מטלה ראשונה - לספר סיפור בעזרת נתונים

כלכלה בעולם ה-**Big Data**

-

ד״ר רועי ששון ; אופיר בצר

-

-

וידאו סיכום - **https://drive.google.com/file/d/1thaT1Tj6B5l53-RSOqhnvKob4osBF6DT/view?usp=drive_link**

-

מגישים :

אריאל חדוות

איתן בקירוב

יובל בקירוב

- הנחה בנוגע למשתנה **available_category** :

עבור **available_category** = 0 : הנכס הוצע על ידי המארח אבל לא הוזמן ע״י לקוח.

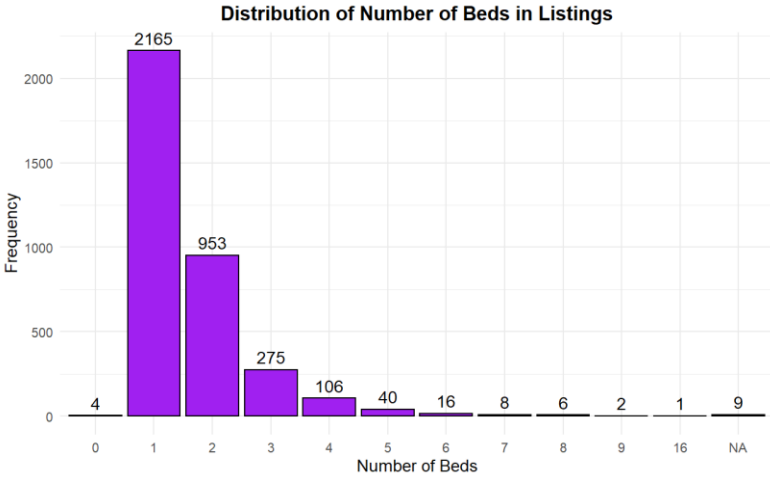
עבור **available_category** = 1 : הנכס הוצע על ידי המארח והוזמן ע״י לקוח.

עבור תאריך שלא מופיע עבור נכס מסוים, נניח כי המארח לא הציע את הנכס שלו להשכרה בתאריך זה.

- נציין כי באופן כללי עושה רושם שכל מארח מציע כמות זהה של לילות, באזור ה 250 לילות.
- הנחה בנוגע למחירים : אנחנו מניחים כי המחיר שמופיע בטבלת **listings_clean** הינו המחיר לילה הראשוני שהמארח הציע על הנכס שלו בעת הכניסה הראשונית שלו לפלטפורמה, בעוד שהמחיר שמופיע בטבלת **calendar_clean** הינו המחיר לילה שהמארח הציע עבור אותו לילה הספציפי ושזה המחיר ששולם בפועל על ידי המתארח.

חלק 1 – סטטיסטיקה תיאורית

1. **התפלגות של מספר המיטות –**

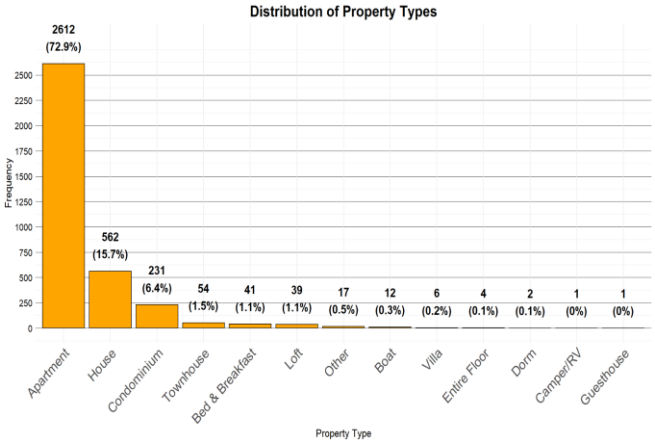


בגרף נראה את התפלגות מספר המיטות בנכס מתוך נכסי העיר בוסטון.

מרבית הנכסים הם בעלי מיטה יחידה. למעשה, יש בדאטה יותר נכנסים עם מיטה 1 מאשר כמות הנכסים עם מספר שונה מ-1 של מיטות.

