**פרויקט**

**מבוא לניתוח נתונים**

**מגישות:**

**סתיו ספקטור - 315818666**

**עידן שריג - 204840003**

QUESTIONS

בחרנו לעשות את הפרויקט על סט הנתונים ‘Google Play Store Apps’.

השאלות העיקריות אשר עודדו אותנו לבצע את עבודת החקר שלנו בנושא זה הינן מהם הגורמים אשר מובילים להצלחה של אפליקציה, ומה משותף לאפליקציות שהצליחו.

הצלחה של אפליקציה נמדדת לרוב לפי כמות ההכנסות שלה אשר להנחתנו תלויות בכמות ההורדות של האפליקציה ובדירוג האפליקציה (ציון Rating). כמובן שיש גורמים נוספים המשפיעים על הצלחת האפליקציה אך בחרנו להתמקד במאפיינים אלו.

שאלות אלו עניינו אותנו כיוון שחברה העוסקת בפיתוח אפליקציות תוכל להיעזר בפרויקט זה וללמוד כיצד לגרום לאפליקציות שלה לחדור אל השוק בצורה טובה, ממוקדת ומהירה יותר, ולהוביל לעלייה בהכנסות החברה. אנחנו מאמינות שבעזרת מחקר מעמיק בנושא ניתן להביא לשינוי אמיתי בקבלת ההחלטות של חברות האפליקציות ושל המפתחים.

שתי שאלות נוספות אשר עניינו אותנו הינן:

1. האם אפליקציה המשוייכת לקטגוריה של משחק בעלת פוטנציאל להצלחה גבוהה יותר לעומת שאר האפליקציות (שאינן משוייכות לקטגוריה זו). שאלה זו מעניינת כיוון שרצינו לתת מידע קונקרטי לחברה המגששת באיזו קטגוריה כדאי לה למקד את משאבי הפיתוח.
2. האם קיימת קורלציה בין ציון הרייטינג של אפליקציה לבין אחוז התגובות החיוביות אשר קיבלה. שאלה זו מעניינת כיוון שברצוננו לאפשר לחברות להסתמך על ציון הרייטינג של אפליקציה בלבד, במקום לבצע ניתוח טקסט ובדיקה האם הביקורות על האפליקציה אכן חיוביות. במידה והקורלציה ביניהם אכן גבוהה, שקלול שני המשתנים לנתון הצלחה אחד הינו מיותר וניתן להזניח את הביקורות.

DATASET

סט הנתונים "Google Play Store Apps" מכיל מידע על אפליקציות שנמצאות בחנות האפליקציות של אנדרואיד. המידע כולל 10841 שורות נתונים על אפליקציות (לאחר סינון וניקיון של סט הנתונים, כולל 9646 שורות נתונים על אפליקציות) ו13 מאפיינים:

* שם האפליקציה
* גודל האפליקציה
* מספר הורדות
* קטגוריה אליה היא משתייכת
* ז'אנר אליה משתייכת האפליקציה חוץ מהקטגוריה הראשית
* דירוג האפליקציה על ידי משתמשים
* מספר הביקורות על האפליקציה
* סוג האפליקציה- האם היא בחינם או בתשלום
* מחיר האפליקציה
* קבוצת גיל אליה האפליקציה מיועדת
* תאריך עדכון אחרון של האפליקציה
* גרסות אנדרואיד שתומכות באפליקציה

בעזרת נתונים אלו ניתן לזהות דפוסים וקורלציות בין מאפייני האפליקציה ובעזרתם לאפשר לחברות לטייב את האפליקציה אותה רוצים לפתח ולכן תורמים למענה על השאלות שלנו. לאורך פרויקט זה אנחנו מניחות שהמידע הינו מייצג (דגימה מייצגת של שוק האפליקציות העולמי).

ראשית, כפי שראינו בסט הנתונים קיים מידע על כמות ההורדות של אפליקציה ועל דירוג האפליקציה על ידי משתמשים (ציון הRating שלה). בעזרת נתונים אלו, כפי שהסברנו לעיל ניתן יהיה לחשב מדד הצלחה, בשילוב עם מספר הביקורות (סטיית תקן) שאפליקציה מקבלת, אשר יוכל לעזור לנו לענות על השאלות העיקריות ששאלנו- מהם הגורמים אשר מובילים להצלחה של אפליקציה, ומה משותף לאפליקציות שהצליחו.

