:

,'artist_name', 'track_name', 'release_date', 'genre'
,'lyrics', 'len', 'dating', 'violence', 'world/life', 'night/time'
,'shake the audience', 'family/gospel', 'like/girls', 'romantic', 'communication'
,'obscene', 'music', 'movement/places', 'light/visual perceptions'
,'family/spiritual', 'sadness', 'feelings', 'danceability', 'loudness'
'acousticness', 'instrumentalness', 'valence', 'energy', 'topic', 'age'

בעיבוד המידע נרצה להסיר את העמודות 'acousticness', 'like/girls' ו-'instrumentalness', כיוון שהמידע שהן נותנות בעל משמעות מעורפלת, או במקרה של עמודת 'instrumentalness', גם הערכים עצמם בעמודה אינם מובנים ואינם עומדים בקנה אחד עם הערכים שנמצאים בעמודות האחרות.
כמו כן, נוסיף עמודה בשם 'word_count' אשר מכילה את ספירת המילים שנמצאת בעמודת 'lyrics' עבור כל שורה. נשתמש במידע שנמצא בעמודה זו על מנת להוכיח הנחות מסויימות כפי שנראה בהמשך.
כעת העמודות באובייקט הדאטא פריים processed_data יראו כך:

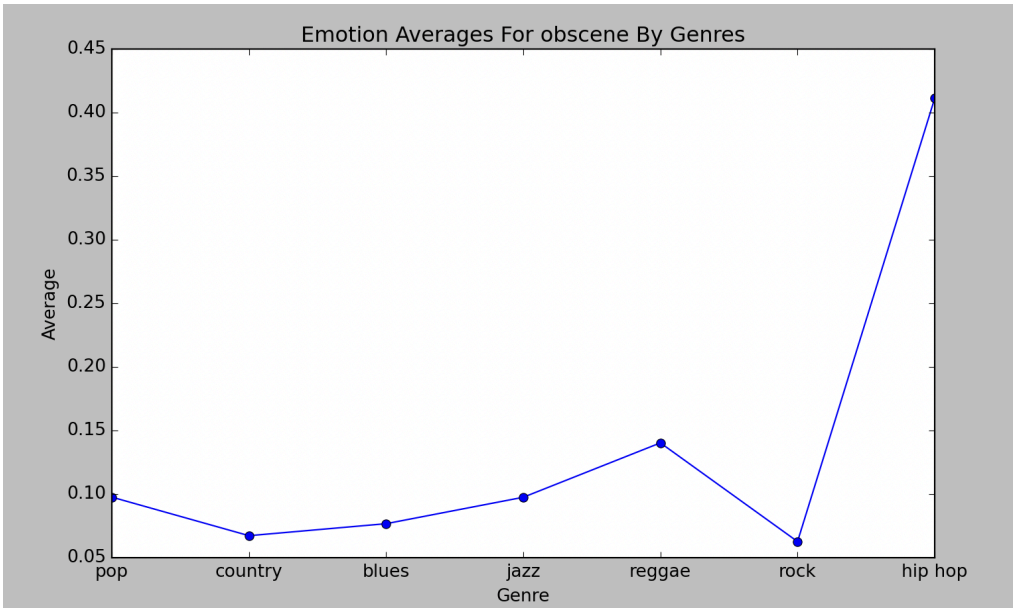
,'artist_name', 'track_name', 'release_date', 'genre
,'lyrics', 'len', 'dating', 'violence', 'world/life', 'night/time'
,'shake the audience', 'family/gospel', 'romantic', 'communication'
,'obscene', 'music', 'movement/places', 'light/visual perceptions'
,'family/spiritual', 'sadness', 'feelings', 'danceability', 'loudness'
,'valence', 'energy', 'topic', 'age'

עמודת הז׳אנרים תהווה עמודת ה-class במהלך העבודה, כיוון שנרצה להבין את הקשר בין ז׳אנרים מסויימים לבין הנושאים המדוברים בהם, או השפעתם על מגמות כלליות במוזיקה.

