מטלה ראשונה - כלכלה בעולם ה **Big Data**

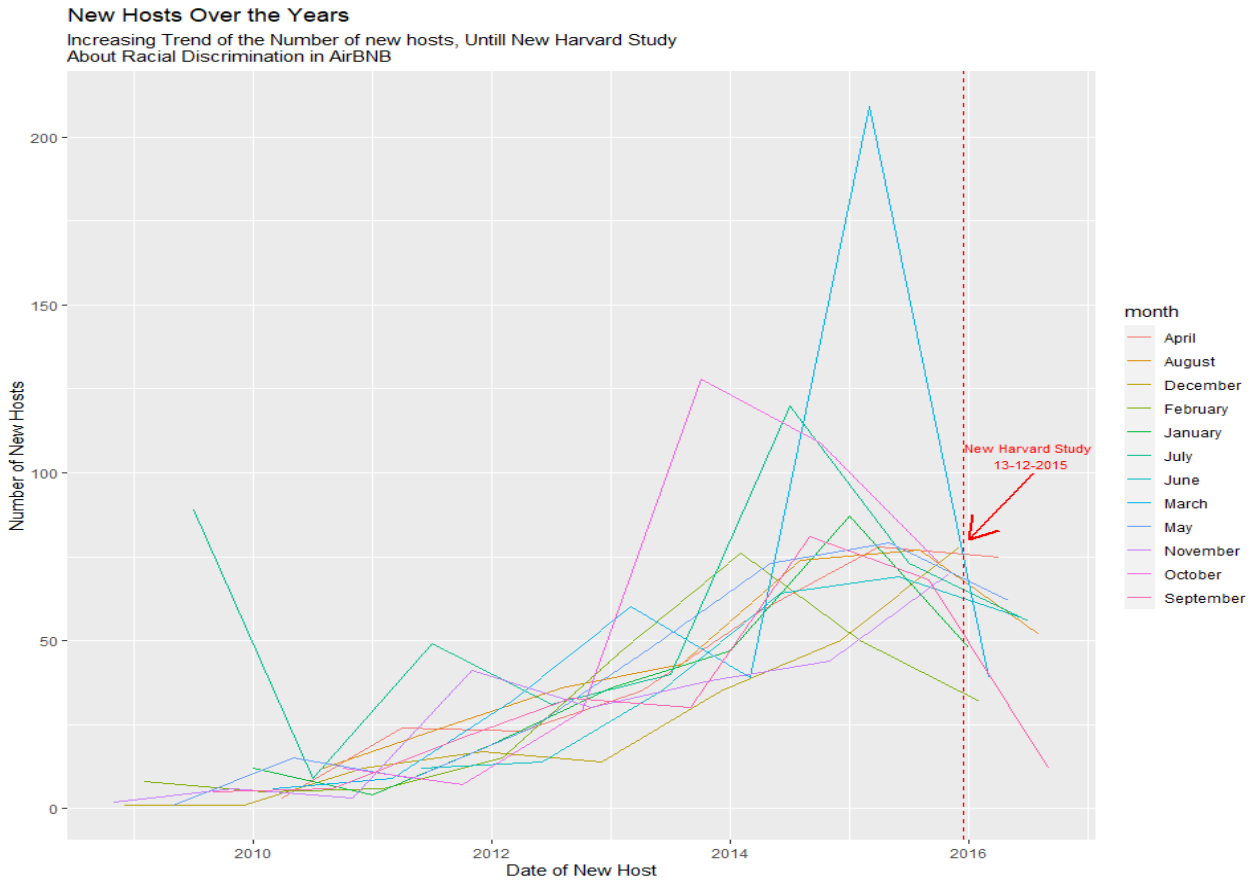
ד"ר רועי ששון ; אופיר בצר

מגישים : אלעד גולן, איתי נחום

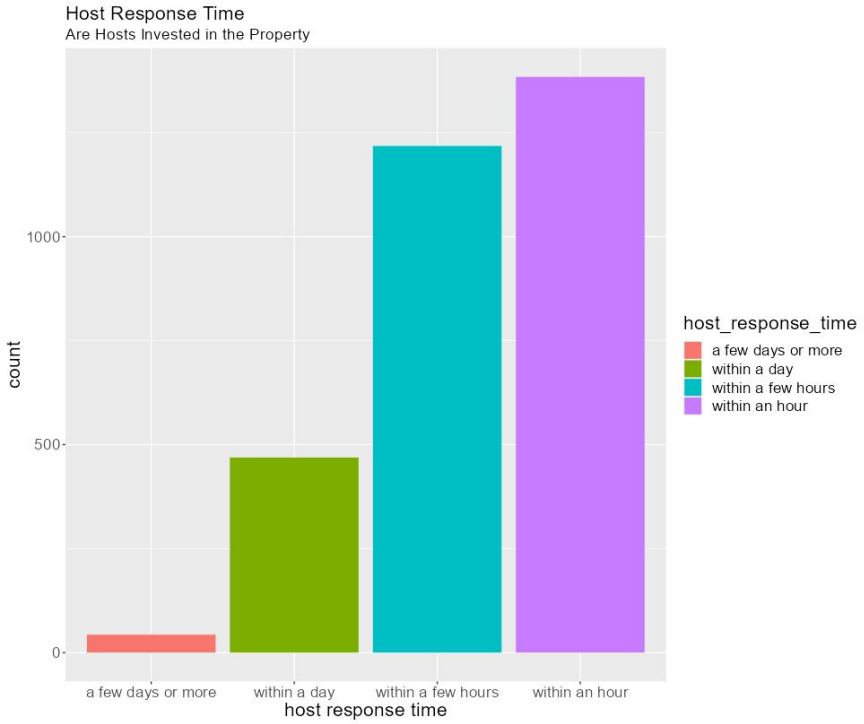
חלק 1 – סטטיסטיקה תיאורית

1. סטטיסטיקה עבור משתנה יחיד :

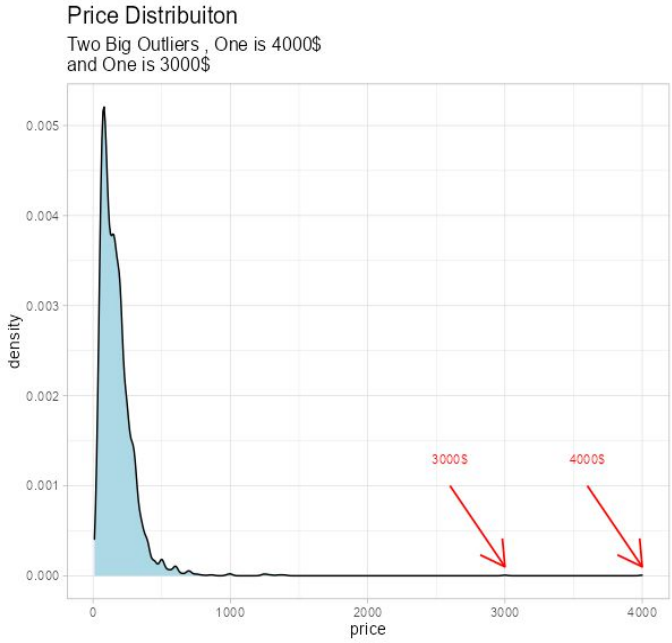
בתרשים "New Hosts…" השתמשנו במשתנה host_since וחילקנו אותו לפי חודשים ושנים בשביל לראות כמה משכירים חדשים הצטרפו ל –AirBNB בכל חודש. הסיבה לחתכים על פני החודשים היא כדי לנסות ולנטרל את ההבדל בפופולריות של כל חודש. ניתן לראות כי ישנה מגמת עלייה במספר המשכירים החדשים משנת 2011, כולל ערך חריג בחודש מרץ 2015 בו מספר המצטרפים היה הגבוה ביותר ובנוסף כי התחילה מגמת ירידה לקראת סוף שנת 2015 היכולה להיות כתוצאה מפרסום מאמר של אוניברסיטת הארוורד אשר תפס תאוצה ברשתות החברתיות על גזענות בקרב המשכירים. ניתן לראות שמגמת העלייה מתרחשת למשך כל חודשי השנה. בעזרת גרף זה אפשר להסיק את הצמיחה של AirBNB בבוסטון, בהנחה שכמות המשכירים גדלה כתוצאה מביקוש סוחרי דירות ל-AirBNB.

