**כלכלה בעולם הביג דאטא – פרויקט סופי**

רקע:

**המשימה:**

בהינתן Dataset של AIRBNB הכולל בתוכו מספר פיצ'רים אשר יפורט אודותיהם בהמשך השייכים ל-user-ים חדשים המצטרפים לשירות, נרצה לחזות האם ה-user-ים הללו יבחרו לעשות booking ראשון בארה"ב (2), booking מחוץ לארה"ב (1) או כלל לא לעשות (0).

**שפת הכתיבה:**

לבחירתנו.

**שפות נבחרות:**

R+Python+SQL

**Features:**

* **id**: user id
* **date\_account\_created**: the date of account creation
* **timestamp\_first\_active**: timestamp of the first activity, note that it can be earlier than date\_account\_created or date\_first\_booking because a user can search before signing up
* **date\_first\_booking**: date of first booking
* **gender**
* **age**
* **signup\_method**
* **signup\_flow**: the page a user came to signup up from
* **language**: international language preference
* **affiliate\_channel**: what kind of paid marketing
* **affiliate\_provider**: where the marketing is e.g. google, craigslist, other
* **first\_affiliate\_tracked**: whats the first marketing the user interacted with before the signing up
* **signup\_app**
* **first\_device\_type**
* **first\_browser**
* **country\_destination**: this is the target variable you are to predict

הוספת משתנה חיצוני:

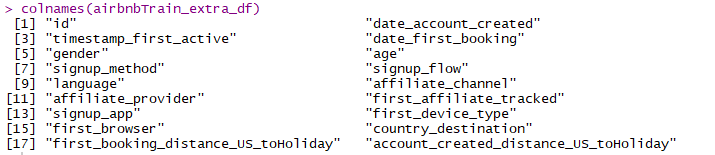
**רקע:**

נתבקשנו להוסיף משתנה חיצוני על מנת להעשיר את ה-Data.

**פיתרון:**

* השתמשנו בסקריפט שנכתב ב-python בשם "script\_for\_adding\_new\_columns.py" עקב נוחות וחבילות פתוחות שהשפה מספקת.
* החלטנו להשתמש בחבילה בשם "Holidays" אשר ניתן בעזרתה לדעת האם תאריך מסויים הוא חג או לא (בהתאם למדינה נבחרת).
* מכיוון שה-users הינם מארה"ב בחרנו חגים שחלים בארה"ב. בנוסף, לא בחרנו חגים במדינות אחרות, מכיוון שהחגים משתנים ממדינה למדינה ותוצאות ה-dataset מופרדות להאם המדינה היא ארה"ב, מחוץ לארה"ב או לא נעשה booking בכלל.
* בהתאם לכך, השתמשנו בחבילה זו (אשר מפורסמת בתור חבילה רשמית כחלק מ-pypi).
* בעקבות זאת, יצרנו 2 פיצ'רים חדשים המתאימים לפיצ'רים קיימים (מופיע בסוגריים):
  + **account\_created\_distance\_US\_toHoliday** (date\_account\_created)
  + **first\_booking\_distance\_US\_toHoliday** (date\_first\_booking)
* המשתנים הללו סופרים מה מספר הימים המינימאלי בין הפיצ'ר הקיים לבין החג הקרוב הקיים בארה"ב. כלומר, למשל עבור כל תאריך dateX בפיצ'ר **date\_account\_created** קיימת מקבילה בפיצ'ר החדש שיצרנו **account\_created\_distance\_US\_toHoliday** ובו ההפרש המינימאלי **בימים** מהתאריך ה- dateX ועד ה-closest holiday.
* הרציונאל העומד מאחוריי כך הוא ההנחה שכאשר מתקרבים לחגים, אנשים רוצים להזמין מקומות לינה ב-airbnb (ייתכן בהתאם לחג). תחת ההנחה כי אנשים נוטים להזמין מראש רצינו למצוא את הפרש הימים המתאים.

רשימת פיצ'רים התחלתית:



נתונים כלליים על ה-Dataset:

* **מספר פיצ'רים התחלתי:** 18
* **מספר תצפיות:** 170760

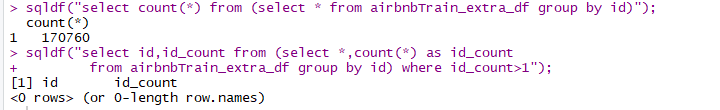
PreProcessing:

**רקע:**

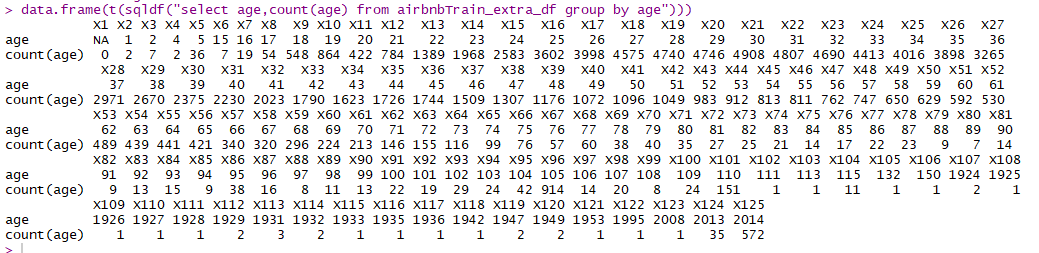
בחלק זה נעבור על הפיצ'רים ונעבד אותם על מנת שנוכל להשתמש בהם במודל עתידי.

**id**:

* להלן בדיקה המראה שכל הנתונים הינם ייחודיים ואין כפילויות ב-IDים:



**Age:**

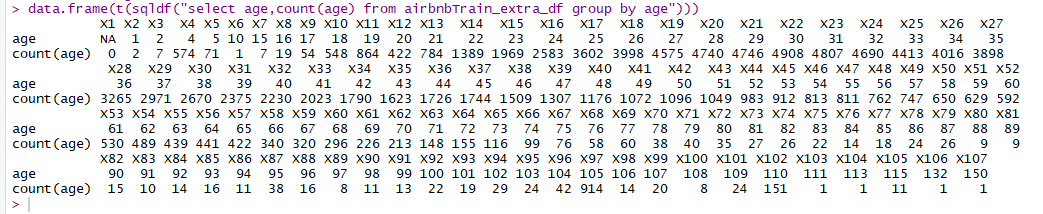
* בבדיקה ראשונית של ה-data על מנת לראות התפלגות, נראה כי ישנם ערכים עפ"י שנת לידה ולא עפ"י גיל:

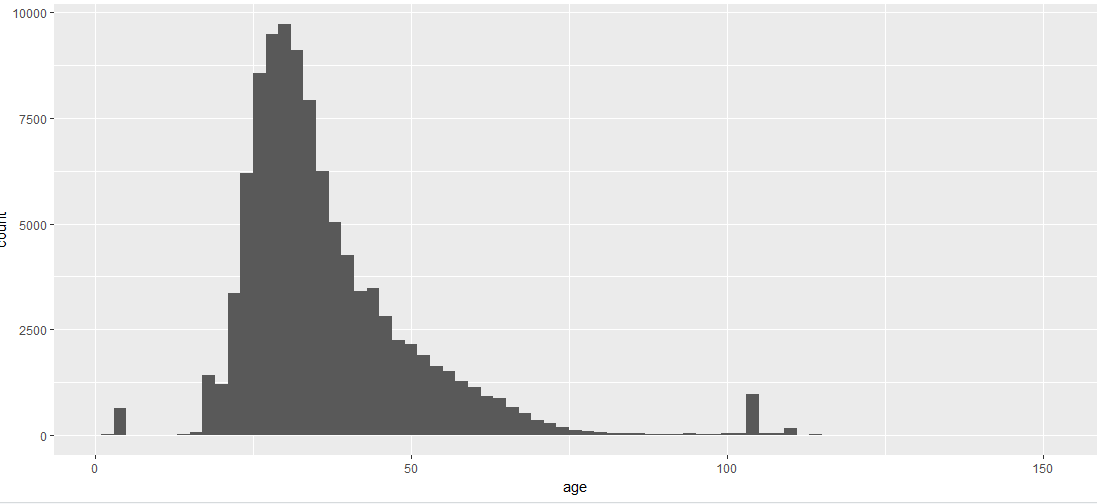


* + קיימים 70398 ערכים חסרים (NA)
  + קיימים גילאים בשנים (החל מ-1900)
  + קיימים גילאים לא הגיוניים כמו 132 ו-150. כמו-כן הסבירות שאנשים בני 100 ומעלה או בני 15 ומטה ישתמשו ב-airbnb ויעשו booking היא נמוכה מאוד.
* לכן, נחליף את כל הנתונים תחת עמודת age שעבורם הגיל הוא מעל 1900 ל-2018-age, אחרת נשאיר כרגיל:

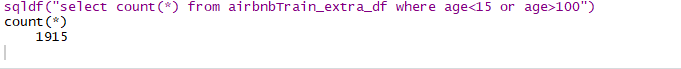


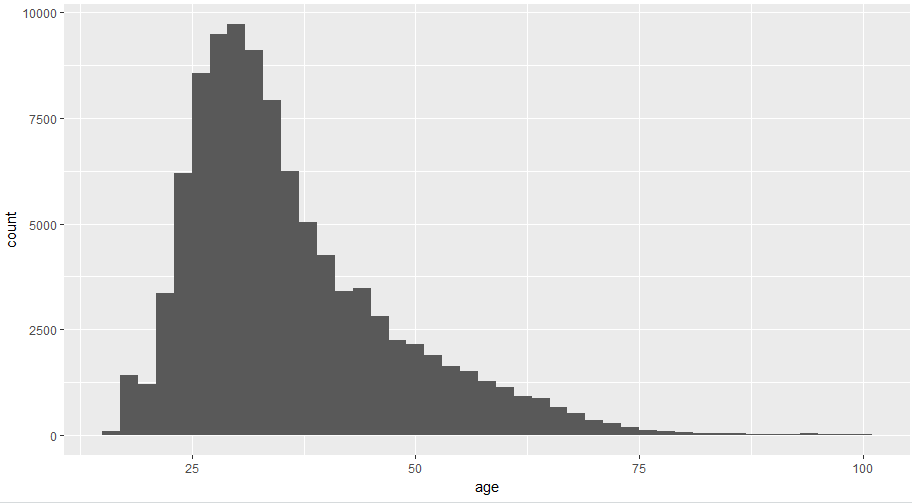
* כעת ההתפלגות יותר הגיונית:

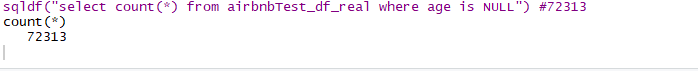




* ניתן לראות שיש כאן התפלגות זנב-עבה יחד עם outliers.
* נבדוק כמה כאלה קיימים שהם מתחת לגיל 15 ומעל גיל 100:



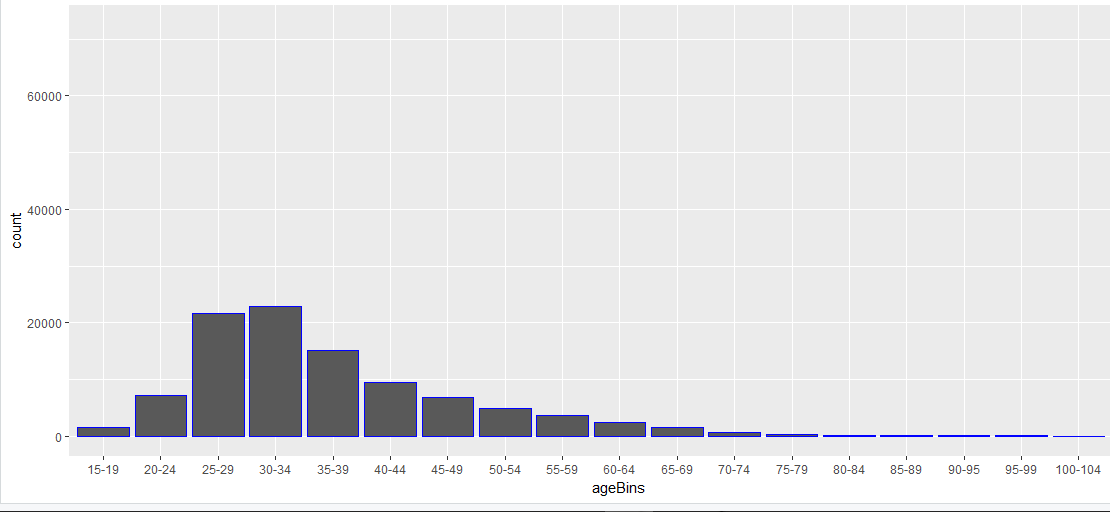
* בחרנו לשנות את כל מי שבטווח הגילאים הנ"ל לערכים שהם NA (אשר קיימים כבר כ-70,000 כאלה) מ-2 סיבות:
  + 1. מכיוון שיש יותר מ-10% מה-data אנשים בגילאים האלה, לחתוך אנשים אלה מה-data עלול להשפיע יותר מדי על ה-data.
    2. ייתכן ואנשים שציינו גיל לא הגיוני, יציינו נתונים נוספים לא הגיוניים. אולם, מבדיקה שנעשתה, שאר הנתונים שלהם הגיוניים לכן נעדיף לא לחתוך את ה-users האלה מה-data.
* אופציה אחרת היא להפוך אותם לערך שאינו בשימוש – למשל את כולם לערך 1-, זאת כי ייתכן וישנה משמעות לערכים חסרים, או משמעות ל-outliers. אך ככל הנראה מדובר בטעויות או שאנשים רשמו או גיל מזוייף ולכן זה סביר והגיוני להפוך אותם ל-NA במקום לערך משלהם.
* בסה"כ קיימים 72313 ערכי NA בפיצ'ר age.



