**מטלה מסכמת –:NEW YORK AIRBNB DATASET**

מגישים: גילעד גילבוע ושגב כהן

# 1. מבוא

**בשנים האחרונות ניכרת עלייה משמעותית ביוקר המחיה, לכן, אנשים רבים נאלצים לצמצם בהוצאותיהם. מטבע האדם, תחומי החיים הראשונים שיצומצמו הם המותרות: מסעדות, בילויים, חופשות וכו'. מכורח המצב, האדם הממוצע נדרש לערוך חיפוש אלטרנטיבה זולה לחופשות, תחליף זול לשהות במלון. כל זאת, יחד עם התממשות אפליקציות ואתרים כמו Booking ו- Airbnb והנגשתם את מקומות האירוח למשתמש, שינו את עולם התיירות לחלוטין, כיום, אין אדם שייצא לחופשה מבלי לבדוק את כלל האופציות והאתרים הללו מהווים שחקן מפתח בשוק התיירות המודרני.**

בעבודה זו אנחנו נבחן כיצד מתומחרות דירות Airbnb בעיר עולם מרכזית ומה הם מרכיבי המחיר.

# 2. שיטה ותוצאות

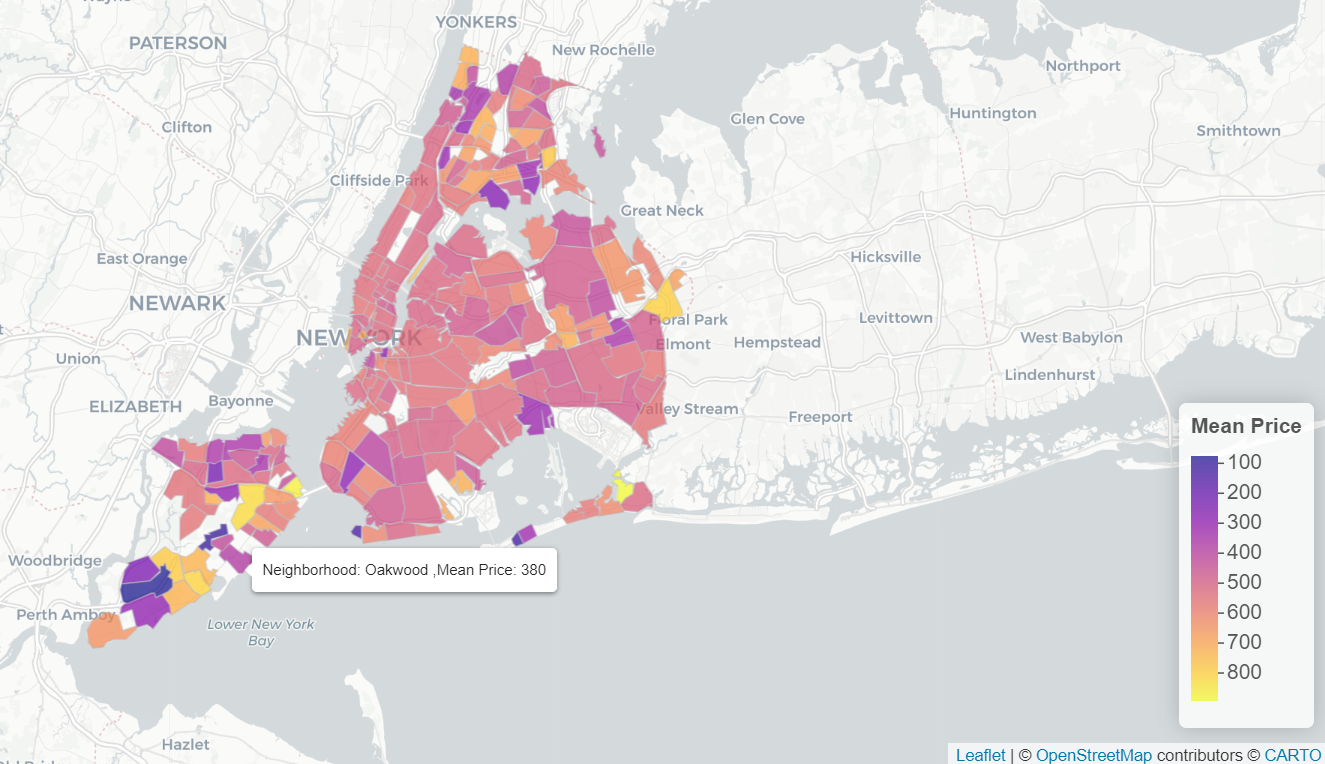
ראשית נייבא את מאגר הנתונים שלנו (Airbnb\_data), נסנן בו את הערכים החריגים, נהפוך את עמודות המחיר ומחיר השירות לעמודה מספרית (הערכים מגיעים כstring), נסנן דירות עם מתחת ל10 דירוגים. לאחר מכן נסנן את מאגר הנתונים אל תוך משתנה Only\_homes כך שנוכל לעבוד עם בתים בלבד (בairbnb מפורסמים גם חדרי מלון וסוגים שונים של מקומות אירוח)

