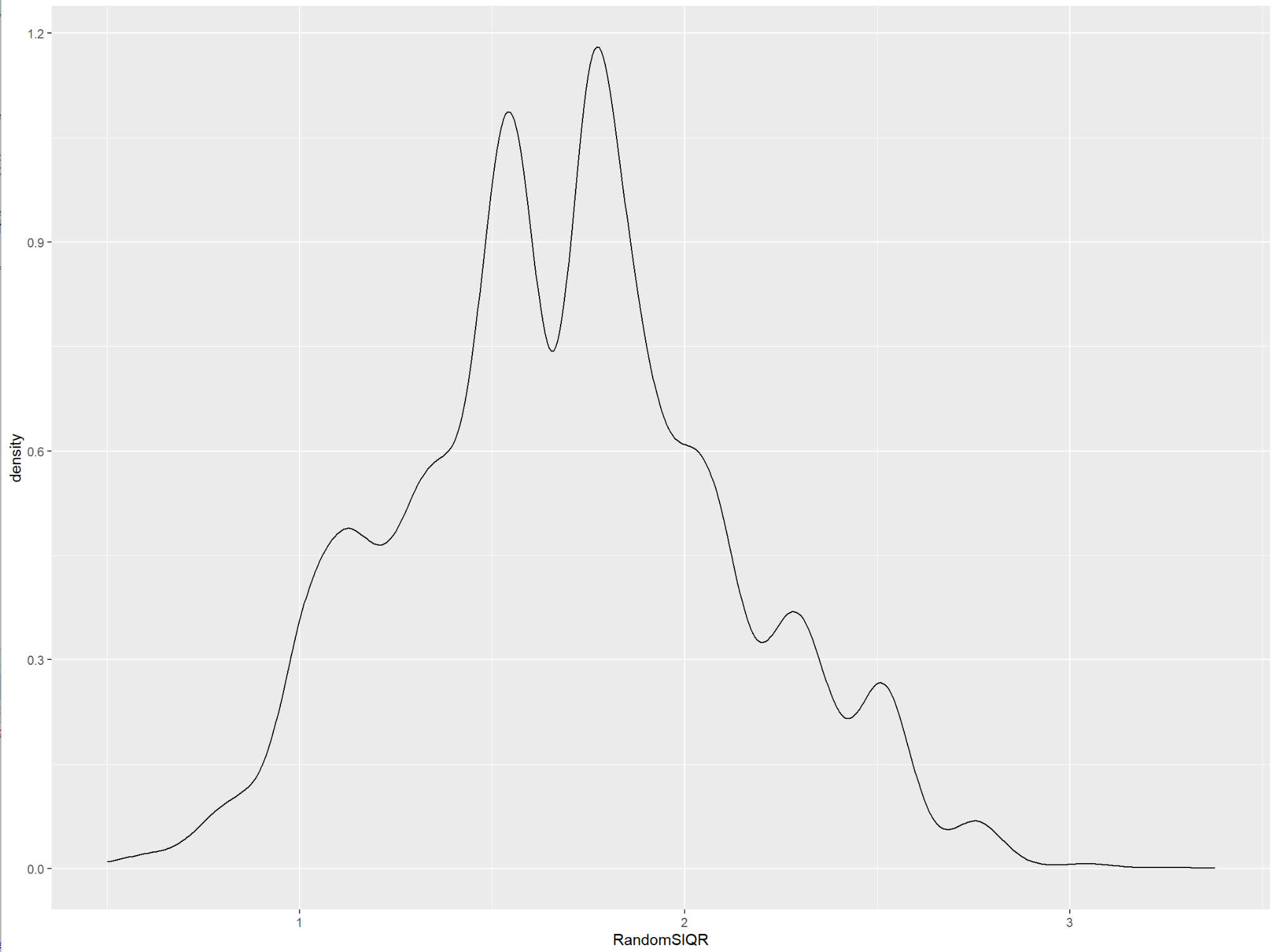
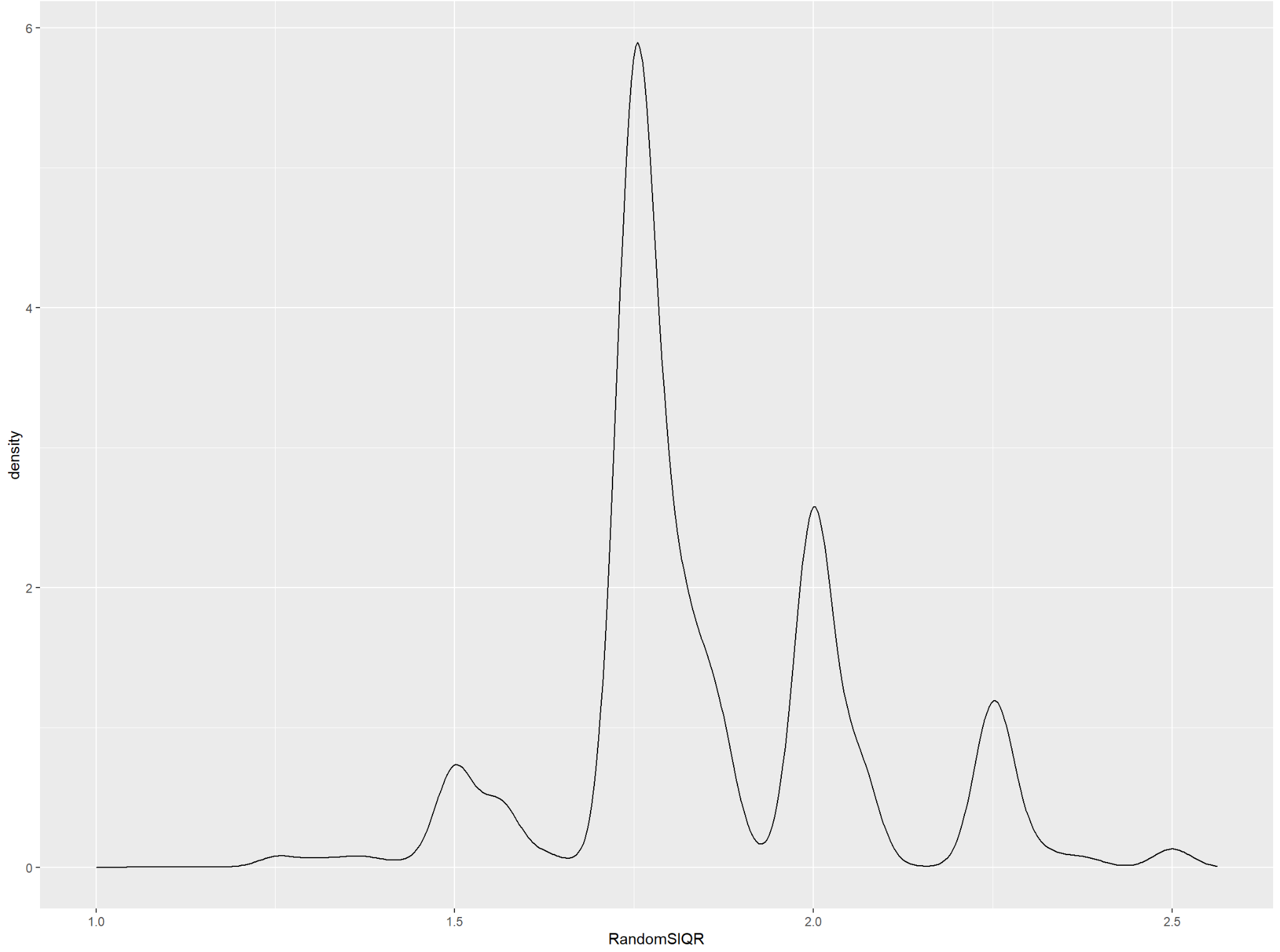
**עבודה ב-R מבוא לסטטיסטיקה**

