הפקולטה להנדסה ע״ש אייבי ואלדר פליישמן

המחלקה להנדסת תעשיה וניהול

# פרויקט Airbnb

## חלק א'- תכנון ו-ETL :

### עיצוב מחסן הנתונים בסכמת כוכב:

### שלבי העיצוב:

**זיהוי התהליך :**

התהליך בן מחסן הנתונים אמור לתמוך הוא ארגון המידע לגבי גורמים אפשריים לשונות בתפוסה של נכסים באתר Airbnb. ארגון הנתונים נועד לניתוח והסקת מסקנות עסקיות.

**בחירת הגרעין:**

בסיס הנתונים הקיים מכיל מידע שימושי רב היכול לשמש לצורך התהליך. הנתונים שנלקחו משם ממנו משקפים snapshot של חלר מבסיס הנתונים השלם, המשקף את ההזמנות לשנה קלנדרית החל מחודש דצמבר 2018 ועד דצמבר 2019. מתוך נתונים אלה נבחר נתונים לגבי הנכסים, דירוג שלהם על ידי לקוחות, דירוג המארחים, אופי הנכס, והזמנות שנעשו על אותם נכסים בפרק הזמן המדובר.

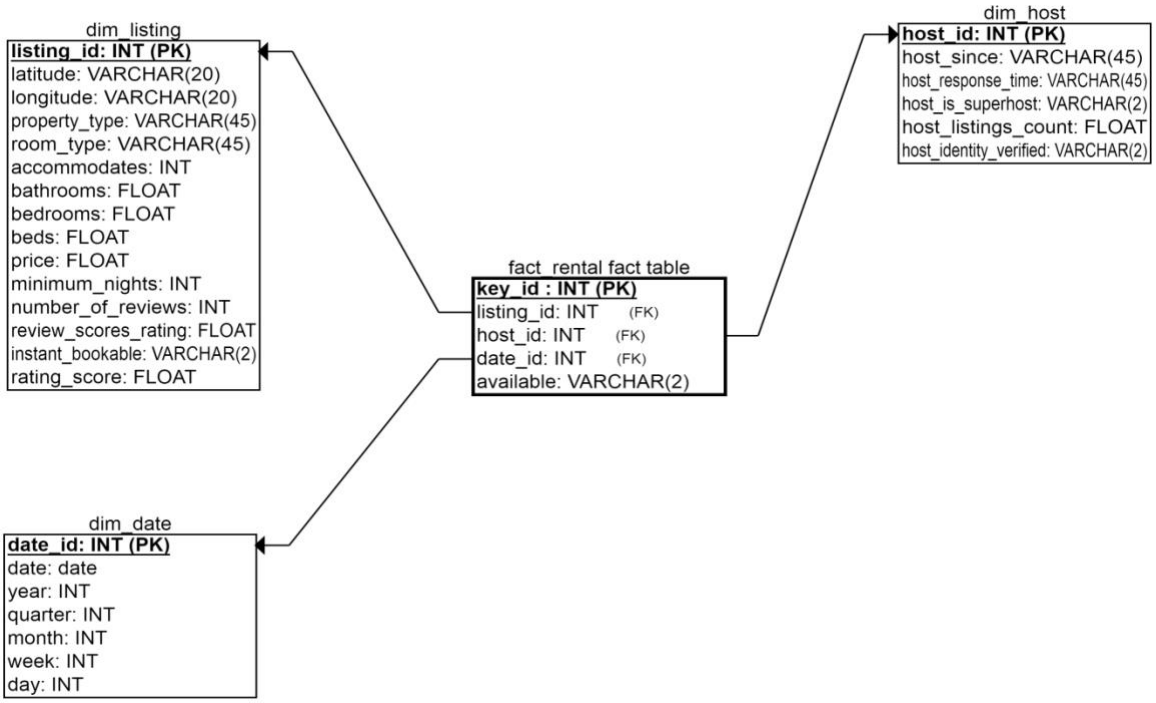
**בחירת המימדים:**

- Listings : מימד זה מכיל מידע לגבי הנכסים עצמם כגון דירוג ואופי הנכס.
- Host : מימד המפרט אודות המארחים מבחינת דירוג הלקוחות, זמן תגובה ועוד.
- זמן : מימד המכיל מידע לגבי התאריכים הרלוונטים לנתונים שנלקחו כגון חודש, רבעון ושנה.

**זיהוי עובודת:**

העובדה מורכת מנכס, מארח, תאריכים ותפוסה. שלושת הראשונים מיוצגים על ידי מספרי זיהוי חד-חד ערכיים לנכס, מארח ותאריך בטבלאות המימדים המתאימים. לכל צירוף של אלו קיימת אינדיקציה לאם הנכס הוזמן או לא. בנוסף, קיים מספר המשמש כמפתח ראשי לזיהוי עובדה בטבלה. על ערך זה נעשית אגרגציה לצורך ניתוח הנתונים והסקת מסקנות.

