**קטלוג סרטונים**

**שאלה עסקית**

כיצד נחלק בין הסרטונים כדי להגדיל את הסיכויים לכניסה ותשלום באתר, ובפרט: איך לשפר את בחירת המיקום להצגת הסרטונים (Home/Bottom Page)?

**משימות**

1. קטלוג 100 הסרטונים הקיימים לקטגוריות הבאות:
   * "Hot"
   * "Stable and Popular"
   * "Everything else"
2. דקירת הפיצ'רים המתאימים לכל קטגוריה
3. שימוש באופן הקטלוג ליצירת שיטה חכמה ומסודרת מוכוונת פרודקט

**הנחות עבודה**

* צוות המוצר ירצה לדרג את הסרטונים העתידיים עפ"י נתוני מסגרת, ללא נתונים על ביצועי עבר (לדוג' שפה או איכות, אך לא כמות צפיות)
* אין יכולת להתאמה אישית של סרטונים עבור כל משתמש (לדוג' להציע סרטונים בשפה שמתאימה לו על בסיס שפות סרטונים אחרים בהם צפה)

**מסקנות**

מאפייני סרטון מוצלח לפי משקל (בסדר יורד):

* אורך קצר
* איכות גבוהה

**תהליך עבודה**

**שלב א': ניתוח עפ"י נתוני הצפיות בלבד וחלוקה לקבוצות**

* + - 1. הגדרת מדדים להצלחה – בהסתמך על נתוני הצפייה בלבד
         * כמות צפיות כוללת
         * ממוצע צפיות ליום
         * כמות ימים בהם הסרטון נצפה
         * ההפרש (באחוזים) בין הממוצע השבועי לכמות הצפיות השבועית (Avg Offset)
         * אחוז הימים בהם הסרטון עקף את הממוצע השבועי
         * כמה מהר הגיע הסרטון לכמות קבועה של צפיות
      2. לאחר ניתוחים של כלל השדות, הוחלט שהשדות שהכי משקפים את הצלחת הסרטון הם:
         * כמות צפיות כוללת (כלל הסרטונים נמדדים על אותה כמות ימים, בדיוק 120 יום מאז העלאתם)
         * Avg Offset
         * אחוז הימים בהם הסרטון עקף את הממוצע השבועי
      3. בניית שדות מחושבים לפי המדדים הנ"ל בתוך טבלת Video Features (דרך Python, מצורף כ-”data configuration”)
      4. דירוג פנימי עבור כל מדד במספר רץ (Rank, 1 – מקום ראשון, 2 – שני וכו')
      5. יצירת כלי לפילוח לקבוצות לפי הדירוגים עפ"י ההגדרות הבאות:
         * כאשר כל המדדים גדולים מ-X, הסרטון יסווג **כמוצלח** (כדי למנוע מצבים חריגים בהם אחד המדדים גבוה והשני נמוך, מעיד על volatility) - **TOP**
         * אם לא, כאשר כל המדדים גדולים מ-Y, הסרטון ידורג **כיציב ופופולרי** (ע"מ לתת דגש על היציבות) **- MIDDLE**
         * אחרת, יסווג הסרטון כ"אחר" **- BOTTOM**