## 2.1 שאלה ראשונה: מה ממוצע המחיר של הדירות פר שכונה במנהטן? וכיצד ממוצעי המחירים מתפרשים על פני ניו יורק?

כדי לענות על שאלה, זו נחשב את ממוצע המחירים בדירות באיזור מנהטן לפי שכונות, כאשר ניקח בחשבון רק דירות בעלות 30 דירוגים ומעלה (על מנת לגשת רק לדירות רלוונטיות אשר שהו בהם כמות סבירה של אנשים), לאחר מכן נייצר מפת חום המובנת על מפת ניו יורק המציגה את כל שכונה בצבע המתאים לממוצע המחיר בה לעומת השכונות האחרות .

**שיטת העבודה:**

* נבנה טבלה (table1) אשר מכילה את שמות השכונות ממאגר הנתונים והמחיר הממוצע בהן
* נייבא את הספריות leaflet, sf אשר יעזרו לנו לעבוד ולייצר את המפה בR
* לאחר מכן נייבא קובץ geojson אשר מכיל את המפה עצמה מהאתר data-beta-nyc, עליה נלביש את ממוצע מחירי הדירות בכל אחת מהשכונות בתצורת צבע.



המפה הנ"ל נוצרה בקוד המצורף לעבודה, מומלץ לעיין בה דרך הRstudios , שם ניתן להחליק את העכבר על גבי כל אחת מהשכונות הצבועות ולהתרשם משם השכונה וממוצע המחיר בה.

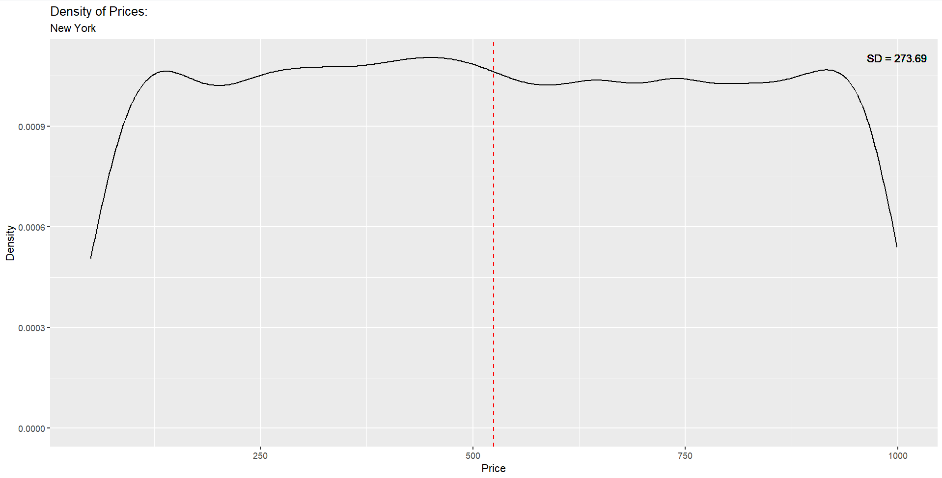
**ממצאים:**

מפת החום מספקת לנו תמונה ברורה הממחישה בצורה ויזואלית את ממוצע המחירים של כל שכונה ואת גובה המחיר הממוצע שלה לעומת שכונות אחרות, כך נוכל לענות בצורה המיטבית על השאלה הנשאלת.

ממפת החום ניתן להסיק כי כלל האזור המרכזי בניו יורק נמצא פחות או יותר באותו טווח מחירים ואילו בקצוות העיר ישנם אזורים יותר ופחות יקרים. על כן ניתן להסיק כי ריבוי ההיצע והביקוש של מקומות אירוח במרכז ניו יורק יצרו תחרות גדולה ועל כן המחירים דומים יחסית ובאופן מעניין נמצאים במרכז הסקאלה של המחיר.