קובץ הנתונים בו בחרנו להשתמש עוסק במחיר הבירה (המשקה) באצטדיונים ערים שונות (בהן משחקות קבוצות שונות), בגודל הבירה באצטדיונים הללו (כמה אונצ' יש בכוס), ובכך גם בקשר בין הגודל של הבירה למחיר שלה. כל זה מקושר לשנה, לעיר ולקבוצה המשחקת.

סעיף א'

בשאלה זו בחרנו להשתמש בשני משתנים מספריים מקובץ הנתונים שלנו: מחיר בירה לאונצ' (PPO) וגודל המשקה. ביצענו בניהם מכפלה וקטורית שנתנה לנו בעצם וקטור חדש- מחיר המשקה (יש לציין שיכולנו לוודא כי המכפלה הוקטורית יצאה לנו נכון היות שיש עמודה של מחיר המשקה בקובץ הנתונים שלנו. במידה והיא תואמת את המידע שבוקטור נדע כי הפעולה שעשינו על מנת לעשות  
מכפלה וקטורית- נכונה). עם הוקטור החדש ביצענו את החישובים בשאלה. לשם ביצוע החישובים יצרנו משתנה וקטורי נוסף: RandomSIQR (Random Selection Interquadral Ratio).   
היה עלינו ליצור לולאה שתרוץ 10000 פעמים, שבכל פעם תדגום מקרית 100 תצפיות מתוך המשתנה החדש (בדגימה עם החזרה), תחשב את הטווח הבין רבעוני שלהן ואת הערך הזה תכניס כערך למשתנה חדש. את החישוב הנדרש ביצענו עם לולאה אחת והכנסנו ישירות את הטווח הבין רבעוני לתוך הוקטור הייעודי באמצעות פונקציית IQR (עליה נרחיב בסעיף העבודה האחרון המתעסק בפונקציות חדשות ודרך מציאתן). לאחר מכן, הצגנו את ההתפלגות באמצעות פקודת GGPLOT.

**מימין-** דגימה של  
 20 תצפיות **משמאל -** דגימה של 100 תצפיות

כפי שניתן לראות, ההתפלגות שיצאה הינה רב-שכיחית וא-סימטרית שלילית (ממוצע גדול מהחציון- ממוצע 1.85 וחציון 1.79). אילו היינו דוגמים מקרית רק 20 תצפיות בכל פעם, ולא 100, ככל הנראה ההתפלגות הייתה משתנה מבחינת השונות והפיזור שלה. הסיבה לכך היא שכל ערך בהתפלגות מייצג פחות תצפיות, מה שישפיע על השונות בין התצפיות וכך גם על הטווח הבין רבעוני. מה שעשוי להשפיע על השונות ועל הטווח הבין רבעוני של כל 20 תצפיות לעומת 100, הוא שככל שיש יותר תצפיות, כך ההתפלגות קרובה יותר להתפלגות התיאורטית (ההתפלגות באינסוף), וככל שיש פחות, כך אנו עשויים להתרחק מהטווח הבין רבעוני של ההתפלגות המקורית (עלול פחות לייצג את המגוון הקיים). במקרה שכזה, נראה סטייה משמעותית יותר בהתפלגות של הטווחים הבין רבעוניים (שפחות משקפים את הקיים). דבר נוסף המשפיע היא העובדה כי מדובר בדגימה עם החזרה, אשר עשויה להביא אף לסטייה משמעותית יותר (ישנו סיכוי נמוך משמעותית לדגימה כפולה של אותו הערך כאשר דוגמים פחות פעמים מאותו סך של תצפיות).   
בעצם, השונות של 20 תצפיות עשויה להיות או קטנה יותר או גדולה יותר, מה שישפיע על הדיוק, ויזיז את הטווח הבין רבעוני בהתאם לערכים הנבחרים. מסתמן כי נקבל התפלגות דומה, אך עם שונות גבוהה היות שתהיה סטייה גדולה יותר בין הטווחים הבין רבעוניים של כל 20 תצפיות.