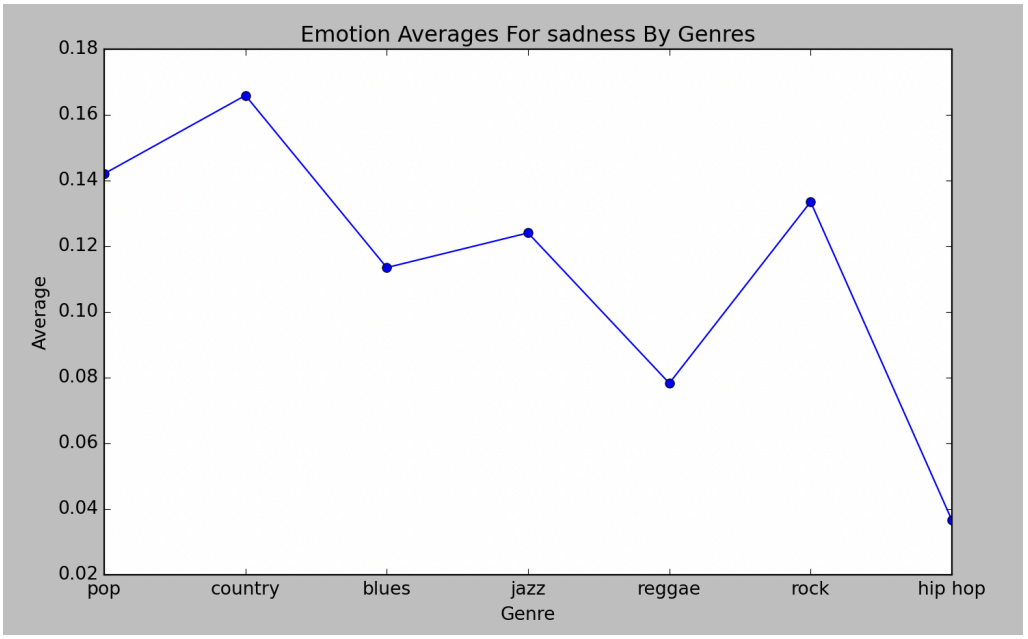
תהליך עיבוד המידע מצורף בקובץ ה-ipynb.

3. פיזור מאפיינים מעניינים:

מתיאור העמודות ניתן לראות כי רוב העמודות מכילות מידע לגבי יחס מסויים של רגש או תכונה המתוארים בשיר, למשל הערך בעמודה 'sadness' יהיה גבוה בשירים עצובים ונמוך בשירים שמחים. כמו כן מתוארים גם מאפיינים שעשויים להיות יותר ספציפיים לז׳אנר מסויים.
למשל, במקרה של עמודת 'obscene' ('בוטה'), נראה כי הז׳אנר הבוטה ביותר הוא היפ הופ:



לעומת זאת, הז׳אנר העצוב ביותר הוא קאנטרי, אך עם פער פחות מובהק:



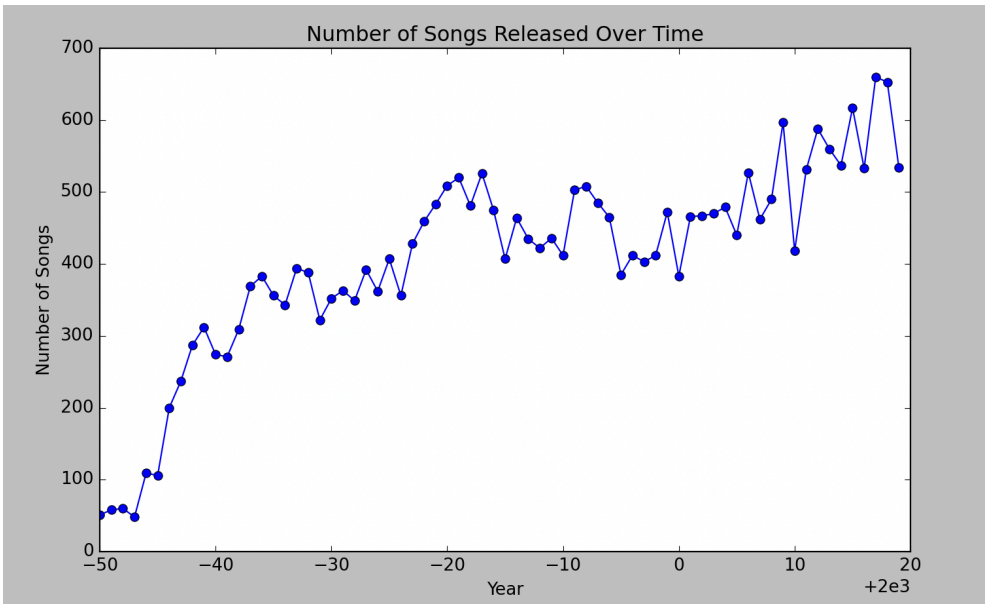
4. **הנחות טרם עיבוד המידע:**

- **כמות שירים ממוצעת לשנה עולה לאורך זמן** - נצפה לראות מגמה עולה של שירים היוצאים בשנה, מתוך הבנה שבחלוף הזמן נוצרים ז׳אנרים חדשים המאפשרים לאמנים רבים יותר לבטא את עצמם, וגם מכיוון שהאמצעים ליצירת מוזיקה הולכים ונעשים נגישים יותר לקהל הרחב.
- **אורך שיר ממוצע יורד לאורך זמן** - נצפה לראות מגמת ירידה באורכם של שירים לאורך זמן, במיוחד בשנים 2010-2019, לאור עליית המדיה החברתית והצורך לתפוס את תשומת הלב של המאזינים בזמן קצר, לעומת שנים בהן הרדיו שלט בקידום אמנים והשמעת מוזיקה.
- **כמות ז׳אנרים גדלה לאורך זמן** - מכיוון שז׳אנרים חדשים נוצרים מתוך ז׳אנרים ישנים (למשל רוק שהגיע מהבלוז, היפ הופ שמושפע מג׳אז ופופ וכו׳), וז׳אנרים ישנים נשארים רלוונטיים, נצפה לראות עלייה בכמות הז׳אנרים בהם שירים יוצאים לאורך השנים.
- **קשרים בין רגשות או נושאים לז׳אנרים מסויימים** - כפי שראינו קודם, נצפה לראות קשרים מובהקים כלשהם בין ז׳אנרים לבין נושאים או מאפיינים המדוברים בהם, כמו במקרה של היפ הופ שהוא הז׳אנר הבוטה ביותר בממוצע במאגר.
- **עלייה בכמות הממוצעת של מילים לשיר** - בתהליך עיבוד המידע הוספנו את עמודת 'word_count', המונה את כמות המילים בעמודת 'lyrics' עבור אותה שורה. נצפה לראות עלייה בכמות המילים הממוצעת של שירים לאורך זמן, בעקבות היווצרות ועלייה בפופולריות של ז׳אנרים שטכניקה ומשחקי מילים הן אבן יסוד בהם, כמו היפ הופ למשל.

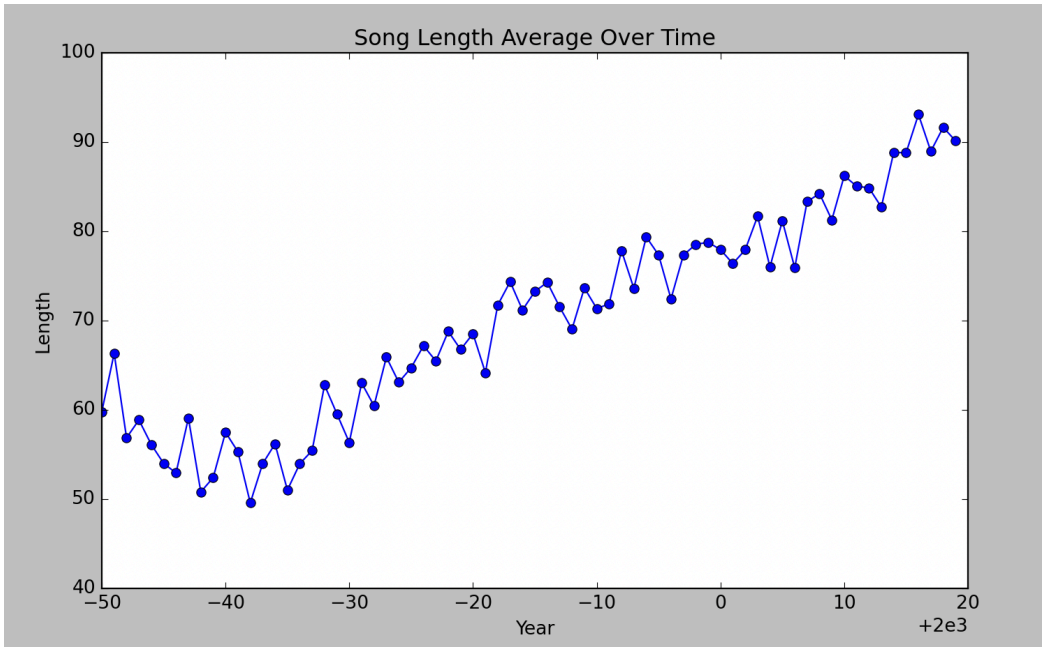
בנוסף, נייצר פונקציות עזר לבדיקת ההנחות ופונקציות עצמאיות מבדיקות אלו, שייתנו לנו מידע כמו כמות אמנים unique בכל ז׳אנר, השנים בהן אמן היה פעיל ועוד.