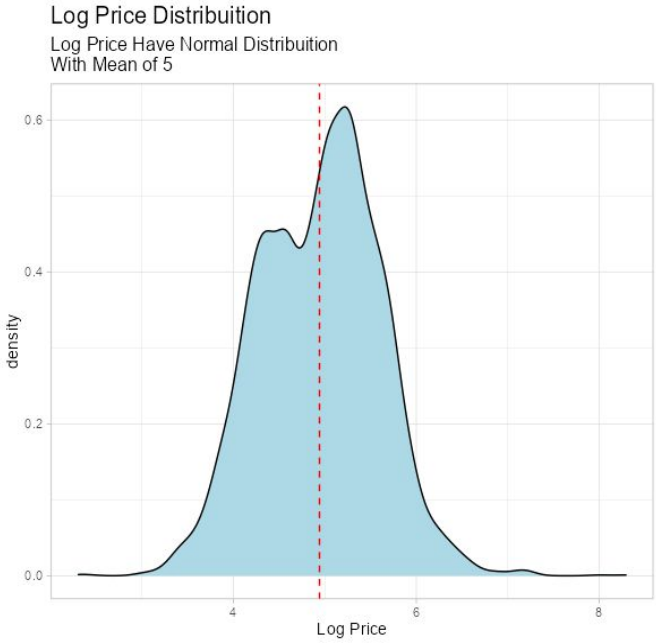


בתרשים "Host Response Time" אנו בוחנים את המשתנה host_response_time ומנתחים כמה המשכירים משקיעים בנכס אשר ברשותם על ידי האינטראקציה עם השוכר. אנו רואים שהרוב הגדול של המשכירים עונים לשוכרים מקסימום תוך מספר שעות, דבר המעיד על חשיבות הנכס למשכיר ועל כך שהנכס המושכר הוא חלק אינטגרלי מהכנסתו של המשכיר, ואינו פרויקט צד אשר אין לו חשיבות גדולה.



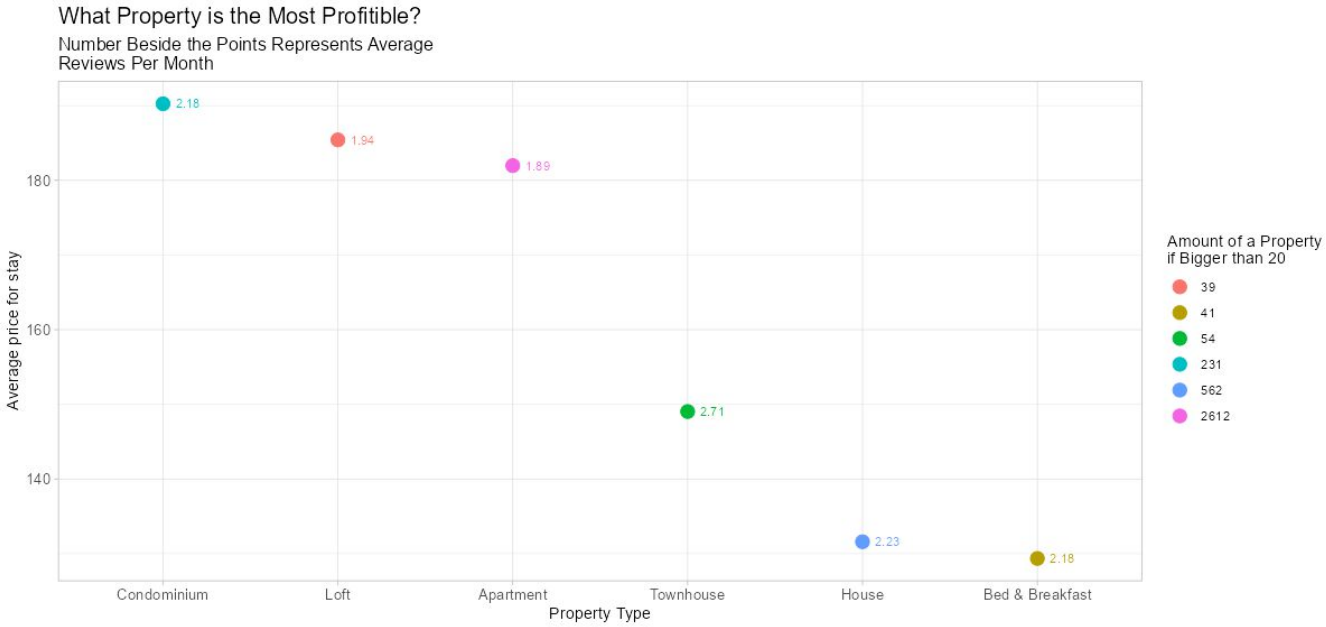
תרשים "Price And Log Price Distribution" אנו מנתחים את משתנה המחיר, בשביל לזהות את ההתפלגות שלו והאם קיימים ערכים חריגים. ניתן לראות בגרף השמאלי כי ישנם 2 ערכים חריגים מאוד, וכי ההתפלגות של המחיר עם זנב ימיני. לכן השתמשנו בפונקציה לוג כפי שניתן לראות בגרף הימני, ובכך ההתפלגות כעת היא נורמאלית. ניתוח זה חשוב לניתוחים ומודלים עתידיים.