**סעיף ב'**

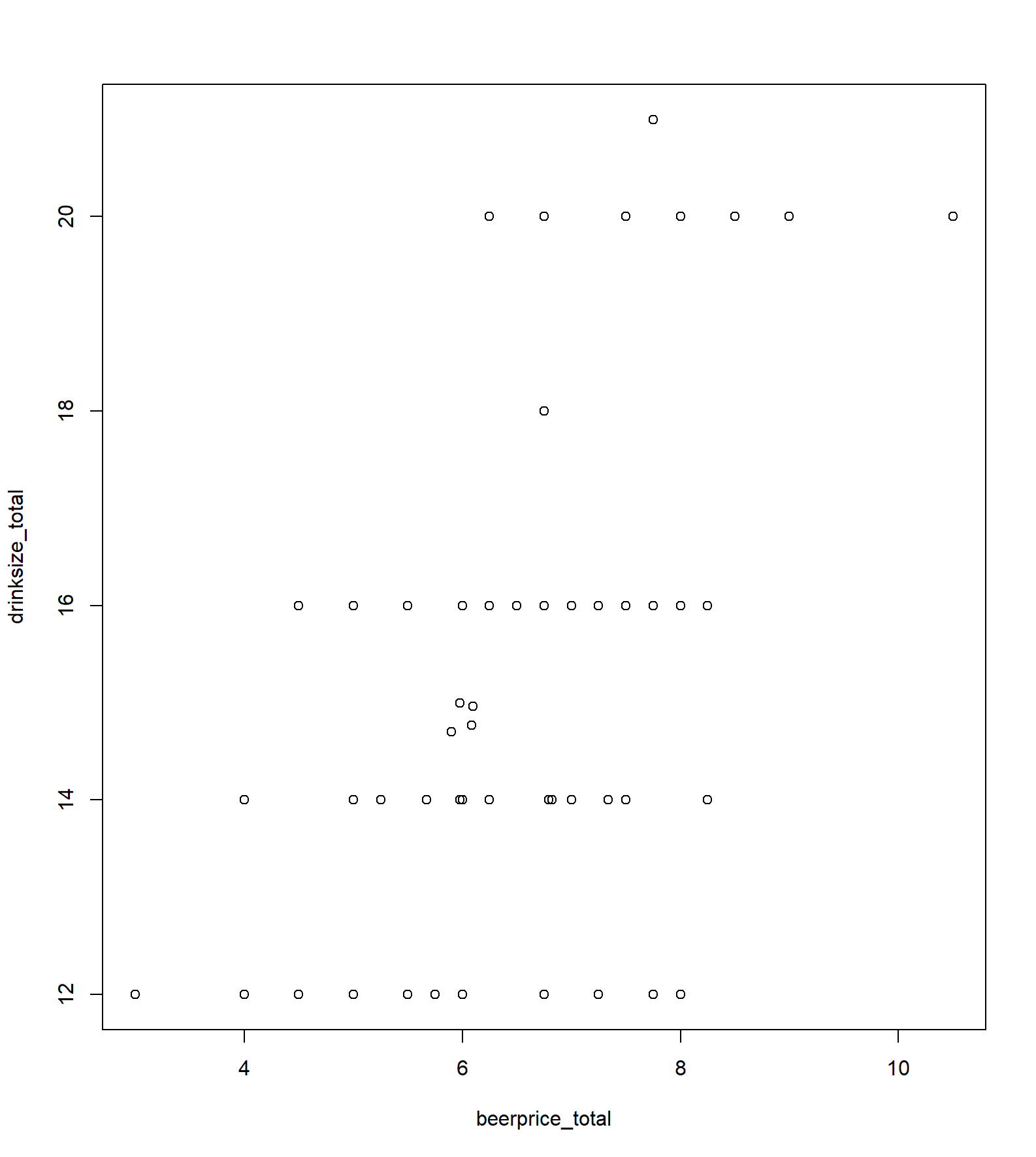
המשתנה בו בחרנו להשתמש בסעיף זה הוא "מחיר לפי כמות" (ביחדות ounce).   
בקובץ הנתונים ובקוד המשתנה מופיע כ-PPO.   
חישבנו את מדד הפיזור AD של משתנה זה על ידי שימוש בלולאה שרצה 150 פעמים (קבענו זאת על פי מספר הערכים שהמשתנה בו בחרנו יכול לקבל). בכל פעם, הלולאה מחשבת את ערכו המוחלט של ההפרש בין הערך במיקום מסוים לבין החציון, וסוכמת אותו יחד עם כל הערכים שהתקבלו עד למיקום זה (לפי הסדר- מהערך הממוקם במיקום הראשון עד למיקום ה-150).   
הלולאה מפיקה את סכום הסטיות המוחלטות מהחציון. היות שעלינו לחשב את ממוצע הסטיות המוחלטות מהחציון, חילקנו את הסכום ב-150 (מספר הערכים אותם סכמנו).   
לאחר מכן, יצרנו לולאה הדוגמת 100 פעמים 50 ערכים מתוך המשתנה שבחרנו (באופן מקרי).   
על ידי שימוש בלולאה פנימית המחשבת כל פעם את סכום הסטיות המוחלטות מהחציון של 50 ערכים שונים שנדגמו, הלולאה "החיצונית" מפיקה וקטור בעל 100 ערכים כאשר כל אחד מהם הוא ה-AD של כל מדגם כזה (פשוט מחלקים ב-50 את כל איבר בוקטור כדי לקבל ממוצע).   
לולאה זו יוצרת וקטור נוסף- ad\_ratio, שבכל אחד מערכיו נמצא היחס בין ה-AD שחישבנו בהתחלה (המקורי), לבין ערך במיקום מסוים בוקטור הקודם (ה-AD של מדגם של 50 ערכים).   
חישבנו את השונות של הוקטור ad\_ratio.   
את כל זה הרצנו מספר פעמים, ובכל פעם שהרצנו השונות הייתה קטנה מ-0.015,   
משמע- נמוכה מאוד. שונות נמוכה מעידה על כך כי הפיזור של הערכים בוקטור ad\_ratio צר, קרי- היחס בין כל אחד מה-AD שחושבו לכל מדגם של 50 ערכים, ובין ה-AD המקורי שחושב עבור מדגם מלא (150 ערכים)- די דומה.   
במקרה כזה, ניתן להסיק כי גם השונות של הוקטור בו מצויים הערכים של המשתנה שבחרנו (PPO) נמוכה, היות שבכל אחת ממאה הפעמים בהן דגמנו 50 ערכים שונים של המשתנה באופן מקרי, עדיין קיבלנו יחס דומה. השונות שחישבנו בעצם מעידה על כך שהשונות המקורית נמוכה ככל הנראה, וכך גם הפיזור צר (רוב הערכים קרובים/ דומים).

**סעיף ג'**

בקובץ הנתונים שלנו ישנם שלושה משתנים מספריים (מחיר בירה, כמות הבירה, ומחיר לפי כמות (ביחידות ounce), בשאלה זו נשתמש במשתנים: מחיר, וכמות בירה.

השלב הראשון- הכנת מרחב העבודה - שלב כללי לרוב הסעיפים והוא כולל את הכנסת המידע לתוך משתנה "DATA", וחלוקתו לפי שנים לטובת תפעול נוח יותר לפי הצורך\עניין.

משתנים נוספים שעלינו ליצור הינם זוג וקטורים המכילים את מחירי הבירות ואת כמות הבירות (DRINKSIZE\_TOTAL ו BEERPRICE\_TOTAL בהתאמה) בנוסף לוקטור "COR\_VECTOR2" אשר בו נשתמש בהמשך השאלה.