### תרשים סכמת הכוכב:




## טבלת dim_host

| שדה | טיפוס נתונים | מקור נתונים | הערות |
|---|---|---|---|
| host_id | int | שדה host_id בטבלת listings | מפתח ראשי, מקושר למפתח זר בטבלת fact rental |
| host_since | string | שדה host_since בטבלת listings | לא השתמשנו בניתוח הנוכחי, ייתכן ויעזור בניתוח עתידי |
| host_response_time | string | שדה host_response_time בטבלת listings | הערה (1) |
| host_is_superhost | string | שדה host_is_superhost בטבלת listings | לא השתמשנו בניתוח הנוכחי, ייתכן ויעזור בניתוח עתידי |
| host_listings_count | float | שדה host_listings_count בטבלת listings | לא השתמשנו בניתוח הנוכחי, ייתכן ויעזור בניתוח עתידי |
| host_identity_verified | string | שדה host_identity_verified בטבלת listings | לא השתמשנו בניתוח הנוכחי, ייתכן ויעזור בניתוח עתידי |

הטבלה נבנתה באופן הבא:

א.לקיחת הנתונים מקובץ listings ובחירת העמודות המפורטות בטבלה בלבד.

ב. הסרת ערכים כפולים (בדומה לפקודת distinct)

ג. הסרת שורות שלהן ערכים ריקים בלבד פרט למספר מארח מתוך הנחה שאלו נתוני "זבל"

ד. ההנחה שנעשתה במשתנה host_response_time היא שנתון חסר מהווה ערך של "לעולם לא". לכן ערכים ריקים מולאו במילה "never" (הערה 1).

## טבלת dim_listing

| שדה | טיפוס נתונים | מקור נתונים | הערות |
|---|---|---|---|
| listing_id | int | שדה listing_id בטבלת listings | מפתח ראשי, מקושר למפתח זר בטבלת fact rental |
| latitude | float | שדה Latitude בטבלת listings | לא השתמשנו בניתוח הנוכחי, ייתכן ויעזור בניתוח עתידי |
| longitude | float | שדה longitude בטבלת listings | לא השתמשנו בניתוח הנוכחי, ייתכן ויעזור בניתוח עתידי |
| property_type | string | שדה property_type בטבלת listings | לא השתמשנו בניתוח הנוכחי, ייתכן ויעזור בניתוח עתידי |
| room_type | string | שדה room_type בטבלת listings | לא השתמשנו בניתוח הנוכחי, ייתכן ויעזור בניתוח עתידי |
| accommodates | int | שדה accommodates בטבלת listings | לא השתמשנו בניתוח הנוכחי, ייתכן ויעזור בניתוח עתידי |
| bathrooms | float | שדה bathrooms בטבלת listings | לא השתמשנו בניתוח הנוכחי, ייתכן ויעזור בניתוח עתידי, הערה (2) |
| bedrooms | float | שדה bedrooms בטבלת listings | לא השתמשנו בניתוח הנוכחי, ייתכן ויעזור בניתוח עתידי |
| beds | float | שדה Beds בטבלת listings | לא השתמשנו בניתוח הנוכחי, ייתכן ויעזור בניתוח עתידי, הערה (2) |
| price | float | שדה price בטבלת listings | לא השתמשנו בניתוח הנוכחי, ייתכן ויעזור בניתוח עתידי, הערה (3) |
| minimum_nights | int | שדה minimum_nights בטבלת listings | לא השתמשנו בניתוח הנוכחי, ייתכן ויעזור בניתוח עתידי |
| number_of_reviews | int | שדה number_of_reviews בטבלת listings | הערה (2) |
| review_scores_rating | float | שדה review_scores_rating | הערה (2) |

| | | | |
|---|---|---|---|
| | | בטבלת listings | |
| instant_bookable | string | שדה instant_bookable בטבלת listings | |
| rating_score | float | מחושב על ידי נוסחא (1) | משקלל את הנתונים מ-number_of_reviews ו-review_scores_rating |

הטבלה נבנתה באופן הבא:

א. בחירת העמודות המופיעות בטבלה מקובץ listings.