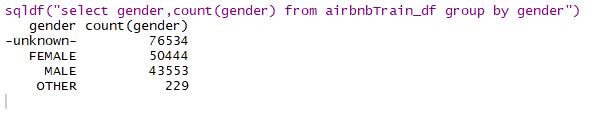
* סיכום של ה-data ב-age:

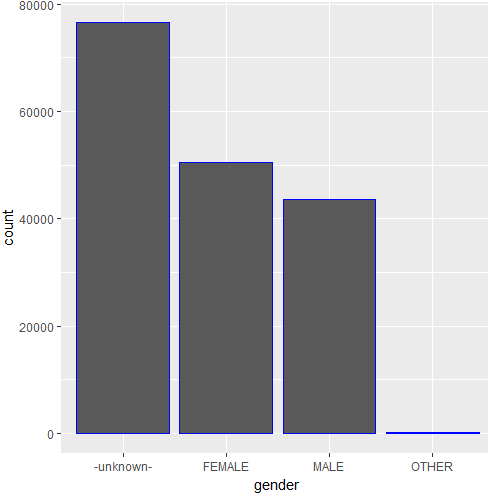
**ageBins**

* החלטנו ליצור משתנה נוסף קטגוריאלי עבור age אשר ייתכן וישמש אותנו בהמשך בשם ageBins.
* משתנה זה מאחד כל 5 שנים לידי קטגוריה אחת . לדוגמא, גילאים 15-19 יאוחדו תחת קטגוריה אחת.
* הייתה דילמה האם לחלק זאת ל-bins עפ"י כמות, כלומר שכל bin יכיל את אותה כמות של משתנים, או לחלק לפי גילאים. לבסוף החלטנו לחלק לפי גילאים. אמנם הדבר יוצר bias יותר גדול ול-bins שהם outliers ישנה יותר חשיבות, אך יותר הגיוני לחלק את הגילאים לפי קבוצות גיל ולא לפי כמות שווה של data.
* **יש לזכור** כי גם בפיצ'ר זה קיימים 72313 ערכי NA.
* להלן גרף ההיסטוגרמה של ageBins:



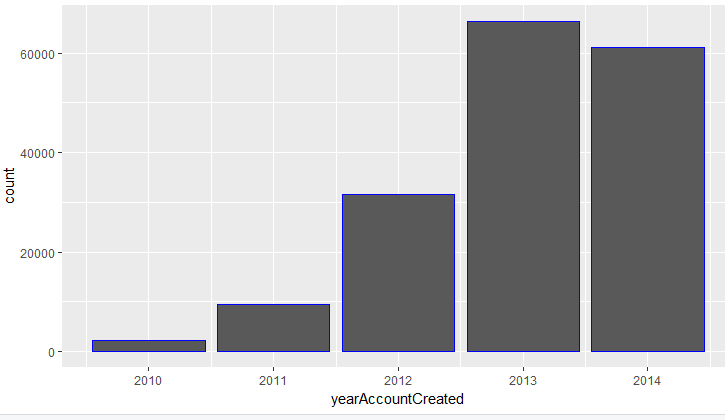
**gender**

* ניתן לראות את החלוקה לפי מגדר באופן הבא:

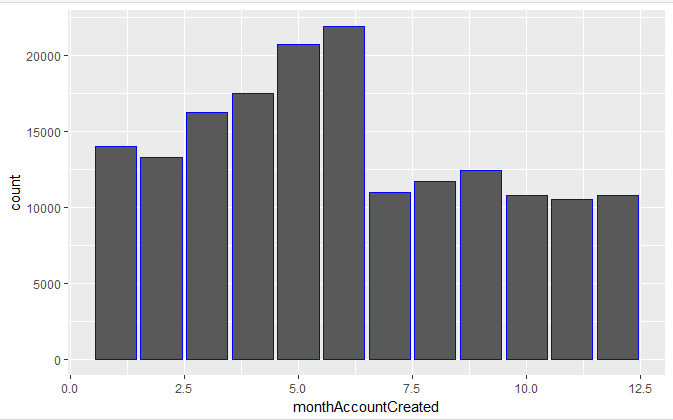


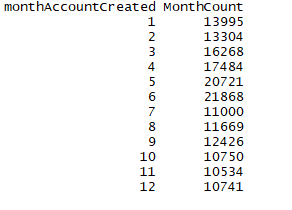
* ניתן לראות שיש כ-75,000 אנשים אשר צויינו כ-"-unknown'" לכן נהפוך אותם מ-string ל-NA.
* כמו-כן ניתן לראות כי יש כ-229 תצפיות של OTHER. אמנם מדובר במספר תצפיות מועט (כ-%0.001 מה-train), אולם ייתכן והוא בעל משמעות עבור אנשים שאינם רואים את עצמם שייכים למגדר. בנוסף, מבדיקה שנערכה על שאר הפיצ'רים של ה-user-ים בעלי Gender של OTHER, לא נראו דברים חריגים. על-כן, הוחלט להשאיר את OTHER כמו שהוא.
* כמו-כן, הוחלט להפוך את GENDER למשתנה קטגוריאלי (ממשתנה מסוג character) בעל 4 קטגוריות.

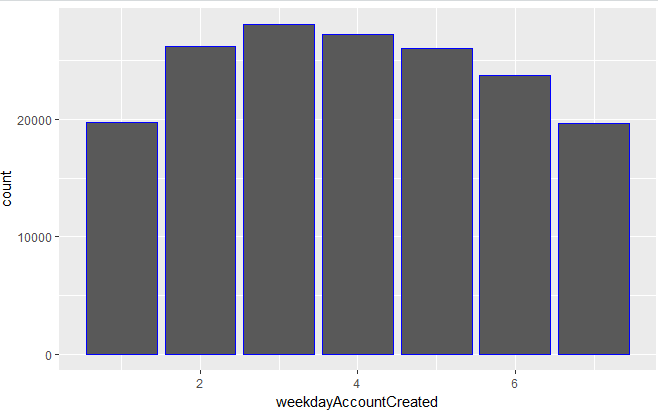
**date\_account\_created**

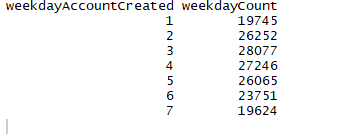
* בפיצ'ר זה לא קיימים כלל ערכי NA.
* יצרנו פיצ'ר חדש זהה בפורמט Datetime בשם **dateAccountCreated**.
* על בסיס פיצ'ר **dateAccountCreated** יצרנו 3 פיצ'רים נוספים בשם:
  + **yearAccountCreated** – השנה בה נוצר החשבון.
  + **monthAccountCreated** – החודש בו נוצר החשבון.
  + **weekdayAccountCreated** – היום בשבוע בו נוצר החשבון.
* **yearAccountCreated**:
  + ניתן לראות שרב המידע הוא בין השנים 2013 ו-2014:
  + ניתן לראות במספרים את כמות הרשומות מהתפלגות השנים:



* **monthAccountCreated**:
  + ניתן לראות שאין אינדקציה לחריגות בחודשים:
  + ומבחינה מספרית:



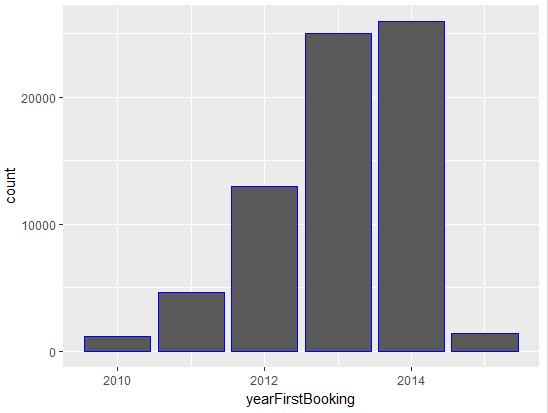
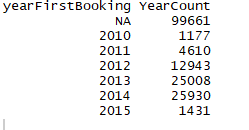
* **weekdayAccountCreated**:
  +  ניתן לראות שאין אינדקציה לחריגות מבחינת ימות השבוע:
  + ומבחינה מספרית:

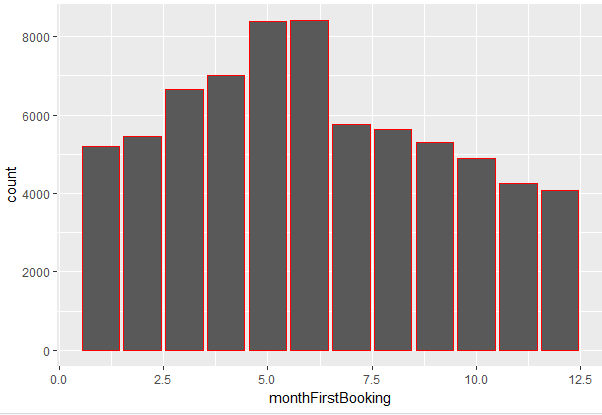


**date\_first\_booking**

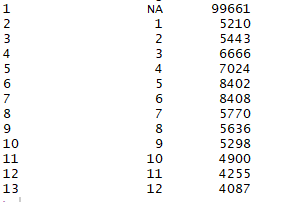
* בפיצ'ר זה כלל לא קיימים ערכי NA אולם מכיוון שהוא מוגדר ב-character, ערכי ה-NA שלו הושמו כמחרוזת ריקה "". מבדיקה של הכמויות עולה כי קיימים 99661 ערכים כאלה. את ערכים אלה נמיר להיות NA.



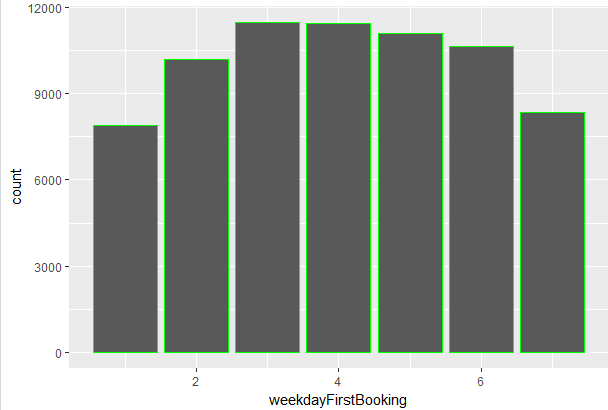
* יצרנו פיצ'ר חדש זהה בפורמט Datetime בשם **dateFirstBooking**.
* על בסיס פיצ'ר **dateAccountCreated** יצרנו 3 פיצ'רים נוספים בשם:
  + **yearFirstBooking**– השנה בה התרחש ה-booking הראשון.
  + **monthFirstBooking**– החודש בו התרחש ה-booking הראשון.
  + **weekdayFirstBooking**– היום בשבוע בו התרחש ה-booking הראשון..
* **yearFirstBooking**:
  + ניתן לראות שרב המידע הוא בין השנים 2013 ו-2014:
  + ניתן לראות במספרים את כמות הרשומות מהתפלגות השנים:
* **monthFirstBooking**:
  + ניתן לראות שאין אינדקציה לחריגות בחודשים NA:



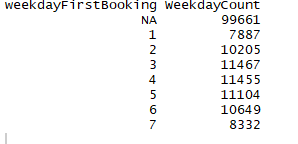
* + ומבחינה מספרית:



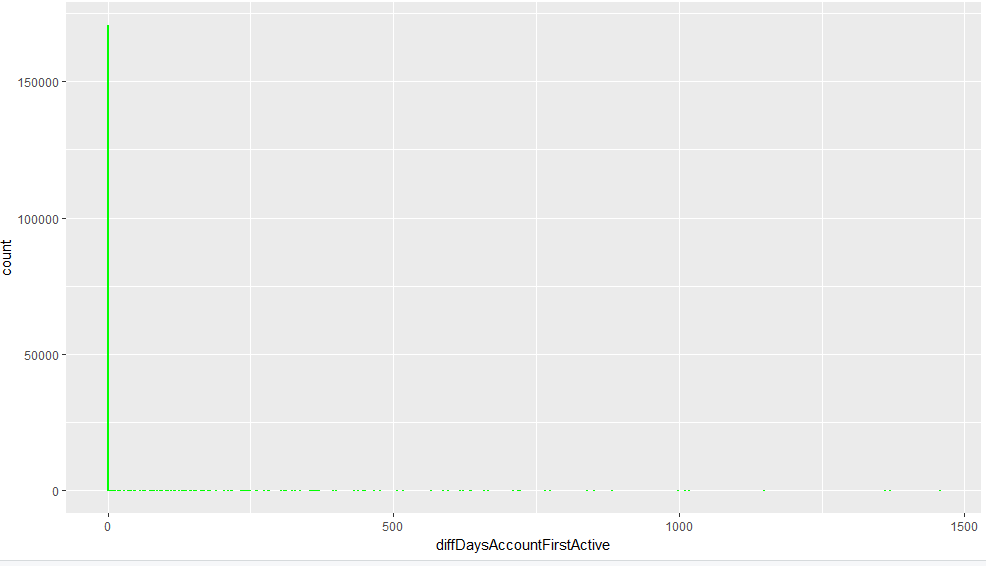
* **weekdayFirstBooking**:
  + ניתן לראות שאין אינדקציה לחריגות מבחינת ימות השבוע:

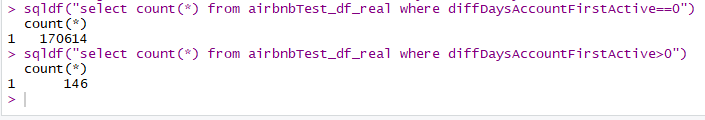


* + ומבחינה מספרית:

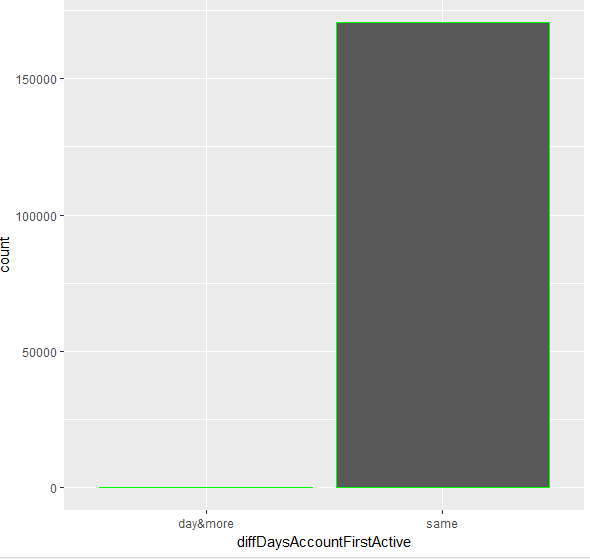


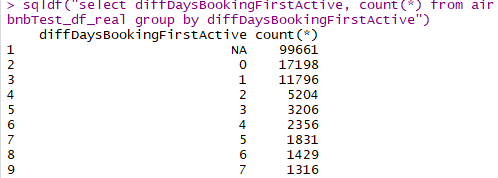
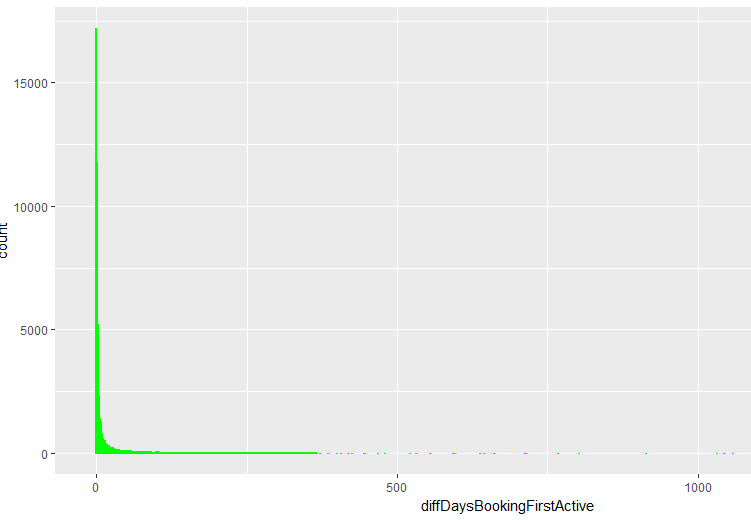
**Timestamp\_first\_active**