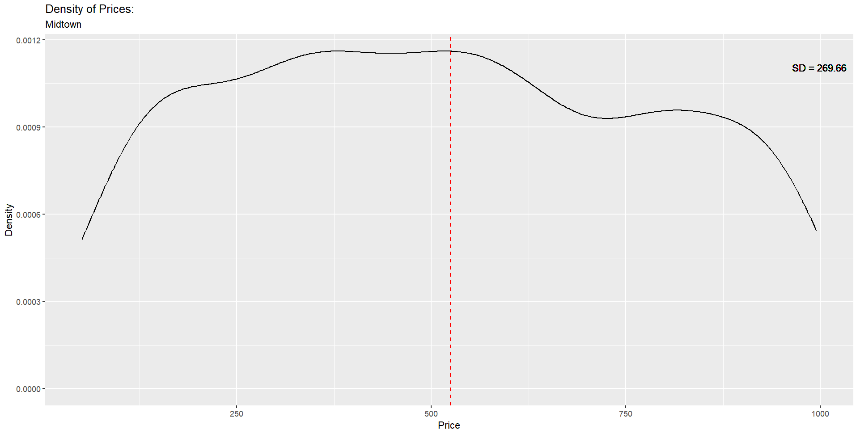
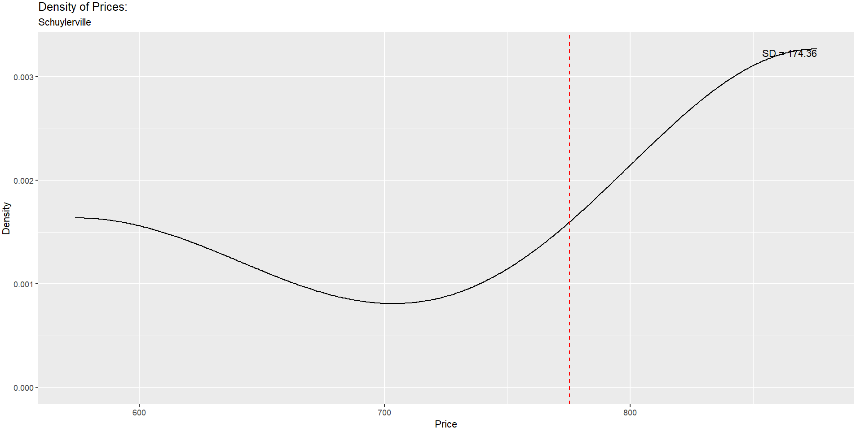
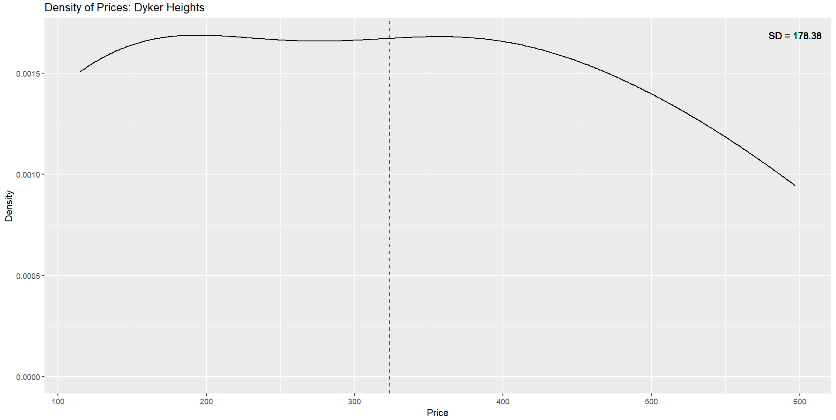
**2.2 שאלה שניה : האם יש דמיון בין התפלגות המחירים בכלל ניו יורק לבין שכונות ספציפיות בניו יורק?**

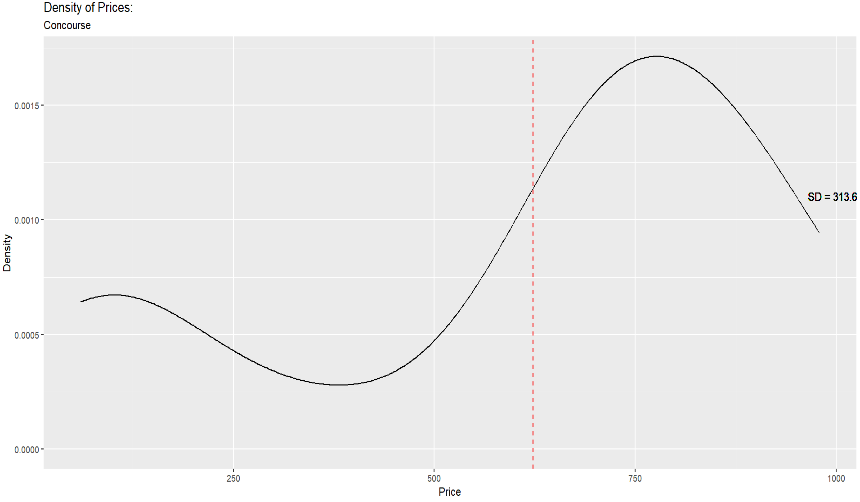
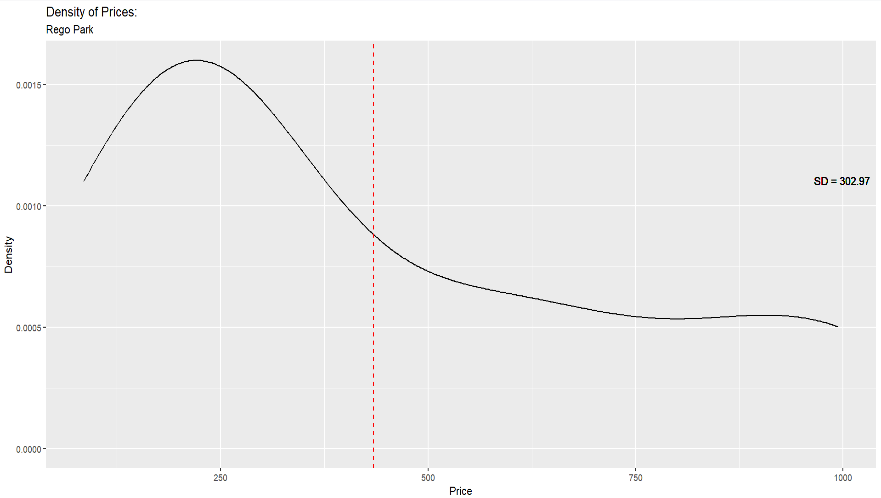
על מנת לענות על שאלה זו ניצור גרפי התפלגות המחירים של כלל ניו יורק ושל שכונות בתוך ניו יורק, לאחר מכן נבחר את היעילות (טעות האמידה הממוצעת) של גרפים בשכונות הנבחרות ואת דמיונם לגרף ההתפלגות של כלל ניו יורק.