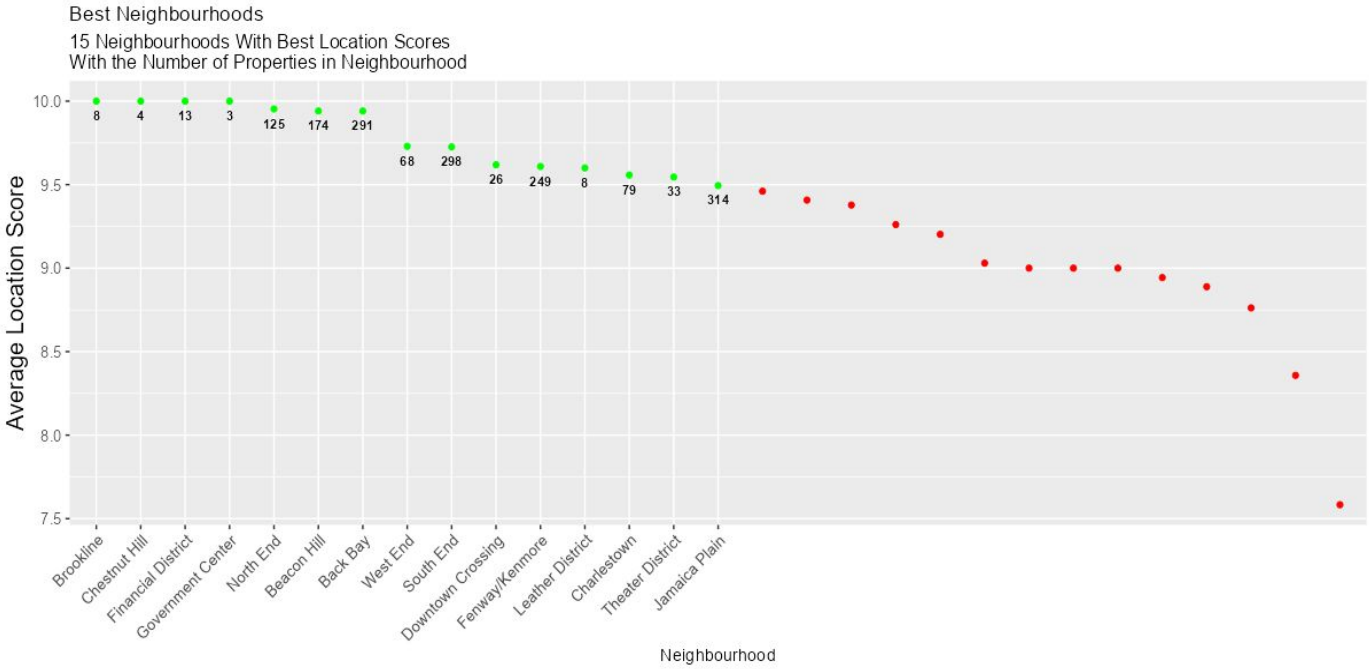


2. <u>סטטיסטיקה עבור 2 משתנים או יותר:</u>

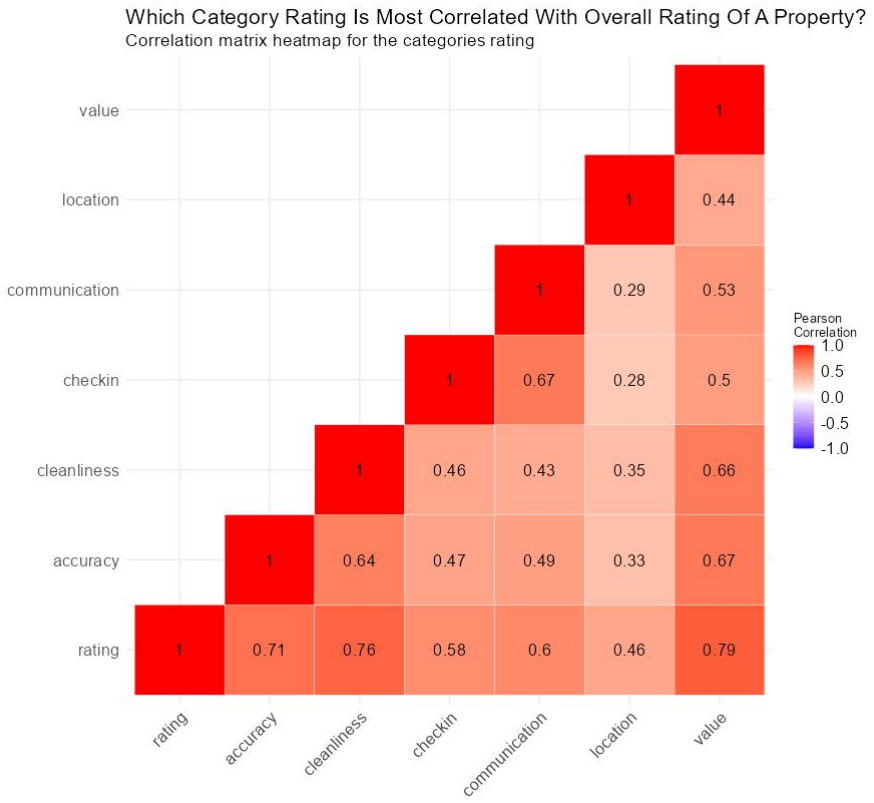
בתרשים "What Property is…" אנו בוחנים איזה מהנכסים יכולים להכניס הכי הרבה רווחים, לשם כך אנו לוקחים את המשתנה של ממוצע דירוגים בחודש הנמצא ליד כל נקודה, בתור אומד מוטה כלפי מטה למספר הפעמים בחודש בממוצע שסוג נכס זה מושכר (לא כולם משאירים דירוגים לכן הוא אומד מוטה). בנוסף אנו מסתכלים על ממוצע התשלום עבור לילה אחד עבור כל נכס, ועל מספר הנכסים מכל סוג. בעזרת ניתוחים אלה נוכל לדעת עבור אילו סוגי נכסים התחרותיות והביקוש גדולים יותר, השהות החודשית גבוהה יותר ומה הבדל המחירים בין הסוגים. כך נוכל להסיק באילו סוגי נכס כדאי לחברה להשקיע יותר ובאילו פחות.



בתרשים "Best Neighbourhoods" אנו בוחנים את ממוצע דירוג המיקום של הנכס מול השכונה בה הוא נמצא. מיקום הנכס במקום מרכזי הוא גורם קריטי בבחירה של הנכס, כאשר אנשים יוצאים לחופשות, ובעזרת גרף זה נוכל לדעת איזה מהשכונות כדאי להשקיע מבחינת המיקום שלהם. בנוסף ליד כל נקודה רשום את מספר הנכסים הנמצאים באותה שכונה, כך שנוכל לדעת באלה מהשכונות יש יותר תחרות בין המשכירים והביקוש גדול יותר.



