ז׳אנה אילין
346186992

דו״ח מסכם בקורס פקרטיקום בנושא ניתוח נתונים של AirBNB

תוכן עניינים

רקע על תחום ה- Airbnb

Airbnb הוקמה בשנת 2008 על ידי בריאן צ'סקי, ג'ו גביה, ונתן בלקרצ'יק, והתחילה כדרך קלה להשכרת מקומות לינה לאנשים שהגיעו לעיר הגדולה כדי להשתתף בכנסי עיצוב. הרעיון המרכזי היה להציע לאנשים מקומות לינה במחירים נוחים, כאלטרנטיבה למלונות יקרים, באמצעות השכרה של חדרים או דירות פרטיות.
עם הזמן, Airbnb התפתחה לפלטפורמה גלובלית שמחברת בין אנשים שמחפשים מקום לינה לבין מארחים מכל רחבי העולם, שמציעים את הנכסים שלהם להשכרה לתקופות קצרות. החברה שינתה את פני תחום התיירות והאירוח בכך שהיא אפשרה לכל אחד להיות מארח ולתיירים למצוא מקומות לינה ייחודיים ומקומיים יותר.

מאז הקמתה Airbnb, שינתה בצורה משמעותית את השוק. היא הפכה לאחד השחקנים המרכזיים בתחום השכרת מקומות לינה לטווח קצר. המודל העסקי החדשני של Airbnb גרם למהפכה בשוק התיירות בכך שהוא אפשר למטיילים ליהנות מחוויות מקומיות במחירים נוחים, ולבעלי נכסים להרוויח כסף נוסף על ידי השכרת הנכסים שלהם.

עם זאת Airbnb, לא רק שינתה את השוק לטובה, אלא גם יצרה אתגרים ודיונים רבים סביב סוגיות כמו רגולציה, השפעה על מחירי השכירות בערים גדולות, ותחרות עם המלונאות המסורתית. חברות רבות בתחום האירוח המסורתי ניסו להתאים את עצמן לשינויים האלה, ולעיתים אף ניסו לקדם חוקים שיגבילו את פעילותה של Airbnb .

מחקרים רבים עסקו בהשפעת הפלטפורמה על מחירי השכירות, כאשר נמצא כי בערים בהן הפעילות של Airbnb גבוהה, ישנה נטייה לעליית מחירי השכירות, בעיקר באזורים מרכזיים ומבוקשים. השפעה זו העלתה שאלות וחששות לגבי היכולת של תושבים מקומיים למצוא דיור בר השגה, בעיקר בערים בהן התיירות היא מקור הכנסה מרכזי. לצד זאת, הפלטפורמה השפיעה גם על תעשיית המלונאות, כאשר מחקרים הצביעו על ירידה בתפוסה ועל לחץ להורדת מחירים בקרב מלונות במקומות בהם Airbnb פופולרית.

מה מיוחד במחקר שלי

בעבודה זו, לא הסתפקתי בניתוח בסיסי של מחירי השכירות ב-Airbnb אלא לקחתי את הניתוח צעד קדימה, ופיתחתי מערכת המלצות מבוססת ניתוח טקסט שנועדה לעזור למשתמשים למצוא נכסים דומים לנכס שמעניין אותם. השימוש במערכת המלצות המבוססת על ניתוח שפה טבעית (NLP) מהווה גישה ייחודית, מאחר והיא מאפשרת זיהוי קשרים מעמיקים בין תיאורי הנכסים עצמם, ולא מסתפקת במאפיינים מספריים כמו מחיר או מיקום.
מעבר לכך, העבודה כוללת גם בניית מודלים לחיזוי מחירי השכירות על בסיס מאפיינים שונים של הנכס והמארח, כגון מספר חדרים, מספר אמבטיות, דמי ניקיון ומיקום. המטרה הייתה לפתח מודלים שיכולים לנבא בצורה מדויקת את מחיר הנכס, וכך לספק למשתמשים כלי לחיזוי מחירים בצורה אמינה ומבוססת.

ביצוע עבודה זו יכול לעזור למשתמשים בפלטפורמה למצוא נכסים שמתאימים להם בצורה מדויקת יותר, ולמארחים להבין כיצד לתאר את הנכס שלהם כדי למשוך שוכרים פוטנציאליים. המערכת יכולה לשפר את חוויית המשתמש בצורה משמעותית בכך שהיא מציעה חיפושים מותאמים אישית יותר. מעבר לכך, הפיתוח של מערכות המלצה מבוססות טקסט יכול להיות מיושם בתחומים נוספים כמו חיפוש עבודה, קניות מקוונות ועוד, דבר שיכול לתרום לשיפור הממשקים וההתאמה האישית של שירותים רבים אחרים.

השערות ושיטות מחקר

במהלך המחקר, הצבתי מספר השערות מחקר שנועדו לבחון את ההשפעה של מאפייני נכסים שונים על המחיר, וכן את היכולת לחזות את המחיר בצורה מדויקת באמצעות מודלים שונים של למידת מכונה. להלן פירוט השערות המחקר המרכזיות, וכיצד ניגשתי לבדוק אותן בעזרת עיבוד הנתונים ופיתוח המודלים.

- **נכסים עם יותר חדרי שינה, מיטות ואמבטיות יובילו למחירים גבוהים יותר**

ניתוח הקשר בין מספר חדרי השינה, חדרי האמבטיה והמיטות לבין המחיר בוצע באמצעות ניתוח מתאם (Correlation Analysis), גרפי פיזור ועוד. נבדק האם יש מתאם חיובי בין המשתנים הכמותיים הללו לבין המחיר. בנוסף, מודלים כמו Decision Tree יכולים לספק מידע על חשיבות המשתנים הללו בתחזיות המחיר.

- **ישנם הבדלים משמעותיים במחירי הנכסים בין ערים שונות**

ניתוח סטטיסטי והשוואת ערים באמצעות גרפי קופסה.

מבחן סטטיסטי ANOVA (ניתוח שונות) כדי לבדוק האם ההבדלים במחירי הנכסים בין הערים הם מובהקים.

- **שכונה מסוימת יכולה להשפיע בצורה משמעותית על המחיר**

גרף המציג את מחירי הלוג הממוצעים של נכסים לפי שכונה. להמחיש בצורה ויזואלית את השונות במחירים בין השכונות ואת ההבדלים בין הערים. גרף עמודות מאפשר לראות בקלות את המחירים הגבוהים בכל שכונה ולהשוות בין ערים שונות.

- **נכסים עם מדיניות ביטול גמישה יהיו זולים יותר מנכסים עם מדיניות ביטול קפדנית**

השתמשתי בגרפי תיבות (box plots) על מנת להשוות בין מחירי הלוג הממוצעים של נכסים לפי סוגי מדיניות הביטול. בנוסף, ביצעתי ניתוח השוואתי במטרה לזהות הבדלים במחיר בין סוגי המדיניות.

- **למארחים עם תמונת פרופיל או אימות זהות תהיה השפעה על המחיר או על ציון הביקורות**

ניתוח השוואת ממוצעים. לבדוק האם יש הבדל ממוצע במחירי הנכסים ובציוני הביקורות בין מארחים עם תמונת פרופיל או אימות זהות לבין מארחים שאין להם.

- **נכסים שכוללים דמי ניקיון יהיו יקרים יותר, כיוון שהשירות הנוסף מתבטא במחיר**

השתמשתי בגרפי תיבות כדי להשוות את התפלגות המחירים בין נכסים שיש להם דמי ניקיון לבין נכסים ללא דמי ניקיון. גרף זה אפשר לראות אם יש הבדל ברור במחירים בין שתי הקבוצות, ואיך דמי הניקיון משפיעים על המחיר הממוצע של הנכסים.

- **תיאורי נכסים בערים מרכזיות יתמקדו במיקום ונגישות**

ניתוח התיאורים הטקסטואליים של נכסים בערים שונות באמצעות ענני מילים (word clouds) כדי לזהות מונחים רלוונטיים שמופיעים בתיאורים וקשורים לקרבה למרכז העיר.

ניתוח הטקסטואלי על עמודת תיאורי הנכנסים באמצעות TF-IDF (Term Frequency – Inverse Document Frequency), אשר נותן משקל גבוה למילים שמופיעות לעיתים קרובות בתיאורים אך לא בכל התיאורים.

נתונים

הנתונים בהם השתמשתי לצורך הפרויקט נלקחו מאתר Kaggle. הפרויקט מבוסס על מאגר נתוני Airbnb, הכולל מגוון רחב של מאפיינים על נכסים להשכרה לטווח קצר. בין המשתנים החשובים במסד הנתונים ניתן למנות את מחיר הנכס (בפורמט לוגריתמי), סוג הנכס (דירה, בית, וכו'), סוג החדר (חדר שלם, חדר פרטי, חדר משותף), ציוני המארחים והנכסים, וכן מידע גיאוגרפי (קו אורך וקו רוחב). כמו כן, קיימים מאפיינים הקשורים ליכולת האירוח של הנכס, כגון מספר המיטות, חדרי השינה, וחדרי האמבטיה.