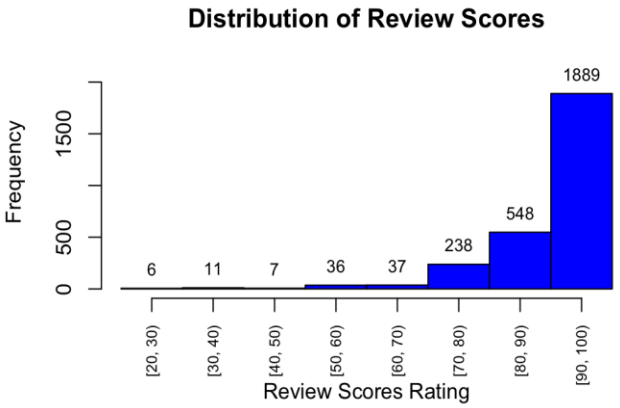
משהו חשוד שתהינו בנתונים הוא שמיטה יחידה יכולה להיות מיטת יחיד ויכולה להיות מיטה זוגית שמתאימה לאירוח של 2 אורחים. אין בנתונים הבחנה בין סוגי מיטות. נתון כזה יכול להעיד על כמות האורחים שיכולים להתארח בנכס = משפיע על הכנסות החברה.

התפלגות של סוגי הנכסים –



בגרף ניתן לראות את התפלגות הנכסים לפי סוג הנכס. רוב הרישומים בדאטה הם דירות, עם כ-2612 דירות (72% מהנכסים). זה מצביע על כך שדירות הן סוג הנכס הנפוץ ביותר הזמין ב-Airbnb בדאטה, מה שיכול אולי לשקף את המאפיינים העירוניים המאפיינים את בוסטון. נוכחותם של סוגי נכסים ייחודיים כמו סירות, מעונות, חניון/קרוואן ובתי אירוח אמנם מעיד על גיוון בסוגי הלינה הזמינים, אבל הכמות יכולה להעיד על כך שנכסים כאלה נדירים יותר. בנוסף, התדירות הגבוהה של דירות עשויה להעיד על חוסר גיוון בסוגי הלינה הזמינים, מה שעלול להוות חיסרון עבור מטיילים המחפשים חוויות ייחודיות.

היסטוגרמה של דירוגי הנכס -



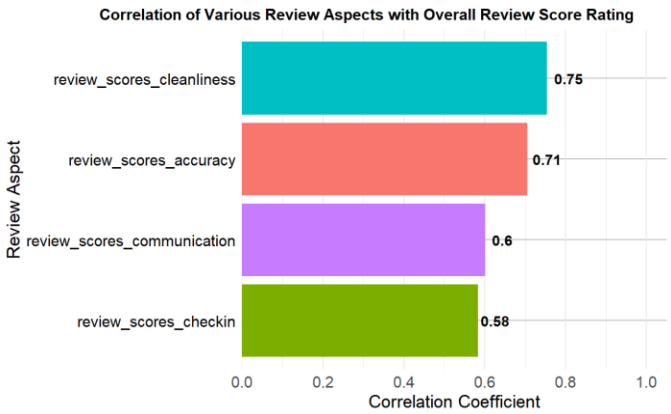
גרף זה מציג את ההתפלגות של ממוצע הציונים שניתנו לכל נכס. ניתן לראות ריכוז גבוה של ציונים גבוהים, הרוב בין 90 ל-100, ומעט מאוד נכסים עם ציונים נמוכים - מה שמצביע על כך שרובם מדורגים גבוה על ידי האורחים - דבר המעיד על חווית אירוח כללית חיובית.

עולה בנו החשד שמא יש מי שמסתיר/מעוות את הציונים לטובה. סביר שיהיו לא מעט נכסים עם דירוגים פחות טובים, שכן, Airbnb מציע מגוון רחב של נכסים, עם מחירים שונים, במיקומים שונים ובעלי מאפיינים ייחודיים. כמובן,

בראייה קדימה, נרצה להמשיך ולתחזק במערכת נכסים עם דירוגים גבוהים. במקביל, להבין מה הוביל לציונים הנמוכים לחלק מהנכסים ומה לעשות כדי לשפר את חווית הלקוחות.

2. גרף קורלציה בין ציוני הביקורות בתחומים שונים עם ציון הדירוג הכללי -

ניתן לראות בגרף זה את הקורלציה בין ציוני הביקורות בתחומים שונים (ניקיון, דיוק להסבר הדירה, תקשורת וצ'ק-אין) עם ציון הדירוג הכללי של הנכס. ככל שהקורלציה גבוהה יותר, יש מתאם חזק יותר בין הדירוג על תחום מסוים עם הדירוג הכללי של הנכס.