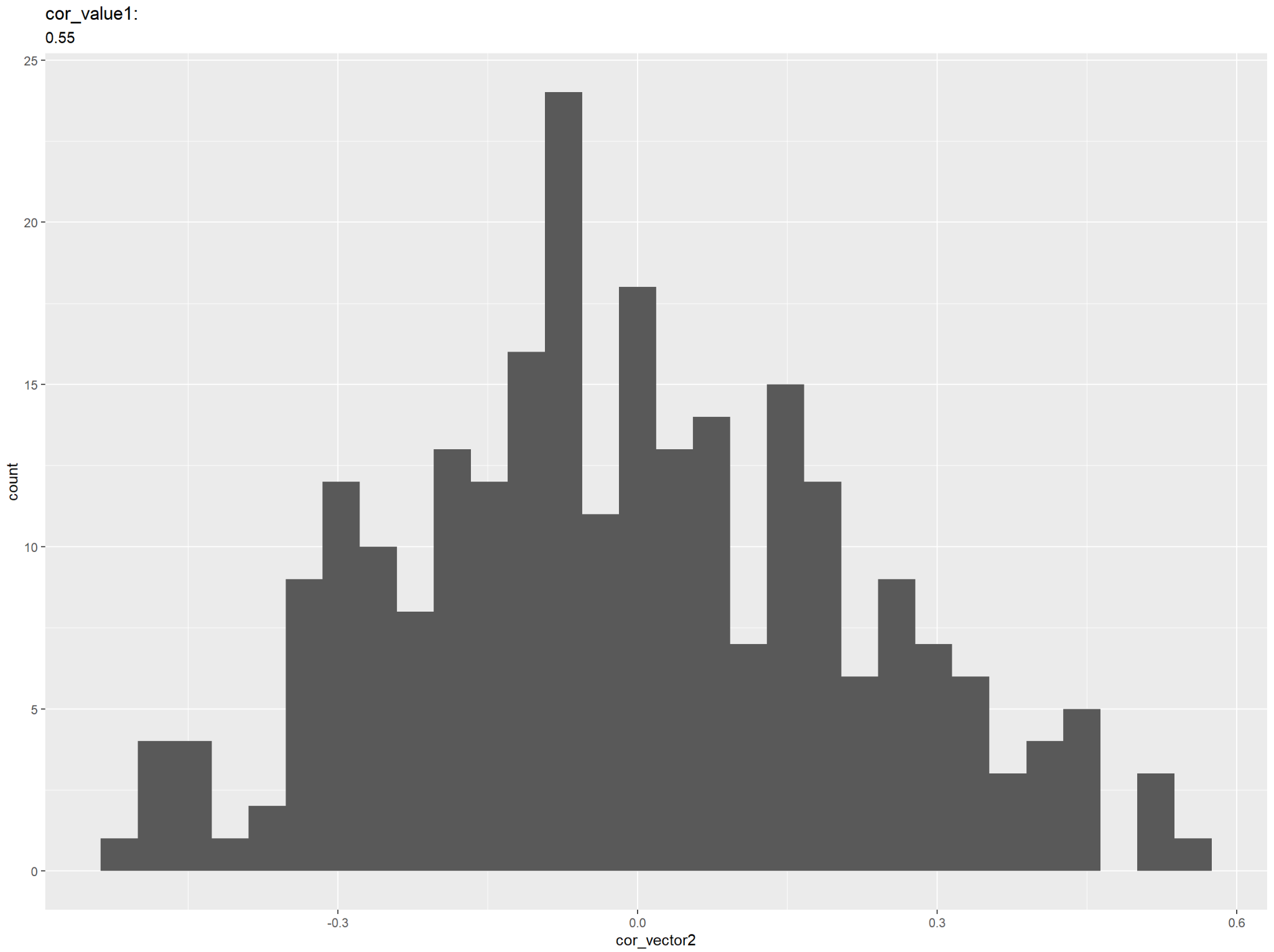
בשלב הראשון של השאלה עלינו לחשב את המתאם בין שני המשתנים - לטובת חישוב זה נשתמש בפקודת "COR" ונכניס למשתנה חדש "COR\_VALUE1". בשימוש במתאם פירסון נקבל מתאם של 0.55 (0.51 ספירמן).

ניתן לראות בדיאגרמת הפיזור, כי במקומות רבים (השונים זה מזה כמובן) יש מחירי בירה זהים - לכן, למרות שזה לא אינטואיטיבי במיוחד, ניתן להסיק כי הקשר הינו אקראי דבר שנראה טוב בחלק הבא של השאלה. (היינו מצפים כי יותר בירה תעלה יותר, בקשר מונוטוני בינוני - חזק).

בחלק זה עלינו להכין לולאה שתפקידה לחשב מתאם על 70% מהערכים אשר יבחרו באקראי מתוך המדגם הכולל ולחשב איתם מדגם 250 פעמים- לשם כך נשתמש בוקטור "COR\_VECTOR2" שיצרנו בתחילת העבודה.

כעת, הוקטור מכיל 250 מתאמים של דגימות אקראיות - והממוצע שלו שווה לאפס, דבר המחזק את הטענה כי אכן המשתנים אינם תלויים אחד בשני.

דיאגרמת פיזור של "COR\_VECTOR2":

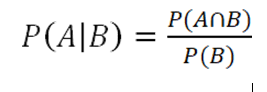
לטובת וידוא התוצאות, ביצענו את אותו התהליך גם עם נתונים אשר נלקחו אך ורק משנה (נבחר שרירותית - 2016) וקיבלנו תוצאות זהות.

**סעיף ד'**

בסעיף זה של העבודה, יצרנו משתנה מקרי בעל 1000 ערכים שנלקחו מתוך התפלגות נורמלית,   
על ידי שימוש בפקודה שדוגמת באופן אקראי 1000 ערכים מתוך התפלגות נורמלית.   
יצרנו לולאת while (עליה נרחיב בסעיף האחרון של העבודה הנוגע לפונקציות שלא נלמדו) הדוגמת בכל פעם 20% מהערכים מתוך המשתנה בעל 1000 הערכים (משמע- 200 ערכים שונים כל פעם, שבתוך כל מדגם הערכים גם שונים זה מזה).   
הלולאה שיצרנו רצה כל עוד הפער בין הממוצע, לחציון של הערכים שנדגמו בהרצה קטן מהערך 0.01. המשמעות היא שהלולאה עוצרת רק כאשר הפער הנ"ל גדול מהערך 0.01 (כפי שנתבקשנו לעשות בעבודה). חשוב לציין כי הוספנו ערך מוחלט היות שמדובר על הפער ולא על ההפרש.   
על מנת לדעת כמה פעמים הלולאה רצה עד שנעצרה, הוספנו מן "מונה".   
יצרנו משתנה "count" שלפני הרצת הלולאה ערכו היה 0.   
בכל פעם שהלולאה רצה, חוץ מלבצע את פעולת הדגימה והחישוב של הפער, היא גם הוסיפה 1 למשתנה "count", וכך- קיבלנו את מספר הפעמים שהלולאה רצה.   
את הלולאה הרצנו מספר פעמים על מנת לקבל רקע טוב יותר למסקנה שלנו. בדקנו על ידי פקודת הדפסה של המשתנה "count" כמה פעמים היא רצה, ובכל פעם יצא כי היא לא רצה הרבה- לעיתים נעצרה לאחר פעם אחת בלבד, כאשר המקסימום בו פגשנו הוא 3.  
  
מנתון זה, ניתן להסיק כי הפער בין הממוצע ,לחציון של כל מדגם כזה (של 20% מתוך הדגימה של 1000 ערכים שנלקחו מתוך התפלגות נורמלית) יהיה לרוב גדול מ0.01, מה שמצביע על כך שכל אחת מההתפלגויות שנוצרו בדגימה של עשרים אחוזים מן הערכים יהיו ברוב המקרים התפלגויות   
א-סימטרית שליליות/ חיוביות (תלוי מי יותר גדול- הממוצע או החציון).

