To BNB or not to be: What Makes an Airbnb Stand Out?

תיאור הבעיה:
Airbnb מציע מגוון רחב של דירות, אך חלקן מוצלחות יותר מאחרות. רצינו לבדוק מה גורם לדירות מסוימות להצליח ולהתבלט, מה התכונות של הדירות המוצלחות, ואיך להימנע מלהיות דירה לא מוצלחת, תוך הסתמכות על 3 מדדי הצלחה: דירוג משתמשים, הערכת תפוסה, הערכת הכנסה. השתמשנו ב-frequent itemset על מנת לראות אילו תכונות שכיחות אצל הדירות הכי מוצלחות ושאינן שכיחות בדירות הכי לא־מוצלחות.
כמו כן, השתמשנו בקלאסטרים, כדי לאפיין את סוגי הדירות באופן unsupervised ולמצוא קשרים נסתרים ולא טריוויאליים. לבסוף יצרנו מערכת המלצה המתבסת על Collaborative Filtering אשר ממליצה על תכונות לשיפור לדירה לא־מוצלחת על בסיס תכונות של דירות דומות שכן מוצלחות.

תיאור הדאטה:
השתמשנו בדאטה סט מקיף שמספק מידע מפורט על רשומות Airbnb בערים שונות מאתר בשם insideairbnb.
על מנת להרחיב את הדאטה, איחדנו את הדאטה של הערים: פריז, אמסטרדם וברצלונה - ערי בירה מתוירות במערב אירופה שמושכות אליהן סוג דומה של תיירות. גודל הדאטה המאוחד: MB144 מספר השורות: 62672, מספר העמודות 80. הדאטה לקוח מהאתר: .insideairbnb
מתוך 80 העמודות השתמשנו ב-19 שחשבנו שהכי משמעותיות והן:

- ID - המספר הייחודי של כל נכס
- Picture url - לינק לתמונה הראשונה שמופיעה באתר של Airbnb
- host_is_superhost ,host_identity_verified ,host_identity_verified host_has_profile_pic - עמודות שמציינות האם למשכיר הוא superhost או יש לו תמונת פרופיל או אומת על ידי Airbnb
- Host_total_listings_count - מספר הנכסים שיש למשכיר
- Property_type - סוג הנכס - בית פרטי, חדר בדירה משותפת וכו'
- Accommodates, bathrooms, bedrooms, beds, minimum_nights - מציינות כמה יש מכל אחד
- Instant_bookable - האם אפשר לקבוע ללא המתנה לתשובה מהמארח
- Description - תיאור הדירה כפי שרשום ב-Airbnb, נכתב מטעם המארח
- Amenities - רשימה של amenities שנכתבה מטעם המארח.
- review_scores_rating - הדירוג של הדירה ב-Airbnb
- Estimated_occupancy_l365d, Estimated_revenue_l365d - הערכה שמחושבת ע"י InsideAirbnb ומפורטת בקישור: Data Assumptions

עיבוד מקדים:
למטרותינו, היינו צריכים להפוך את הדאטה לעמודות בינאריות, על מנת להשתמש בו לביצוע frequent itemset וכן לשאר הניתוחים שביצענו. ראשית מחקנו את כל השורות בהן היה חסר מידע בעמודה מסוימת שקבענו כקריטית ובנוסף מחקנו את העמודות שקבענו כלא רלוונטיות. את העמודות image_url,description,amenities העברנו עיבוד מורכב יותר באמצעות מודלים ואלגוריתמים שונים עליהן נפרט בנפרד. השתמשנו בעמודה bathrooms_text כדי למלא ערכים חסרים בעמודה של bathrooms וגם כדי ליצור עמודה ביניארית חדשה שמציינת האם השירותים הם פרטיים או לא. את שאר העמודות המרנו לעמודות ביניאריות ב-2 דרכים שונות, הראשונה היא שכל עמודה שהייתה בעלת ערכים בוליאנים המרנו לערכים בינאריים.
השנייה הייתה דרך המרה של כל עמודה ל 3-5 עמודות בהתאם לסלים שקבענו לפי שיקולים של ידע כללי או חלוקה לאחוזונים. למשל את העמודה accommodates חילקנו לסלים 1,2,3-4,5-6,7+ מתוך המחשבה שמדובר בקטגוריות אדם יחיד, זוג, משפחה קטנה, משפחה בינונית ומשפחה/קבוצה גדולה. לעומת זאת, בעמודת ה-price השתמשנו ברבעונים: 0-24%, 25-49%, 50-74% ו-75-100%.