כאן ניתן לראות את גרף התפלגות המחירים בניו יורק, כעת, נשווה ויזואלית את הגרפים של השכונות הבאות לעומת גרף זה:

* Schuylerville – שכונה עם מחיר ממוצע 775, גבוה יחסית לניו יורק
* Dyker Heights– שכונה עם מחיר ממוצע 284, נמוך יחסית לניו יורק
* Midtown – שכונה מרכזית מאוד בעלת מחיר ממוצע 517, מחיר ממוצע ביחס לניו יורק
* Concourse – שכונה בעלת מחיר ממוצע 622, מחיר גבוה יחסית אך קרוב לממוצע
* Rego park – שכונה עם מחיר ממוצע 434, מחיר נמוך יחסית אך קרוב לממוצע

\* את כל המידע ניתן לראות במפה המצורפת משאלה 1





**ממצאים:**

מהגרפים הנ"ל ניתן להסיק כי רוב גרפי התפלגות המחירים של השכונות השונות בניו יורק שונים מאוד מהגרף הכולל של ניו יורק, יתרה מכך, הגרפים שונים מאוד ויזואלית גם האחד מהשני, כל זאת מלבד השכונה Midtown, בה מספר גבוה מאוד של דירות, ניתן לראות גם כי צפיפות המחירים בגרף שלה גבוהה ביחס לשאר הגרפים, גם כי ממוצע המחירים קרוב יותר לממוצע המחירים בכלל ניו יורק ובכלל גרף התפלגות המחירים של השכונה דומה יחסית לגרף של כלל ניו יורק, על כן ניתן להניח כי ככל שמספר הדירות רב ושואף למספר הדירות של כלל ניו יורק, כך גרף התפלגות המחירים (האומדן) שואף לגרף התפלגות המחירים בניו יורק (האמוד), מכאן, שגרף התפלגות המחירים בניו יורק הינו אומד עקיב.

מעצם העובדה שהגרפים הפרטניים של השכונות שונים ויזואלית אחד מהשני ומהגרף של כל ניו יורק ניתן להניח כי במדגמים קטנים (בעלי מעט דירות) יגדל השוני ביחס לאמוד, דבר שמחזק את ההשערה כי גרף התפלגות המחירים הינו אומד עקיב. יתרה מכך, ניתן לשער כי היעילות של גרף התפלגות המחירים של כלל ניו יורק הינו אומד בעל יעילות נמוכה.

**2.3 שאלה שלישית : מהן 1,000 הדירות בעלות הערך הגבוה ביותר ביחס למחיר בניו יורק, בהתחשב בגורמים כמו אימות זהות המארח, אפשרות הזמנה מיידית, מדיניות ביטולים, שנת בנייה, מספר ביקורות ושיעור ביקורות?**

על מנת לענות על שאלה זו נייצר אלגוריתם אשר יסנן וימצה את הנתונים כך שנמצא את הדירות הכי קוסט אפקטיביות בניו יורק, אלה אשר להן העלות למול התועלת הינה מרבית, ע"פ קריטריונים שבחרנו.

האלגוריתם ייצור טבלה ובה 1000 הדירות עם הדירוג הגבוה ביותר בניו יורק. הדירוג יתבצע על ידי ממוצע שני מבחנים הכוללים שימת דגש יחסי לכל מאפיין רלוונטי הקיים במאגר הנתונים. כך, נוכל לקבוע קוסט אפקטיביות של דירה על פי הציון הנקבע לה על ידי האלגוריתם.