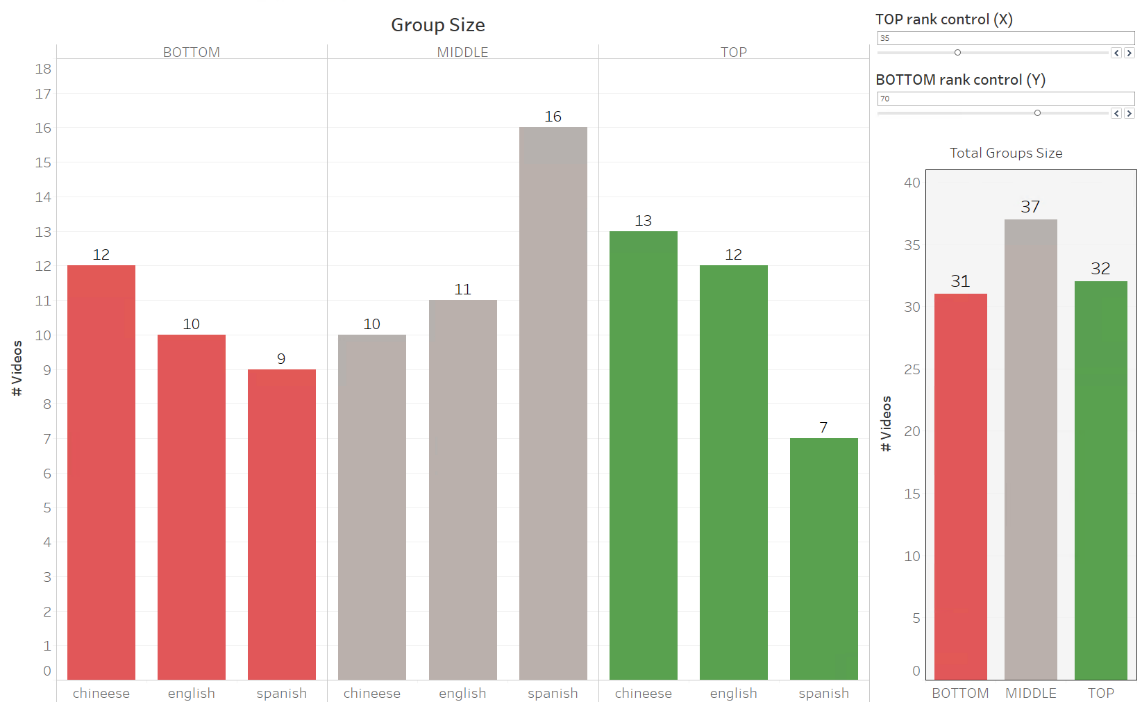
**שלב א': תוצאות**

**בניית הקבוצות**

הדירוג החלש ביותר (בין 1 ל-100) עבור הקבוצה העליונה – 35 || **32%** מהסרטונים