עיבוד תמונות:
העמודה picture_url בדאטה מייצגת את התמונה הראשונה של הנכס כפי שהיא מופיעה באתר Airbnb.
עבור כל תמונה יצרנו 4 מאפיינים שונים - רזולוציה, בהירות, mood וסוג החדר. את המאפיינים רזולוציה, בהירות ו-mood חישבנו באמצעות אלגוריתמים של עיבוד תמונה מהספרייה 2CV ולאחר מכן פיצלנו אותם

לעמודות ביניאריות שונות כאשר הרזולוציה והבהירות פוצלו לפי רבעונים ו-mood לפי אם התמונה קרה או חמה. את הסיווג לאיזה חדר מופיע בתמונה עשינו באמצעות מודל בשם CLIP ורשימת תוויות שהכנו מראש. בנוסף סיננו את הנכסים שהביטחון של המודל בסיווג התמונה שלהם היה נמוך מ-70 אחוז או אלה שהתמונה שלהם כבר לא קיימת. לאחר מכן המרנו את התוויות לקבוצה מצומצמת יותר שכוללת את החדרים שהתגלו כנפוצים ביותר: סלון, חדר שינה, מטבח, חדר אמבטיה, מרפסת וגם קטגוריה בה התמונה היא לא של הבית.

עיבוד ה**amenities**:
בדאטה המקורי הamenities הופיעו בעמודה אחת כרשימה שהוזנה ידנית ע"י בעל הדירה ולכן אין אחידות במינוחים בין הרשומות. רצינו להפוך את העמודה לעמודות ביניאריות, שמציינות האם הamenity קיים ברשומה או לא. לשם כך, יצרנו מילון עם קטגוריות שונות של הamenities, ובנוסף פיצלנו כל קטגוריה לקטגוריית בסיס וקטגוריית אקסטרה על מנת להבדיל בין אבזור חיוני לבין אבזור שהינו מותרות. לדוגמה, אם בדירה אין מקרר (מוצר בסיסי) אך יש בלנדר (מוצר אקסטרה), נוכל להבין שבדירה חסר מוצר בסיסי, למרות שכביכול שניהם בקטגוריה של כלי מטבח. עבור כל קטגוריה רשמנו מילים מייצגות לקטגוריה.
עבור כל רשומה, עברנו על רשימת הamenities ונרמלנו אותם באופן הבא:

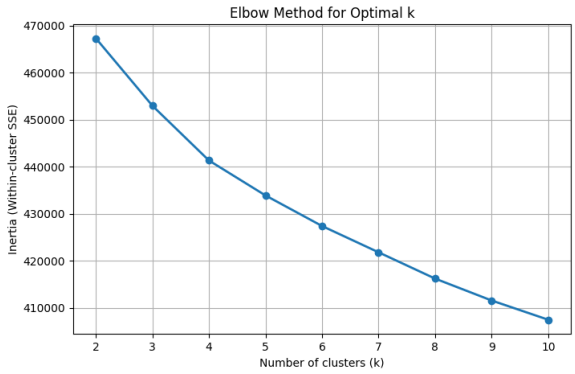
- השארנו רק אותיות קטנות באנגלית, ומחקנו stop words (בין היתר מחקנו את המילים ' listing', 'extra', 'cost', 'included', 'private').
- כמו כן, שמנו לב שהרבה מהamenities צויינו עם המותג של המוצר, ולכן הקוד לא סיווג אותם כאותו amenity. לכן השתמשנו בספריית Flair, אשר יודעת לתייג מילים כשמות של ארגונים או מותגים, וכך ידענו למחוק את שמות המותגים טרם הסיווג.

לאחר מכן אם המילה נמצאת ברשימה המייצגת של אחת הקטגוריות, נסווג אותה לשם. אחרת, נמצא את הקטגוריה שהembeddings של המילים בה הכי קרובות לamenity שאנו רוצים לסווג. המטריקה בה השתמשנו היא cosine similarity.