שנית, כפי שראינו בסט הנתונים קיים מידע על הקטגוריה אליה אפליקציה משתייכת ובכך ניתן להסיק מסקנות עבור השאלה- האם אפליקציה המשוייכת לקטגוריה של משחק בעלת פוטנציאל להצלחה גבוהה יותר לעומת שאר האפליקציות (שאינן משוייכות לקטגוריה זו).

בנוסף, עבור השאלה- האם קיימת קורלציה בין ציון הרייטינג של אפליקציה לבין אחוז התגובות החיוביות אשר קיבלה, הוספנו סט נתונים חיצוני אשר מתקשר לסט המקורי המכיל את התגובות של המשתמשים לאפליקציות מתוך סט הנתונים המקורי שלנו. בסט נתונים זה קיים מידע על הביקורות עצמן שהאפליקציות קיבלו והאם הביקורת היתה חיובית/שלילית/ניטרלית.

ANALYSIS & FINDINGS

ראשית, ביצענו תהליך מעמיק של ניקוי סט הנתונים.

גילינו שיש מספר שורות כפולות, כלומר מספר שורות שמתארות את אותה האפליקציה (ביצענו חיפוש לפי שם) אך עם כמות ביקורות שונה. הנחנו שמדובר בדגימות שקרו בזמנים שונים על אותה האפליקציה.

עמדו בפנינו שתי אופציות:

1. מחיקת כל השורות הכפולות:

יתרון - הדאטה נקי ואין סיכוי להטיית מדידה (הסיכוי למחוק את השורה הלא נכונה).

חסרון – אובדן של 8% מהמידע. מחיקה של בערך 1000 אפליקציות מסט הנתונים, דבר שישפיע על המסווג שלנו בהמשך. במקרה שהמידע המוכפל הוא לא רנדומלי אנחנו יכולות בטעות לגרום לדאטה להיות לא מייצג.

1. בכל אפליקציה מוכפלת – מחיקת השורות הכפולות פרט לשורה בעלת מספר הביקורות הגבוה ביותר:

יתרון – לא מאבדים 8% מהאפליקציות, שומר על המידע שלנו מייצג כי כל אפליקציה יכולה להשפיע על המסווג שלנו במחקר.

חסרון – למחוק בטעות את השורה העדכנית ביותר (ההנחה לגבי מספר הביקורות אינה נכונה וקיימת טעות אקראית במדידה) ולגרום להטיית מדידה.

בחרנו באופציה מספר 2. אנחנו מאמינות שהיתרונות גוברים על החסרונות מפני שאם בחרנו באופציה זו ואכן מחקנו את השורות הלא נכונות אז הטיית המדידה ככל הנראה תהיה אקראית ולא מגמתית. ואילו אם נמחק את כל האפליקציות אשר מופיעות פעמיים נאבד כמות גדולה של מידע מה שעלול ליצור הטיית בחירה ולהפוך את כל המחקר להיות לא רלוונטי כיוון שהנתונים לא מייצגים.

שנית, מצאנו שורה שכל ערכיה הוזזו שמאלה, שורה אחת מתוך 9660 שורות נוספות הינה זניחה ולכן החלטנו למחוק אותה.

Calendar

Description automatically generated with medium confidence

בנוסף, גילינו כי היו 1474 שורות בעלות ערכי NA.

בעמודת 'Rating' היו 1463 ערכי NA אשר מהווים מעל 15% מהמידע לכן מחיקת שורות אלו ישפיעו בהכרח על המחקר שלנו ויגרמו למידע להיות לא מייצג, לכן על מנת לגרום להשפעה מינימלית על הסיווג שלנו החלפנו ערכים אלו בערך החציון.

בשאר העמודות כמות הNAים הייתה זניחה לכן מחקנו שורות אלה.

