

## תרגיל בית 1 - Yelp Dataset

1. לקראת קבלת החלטות לגבי עתיד החברה בישיבת הדירקטוריון הקרובה, נוסחו **שלוש שאלות עסקיות** לניתוח נתונים על אודות אופי פעילות החברה וטרנד שימוש בשירותיה.

**שאלה עסקית 1 - מי הם מבקרי ה"עלית" וכיצד הם נחלקים לרבעונים?**

על ידי פילוח מבקרי ה"עלית" בשנה האחרונה (העדכנית נכון ל-data הקיים), נוכל להבין מי הם מבקרי העלית המובילים שמומלץ להשקיע בהם ולקדם אותם, ומי מבקרי ה"עלית" ה"זוטרים" שאותם יש לחזק באמצעות הדרכה ו/או לבחון את התאמתם כמבקרי עלית. הפילוח יבוצע על בסיס שני פרמטרים עיקריים:

1. דירוג מבקר - כל מבקר מקבל ציון בין 0 ל-1 על בסיס מספר הביקורות שכתב, מספר התגובות והמחמאות שקיבל, וכן מספר המעריצים לפי החישוב המנורמל הבא:

a. 30% לפי מספר ההצבעות החיוביות (המבקר בעל מספר ההצבעות המרבי מקבל 0.3).

b. 20% לפי מספר הביקורות (המבקר בעל מספר הביקורות המרבי יקבל 0.2).

c. 30% לפי מספר המעריצים (המבקר בעל מספר המעריצים המרבי מקבל 0.3).

d. 20% לפי מספר המחמאות (המבקר בעל מספר המחמאות המרבי מקבל 0.2).

2. מספר בתי העסק עליהם המבקר כתב ביקורות בתקופת הזמן הנבחרת.

**שאלה עסקית 2 - באילו מדינות (States) נפח הפעילות גבוה, ובאילו מדינות נפח הפעילות דווקא נמוך, בשנת 2017?**

בעזרת שאלה זו נמקד את המאמץ השיווקי של Yelp. במדינות בהן קיים נפח פעילות גבוה, נרצה לפעול לשימור המצב הקיים, ובמדינות בהן נפח הפעילות נמוך נרצה לבחון קידום מאמץ שיווקי ייעודי עבור שיפור המצב הקיים.

בחינת נפח הפעילות בשנת 2017, יבוצע באופן באופן הבא:

- ראשית, נבצע את הבדיקה על אודות בתי עסק פעילים בתקופת הזמן הנבחרת.
- נפח הפעילות במדינה מסוימת יחושב כיחס הבא -

מספר בתי העסק במדינה מסוימת שבוצעה ביקורת לגביהם בתק' הזמן הנבחרת (שנת 2017)

מספר בתי העסק הקטנים הקיימים במדינה המסוימת בתקופת הזמן הנבחרת (שנת 2017)

ככל שיחס זה גבוה יותר, נוכל לומר כי נפח הפעילות באותה מדינה גבוה יותר, וזאת כי קיים אחוז גבוה יותר של בתי עסק מבוקרים במדינה.

הערה: הנתון המצוין במכנה, אינו שייך לנתוני חברת Yelp, אלא מיובא מאתר INSURANCE INFORMATION INSTITUTE.<sup>2</sup>

**שאלה עסקית 3 - מהי הקורלציה בין אורך הביקורת לבין מספר הצבעות שקיבלה, עבור ביקורות שנכתבו על בתי עסק המדורגים 4 כוכבים ומעלה, בשנת 2017?**

נרצה להגדיר את אורך הביקורת המומלץ עבור שתי מטרות עיקריות:

1. העלאת אפקטיביות הביקורות עבור המבקרים ומשתמשי המערכת. בעידן הנוכחי אנו נחשפים לכמויות מידע עצומות בכל יום, שעה ודקה, ומשתמשים נוטים להשקיע זמן קצוב למידע שצורכים, אם בכלל. נרצה להמליץ על אורך הביקורת אפקטיבי, כך שנגדיל את הסיכוי שהביקורת תיקרא ותבוקר על ידי משתמשים אחרים.