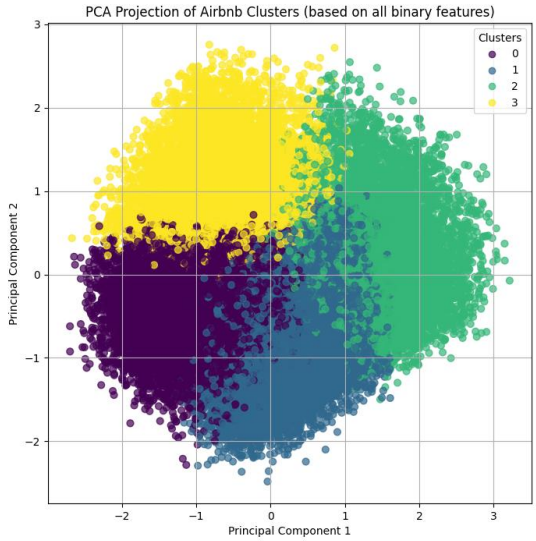
עיבוד מקדים תיאורי הדירות (**description**):
בעיבוד המקדים של תיאורי הדירות (description) הגדרנו מראש סט של תגיות שמייצגות מאפיינים שונים של הדירה, כגון ניקיון, מטבח, מיקום, סוג הדירה, אווירה ותכונות נוספות. לכל תגית הגדרנו מילון מונחים אופייניים, וחישבנו cosine similarity בין המשפטים בתיאור הדירה לבין המונחים המאפיינים כל תגית. על מנת לייצג באופן וקטורי את המשפטים והמילים השתמשנו במודל שפה (SentenceTransformer). ראשית, עבור כל דירה תיאור הדירה חולק למשפטים והם הועברו למודל להפקת embeddings. לאחר מכן, עבור כל תגית נבדק אם המרחק בין המשפטים לבין המונחים המאפיינים את התגית עובר סף מוגדר. אם כן, התגית סומנה כ-1 עבור אותה דירה, אחרת כ-0. התוצאה הסופית היא שלכל תיאור דירה יש רשימת תגיות המייצגות את התכנים שעליהם מדבר המארח בתיאור הדירה.

clustering

ביצענו קלאסטרינג כדי לגלות דפוסים במידע. האלגוריתם מאפשר לנו לחלק את כל הדירות לתתי־קבוצות (קלאסטרים), כך שבתוך כל קבוצה הדירות דומות זו לזו, אך שונות מדירות בקבוצות אחרות. באמצעות הקלאסטרינג הצלחנו להפוך את הנתונים הגולמיים לתמונה ברורה יותר של סוגי הדירות בשוק.
שיטת הקלאסטרינג: לצורך חלוקת הדירות לקבוצות השתמשנו באלגוריתם K-Means.
בחירת מספר הקלאסטרים - **K**: לבחירת מספר הקלאסטרים הרצוי השתמשנו ב-Elbow Method שבה חישבנו את ה-*inertia* עבור ערכי K שונים (2–10). נקודת ה"Elbow" מייצגת את המקום שבו השיפור בהוספת קלאסטר נוסף מתחיל להיות זניח. בהתאם לגרף שהתקבל, נבחר הערך K=4.



משתנים שנלקחו בחשבון בקלאסטרינג: לצורך החישוב השתמשנו בכל המשתנים הבינאריים כגון: סוג הנכס, amenities, אפשרות להזמנה מיידית, מספר מקומות לינה, טווחי מחירים, סוג חדרים, מאפייני התמונה ומאפייני תיאור הדירה על ידי המארח. בנוסף, נלקחו בחשבון מספר משתנים מספריים לצורך ניתוח והשוואה בין הקלאסטרים, אך לא כחלק מחישוב הקלאסטרים עצמו: דירוג ממוצע, וימי תפוסה משוערים בשנה.



השתמשנו ב-PCA כדי לצמצם את המידע למרחב דו־ממדי, ולאחר מכן יצרנו scatter plot כאשר כל קלאסטר מיוצג בצבע אחר. ניכר כי הדירות התחלקו לקבוצות מובחנות, דבר המרמז על כך שהאלגוריתם הצליח לזהות מבנים מסוימים בנתונים, גם אם ההפרדה אינה מוחלטת.

איפיון הקלאסטרים: לאחר ביצוע הקלאסטרינג, ניתחנו כל קלאסטר באמצעות חישוב הממוצעים בעמודות נבחרות בכל קלאסטר. כמו כן חישבנו והשוונו את הדירוג ממוצע וימי תפוסה משוערים בכל קלאסטר על מנת להבין האם יש תבניות בעלות משמעות. בחרנו שלא לכלול בניתוח את המדד של ההכנסה כי הוא מוטה לכיוון דירות גדולות ויקרות ולכן לא הוסיף ערך.

