פרויקט רגרסיה חלק א'

קורס : מודלים של רגרסיה לינארית

תאריך : 05/12/2021

מרצה : הלל בר גרא

מגישים : קבוצה 14

315695643

316161694

תוכן עניינים:

1. בחירת מאגר נתונים:

המאגר הנבחר - Life expectancy in countries

מאגר הנתונים המסביר באמצעות משתנים שונים את תוחלת החיים ב-101 ממדינות העולם. כל רשומה בטבלה מייצגת מדינה עליה נאספו כלל הנתונים במדדים השונים. ברשומה אודות המדינה south Sudan לא קיים ערך בשדה "Alcohol consumption per person", ולכן נבצע את כלל הניתוח ללא התייחסות לרשומה זו.

2. יצירת טבלת משתנים:

משתנה	סוג המשתנה - מוסבר/מסביר	סימון	יחידת מידה	סוג המשתנה – רציף / קטגוריאלי	הסבר קצר על המשתנה
Outdoor air pollution (%)	מסביר	X1	אחוז (%)	רציף	אחוז זיהום האויר במדינה הנבדקת המייצג את אחוז הגזים המסוכנים באוויר.
HIV - Estimated number of people that have been infected	מסביר	X2	איש	בדיד	מספר האנשים במדינה הנבדקת שנדבקו בנגיף ה-HIV מתוך כלל האוכלוסייה.
malaria - Estimated number of people that have been infected	מסביר	X3	איש	בדיד	מספר האנשים במדינה הנבדקת החולים במלריה מתוך כלל האוכלוסייה.
Average income per person ($)	מסביר	X4	דולר ($)	רציף	הכנסה שנתית ממוצעת לאדם בדולרים - ההכנסה הממוצעת המחושבת ע"י סך הכנסות של כל אזרחי המדינה לחלק בכמות האנשים בה.
Alcohol consumption per person (liters, year)	מסביר	X5	ליטר (L)	רציף	ממוצע צריכת האלכוהול השנתית לאדם בליטר - מחושב ע"י כמות האלכוהול הנצרך בשנה במדינה הנבדקת לחלק כמות האנשים בה.
density per square (km)	מסביר	X6	איש	רציף	צפיפות אוכלוסין לקילומטר מרובע - מייצג את כמות האנשים המתגוררים בקילומטר מרובע, מחושב ע"י סך האנשים במדינה לחלק בשטח שלה.
Cigarette consumption (%)	מסביר	X7	אחוז (%)	רציף	אחוז צרכני הסיגריות - מספר האנשים המעשנים במדינה ביחס לכלל האוכלוסייה.
Continent	מסביר	X8	-	קטגוריאלי	היבשת בה ממוקמת המדינה - היבשות מיוצגות ע"י מספרים חד ערכיים, כך שכל יבשת מיוצגת ע"י מספר שונה. 1-אסיה, 2- אפריקה והמפרץ הפרסי, 3- דרום אמריקה, 4- אירופה, 5- מרכז אמריקה.
Member of OECD	מסביר	X9	-	קטגוריאלי	שייכות לארגון ה-OECD - האם המדינה חברה בארגון,כך שמיוצג ע"י משתנה בינארי (1-כן, 0-לא).
Life expectancy (year)	מוסבר	y	שנה	רציף	תוחלת החיים במדינה - אומדן למספר השנים הממוצע שבני אדם חיים במדינה הנבדקת.

3. תיאור המשתנים:

Outdoor air pollution (X1) - אחוז גבוה של זיהום אויר במדינה משפיע לרעה על בריאותם של האזרחים בה, ולכן ייתכן ויוביל לתוחלת חיים נמוכה ביחס למדינות בהן זיהום האויר נמוך יותר.