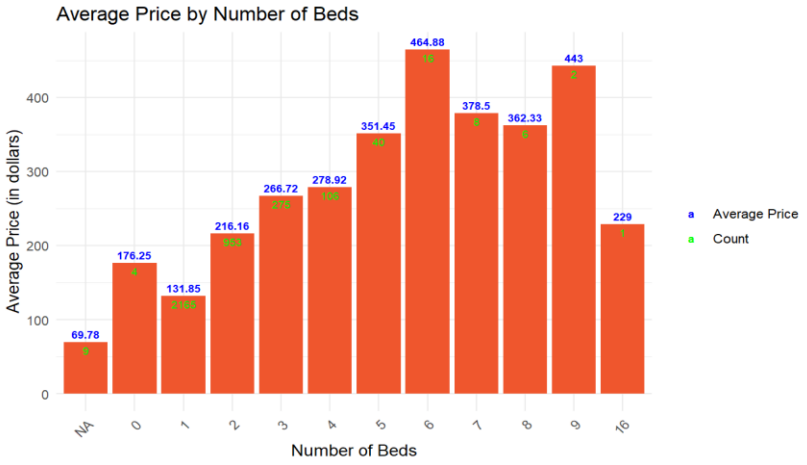


רצינו להסתכל על תחומים שלמארח יכולה להיות עליהם השפעה ובכך יוכל בעתיד לשפר לעצמו את ציון הדירוג של הנכס שאותו מציע להשכרה ולהצליח למשוך יותר לקוחות.

הגרף מראה שבאופן כללי יש קורלציה חיובית וגבוהה מחצי עבור כל הפרמטרים, כך שאפשר להסיק שאם המארח ישקיע בחוויית האירוח שלו (בניקיון הדירה, בהסברים יותר מדויקים, תקשורת טובה עם הלקוחות וחווית צ'ק-אין טובה) הוא ישפיע לטובה על דירוג הנכס שלו. הגרף מראה שניקיון משפיע הכי הרבה על הדירוג הכללי של הנכס, מה שמצביע על כך שאם מארחים ישקיעו יותר בניקיון הנכס, עשויה להיות לזה השפעה חיובית על הציון הכללי של הנכס.

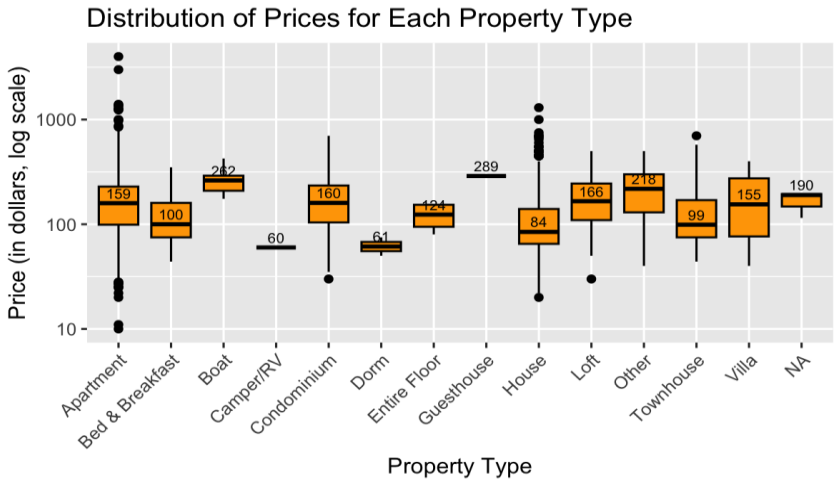
גרף מחיר ממוצע לנכס ליום לפי מספר מיטות בנכס -

גרף של מחיר ממוצע לנכס ליום לפי מספר מיטות בנכס, יחד עם כמות הנכסים עבור כל מספר של מיטות בנכס. ישנה מגמה כללית של עליית המחיר הממוצע ככל שמספר המיטות גדל. זה מצביע על כך שנכסים גדולים יותר, כצפוי, דורשים מחירים גבוהים יותר. המחיר אינו עולה באופן לינארי עם כמות המיטות, מה שמעיד שגם גורמים אחרים משפיעים על התמחור. יש ירידה חדה בזמינות עבור נכסים עם יותר מ-3 מיטות. מעניין לציין כי לנכסים עם 0 מיטות (כנראה דירות סטודיו) יש מחיר ממוצע גבוה יותר ($176.25) מאשר נכסים עם מיטה אחת ($131.85). המחיר הממוצע הגבוה ביותר הוא עבור נכסים עם 6 מיטות ($464.88), ולא עבור הנכסים הגדולים ביותר. זה יכול להצביע על נקודה אופטימלית עבור השכרות יוקרה.



סה"כ, ניתן לראות שמספר מיטות בנכס הוא לא המשפיע העיקרי בהכרח על מחיר הנכס.