בתרשים "Which Category Rating…" ניתן להבחין בקורלציות בין סוגי הדירוגים השונים של הנכס. ניתן לשים דגש עיקרי בגרף זה על הקורלציה בין הדירוג הכללי לנכס עם שאר הדירוגים. הבחנה זו מאפשרת הבנה על אילו פרמטרים חשוב שבעל הנכס ישים דגש על מנת שהלקוחות יהיו מרוצים וימשיכו להשכיר נכסים בפלטפורמה. לדוגמה :ניתן לראות שתיאור הנכס והניקיון שלו חשובים מאוד ללקוחות ולעומת זאת המיקום של הנכס פחות משפיע על הדירוג הסופי. כלומר, תיאום ציפיות וניקיון הינם הכרחיים ללקוח (שוכר) כדי שיחזור.



חלק 2–מטריקות

צמיחה – המטריקה אותה אנו מגדירים לצמיחה היא ממוצע השינוי באחוז התפוסה החודשי עבור נכסים, כלומר אחוז השינוי בין תפוסת הנכס בחודש לעומת חודש קודם לו (תפוסה – מספר הימים מכל חודש בו הנכס היה תפוס). המטריקה תופסת את הצמיחה הכוללת בביקוש החודשי לשימוש בנכסי AirBNB באזור בוסטון, ובכך מתארת את הצמיחה התפיסתית כארגון הנותן פתרון לדיור וחופשות. המטריקה מפספסת שינויים לפי חודשים ספציפיים או לפי שכונות ספציפיות באזור, כמו כן, המטריקה לא מתארת האם ישתנה מספר הנכסים באזור, כלומר לא מתחשבת בגידול בהיצע. אם כי, הנחה סבירה היא שהיצע הדירות גדל ועל כן הצמיחה בתפוסה גדולה יותר מאומדן שנקבל על פי המטריקה. תוצאת האומד למטריקה היא 0.02362788 , כלומר בממוצע השינוי באחוז התפוסה החודשי בנכסי AirBNB גדל ב - 2.3% מדי חודש. מכיוון שהנתונים מורכבים רק משנה אחת, בהינתן יותר נתונים המטריקה הייתה יותר מדויקת ותופסת גם שינויים בין שנים ועל פני חודשים ספציפיים משנה לשנה (סביר להניח שחודשים מסוימים יותר פופולריים מאחרים, לדוגמה חודשי הקיץ).

צמיחה לפי שכונה – המטריקה אותה אנו מגדירים היא ממוצע השינוי השנתי עבור מספר הנכסים החדשים עבור כל שכונה. המטריקה תופסת את השינוי בהיצע הנכסים עבור כל שכונה בממוצע, ומכך ניתן להסיק גם את השינוי השנתי בביקוש להשכרת נכסי AirBNB . המטריקה מפספסת את השינוי עבור סוגי נכסים שונים, אלא רק מתארת ממוצע עבור כל שכונה. לצפייה בתוצאות האומדנים של המטריקה יש להסתכל *בטבלה 1*. לדוגמא בשכונה Financial District בממוצע גדל מספר הנכסים החדשים בשכונה פי 1.5 מהשנה הקודמת (אם למשל בשנה ספציפית יש 6 נכסים חדשים שהצטרפו בשכונה, אז בשנה הבאה יש 9 נכסים חדשים). נראה כי מצב העסק משתפר משנה לשנה, יותר ויותר נכסים חדשים מצטרפים לעסק כתוצאה מהשלכות הרווח הטמונות בו. נציין שהנתונים מציגים לנו רק נכסים ומצטרפים ולא נכסים שעזבו את הפלטפורמה, עוד נקודה שעלולה להיות קריטית מבחינת הניתוח.