(קישור: https://www.kaggle.com/datasets/lovishbansal123/airbnb-data/data)

מסד הנתונים כולל 29 עמודות ו- 74111 שורות. להלן פירוט העמודות וסוגי הנתונים:

```
RangeIndex: 74111 entries, 0 to 74110
Data columns (total 29 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   id                      74111 non-null  int64
 1   log_price               74111 non-null  float64
 2   property_type           74111 non-null  object
 3   room_type               74111 non-null  object
 4   amenities               74111 non-null  object
 5   accommodates            74111 non-null  int64
 6   bathrooms               73911 non-null  float64
 7   bed_type                74111 non-null  object
 8   cancellation_policy     74111 non-null  object
 9   cleaning_fee            74111 non-null  bool
 10  city                    74111 non-null  object
 11  description             74111 non-null  object
 12  first_review            58247 non-null  object
 13  host_has_profile_pic    73923 non-null  object
 14  host_identity_verified  73923 non-null  object
 15  host_response_rate      55812 non-null  object
 16  host_since              73923 non-null  object
 17  instant_bookable        74111 non-null  object
 18  last_review             58284 non-null  object
 19  latitude                74111 non-null  float64
 20  longitude               74111 non-null  float64
 21  name                    74111 non-null  object
 22  neighbourhood           67239 non-null  object
 23  number_of_reviews       74111 non-null  int64
 24  review_scores_rating    57389 non-null  float64
 25  thumbnail_url           65895 non-null  object
 26  zipcode                 73145 non-null  object
 27  bedrooms                74020 non-null  float64
 28  beds                    73980 non-null  float64
dtypes: bool(1), float64(7), int64(3), object(18)
```

פעולות ניקוי, השלמה ונרמול

בסט הנתונים קיימים ערכים חסרים במספר עמודות. להלן פירוט העמודות בהן יש ערכים חסרים וכמות הערכים החסרים בכל עמודה:

id	0
log_price	0
property_type	0
room_type	0
amenities	0
accommodates	0
bathrooms	200
bed_type	0
cancellation_policy	0
cleaning_fee	0
city	0
description	0
first_review	15864
host_has_profile_pic	188
host_identity_verified	188
host_response_rate	18299
host_since	188
instant_bookable	0
last_review	15827
latitude	0
longitude	0
name	0
neighbourhood	6872
number_of_reviews	0
review_scores_rating	16722
thumbnail_url	8216
zipcode	966
bedrooms	91
beds	131

השלמת ערכים חסרים:

- <u>ערכי neighbourhood</u> החסרים הושלמו בעזרת מודל KNN מבוסס מיקום גיאוגרפי (קו אורך ורוחב).

תחילה, המודל אומן על נכסים שבהם אין ערכים חסרים ב-neighbourhood, תוך שימוש בקואורדינטות גיאוגרפיות. לאחר מכן, המודל חזה את השכונה עבור הנכסים עם ערכים חסרים והשלימם על בסיס הקשרים הגיאוגרפיים.

- <u>host_since</u>

מכיוון שתאריך הביקורת הראשונה first_review מהווה עדות לכך שהמארח היה פעיל לפחות מאז אותו תאריך, השלמתי את הערכים החסרים בעמודת host_since על בסיס ערך עמודת first_review. כלומר, אם תאריך ה-host_since היה חסר, השתמשתי בתאריך הביקורת הראשונה להשלמה. לאחר ההשלמה, מחקתי 42 שורות שבהן עדיין לא היה ערך בעמודת host_since, שכן לא ניתן להשלים את המידע הנדרש במקרים אלו.

יצרתי משתנה חדש host_active_years המייצג את הניסיון של המארח (כמה זמן עבר מאז שהמארח החל לפעול), על ידי חישוב ההפרש בין תאריך הניתוח לתאריך ההתחלה.

- beds/bedrooms/bathroom

תחילה, חישבתי את מטריצת הקורלציה בין המשתנים הללו לבין כמות האורחים (accommodates) כדי להבין את הקשרים ביניהם ולוודא שישנם קשרים משמעותיים שיאפשרו השלמה מדויקת. את הערכים החסרים במשתנים אלו השלמתי באמצעות החציון של כל קבוצה. לדוגמה, עבור עמודת המיטות, הערכים החסרים הושלמו לפי החציון בקבוצת הנכסים שמכילה את אותו מספר אורחים, בעוד שבחדרי השינה הערכים הושלמו על פי החציון בקבוצות שמכילות את אותו מספר מיטות ואורחים. תהליך דומה בוצע להשלמת חדרי האמבטיה בהתבסס על מספר המיטות וחדרי השינה. גישה זו מבטיחה שהשלמת הערכים תתבצע בצורה שמבוססת על קשרים סטטיסטיים קיימים בין הנתונים, ובכך תשפר את דיוק המודל הסופי.

- review_scores_rating

ערכי הציון החסרים הושלמו בשתי רמות: ראשית, עבור נכסים שאין להם כלל ביקורות, הושלמו הציונים החסרים באמצעות חציון הציונים של נכסים ללא ביקורות. לאחר מכן, יתרת הערכים החסרים הושלמו באמצעות החציון הכללי של העמודה.

- host_has_profile_pic ו-host_identity_verified

עמודות קטגוריות כמו host_has_profile_pic ו-host_identity_verified הושלמו על ידי מילוי הערך החסר ב-"f" (שמתאר כי למארח אין תמונת פרופיל או שהזהות שלו לא אומתה).

המרת נתונים קטגוריים לערכים בוליאניים:

במסד הנתונים, ישנן מספר עמודות שהכילו ערכים טקסטואליים ("t" ו-"f") שציינו האם נכס או מארח עומדים בקריטריונים מסוימים. כדי להקל על הניתוח ולהשתמש בנתונים במודלים סטטיסטיים, ביצעתי המרה לערכים בוליאניים (True/False) לעמודות הבאות:

instant_bookable

host_has_profile_pic

host_identity_verified

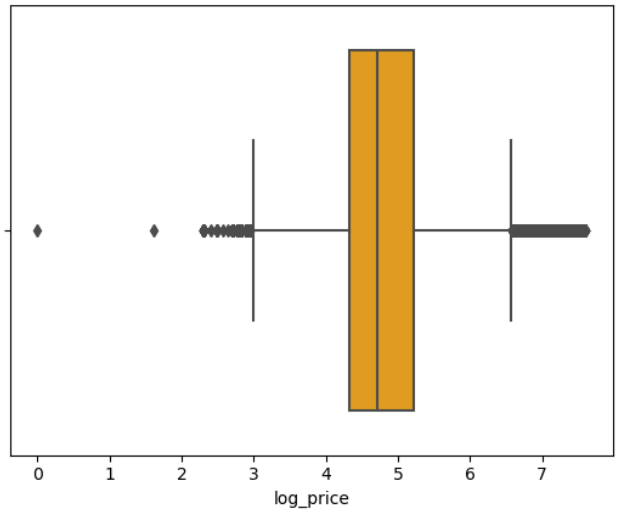
טיפול בעמודת amenities :

עמודת amenities הכילה רשימה של מתקנים בנכס, אשר היו מיוצגים כמחרוזת עם ערכים מוקפים בסוגריים מסולסלים. כדי להמיר את המידע הזה לרשימה שמישה ונקייה, נדרשה המרה של המחרוזת לרשימה של מתקנים. הפעולה כללה הסרה של הסוגריים והמרכאות, ופיצול כל מתקן לרשימה בודדת.

הסרת עמודות לא רלוונטיות:

הסרתי עמודות מסוימות שהן פחות רלוונטיות לניתוח (כגון 'id', 'zipcode', 'thumbnail_url', ותאריכים הקשורים לביקורות ראשונות ואחרונות) כדי לפשט את הדאטה ולהתמקד במאפיינים המרכזיים.

נתונים חריגים:



במהלך ניתוח הערכים החריגים במשתנה log_price נמצא כי קיימים 1,531 ערכים חריגים. הערכים שמעל הגבול העליון של 6.57, המהווים 1.85% מהנתונים, נחשבים כתקינים בהקשר של ניתוח המחירים, שכן מדובר בנכסים יוקרתיים וגדולים, שמחיריהם גבוהים מהרגיל אך אינם שגויים .לעומת זאת, נמצא ערך אחד חריג במיוחד של 0 ב-log_price, שהוא אינו תקין. מחיר של 0 אינו סביר לנכסים ב-Airbnb, ולכן ערך זה יש להסיר מהנתונים על מנת להבטיח ניתוח מדויק ולמנוע הטיה של התוצאות.