התפלגות המחירים של נכסים לפי סוג הנכס -

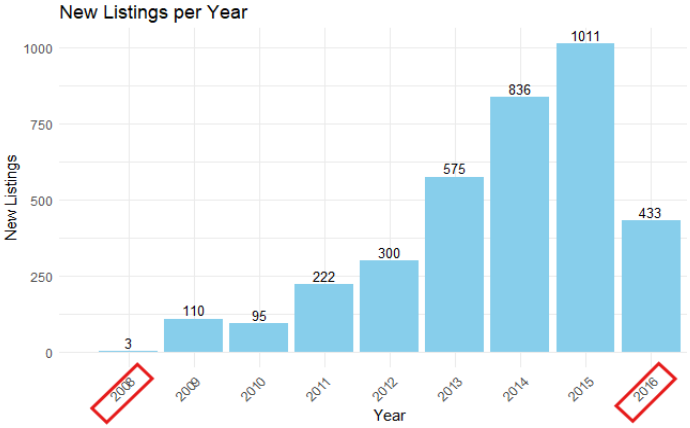


נראה בגרף את התפלגות המחירים של נכסים לפי סוג הנכס. נעזרנו בלוג כדי להתמודד עם ערכים חריגים שמקשים על תצוגת הנתונים. סוגי נכסים כמו דירות, בתים, בתים עירוניים, וילות ועוד מציגים שונות משמעותית במחיריהם, זה מצביע על כך שסוגים אלה של נכסים יכולים להשתנות מאוד באיכות, בגודל, במיקום או בשירותים. ישנם מספר נכסים עם שונות נמוכה, אך במקרים אלו הדבר נובע ממעט התצפיות שלהם בדאטה. לבתי אירוח, לסירות ול"אחר" יש מחירים חציוניים גבוהים יותר, מה שמרמז שהם עשויים להיות מפוארים יותר או בעלי ביקוש גבוה יותר. למעונות, לבתים, לקמפינג, ללינה וארוחת בוקר, ולבתים עירוניים יש מחירים חציוניים נמוכים יותר, מה שמצביע אולי על אפשרויות ידידותיות יותר לתקציב. נוכחות חריגים ברוב הקטגוריות, במיוחד עבור בתים ודירות, יכולים לנבוע מנכסים יוקרתיים במיוחד או בסיסיים מאוד.

<u>חלק 2 – מטריקות:</u>

מימד הצמיחה - הגדרנו מטריקה שמטרתה לראות את הצמיחה השנתית של היצע הנכסים על ידי ספירת כמות נכסים המצטרפים לפלטפורמה מדי שנה.

הנוסחה מוגדרת כך : **Annual Growth (New Listings) = New Listings in the current year**



המדד הזה לוכד את המספר הכולל של נכסים חדשים שנוספו מדי שנה, ומספק תמונה ברורה של האופן שבו היצע הנכסים בפלטפורמה גדל מדי שנה. עם זאת, המדד הזה מפספס מימדים כמו איכות העלייה של הנכסים בפלטפורמה. כלומר, האם הנכסים שהצטרפו לפלטפורמה הוזמנו על ידי הלקוחות, והאם הם מצליחים לענות על דרישות הביקוש בשוק.

מאמידת המטריקה על סמך הנתונים, ניתן לזהות עלייה במספר הנכסים החדשים שנוספו למערכת בכל שנה. נשים לב כי שנת 2016 מסתיימת בספטמבר, כך שהנתונים המופיעים בגרף לא מכסים את כל השנה.

בטבלת התאריכים מתועדת פעילות הנכסים בין ספט' 16 - ספט' 17 בלבד! כלומר, לא ניתן לאמת שהצמיחה שזיהינו במספר הנכסים המוצעים מידי שנה בטבלת הנכסים, אכן הביאה איתה גם יותר לקוחות חדשים מדי שנה (חסרים לנו נתונים על הזמנות של נכסים לפני שנת 2016).

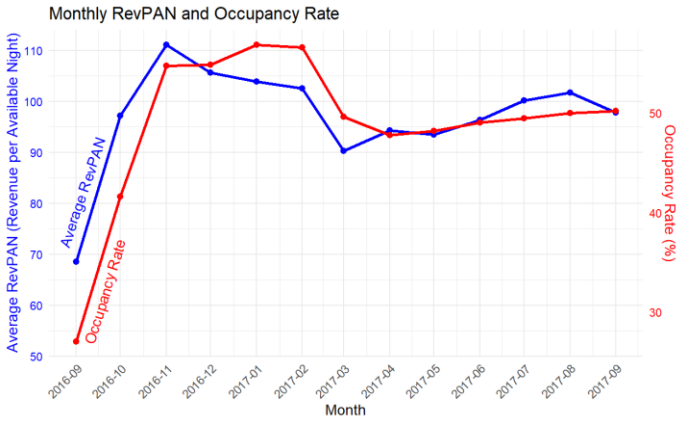
מימד שינוי המחירים - הגדרנו מטריקה שמטרתה להראות את ההכנסה היומית הממוצעת על נכס בחודש. בחרנו להשתמש בממוצע חודשי של מדד RevPAN שמשמעותו Revenue Per Available Night על פני כל הנכסים שנתונים, הוא יכול להיות מחושב באמצעות 2 אופנים:

1. **RevPAN = ADR × Occupancy Rate per month**
   - ADR (Average Daily Rate): Total Listing Revenue / # Nights Booked
   - Occupancy Rate: # Nights Booked / # All Nights per Listing
   - Total Listing Revenue: Sum of price_dollars for a listing where available_category = 1
   - # Nights Booked: Nights where the listing was rented out (available_category = 1)
   - # All Nights per Listing: Both booked and unbooked nights within the month
2. **RevPAN = Total Listing Revenue / # All Nights per Listing**

מדד זה מצליח לתפוס את יעילות ההכנסה הממוצעת של נכסים על ידי התחשבות בתמחור ובשיעורי התפוסה. שימוש ב-RevPAN ממוצע בכל הנכסים בכל חודש מספק תמונת מצב של יעילות ההכנסה הכוללת עבור Airbnb על מנת להבין כיצד לשנות את המחירים נכון. ביצועים יוצאי דופן או גרועים יכולים להטות את הממוצע ולהסתיר בעיות ספציפיות. זה גם לא לוקח בחשבון השפעות שוק חיצוניות על התפוסה או המחירים באופן ישיר.

1X1000 = **1000** -> RevPAN = 1000/1 = **1000** vs. 10X100 = **1000** -> RevPAN = 1000/10 = **100**

עקומת ה RevPAN מציגה ירידה ב-RevPAN מסוף 2016 לתחילת 2017, מה שמצביע על יעילות הכנסה מופחתת ללילה זמין במהלך תקופה זו. עם זאת, יש התאוששות הדרגתית החל מאמצע 2017, מה שמצביע על שיפור באסטרטגיית התמחור ו/או בשיעורי התפוסה. בהקשר של אינטואיציה לשינוי מחירים, נרצה לשנות מחיר במקומות בהם ה- RevPAN נמוך כדי להעלות את פוטנציאל הרווח.



כדי לעשות זאת, נעזר בגרף שיעור התפוסה הממוצעים פר חודש כדי לתת המלצה עסקית ונתמקד באזורים בהם ה- RevPAN היה נמוך ביחס לשאר ונבדוק: עבור אחוזי תפוסה גבוהים, נרצה להעלות את המחיר ועבור אחוזי תפוסה נמוכים, נמליץ להוריד מחירים על מנת למשוך מבקרים, להעלות את אחוזי התפוסה וכך גם את הרווח הכולל.

בעקומת התפוסה, אנו רואים ירידה בתפוסה מסוף 2016 ועד תחילת 2017, אשר מתיישרת עם הירידה ב-RevPAN במהלך אותה תקופה. כשהתפוסה מתחילה להתייצב ולעלות בהדרגה מאמצע 2017 ואילך, היא תומכת במגמת ההתאוששות ב-RevPAN.

המטריקה הזו יכולה להיות יותר מדויקת במידה והיינו יודעים בוודאות את המחיר ששולם בפועל מבלי להניח כי המחיר ששולם בפועל זהה למחיר המוצע ע"י המארח.