האמוד בשאלה זו הינו מדידה של דירות בניו יורק על פי מבחן עלות – תועלת. האלגוריתם שמבצע את החישוב הנ"ל הינו האומד והטבלה הסופית מהווה את האומדן.

**להלן שיטת החישוב של האלגוריתם:**

\*ראשית, אנו מסירים כל שורות עם נתונים חסרים או לא חוקיים בעמודות הרלוונטיות.

* מבחן 1 – מבחן המחיר:

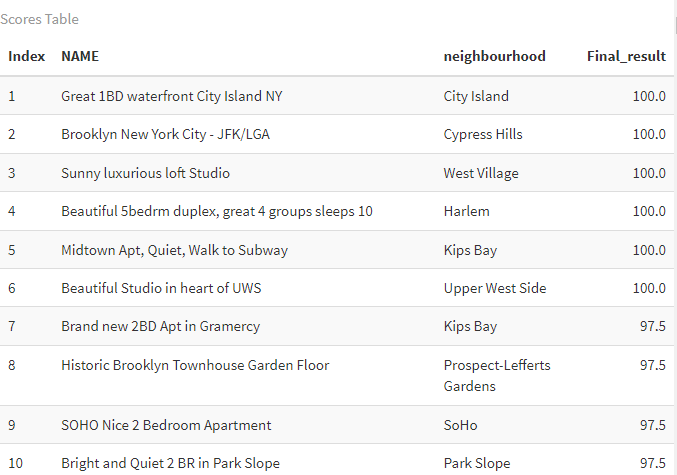
מבחן זה כולל בחינה בודדת של מחיר הדירה, כאשר הציון המירבי הניתן הינו 100

* + הדירות המדורגות יקבלו ציון מ-100 ועד 50 בקפיצות של 10 כך שאם מחיר הדירה נמוך מ-200, הדירה תקבל ציון 100, אם מחיר הדירה בין 200 ל-300 הדירה תקבל ציון 90 וכך הלאה עד המחיר 600 שלאחריו כל דירה בעלת מחיר גבוה ממנו תקבל ציון 50.
* מבחן 2 – מבחן קריטריונים מעורב:

מבחן זה כולל סכימה של 6 תתי מבחנים, אשר עמידה בכולם בצורה המרבית מקנה סכום כולל של 100.

* + דירה אשר המארח בה מאומת (verified) תקבל 15 נק' ו-0 אחרת.
  + דירה אשר ניתנת להזמנה באופן מיידי (instant bookable) תקבל 10 נק' ו-0 אחרת.
  + דירה אשר מדיניות הביטול שלה הינה קפדנית (strict) תקבל ציון 0. עבור מדיניות מתונה (moderate) הדירה תקבל ציון 5, אחרת תקבל ציון 10 כשהמדיניות ביטול שלה גמישה (flexible).
  + דירה אשר שופצה (Construction.year) במהלך 5 השנים האחרונות, החל מ-2018 כולל, תקבל 10 נק' ו-0 אחרת.
  + דירה אשר יש לה 30 (number.of.reviews) מדרגים ומעלה תקבל 15 נק' ו-0 אחרת.
  + דירה תקבל ציון בכפולות של 8 בהתאמה לדירוג (review.rate.number) אשר קיבלה. לדוגמה אם הדירוג של הדירה הינו 3, יינתן לה ציון 24. על כן, 40 הינו הדירוג המקסימלי שדירה יכולה לקבל שכן הדירות מדורגות בסקאלה של 1-5.

להלן עשרת הדירות בעלות הדירוג הגבוה ביותר:

****

הטבלה הנ"ל מספקת מעין סנן חכם אשר יכול לסייע לצרכנים בקבלת החלטה על אודות הדירה אשר בה הם ירצו להתאכסן בעיר ניו יורק בהתחשב במספר פרמטרים חשובים עבורם כגון המחיר, הדירוג ושאר הפרמטרים המפורטים מעלה. מהטבלה ניתן להסיק כי, על אף מספר הדירות הרב הקיים במאגר הנתונים, ישנן מספר דירות בודדות [6] אשר עונות על הקריטריונים באופן מלא וקיבלו את הציון העגול 100.

הציונים בטבלה מתפרסים בין 100 ל -83 וניתן לראות כי עשרת המדורגות הראשונות מגיעות משכונות נפרדות מלבד הדירות אשר דורגו במקומות 5 ו-7 אשר נמצאות בשכונה Kips bay.

חשוב לציין שהחסכוניות בהקשר זה מוגדרת על סמך הגורמים הספציפיים ושיטת הניקוד שהשתמשנו בהם. גורמים שונים או מערכות ניקוד שונות עלולים לגרום לקבוצה שונה של רישומים משתלמים ביותר.