לאחר תהליך ההשלמה והניקוי, מסד הנתונים כעת מכיל 74068 שורות ו 23 עמודות.

```
Index: 74068 entries, 0 to 74068
Data columns (total 23 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   log_price               74068 non-null  float64
 1   property_type           74068 non-null  object
 2   room_type               74068 non-null  object
 3   amenities               74068 non-null  object
 4   accommodates            74068 non-null  int64
 5   bathrooms               74068 non-null  float64
 6   bed_type                74068 non-null  object
 7   cancellation_policy     74068 non-null  object
 8   cleaning_fee            74068 non-null  bool
 9   city                    74068 non-null  object
 10  description             74068 non-null  object
 11  host_has_profile_pic    74068 non-null  bool
 12  host_identity_verified  74068 non-null  bool
 13  instant_bookable        74068 non-null  bool
 14  latitude                74068 non-null  float64
 15  longitude               74068 non-null  float64
 16  name                    74068 non-null  object
 17  neighbourhood           74068 non-null  object
 18  number_of_reviews       74068 non-null  int64
 19  review_scores_rating    74068 non-null  float64
 20  bedrooms                74068 non-null  float64
 21  beds                    74068 non-null  float64
 22  host_active_years       74068 non-null  int32
dtypes: bool(4), float64(7), int32(1), int64(2), object(9)
```

ניתוח נתונים

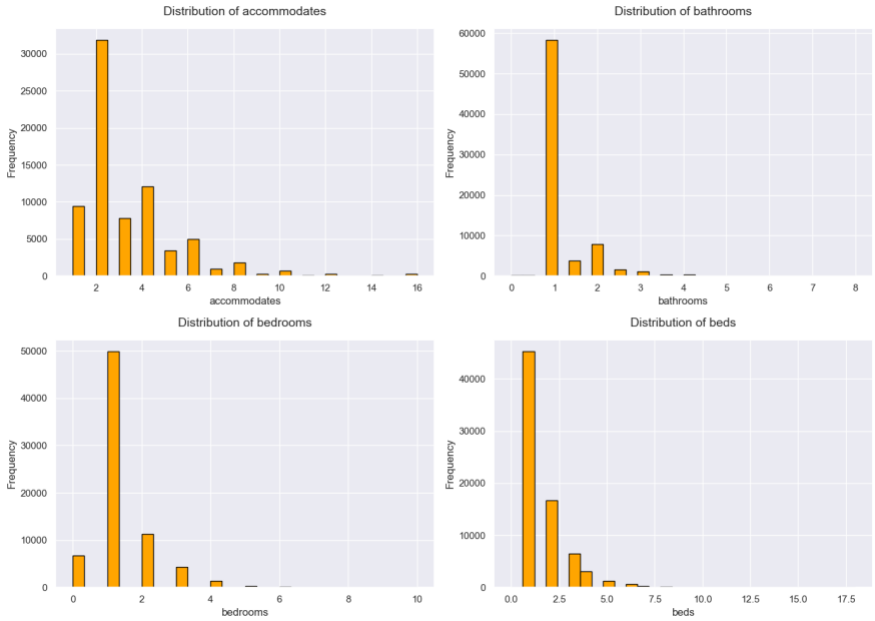
בפרויקט זה, ביצעתי ניתוח נתונים מקיף במטרה להבין את הגורמים שמשפיעים על מחיר הנכסים ב-Airbnb, לחזות את המחיר בצורה מדויקת, ולפתח מערכת המלצות לנכסים דומים. להלן התובנות המרכזיות

- **ההתפלגות של log_price**

במהלך הניתוח הראשוני של משתנה המחיר הלוגריתמי (log_price) בחנתי את התפלגותו כדי לקבל הבנה ראשונית על מבנה המחירים בסט הנתונים. התפלגות המחירים מציגה התפלגות הקרובה לנורמלית עם ממוצע של 4.78, חציון של 4.71, וסטיית תקן של 0.72. המחיר הנמוך ביותר בנתונים הוא 1.60 שזה מקביל למחיר של כ 5 דולר, בעוד שהמחיר הגבוה ביותר מגיע לערך לוגריתמי של 7.6, כלומר מחיר של כ 2000 דולר ללילה .

בנוסף, ערכי הרבעונים מראים כי 25% מהמחירים נמצאים מתחת לערך של 4.31, בעוד ש-75% מהמחירים נמצאים מתחת ל-5.22, מה שמעיד על פיזור רחב של המחירים.

- **התפלגות משתנים :**



מספר האנשים שהנכס יכול להכיל

רוב הנכסים יכולים להכיל בין 2 ל-4 אנשים, כשההתפלגות מציגה ירידה ברורה במספר הנכסים שיכולים לאכלס יותר אנשים. זה מתאים למציאות של הרבה דירות קטנות או דירות במרכזי ערים, שמיועדות לזוגות או קבוצות קטנות.

מספר חדרי האמבטיה:

כפי שניתן לצפות, רוב הנכסים כוללים חדר אמבטיה אחד בלבד, וזה מסביר את המספר הגבוה סביב 1 בחלוקה. נכסים עם יותר מ-2 חדרי אמבטיה נדירים יחסית, אך הם קיימים, וכנראה מדובר בנכסים יוקרתיים או גדולים יותר המיועדים לקבוצות גדולות.

מספר חדרי השינה

הגרף מראה כי נכסים עם חדר שינה אחד הם הנפוצים ביותר, מה שמסביר את העובדה שרוב הדירות בAirbnb מיועדות לזוגות או יחידים.

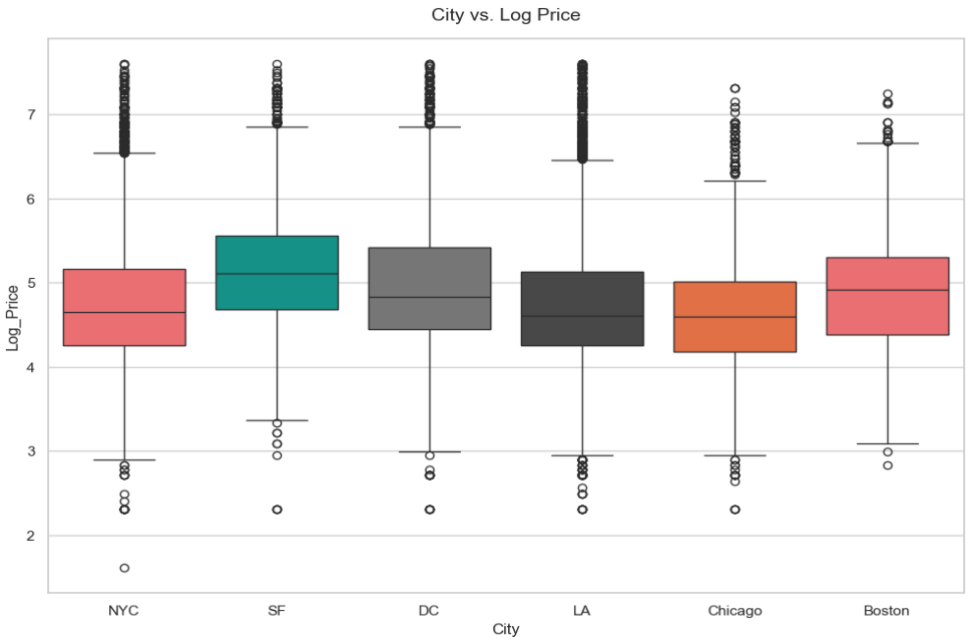
מספר המיטות

בדומה למספר חדרי השינה, מספר המיטות בנכסים הוא ברובו קטן, עם רוב הנכסים הכוללים מיטה אחת או שתיים. נכסים עם מספר רב של מיטות מיועדים לקבוצות גדולות יותר, אך אלה נדירים יחסית.

- **ניתוח השפעת הערים**

בהתבסס על הניתוח של כמות הנכסים ומחירי הלוג הממוצעים לפי ערים, מתקבלת תמונה מעניינת על שוק ה-Airbnb בערים המרכזיות בארצות הברית.

הערים המובילות מבחינת כמות הנכסים הן ניו יורק(43%) ולוס אנג'לס (30%) , מה שמצביע על ביקוש גבוה בשוק י הנדל"ן העירוניים הגדולים. סן פרנסיסקו (8.6%) ווישינגטון די סי (7.6%) גם מציגות מספר גבוה יחסית של נכסים , כאשר בדרך כלל אלה הן ערים יקרות, מה שמשפיע ישירות על המחיר הממוצע לנכס.