HIV infected (X2) - עבור אנשים שנדבקו בנגיף ישנה השפעה לרעה על בריאותם, מקרה שעלול להוביל לתמותה בגיל צעיר. לא ניתן לדעת באמצעות משתנה זה בלבד על ההשפעה על תוחלת החיים מכיוון ומדובר במספר האנשים שנדבקו ולא באחוז שלהם מסך האוכלוסייה. בעת ניתוח נתונים אלו לא ניתן לדעת האם כמות הנדבקים היא גבוהה או נמוכה, ולכן לא ניתן לבצע השוואה בין מדינות שונות.

Malaria infected (X3) - עבור אנשים החולים במלריה ישנה השפעה לרעה על בריאותם, מקרה שעלול להוביל לתמותה בגיל צעיר. לא ניתן לדעת באמצעות משתנה זה בלבד על ההשפעה על תוחלת החיים מכיוון ומדובר במספר האנשים החולים ולא באחוז שלהם מסך האוכלוסייה. בעת ניתוח נתונים אלו לא ניתן לדעת האם כמות החולים היא גבוהה או נמוכה, ולכן לא ניתן לבצע השוואה בין מדינות שונות.

Average income per person (X4) - ההכנסה גבוהה מאפשרת איכות חיים טובה הכוללת מזון בריא, פעילות ספורטיבית, טיפול רפואי איכותי ושמירה על אורח חיים בריא. מרבית האנשים שהכנסתם נמוכה מהממוצע ולא יכלו להרשות לעצמם את אמצעים אלה, ולכן ייתכן ותהיה לכך השפעה על בריאותם ותוחלת חייהם.

Alcohol consumption per person (X5) - ידוע שצריכת אלכוהול מוגברת מזיקה לבריאות, ולכן ייתכן ובמדינות בהן כמות צריכת האלכוהול לאדם היא גבוהה, תוחלת החיים תהיה נמוכה.

Density per square (X6) - כאשר ישנה צפיפות גבוהה יש פחות מרחב אישי, ישנם פחות מרחבים פתוחים ושטחים ירוקים, קל יותר להידבק במחלות וזיהום האויר גבוה יותר. לכן, ייתכן ותוחלת החיים במקומות אלו תהיה נמוכה יותר ביחס למדינות בהן הצפיפות נמוכה יותר.

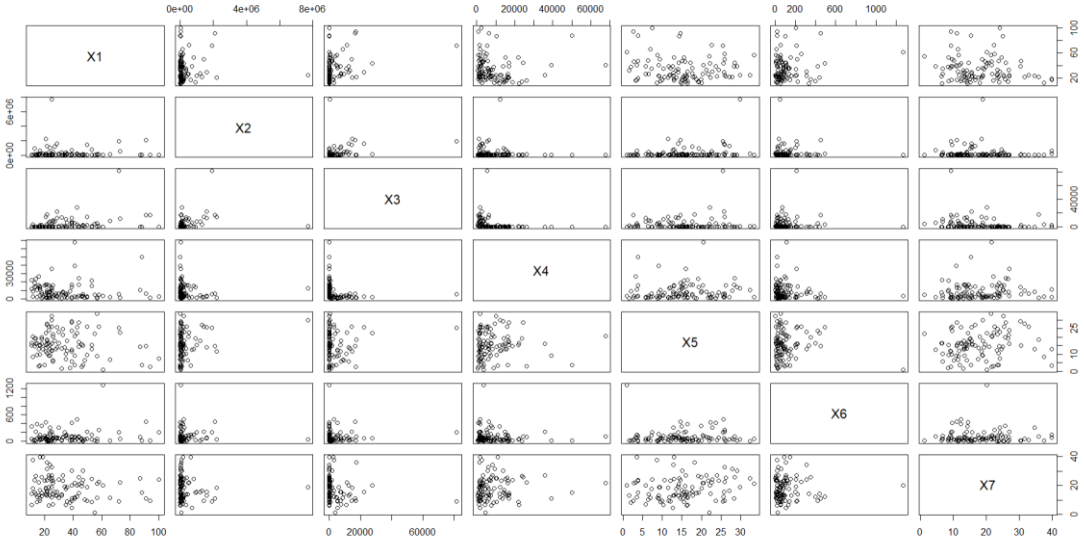
Cigarette consumption (X7) - ידוע שעישון סיגריות מזיק לבריאות, ולכן ייתכן ובמדינות אחוז המעשנים הוא גבוה, תוחלת החיים תהיה נמוכה.