**2.4 שאלה רביעית: מהו המחיר הממוצע של הדירות במערך הנתונים, ועד כמה אנחנו יכולים להיות בטוחים לגבי הערך הממוצע הזה?**

כדי לענות על שאלה זו, נשתמש בטכניקה סטטיסטית הנקראת Bootstrapping.

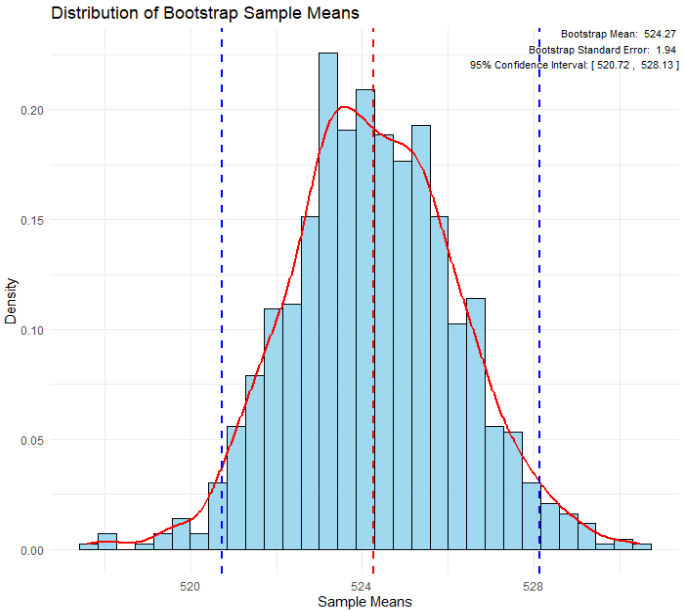
Bootstrapping היא שיטת דגימה מחדש המאפשרת לנו להעריך את השונות של קריטריון המחיר מהנתונים שבידינו. במקרה שלנו, אנחנו מעוניינים במחיר הממוצע של הדירות.

האמוד בשאלה זו הינו המחיר הממוצע האמיתי של כל הדירות הקיימות במערך הנתונים. האומד הוא תהליך של לקיחת דגימת bootstrap (שהיא דגימה בגודל זהה למערך הנתונים המקורי, עם החלפה) וחישוב הממוצע שלו. תהליך זה חוזר על עצמו מספר רב של פעמים (1,000 פעמים) כדי ליצור התפלגות של ממוצעי המדגם. האומדן הוא ממוצעי המדגמים הבודדים המחושבים מכל מדגם bootstrap. אלה מאוחסנים בווקטור bootstrap\_sample\_means. הממוצע של אמצעי הדגימה הללו (bootstrap\_mean) משמש כאומדן סופי של המחיר הממוצע של כל הדירות הקיימות במערך הנתונים. סטיית התקן של אמצעי הדגימה הללו (bootstrap\_standard\_error) משמשת כאומדן של טעות התקן של אומדן זה.

**להלן הסבר שיטת החישוב:**

* אנו לוקחים את מערך הנתונים (only\_homes) המקורי שלנו ובוחנים את קטגוריית מחירי הדירות (price), מוציאים ממנה מדגם אקראי עם החלפה. המשמעות היא שרשומת נתונים מסוימות עשויות להיבחר מספר פעמים ואחרות לא ייבחרו כלל.
* אנו מחשבים את הממוצע של המדגם הזה.
* אנו חוזרים על שני השלבים הראשונים אלף פעמים, בכל פעם שומרים את הממוצע של המדגם.
* בסופו של דבר אנחנו מקבלים אלף הערכות שונות של המחיר הממוצע. חלקן יהיו גבוהות יותר וחלקן יהיו נמוכות מהממוצע של מערך הנתונים המקורי, אך רובן יהיו קרובות לממוצע המקורי.
* אנו מחשבים את הממוצע של אלף ההערכות הללו. זה מספק לנו את ממוצע הדגימה הכולל, שהוא אומדן טוב של הממוצע האמיתי.
* אנו גם מחשבים את סטיית התקן של אלפי האומדנים הללו. זה נותן לנו את שגיאת ה-bootstrap הסטנדרטית, שאומרת לנו עד כמה ההערכה שלנו לגבי הממוצע עשויה להשתנות.
* לבסוף, אנו מחשבים את רווח הסמך של 95% עבור הממוצע. זה מספק לנו טווח של ערכים שאנחנו יכולים להיות בטוחים ב-95% שהם מכילים את הממוצע האמיתי.