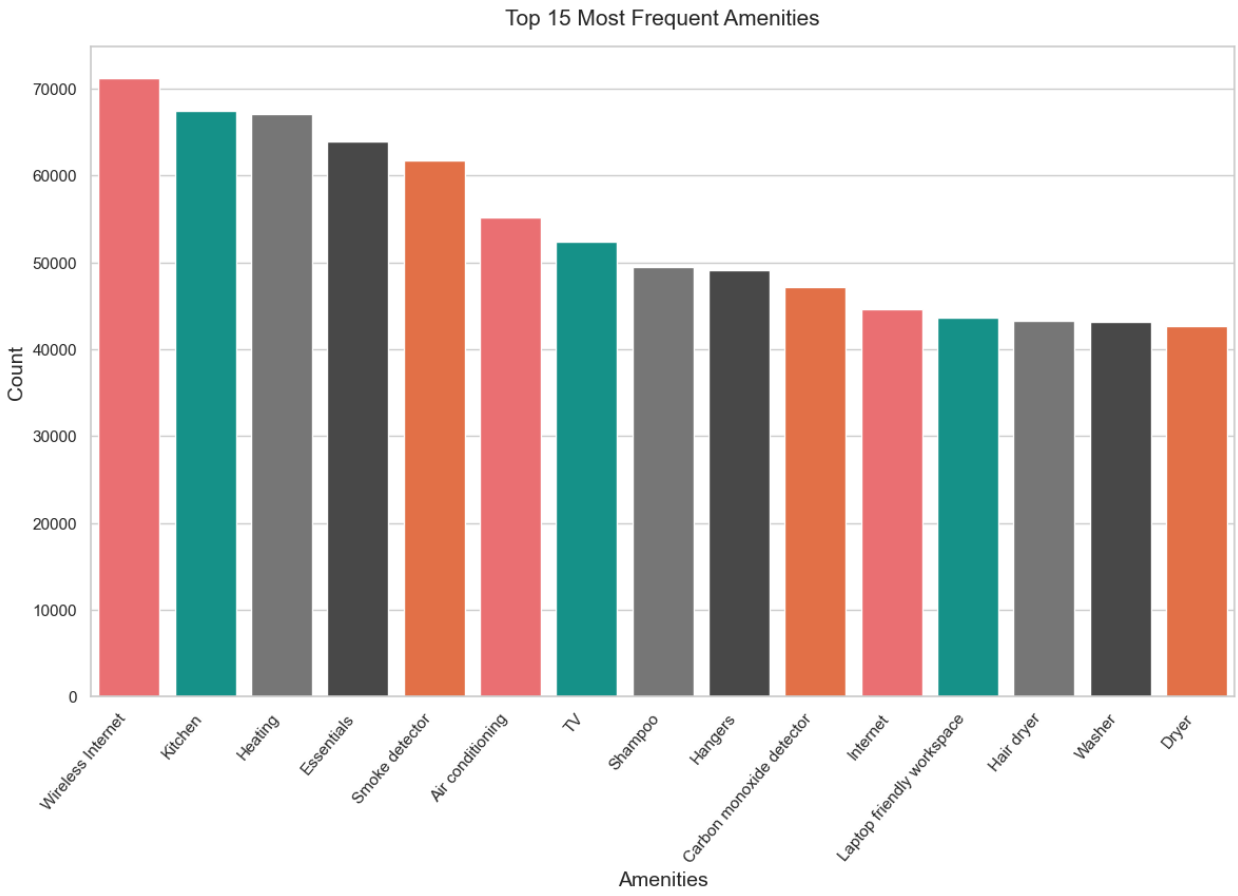


SF	5.170014
DC	4.986798
Boston	4.884035
LA	4.720492
NYC	4.719237
Chicago	4.620079

כשמסתכלים על מחירי הלוג הממוצעים, סן פרנסיסקו בולטת עם המחיר הגבוה ביותר (5.17), וושינגטון די.סי ובוסטון מציגות מחירים גבוהים גם הן, עם ערכי לוג של 4.99 ו-4.88 בהתאמה. לוס אנג'לס וניו יורק כמעט זהות במחירי הלוג הממוצעים שלהן, סביב 4.72, בעוד ששיקגו מציגה את המחיר הממוצע הנמוך ביותר מבין הערים, עם ערך של 4.62. הניתוח מראה שיש פערים משמעותיים הן בכמות הנכסים והן במחירים בין הערים השונות. ניו יורק ולוס אנג'לס מציגות את הכמות הגדולה ביותר של נכסים, בעוד סן פרנסיסקו מובילה במחירי הלוג הממוצעים.

כדי לבחון את השערה שלי שישנם הבדלים משמעותיים במחרי הכנסה בין ערים שונים, ביצעתי ניתוח ANOVA שהראה ערך F-statistic של 629.46 עם p-value של 0.0, מה שמצביע על כך שישנם הבדלים מובהקים סטטיסטית במחירי הנכסים בין הערים השונות.

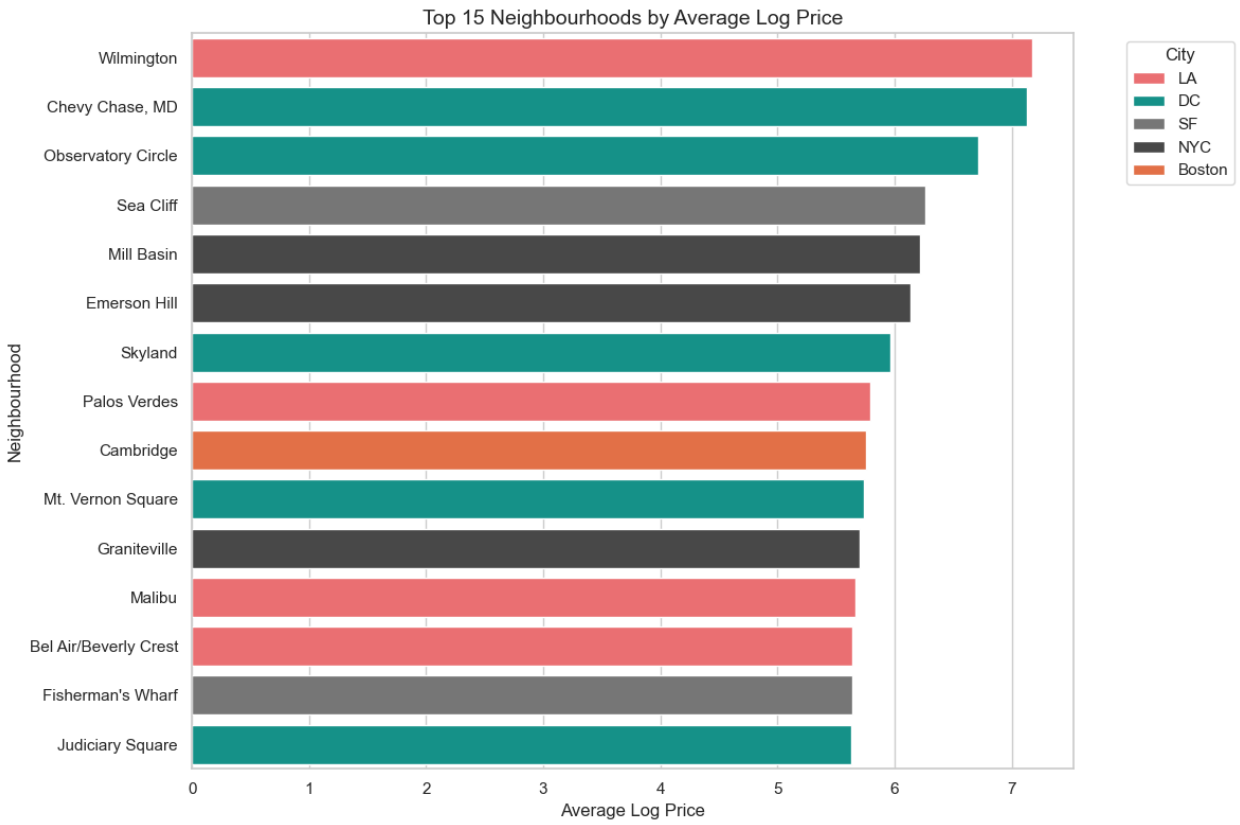
- **מתקנים**



הגרף המציג את טופ 15 המתקנים הנפוצים ביותר ב-Airbnb מספר סיפור ברור על הציפיות והצרכים של האורחים, כמו גם על הסטנדרטים שמארחים שואפים לספק. אינטרנט אלחוטי הוא המתקן הנפוץ ביותר, מה שמראה את חשיבות הקישוריות עבור תיירים ואנשי עסקים. מטבח מגיע למקום השני, דבר המצביע על הצורך ביכולת לבשל ולחסוך

בעלויות, במיוחד בשהיות ארוכות. מתקנים כמו חימום, מיזוג אוויר, וצרכים בסיסיים מדגישים את המחויבות לנוחות בסיסית עבור האורחים. בנוסף, גלאי עשן וגלאי פחמן חד-חמצני מראים כי בטיחות האורחים נמצאת בעדיפות גבוהה. טלוויזיה ושמפו הם מתקנים המספקים נוחות נוספת, בעוד מתלה בגדים נותן מענה לצורך בארגון חפצים. הגרף מציג תמונה של דאגה לפרטים הקטנים, שמשפיעים על חוויית השהייה הכוללת, וגורמים לנכס ב-Airbnb להיות אטרקטיבי יותר בעיני האורחים.

