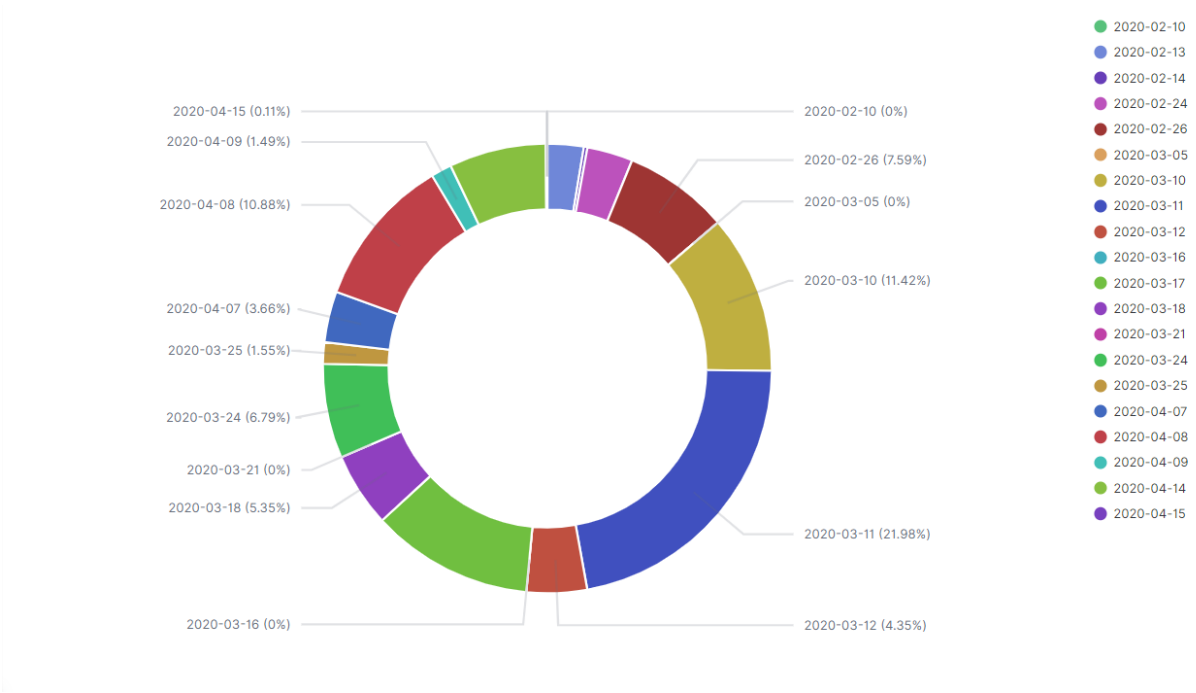
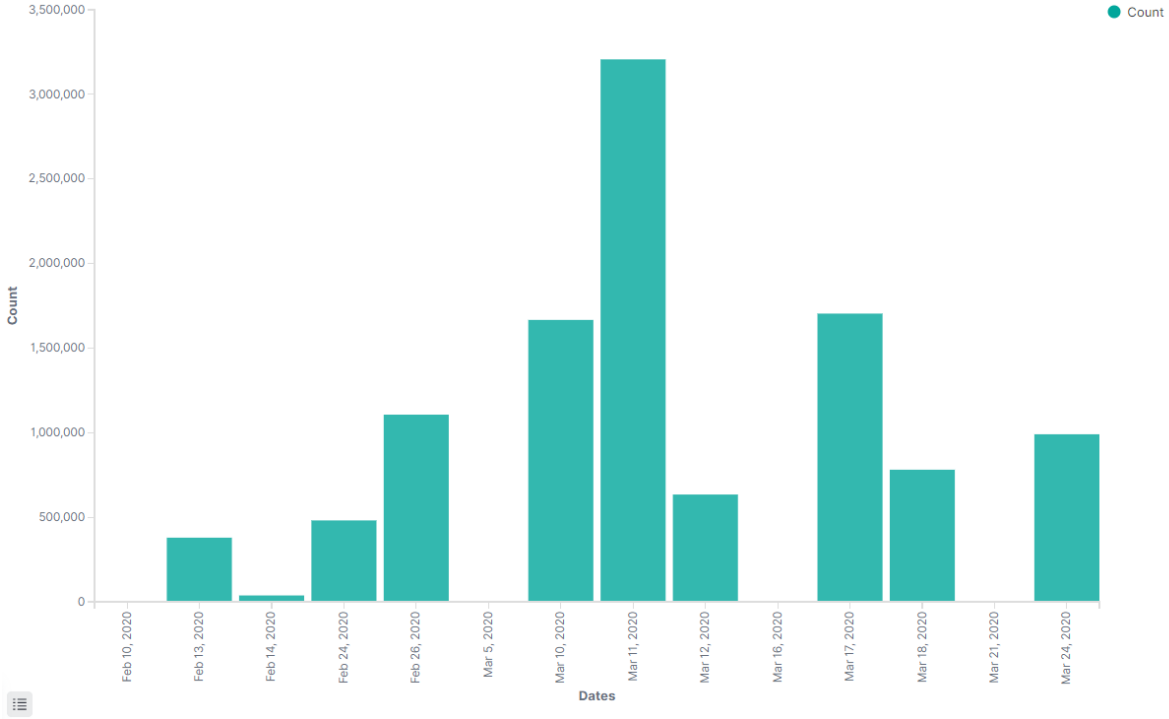
Visualizations