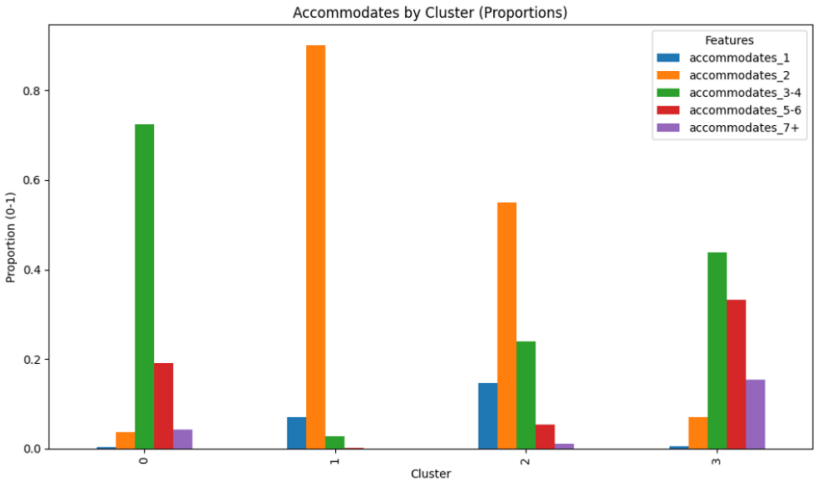
קלאסטר 0 - "דירות ל־3-4 אנשים, מאובזרות היטב"
נכסים איכותיים, עם הדירוג הממוצע הגבוה ביותר מבין ארבעת הקלאסטרים (4.83) ושיעור גבוה יחסית של "סופר־מארחים" (33%). הדירות מאובזרות היטב, מוצעות בטווח מחירים בינוני-גבוה, ומיועדות בעיקר ל־3-4 אורחים (72.5%).

קלאסטר 1 - "דירות לזוג במחיר נגיש"
דירות קטנות ונגישות במחיר, בעלות דירוג ממוצע גבוה (4.78) ותפוסה ממוצעת גבוהה (120 ימים). כמעט כל הדירות בקלאסטר זה (96%) הן מסוג "דירה שלמה". 90% מהדירות בקלאסטר זה מיועדות לזוגות, ורובן המכריע (74%) מתומחרות עד 160 דולר ללילה.

קלאסטר 2 - "דירות בסיסיות עם אבזור חלקי וחדרים משותפים/הוסטלים"
זהו הקלאסטר עם הביצועים הנמוכים ביותר מבחינת דירוג ותפוסה: הדירוג הממוצע הנמוך ביותר (4.61) לצד התפוסה הממוצעת הנמוכה ביותר (76 ימים). שיעור ניכר מהנכסים הם חדרים משותפים או הוסטלים (כ־41%),

ורבים מהנכסים בקלאסטר זה חסרים אבזור חיוני כגון חימום, מטבח או מקומות אחסון. יתר על כן, שיעור ה"סופר־מארחים" בקלאסטר זה נמוך (17%).

קלאסטר 3 - "דירות גדולות ומאובזרות היטב של מארחים מרובי נכסים"
זהו הקלאסטר עם הביצועים הטובים ביותר מבחינת תפוסה ממוצעת (123 ימים) אך דירוג ממוצע נמוך (4.62). רוב הנכסים שייכים למארחים מקצועיים - 61% מהמארחים מחזיקים ביותר מ־21 נכסים, ול־82.6% מהם יש מעל 6 נכסים. הדירות עצמן גדולות ויקרות: 80% עולות מעל 160 דולר ללילה, 92.3% מיועדות ל־3 אורחים ומעלה, כאשר כ-50% מהדירות מתאימות ל־5 אורחים ומעלה. הדירות מאובזרות היטב, וב־69.1% מהן קיימת אפשרות להזמנה מיידית (לעומת ממוצע של 24.6% בדירות השייכות לקלאסטרים האחרים).



Frequent Item-Set

ראשית התמקדנו רק בנכסים בהם משכירים את כל הנכס כי העמודה entire_house שמציינת כי סוג הנכס הוא בית פרטי נמצאת בכ-82 אחוז מהשורות, כלומר יש הטייה לכיוונו.
כדי לוודא שלא מדובר במקריות הסתכלנו על ההפרש ב-support שבין הדירות הכי מוצלחות לבין ההכי לא מוצלחות. דירות מוצלחות הוגדרו ככאלו שנמצאות ב-20% העליונים בכל אחד משלושת המדדים review_scores_rating, Estimated_occupancy_l365d, Estimated_revenue_l365d. באופן דומה דירה לא מוצלחת נחשבת ככזאת כאשר היא ב-10% התחתונים בכל אחד מהמדדים. כיוון שבחיתוך של 10% הרשומות התחתונות בכל המדדים יש מספר רב יותר בצורה משמעותית של דירות מאשר בחיתוך של 10% העליונים בכל המדדים, לקחנו את ה-20% הרשומות העליונות. בחרנו להתייחס לקבוצות בגדלים של 1 או 2 שכן לאחר בדיקה ההבדל בsupport בין קבוצה בגדול 2 לבין סט בגודל 3 הוא לא משמעותי. את ההבדל ב-support חישבנו באופן הבא:

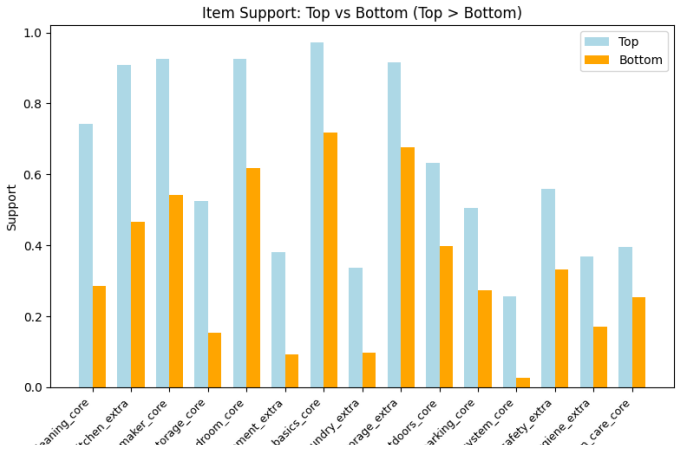
added_value=new_duo_diff-max(item1,item2)

עבור קבוצות בגודל 2 התווספו הפרשים ניכרים ב-support (כ10%) ולכן החלטנו להתייחס אליהם, אם כי הקבוצות בגודל 1 הכי משמעותיות שכן הן מסבירות את רוב ההפרש ב-support בין הדירות המוצלחות ללא מוצלחות.
בנוסף, מעבר לחישוב ה-Frequent Item-Set שהרצנו על הדאטה יחד, גם חישבנו frequent item-set עבור כל אחד מהתחומים - תמונה, תיאור ו-amenities בנפרד כדי לנסות להגיע לתוצאות משמעותיות יותר. נציג את ההתרשמויות ממבט על כל תחום בנפרד ולאחר מכן הדאטה הכולל, כאשר נתמקד במאפיינים בודדים וזוגות.

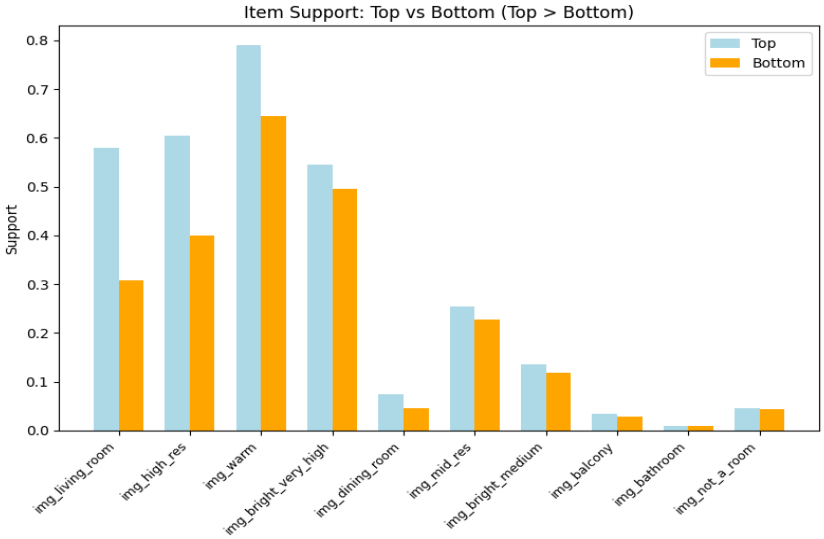
Amenities: בגרף ניתן לראות את השכיחות (support) של 15 ה-Amenities בהם הפער היה הכי גדול בין הרשומות המוצלחות לרשומות הלא מוצלחות. ניתן לראות למשל כי קיומה של מכונת קפה מופיע בפער גדול בדירות הטובות (93% מול 54%). ייתכן שזהו מתקן בעל ערך גבוה לאורחים שיכול לשדרג את החוויה. מנגד, ייתכן שמדובר בסימן כללי לכך שבעל הדירה משקיע יותר בתחזוקה ובאבזור, ולאו דווקא בהשפעה הישירה של

מכונת הקפה עצמה. עוד כמה בעלי הפרש משמעותי הם אביזרי ניקוי בסיסיים (74% מול 29%) ואביזרי מטבח אקסטרה (91% מול 47%).