- **השכונות המובילות במחירי הלוג**



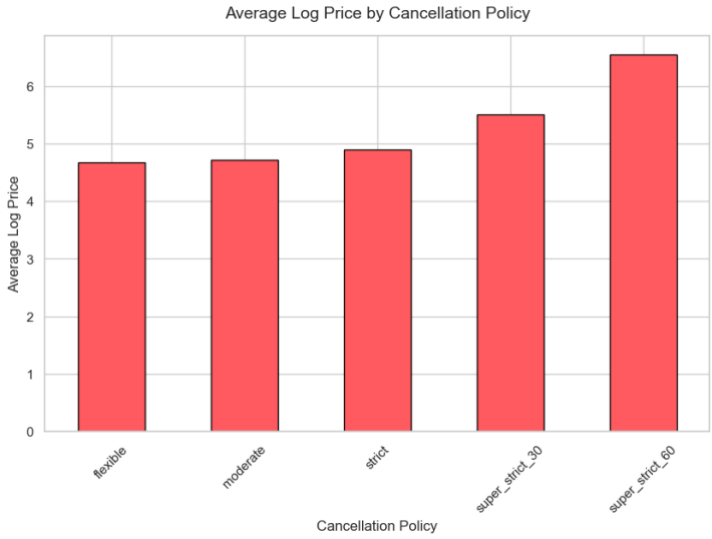
הגרף מציג את 15 השכונות עם מחירי הלוג הממוצעים הגבוהים ביותר, כאשר כל עמודה מייצגת שכונה מסוימת והצבעים מסמלים את העיר המתאימה. ניתן לראות כי השכונות Wilmington (לוס אנג'לס) Chevy Chase (וושינגטון), מובילות עם מחירי הלוג הממוצעים הגבוהים ביותר, סביב 6.5. לאחריהן, השכונות Observatory Circle (וושינגטון) ו Sea Cliff (סן פרנסיסקו) מציגות גם הן מחירים גבוהים מהממוצע.

מסקנה זו מראה כי שכונות מסוימות בערים כמו לוס אנג'לס, וושינגטון די.סי., וסן פרנסיסקו מכילות נכסים בעלי ערך גבוה מאוד. המשמעות היא שישנה השפעה משמעותית של השכונה על מחירי הנכסים. השערה זו מתחזקת לנוכח

המחירים הגבוהים המופיעים בעיקר בערים כמו לוס אנג'לס, וושינגטון, וסן פרנסיסקו, בהן ישנן שכונות יוקרתיות במיוחד.

- **קשר בין סוגי מדיניות הביטול לבין log_price**

הגרף מציג את הקשר בין סוגי מדיניות הביטול לבין המחיר הלוג של נכסים ב Airbnb. לפי הנתונים, נכסים עם מדיניות ביטול super_strict_60 ו -super_strict_30 מציגים את המחירים הגבוהים ביותר, בעוד נכסים עם מדיניות Flexible מתומחרים נמוך יותר.



בהשערה שלי, חשבתי שמדיניות ביטול מחמירה תגרום למחירים נמוכים יותר בגלל החשש של האורחים מאובדן כספי במקרה של ביטול. עם זאת, הממצאים דווקא מראים שנכסים יקרים נוטים להציע מדיניות מחמירה כדי להגן על עצמם מפני ביטולים של הרגע האחרון.

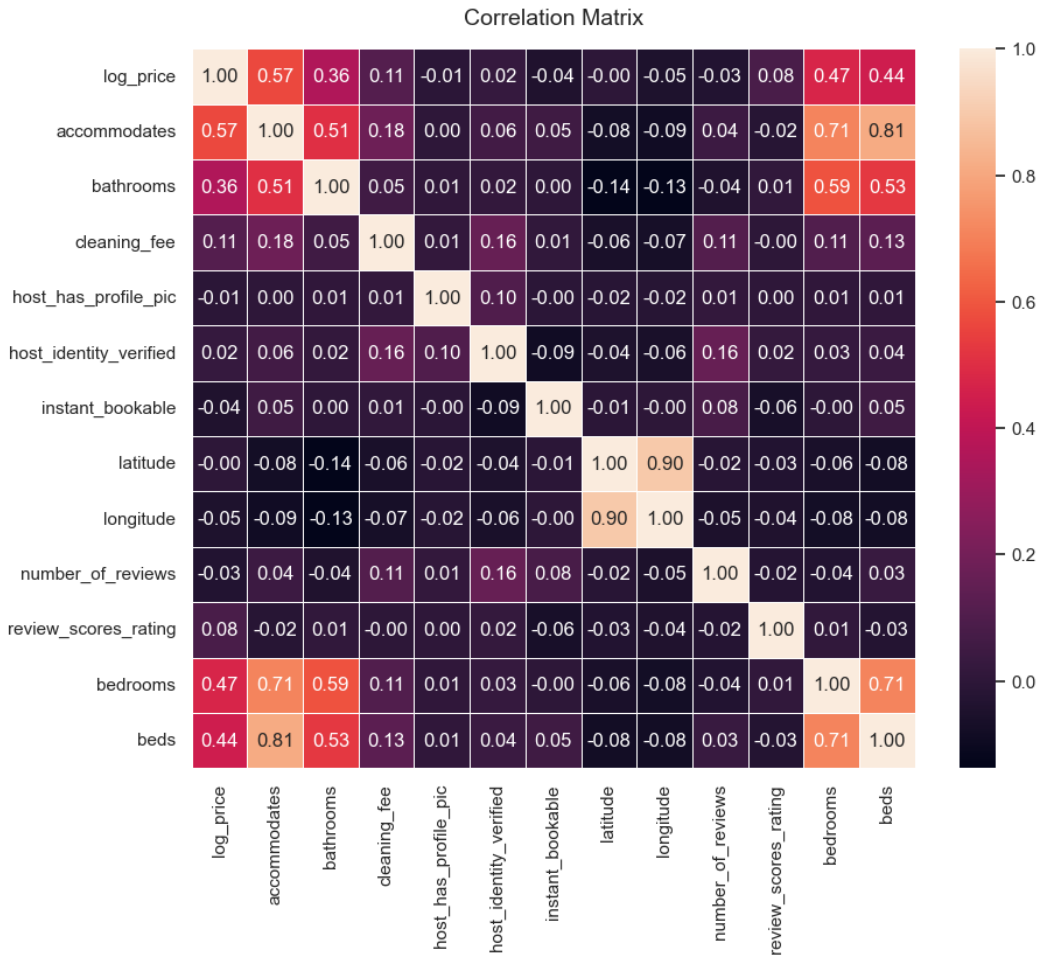
הקשר בין מדיניות הביטול למחיר קשור גם להשפעה שלה על שיעורי ההזמנות. מדיניות גמישה מושכת אורחים שלא בטוחים בתוכניות שלהם, כי היא מאפשרת ביטול ללא קנס ומעלה את שיעור ההזמנות. לעומת זאת, מדיניות מחמירה כמו super_strict_60 עשויה להרתיע אורחים שחוששים מביטול והפסד כספי, אך היא שומרת על יציבות ההזמנות בנכסים יוקרתיים שמכוונים לקהל שמוכן לשלם מראש ולהתחייב לשהות.

- **מאפייני המארח**

ההשערה הייתה שלמארחים עם תמונת פרופיל או אימות זהות תהיה השפעה על המחיר או על ציון הביקורות, מתוך ההנחה שמארחים שנראים אמינים יותר יוכלו לדרוש מחירים גבוהים יותר ויקבלו ציוני ביקורות טובים יותר. אולם, מהנתונים עולה כי ההשפעה של מאפיינים אלו היא קטנה מאוד, ואף זניחה.

לפי הניתוח, ההבדל הממוצע במחירי הנכסים בין מארחים עם תמונת פרופיל לבין מארחים ללא תמונת פרופיל הוא כמעט אפסי. מארחים עם תמונת פרופיל דורשים בממוצע מחיר של 4.78 בלוג (log price), בעוד שמארחים ללא תמונת פרופיל דורשים בממוצע 4.85. גם לגבי ציוני הביקורות, ההבדל הממוצע בין מארחים עם תמונת פרופיל (94.5) לבין אלו ללא תמונה (94.3) הוא כמעט לא קיים. ההבדלים בין מארחים מאומתים לכאלו שאינם מאומתים גם הם קטנים מאוד. המחיר הממוצע של מארחים עם אימות זהות עומד על 4.79, לעומת 4.76 למארחים ללא אימות זהות. ההבדל הקטן הזה מעיד שאימות זהות גם הוא לא מהווה גורם משמעותי בקביעת המחיר.