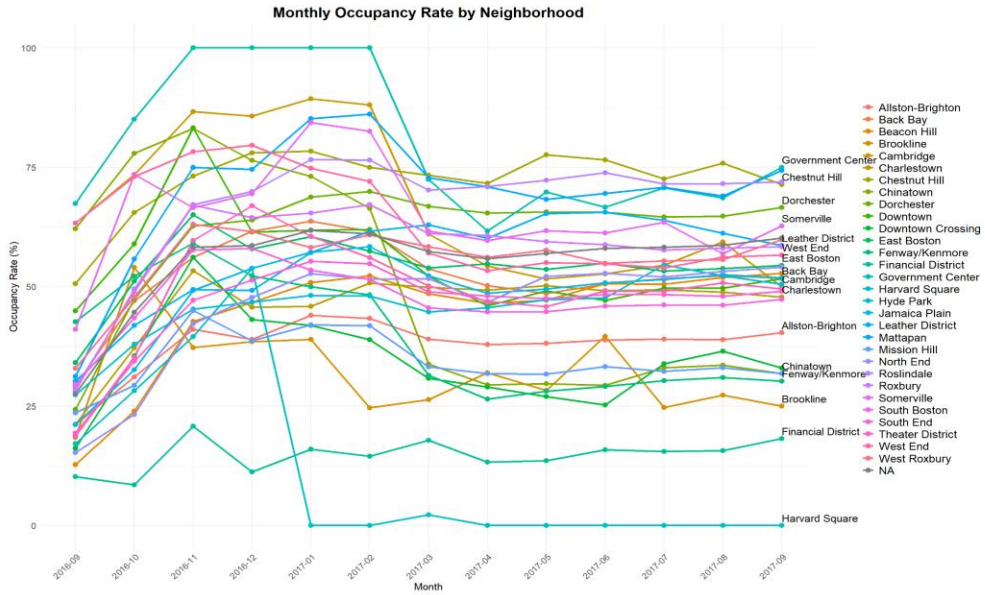
מימד הצמיחה לפי שכונה- הגדרנו מטריקה זהה ל- RevPAN רק שהפעם הסתכלנו על ההכנסה הממוצעת שנוצרה ללילה פנוי בתוך כל שכונה. מדד זה, הנקרא Monthly Neighborhood RevPAN ומוגדר כך :

Monthly Neighborhood RevPAN = Total Revenue in Neighborhood / Total Available Nights in Neighborhood

- Total Revenue in Neighborhood: סכום ההכנסות שנוצרו מנכסים בשכונה מסוימת עבור חודש מסוים.
- Total Available Nights in Neighborhood: המספר הכולל של הלילות הזמינים ע"י המארחים בשכונה מסוימת עבור חודש מסוים.

המדד לוכד הביצועים הכלכליים ופוטנציאל הצמיחה של שכונות שונות, ובכך יכול לספק תובנות לגבי שכונות מרוויחות יותר ופחות, ובהתאם להציע אסטרטגיה כלכלית עתידית. עם זאת, הוא אינו מתייחס לגודל כל שכונה, איכות השכונה ועד כמה היא תיירותית, מידע שחסר בנתונים שיש לנו ועשוי לעזור לנו להבין עד כמה הצמיחה של השכונה הוא ריאלי בהינתן למציאות.

הגרף המוצג מצביע על שונות ב-RevPAN בין השכונות, כאשר אזורים מסוימים עולים באופן עקבי על אחרים וכאלה במגמת ירידה.



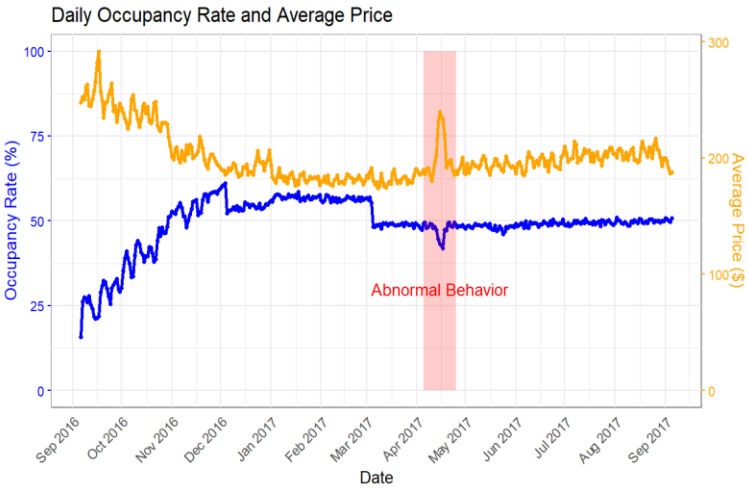
נסתכל על כמות הנכסים שיש (נכון לספטמבר 2016 - חסר מידע בנוגע לנכסים שהצטרפו מאוחר יותר דבר שיכל לדייק את הבנת המצב העסקי) בכל שכונה ולפי זה נבין את מצב העסק שלנו. ניכר בגרף שיש שכונות עם מעט נכסים ו- RevPAN נמוך, ושכונות עם הרבה נכסים ו- RevPAN גבוה. היינו רוצים שזה יהיה ההפך על מנת להגדיל את הרווחים של Airbnb, כך שמצבנו יכל להיות יותר טוב.