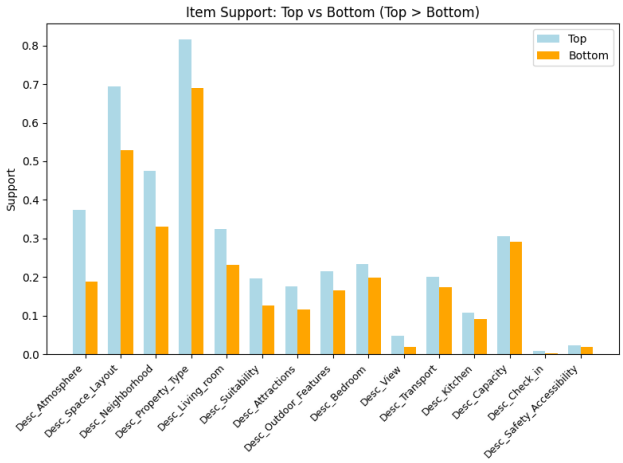
בנוסף ישנם גם מאפיינים שכאשר מצרפים אותם מקבלים הפרש גבוה אף יותר בשכיחות ולכן משמעותיים. למשל אביזרי מטבח בסיסיים בצירוף שטח אחסון מוסיף כ-10% הפרש בsupport בין המוצלחות ללא מוצלחות ובאופן כללי נמצא בכ-90% מהדירות המוצלחות לעומת 54% בלא מוצלחות. עוד צירוף משמעותי הוא של אביזרי חדר שינה בסייסים ומכונת קפה, מגדיל את הפער בכ-6% נוספים. צירוף זה נמצא בכ88% מהדירות המוצלחות ורק ב42% מהדירות הלא מוצלחות.



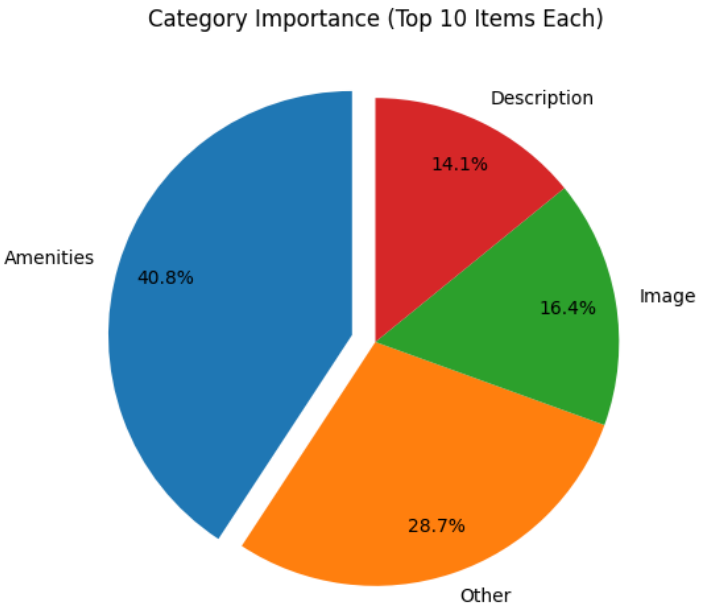
תמונות: בגרף ניתן לראות את השכיחות של כל המאפיינים של התמונה שמופיעים יותר בדירות המוצלחות מאשר בלא מוצלחות. ניתן לראות למשל כי התמונה הראשית היא של סלון בכ-60% מהנכסים המוצלחים ביותר לעומת כ-30 אחוז מהנכסים הלא מוצלחים. וכי התמונה היא באיכות גבוהה בכ-60% מהנכסים המוצלחים לעומת כ-40% בלא מוצלחים. ייתכן שאיכות התמונה הגבוהה נובעת מלכתחילה מהשקעה גבוהה יותר של בעל הנכס ולכן זה נפוץ יותר בדירות המוצלחות ולא מתוך כך שהתמונה מושכת יותר את העין. הסבר אפשרי לכך שהתמונה של הסלון נפוצה יותר בדירות המוצלחות הוא שהוא החדר ייצוגי בבית או שבדירות הלא מוצלחות לעיתים אין סלון. הצירוף היחיד של מאפיינים שמגדיל את ההפרש ב-support הוא של תמונה חמה עם תמונה באיכות גבוהה (49% במוצלחות לעומת 25% בלא מוצלחות) שמוסיף כ-2% בלבד.