- **מטריצת הקורלציה**

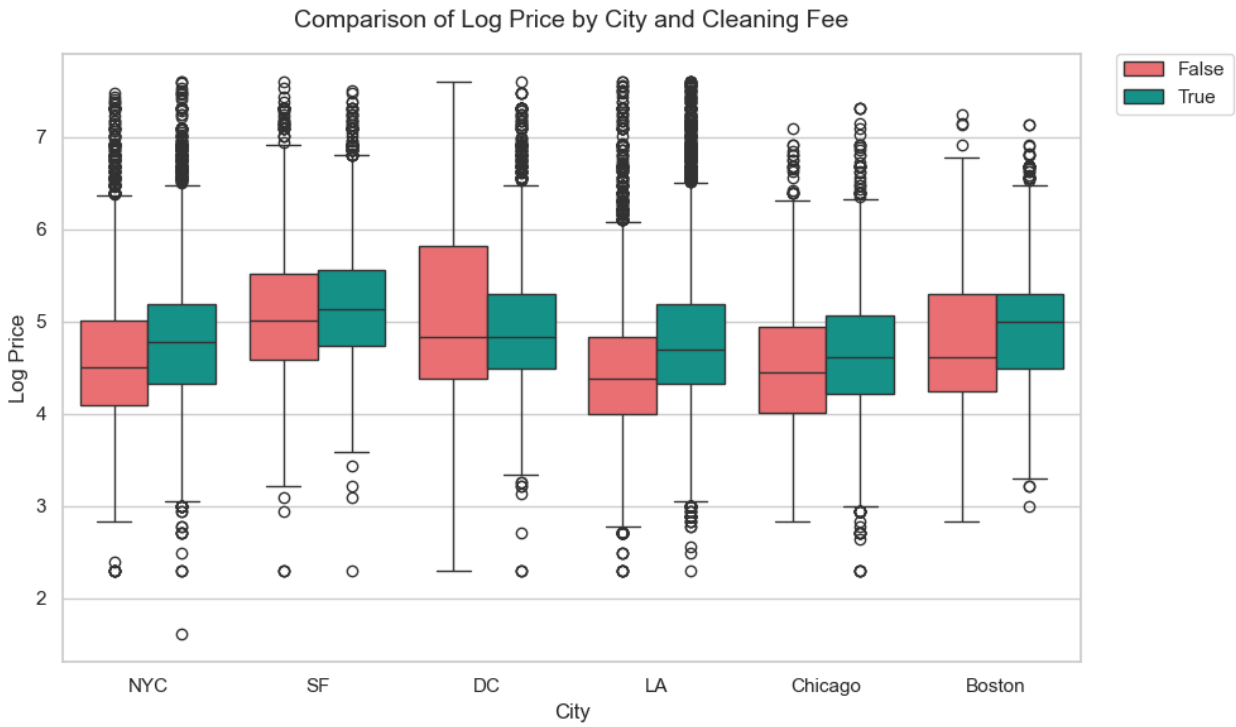


מטריצת הקורלציה חושפת כמה קשרים מעניינים בין התכונות השונות של הנכסים ב Airbnb . ראשית, ניתן לראות שמספר האנשים שהנכס יכול להכיל (accommodates) הוא הגורם המשפיע ביותר על המחיר (log_price) עם מתאם חיובי של 0.57. ככל שהנכס יכול לארח יותר אורחים, כך המחיר עולה. קשרים חזקים נוספים נמצאו עם מספר חדרי השינה (0.47) ומספר המיטות (0.44) מה שמדגיש את חשיבות גודל הנכס בתמחור.

בנוסף, גם מספר חדרי השירותים מציג מתאם חיובי (0.36) עם המחיר, מה שמרמז על כך שנכסים עם יותר שירותים מתומחרים גבוה יותר, שכן הם מציעים יותר נוחות למספר גדול של אורחים.

לעומת זאת, תכונות כמו האפשרות להזמנה מיידית (instant_bookable) והדירוגים (review_scores_rating) משפיעות הרבה פחות על המחיר. למשל, המתאם בין log_price לinstant_bookable- הוא שלילי וחלש (-0.04), מה שמעיד על כך שהאפשרות להזמנה מיידית אינה משפיעה משמעותית על המחיר. גם הדירוגים מראים קשר חלש מאוד (0.08), מה שמצביע על כך שהמחיר לא בהכרח נקבע על סמך איכות החוויה שדירגו האורחים.

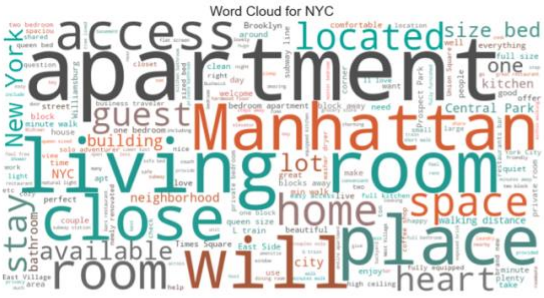
- **השפעת דמי הניקיון על מחירי הנכסים**

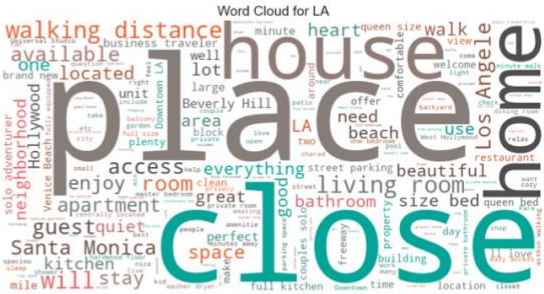


הגרף מציג תמונה ברורה על השפעת דמי הניקיון על מחירי הנכסים בפלטפורמת Airbnb בערים שונות. הגרף מאשר את ההשערה שנכסים שכוללים דמי ניקיון יהיו יקרים יותר. ניתן לראות שבכל אחת מהערים, התפלגות המחירים עבור נכסים עם דמי ניקיון (True) גבוהה יותר ברוב המקרים בהשוואה לנכסים ללא דמי ניקיון (False). בעיקר בערים כמו ניו יורק ולוס אנג'לס, נכסים עם דמי ניקיון נוטים להראות מחירים גבוהים יותר. בניגוד לערים אחרות, בוושינגטון די.סי. המחיר החציוני של נכסים ללא דמי ניקיון דומה מאוד למחיר החציוני של נכסים עם דמי ניקיון. זה מראה שבוושינגטון אין השפעה משמעותית לדמי ניקיון על המחיר.

- **ענן המילים**

תיאורי נכסים מהווים כלי חשוב עבור מארחים ב-Airbnb כדי למשוך שוכרים פוטנציאליים. בעזרת ניתוח ענן המילים עבור תיאורים אלו, ניתן לזהות את הנושאים המרכזיים שמארחים בוחרים להדגיש בערים שונות ואת השפעת המיקום על המחיר הנתפס של הנכס.





בגרף ענן המילים המצורף, ניתן לראות אילו מילים חוזרות בתיאורי הנכסים בערים שונות:

ניו יורק (NYC): בעיר ניו יורק, בולטות מילים כמו "room", "Manhattan", "apartment", ו-"close". המונח "Manhattan" מהווה נקודת מפתח בתיאורים, ומדגיש את הקרבה של הנכסים למרכז הכלכלי והתרבותי של העיר. "Apartment" מצביע על סוג הנכסים השכיח ביותר בניו יורק – דירות קטנות וקומפקטיות. נראה שהמארחים בניו יורק שמים דגש על הנגישות (close) והקרבה למרכזי העיר הגדולה, כמו גם על סוג הנכס (room, apartment). בניו יורק, היכן שהמיקום נחשב קריטי והשוק הנדל"ני יקר וצפוף, חשוב למארחים להבליט את היתרון היחסי של הקרבה לאזורים מרכזיים בעיר.

סן פרנסיסקו (SF): בסן פרנסיסקו, בולטים מונחים כמו "place", "kitchen", "Golden Gate" ו "house". בניגוד לניו יורק, התיאורים בסן פרנסיסקו כוללים גם דגש על מאפייני הנכס עצמו, כגון מטבח (kitchen) ומרחב (house) זה מצביע על כך שבסן פרנסיסקו, שוק הדיור יותר מגוון וכולל נכסים עם מאפייני מגורים איכותיים. תיאורים אלו מבליטים את היתרונות של מרחב פנימי נוח ונגישות טובה. כמו כן, המונח "Golden Gate" מייצג את

האטרקציות המקומיות שמהוות פקטור חשוב בשיווק הנכס. שוכרים עשויים להיות מעוניינים בנכסים עם נוף לאטרקציות מרכזיות כמו גשר הזהב, מה שמוסיף ערך לנכס. בסן פרנסיסקו, אם כן, ישנה גישה שמשלבת בין מיקום לאטרקציות לבין נוחות ואיכות חיי היומיום בנכס.

וושינגטון די סי(DC): בוושינגטון, בולטות המילים "DC", "apartment", ו "place" בתיאורים רואים דגש על מיקום (DC), אך פחות ניכר דגש על קרבה לתחבורה או אטרקציות ספציפיות, כפי שנראה בניו יורק או בסן פרנסיסקו. התיאורים בוושינגטון מדגישים יותר את סוג הנכס (apartment) ואת הנוחות הכללית של המיקום (place) . ניתן להבין שהשוכרים בוושינגטון, בעיקר שוכרים לתקופות קצרות לצורכי עסקים או ממשל, מחפשים קרבה כללית למרכזי עבודה ולפעילות בעיר.