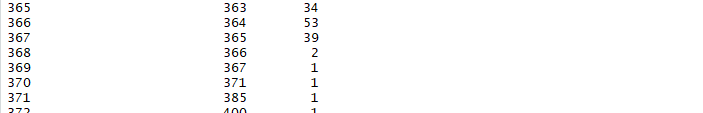
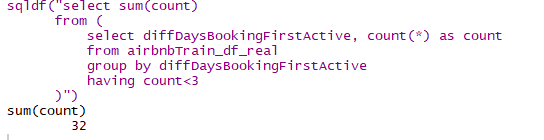
* בפיצ'ר זה מדובר על מתי ה-User היה active בפעם הראשונה.
* נשים לב כי אין כלל ערכים חסרים בפיצ'ר זה.
* ניצור 2 משתנים חדשים ממנו:
  + **timestampFirstActive ­**– נמצא בפורמט רגיל של תאריך
  + **dateTimestampFirstActive** – חילוץ של התאריך ללא מרכיב ה-time
* הוספת פיצ'רים נוספים רלוונטיים:
  + **diffDaysAccountFirstActive** – ההבדל במספר הימים מהרגע שהחשבון נפתח ועד לרגע שהייתה בו פעילות ראשונה.
    1. ניתן לראות שרב פתיחת החשבונות התרחשה באותו יום של הפעילות הראשונה, אך ישנם בודדים שהדבר אינו כך עבורם, כלומר פתיחת החשבון התרחשה יום ומעלה לאחר הפעילות הראשונה באתר:
    2. ניתן לראות שישנו רב מוחץ. ישנם 170614 ערכים עבורם פעילות ראשונה ופתיחת חשבון התרחשן באותו היום ו-146 בימים שאחרי:



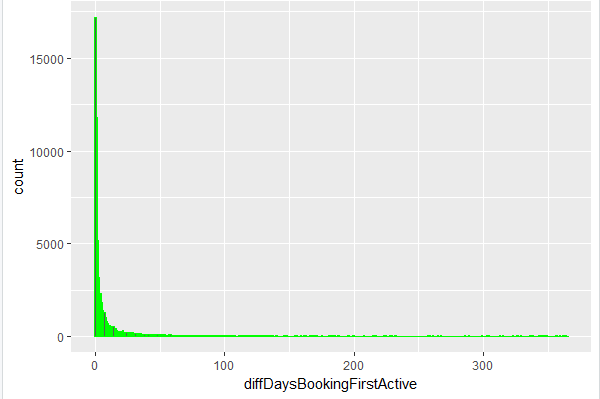
* + 1. לכן, החלטנו להפוך את המשתנה למשתנה קטגוריאלי בינארי אשר יכיל עמודת אנשים שנרשמו באותו יום של הפעילות הראשונה וכאלה של יום אחרי ומעלה.



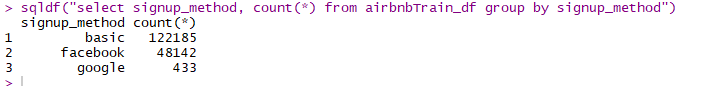
* + **diffDaysBookingFirstActive –** ההבדל במספר הימים מהרגע שהיה booking ראשונה ועד לפעילות הראשונה בחשבון.
    1. במקרה זה יש לנו התפלגות זנב, לכן החלטנו להשאיר זאת באותו אופן שהתקבל ולא להפוך למשתנה קטגוריאלי:
* נשים לב שבהתפלגות הזנב יש הרבה outliers. מכיוון שכך ולאחר בדיקות נראה כי לאחר 365 יום הכמויות מתחילות להיות 1 או 2 ובסה"כ קיימות 32 רשומות כאלה. מכיוון שאלו הן לא הרבה רשומות, נסיר אותן.

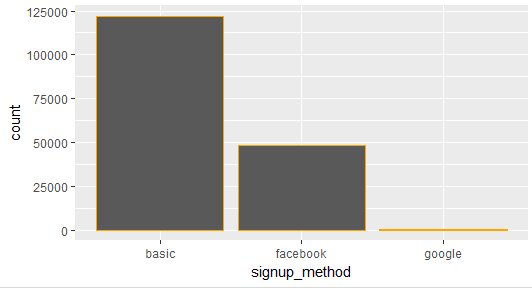
 

* ולאחר השינוי:



**Signup method:**

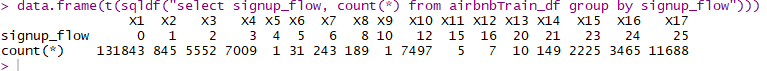
* בפיצ'ר זה מדובר דרך איזו שיטה נרשם ה-user לאתר Airbnb.
* נראה את ההתפלגות של הערכים:
* נשים לב כי אין כלל ערכים חסרים בפיצ'ר זה.
* כמו-כן נשים לב שהשיטה הכי נפוצה היא דרך האתר של Airbnb.

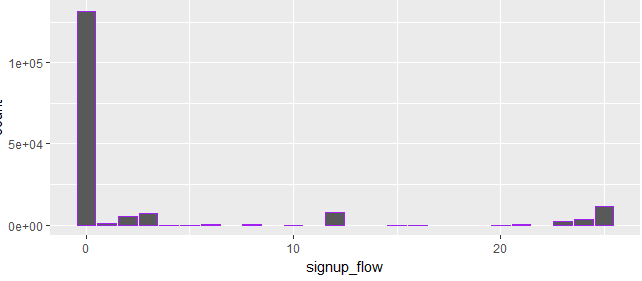


* מכיון שישנם רק 3 ערכים, נהפוך פיצ'ר זה לקטגוריאלי.

**Signup flow:**

* בפיצ'ר זה מדובר בכמה עמודים ה-user עבר עד שנרשם לאתר Airbnb.
* בפיצ'ר זה אין ערכי NULL.
* ניתן לראות את ההתפלגות:

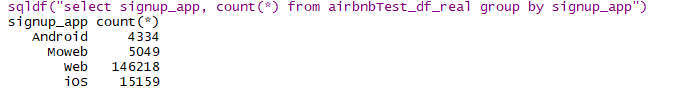


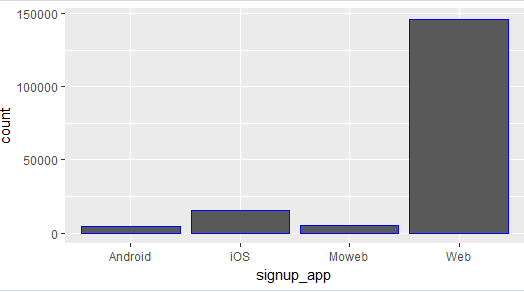


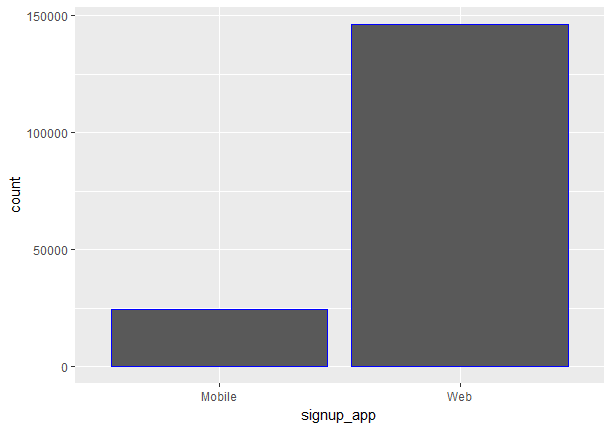
* נשים לב שהרב המוחלט עבר 0 דפים וכן שישנם מספרים בטווח שאין בהם כלל userים.

**נתוני signup app :**

* בפיצ'ר זה אין ערכי NA.
* ניתן לשים לב שישנם 4 קטגוריות: 3 השייכים לסוג ה-mobile ואחד ל-web:

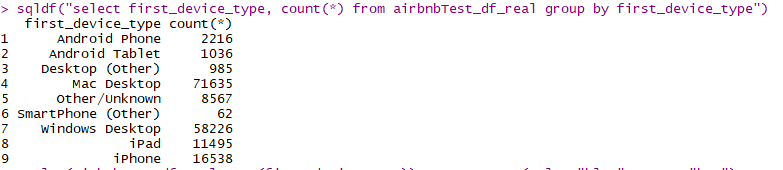


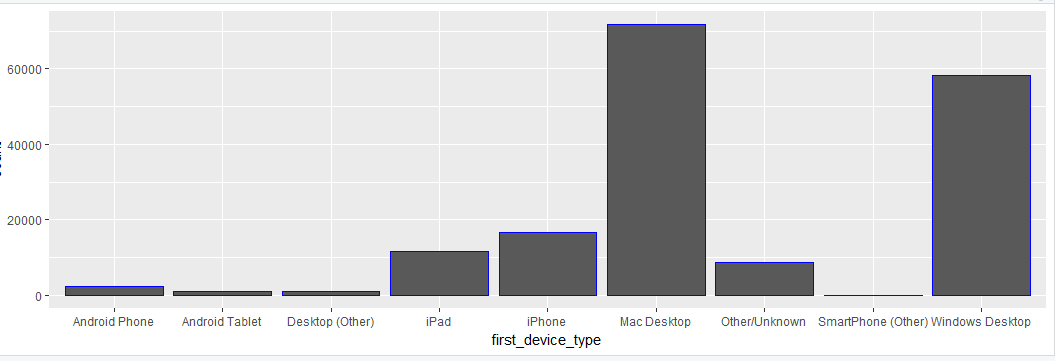


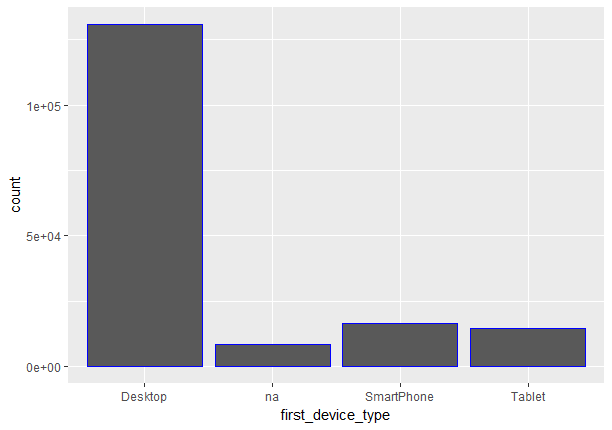
* איחדנו את iphone, android, moweb תחת mobile כי נראה לנו שההבדל המהותי הוא בסוג האינטראקציה של המשתמש עם airbnb – mobile או web.
* נראה שרוב המשתמשים נכנסים airbnb דרך web:
* בנוסף הפכנו את המשתנה לקטגוריאלי.

**first device type :**

* להלן התפלגות הנתונים בפיצ'ר זה:

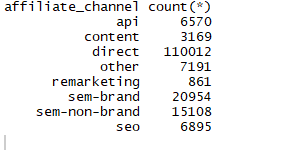


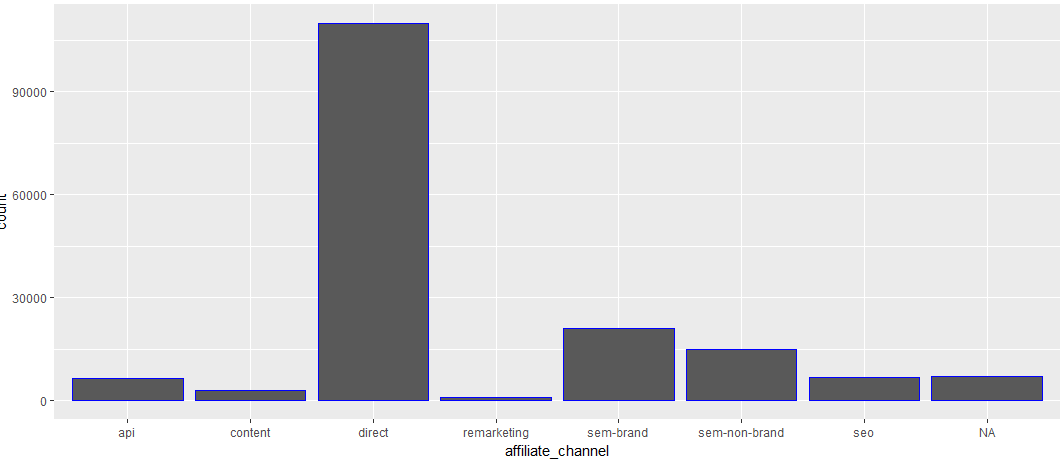


* ניתן לשים לב שיש כ-8567 ערכים שהם לא ידועים, אולם רשומים בתור String. לכן, נהפוך אותם ל-NA.
* איחדנו את iphone, android, smartphone תחת SmartPhone, ואת ipad, androidtablet תחת Tablet כי נראה לנו שההבדל המהותי הוא בסוג המכשיר. נראה שרוב המשתמשים נכנסים לAIRBNB דרך הDesktop, כלומר לא מהכשירים שזמינים לנו בכל סביבה.
* יתכן שזה מעיד על כך שכשאנשים מזמינים דירה הם עושים זאת אל מול מחשב נוח.
* הפכנו משתנה זה למשתנה קטגוריאלי.

**affiliate channel**:

* נראה שסוג השיווק בתשלום שממנו מגיעים הכי הרבה משתמשים הוא ישיר.
* ישנם 7191 ערכים שמוגדרים כ-Other. הפכנו אותם ל-NA.

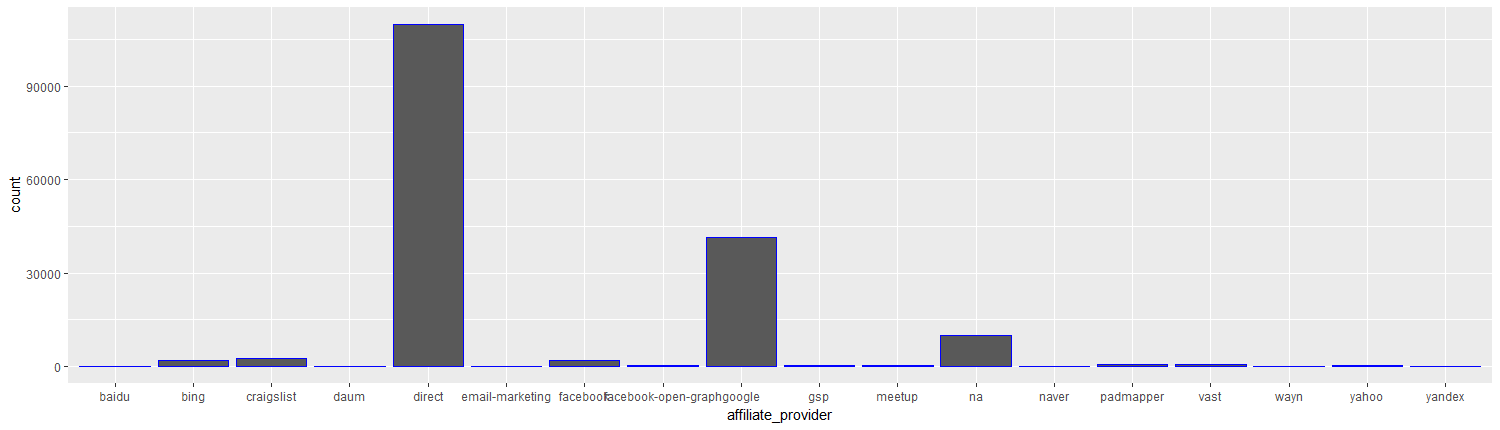




* גם משתנה זה הפכנו לקטגוריאלי.