Continent (X8) - ביבשות שונות ישנו סגנון חיים שונה המבוסס על תרבות, ותנאי שטח ואקלים. מאפיינים אלו משפיעים על הטכנולוגיה (תקשורת, מזון, רכבים), תשתיות (מים, חשמל, כבישים), ההשכלה וחינוך ותחומי חיים נוספים. לכן ייתכן שבמדינות שונות ביבשת מסוימת תוחלת החיים תהיה דומה.

Member of OECD (X9) - בארגון ה-OECD משתתפות מדינות שונות המקיימות שיתוף פעולה כלכלי. מדינות אלו הן מדינות דמוקרטיות, ליברליות, היוצרות תוצר כלכלי עולמי. השתתפות מדינה בארגון זה מעיד על היותה מדינה מפותחת, דבר המשפיע על תנאי המחייה של האזרחים בה וייתכן גם על תוחלת החיים שלהם.

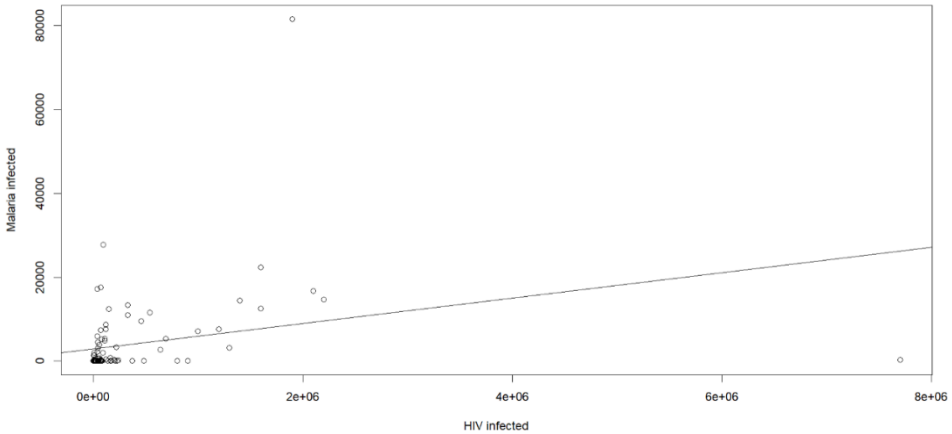
4. תיאור קשרים בין משתנים:

- גרף המקשר בין כל זוג משתנים מסבירים רציפים:



- <u>קשרים סיבתיים - השערות ותוצאות:</u>

1. ייתכן וקיים קשר (קורלציה חיובית) בין מספר האנשים הנדבקים ב-HIV (X2) לבין מספר החולים במלריה (X3). כלומר, במדינות בהן יש מספר גבוה של נדבקים ב-HIV יהיו גם מספר רב של חולים במלריה. זאת, משום שייתכן ובמדינות אלו לא קיים ידע רפואי מתאים, או שהתנאים הסניטריים לא מתאימים לטיפול בנגיפים אלו. כמו כן, בסבירות גבוהה אין מודעות להתגוננות מפני הדבקה באותן מדינות.