לוס אנג'לס(LA) : בלוס אנג'לס מופיעות מילים כמו "house", "place" ו "close"-העיר, המוכרת במבנה הפיזי המפוזר שלה, מתאפיינת בנכסים עם דגש על מרחב, כפי שניתן לראות בשימוש במונח "house" .כמו כן, ניתן לראות שימוש במילה "close" שמדגישה את החשיבות של קרבה לאטרקציות או לאזורים מרכזיים בתוך העיר המפוזרת. לא נראית התמקדות בתחבורה ציבורית, אלא יותר בנכסים מרווחים ונוחים.

שיקגו ובוסטון: בערים אלו בולטות המילים "room", "apartment", ו-"close", בדומה לניו יורק. עם זאת, התיאורים לא כוללים אזכור של אטרקציות מרכזיות, אלא מתמקדים בסוג הנכס (apartment) והקרבה (close). זה מרמז שהמארחים בערים אלו מתמקדים יותר בנוחות הפיזית של המגורים והנגישות למרכזי הערים, ופחות באטרקציות תיירותיות ספציפיות.

הניתוח של תיאורי הנכסים בערים השונות מראה דפוסים ברורים שמשקפים את אופי השוק המקומי והעדפות השוכרים. בערים כמו ניו יורק וסן פרנסיסקו, הקרבה למרכז העיר ולאטרקציות תיירותיות מהווה גורם מרכזי בתיאורי הנכסים, מה שמרמז על חשיבות המיקום במחיר' מאשר את ההשערה שתיאורי נכסים מתמקדים במיקום ובנגישות. לעומת זאת, בערים כמו לוס אנג'לס ושיקגו, התיאורים מתמקדים יותר במאפיינים הפנימיים של הנכס, כמו גודל הבית והנוחות, ופחות בנגישות או במיקום גיאוגרפי.

מערכת המלצות

בפרויקט זה החלטתי לפתח מערכת המלצות לנכסי Airbnb המבוססת על ניתוח התיאורים הטקסטואליים של הנכסים. המערכת שלי נועדה לסייע למשתמשים למצוא נכסים דומים לנכס מסוים בו הם מתעניינים.

בשלב הראשון יצרתי מטריצת TF-IDF מתיאורי הנכסים. מטריצת TF-IDF הופכת את התיאורים לוקטורים מספריים המייצגים את שכיחות המונחים השונים בכל תיאור. בנוסף, היא מקצה משקל נמוך למילים נפוצות מאוד, ומשקל גבוה יותר למילים שמופיעות במינון נכון – כלומר, מונחים שמייחדים כל תיאור.

עשיתי הגבלת מספר המונחים ל-5000 מכיוון שמספר גדול מדי של מונחים גרם לבעיות זיכרון והכביד על החישוב.
בנוסף, הגדרתי את הפרמטרים min_df ו-max_df כדי לסנן מונחים שמופיעים לעיתים נדירות מדי (כדי לא להכליל מונחים לא רלוונטיים) או לעיתים קרובות מדי (כדי לא להכליל מילים נפוצות מדי).

בשלב השני, חישבתי את הדמיון הקוסינוסי בין כל זוג תיאורי נכסים, על סמך וקטורי ה-TF-IDF שחושבו קודם לכן.
מכיוון שחישוב הדמיון על כל המטריצה בבת אחת גרם לבעיות זיכרון, ביצעתי את החישוב בבלוקים של 1000 שורות ושמרתי את התוצאות בקובץ HDF5. השתמשתי ב-HDF5 מכיוון שהוא מאפשר גישה מהירה לנתונים מבלי להחזיק את כל המידע בזיכרון בבת אחת.
לאחר חישוב הדמיון הקוסינוסי, בנייתי פונקציית המלצות שמקבלת שם של נכס ומחזירה רשימה של נכסים דומים לו.
הפונקציה משתמשת במטריצת הדמיון הקוסינוסי כדי למצוא את הנכסים הקרובים ביותר לנכס הנתון.
בשלב האחרון, ניגשתי למטריצת הדמיון הקוסינוסי מתוך קובץ HDF5 וביצעתי את ההמלצות בפועל.

THE LIBRARY LOUNGE	**SUNLIGHT SPECIAL** (recommendation)
Cozy room in my big private house in Brooklyn. The room is smaller but guests have access to all of the common areas in the house. Its is great if your visiting the city and want some privacy and peacefulness in your New York travels. Check in time is flexible please contact me if you are interested in long term rental Guest have acces to the room and all the common areas including the back yard We can chat sometimes and i will tell you more about the neighborhood other than that you are on your own Jamaican / West Indian neighborhood amazing food, cool people, irie (URL HIDDEN) entire street is private houses. Train stop: Utica Avenue & Sutter Avenue - 8minute walk or 3 minute bus ride, the bus that drops you from the train station runs very often so it really is a quick ride to	This sun filled room has a queen size bed, and a love seat. There is plenty of storage space, privacy and an awesome artsy vibe. There is even a space to set up a work desk if you need to. There is a bus stop 1 block away, Reach manhattan in 30 mins. The space is comfortable and quite, gets amazing sunlight in the daytime and has a huge queen size bed. It also has love seat. We have a backyard, full kitchen,full bathroom and a half bath downstairs. guest are free to enjoy them all. We can chat about things to do in the city and the neighborhood, but mostly you will be on your own to explore nyc. Jamaican / West Indian neighborhood amazing food, cool people, irie (URL HIDDEN) entire street is private houses. Train stop: Utica Avenue & Sutter Avenue - 8minute walk or 3 minute bus

the manhattan. the bus stop is one block away'.	ride, the bus that drops you from the train station runs very often so it really is a quick ride to the manhattan. the bus stop is one block away'.

לפי התיאורים, הנכסים נראים דומים מאוד מבחינת החוויה הכוללת שמוצעת לאורחים. הדמיון בין הנכסים נוגע למאפייני הנכס והשכונה, תחבורה נוחה, גישה למרחבים משותפים, וחוויית אירוח נוחה ופרטית. כל אלו מרכיבים שמייחדים את הנכסים ויכולים להצדיק את תוצאות ההמלצה של המערכת שלך.

המערכת פעלה בצורה טובה בזיהוי נכסים דומים על בסיס התיאורים הטקסטואליים, כאשר היא הצליחה לשים דגש על מילות מפתח ומאפיינים חשובים בתיאורים

אימון מודלי למידת מכונה והשוות ביצועיהם

במסגרת פרויקט חיזוי המחיר ב-Airbnb, השתמשתי במודלים שונים על מנת לנבא את המחיר בלוגריתם (log price) של הנכסים. תהליך העבודה כלל מספר שלבים חשובים שכללו הכנת הנתונים, שימוש במודלים מתקדמים ללמידת מכונה, ובחינת הביצועים שלהם.

הכנת הנתונים:

- **הסרת עמודות לא רלוונטיות :**לאחר חקירת הנתונים, בחרתי למחוק עמודות מסוימות שאינן תורמות לחיזוי המחיר, עמודות כמו description, host_has_profile_pic, amenities, host_identity_verified, latitude, longitude, ו name
- **קידוד משתנים קטגוריאליים :**ביצעתי קידוד למשתנים קטגוריים באמצעות LabelEncoder
- **נרמול :**נרמלתי את הנתונים באמצעות StandardScaler כדי להבטיח שהמודלים יקבלו נתונים בסקאלה אחידה. הנרמול המיר כל עמודה לטווח ערכים סטנדרטי, עם ממוצע 0 וסטיית תקן 1.
- **חלוקת הנתונים :**הנתונים חולקו ל-80% סט אימון ו-20% סט בדיקה באמצעות train_test_split.

בכל אחד מהמודלים נבחנתי שני מדדים עיקריים:

- **MSE** – (Mean Squared Error) מדד שמודד את ממוצע השגיאות המרובעות בין הערכים החזויים לערכים האמיתיים.
- **R^2** –מדד שמראה את מידת הדיוק של המודל בהסבר הנתונים, כאשר 1 משמעו דיוק מלא.

חיזוי בעזרת מודלים שונים:

1. **Decision Tree Regressor**

השתמשתי במודל עץ החלטה עם עומק מוגבל (max_depth=10) ומינימום דוגמאות עלה (min_samples_leaf=5). את הפרמטרים הללו בחרתי לעץ החלטות באופן ניסיוני.

תוצאות המודל:

MSE test: 0.21200162669869566

R2: 0.5834441944269175

ביצועים בינוניים יחסית, מה שמראה שהמודל מתקשה להתמודד עם מורכבות הנתונים.

לאחר הכשרת המודל, חישבתי את החשיבות היחסית של כל משתנה והשפעתו על חיזוי המחיר. התוצאות שקיבלתי מצביעות על כך שסוג החדר (0.58) וחדרי האמבטיה (0.18) הם גורמים יותר משמעותיים בחיזוי המחיר, בעוד שמספר האנשים שהנכס יכול להכיל תורם פחות לחיזוי המחיר(0.03) מכפי שנצפה בהשערות הראשוניות.