היינו מציעים ל Airbnb לבדוק באיזה שכונות יש פוטנציאל צמיחה נמוך ולהבין האם מדובר בחוסר היצע, חוסר פרופורציה בין אחוזי תפוסה למחיר (גם פה כדאי להסתכל על אחוזי התפוסה בשכונות ובהתאם להציע הורדה או העלאת מחירים), שכונות שנחשבות לפחות טובות או פחות תיירותיות ואז בהתאם להציע לבעלי הנכסים להוזיל מחירים כדי למשוך מבקרים. בנוסף, כדאי ל Airbnb להתמקד בשיפור ההיצע בשכונות בעלות ביצועים גבוהים ולשמר את הצמיחה בשכונות אלו ומנגד לפתח אסטרטגיות לשיפור הביצועים בשכונות עם ביצועים נמוכים כדי להגדיל את הצמיחה שם.

חלק 3 – זיהוי ערכים חריגים:

טווח זמנים עם התנהגות חריגה -

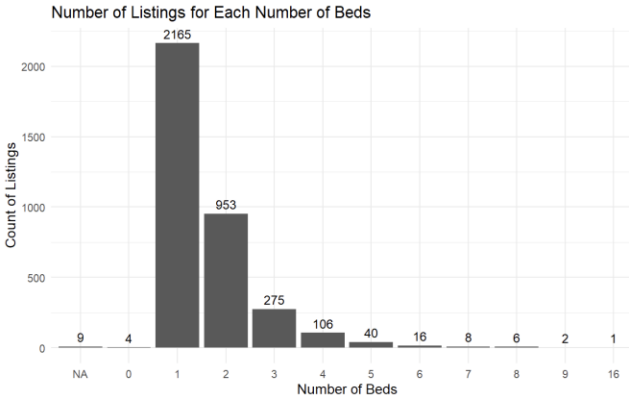
נתבונן בגרף המכיל את ה - Occupancy Rate (דנו בו בחלק 2 - מימד שינוי במחיר) והמחיר הממוצע בכל יום:



נשים לב שבתקופה 12/04/2017 עד 19/04/2017 יש ירידה ב occupancy rate ועליה בממוצע המחירים.

מחקר קצר שעשינו מצאנו שבתאריכים אלו התקיים מרתון בוסטון. לדעתנו, עקב המרתון שתוכנן לאפריל והביקוש הגדל להמצאות באזור, המארחים העלו מחירים באופן דרסטי, דבר שגרם לבסוף לירידה בהזמנות בפועל. אלו נתונים אמיתיים והשמטה שלהם תפגע ברציפות ואחידות הדאטה. נרצה לסמן ערכים חריגים אלו, בשביל לדעת להפיק מהם מסקנות להמשך. דרך התמודדות שנציע היא להוסיף משתנה בוליאני 'Event', שיאותת על תקופה שעשויה להשפיע באופן חריג על הנתונים.

מספר מיטות -

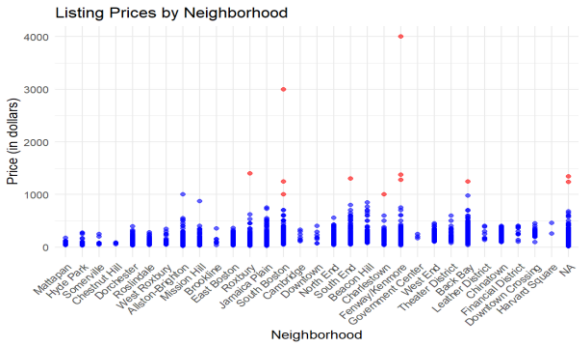


נשים לב כי קיימים כמה ערכים חריגים: נכסים עם 16 מיטות, 0 מיטות ו-NA.

לפי תיאור החדר בעל 16 המיטות, ניתן להבין שהוא אכן חדר גדול ולפי הגדרת המארח, ניתן להכניס 16 אנשים\מיטות. נשאיר את הנתון כדי לא לפגוע בגיוון הדאטה.

לפי תיאור החדרים עבור הנכסים עם 0 מיטות - כולן דירות סטודיו. בדרך כלל חדרי סטודיו מכילים מיטה אחת לכן נמיר את כמות המיטות ל1. קיימים 9 נכסים שבהם מספר לא ידוע של מיטות (NA). מחקירה קצרה, למרות שחלקם רשימות כפולות וללא ביקורות, הם אכן הושכרו לאורך התקופה, לכן נמיר את כמות המיטות שלהם לממוצע של הנתונים, כלומר, מיטה אחת.