5. **הוכחה/הפרכה של הנחות קודמות:**

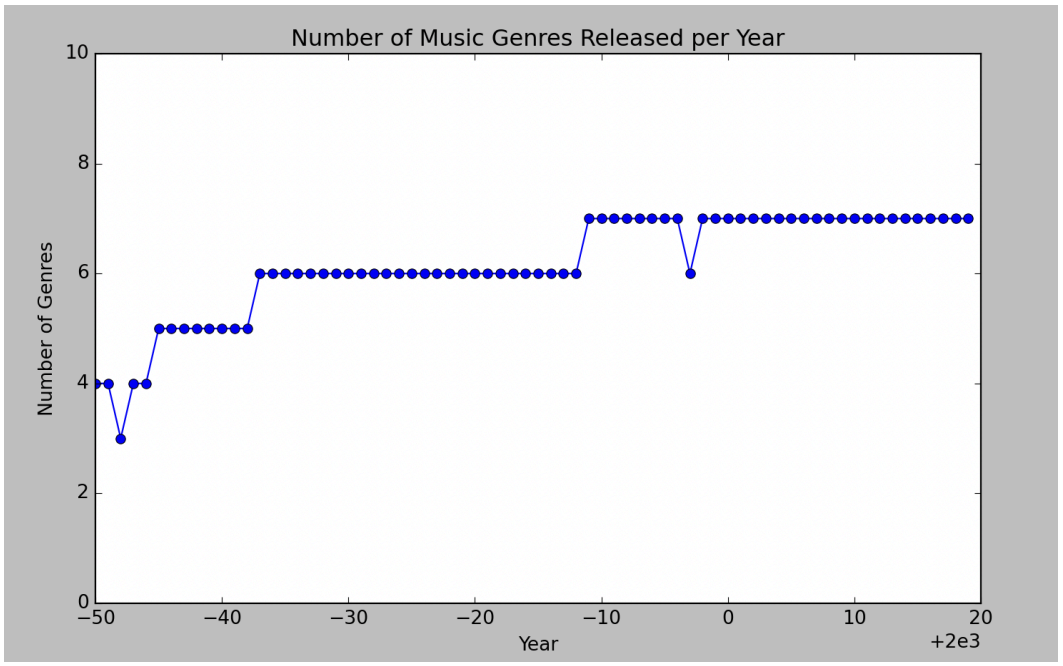
- **כמות שירים ממוצעת לשנה עולה** - ניתן לראות כי כמות השירים הממוצעת לשנה אכן במגמה עולה לאורך זמן:



- **אורך שיר ממוצע עולה לאורך זמן** - בניגוד לציפייה שלנו, אורך השירים דווקא עולה בממוצע לאורך זמן:

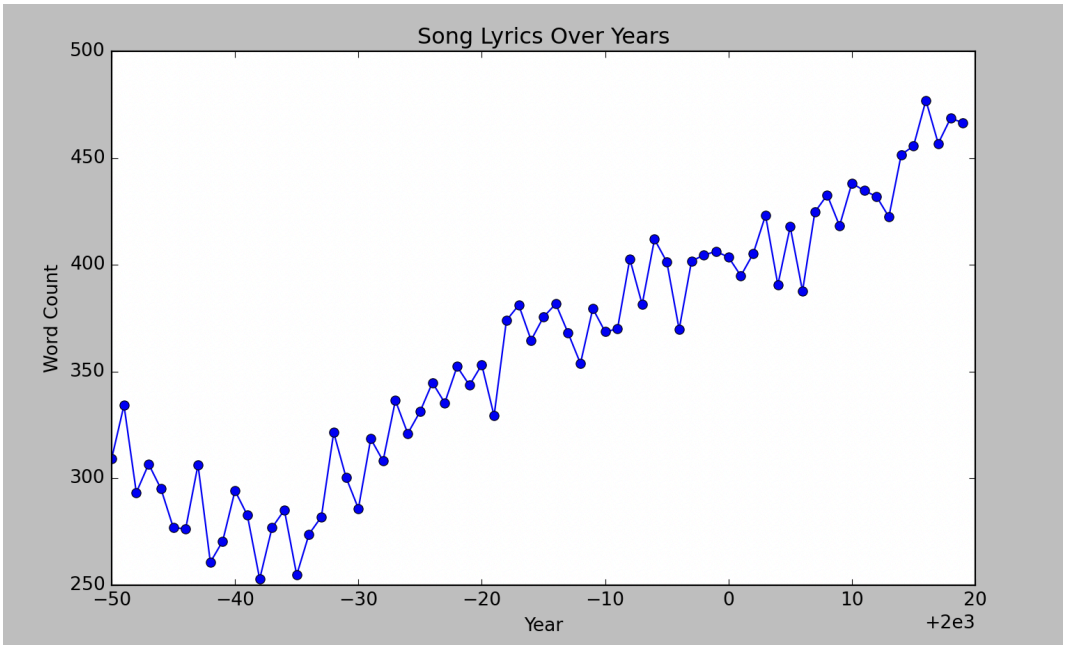


- **מספר ז׳אנרים עולה לאורך זמן** - כמות הז׳אנרים בהם יוצאים שירים עולה לאורך זמן, כפי שציפינו לראות:



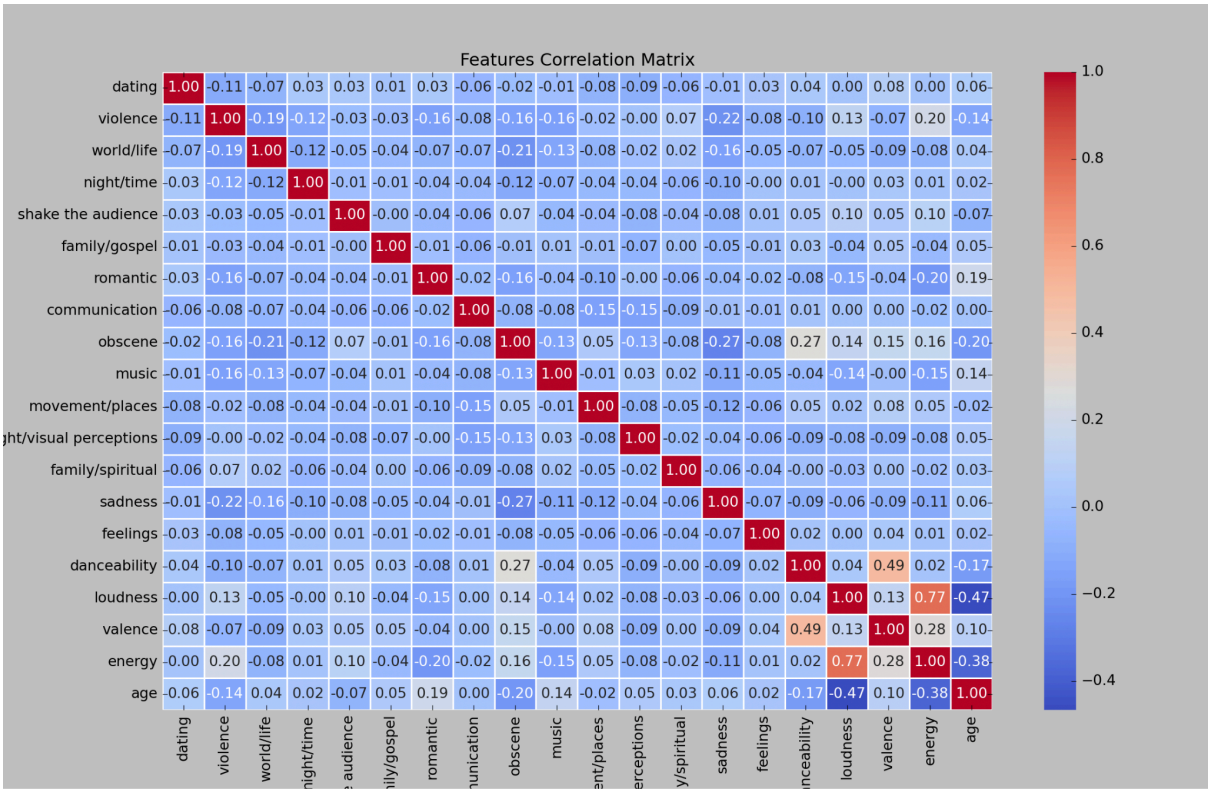
- **קשרים בין רגשות/מאפיינים לז׳אנרים שונים** - ראינו בסעיפים קודמים הוכחות לקשרים בין ז׳אנרים לבין הרגשות או הנושאים המובעים בהם.