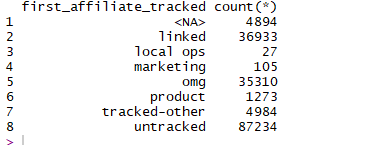
**affiliate provider :**

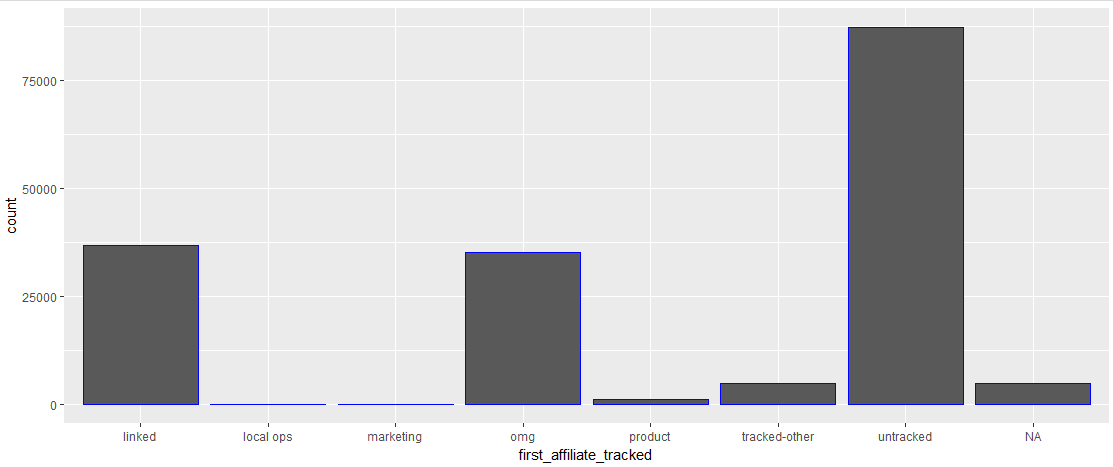
* נראה שמיקום השיווק של רוב המשתמשים הוא ישיר. יתכן שזה מעיד על כך שמי שנכנס עשה זאת במכוון ולא במקרה, כלומר יש בכוונתו להשכיר דירה או לפחות לבחון את האופציה.
* בוצע אותו טיפול עבור ערכים שהם "other" (הפיכה ל-NA) בדומה ל-affiliate channel.
* גם משתנה זה הפכנו לקטגוריאלי.



**first affiliate tracked:**

* ל-4894 מהתצפיות אי אפשר לדעת מה היא האינטראקציה השיווקית שגרמה להם להגיע לairbnb (תצפיות אשר הוגדרו כ-string ריק). לכן הפכנו אותן ל-NA.
* נשים לב כי קיימים בפיצ'ר זה גם untracked, גם NA וגם tracked-other. החלטנו להשאירן ב-3 קטגוריות נפרדותלמקרה שהן בעלות משמעות שונה.
* להלן התפלגות הנתונים:

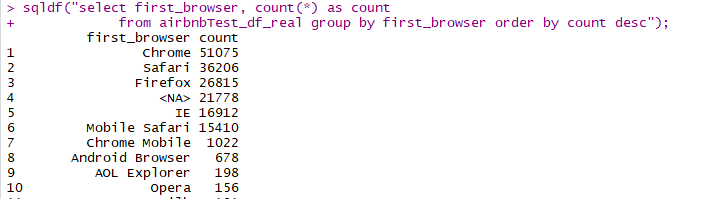




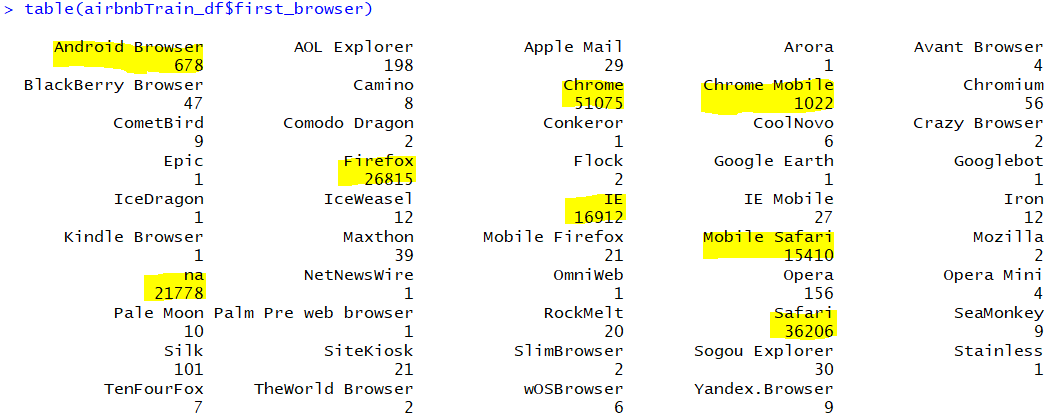
* אמנם יש משתנים שמופיעים מעט פעמים כמו local ops או marketing, אך בחרנו לא להשמיטם כי זוהי קטגוריה שבה יש מספר תצפיות מועטות ולא חריגה כלשהי מה-data.
* גם משתנה זה הפכנו לקטגוריאלי.

**first browser :**

* ישנם כ-21,778 משתנים שהם unknown--. הפכנו אותם ל-NA.
* נראה שהאפליקציות המובילות הן chrome, safari, firefox ו-IE:



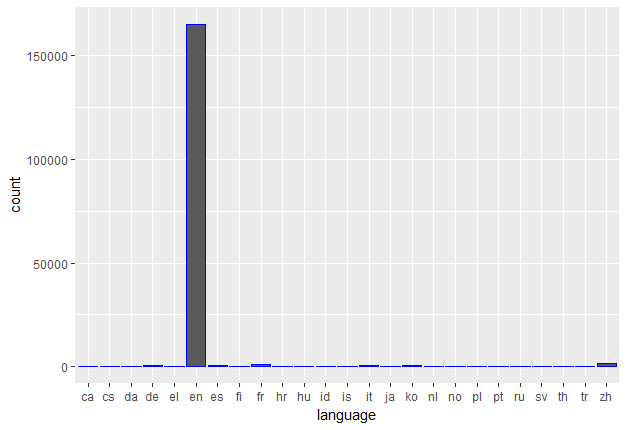
* על פלטפורמת desktop נראה שאין הבדל משמעותי בין PC ל Mac, אולם בsmartphone נראה שהרבה יותר משתמשי iphone נכנסים לairbnb מאשר משתמשי android.



* גם משתצנה זה הפכנו לקטגוריאלי.

**language:**

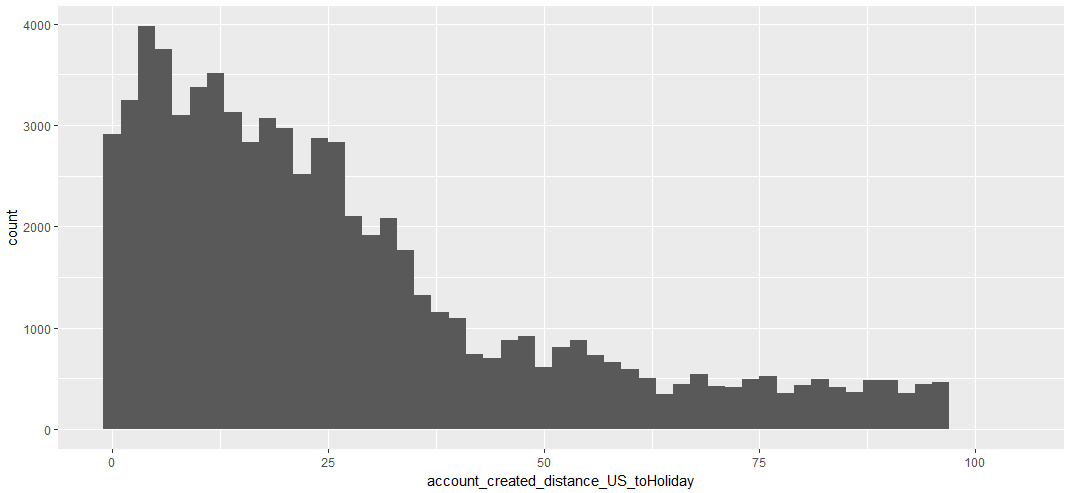
* ניתן לראות שהרוב המוחלט של המשתמשים הינו דובר אנגלית:



* נשים לב גם שאין ערכים לא ידועים בפייצ'ר זה.
* אמנם יש שפות עם מעט תצפיות, אך המשמעות הינה חשובה ולכן לא נשמיט תצפיות אלה.
* בנוסף ייתכן כי השפה משפיעה על ה-destination הסופי ולכן החלטנו לא לאחד ביחד את כל השפות ש**אינן** אנגלית לידי קטגוריה אחת.
* גם פיצ'ר זה נהפוך לקטגוריאלי.

**account\_created\_distance\_US\_toHoliday:**

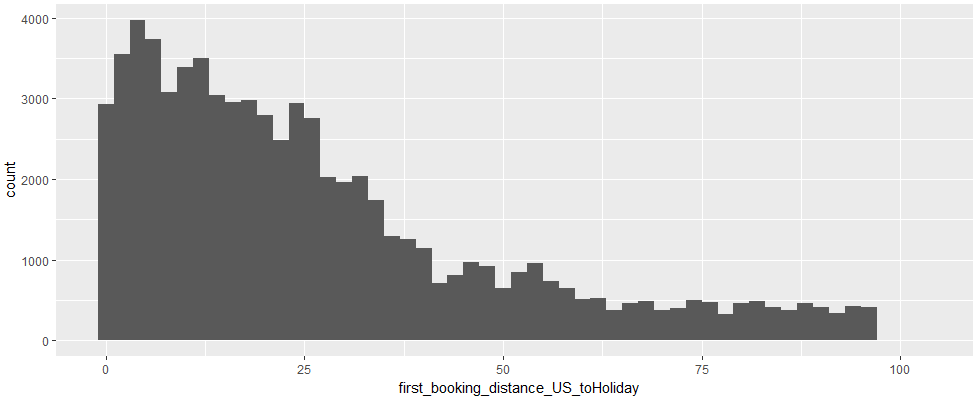
* למשתנה זה אין ערכי NULL.



* ניתן לראות כי ישנה התפלגות מעניינת ביחסת לכמות מבחינת מרחק הימים לחג הכי קרוב כאשר ניתן לראות מגמה שככל שהחג יותר קרוב, יותר חשבונות נפתחו.

**first\_booking\_distance\_US\_toHoliday**:

* למשתנה זה יש 99661 ערכיNULL, כמובן בהתאם לערכי ה-date\_first\_booking.

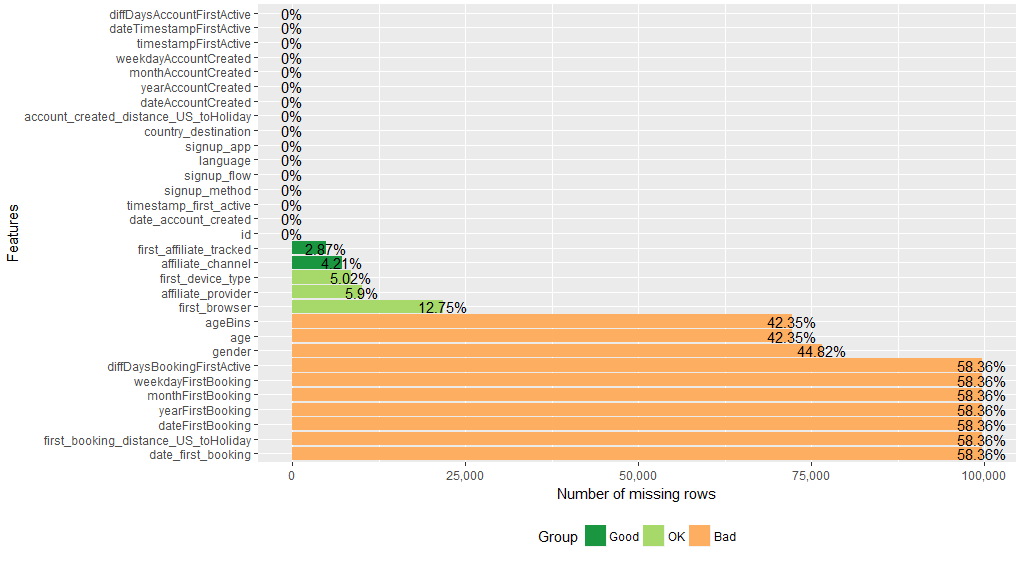


* גם כאן נית לראות קשר ברור לככל שהמרחק לחג הכי קרוב קצר יחסית, יש מגמה של יותר הזמנות (בדומה ליצירת חשבון).

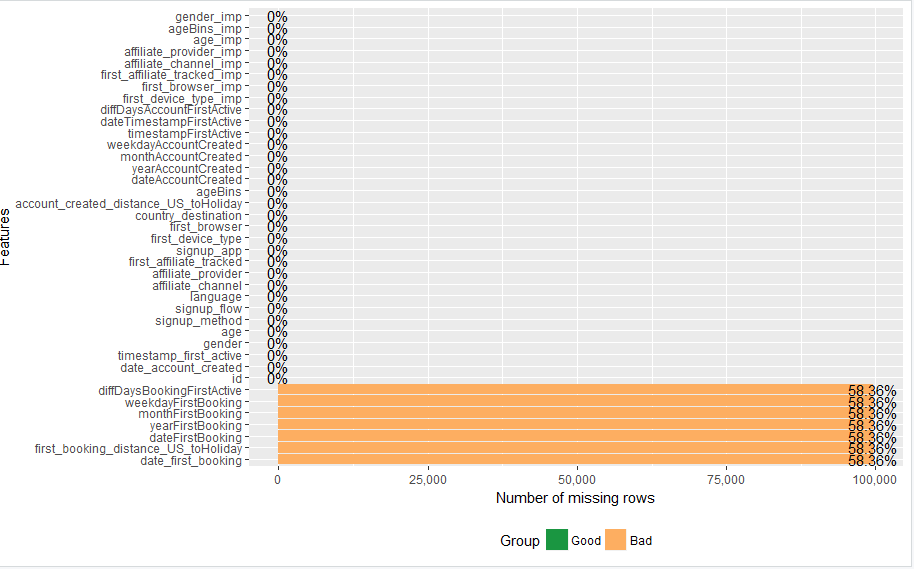
**נדגיש:**

כל המשתנים שהיו NA נבדקו על מנת לראות האם יש משהו מחשיד בתצפיות על מנת שיהיה ניתן להסירן במידת הצורך. אולם, לא התגלה דבר חריג בנוגע לתצפיות שבהן היו ערכי NA.