תחילה נציג מידע אודות כלל הדאטא שאספנו:

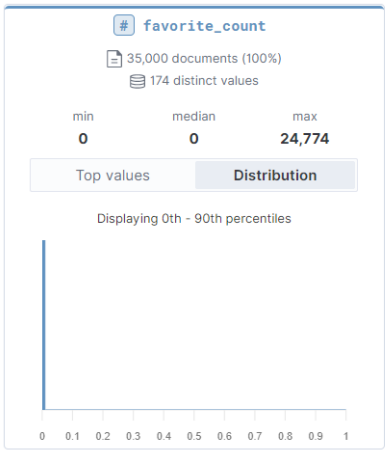
מספר הטוויטים בכל אחד מהימים שבחרנו:

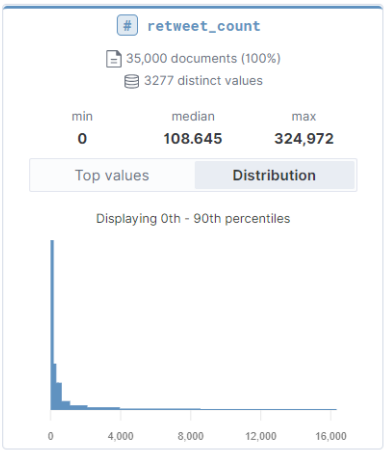


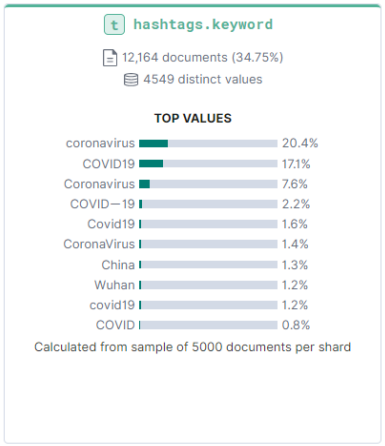