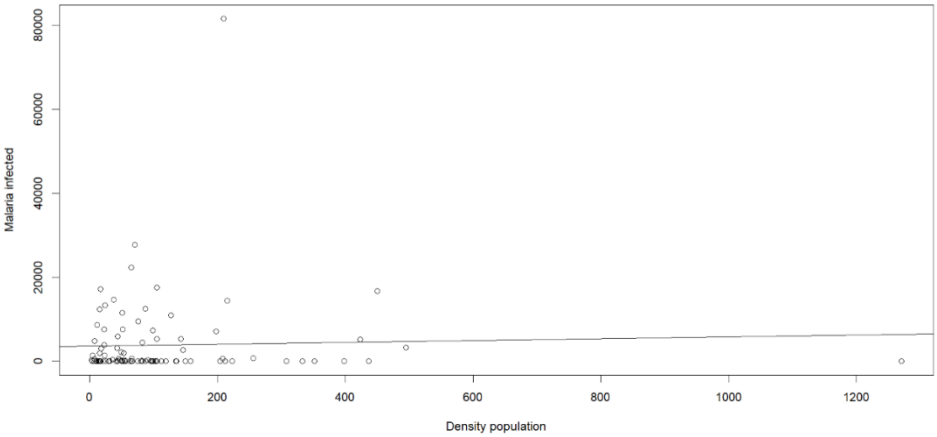


כפי שניתן לראות ערך הקורלציה הוא 0.278 והינו הגדול ביותר מבין הקשרים שנבדקו. לכן, קיים קשר סיבתי חיובי בין מספר האנשים הנדבקים ב-HIV לבין מספר החולים במלריה.

```
> cor(subdata$X2, subdata$X3) %>% print()
[1] 0.27891
```

2. ייתכן וקיים קשר (קורלציה חיובית) בין צפיפות אוכלוסין (X6) לבין מספר החולים במלריה (X3). כלומר, במדינות בהן צפיפות האוכלוסין גבוהה נצפה לראות מספר גבוהה יותר של אנשים החולים במלריה. בעקבות היעדר מרחב אישי הסיכוי להידבק במחלות מהאוכלוסייה שבסביבתך גבוה יותר.

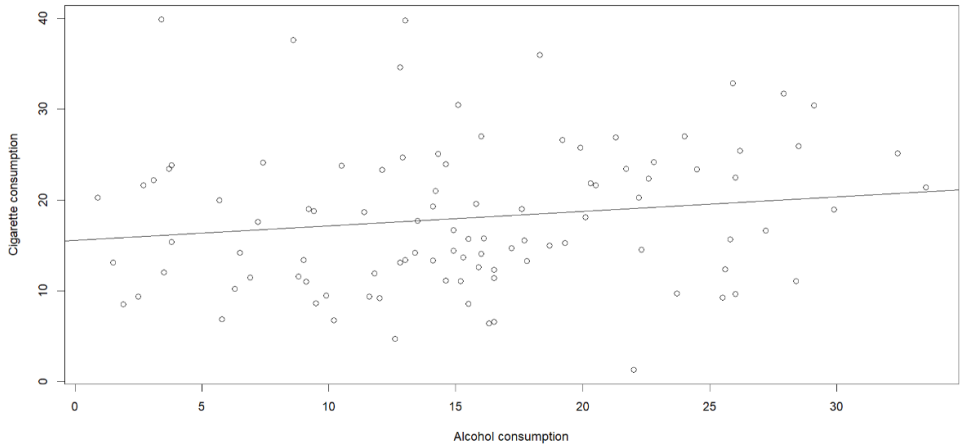
כפי שניתן לראות ערך הקורלציה הוא 0.037 ולכן ניתן להסיק כי לא קיים קשר בין צפיפות האוכלוסין לבין מספר החולים במלריה. כלומר, בניגוד להשערתנו לא קיים קשר בין משתנים אלה.




```
> cor(subdata$X6, subdata$X3) %>% print()
[1] 0.03723625
```

3. ייתכן וקיים קשר (קורלציה חיובית) בין צריכת האלכוהול (X5) לצריכת סיגריות (X7). כלומר, במדינות בהן צריכת אלכוהול השנתית עבור אדם היא גבוהה ייתכן ואחוזי המעשנים גבוהה בהתאמה. במחקרים שונים (לדוגמה, Alcoholism: Clinical & Experimental Research) נמצא כי אלכוהול משפיע על המוח בכך שהוא מגביר את הדחף לעישון. כמו כן, משום שמדובר בשני צריכות המזיקות לבריאות הגורמות להתמכרות, ייתכן ואנשים הצורכים אחת מהן יצרכו גם את השנייה.