תיאור הדירה: בדומה לסעיפים הקודמים, בגרף ניתן לראות את השכיחות של תיאורי דירה בדירות מוצלחות ולא מוצלחות. ניתן לראות למשל כי ההפרש הכי גדול נמצא בתיאור האווירה (38% מול 19%) ואחריו תיאור המרחב בדירה (70% מול 52%). גם כאן ההפרשים יכולים להיות מוסברים כסימן לכך שבעל הדירה משקיע יותר בנכס ובין היתר בתיאור הדירה באופן כללי, מבחינת הפירוט ומבחינת הניסוח, ולאו דווקא בכך שדגשים שונים או נוספים של התיאור מושכים יותר מאחרים.



כלל הדאטה: בתרשים ניתן לראות את החלק היחסי של עשרת המאפיינים עם ההפרש הכי גבוה לפי קטגוריה. בחרנו להתייחס רק ל-10 המובילים כי מספר המאפיינים בכל קטגוריה שונה וזה מטה את המשקל של כל אחת על ההפרש הכללי. ניתן לראות כי ה-amenities היא הקטגוריה המשפיעה ביותר על סכום ההפרשים של הsupport בין המוצלחות ללא מוצלחות ולכן אנחנו מסיקים כי תוספת של amenities לדירה כנראה תניב הצלחה גדולה יותר משאר הקטגוריות. עם זאת ברמה הפרקטית השינוי של התמונה והתיאור של הדירה הוא מאוד פשוט ונגיש, וביחד כן מדובר בהפרש משמעותי של כ-30% ולכן, לדעתנו, לשנות את אותם זה הכי משתלם.



תופעה אחת מעניינת שנתקלנו בה היא שדירות שבעליהם מחזיקים ביותר מ-21 נכסים שכיחות יותר בדירות לא מוצלחות (14% במוצלחות לעומת 55% בלא מוצלחות). לעומת זאת ראינו בקלאסטרינג שדווקא הקלאסטר עם התפוסה הגבוהה ביותר היה הקלאסטר שבו האחוז הגבוה ביותר של מארחים המחזיקים ביותר מ-21 נכסים אך בעל ממוצע דירוג נמוך יחסית. ייתכן שזה נובע מכך שבfrequent item-set אנחנו מסתכלים על החיתוך בין שלושת המדדים מראש ולכן לא מתייחסים לדירות בהן רק חלק מהמדדים גבוהים. הסבר אפשרי לתופעה הוא שחברות נדל"ן פחות מתעניינות בחווית האורחים האינדיבידואליים ממארחים פרטיים אך משקיעות יותר במציאת לקוחות.

מערכת המלצה

יצרנו מערכת המלצה אשר מוצאת דירות דומות מוצלחות לדירה שאינה מוצלחת ע"פ מדד אחד, low rating/low estimated occupancy/ low estimated revenue, ומוצאת אילו דברים משותפים לדירות המוצלחות, שחסרות בדירה הלא־מוצלחת.

תחילה הגדרנו תכונות שחייבות להיות משותפות בין הדירה הלא־מוצלחת הנתונה, לבין הדירות המוצלחות שנשווה אליהן. לדוגמה, צריכות להימצא באותה העיר, להיות עם אותה כמות חדרים ומיטות ולהיות מיועדת לכמות אורחים שווה, שכן אחרת אנו לא מתמודדים עם אותו סוג של דירה. דברים שטובים לבית גדול ומרווח, כנראה לא טובים לדירת סטודיו המיועדת לאירוח של זוג. כמו כן, קבענו תכונות שאנו רוצים להתעלם מהן ולא להחשיב אותן במציאת הדירות הדומות, כגון מספר הדירות שיש לבעל הדירה בנוסף, שכן להציע לבעל דירה בעל רשומה לא־מוצלחת לקנות עוד 5 דירות, אינה הצעה ריאלית.