**סעיף ה'**בסעיף זה נתבקשנו לבחור שני משתנים קטגוריאליים (=בסולם שמי) מהקובץ ולחשב את ההסתברות המותנית עבור כל אחת מהקטגוריות של המשתנה הראשון בכל אחת מהקטגוריות של המשתנה השני. קובץ הנתונים עליו אנו מתבססים בעבודה אכן מכיל משתנים קטגוריאליים, אך כאלה המכילים המון קטגוריות. דוגמא לכך היא שישנן 30 קטגוריות באחד מן המשתנים, משמע- במידה שנרצה לחשב את ההסתברות המותנית עבור כל אחת מהקטגוריות של המשתנה הראשון בכל אחת מהקטגוריות של המשתנה השני, יהיה עלינו לחשב ולכתוב 900 הסתברויות מותנות (בכתב/ בקוד).

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי  
היות שרוב הסתברויות החיתוך שוות ל-0 (כי הן 0/30), כך גם רוב ההסתברויות המותנות יהיו שוות ל-0.  
ההסתברויות המותנות שלא יהיו שוות ל-0 הן אלו שידוע כי יש סיכוי לחיתוך ביניהן (ניתן לראות את זה די בבירור בקובץ הנתונים היות שכפי שכבר ציינו קודם לכן- המשתנה מחולק להרבה מאוד קטגוריות).  
כל הסתברויות החיתוך שאינן שוות ל-0, שוות ל-1/30 (ניתן גם לראות זאת בטבלה).  
ההסתברות לכל אחת מהקטגוריות של המשתנה "כינוי" היא 1/30.  
ההסתברות לכל אחת מהקטגוריות של המשתנה "עיר" היא 1/30, חוץ מניו יורק ושיקגו שלהן יש הסתברות של 2/30=1/15, היות שבשונה מיתר הערים (בהן יש קבוצה אחת המצוינת על ידי קובץ הנתונים), בערים אלו ישנן שתי קבוצות שמצוינות.  
על כן, רוב ההסתברויות המותנות שאינן שוות ל-0 יהיו שוות ל: (1/30)/(1/30)=1.  
ההסתברויות המותנות שקובעות את ניו יורק ו/או שיקגו כ"מאורע שבהכרח מתקיים" יהיו שוות ל: (1/15)/(1/30)=0.5.  
על מנת לבדוק האם קיימת תלות בין המשתנים, יש לבדוק האם השוויון מטה מתקיים (אם מתקיים, משמע מאורע A בלתי תלוי במאורע B):  
  
  
אך, במקרה זה, אפילו לא נצטרך לערוך השוואה, היות ש:  
א. כבר צוין למעלה כי ההסתברות האפשרית היחידה לכל אחת מהקטגוריות של המשתנה "כינוי" היא 1/30.  
ב. למדנו כי ישנן רק 3 אופציות להסתברות מותנית שתתקבל:   
0, 1, 0.5.  
אף אחת מההסתברויות הנ"ל אינה שווה להסתברות לכל אחת מן הקטגוריות, ומכך ניתן להסיק שהשוויון לעולם לא יתקיים ולכן כל המאורעות תלויים.  
  
מעבר לכך, נציין כי בסעיף זה לא השתמשנו בקוד היות שבמקרה זה (בו יש ריבוי קטגוריות), על מנת לבדוק האם המאורעות תלויים, פשוט יותר לעשות זאת כך.

סעיף ו'  
בשאלה זו בחרנו להביא אינטרפרטציה משלנו לפרדוקס יום ההולדת המפורסם.  
השאלה עוסקת בקומבינטוריקה, אך משלבת גם הסתברות.  
השאלה היא כדלהלן:   
כמה אנשים אנחנו צריכים על מנת שנקבל סיכוי של 90% שלפחות אדם אחד נולד בכל חודש בשנה   
(**יש לציין:** פתרון שאלה זו הינה תחת ההנחה כי אינה תלות בין ההצלחות השונות על אף שהדבר אינו בהכרח נכון).

נסתכל על בעיה זו כבעיה המתפלגת בינומית, שבה הסיכוי להצלחה הינו הסיכוי ש12 אנשים אקראיים נולדו כל אחד בחודש אחר (יוצא - 0.0000537). מכאן יש לחשב את ההסתברות המשלימה לאפס הצלחות בn ניסויים. על מנת למצוא את n יש לחשב את כמות הקומבינציות האפשריות של 12 אנשים מתוך I סה"כ - מטרת השאלה הינה למצוא עבור אילו I נקבל הסתברות השווה או גבוהה מ0.9. על מנת לבצע את החישוב בR, בחרנו להשתמש בלולאת WHILE אשר תרוץ בשימוש של אינדקס I עד לחישוב ההסתברות המתאימה - כך חישבנו את הערך המינימלי של טווח האנשים הנדרש (19 אנשים).

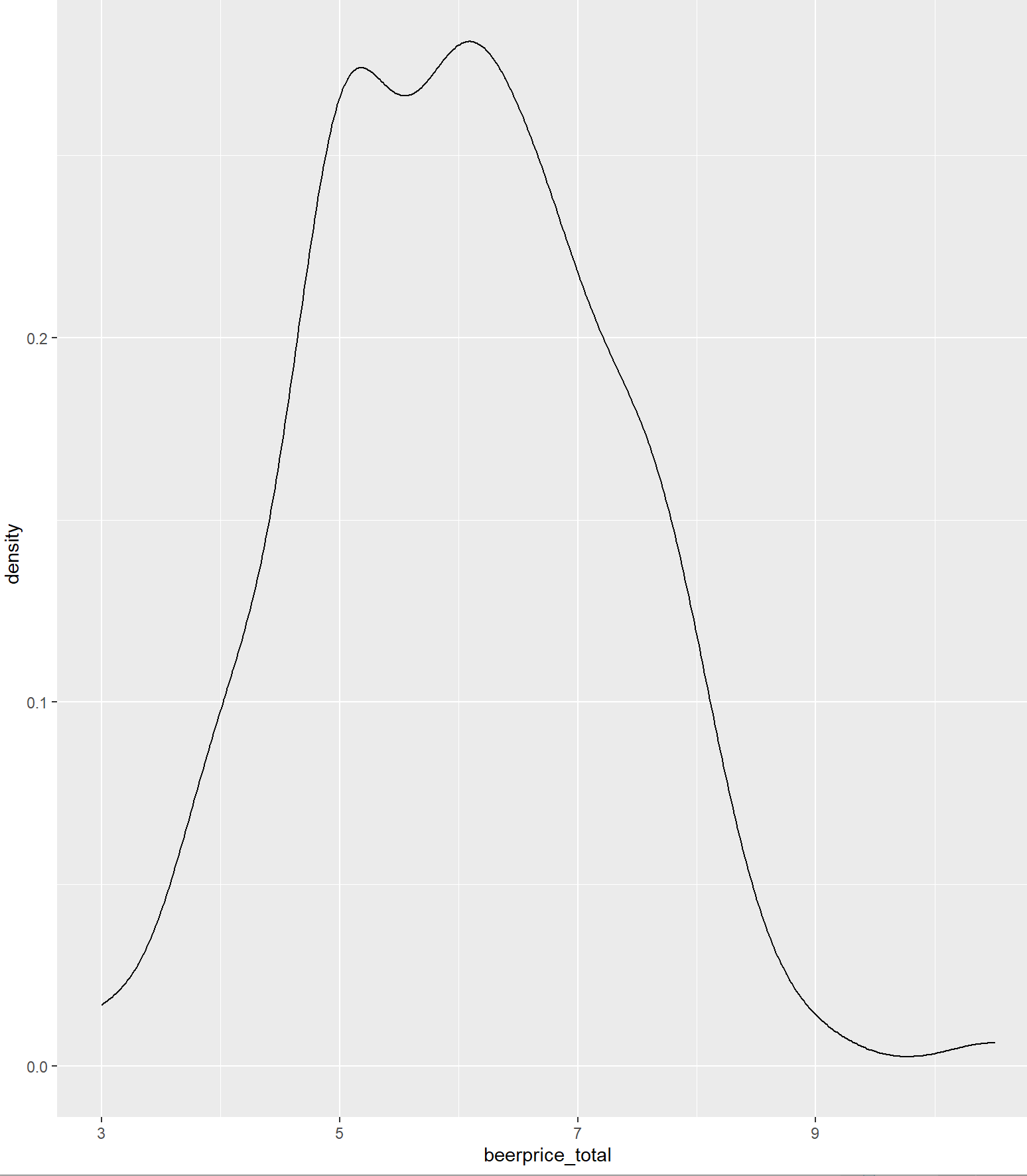
סעיף ז'