להלן צילום מסך של אחת מהרצות הקוד שבוצעו עבור שאלה זו:

****

**ממצאים:**

בהסתכלות על הגרף, אנו יכולים לראות שההתפלגות של אמצעי דגימת ה-bootstrap היא נורמלית בערך, מרוכזת סביב ממוצע הדגימה. הקו האדום המקווקו מייצג את ממוצע הדגימות הכולל, בעוד שהקווים הכחולים המקווקוים מייצגים את הגבול התחתון והעליון של רווח הסמך של 95%.

ממוצע הדגימות הכולל ,ספציפית בדגימה זו (בכל הרצת קוד נקבל תשובה מעט שונה), הוא 524.27 עם שגיאת תקן של 1.94. משמעות הדבר היא שאם היינו חוזרים על הניסוי שלנו פעמים רבות, 95% מהמקרים האומדן שלנו של הממוצע היה נופל בטווח הערכים 520.72-528.13 (95% רווח בר-סמך).

לסיכום, באמצעות שיטת bootstrap, אנו מסוגלים להעריך את המחיר הממוצע של הדירות במערך הנתונים של Airbnb ולכמת את אי הוודאות סביב אומדן זה.

**2.5 שאלה חמישית: מה המתאם הלינארי בין המחיר של הדירות במערך הנתונים only\_homes לבין משתנים אחרים כמו מספר הביקורות(number of reviews), דירוג הדירה (review rate number), זמינות במשך 365 ימים (availability over 365 days), מינימום לילות (minimum nights) ומספר הדירות הכולל הקיים ברשות המארח (calculated host listings count)?**

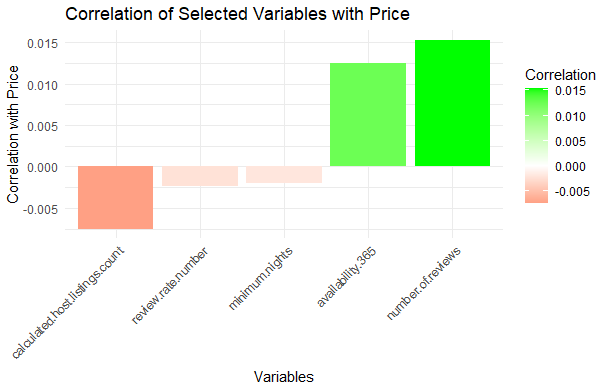
כדי לענות על שאלה זו, אנו הולכים לחשב את מקדמי המתאם בין המחיר לכל אחד מהמשתנים האחרים. מקדם המתאם הוא מדד לחוזק ולכיוון של הקשר בין שני משתנים. הוא נע בין -1 ל-1, כאשר -1 מציין מתאם שלילי מושלם, 1 מציין מתאם חיובי מושלם ו-0 מציין שאין מתאם.

האומד בשאלה זו הוא הקורלציה בין משתנה המחיר לבין שאר המשתנים. האמוד הינו הפונקציה cor ב-R, אשר מחשבת את מקדם המתאם של פירסון בין שני משתנים והאומדן הינו מקדמי המתאם בפועל המחושבים עבור כל זוג משתנים.

**להלן שיטת החישוב:**

* אנו מחשבים את מקדמי המתאם בין המחיר לכל אחד מהמשתנים האחרים באמצעות פונקציית cor ב-R. פונקציה זו מחזירה מטריצה של מקדמי מתאם.
* אנו מחלצים את המתאם עם המחיר (למעט המתאם של המחיר עם עצמו).
* אנו ממירים את הנתונים האלה ל-dataframe עבור יצירת גרף עם ggplot.
* אנו יוצרים גרף ברים, כאשר כל פס מייצג משתנה וגובה הפס מייצג את המתאם של המשתנה הזה עם המחיר. הפסים מקודדים בצבע כדי לציין את חוזק וכיוון המתאם, כאשר ירוק מייצג מתאמים חיוביים, אדום מייצג מתאמים שליליים ולבן מייצג שום מתאם.

להלן גרף העמודות המתאר את הקשר הלינארי בין המשתנים המפורטים מעלה לבין המחיר:



**ממצאים:**

בהסתכלות על הגרף, אנו יכולים לראות את המתאם של כל משתנה שנבחר עם המחיר. המתאמים נעים בין -0.005 ל-0.015 בקירוב. זה מצביע על כך שהקשרים בין המחיר למשתנים האחרים חלשים יחסית, שכן מתאמים קרובים יותר ל-1 או 1- יצביעו על קשרים חזקים יותר.

להלן המתאמים הספציפיים לכל משתנה:

* מספר ביקורות: [0.01532]
* דירוג הדירה [0.00235-]
* זמינות לאורך 365 ימים: [0.01247]
* מינימום לילות: [0.00197-]
* מספר הדירות הכולל הקיים ברשות המארח: [0.00750-]

לדוגמה, מתאם של [0.01532] עבור מספר הביקורות אומר שיש קשר חלש וחיובי בין מספר הביקורות למחיר הדירה.