**טיפול בערכי NA (missing data):**

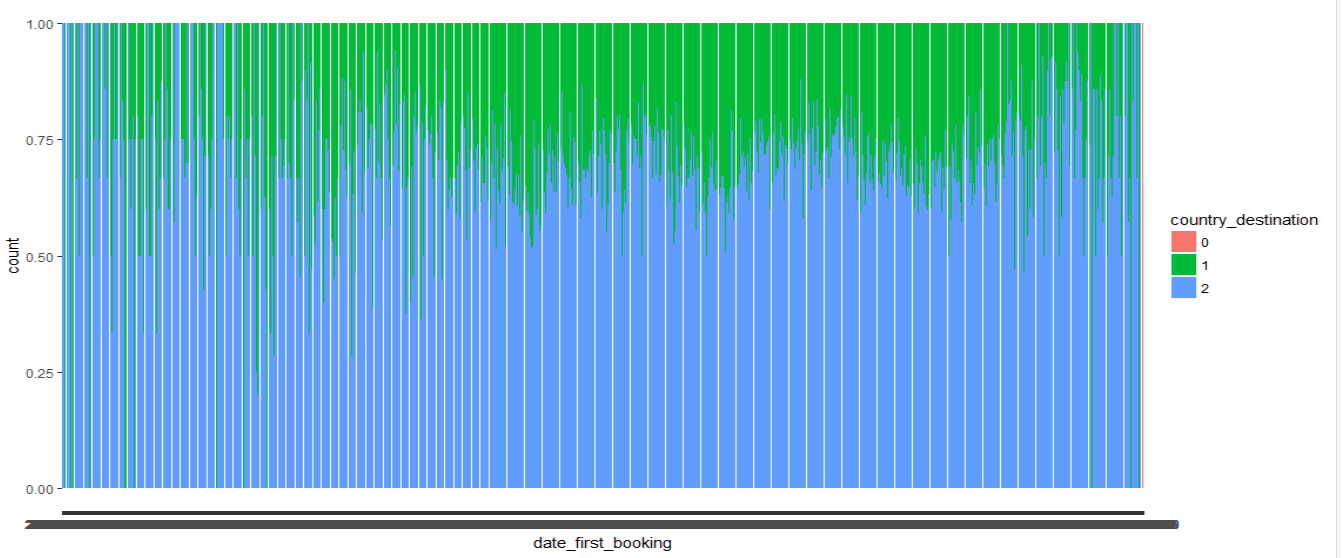
* קיימות מספר שיטות לטיפול בערכי NA:
  + במשתנים נומריים – לקחת את החציון/ממוצע – עלול ליצור bias גבוה עבור המשתנים אם יש הרבה data (כמו במקרה שלנו).
  + ניתן להשמיט – ב-data set שקילבלנו קיימות יותר מדי תצפיות מכדי להשמיט.
  + להתייחס אליו כאלה משתנה קטגורי (במקרים בהם ניתן).
  + לקבלו כחלק מה-data.
  + KNN imputation – שימוש באלגוריתם KNN על מנת להשלים את הערכים החסרים.
  + Sampling – לזהות את התפלגות הפיצ'ר בו יש ערכים חסרים ומהם, ליצור פונקציה שממפה את הערכים להתפלגות, לבחור ערך רנדומאלי מההתפלגות ולהשתמש בפונקציה ההופכית כדי לחשב את ערך המשתנה החסר.
  + שימוש ב-EM לצורך הערכה.
* ב-data שלנו יש משתנים רבים אשר מכילים ערכי NA. להלן פירוטם ויחסיהם ב-data:
* נשים לב שישנם משתנים שקשה או לא רצוי להשלימם בצורה זו – למשל כמו כל המשתנים המבוססים על הפיצ'ר date\_first\_booking (כל המשתנים בעלי 58.36% מדיע חסר). מכיוון שמעולם לא בוצע booking עפ"י נתוני ה-train, נכון לאותו session, אז המידע הזה חסר ונרצה להשאירו כך.
* לעומת זאת משתנים אחרים כמו פיצ'רי ה-affiliate או משתני ה-first או משתנה ה-age נרצה לטפלם בדרך כלשהי.
* נסיונות לבצע knn imputation כשלו עקב מגבלות טכניות של האמצעים אשר עמדו לרשותנו ומשך זמן החישוב הארוך הדרוש לכך.
* נסיונות לבצע EM גם כשלו בהיעדר חבילות מספיק טובות אשר ישמשו לדבר.
* על-כן, הוחלט להשתמש בשיטת imputation בשם **hotdeck**. שיטה זו אינה גזלה זמן.
* אולם, כיוון שאין אנו בטוחים בתוצאותיה, נבצע את המודלים על 2 סוגי dataset.
  + 1. ה-dataset עליו בוצע hoteck ובו לא קיימים ערכי NA.
    2. Dataset בו ערכי ה-NA נחשבים ל"label" כחלק ממשתנה קטגוריאלי לכל דבר. בכל המשתנים אין עם הדבר בעיה מלבד בפיצ'ר age. זאת ניתן לפתור ב-2 דרכים:

1. נשלים את הערכים באמצעות ממוצע.
2. נשתמש במשתנה ה-ageBins במקום במשתנה age אשר מחלק את ה-age לקטגוריות לפי גילאים ובו NA נחשב לעוד label (הדרך שנבחרה לבסוף).

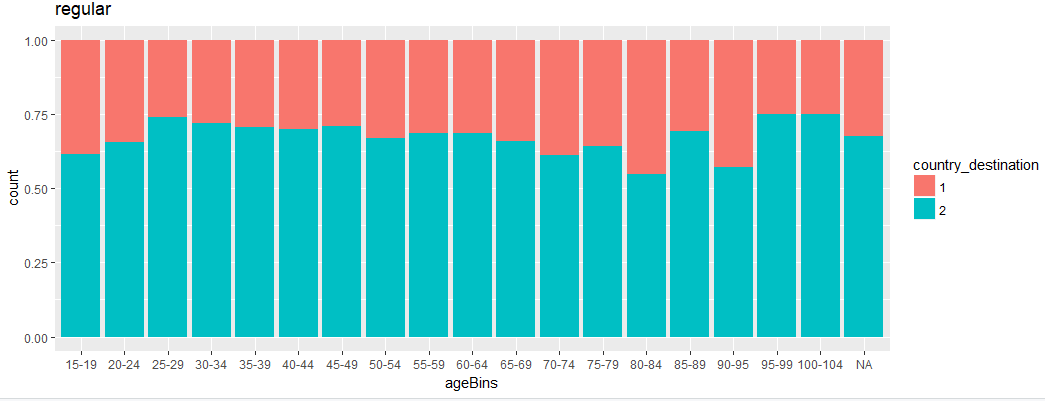
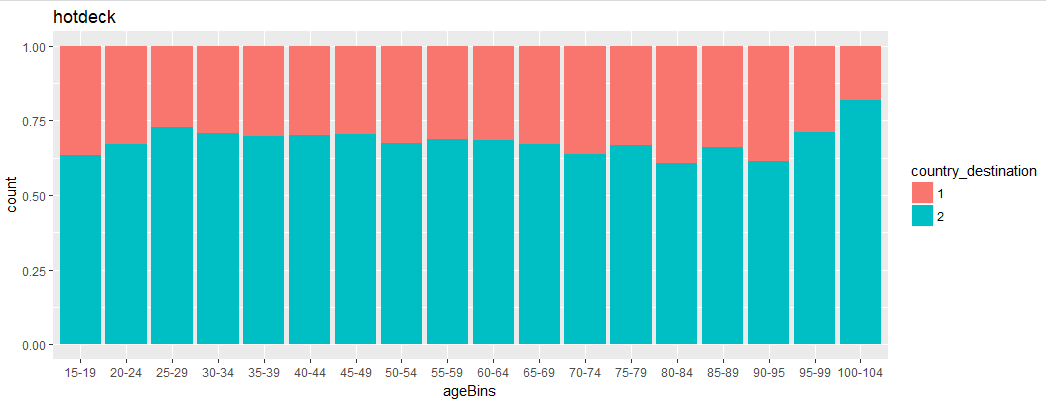
* כך נראה המידע לאחר ה-hotdeck:
* יש לשים לב כי בכוונה השארנו את ה-booking עם ערכי NA. זאתף כיוון שבעתיד אנו מתכוונים להשתמש באופן ישיר בעובדה זו על מנת לתת "0" (כלומר אי הזמנה) למי שלא עשה booking ולהשתמש בכל מי שכן יש לו booking date בשביל להכריע אם הוא יזמין לארה"ב (2) או אל מחוץ לארה"ב (1) .

**קורלציה וקשרים בין המשתנים:**

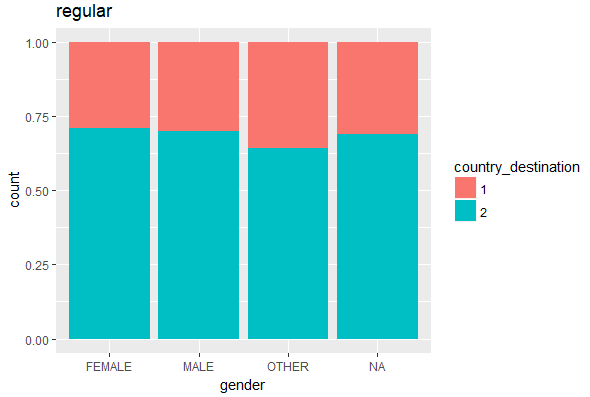
* ראשית ננסה באמצעות bar plots שונים לראות קשר בין משתנים מסויימים מובהקים לבין התוצאה המיועדת.
* לאחר מכן נבדוק קורלציות ולפיהן נחליט אילו משתנים נרצה יותר להכניס למודל בהתאם למודל.
* כמובן שיש שלל דרכים לבחור בין משתנים שונים ולעשות feature selection ויכולנו גם להשתמש בMutual Information על מנת להחליט זאת.
* Date first booking:

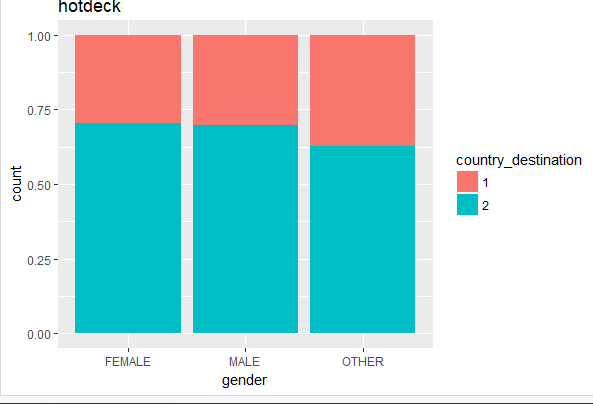


* ניתן לראות שרק ערכים שהם NA (מצד ימין קיצון) לאורך השנים (כלומר לא בוצע בהם כלל booking), החזירו ב-100% 0 שזה הגיוני. לכן יש פה קשר ישיר בין משתנה זה לבין התוצאה 0 בתחזית. לכן כשנעשה תחזית נשתמש רק בנתונים שה-date first booking בהן הוא לא NA ונרצה לחלקן לקלאסיפקציה להאם הבוקינג בוצע בארה"ב או מחוצה לו.
* מכיוון שאנו רואים שקשר זה משפיע, נמשיך כעת לבצע את הבדיקות שלה הקשר על subset עם הנתונים עבורם לא מופיע NA ב- date first bookin.
* ageBins:

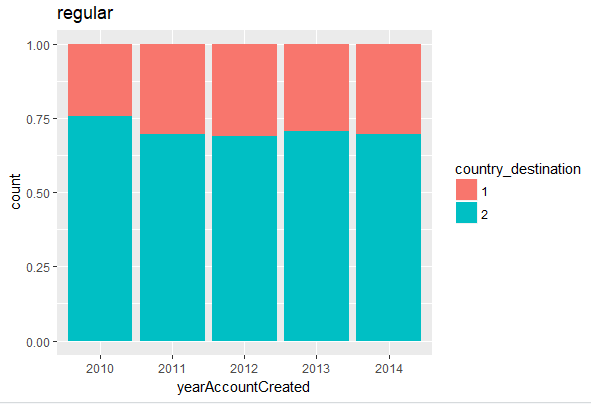
 

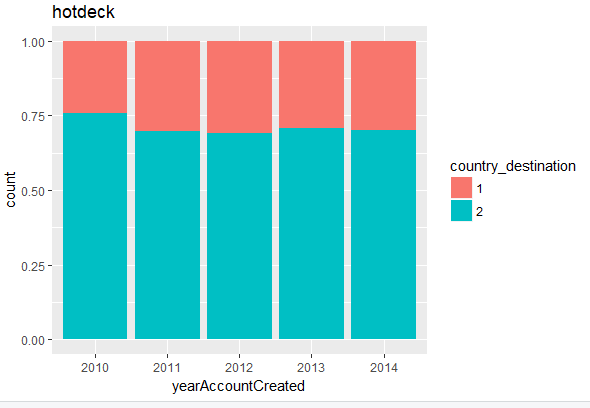
* ניתן לראות שעבור ageBins הן ללא ה-Imputation והן עם ה-Imputatation, ניתן לראות כי הגיל לא משפיע בצורה ניכרת ודי דומה בשניהם.
* gender:



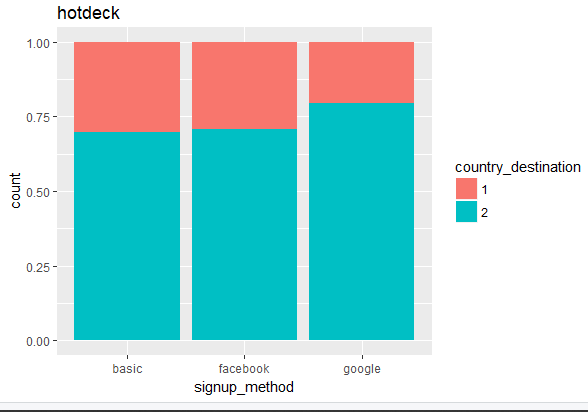
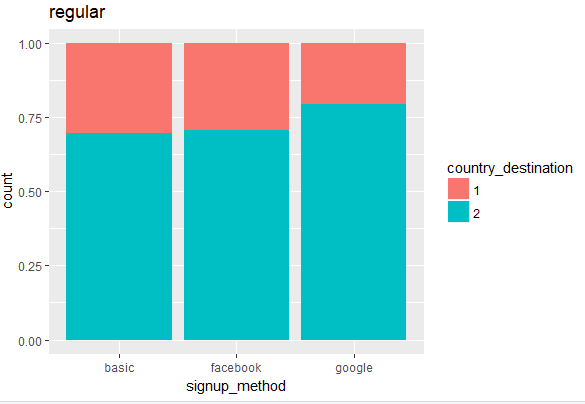