הדירוג החלש ביותר (בין 1 ל-100) עבור הקבוצה האמצעית – 70 || **37%** מהסרטונים

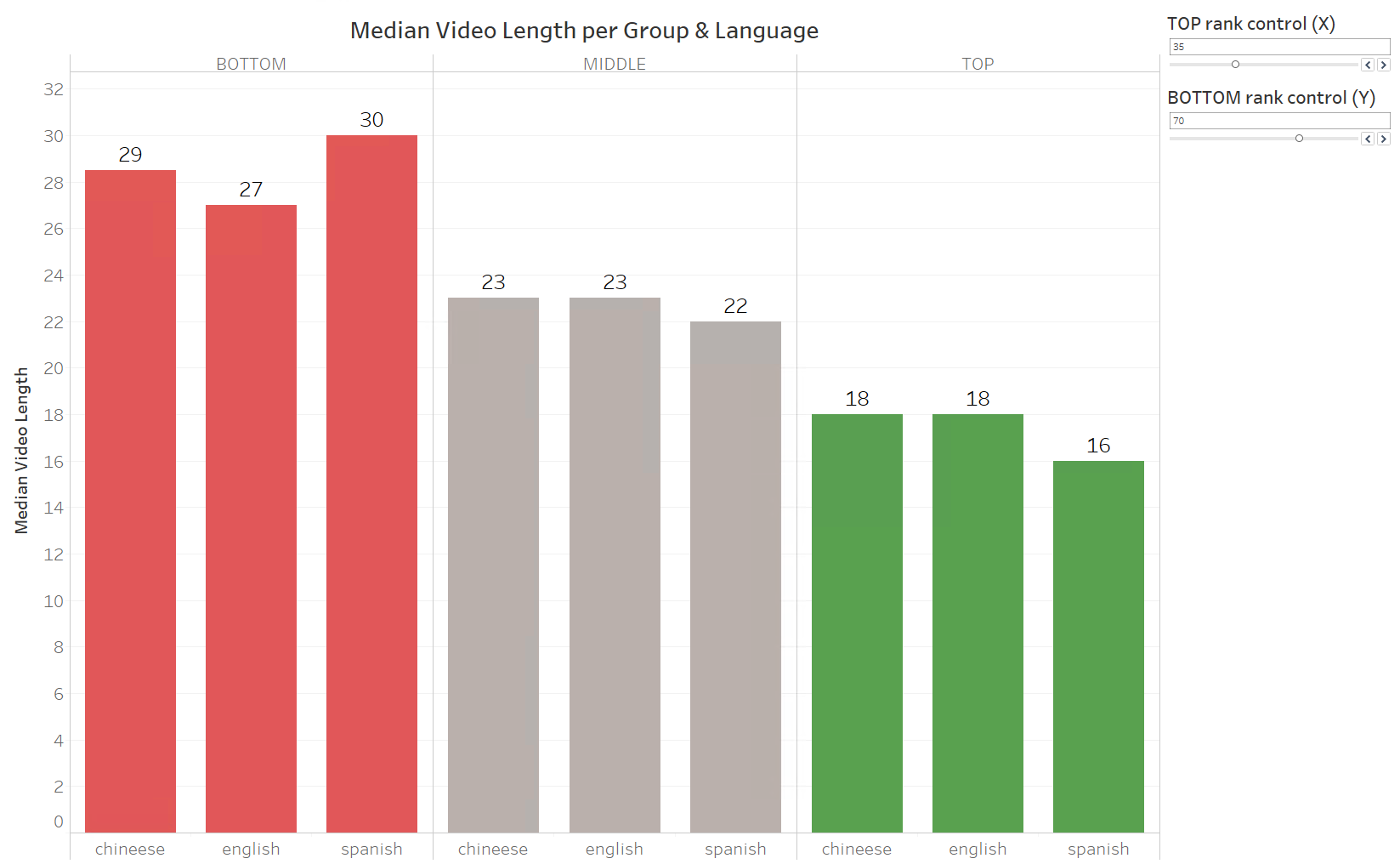
כל השאר || **31%** מהסרטונים