ב. נעשה עיבוד מקדים להסרת התו "$", מאחר ומובן שהערך מייצג מחיר ולא ניתן לבצע עליו פעולות מתמטיות כ-string. הסרת התו איפשרה להפוך אותו ל-float לצורך ניתוח עם כלים מתמתטיים (הערה 3)

ג. חישוב עמודת rating_score על סמך כמות הדירוגים שהוזנו לנכס ואיכותם. החישוב מבוסס על נוסחא (1):

$$\text{score} = \text{Pp} + 100(1-\text{P})(1-\text{e}^{-q/Q}) \qquad (1)$$

הנוסחא נלקחה ממאמר המופיע בכתובת האינטרנט הבאה:

https://math.stackexchange.com/questions/942738/algorithm-to-calculate-rating-based-on-multiple-reviews-using-both-review-score

ד. אנו מניחים כי ערך חסר למספר בתי שימוש בנכס מעיד על כך שאין חדר שירותים, כמו למשל בדירת סטודיו של חדר אחד. בנוסף אנו מניחים כי ערך חסר במספר מיטות משמע שאין מיטות כגון בנכס מסוג מחסן. אותן הנחות נעשו גם לגבי מספר ביקורות ודירוג, כאשר ערכים חסרים משמעם 0 ביקורות. לכן בכל המקרים האלו הזנו את הערך 0 במקום NA במקומות הרלוונטים.

## טבלת dim_date

| שדה | טיפוס נתונים | מקור נתונים | הערות |
|---|---|---|---|

| | | | |
|---|---|---|---|
| date_id | int | ניתן אינדקס רץ בעזרת פונקצית reset_index של pandas לאחר שליפת כל התאריכים הרלוונטים מטבלת calendar | מפתח ראשי, מקושר למפתח זר בטבלת fact rental. |
| date | string | חושב מתוך שליפת הערכים היחודיים בשדה date בטבלת calendar | |
| day | int | חושב ב-pandas מתוך התאריך ע"י פונקית DatetimeIndex | ערכים בין 1-31 |
| month | int | חושב ב-pandas מתוך התאריך ע"י פונקית DatetimeIndex | |
| quarter | int | חושב ב-pandas מתוך התאריך ע"י פונקית DatetimeIndex | |
| week | int | חושב ב-pandas מתוך התאריך ע"י פונקית DatetimeIndex | ערכים בין 1-52 |
| year | int | חושב ב-pandas מתוך התאריך ע"י פונקית DatetimeIndex | |

הטבלה נבנתה באופן הבא:

א. נלקחה עמודת התאריך מקובץ listings והוסרו כפילויות (בדומה לפקודת distinct).

ב. בעזרת פקודת DatetimeIndex של pandas נבנו שאר העמודות אשר מהוות את המאפיינים השונים של כל תאריך כגון יום, חודש, שבוע בשנה, ושנה.

ג. נבנה אינדקס מחדש לתאריכים.

## טבלת Fact_rental

| שדה | טיפוס נתונים | מקור נתונים | הערות |
|---|---|---|---|

| | | | |
|---|---|---|---|
| listing_id | int | שדה listing_id בטבלת listings | מפתח ראשי, מפתח זר לטבלת dim_listing |
| available | int | שדה available בטבלת calendar | נתון לשינוי, יש לעדכן מבסיס הנתונים לפני חישובים |
| host_id | int | שדה host_id בטבלת listings | מפתח ראשי, מפתח זר לטבלת dim_host |
| date_id | int | נעשה merge בין שדה date בטבלת listings שנלקח יחד עם ה-listing_id לבין שדה date_id בטבלת dim_date | מפתח ראשי, מפתח זר לטבלת dim_date |

הטבלה נבנתה באופן הבא:

א. מתוך טבלת מימד listing נבחרו עמודות listing_id ו-host_id כבסיס.

ב. נעשה left join על הטבלה שהתקבלה ב-א עם נתוני קובץ calendar על סמך עמודת listing_id לצורך חיבור זמינות הנכסים

ג. נעשה left join עם טבלת מימד הזמן על סמך עמודת date לצורך חיבור מזהה התאריך date_id.

ד. שינוי הערכים של t ו-f ל-0 ו-1 לצורך שימוש בכלים אגרגטיבים ומתמטים בניתוח הנתונים.

## חלק ב' - יצירת דוחות :

### שאלת מחקר:

*האם אחוז התפוסה של נכס מושפע מביקורות על הנכס (בכמות ובאיכות), זמן תגובה של המארח לפניה, והאופציה לשריין את הנכס באופן מיידי?*