* באופן דומה לעיל ב-gender.
* yearAccountCreated:

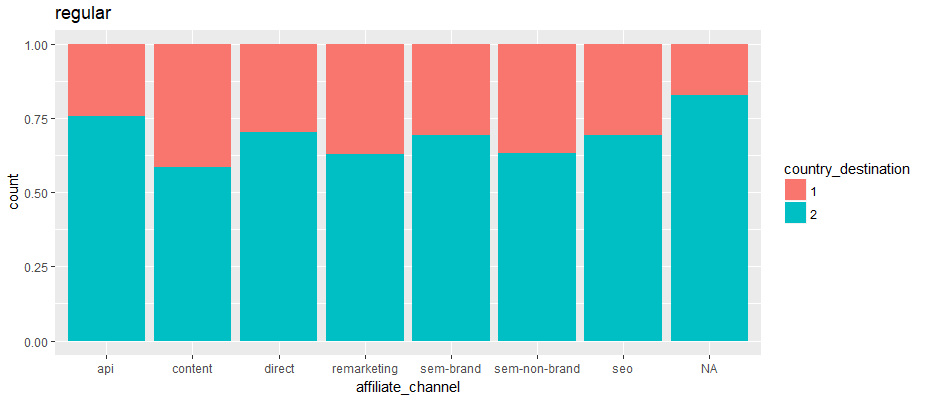
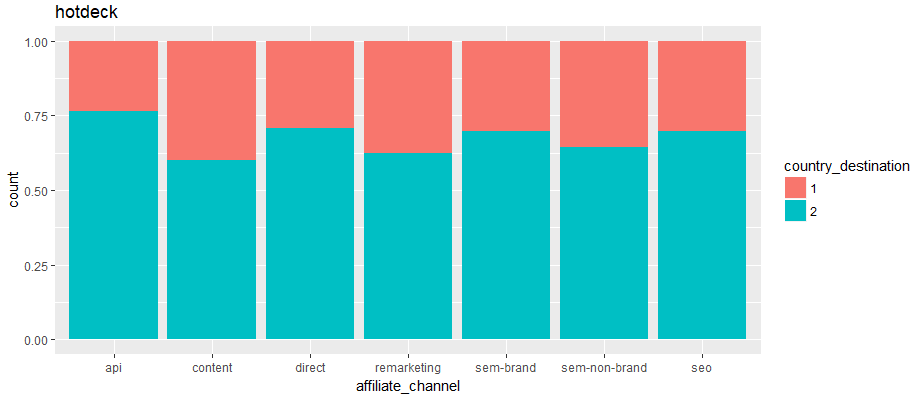




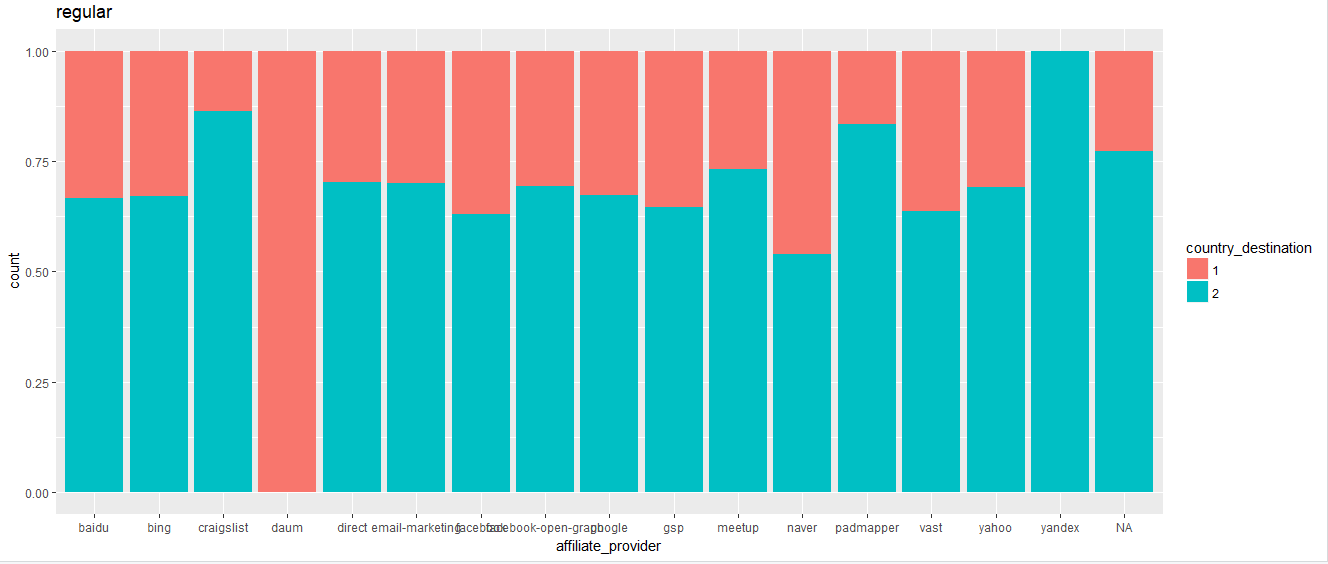
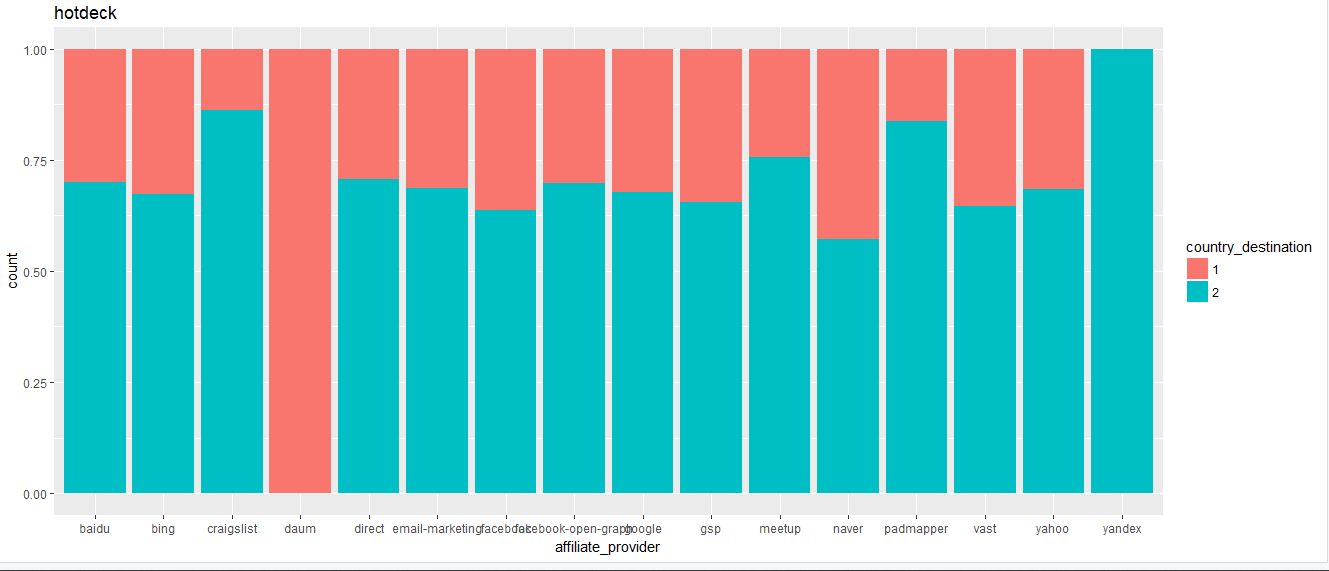
* באופן דומה לעיל ב-gender.
* תוצאות דומות נראו גם ב- monthAccountCreate, weekdayAccountCreate, yearFirstBooking, monthFirstBooking, weekdayFirstBooking, diffDaysAccountFirstActive, diffDaysBookingFirstActive.
* signup\_method:



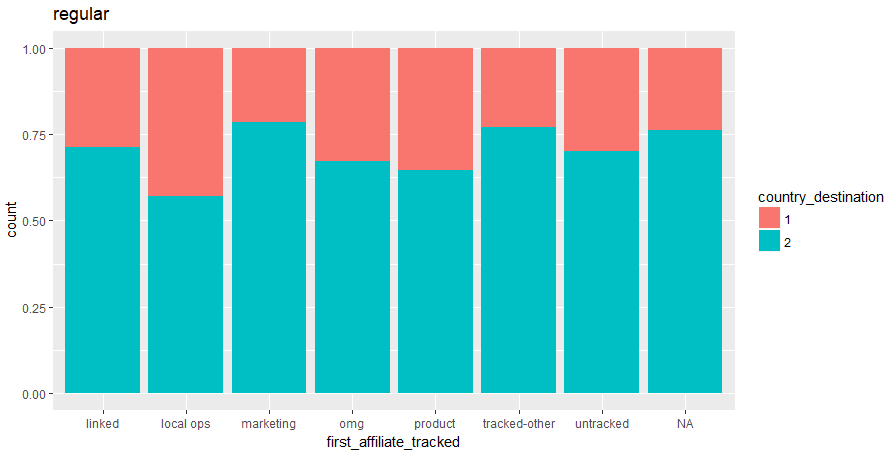
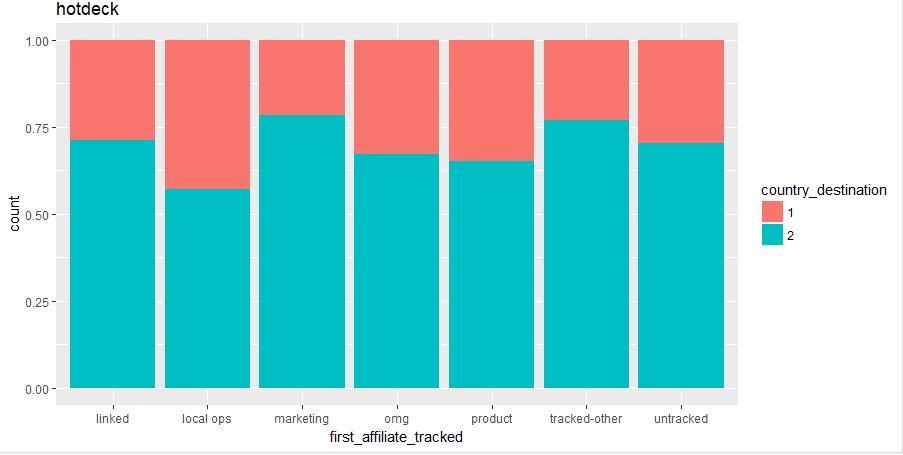
* ניתן לראות שב-signup method ישנה העדפה קלה למי שמשתמש ב-google להזמין יותר בארה"ב ביחס ל-facebook או ב-basic.
* ב-signup\_app זה די דומה גם עם העדפה קלה ב-mobile על פני web להזמין יותר בארה"ב.
* ב-first\_device\_type זה גם פחות או יותר דומה.
* Affiliate channel:

* ניתן לראות שיש העדפה לאנשים שמשתמשים ב-api להזמין יותר בארה"ב.
* Affiliate\_provider:

* ניתן לראות שבאופן די גורף ב-daum ישנה העדפה להזמין מחוץ לארה"ב ולמי שב-yandex ישנה העדפה להזמין בתוך משמעית ארה"ב. בנוסף גם למי שמשתמש ב-padmapper ישנה נטייה להזמין יחסית בתותך ארה"ב. מצד שני, אלו עמודות אשר יש בהן מספר בודד של נתונים לכן נוצרת עבורם הטייה (ב-daum יש נתון אחד וב-yandex 4). לעומת זאת ב-paddmapper יש כבר 200 ושם המצב קצת יותר טוב. באופן כללי ניתן לראות שב-labelHים בהם יש הרבה נתונים, המצב יחסית מאוזן.
* First\_affiliate\_tracked:

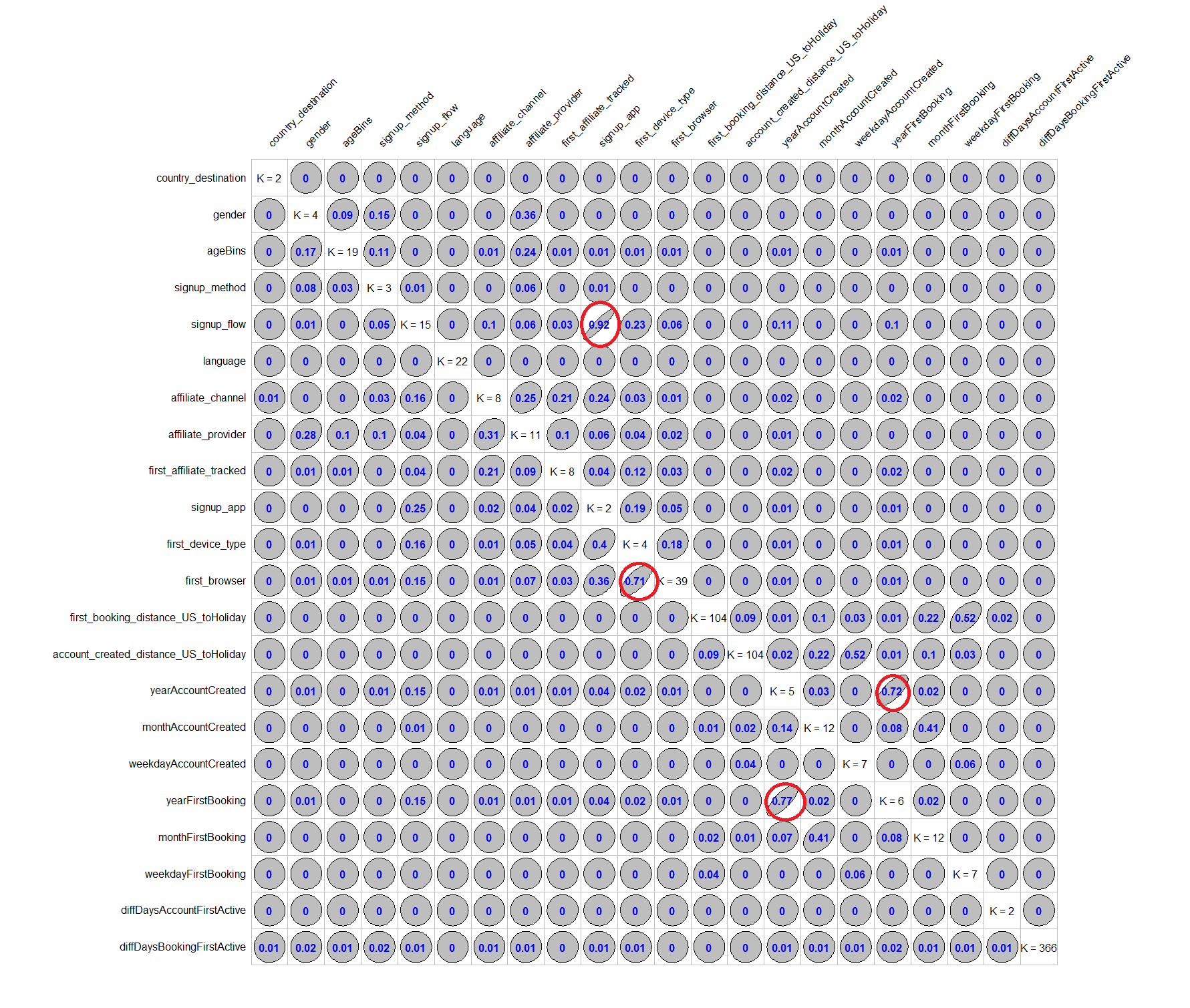
* ניתן לראות כי ב-marekting וב-tracked\_other יש נטייה עבור ארה"ב כאשר ב-marketing יש יחסית פחות נתונים (כ-105 תצפיות). בסה"כ גם כאן המצב מאוזן יחסית.
* First\_browser:
* גם כאן המצב יחסית מאוזן עבור lableים עם נתונים דומים (רמות דומות) ורמות לא מאוזנות עבור lableים עם מעט תצפיות. כנ"ל לגביי language, account\_created\_distance\_US\_toHoliday ו- first\_booking\_distance\_US\_toHoliday.