בתור המשתנה הרציף שלנו, בחרנו לעבוד עם מחירי הבירות (לאחר ניסיון עם מחירים לפי אונצ', אשר הפיקו תוצאות לא טובות לשאלה - ממוצע, חציון ושכיח שווים, זאת על אף שההתפלגות לא נורמלית. מניחים כי הדבר נוצר משונות קטנה מאוד בין ערכים). תחילה נדרשנו לחשב מדדי מרכז של המשתנה - הכנסנו את הממוצע, החציון והשכיח למשתנים נפרדים. חישוב הממוצע והחציון בוצע באמצעות פונקציות בR עצמו ולשם חישוב השכיח (שבR אין דרך קלה לחשב זאת) כתבנו פונקציה קצרה לטובת חישוב זה (פונקציה זו נלקחה מהאינטרנט מאתר: <https://www.tutorialspoint.com/r/r_mean_median_mode.htm#:~:text=R%20does%20not%20have%20a,the%20mode%20value%20as%20output> )

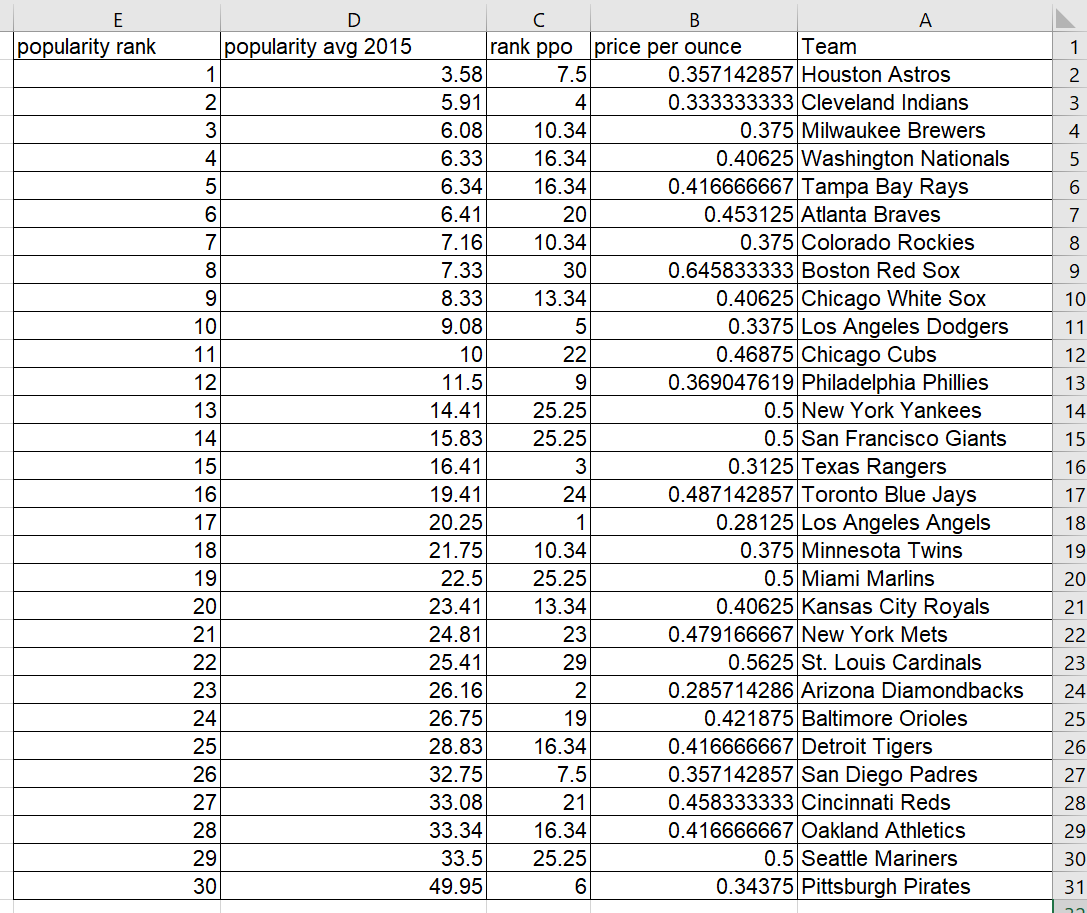
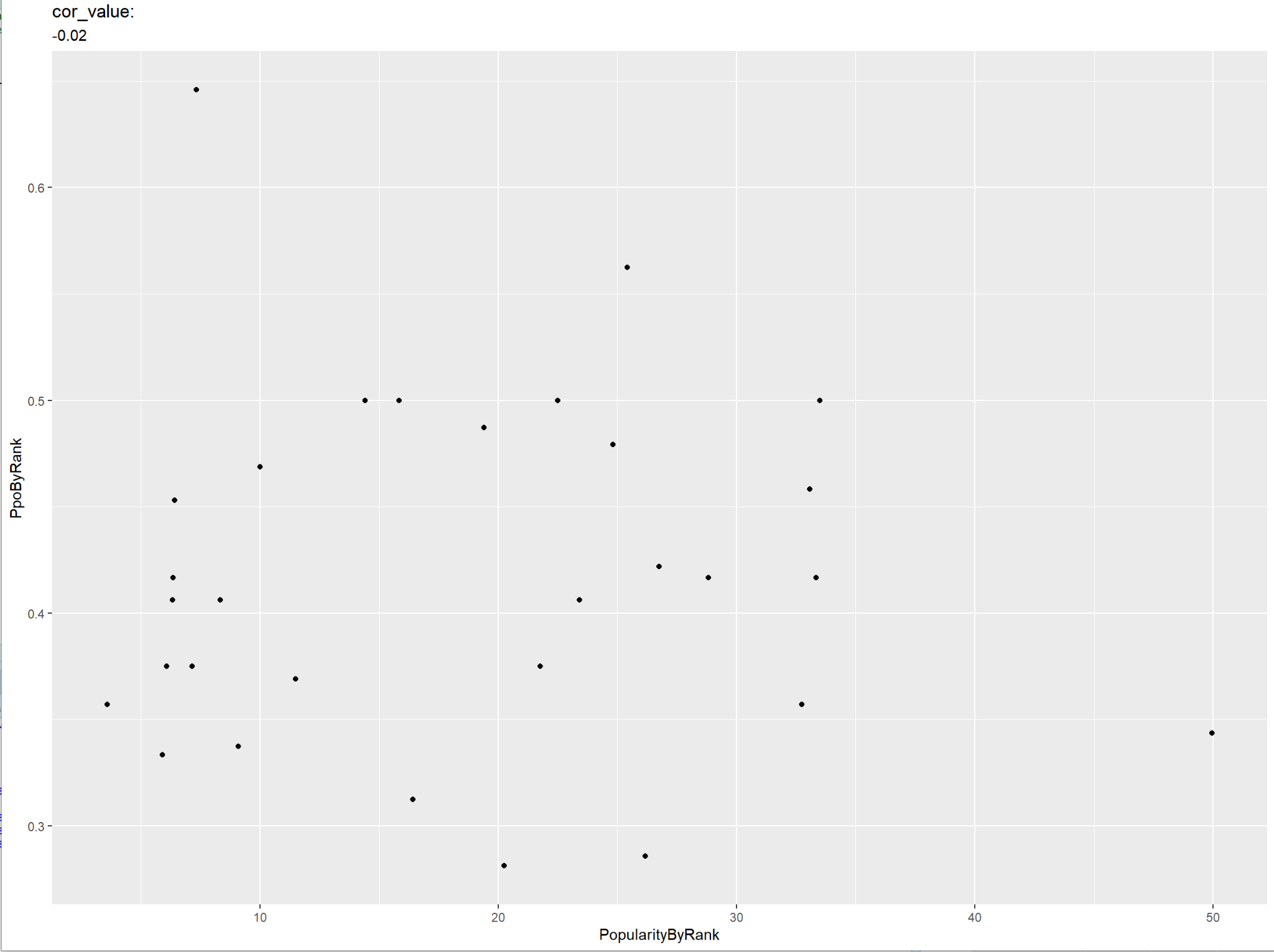
פונקציה זו מזהה ערכים חוזרים באמצעות פונקצית "UNIQUE", ומזהה איזה מהם חוזר במספר הפעמים הרב ביותר באמצעות "WHICH.MAX".