2. היבט כלכלי - חיסכון באחסון. במידה שנגלה כי ביקורות ארוכות "מדי" (מעל כמות תווים מסוימת), אינן מבוקרות מספיק, נמליץ על הגבלת מספר התווים האפשרי עבור חיסכון באחסון.

אפקטיביות הביקורת תיבחן על ידי סך מספר ההצבעות שקיבלה (useful, funny, cool), בשנת 2017.

<sup>1</sup> ממצוי הנתונים עלה כי הביקורות העדכניות ביותר הינן מ-2017, ולכן נתייחס לשנה האחרונה האפשרית.  
<sup>2</sup> להרחבה, (1) [Small Businesses By State, 2017](#).

## 2. אפיון מחסן הנתונים בסכמת כוכב

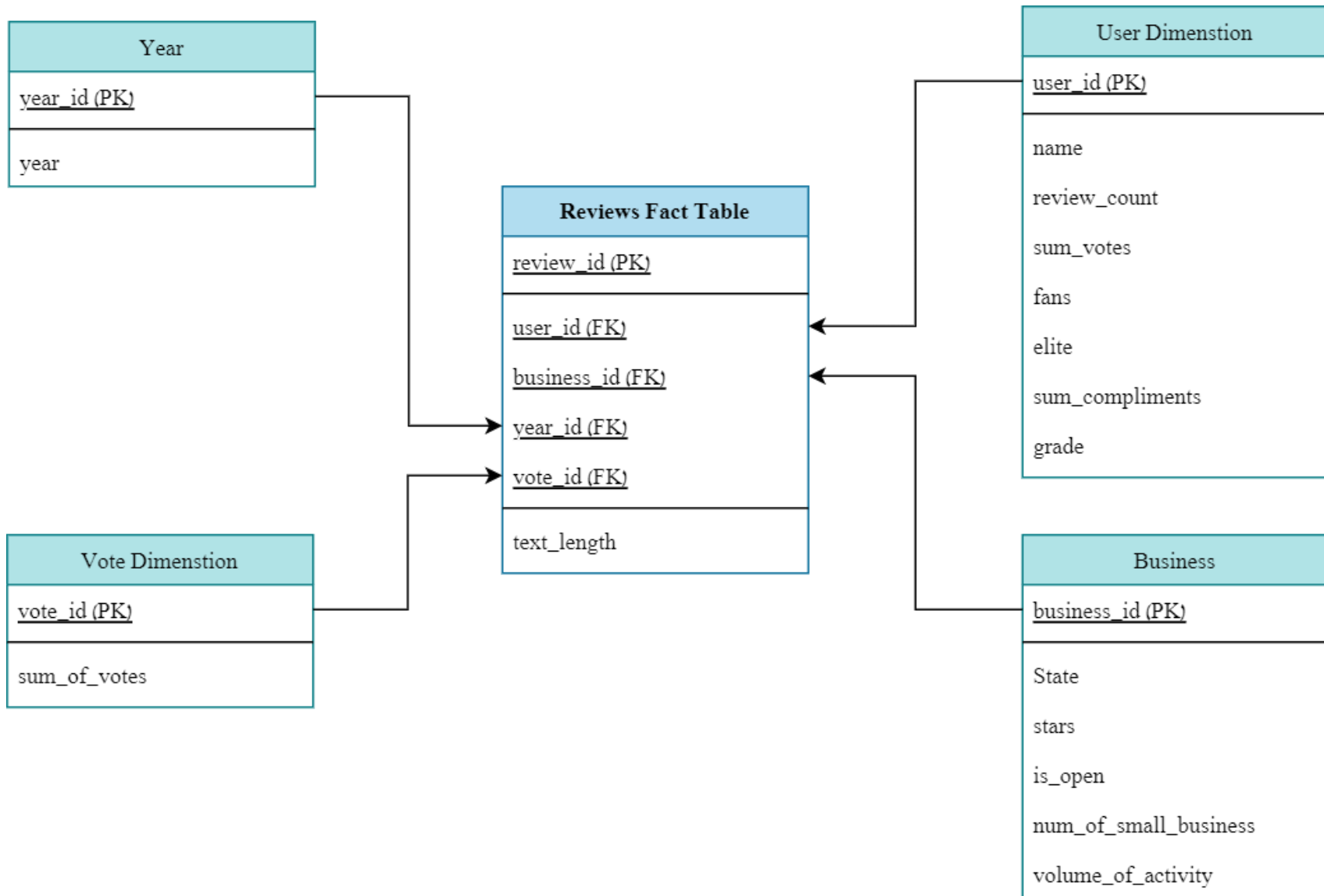
- a. זיהוי התהליך בו מחסן הנתונים מתמקד: מחסן הנתונים מתמקד בביקורות הנכתבות באתר "Yelp".
- b. בחירת הגרעין (grain): נפעל בהתאם ל-transaction design, כאשר נבחר גרעין של ביקורת. לפיכך, כל שורה בטבלת העובדות תייצג ביקורת אחת מסוימת.
- c. בחירת ממדי מחסן נתונים: לפי ה-bus chart שלהלן, המימדים הקיימים במחסן הם:

	Business	User	Year	Vote
Reviews				

כאשר:

- **מימד בית עסק:** כולל את כל המידע הרלוונטי על בתי העסק, בתוספת שני שדות ייעודיים - נתונים חיצוניים על אודות בתי עסק קטנים בכל מדינה, וכן נפח הפעילות בכל מדינה.
  - **מימד המשתמש:** כולל את כל המידע הרלוונטי על המשתמשים, בתוספת שלושה שדות ייעודיים - סכימה של מספר הצבעות חיוביות למשתמש, סכימה של מספר המחמאות למשתמש, ציון משתמש.
  - **מימד הזמן - שנה:** כולל את השנים הרלוונטיות.
  - **מימד ההצבעות:** כולל את סך ההצבעות לכל ביקורת.
- d. זיהוי העובדות:
- אורך הביקורת
  - מספר בתי העסק עליהם המבקר כתב ביקורות בתקופת הזמן הנבחרת.
  - מספר בתי העסק במדינה מסוימת שבוצעה ביקורת לגביהם בתקופת הזמן הנבחרת

סכמת הכוכב המלאה למחסן הנתונים מצורפת בעמוד הבא.



### 3. תרשים קונספטואלי של מחסן נתונים

טבלת העובדות - ביקורות:

שדה	טיפוס נתונים	מקור נתונים	הערות
<u>review_id</u>	int	שדה auto increment	מפתח ראשי, ייחודי
text_length	int	מחושב ע"י אורך הביקורת ממאגר הנתונים הקיים	אורך ביקורת שנכתבה ע"י משתמש מסוים, על בית עסק מסוים, בתאריך מסוים
user_id	int	מפתח זר	מפתח זר למימד משתמש
business_id	int	מפתח זר	מפתח זר למימד בית עסק
year_id	int	מפתח זר	מפתח זר למימד שנה
vote_id	int	מפתח זר	מפתח זר למימד הצבעות

טבלת מימד - בית עסק:

שדה	טיפוס נתונים	מקור נתונים	הערות
<u>business_id</u>	int	שדה auto increment	מפתח ראשי, ייחודי
state	varchar	מיובא מהמאגר הנתונים הקיים	מדינת בית העסק
stars	float	מיובא מהמאגר הנתונים הקיים	דירוג בית העסק
is_open	tinyint	מיובא מהמאגר הנתונים הקיים	האם בית עסק פעיל
num_of_small_business	int	מיובא ממאגר נתונים חיצוני מאתר INSURANCE INFORMATION INSTITUTE. <sup>3</sup>	מספר בתי העסק הקטנים הקיימים במדינה המסוימת בתקופת הזמן הנבחרת (שנת 2017)
volume_of_activity	float	מחושב על ידי חלוקה של מספר בתי העסק במדינה שבוצעה עליהם ביקורת ב-2017, במספר בתי העסק הקיימים במדינה בשנת 2017	נפח פעילות של כל מדינה בשנה האחרונה (שנת 2017)

<sup>3</sup> להרחבה, (1) [Small Businesses By State, 2017](#).