שינוי במחיר – המטריקה אותה אנו מגדירים היא הממוצע השינוי החודשי במחיר ללילה עבור כל סוג נכס. המטריקה מתארת את השינוי החודשי של המחיר המבוקש ללילה עבו כל סוג נכס, כלומר המחיר אנו מסיקים כמחיר שיווי משקל בין הביקוש וההיצע של נכסים בבוסטון. המטריקה מפספסת מאפיינים ספציפיים עבור כל חודש מכיוון שהנתונים נלקחו רק משנה אחת, למשל חודשים בהם יש יותר אנשים שיוצאים לחופשות. התוצאות מוצגות *בטבלה 2*. לדוגמא עבור סוג נכס Loft ממוצע המחירים ללילה עבור כלל הנכסים מסוג זה עולה בכל חודש ב - 1.6%. בהינתן כך שהנתונים הם מייצגים רק שנה אחת, הוספה של נתונים הייתה יכולה לעזור לאמוד את המטריקה באופן יותר מדויק כאשר מוסיפים עוד מספר חודשים לחישובים.

טבלה 1:

neighborhood	Mean lag
Allston-Brighton	1.900448469
East Boston	1.921201814
Back Bay	2.073023416
Beacon Hill	1.277889732
Chinatown	1.646111111
Fenway/Kenmore	1.726967241
Jamaica Plain	2.008105721
Mission Hill	1.639661925
North End	1.603132029
Roslindale	1.760185185
Roxbury	1.356122449
South Boston	2.906492618
South End	1.348451048
Theater District	1.663492063
West End	1.535950081
Charlestown	1.14265873
Dorchester	1.216074909
Financial District	1.5
West Roxbury	1.642171717
Chestnut Hill	1.5
Mattapan	2.691666667
Downtown	0.777777778
Downtown Crossing	4.472222222
Government Center	0.5
Hyde Park	2.033730159
Leather District	1.388888889
Somerville	1.666666667
Brookline	1.25
Cambridge	0.833333333
Harvard Square	NA

טבלה 2:

Property kind	mean_lag_prop
Apartment	0.974883773
Bed & Breakfast	0.984071987
Boat	0.987456787
Camper/RV	1
Condominium	0.998693225
Dorm	0.958333333
Entire Floor	0.990383061
Guesthouse	1.01749379
House	1.007811641
Loft	1.016520523
Other	0.992716787
Townhouse	1.010820871
Villa	0.940104622