לטובת חישוב אחוזי הסטיות ממדדי הפיזור שחישבנו, כתבנו שלוש פונקציות המאפשרות חישוב מהיר שלהם. כקלט הכנסנו את כלל מחירי הבירות, וכפלט נקבל את אחוז הסטיות הרצוי.

עקרון עבודת הפונקציות הינה לולאה שרצה 150 פעמים (לפי כמות הערכים במחירי הבירות).   
כל הפעלה משווים בין הערך במקום הI של מחירי הבירות למדד המרכז הרצוי. במידה ושווה אליו, יוכנס לוקטור אינדקס ערך "1" במקום הI ובמידה ולא יוכנס הערך "NA". לאחר סיום פעולות הלולאה, מסירים ערכים חסרים מוקטור האינדקס באמצעות פונקציית "NA.OMIT", מחלקים את האורך החדש של הוקטור (לאחר ההשמטה) באורך הוקטור המקורי, ומכפילים במאה לקבלת התשובה באחוזים.   
בסוף הפונקציה של הסטיות מהממוצע הוספנו תנאי נוסף: אילו התוצאה הסופית הינה "0", התוצאה תשתנה למאה. בעצם, משמעות התנאי היא שאין שום ערך השווה בדיוק לממוצע ושכל הערכים סוטים ממנו.

על פי הניתוח, ניתן לראות שהערך הנמוך ביותר מתוך מדדי המרכז הינו השכיח (5) והערך הגבוה ביותר הינו הממוצע (~6.03).   
נסתכל על פונקציית הצפיפות על מנת להסביר טוב יותר את הממצאים. מדובר בהתפלגות דו-שכיחית וא-סימטרית חיובית.   
הערכים שגבוהים מהממוצע שכיחים יותר מאלה שמתחתיו, ובכך מושכים את הממוצע כלפי מעלה. (הממצא גם הגיוני, אין הרבה אצטדיונים שיימכרו בירות מתחת למחיר השוק).

סעיף ח'

לאחר בחינת הנתונים, עניין אותנו לחקור על הבדלי המחירים בין המקומות השונים, בדגש על ההשפעה שיש לפופולריות הקבוצות ומחירי הבירה במגרשים שלהם.   
על מנת לעשות זאת, החלטנו עשות חישוב מתאם בין מחיר הבירה לאונצ' (ounce) ובין מדד לפופולריות של הקבוצות. מדד פתוח המאפשר לבדוק פופולריות של משהו בזמן מסוים (על אף שכמובן אינו מושלם), הם הנתונים שניתן לייצא על חיפושי גוגל בזמנים שונים באתר "גוגל טרנדס". (הערה: לא יכולנו להוסיף נספח עם הנתונים הללו לקובץ הנתונים המקורי, היות שבמקרה כזה לא תוכלו למשוך את של כל העבודה בצורה טובה. לכן, מצרפים קישור- נשמח אם תוכלו להפוך לייצא את הנתונים לקובץ אקסל ולמשוך משם את הנתונים. נוסף על כך, מצרפים את טבלת הנתונים כתמונה כאן). <https://trends.google.com/trends/?geo=IL>.   
ניתן לראות כי השונות במחירי הבירה נמוכה מאוד בין השנים (רוב המגרשים שומרים על מחיר זהה לאורך השנים, גם אם הם משנים כמות בירה).  
לאור זאת, בחרנו לבדוק ממוצעי חיפוש לשנת 2015 (שרירותית) בתור מדגם לשאלה.   
ניתן להבחין מדיאגרמת הפיזור ומחישוב המדגם כי שונות המחירים ככל הנראה איננה נגזרת של פופולריות הקבוצה. (חישוב מתאם על פי R)

דיאגרמת פיזור:

השערה נוספת להסבר השונות במחירי הבירה היא הבדלים   
במצב הסוציו-אקונומי בין מדינות שונות בארה"ב,   
אך לא מצאנו מאגר נתונים מספק שיוכל לשמש כמדגם לבדיקת הקשר.