טבלת מימד - משתמש:

שדה	טיפוס נתונים	מקור נתונים	הערות
<u>user_id</u>	int	שדה auto increment	מפתח ראשי, ייחודי
name	date	ייבוא ממאגר הנתונים הקיים	תאריך
review_count	int	מחושב ע"י ספירת המשתמשים הייחודיים הקיימים ב-data.	מהווה תמונת מצב לגבי מס' המשתמשים הרשומים ל-Yelp עד אותו תאריך ספציפי.
sum_votes	int	סכימת כל ההצבעות שהמשתמש קיבל עבור כל הביקורות	סך הצבעות - רלוונטי ל-grade
fans	int	ייבוא ממאגר נתונים הקיים	מעריצים - רלוונטי ל-grade
elite	mediumtext	ייבוא ממאגר נתונים הקיים	באילו שנים המשתמש היה עלית
sum_compliments	int	סכימת כל המחמאות שהמשתמש קיבל	סך המחמאות - רלוונטי ל-grade
grade	float	הציון שאנחנו מודדים	ציון משתמש

טבלת מימד - שנים:

שדה	טיפוס נתונים	מקור נתונים	הערות
<u>year_id</u>	int	שדה auto increment	מפתח ראשי, ייחודי
year	int	מיוצר באופן אוטומטי. מתחיל משנה MIN ומסתיים MAX מקובץ time עמודה date	שנה מהתאריכי ביקורות

טבלת מימד - הצבעות:

שדה	טיפוס נתונים	מקור נתונים	הערות
<u>vote_id</u>	int	שדה auto increment	מפתח ראשי, ייחודי
sum_votes	int	סכימת כל ההצבעות בעבור כל ביקורת	סך ההצבעות החיוביות

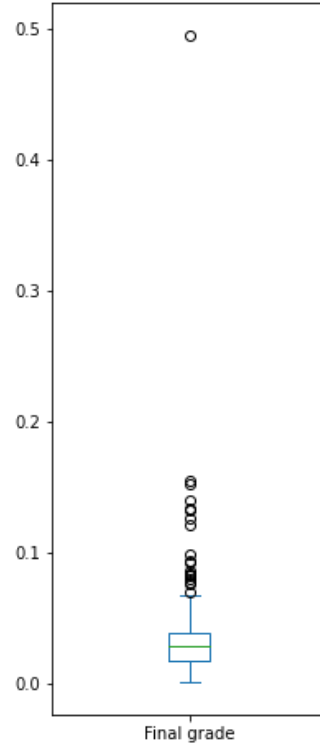
4. בניית מחסן נתונים ב-MYSQL - בוצעה באמצעות קובץ ipynb העונה לשם: team\_8\_notebook.ipynb.

## 5. מענה על השאלות העסקיות באמצעות SQL

i. השאלה העסקית הראשונה עוסקת במדידה של מבקרי ה"עלית" וחלוקה שלהם לרבעונים. שני הנתונים העיקריים עבור המדידה - ציונו של המשתמש, ומספר בתי העסק בהם הוא ביצע ביקורת במהלך שנת 2017. הציון הסופי חושב כממוצע משוקלל של נתונים אלו, כאשר ציון המשתמש הקיים בעל משקל של 90%, ו-10% נוספים עבור מספר בתי העסק המנומל שביקר. לאחר ביצוע שליפה מתאימה (מצורפת במחברת ipynb), התקבלה טבלת התוצאות הבאה, בעלת 226 ערכים, המסודרים לפי "Final grade" בסדר יורד:

User id	Name	Final grade
247552	Stephanie	0.4949566662
321177	Jennifer	0.1546981901
1074033	Breanna	0.1518198982
405206	Evelyn	0.1395302713
737989	Edwin	0.133781907
...		
432126	Ryan	0.001475000067
10325	Charles	0.001443321921
358530	Sar	0.001416646829
372408	Steve	0.001398686401
184301	Jean-Philippe	0.0006752816902

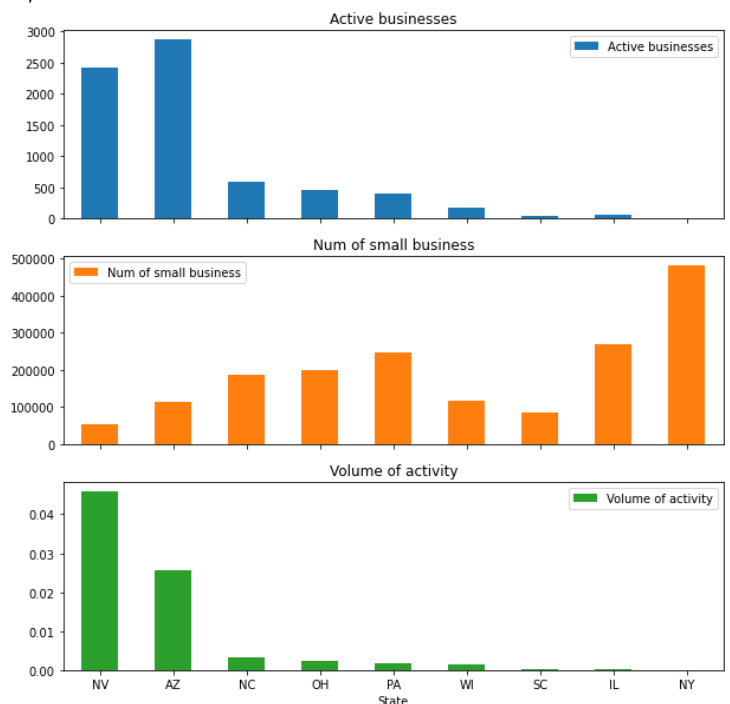
Distribution of Final Grades Amongst Elite Users of 2017



בעזרת שימוש ב-Pandas, הוצא תרשים BoxPlot המציג את התפלגות הניקוד והחריגים. נשים לב כי קיים מספר קטן של מבקרים אשר קיבלו ציון גבוה באופן משמעותי מרוב המבקרים, וקיים מבקר אחד בעל ציון חריג במיוחד, הגדול פי שלושה מהבא אחריו. לכאורה, נרצה לזקק מסקנות כלשהן מתוך התוצאות, אך בפועל אין זה מעשי בהתחשב הנסיבות, בהן הנתונים שקיבלנו במקור הינם קטומים. לפיכך אין לייחס לתוצאות אלו חשיבות רבה, הרי שהן בעלות שגיאות והטיות.

ii. השאלה העסקית השנייה עוסקת בנפח הפעילות במדינות השונות בארה"ב במהלך שנת 2017. השאלתא בוצעה רק עבור בתי עסק פעילים ב-2017, אשר בוצעו עליהם ביקורות. נפח הפעילות חושב בהתאם להסבר שצוין לעיל, ולהלן התוצאות שהתקבלו:

State	Active businesses	Num of small business	Volume of activity
NV	2,420	53,908	0.0460043
AZ	2,872	114,076	0.0256583
NC	584	187,749	0.0031638
OH	448	198,301	0.00231466
PA	399	245,259	0.00166762
WI	169	117,354	0.00148269
SC	33	86,695	0.000392179
IL	55	268,674	0.000208431
NY	1	481,792	0.000002075