***מסקנה נוספת: סרטונים בספרדית פחות נפוצים בקבוצת הלהיטים***

**אורך הסרטון**

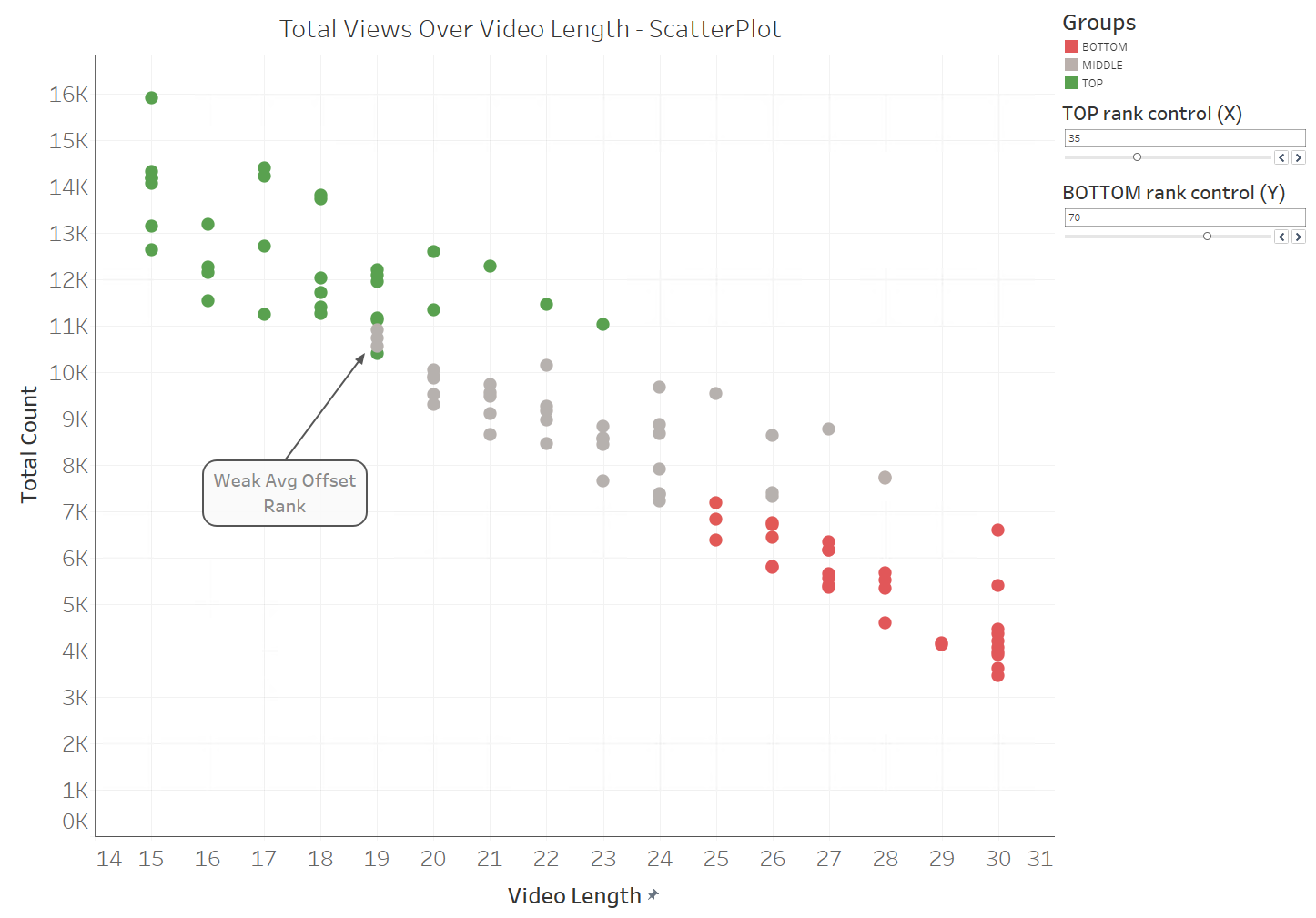
להלן חישוב אורך הסרטון החציוני עבור כל שפה וכל קבוצה