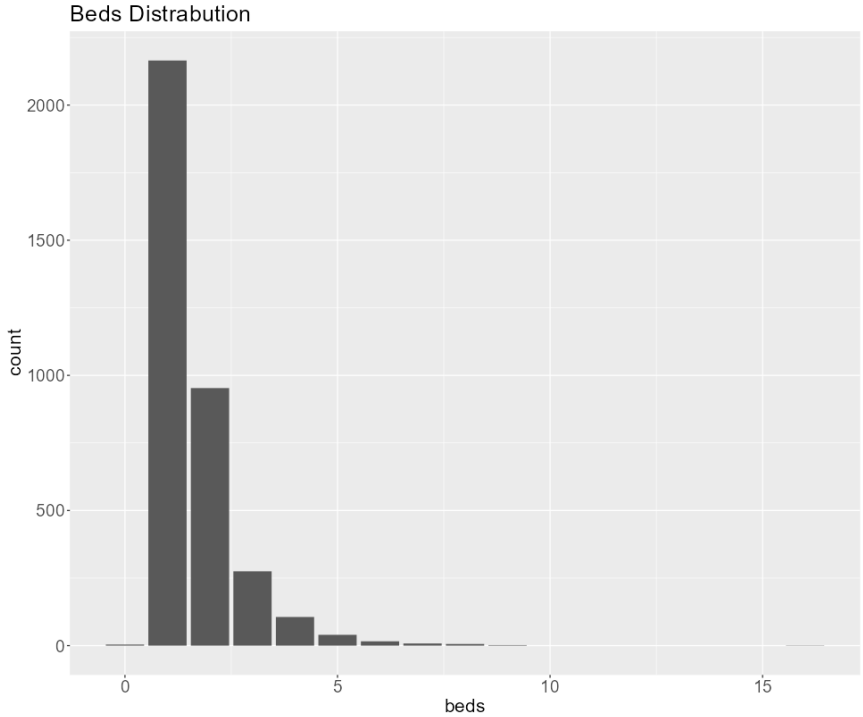
חלק 3 - זיהוי ערכים חריגים

ערכים חריגים עם טווח זמנים: *בתרשים "New Hosts…"* בו אנו מנתחים את מספר המשכירים החדשים שהצטרפו לארגון, ניתן לראות טווח זמנים בהם יש התנהגות חריגה. למשל, בחודש מרץ של 2015 ישנה את הכמות הגבוהה ביותר בפער ניכר של משכירים חדשים. אף ניתן להבחין שההבדל בין מספר המשכירים החדשים בין 2015 לבין 2016 עבור כל חודש הוא המקרה היחיד בו בכל החודשים ישנה ירידה של מספר המשכירים החדשים. הסבר אחד לתופעה זו היא המחקר שהוציאה אוניברסיטת הארוורד בדצמבר 2015[1]. במחקר זה טענו כי קיימת גזענות בקרב המשכירים והשוכרים של AirBNB , וטענות אלה פורסמו וצברו תאוצה ברשתות החברתיות באותה תקופה. הדבר מעיד על הערך ביצירת מוניטין חיובי של העסק ברשתות החברתיות למען הצלחתו ומדגיש את חשיבות

[1] https://www.benedelman.org/publications/airbnb-guest-discrimination-2016-09-16.pdf

האופק החברתי, הפוליטי והתרבותי של החברה וכיצד גורמים אלה משפיעים על העסק. מבחינת הזווית של ניתוח הנתונים, תצפיות חריגות אלה מוסיפות לנו עוד מידע עם מאפיינים שונים אשר לא נמצאים בנתונים, כפי שציינו לעיל, לפיכך נשאיר את תצפיות אלה.

<u>זיהוי חריגים מספר מיטות</u>: *תרשים "Number Of Beds Distribution"* מציג את התפלגות מספר המיטות בנכסים השונים, ע"פ הגרף ניתן לזהות ערכים חריגים ב0 ו-16. 0 מיטות: בהסתכלות בנותנים נראה לפי תיאור הנכסים שמדובר בטעות הקלדה. שניים מהנכסים הינם דירות סטודיו וסביר שיש מיטה אחת, אלא אם כן קיים מזרן בלבד ועל כן ציינו 0 במספר מיטות. 16 מיטות: בתיאור הנכס רשום שהדירה ללא ריהוט כלל ויש להביא את ציוד השינה (מזרן מתנפח, שקי שינה), יתרה מזאת רשום שהנכס גדול. עקב זאת, הגיוני לסבור שבעל הנכס התכוון שהדירה מיועדת לאכלס מספר רב של אנשים/טעות הקלדה. מזווית של ניתוח נתונים אנו לא יכולים לקבוע מה מספר המיטות המדויק בנכסים אלה, לכן בניתוחים עדיף להוריד תצפיות אלה או להשתמש בממוצע שיש בנכסים דומים.



<u>זיהוי חריגים מחיר</u>: בחלק של הסטטיסטיקה התיאורית הצגנו את התפלגות המחיר. סימנו שני ערכים חריגים של המחיר ($3000, $4000) היוצרים זנב ימני ארוך. ניתן לראות בטבלת ה-listings שחסרים ערכים רבים בנכסים אלו ובנוסף תיאור הנכס לא מעיד על נכס ייחודי שמצדיק מחיר מופקע. משום כך, סביר להניח שזוהי טעות הקלדה. טיפול נכון בערכים אלה יכול להתבצע על ידי טרנספורמציה של לוג נראות/מחיקת נכסים אלה/השלמת הממוצע של הנכס על פי סוג נכס, מספר מיטות והשכונה שבה ממוקם. בחירת האסטרטגיה תהיה בהתאם לניתוח אותו נרצה לבצע.

<u>חלק 4– המלצה עסקית</u>

קישור לסרטון

https://drive.google.com/file/d/1G6sHiXc5OkdhnrrbJcxNsPGccTnWUZsI/view?usp=sharing