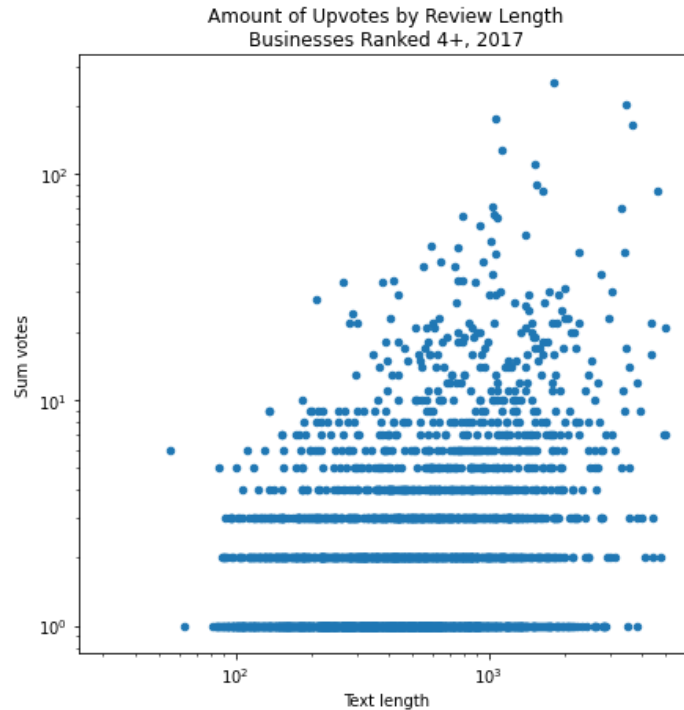
על פי התוצאות המוצגות מעלה, לכאורה ניתן להסיק כי במדינת ניו יורק (NY) מספר בתי העסק הכולל הוא הגדול ביותר, אך מספר בתי העסק בהן בוצעה ביקורת הוא הקטן ביותר, ולכן נתח השוק שם הוא כמעט "אפסי". מסקנה נוספת היא המלצה לביצוע קמפיינים המציעים לבעלי עסקים בנבאדה (NV) לפרסם את העסק שלהם ב-Yelp, וזאת כי שם קיים נתח השוק "הגדול ביותר".

על אף התוצאות, יש להביא בחשבון את העובדה כי הנתונים שגויים ואינם מייצגים כראוי את המציאות. סיבה ראשונה לכך היא קטימת הנתונים, ולצד זאת יצוין כי בנתונים של Yelp קיימות הגדרות רבות למדינות אשר אינן מזהות עם ראשי התיבות של המדינות השונות בארה"ב, ולפיכך אין אנו יכולים לשייך אותם בצורה נכונה. לפיכך - בפועל, לא ניתן לענות על השאלה העסקית על סמך נתונים אלה.

iii. השאלה העסקית השלישית עוסקת בקורלציה בין אורך הביקורות לבין מספר ההצבעות שהביקורת קיבלה. הבדיקה נעשתה אך ורק על ביקורות שבוצעו במהלך 2017, עבור בתי עסק המדורגים 4 כוכבים ומעלה. על בסיס פלט הרצת השאילתא המתאימה, נבנו שני תוצרים: האחד - מטריצת קורלציה (בעזרת שימוש ב-Pandas), והשני - תרשים Scatter תואם של התוצאות, אשר ציריו מוצגים בסקאלה לוגריתמית.

Text length	Sum votes
1	0.2687662191
0.2687662191	1

Text Length	Sum Votes
4970	7
4949	21
4948	0
4892	7
...	...
54	0
53	0
31	0



ניתן לראות כי קיימת קורלציה חלשה (0.2687) בין מספר התווים לבין כמות ההצבעות. לפי צורת הפיזור של התוצאות בגרף, מצד אחד ניתן לומר כי ביקורת יכולה להיות ארוכה מאוד ולא לקבל הצבעות רבות, ומצד שני, הביקורות בעלות מספר ההצבעות הגבוה ביותר הן הארוכות ביותר. לפיכך נוכל להסיק כי ביקורת ארוכה היא תנאי הכרחי אך לא מספיק כדי לקבל מספר הצבעות גבוה.

עם זאת, נסייג את המסקנה הקודמת, מכיוון שהניתוח הינו פשטני למדי, מאחר שלא כולל התייחסות למימדים נוספים כדוגמת מספר המעריצים והחברים שיש למבקר, אשר להערכתנו יכולים להשפיע על כמות ההצבעות.