**סעיף ט'**

בסעיף זה, עלינו להציג שתי פונקציות שלא נלמדו בקורס ולהסביר את השימוש בהן.   
לשם כך, בחרנו להציג שתי פונקציות כאלו, בהן השתמשנו בעבודה שלנו בעת כתיבת הקוד (לשם פתרון שאלות כלשהן).  
**הפונקציה הראשונה בה השתמשנו היא הפונקציה: "IQR".**   
השתמשנו בה לשם חישוב הטווח הבין רבעוני, אותו נתבקשנו לחשב בסעיף א'.   
יש לציין כי "טווח בין רבעוני", הוא חלק מן החומר הנלמד בשיעורים, אך הפונקציה- כפונקציה בה ניתן להשתמש ב-R, לא נלמדה, ולכן בחרנו להציג אותה.  
טרם השימוש וההיכרות עם הפונקציה החדשה, ובהתבסס על הידע שרכשנו בשיעורים, ידענו ניתן לחשב טווח בין רבעוני על ידי שימוש בפקודה summary, שמציגה לך מן סיכום נתונים על הערכים בהתפלגות. בין היתר, הפקודה מציגה את הערך במאון ה-75 ובמאון ה-25, ואז ניתן לחסר בין הערכים ידנית ולקבל את הטווח הבין רבעוני.  
דרך חישוב שכזו היא לא בהכרח נוחה או אידיאלית, בעיקר כאשר מדובר בקוד הכולל שימוש בלולאה (מסורבל ולא אסתטי במיוחד).  
לכן, כאשר עבדנו על הסעיף, חיפשנו חלופה טובה יותר באינטרנט. מהר מאוד מצאנו את האופציה של שימוש בפונקציה IQR שהייתה יעילה בהרבה ונוחה לעבודה. את הדרך שבה השתמשנו בפונקציה, ניתן לראות בסעיף א' בקוד, ואת ההתייחסות בקובץ הוורד.

**הפונקציה השנייה אותה נציג היא: "לולאת while".**  
בפונקציה זו השתמשנו פעמיים- בסעיף ו' ובסעיף ד', היות שהיוותה חלופה מתאימה יותר ללולאת for שנלמדה בשיעור (ספציפית במקרים ההם).  
במידה ונרצה לבצע את אותה הפעולה הרבה פעמים, אנחנו יכולים להשתמש בלולאת for לשם כך.

יהיו מקרים בהם נרצה לבצע לולאה כל עוד תנאי כלשהו מתקיים, נוכל להשתמש בלולאת for עם הוספת תנאי (if) כפי שלמדנו בשיעור (אם התנאי כבר לא מתקיים הלולאה עוצרת לדוגמא), אך במקרה זה ניתן גם להשתמש בלולאת while (יכול להיות אף נוח יותר היות שהלולאה רצה כל עוד התנאי מתקיים- לא עובד על דרך השלילה כמו בלולאה הקודמת).  
זוהי לולאה פחות נפוצה, אך לעיתים שימושית (כפי שהייתה עבורנו בסעיפים המצוינים מעלה).  
לשם ההסבר עליה, נתמקד בסעיף ד' של העבודה (אותו עשינו לפני סעיף ו', ושם בעצם גילינו את אופציית השימוש בלולאת while). הסיבה שבחרנו להשתמש בלולאה זו, היא גם הסיבה שגילינו אותה. למען האמת, כשניגשנו לסעיף ד', כתבנו תחילה לולאת for עם תנאי (if), שעוצרת רק אם הפער בין הממוצע, לחציון של הערכים שדגמנו גדול מהערך 0.01. על מנת לעשות זאת (וכפי שתמיד נוהגים לעשות כאשר משתמשים בלולאת for), היה עלינו לקבוע כמה פעמים תרוץ הלולאה (אנחנו מקווים שלומר "לקבוע כמה פעמים תרוץ הלולאה" זוהי הדרך הטובה ביותר לתאר את מה שקורה בתחילת הפקודה של הלולאה: "בשביל 'איי' שנמצא במקום הראשון עד המקום ה...").   
באופן אישי הקביעה הזו (לה אנו זקוקים בעת השימוש בלולאת for) קצת עצבנה אותנו, ונראתה לנו לא מתאימה לשם מענה על הסעיף, ומכמה סיבות:  
ראשית, אין ביכולתנו לדעת כמה פעמים הלולאה תרוץ עד שתיעצר. יכול להיות שהלולאה תרוץ מספר רב מאוד של פעמים עד שתיעצר, ושהמספר שנקבע (גם אם יהיה גדול מאוד) לא יספיק.   
אך, יחד עם זאת, הרגיש לנו מיותר לקבוע מספר גדול מאוד אם הלולאה תרוץ רק מספר ממש מועט של פעמים (מה שקרה בפועל).  
נוסף על כך, למה שלמרות שהוספנו תנאי למתי הלולאה תיעצר, נצטרך בכל זאת לקבוע כמה היא תרוץ? אין משמעות למספר שנקבע. רצינו לכתוב שהיא פשוט תרוץ עד כמה שצריך על פי התנאי, אך לא מצאנו דרך לעשות זאת (כנראה כי זה בעיקר יוצר מן לופ). חשבנו אולי לקבוע שהלולאה תרוץ עד "inf", משמע- שתרוץ עד "אינסוף", כי זה לא משנה- היא בכל מקרה נעצרת ואנחנו לא יודעים כמה היא תרוץ עד שהיא תיעצר (יכול להיות שתרוץ מספר רב מאוד של פעמים ולכן לא ידענו כמה המספר ששמים צריך להיות גבוה). אך, מעבר לזה שלא ידענו האם זו פעולה שניתן לעשות בלולאות, גם הרגשנו שהפעולה קצת מסורבלת מידי. תהינו לעצמנו איך זה שאין לולאה (שאנחנו מכירים) שרצה כל עוד התנאי מתקיים, ולמה אנחנו צריכים להוסיף "ידנית" את התנאי.  
בעצם, הלולאה בכל זאת תיעצר מתישהו (בתקווה שלפני המספר שקבענו), למה שנקבע מספר סתם בשביל שהלולאה בכלל לא תרוץ את כל הפעמים הללו? למה לא לתת ללולאה "להחליט" כמה עליה לרוץ? רצינו למצוא משהו שתמיד יעבוד ושאפשר לומר "יפעיל את עצמו" כמה שצריך.  
הרגשנו עלינו לצמצם את שורות הקוד ולמצוא פקודה אסתטית ומתאימה יותר, בלי פרטים מיותרים (כמו מספרים חסרי ערך). כך, לאחר חיפוש באינטרנט, מצאנו את אפשרות השימוש בלולאת while שבאמת רצה כל עוד התנאי מתקיים, מה שנתן לנו מענה בדיוק על מה שחיפשנו- במקום לקבוע כמה תרוץ, קבענו מה התנאי שכל עוד הוא מתקיים, היא תרוץ. את הדרך שבה השתמשנו בפונקציה, ניתן לראות בסעיפים ו' ו-ד' בקוד, ואת ההתייחסות בקובץ הוורד.