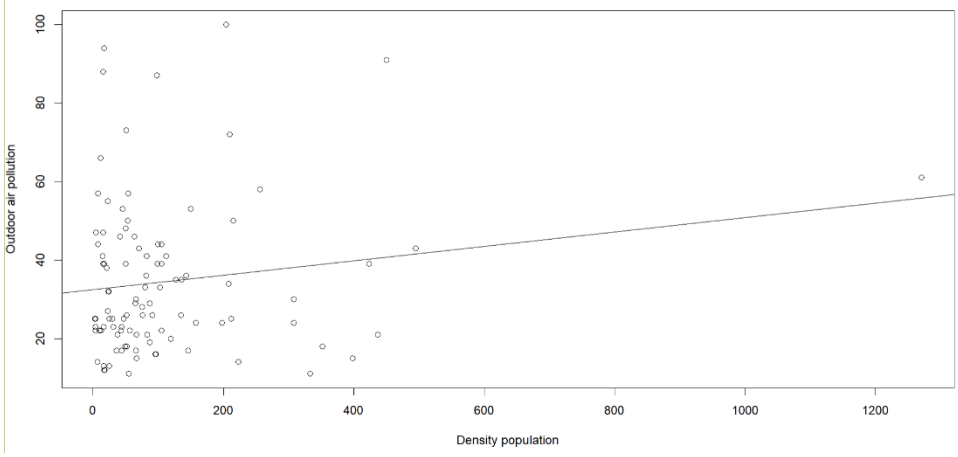
כפי שניתן לראות ערך הקורלציה הוא 0.155, ולכן ניתן להסיק שישנה קורלציה חיובית מזערית בין צריכת אלכוהול לצריכת סיגריות.




```
> cor(subdata$X5, subdata$X7) %>% print()
[1] 0.1550073
```

4. יתכן וקיים קשר (קורלציה חיובית) בין צפיפות אוכלוסין (X6) לבין זיהום אוויר (X1). כלומר, במדינות בהן צפיפות האוכלוסין גבוהה נצפה לראות אחוז זיהום אוויר גבוה יותר. בהנחה שאדם ממוצע גורם לפליטה אחידה של פחמן דו-חמצני ללא קשר למדינה בה הוא מתגורר, במדינות בהן צפיפות גבוהה יותר, אחוז זיהום האוויר במדינה יהיה גבוה יותר. כמו כן, במדינות צפופות קיימים פחות שטחים פתוחים וירוקים המשפיעים לטובה על איכות האוויר.

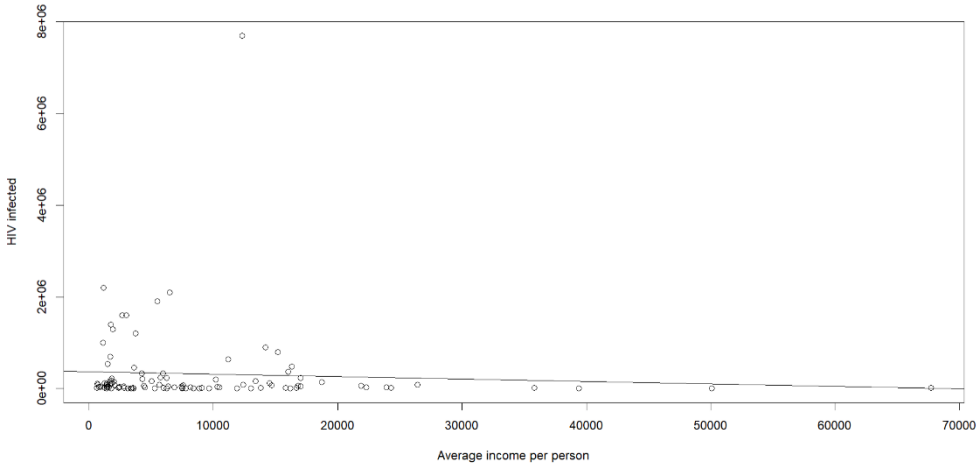
כפי שניתן לראות ערך הקורלציה הוא 0.151, ולכן ניתן להסיק שקיימת קורלציה חיובית מזערית בין צפיפות האוכלוסין לאחוז זיהום האוויר.




```
> cor(subdata$X6, subdata$X1) %>% print()
[1] 0.151569
```