Table

Description automatically generated

בעמודת 'Content Rating' גילינו שהיו 2 שורות בהן לא נכתב למי מיועדת האפליקציה (Unrated). 2 הינו מספר זניח לעומת כמות הדאטה ולכן החלטו למחוק שורות אלה.

A picture containing text

Description automatically generated

לאחר מכן בדקנו כמה NAים נשארו לנו בסט הנתונים. גילינו שיש 11 ומפני שזה זניח ביחס לכמות השורות שיש לנו בסט הנתונים החלטנו למחוק שורות אלו.

בהמשך, סידרנו את המידע בסט הנתונים:

המאפיינים הקטגוריאליים: Installs, Size, Reviews ו Priceהינם מקוטלגים כקטגוריאליים אך בפועל מייצגים משתנים נומריים. בנוסף Last Updated מקוטלג כקטגוריאלי גם כן, אך מייצג תאריך.

לכן ערכנו את העמודות כך שכל ערך ייוצג בהתאם בזיכרון באופן הבא:

* **Installs** – מחיקת הסימנים '**+**' , '**,**' והמרה למספר.
* **Size** החלפנו את הערך 'Varies with device' בערך החציון (16). הערך 'Varies with device' מופיע ב1226 שורות. אחוז 1226 השורות מתוך 9646 הינו מעל 12% לכן מחיקת שורות אלו תגרום להשפעה מהותית על המידע ועלולה להפוך את המידע ללא מייצגת לעומת זאת החלפת השורות בערך החציון הינה ניטרלית יותר מבחינת תהליך הסיווג. נוסף על כך המרנו את כלל הערכים ליחידות מידה מ - Kbytes ל Mbytesעל ידי מחיקת האות 'k' וחלוקה ב1024. מחקנו את הסימונים K,M ולאחר כלל השינויים העמודה מייצגת גודל אפליקציה בMbytes.
* **Price** מחקנו את הסימן '$' והמרנו לנומרי.
* **Last Updated** המרנו לתאריך, לאחר מכן לצרכי הסיווג המרנו את העמודה למילי שניות (מספר).
* **Category** החלפנו את הסימונים '\_' ברווח, והמרנו לlowercase על מנת שיהיה ניתן לבצע השוואות עם עמודת Genres.
* **Genres** המרנו את העמודה למספר תתי הקטגוריות אליהן משתייכת האפליקציה. החלטנו לבצע שינויים כיוון שכל אפליקציה יכולה להשתייך למספר ז'אנרים נוסף על הקטגוריה הראשית שלה. גילינו שרוב האפליקציות שייכות לקטגוריה אחת ורק מספר מצומצם של אפליקציות שייכות לעוד קטגוריות חוץ מהקטגוריות הראשיות שלהן. בחרנו בשיטה זו כי לרוב האפליקציות אין קטגוריות משניות לכן לדעתנו יש משמעות רבה יותר לכמות הקטגוריות אליה משתייכת כל אפליקציה מאשר השמות של הקטגוריות המשניות. כהמשך המחקר, ניתן לבצע בדיקה על סט הנתונים המצומצם ולבדוק האם קיימת משמעות לתת הקטגוריה.
* לאחר השינויים סט הנתונים שלנו הכיל:
* 9646 שורות ו14 מאפיינים, כאשר העמודות הן:
* App – קטגוריאלי
* Category – קטגוריאלי
* Rating - נומרי
* Reviews - נומרי
* Size - נומרי
* Installs – נומרי
* Type – קטגוריאלי
* Price – נומרי
* Content Rating – קטגוריאלי
* Genres – קטגוריאלי
* Last Updated – נומרי (תאריך)
* Current Ver – קטגוריאלי
* Android Ver – קטגוריאלי
* Number of Categories – נומרי

**ניתוח נתונים אקספלורטרי:**

**היסטוגרמות:**

* **התפלגות קטגוריות:**

Chart, bar chart

Description automatically generated

ניתן לראות מהגרף כי הקטגוריה השכיחה ביותר הינה ‘family’ ואחריה הקטגוריה ‘game’.

* **התפלגות סוג (חינם/בתשלום):**

Chart, bar chart

Description automatically generated

ניתן לראות כי מספר האפליקציות החינמיות גדול משמעותית ממספר האפליקציות בתשלום.