*בחלק מהנושאים שחילצנו מ-Kafka נמצאו גם רשומות מתאריכים סמוכים לנושא

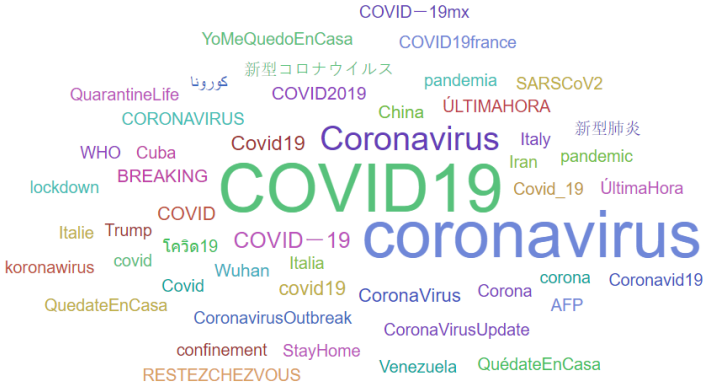
סטטיסטיקות הנוגעות לתכונות מהדאטא שאגרנו:







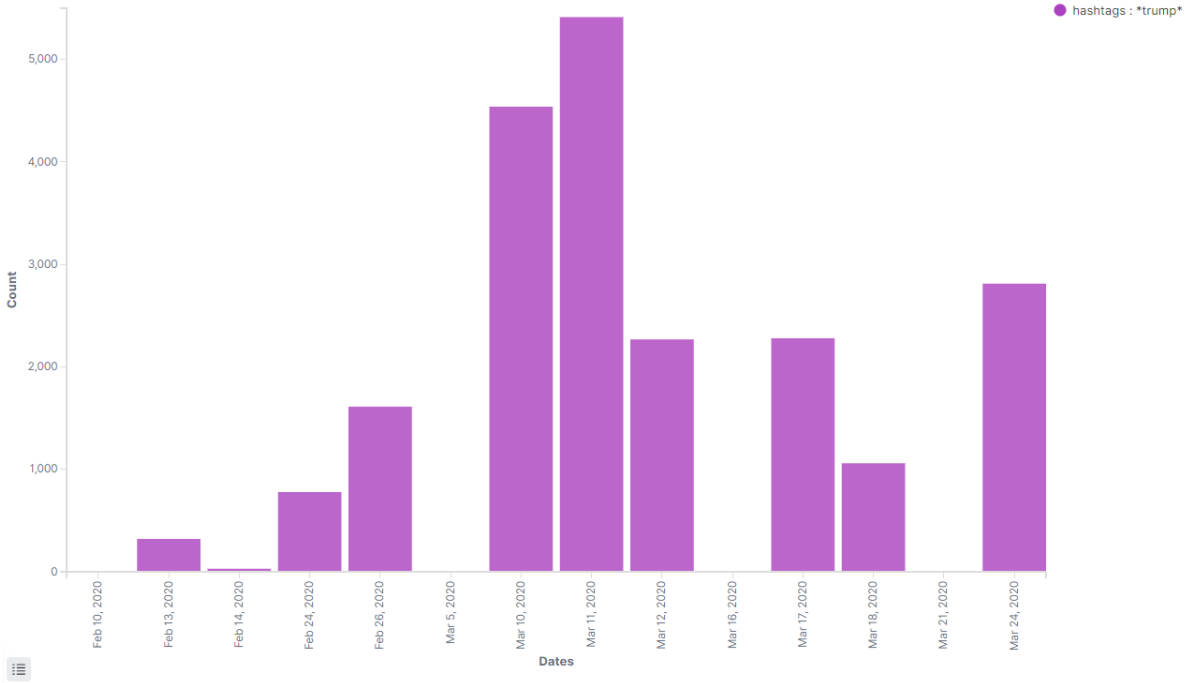
ענן מידע המציג את 20 התיוגים הנפוצים ביותר בקרב הטוויטים:



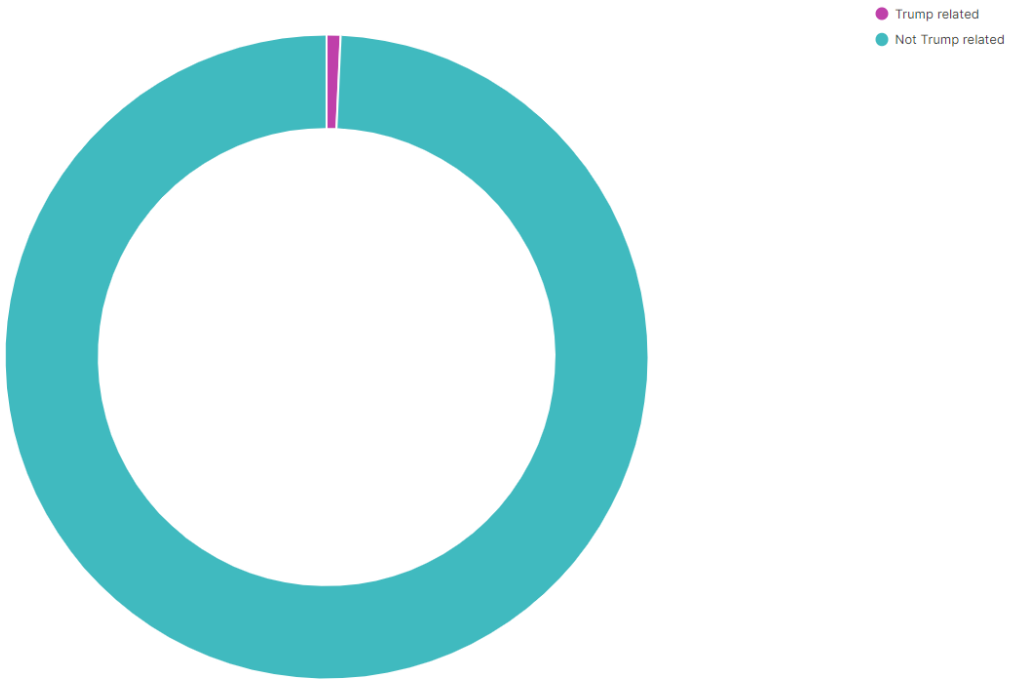
ניתן בבירור לראות שכפי שהיינו מצפים, התיוגים הנפוצים ביותר קשורים לנושא מחלת הקורונה.
ננסה לסנן את ה"רעש" שהתיוגים שקשורים לקורונה מייצרים ונקבל:

ניתן לראות שישנם מספר תיוגים המכילים את המילה Trump בקרב 20 התיוגים הנפוצים ביותר שאינם קשורים ישירות לקורונה.
כעת נזכר כי במשימת הלמידה שלנו, נרצה לחזות האם טוויטים מכילים תיוג המכיל את המילה Trump בתוכו (האם הם טוויטים המדברים על הנשיא Trump). לכן, נסתכל גם על ענן מידע המציג את התיוגים הנפוצים ביותר בקרב טוויטים שברשימת התיוגים שלהם המילה Trump נמצאה כחלק מהתיוגים.

מענן המידע הנ"ל ניתן להסיק כי בקרב הטוויטים שנוגעים אל הנשיא Trump, ישנם המון תיוגים שונים שמשתמשים בשמו וכי בראש התיוגים עומד לבד השם Trump כתיוג המוביל ביותר בקבוצה זו.
ננסה להבין כמה טוויטים אכן קשורים אל הנשיא Trump לאורך התאריכים השונים:

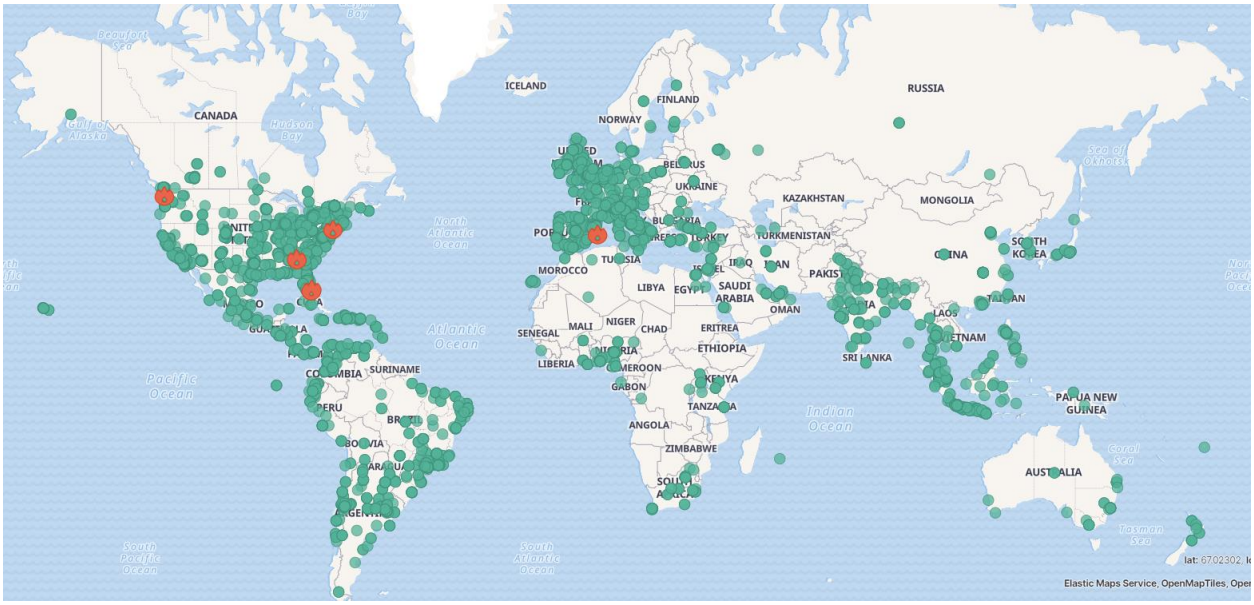


נראה כי על פי הסקאלה המוצגת בציר ה- y, מדובר בכמות טוויטים נמוכה משמעותית מכמות הטוויטים הכוללת שאספנו. ננסה להבין בדיוק עד כמה הנושא היה נפוץ בעזרת התרשים הבא:



אכן ניתן לראות שמתוך כלל התיוגים, כמות התיוגים הנוגעים אל הנשיא Trump אינה גבוהה במיוחד.
תוצאה זו מפתיעה מכיוון שהיינו מצפים שהרשתות החברתיות ובפרט טוויטר יגיבו בצורה גבוהה אל הערותיו הפומביות שהצהיר Trump בתאריכים שבחרנו.

נמשיך וננסה להבין גם את המיקומים השונים בעולם בהם היה ריכוז גבוה יותר של טוויטים הנוגעים אל Trump. בתמונה הבאה ניתן לראות מיקומים שונים של טוויטים בעולם. לשם המשימה, סימנו בסימון להבה את המקומות בהם היה ריכוז גבוה יותר של טוויטים הנוגעים אל הנשיא Trump:

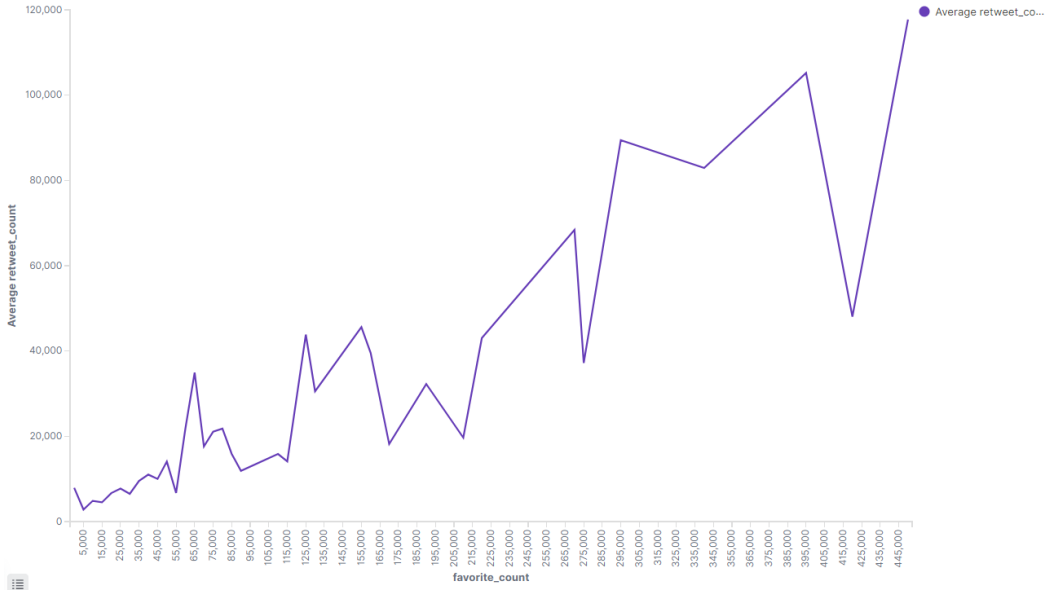


מהמפה לעיל ניתן להבין שהכמות הגדולה ביותר של הטוויטים מגיעה מאיזורי ארה"ב ומדינות מערב אירופה. בנוסף, ניתן לראות כי האיזורים ה"חמים" (pan intended), הנוגעים אל הנשיא Trump, הם בעיקר מארה"ב ומעט יותר ממערב אירופה. מסקנה זו אינה מפתיעה שכן Trump הוא נשיאה של ארה"ב.

כעת, לביצוע המשימה נרצה להבין גם את הקשר בין התכונות retweet_count, favorite_count וטוויטים הקשורים אל הנשיא Trump. בפרט נרצה לענות על השאלה:

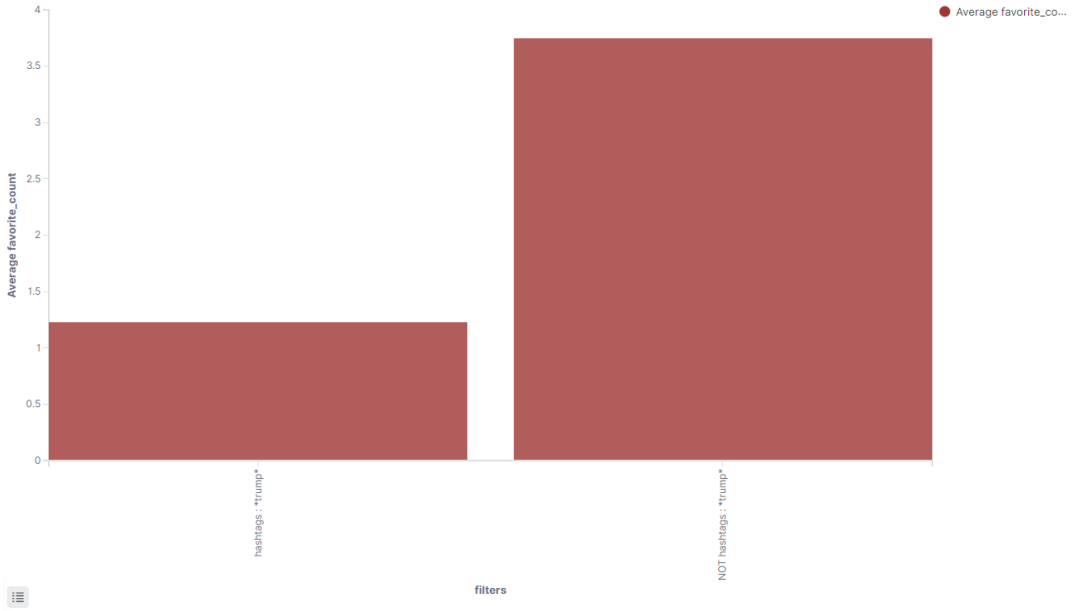
האם טוויטים הנוגעים אל הנשיא ,בימים בהם יצא בהצהרות מעוררות מחלוקת, הם טוויטים פופולריים יותר, כלומר, טוויטים שמקבלים יותר סימוני favorite או ריטוויט?