מחיר -

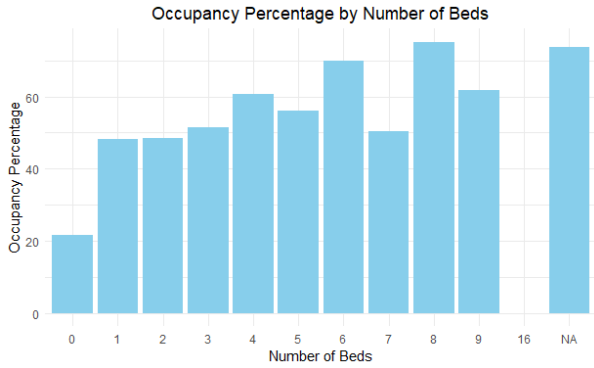


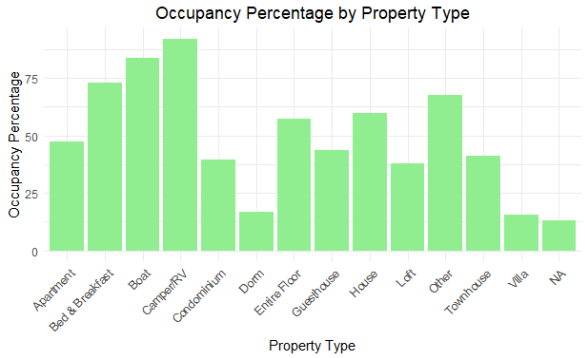
מהסתכלות רוחבית על המחירים בכל שכונה, ניתן לראות שרובם נעים בטווח מחירים שבין 0 ל-1000, עם כמה מחירים חריגים כמו 3000 ו-4000. יש 12 נכסים ששווים 1000 ומעלה, 10 מהם דירות, ו-2 בתים. 8 מתוך 12 הנכסים האלה הוזמנו - כלומר היה ביקוש למרבית נכסים אלו למרות המחיר הגבוה.

הערכים החריגים של מחירים אלו יכולים לנבוע מכמה סיבות שקשה להבין במפורש מהנתונים. מהסתכלות על הנכסים האלה ניתן לראות שכולם בשכונות מרכזיות. מבחינת טיפול בערכים אלו, לא ניגע כדי לא לפגוע בגיוון הדאטה.

חלק 4 - המלצה עסקית:

Airbnb מכניסה כסף לחברה ע״י לקיחת ~3% מסך המחיר לילה שהמארח גובה. היינו ממליצים לחברה ליצור תמריצים ע״י הטבות (למשל, לתמרץ מארחים להפוך להיות superhost ולהוסיף הטבה גם עבור המתארחים שמשאירים ציונים שעוזרים לנו ליצור ניתוח יותר מדויק עבור הנכס), קמפיינים ומסעי שיווק כדי להגדיל את מספר הנכסים המוצעים בכל שכונה ובעיקר שכונות בעלות פוטנציאל רווח גבוה או שכונות שקיבלו ציון שביעות רצון גבוה (כמו Financial District). כך שבגדול המטרה היא ליצור עליה גם בהיצע וגם בביקוש שתגרום לעליית ההשכרות ובכך לעליה בהכנסות של המארח.





על פי האנליזה שביצענו, ראינו כי אין גיוון בהיצע הנכסים המוצעים, היינו רוצים לשנות זאת.

נראה כי היצע הנכסים עם מיטה 1 הוא הכי גבוה אבל דווקא אחוזי התפוסה היו גבוהים יותר עבור נכסים עם מספר מיטות גדול יותר (זה פחות טוב גם כי ראינו שבממוצע נכס עם מיטה 1 הכי זול). בנוסף, נראה כי היצע נכסי הדירות (Apartment) הוא הכי גבוה אבל רק עם 50% תפוסה, לכן היינו מציעים אולי להגדיל את ההיצע בנכסים עם אחוזי תפוסה גבוהים יותר (כמו סירות ובתים) ולראות אם יש שינוי. זאת על מנת להגדיל את הגיוון כדי להגיע לאוכלוסיות שונות.

בנוסף, נרצה שכל מארח יגיע למקסימום פוטנציאל הרווח שאליו הוא יכול להגיע (זאת אנחנו נוכל לדעת לפי ההסתכלות על ה RevPAN בשילוב אחוזי התפוסה שלו) ולגרום לשינויי מחירים איכותיים. על כן, נמליץ שיהיה ל Airbnb הסתכלות כזו שיכולה להתריע למארח על סמך נתונים קודמים האם כדאי לו להעלות או להוריד מחירים על מנת ליצור את האיזון המושלם בין המחיר פר לילה לכמות מבקרים.

ממה שנראה כרגע היינו מציעים ל Airbnb להמליץ בכלליות למארחים בשנה הבאה להוריד מעט את המחיר ללילה בחודש מרץ. במידה ואחוזי התפוסה יעלו וממוצע ההכנסות יעלה, נמליץ להם להמשיך עם מחירים אלו גם בשנה שאחריה.