* **התפלגות קבוצת הגיל אליה האפליקציה מיועדת:**

Chart

Description automatically generated

ניתן לראות כי קבוצת הגיל השכיחה ביותר הינה Everyone.

* **התפלגות ציוני הרייטינג:**

Chart, histogram

Description automatically generated

ניתן לראות כי מרבית ציוני הרייטינג נמצאים בטווח בין 4 ל 5.

* **התפלגות גודל האפליקציות (בMB):**

Chart, histogram

Description automatically generated

ניתן לראות כי מרבית האפליקציות שוקלות בין 0 ל- 20MB. צפינו שיופיעו סביב 16MB כמות גדולה של אפליקציות כיוון ששינינו את 'Varies with device' להיות ערך החציון – 16.

* **התפלגות מספר הקטגוריות לאפליקציה:**

Chart, bar chart

Description automatically generated

ניתן לראות כי מרבית האפליקציות משויכות לקטגוריה אחת בלבד.

* **ציון הרייטינג כנגד מספר הביקורות:**

Chart

Description automatically generated

לא קיימת קורלציה בין מספר הביקורות וציון הרייטינג לאפליקציה. עם זאת ניתן לראות כי בסביבת ציון הרייטינג 4.5 קיימת עליה במספר הביקורות.

* **קטגוריה כנגד מספר ההורדות:**

Chart

Description automatically generated

האפליקציות שלהן מספר ההורדות הממוצע הגבוה ביותר הינן מקטגוריה "Communication".

* **סוג (חינם/בתשלום) כנגד מספר ההורדות:**

Chart

Description automatically generated

ניתן לראות בבירור כי לאפליקציות בחינם יש כמות הורדות גבוהה בהרבה מלאפליקציות בתשלום.

ניתן להסיק כי יש קורלציה חיובית בין סוג – חינם לבין המשתנה ‘Installs’.

* **קטגוריה כנגד ציון הרייטינג:**

Chart, bar chart, histogram

Description automatically generated

ניתן לראות כי לכל הקטגוריות יש ציון גבוה מ-4. אין קורלציה בין שני המשתנים.

**Estimation and hypothesis testing:**

חישבנו מדד הצלחה על ידי העמודות

**בונוס:**

בחלק זה, מצאנו סט נתונים נוסף המתקשר לסט הנתונים המקורי שלנו. סט נתונים זה מכיל מידע על ביקורות המשתמשים, כולל הסמנטיקה של הביקורות (Positive/Negative/Neutral).

החלטנו לחקור לעומק ולבדוק האם ישנה קורלציה בין הסמנטיקה של הביקורת לבין ציון הרייטינג של האפליקציה.

על מנת לעשות זאת, לכל אפליקציה חישבנו את כמות התגובות החיוביות שלה ואיחדנו בין שני הdatasets. לאחר מכן, חישבנו את אחוז הביקורות החיוביות מתוך כלל הביקורות והראנו את התוצאה בגרף.

Chart, scatter chart

Description automatically generated

כפי שניתן לראות, יש קורלציה חיובית מעטה בין שני המשתנים –Positive Percent ו Rating Score.

LIMITATIONS

For example, notice that in row 9656 the main category is 'family' while the genre is 'Education', which means this app belong to 2 categories, therefore in the 'Number Of Categories' column the value is 2. <br>

Another example is in row 9658. This app's main category is 'books and reference' and the genre is 'Books & Reference', which is the same category, therefore in the 'Number Of Categories' column the value is 1.

FUTURE DIRECTIONS

* אין לנו מידע על זמן שהייה של אפליקציה אצל הלקוח. לאחר שהוריד יכול להיות שהוא מחק את האפליקציה יום למחרת ואז זה משפיע על טיב האפליקציה
* אין לנו מידע על אם יש באפליקציה פרסומות או דברים שמפריעים
* אין לנו מידע על איך האפליקציה היתה לפני כמה שנים מבחינת כמות הורדות ולגבי הרייטינג נגיד ואז להסיק מסקנות לגבי שיפור או שימור (מגמות)