בפרט, אף אחד מהמשתנים לא מראה מתאם חזק עם המחיר, כלומר ששינויים במשתנים אלו לא צפויים לגרום לשינויים גדולים במחיר.

חשוב לציין שמתאם אינו מרמז על סיבתיות. אמנם משתנים אלו עשויים להיות קשורים למחיר, אך אין זה אומר בהכרח ששינוי אחד יוביל לשינוי בשני. גורמים אחרים שאינם נכללים בניתוח זה עשויים אף הם להשפיע על המחיר.

נתונים אלו מפתיעים כיוון שהיינו מצפים לראות קשר לינארי חזק בין חלק ממשתנים אלו למחיר, לדוגמה שיערנו טרם ביצוע החישוב שככל שהדירוג שהדירה קיבלה הוא גבוהה יותר כך גם המחיר שבעל הדירה גובה יהיה יקר יותר, אך הממצאים מראים שלאו דווקא קיים קשר כזה. אנו מניחים כי אחת הסיבות המרכזיות לחוסר הלינאריות הכללית הינה מספר הרשומות הגדול יחסית במערך הנתונים [102,592] אשר גורם לפיזור הנתונים.

# 3. דיון

ניתוח מחירי שכונה: הניתוח מראה שהמחיר הממוצע של הדירות ב-NY משתנה באופן משמעותי בין השכונות. זה יכול לנבוע מגורמים רבים, כולל מיקום, שירותים, זמינות תחבורה, והאטרקטיביות הכללית של האזור. אף על פי כן ניתן למצוא דמיון במחירי הדירות באיזור המרכזי ביותר בניו יורק.

התפלגות המחירים בכלל NY ובשכונות ספציפיות: עבור כל שכונה ספציפית, התפלגות המחירים עשויה להשתנות בהתאם למאפיינים והדרישות הייחודיות של אותו אזור. הבדלים אלו יכולים לנבוע ממגוון גורמים כגון הבדלים בביקוש, שירותים מקומיים, נגישות לתחבורה, קרבה לאטרקציות תיירותיות פופולריות, והאיכות הממוצעת של דירות ה-Airbnb בשכונות אלו. התובנות הספציפיות פר שכונה יכולות להיות שימושיות במיוחד עבור מארחים פוטנציאליים בשכונות אלה כדי לתמחר את הנכסים שלהם בצורה תחרותית.

לינה חסכונית: ניתוח העלות-תועלת של הדירות מאפשר למשתמשים לזהות את האפשרויות המשתלמות ביותר על סמך ציון מורכב בהתחשב בגורמים כגון מחיר, אימות זהות המארח, אפשרות הזמנה מיידית, מדיניות ביטולים, שנת שיפוץ, מספר ביקורות והדירוג הכללי. ניתוח זה יכול לסייע רבות לנוסעים פוטנציאליים בקבלת החלטות הזמנה מושכלות.

ניתוח Bootstrap: ניתוח Bootstrap מספק אומדן חזק של המחיר הממוצע ברחבי ניו יורק על ידי התחשבות בשונות בנתונים. רווחי הסמך הנגזרים נותנים אינדיקציה לדיוק של אומדן זה. מסקנה סטטיסטית זו מספקת הבנה אמינה של טווח המחירים הכללי, אשר מועילה לכל בעלי העניין.

מתאם מחיר: ניתוח המתאם מגלה שלמחיר הדירות בניו יורק יש מתאם זניח עם שאר המשתנים המרכזיים. זה מרמז על כך שגורמים אלה אינם משפיעים באופן משמעותי על התמחור, מה שמצביע על אסטרטגיית תמחור מורכבת, שאולי מערבת גורמים אחרים שלא נחקרו. ממצא זה הוא קריטי מכיוון שהוא מאתגר את ההנחה המקובלת לפיה גורמים כמו דירוג הדירה או הזמינות שלה משפיעים ישירות על התמחור. לבסוף, עבור מארחים, זה מרמז שפשוט אופטימיזציה של משתנים אלה לא עשויה להעלות משמעותית מחיר הדירה אותו הם ידרשו, ומכאן שתדרוש אסטרטגיית תמחור הוליסטית יותר.

לסיכום, הניתוח המקיף של מערך הנתונים של דירות Airbnb בניו יורק מספק תובנות חשובות לגבי דינמיקת התמחור של הדירות ברחבי ניו יורק. הוא מסייע לבעלי עניין שונים, כולל מארחים, מטיילים וקובעי מדיניות, בקבלת החלטות מושכלות. עם זאת, ייתכן שיהיה צורך בניתוח מעמיק יותר כדי להבין היטב את הגורמים המניעים את ההבדלים בהתפלגות המחירים בין השכונות וכדי לבצע את חיזוי המחיר בצורה מדויקת יותר.