מסקנה: **הצלחת הסרטון – ביחס הפוך לאורכו!**

***מסקנה נוספת: אין קורלציה מובחנת בין שפת הסרטון לאורכו***

גרף פיזור – מספר צפיות על אורך הסרטון

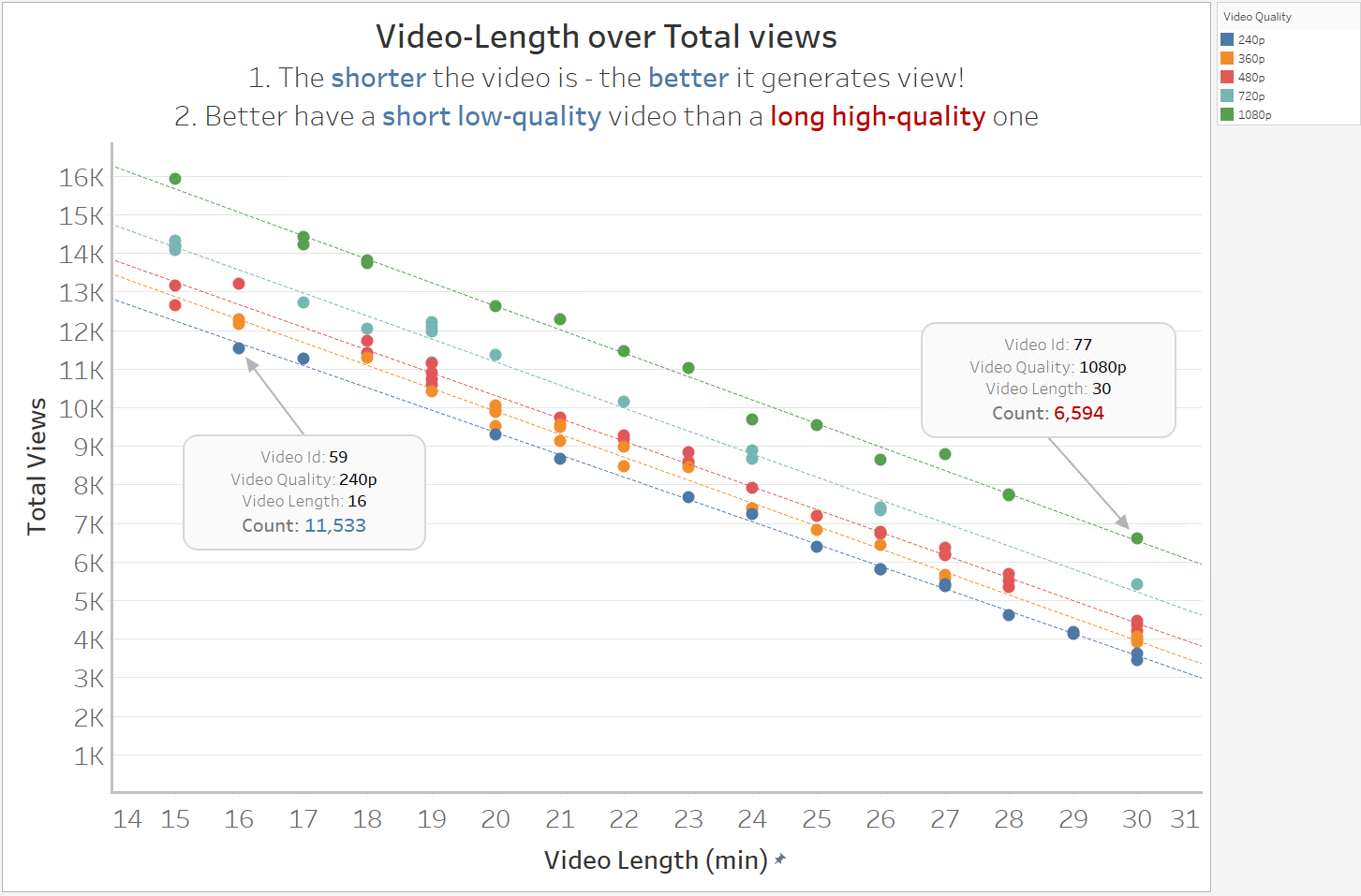
ניתן לראות את המגמה מהסעיף הקודם בבירור



***מסקנה נוספת: שיטת הסיווג משקפת את המציאות והקבוצות פרושות כפי שניתן היה לצפות.*** *גם מעט הסרטונים מהקבוצה האמצעית שנראים שייכים לקבוצה העליונה הם בעלי מדד חלש יחסית של סטייה ממוצע שבועי*