**קורלציות:**

* על מנת לראות קשר יותר ברור בין המשתנים האחד לשני ובין המשתנים למשתנה התלוי, נרצה לראות קורלציות בין המשתנים השונים.
* מכיוון שקיימים לנו משתנים בעלי ערכים שהם NA, נשנה אותם למשתנה "NA" במקום הערך NA ובכך יהפכו לקטגוריאלים באופן סופי ונוכל לבצע עליהם חישובים.
* שינינו ערכים אלה רק ב-dataset הרגיל כי ב-hotdeck אין כלל ערכי NA בזכות ה-imputation.
  + המשתנים ששונו הם:
    1. Affiliate\_channel
    2. first\_device\_type
    3. first\_affiliate\_tracked
    4. affiliate\_channel
    5. first\_browser
    6. ageBins

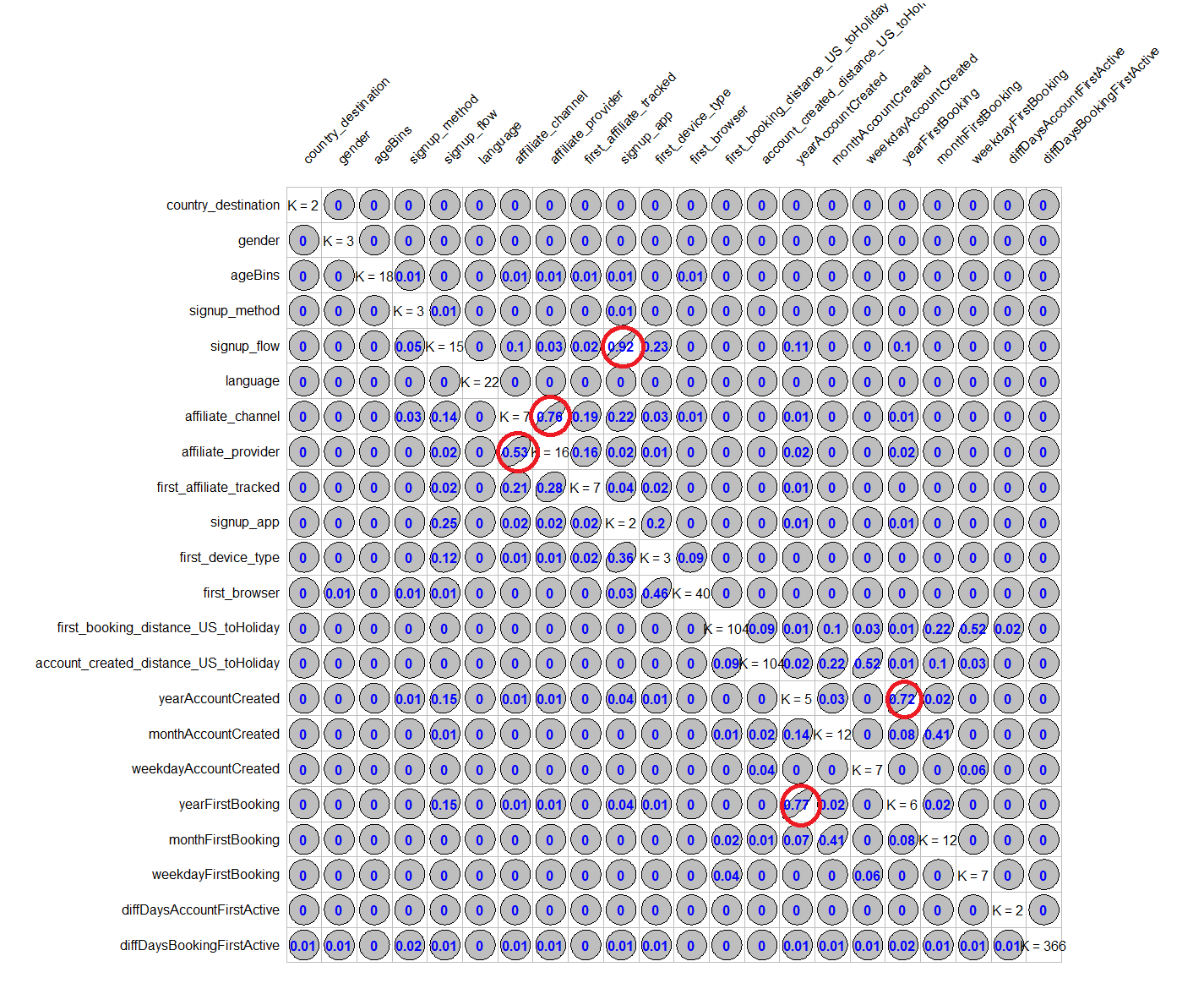
ב-dataset הרגיל:

* מבחינת קורלציה למשתנה החזוי – לא הצלחנו למצוא (שימוש בחבילת kruskal goodman).
* אולם, לעומת זאת הצלחנו למצוא קורלציה בין פיצ'רים יחסית גבוהה:
  + בין signup\_app ל-signup\_flow – 0.92
  + בין first\_browser ל-first\_device\_type – 0.71
  + בין year\_account\_created ל-year\_first\_booking – 0.77
* הקשר השלישי פחות, אולם בשניים הראשונים נתחשב במודל ובמיוחד בקשר ההדוק הראשון.



ב-dataset עם ה-imputation של hotdeck:

* מבחינת קורלציה למשתנה החזוי – גם כאן לא הצלחנו למצוא (שימוש בחבילת kruskal goodman).
* נמצאו קשורים דומים לעיל, אולם הפעם גם נמצא קשר בין affiliate\_provider ל-affiliate\_channel.



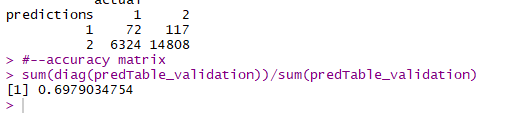
* ניתן כאמור להשתמש גם ב-Mutual Information כפי שציינו, כי שיטת הקורלציה הנ"ל לאו דווקא הכי טובה והכי מתאימה. אולם מפאת חוסר הזמן לא ביצענו זאת.
* 2 מטריצות הקורלציה מצורפות כקבצי png.

**מודלי תחזיות:**

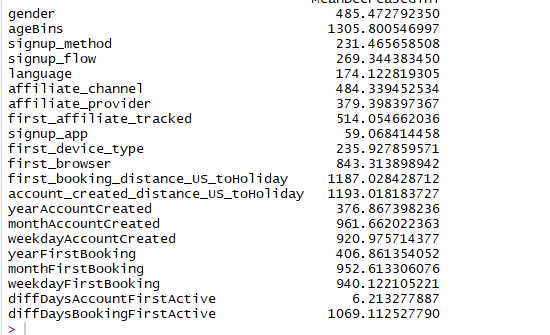
* ראשית, לאור כך שבדקנו וראינו כי עבור כל תצפית שבה לא היה תאריך ב-date first booking אז בתוצאות קיבלנו 0, אזי עבור כל המודלים שנריץ מעתה והלאה, מראש נכתוב במשתנה החזוי 0 עבור תצפיות מסוג זה ואת שאר המודלים נריץ על שאר ה-data.
* מכיוון שיש להכריע בין 2 תוצאות אפשריות, אנחנו נמצאים במצב של מודלי קלאסיפיקציה אשר נרצה להשתמש בהם בשביל לחזות האם ה-user יעשה booking בארה"ב או מחוץ לארה"ב.

**מודל 1: Random Forest**

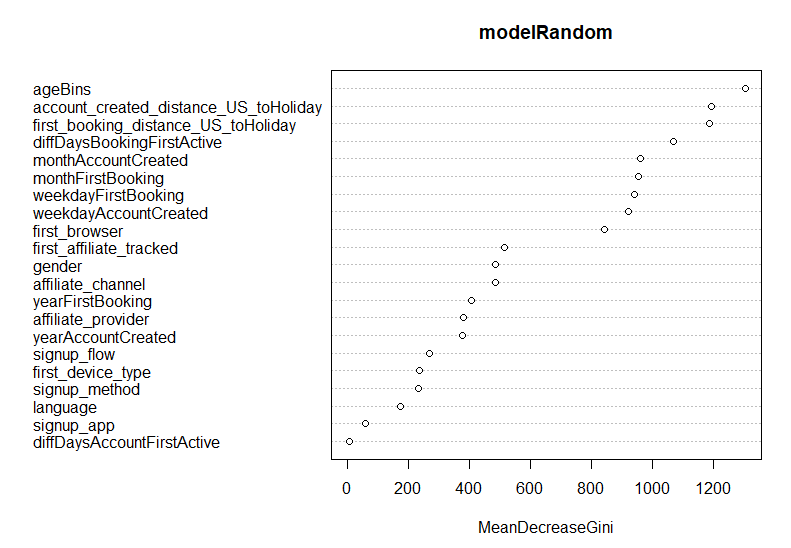
* המודל הראשון שבחרנו להתבסס עליו הינו **Random Fores**t.
* מודל זה מבוסס על עצי החלטה כאשר במהלכו מגדלים עצים ובוחרים משתנים באופן רנדומאלי אשר מוגרלים כאשר לבסוף לוקחים את הממוצע של מספר העצים שמגדלים.
* בכל שלב מחושבת פונקציית Loss אשר האנטרופיה שלה ממוזערת מבין כל המשתנים האפשרים שנבחרים (או לחילופין ממוקסם ה-information gain).
* הפרמטרים העיקריים שצריך להזין הם הפיצ'רים של ה-Dataset שברצוננו להתבסס עליהם, מספר המשתנים הרנדומאליים שנבדקים בכל פעם וכן מספר העצים שנגדל.
* נרצה לא לקחת מעט מדי עצים בשביל שלא נקבל under fitting אולם לא נרצה יותר מדי עצים בשביל over fitting.
* על מנת לחשב את התוצאה הטובה ביותר נרצה להתייחס למדד ה-Accuracy אשר מודד את ה-TP+TN/(TP+TN+FP+FN), כלומר מה מידת הדיוק שלנו בסה"כ עבור 2 התחזיות שלנו – גם בתוך ארה"ב וגם מחוצה לה.
* נריץ את המודל בלולאה על מנת לבחור את הפרמטרים הכי טובים עבורנו.
* כמו-כן נזכיר כי נמצא מודל טוב פעם עבור ה-hotdeck ופעם עבור הרגיל.
* ה-pool של המשתנים שהוזן כולל בתוכו את כלל המשתנים הפקטוריאלים (ageBins במקום age) ומשתנים רציפים כמו ההפרש בין הזמנים שהוצא באמצעות timestamp וכן המשתנים החדשים שנוספו החיצוניים.
* **מהרצת המודל בעל ה-Dataset הרגיל בלולאה כדי לראות מה הפרמטרים הכי טובים התקבל כי:**
  + החלוקה של המשתנים הרנדומאליים היא 2
  + מספר העצים שנבחר הוא 45



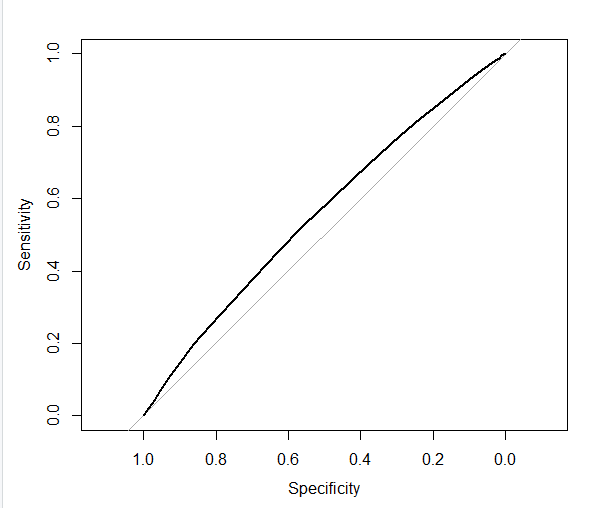
* + התקבל accuracy של 0.697.
  + להלן טבלת החשיבות:



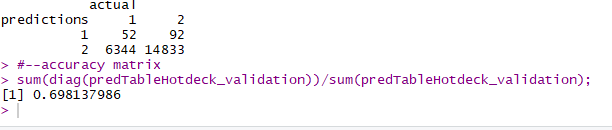
* + להלן המחשה לפי סדר חשיבות:



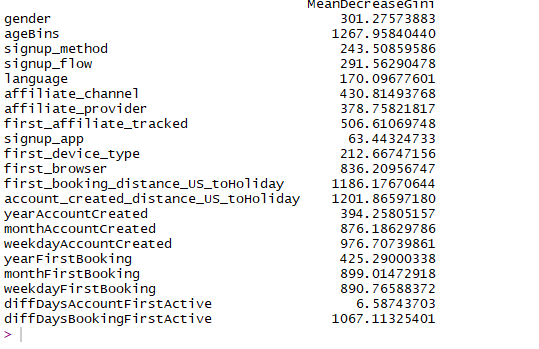
* + נשים לב כי הגורמים המשפיע ביותר הם **המרחקים מהחגים** של תאריך ה-Booking וה-ageBins.
* מבחינת AUC, להלן התוצאות:



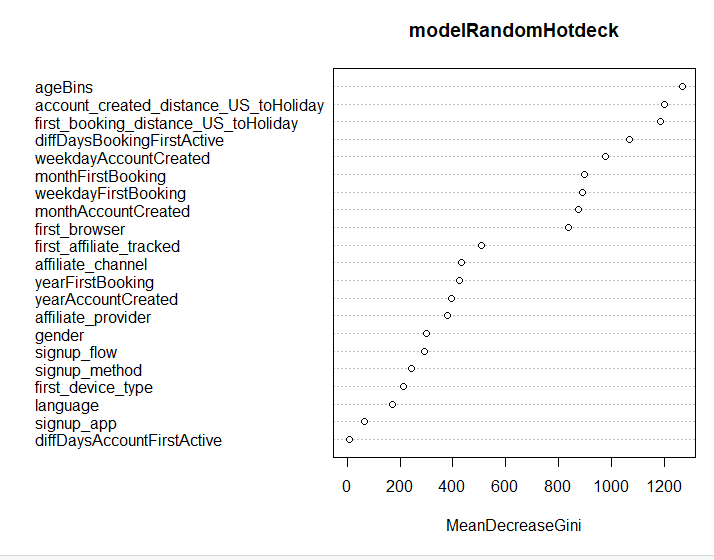
* מודל זה נבחר ביחס לשאר המודלים אחרי הרצות רבות והשוואות מדד ה-accuracy ומדד ה-AUC בין העצים השונים.
* **מהרצת המודל בעל ה-Dataset ה-hotdeck בלולאה כדי לראות מה הפרמטרים הכי טובים התקבל כי:**
  + מספר המשתנים הרנדומאליים הוא 2
  + מספר העצים הוא 45



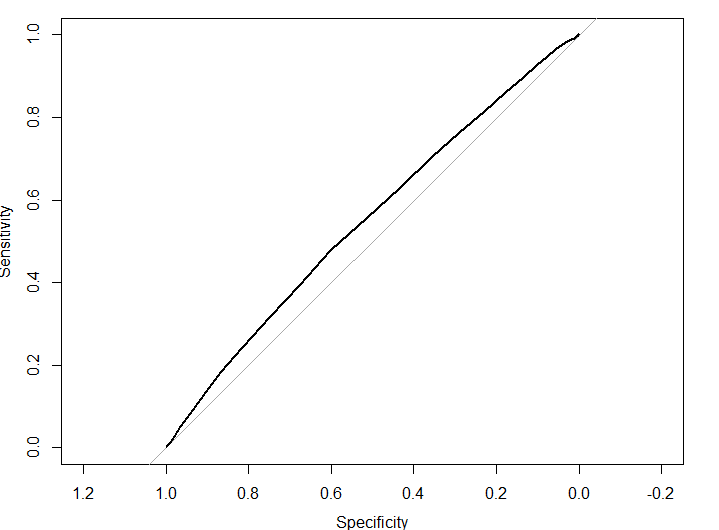
* + התקבל accuracy של 0.698.
  + להלן טבלת החשיבות:



* + להלן המחשה לפי סדר חשיבות:



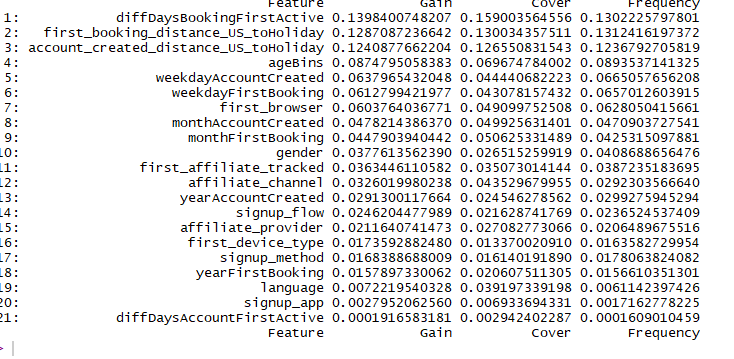
* + נשים לב כי הגורמים המשפיע ביותר הם **המרחקים מהחגים** של תאריך ה-Booking וכן ה-ageBins, בדומה למקודם.
* מבחינת AUC, להלן התוצאות:

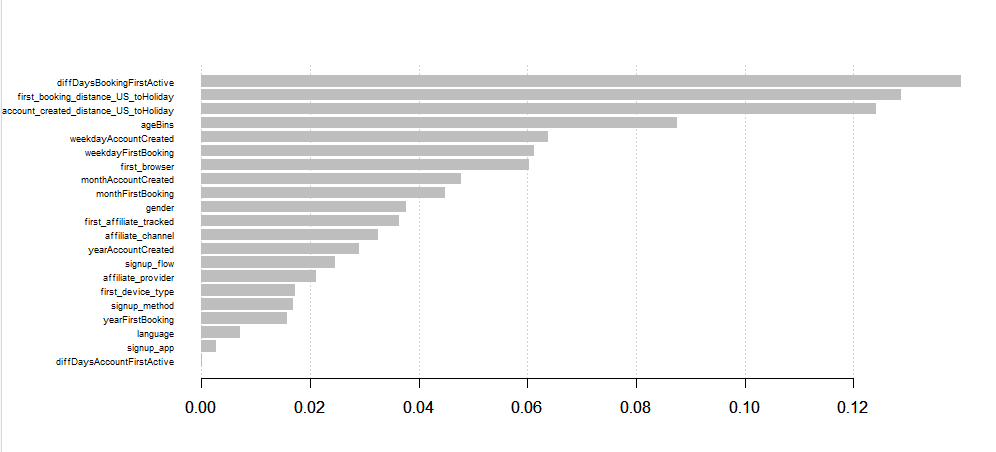


* נשים לב בסה"כ שמכיוון שזהו מודל רנדומאלי התוצאות יוצאות שונות בכל פעם, אך לב הן יוצאות בסביבות accuracy של 0.69.
* יש לציין כי נסיונות להוריד משתנים שביניהם הייתה קורלציה גבוהה הורידו יחסית הרבה את התוצאות – לכן השתמשנו בכל המשתנים האפשריים במודל.

**מודל 2: XGBoost**

* המודל השני שבחרנו להתבסס עליו הינו XGBoost - eXtreme Gradient Boosting.
* זהו מודל נוסף המבוסס עצי החלטה שעוזר לקלאסיפיקציה ורגרסיה באמצעות שימוש בגרדיאנטים.
* לקריאה נוספת אודותיו: https://xgboost.readthedocs.io/en/latest/model.html
* במודל שלנו אנו נעשה שימוש באלגוריתם זה עפ"י שיטת רגרסיה לוגיסטית על מנת שיתמוך בקלאסיפיקציה בינארית.
* הפרמטרים אותם מזינים הם:
  + Data – המידע שאנחנו רוצים לפעול עליו
  + Label – המשתנה אותו נרצה לחזות
  + Max.depth – עומק העץ
  + Nthread – מספר ה-threadים במחשב שנשתמש בהם.
  + Nround – מספר הפעמים שנעבור על ה-data.
* ראשית, לצורך שימוש בחבילה נרצה להפוך את ה-train שלנו למטריצה.
* את הבדיקות עשינו על הרבה פרמטרים שונים ולבסוף הגענו למספר של 55 במספר החזרות ועומק עץ של 10.
* גם במקרה זה השתמשנו במדד ה-Accuracy, במדד err של שגיאה וב-AUC.
* השגיאה שהתקבלה הינה: 0.3739036631.
* להלן מטריצת החשיבות:

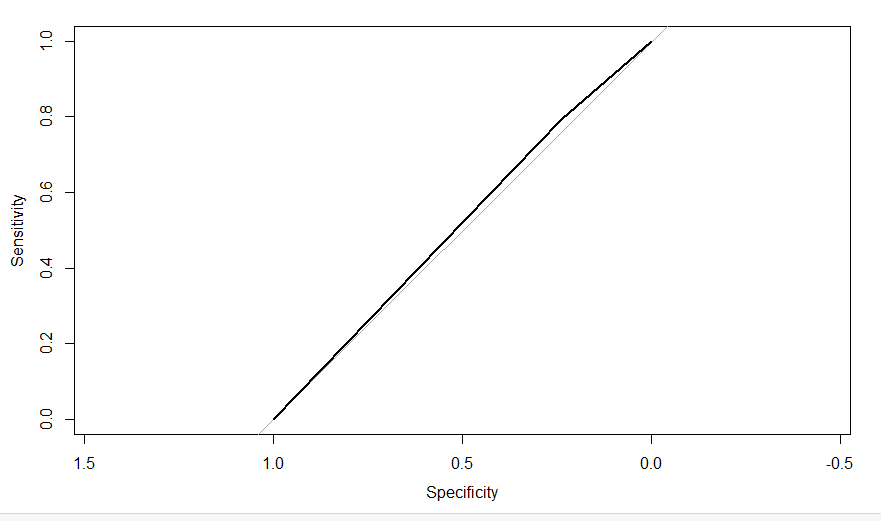




* ניתן לראות שהמדדים המובילים, בדומה לrandom forest הם המרחקים מהחגים ו-ageBins, אולם כאן גם הפרש מספר הימים מהבוקינג הראשון גם היה משמעותי. כמו-כן ניתן לשים לב ששוב **בדומה** ל-random forest, הפיצ'רים של של מספר הימים מאז הפעם הראשונה שה-user היה פעיל, ה-signup\_app וה-language הם הכי פחות משמעותיים.
* מבחינת ה-accuracy, הושג כזה של 0.63. פחות טוב משל ה-random forest:

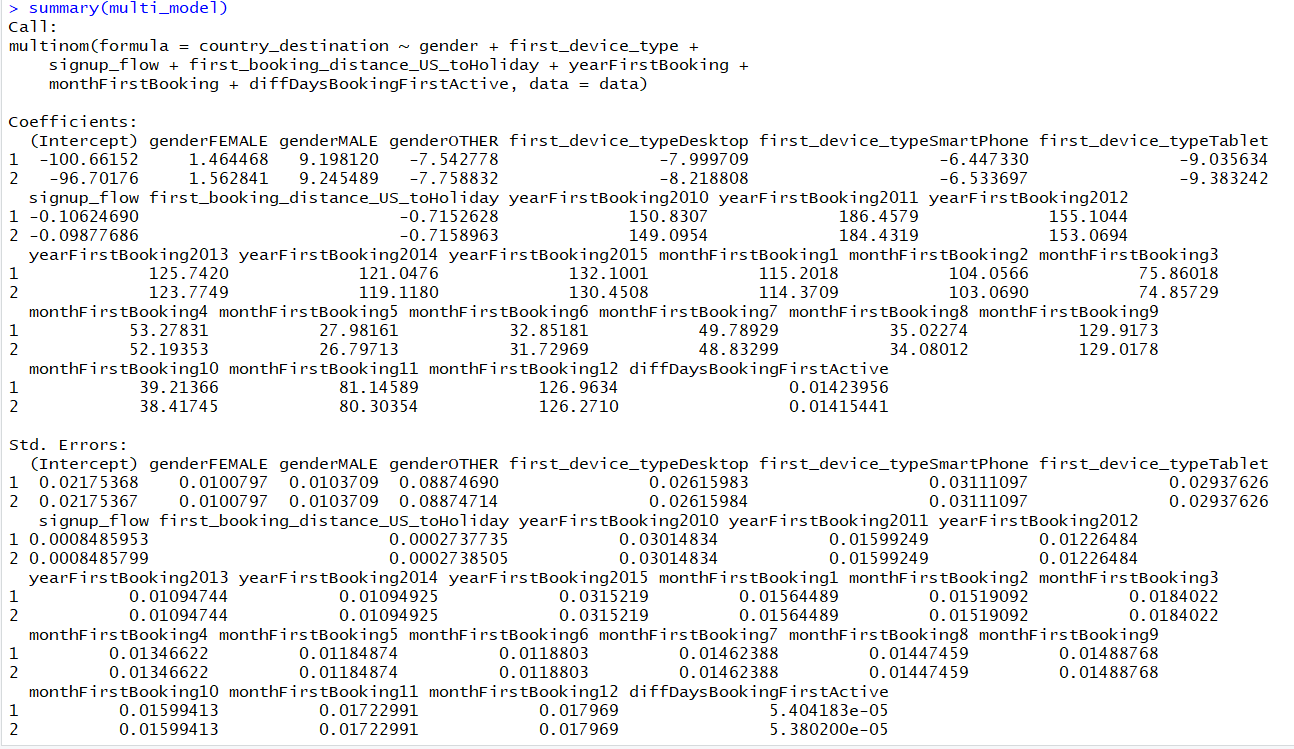


* כמו-כן, גם ה-AUC שהתקבל פחות טוב וקיבלנו בסביבות ה-0.5:

- תוצאות דומות התקבלו גם עבור ה-**hotdeck**.

**מודל 3: רגרסיה לוגיסטית ומולטינומית**

* הרציונאל לבצע אותן הוא עקב קלאסיפיקציה כאשר במולטינומית השתמשנו על כל ה-data (בלי להוריד את האפסים הברורים) או רגרסיה לוגיסטית לאחר הורדת האפסים.
* בגלל שהמשתנה התלוי שלנו הוא קטגוריאלי, רגרסיה לינארית לא אפשרית. לכן בחרנו ברגרסיה לוגיסטית אשר מתאימה למשתנה תלוי בסולם קטגוריאלי.
* במקרה שלנו למשנה התלוי הקטגוריאלי קיימות 3 רמות, ולכן נשתמש בספריית nnet כדי ליצור רגרסיה לוגיסטית multinom.
* בהתאם לקורלציות שראינו מעלה, בחרנו להשתמש במשתנים מסבירים בעלי קשר למשתנה התלוי:
* country\_destination=gender+first\_device\_type+signup\_flow+first\_booking\_distance\_US\_toHoliday+yearFirstBooking+monthFirstBooking+diffDaysBookingFirstActive



* עקב תוצאות פחות טובות בבירור אל מול שאר המודלים הוחלט שלא להשתמש במודלים אלו ועל כן לא נרחיב בנושא.

**מודל סופי:**

* ה-**random forest** הניב את התוצאות הטובות – קרוב ל-0.7 אחוז ב-validation (בלי לכלול את התוצאות שהן booking=0) אשר תוצאות ודאיות ובאופן יחסית יעלו מאוד את האחוז דיוק של כל ה-data.
* מכיוון שהנתונים של booking היוו כ-**58% מה-data** – אזי בסה"כ הגענו בתוצאה הסופית **לאחוז דיוק באמצעות Random Forest של 87.35.**
* יש לשים לב כי אחוז זה עלול להשתנות בהתאם לתוצאות ה-random forest, אך הדבר נע באזור האחוזים הללו.
* מצורפים **2 קבצים עם תוצאות** (עבור כל dataset):
  + airbnb\_test\_final.csv
  + airbnb\_test\_hotdeck\_final.csv