לאחר מכן מצאנו את k הדירות המוצלחות שהכי קרובות לדירה הלא־מוצלחת (באמצעות cosine similarity), לכל תכונה, אם אחוז מסוים (מועבר כהיפר־פרמטר) של הדירות המוצלחות לא מסכימות עם הדירה הלא־מוצלחת על תכונה מסוימת, אנו מציעים לדירה הלא־מוצלחת להוריד או להוסיף תכונה זו. למשל אם ברוב הדירות המוצלחות יש מזגן ובדירה הנתונה אין מזגן, נציע להוסיף מזגן. עבור התכונות שנמצאות בbuckets המלצנו להיות בbucket הכי פופולרי מבין k הדירות המוצלחות. בחרנו לא להציע להוריד amenities שנמצאות בדירה הנתונה אך לא נמצאות בדירות המוצלחות, שכן לא מצאנו סיבה שקיום amenity כלשהו ישפיע לרעה על הדירה. כמו כן, קבענו כי לא נמליץ על הוספה של הamenities הבאים: נוף, נגישות, חנייה, גינה, אטרקציות, שכן אלו דברים שבעל הדירה לא יכול לשנות.

יצרנו demo (סרטון) אינטראקטיבי למערכת ההמלצה. בdemo זה ניתן לבחור את המטריקה בה נמדדת ההצלחה, את האחוזון על מנת להיחשב דירה מוצלחת, האחוזון על מנת להיחשב דירה לא מוצלחת, וכן לשלוט בכמות השכנים המוצלחים של הרשומה הלא מוצלחת, ולבחור את הסף להמלצה להוספה ולהורדה.

add threshold - אחוז הדירות השכנות המוצלחות המינימלי שיש להן תכונה שאין לרשומה הנתונה, על מנת שנמליץ להוסיף אותה.

Remove threshold - אחוז הדירות המינימלי שאין להן תכונה שיש לרשומה הנתונה, על מנת שנמליץ להוריד אותה.

ניתן לבחור אם הקוד יבחר דירה באקראיות או מספר id של דירה ספציפית אליה נשווה.

הdemo מציג את הדירה הלא־מוצלחת שנבחרה, מה הן העמודות שעל פיהן נסנן את הדירות המוצלחות, את ההמלצות להוריד ולהוסיף תכונות, המלצות לשנות תכונה לbucket אחר, וויזאליזציה של היסטוגרמה, כמה אחוזים מהדירות המוצלחות השכנות מסכימות על תכונה מסויימת עם הדירה הלא־מוצלחת. בהיסטוגרמה אנו רואים כי הדירות השכנות מסכימות על רוב התכונות עם הדירה הלא־מוצלחת כך אנו יודעים כי ההמלצות מגיעות אכן מדירות שדומות לדירה הלא מוצלחת.

עבודה עתידית

- הרחבת הדאטה: ניתן להעשיר את הדאטה הנקי בתכונות נוספות, למשל שילוב מידע על אזורי העיר או ניתוח ביקורות המשתמשים. כך נוכל להוסיף מורכבות לדאטה ולקבל תוצאות מדויקות יותר.
- מודל חיזוי: נוכל לפתח מודל חיזוי שיקבל כקלט את מאפייני הדירה וינבא את רמת הצלחתה במדד שנבחר.
- שיפור מדדי הצלחה: מדדי ההצלחה בהם השתמשנו הינם הערכה לתפוסה של דירה ולרווח שלה. נוסיף דאטה ממקורות נוספים בהם מידע זה הוא עובדתי לגבי מדדים אלו.

סיכום

בפרויקט בדקנו אילו תכונות צריכות להיות לדירה מוצלחת. מצאנו באמצעות הfrequent itemset כי הamenities בעלות משקל הכבד ביותר עבור הצלחה של דירה, אך הדרך הקלה ביותר לשפר את הרשומה היא לערוך את התיאור של הדירה ואת התמונה המייצגת שלה.

מצאנו באמצעות הclustering שהדירות הכי פחות מוצלחות מכל הבחינות היו באותו הקלאסטר, אלה היו דירות עם אבזור חלקי ואחוז גבוה של חדרים משותפים/הוסטלים. הקלאסטר עם הדירות עם הדירוג הכי גבוה היו ל3-4 אנשים, מאובזרות היטב ושיעור גבוה של סופר־מארחים, טווח המחירים היה בינוני־גבוה. הקלאסטר עם הoccupancy הגבוה ביותר היה של דירות גדולות ומאובזרות היטב של מארחים מרובי נכסים. לבסוף יצרנו

מערכת המלצה שמתבססת על דירות שכנות מוצלחות (collaborative) ומציעה אילו תכונות כדאי לשפר על מנת להפוך את הדירה למוצלחת יותר.