### רציונל לשאלת מחקר:

מנקודת מבט עסקית, אחוז התפוסה של נכס משקף את רמת הפעילות שלו, ומנתון זה אנו מסיקים על טיבו כהכנסה לאתר AirBNB. אחוז תפוסה גבוה יניח את הדעת בעוד אחוז תפוסה נמוך דורש בחינה מעמיקה יותר. הקריטריונים על פיהם נבחן את הנכס כאשר רמת פעילותו נמוכה הם כאמור מספר ביקורות והציון שניתן בהן, זמן התגובה של המארח ואופציה לשריון מיידי. בחרנו משתנים אלו משום שאנו מאמינים כי הם מהווים גורמים בעלי השפעה מהותית על חווית המשתמש באופן מהימן. בשונה מחופשה במלון, משתמש שנכנס לאתר לעיתים קרובות רוצה תחושה של יחס אישי, ממש כמו הדירה שישכור. תגובה מהירה לפנייתו, אפשרות להרגיש בטחון שהנכס משוריין, וכמות גדולה של ביקורות (שלא ניתן לזייף על ידי חברים קרובים ומשפחה) חמות יקנו תחושה זו של קשר ישיר ואינטימי יותר. קשר זה מהווה בסיס לתחושת בטחון והנאה מהחופשה/נסיעה כולה. לכן, אנשים

בוחנים משתנים אלו באתר, ומחפשים לשכור נכס בעל נתונים חיוביים, בכדי להיות בטוחים עד כמה שאפשר שהחופשה/נסיעה תעבור בצורה חלקה.

## תוצאות ומסקנות:

מבחינת **rating** אותרו 3 מגמות:

1. מתחת לדירוגים של 60 נראים מספר מקרים בהם אחוז התפוסה נמוך במיוחד.
2. מעל דירוגים של 60 יש מגמת עליה יחסית יציבה. אולם, ברמות הדירוג הגבוהות, יש הבדל בין חודשי החורף (דצמבר, ינואר, פברואר) לבין חודשי הקיץ ברמת התפוסה, כאשר בחורף התפוסה מעט נמוכה יותר. נסביר ממצא זה יחד עם ממצא 2 לגבי זמן תגובה.

שתי המגמות עולות בקנה אחד עם ההשערה כי דירוג לקוחות נמוך של listing נוטה להיות פנוי יותר. כלומר, ישנן פחות הזמנות. כנראה שלקוחות נמנעים במידת מה "לקחת את הסיכון" וללון במקום שקיבל דרוגים נמוכים. ממצא זה חוזר על עצמו בכל חודשי השנה וגם בניתוח בשנתי.

מבחינת **response time** אותרו 2 מגמות:

1. נראה כי באופן עקבי זמן חזרה השואף ל"אף פעם" מביא לירידה משמעותית ברמת התפוסה של נכסים אשר בעליהם מוגדרים ככאלה
2. נראה כי בחודשי החורף (דצמבר, ינואר, פברואר ואפשר לומר מרץ) בשנה, זמני תגובה קצרים יותר הכוללים עד שעה, עד מספר שעות ועד יממה תפוסים פחות מאשר אותם זמני תגובה בשאר חודשי השנה. בחודשי השנה שאינם חודשי החורף שהזכרנו, התפוסה דומה לזו של זמן תגובה עד מספר ימים.

המגמה השנייה היא ממצא לא צפוי, אך היא מתכתבת עם ממצא 2 לגבי rating. אולי הסבר אפשרי להם הוא כי בחודשי החורף ישנן באופן כללי פחות הזמנות לנסיעות ולינות, ואלו הן כנראה נסיעות לצורך עסקים ולא חופשות פנאי (מחוץ ל"עונה"). במקרים אלו הן לרוב ידועות מראש ולא ספונטניות, כך שזמני התגובה מקבלים פחות חשיבות. בשילוב עם זמני התגובה, גם דירוג הנכס פחות משנה, משום שאלו הן נסיעות עסקים אשר הפנאי והחדר שוליים למטרת הנסיעה, וכך גם אולי איכות החדר (החברות ששולחות את העובדים לא מבזבזות כסף על איכות החופשה).

מבחינת **instant bookable** נראה כי ההבדלים קטנים מאוד ומתחלפים בכיוונם על פני צירי זמן התגובה והחודש בשנה. לכן, לא נתייחס אליהם כבעלי משמעות.

**<u>לסיכום:</u>** המלצתנו לאתר תהיה לנסות ולשווק בחודשי החורף הכוללים את דצמבר, ינואר, פברואר ומרץ נכסים המתאימים יותר לנסיעות עסקיות. נכסים אלו מתאימים לאדם אחד ותקופות קצרות. עם זאת, ישנו צורך בניתוחי המשך בנסיון לאמת את השערתנו לגבי התוצאות. למשל, לבחון אם קיים הבדלים בסוג הנכס בין חודשי החורף לשאר השנה על ידי בחינת השפעות סוג הנכס (property_type),סוג החדר (room_type), לילות מינימום שניתן להזמין (minimum_nights) על תפוסת החדרים לאורך חודשי השנה.