2. RandomForest Regressor

בחרתי לאמן מודל זה מכיוון שהוא משלב תוצאות של מספר עצים ומקטין את השגיאה. יישמת יער רנדומלי עם 100 עצים. המודל הצליח להשיג:

MSE: 0.18528654229638197

R^2: 0.635935883653315

השיפור בביצועים בהשוואה לעץ החלטה יחיד מצביע על כך שיער רנדומלי מצליח לחזות טוב יותר את המחיר, בזכות השימוש בהרבה עצים המקבלים החלטות מבוזרות.

3. CatBoost Regressor

בשלב הבא, השתמשתי בCatBoost מודל מתקדם במיוחד שמתמודד טוב עם נתונים קטגוריאליים:

הפרמטרים כללו 600 איטרציות, קצב למידה של 0.1 ועומק של 8.

המודל הצליח להשיג תוצאה:

MSE: 0.15738962889667724

R2: 0.6907497141657204

זה מצביע על כך שהמודל הצליח לחזות את המחירים בצורה מדויקת יותר בהשוואה ליערות הרנדומליים ולעצי ההחלטה.

4. XGBoost Regressor

לבסוף, ביצעתי חיזוי בעזרת XGBoost לאחר אופטימיזציה של הפרמטרים באמצעות GridSearchCV.

האופטימיזציה הצביעה על הערכים הטובים ביותר: עומק 5, 1000 איטרציות וקצב למידה של 0.069.

תוצאות המודל:

MSE: 0.1579446352483942

R^2: 0.6896591983921494

הביצועים של ה XGBoost-היו דומים מאוד לאלו של CatBoost עם דיוק גבוה בחיזוי.

השוואת ביצועים:

מהתוצאות עולה שהמודלים המורכבים יותר, כמו CatBoost ו XGBoost מצליחים יותר במתן תחזיות מדויקות, עם R2 גבוה יותר ו MSE נמוך יותר בהשוואה למודלים פשוטים יותר

Model	MSE	R2
DecisionTree Regressor	0.212	0.58
RandomForest Regressor	0.185	0.635
CatBoost Regressor	0.157	0.690
XGBoost Regressor	0.157	0.689

למרות שמודל CatBoost ו-XGBoost הגיעו לתוצאות קרובות מאוד מבחינת הביצועים, נראה כי CatBoost מעט עדיף מבחינת דיוק החיזוי. בנוסף, המודלים המשתמשים ב Boosting מציגים ביצועים טובים יותר מאשר עץ החלטה או יער רנדומלי, מה שמעיד על החשיבות של שימוש במודלים חזקים יותר כאשר יש צורך בחיזוי מחירים מדויק.

דיון ומסקנות

המחקר שבוצע בפרויקט זה נועד לבחון את הגורמים המשפיעים על מחירי נכסי Airbnb, לפתח מערכת המלצות מבוססת ניתוח טקסט, ולהשוות בין ביצועי מודלים שונים לחיזוי מחירי הנכסים. תוצאות המחקר מאפשרות להסיק מספר תובנות מעניינות על תחום השכרת הנכסים לטווח קצר ומספקות כלים מועילים הן למשתמשים בפלטפורמה והן למארחים.

האם ההשערות עמדו במבחן?

נכסים עם יותר חדרי שינה, מיטות ואמבטיות יובילו למחירים גבוהים יותר: הממצאים מראים כי ישנה קורלציה חיובית בין מספר חדרי השינה, המיטות, והאמבטיות לבין המחיר, אך משתנה אחר התגלה כמשפיע מרכזי על המחיר: סוג החדר (room_type). לפי ניתוחי המודלים (**Decision Tree Regressor**), סוג החדר היה המשתנה המשפיע ביותר, ואחריו מספר חדרי האמבטיה. השערה זו אוששה באופן חלקי, אך ישנם מאפיינים נוספים המשפיעים על המחיר בצורה חזקה יותר ממה שציפיתי.

שנם הבדלים משמעותיים במחירי הנכסים בין ערים שונות: השערה זו אושרה במובהק בעזרת מבחן ANOVA, שהראה הבדל מובהק סטטיסטית במחירי הנכסים בין הערים השונות. כמו שנראה מהניתוחים, סן פרנסיסקו מציגה את

מחירי הלוג הגבוהים ביותר, ואחריה וושינגטון די.סי. ובוסטון. הערים ניו יורק ולוס אנג'לס היו קרובות במחירים, בעוד ששיקגו הציגה את המחירים הנמוכים ביותר. התוצאות מראות בבירור שישנם פערים במחירי הנכסים בין הערים, מה שמצביע על כך ששוק השכירות לטווח קצר מושפע במידה ניכרת מהמיקום הגיאוגרפי של הנכס.

שכונה מסוימת יכולה להשפיע בצורה משמעותית על המחיר: גם כאן, ההשערה אושרה. גרף מחירי הלוג הממוצעים לפי שכונה מראה הבדלים ניכרים בין שכונות יוקרתיות בשווקים כמו לוס אנג'לס, וושינגטון וסן פרנסיסקו. שכונות יוקרתיות בערים אלו הציגו מחירים גבוהים באופן ניכר משכונות אחרות, מה שמצביע על כך שהשכונה אכן משחקת תפקיד חשוב בקביעת המחיר, בדומה למיקום העירוני.

נכסים עם מדיניות ביטול גמישה יהיו זולים יותר מנכסים עם מדיניות ביטול קפדנית: ההשערה אושרה במידה מסוימת, אם כי ישנם יוצאים מן הכלל. גרפי הקופסה מראים שנכסים עם מדיניות ביטול גמישה מתומחרים בממוצע נמוך יותר מנכסים עם מדיניות קפדנית. נכסים עם מדיניות super_strict_30 ו-super_strict_60 הציגו את המחירים הגבוהים ביותר, מה שמעיד על כך שנכסים יוקרתיים נוטים להציע מדיניות ביטול מחמירה כדי להגן על עצמם. בניגוד להנחתי, נראה שמדיניות ביטול גמישה לא תמיד תורמת להורדת המחיר, במיוחד בנכסים יוקרתיים.

למארחים עם תמונת פרופיל או אימות זהות תהיה השפעה על המחיר או על ציון הביקורות: ההשערה לא אושרה. מהנתונים עלה כי ההשפעה של תמונת פרופיל ואימות זהות על המחיר או על ציוני הביקורות היא מינורית ואף זניחה. ניתוח השוואת הממוצעים הראה שאין הבדל משמעותי במחיר בין מארחים עם או בלי תמונת פרופיל או אימות זהות.

נכסים שכוללים דמי ניקיון יהיו יקרים יותר: ההשערה אוששה. ניתוח גרפי התיבות מראה בבירור שנכסים עם דמי ניקיון מתומחרים בממוצע גבוה יותר מנכסים ללא דמי ניקיון. בערים כמו ניו יורק ולוס אנג'לס הפער היה משמעותי, מה שמחזק את ההשערה שדמי ניקיון מתבטאים במחיר הנכס.

תיאורי נכסים בערים מרכזיות יתמקדו במיקום ונגישות: ההשערה אושרה בחלקה. ניתוח ענני המילים הצביע על כך שבניו יורק וסן פרנסיסקו תיאורי הנכסים מתמקדים יותר בקרבה לאטרקציות מרכזיות ולמרכזי העיר, ואילו בערים כמו לוס אנג'לס ושיקגו ישנו דגש רב יותר על מאפייני הנכס הפיזיים כמו מרחב הבית והנוחות, ופחות על מיקום גיאוגרפי ונגישות תחבורתית.

במהלך העבודה אושרו חלק מההשערות, בעוד שאחרות נדחו או אושרו חלקית. הניתוח מדגיש את המורכבות בשוק ה-Airbnb, כאשר מיקום הנכס, גודלו, סוגו ומדיניות הביטול מתבררים כגורמים מרכזיים המשפיעים על המחיר. לעומת זאת, מאפיינים כמו תמונת פרופיל ואימות זהות של המארחים התגלו כמשתנים פחות משמעותיים מכפי שציפיתי בתחילת המחקר.

ההמלצה העיקרית למארחים היא להשקיע בתיאור מפורט וברור של הנכס, במיוחד בערים מרכזיות עם תחרות גבוהה. עבור משתמשי הפלטפורמה, מערכת ההמלצות שפיתחתי תוכל לסייע להם למצוא נכסים דומים בהתבסס על העדפותיהם האישיות ולשפר את חוויית החיפוש ב-Airbnb.

המודלים המתקדמים של למידת מכונה, כמו CatBoost ו-XGBoost, הצליחו לספק תחזיות מדויקות יחסית של המחיר, ומערכת ההמלצות המבוססת על ניתוח טקסטואלי הראתה תוצאות מועילות. הרחבה עתידית למחקר עשויה לכלול שיפור דיוק התחזיות על ידי שימוש במודלים מתקדמים יותר או שילוב משתנים נוספים, כמו מגמות עונתיות ודפוסי ביקוש, וכן שיפור מערכת ההמלצות באמצעות נתוני קונטקסט נוספים.