מעבר לכך, יצוין שוב כי לא ניתן לענות ברמת מהימנות גבוהה על השאלה העסקית על בסיס הנתונים הקיימים, וזאת לאור קטימת הנתונים אשר פוגעת באיכות שאלה זו גם כן.

6. נגדיר את ה-KPIs הבאים:

i. **KPI פעילות מבקרי "עלית"** (נקודת מבט של צמיחה) - לפי ראות עיננו, היעד התפעולי הוא כי מבקר "עלית" יבצע ביקורת בממוצע אחת לשבועיים לפחות, ולפיכך היעד השנתי למבקר "עלית" הוא 26 ביקורות בשנה. מתוך מחשבה כי יכולים להיות שבועות בהם הבמקר לא יעמוד ביעד, ומתוך רצון למצוא KPI "יפה" יותר, נגדיר את היעד השנתי כ-25 ביקורות בשנה. **ה-KPI הוא אחוז המבקרים העומדים ביעד.** כלומר, אם 60% מהמבקרים יעמדו ביעד השנה, אז המדד שיוצג הינו 60/100. כך הארגון ידע לתמרץ את המבקרים הפעילים, או לחילופין להדריך ולמשב את מבקרי העלית הפחות פעילים. מדד זה חיוני מאוד לאסטרטגיה הארגונית של Yelp, כי כך נפח הפעילות באתר יהיה גדול יותר.

ii. **KPI נתח שוק** (נקודת מבט פיננסית) - בכל שנה נרצה לבדוק מהו נתח השוק הכולל של חברת Yelp בכל מדינות ארה"ב, וזאת כדי להבין את דריסת הרגל של החברה בתחום. על מנת לעשות זאת, נמדוד את מספר המדינות בהן נפח הפעילות גבוה מ-20%. כלומר, בשנה האחרונה מבקרים באפליקציה ביקרו יותר מ-20% מבתי העסק במדינה מסוימת. את הבדיקה הזו נבצע על כל אחת מ-50 המדינות בארה"ב, כך שבסופו של דבר ה-KPI הינו **מספר המדינות מתוך ה-50 בהן אנו עומדים ביעד.** לדוגמה, אם ב-15 מדינות קיים נפח פעילות הגבוה מ-20%, המדד יציג 15/50 (נשים לב כי בהכרח התוצאה של מדד זה לפי הנתונים הקיימים יהיה 0/50).

7. קוד ליצירת VIEWS לחישוב ה-KPIS מהסעיף הקודם:

i. Elite Engagement KPI:

```
CREATE VIEW elite_engagement AS (
    SELECT reviews_per_user.user_id, elite, businesses_reviewed_in_2017
    FROM reviews_per_user JOIN user_dim on user_dim.user_id =
reviews_per_user.user_id
    WHERE businesses_reviewed_in_2017 >= 25);

CREATE VIEW count_elite_engagement AS(
    SELECT COUNT(*) AS count_engage
    FROM elite_engagement);

CREATE VIEW count_elite AS(
    SELECT COUNT(*) AS count_elite, elite
    FROM user_dim
    WHERE elite LIKE '%2017%');

CREATE VIEW kpi_1 AS (
    SELECT count_engage/count_elite
    FROM count_elite_engagement, count_elite);
```

פלט השאילתות:

```
count_engage/count_elite
0
```



ii. Market Share KPI:

```
CREATE VIEW market_share_kpi AS
SELECT
  COUNT(state) / 50 AS Market_Share_KPI
FROM
  yelp_dw.business_dim
WHERE
  volume_of_activity BETWEEN 0.2 AND 1
GROUP BY state;
```

פלט השאילתות:

```
Market_Share_KPI
0
```

מלבד השימוש ב-VIEWS עבור חישוב ה-KPIs, ניתן לראות שימוש מהותי בשאילתות view גם במסגרת המענה על השאלות העסקיות.