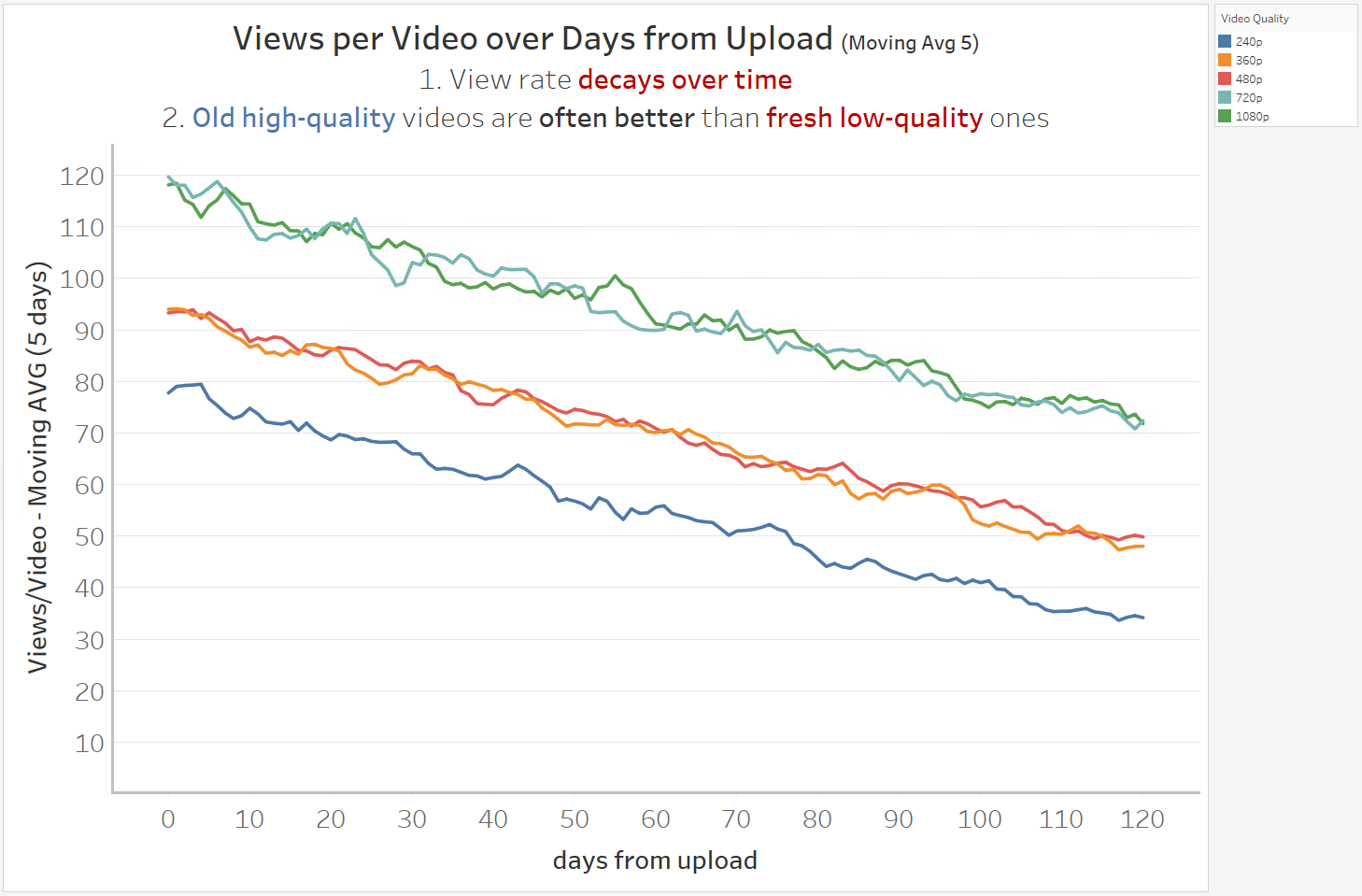
**איכות הסרטון:**

מסקנה: **פרמטר אורך הסרטון הוא גורם בעל משקל רב יותר מאשר האיכות**

 ***מסקנה נוספת: כצפוי, יש יחס ישר בין איכות הסרטון להצלחתו אולם יש להתחשב גם בשאר המשתנים, כמודגם בתרשים***

**זמן העלאה:**

מסקנה: ככל שעובר הזמן מהעלאה – יצפו פחות בסרטון

**מסקנה נוספת:**

* + - 1. אקספלורציה – פילוח הנתונים לפי השדות השונים כדי למצוא דפוסים וקורלציות בין מימד לנתוני הצפיות
         * יחס בין אורך לצפיות
         * יחס בין איכות לצפיות
         * ממוצע צפיות לסרטון לאורך זמן מעלייתו
      2. מציאת תכונות משותפות בתוך קבוצת הסרטונים הבולטים בנתוני הצפיה, וכן בתוך בעלי הנתונים החלשים
         * קיבוץ שבועי, דירוג הסרטונים לכל שבוע
         * פיזור הסרטונים על ממוצע סטייה שבועית (כמות צפיות שבועית לעומת הממוצע השבועי לסרטון) על כמות צפיות כוללת
      3. מתוך התוצאות, נדרג את הסרטונים שלנו ונאמן מודל סיווג שיקבל נתוני מסגרת ויתן פלט של סיווג לפי הקטגוריות הרצויות
         * Random Forest
         * לאמן עם קלט הכולל נתוני צפיה
         * לאמן עם קלט ללא נתוני צפיה