- **עלייה בכמות מילים ממוצעת לשיר** - לאחר ההוספה של עמודת word_count, ניתן לחשב את המגמה עבור כמות מילים ממוצעת לשיר לאורך זמן, ולראות כי המגמה אכן עולה:



6. קשרים בין פיצ׳רים:

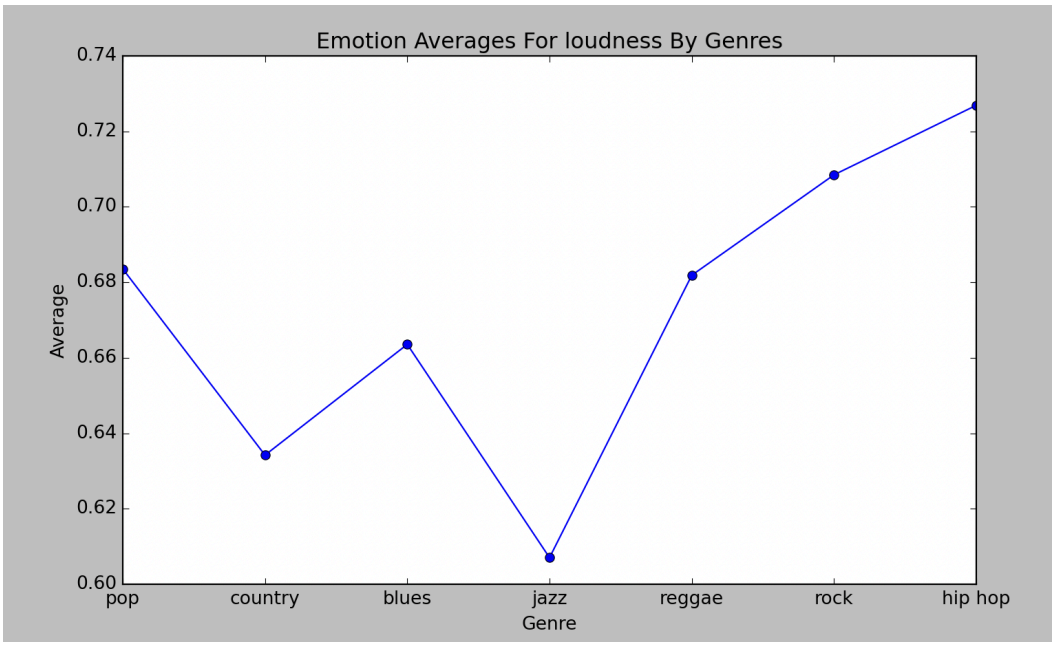
במטריצת הקורלציות הבאה:



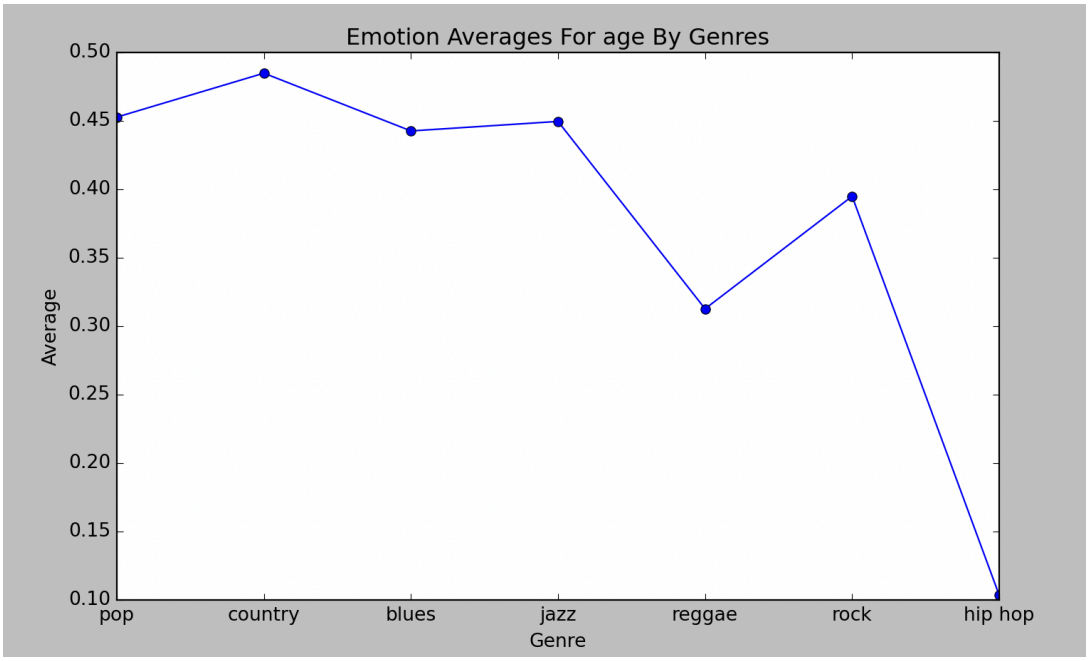
ניתן לראות כי הקשר הישיר החזק ביותר הוא בין עמודת 'energy' לעמודת 'loudness', כלומר ככל ששיר יותר אנרגטי כך הוא יותר רועש ולהיפך. בנוסף אנחנו רואים גם קשר הפוך בין עמודות אלו לעמודת 'age', המסמלת את 'גילו' של שיר מסויים, לפי השנה שבה יצא.

7. **פיצ׳רים המשפיעים על עמודת הקלאס:**

כפי שראינו מקודם, ישנם מאפיינים המופיעים יותר בז׳אנרים מסויימים ופחות באחרים, למשל עמודת 'obscene' שגבוהה במיוחד בהיפ הופ. נוכל להשתמש באותה פונקציונליות על מנת למצוא מאפיינים של ז׳אנרים אחרים:



הז׳אנרים הרועשים ביותר הם היפ הופ, רוק ורגאיי, בעוד שהז׳אנר השקט ביותר הוא ג׳אז.
נשתמש באותה פונקציונליות על עמודת 'age' ונקבל את הגרף הבא:



הז׳אנרים הותיקים ביותר במאגר המידע הם אלו עם ממוצע הגיל הגבוה ביותר, בעוד שהחדשים יותר הם בעלי ממוצע גיל נמוך. בנוסף, ניתן להסיק מכך על מגמות פופולריות של ז׳אנרים:
מכיוון שהז׳אנרים קאנטרי, בלוז וג׳אז אינם פופולריים היום כפי שהיו בעבר, יוצאים בהם פחות שירים חדשים, מה שגורם לשירים ה״זקנים״ יותר להשפיע בצורה מכרעת על ממוצע גילאי השירים.

8. סיכום:

בעבודה מול מאגר המידע הצלחנו להוכיח ולהפריך הנחות שהיו לנו טרם לעבודה, ואפילו להסיק מסקנות על פופולריות של ז׳אנרים מסויימים לאורך השנים, אפילו כאשר אין עמודה בפועל שמצביעה על כמות מכירות או השמעות של שיר מסויים.

למדנו לחשוב בצורה מחקרית כאשר אנחנו מקבלים סט גדול של מידע, ולשאול שאלות שאנחנו מעוניינים לקבל עליהן תשובות מתוך המידע עצמו, ובכך למעשה לא לעבוד ״על עיוור״ ולייצר פונקציות שאין בהן ערך. שאלות אלו עזרו גם להחליט אילו עמודות אינן נחוצות במידע שלאחר העיבוד, ואילו עמודות נצטרך להוסיף על מנת לקבל פיסות מידע שאנחנו מחפשים.