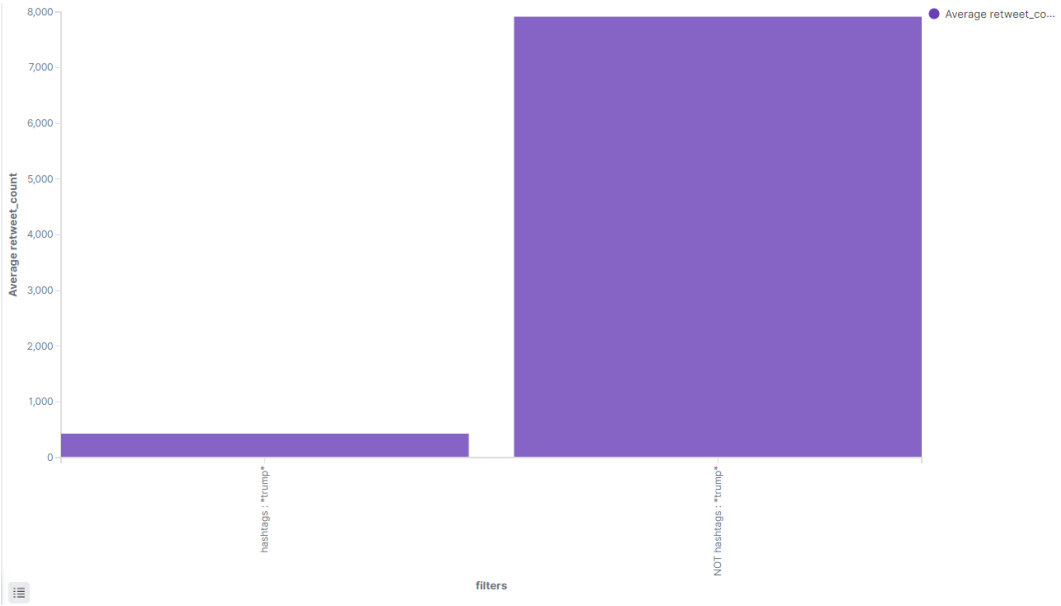
תחילה נשים לב לקשר הבא בין retweet_count ו- favorite_counte:



הגרף לעיל מציג את הקשר בין כמות סימוני ה favorite וכמות ה retweets שמקבל טוויט. ניתן לראות ממגמת העלייה כי באופן כללי, ככל שטוויט מקבל יותר סימוני favorite הוא מקבל גם יותר retweets.
מסקנה זו אינה מפתיעה שכן 2 המדדים הללו הם מדדים לפופולריות הטוויט ולכן הגיוני שקיים ביניהם יחס חיובי.

בניסיון לענות על השאלה, נבדוק האם ממוצע ה retweets וסימוני ה favorite שמקבלים טוויטים שקשורים לTrump גבוה מטוויטים שאינם קשורים אליו:





בצורה מפתיעה קיבלנו כי המסקנה היא דווקא הפוכה והטוויטים אינם פופולריים כפי שציפינו. ייתכן מאוד כי תוצאה זו נובעת מ Outliers שמושכים את ממוצע הטוויטים שאינם קשורים לTrump למעלה.
בכל אופן, כן ניתן להסיק כי יש הבדל ניכר בין טוויטים הנוגעים אל הנשיא Trump לבין טוויטים שאינם נוגעים אליו.
תחת מסקנות אלו, נעבור למשימת הלמידה ונראה האם נצליח לחזות בצורה טובה טוויטים שקשורים אל הנשיא בעזרת התכונות שבחרנו.