5. ייתכן וקיים קשר (קורלציה שלילית) בין ההכנסה השנתית הממוצעת לאדם (X4) לבין מספר הנדבקים ב-HIV (X2). כלומר, ככל שההכנסה עולה כמות הנדבקים יורדת. במדינות בהן ההכנסה גבוהה יותר יש יכולת לאוכלוסייה לקנות אמצעי מניעה דבר שיקטין את כמות הנדבקים במחלות מין בכלל וב-HIV בפרט.

כפי שניתן לראות ערך הקורלציה הוא -0.064 אשר מראה על קורלציה שלילית כפי ששיערנו, אך היא איננה משמעותית, ולכן בניגוד להשערתנו ניתן להסיק כי לא קיים קשר בין ההכנסה השנתית הממוצעת למספר הנדבקים ב-HIV.




```
> cor(subdata$X4, subdata$X2) %>% print()
[1] -0.06411276
```

לסיכום, בכלל הקשרים שהצגנו לא נמצאו קשרים סיבתיים עם מדדי קורלציה מובהקים, ולכן ניתן להסיק שאינם משפיעים אחד על השני. בנוסף, גם בתרשים הפיזור הראשי לא נצפו קשרים מדגמיים נוספים בעלי קורלציה מובהקת.

5. ניתוח תיאורי של המשתנים:

X1 - אחוז זיהום האוויר:

ממוצע: 34.48
חציון: 28.5
תחום בין רבעוני: 21.75-43.25
סטיית תקן: 19.38628
א-סימטריה: 1.394302

```
> summary(Dataset$X1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  11.00   21.75   28.50   34.48   43.25  100.00
> sd(Dataset$X1)
[1] 19.38628
> skewness(Dataset$X1)
[1] 1.394302
```

ברוב המדינות שנבדקו במדגם אחוז זיהום האוויר הינו נמוך מהממוצע, כלומר קיימות מדינות מעטות עם זיהום אוויר גבוה המעלות את הממוצע כלפי מעלה ביחס לחציון (התפלגות א-סימטרית חיובית עם זנב ימני).

X2 - כמות הנדבקים ב-HIV:

ממוצע: 323,178
חציון: 56,000
תחום בין רבעוני: 14,596-202,500
סטיית תקן: 881656.9
א-סימטריה: 6.18098

```
> summary(Dataset$X2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    200   14596   56000  323178  202500 7700000
> sd(Dataset$X2)
[1] 881656.9
> skewness(Dataset$X2)
[1] 6.18098
```

ניתן לראות פער משמעותי בין החציון לממוצע ושונות גבוהה בתצפיות, מכך ניתן להסיק שברוב המדינות שנבדקו במדגם כמות הנדבקים ב-HIV הינו נמוך מהממוצע. כלומר, קיימות מספר מדינות עם כמות נדבקים גבוהה משמעותית ביחס לשאר (התפלגות א-סימטרית חיובית עם זנב ימני). בנוסף, התחום הבין רבעוני הוא קטן ביחס לטווח כל התצפיות, ומכך ניתן להסיק שפיזור הנתונים הינו מצומצם.

X3 - כמות החולים במלריה:

ממוצע: 3,828.9
חציון: 33.5
תחום בין רבעוני: 0-4,004.5
סטיית תקן: 9,607.835
א-סימטריה: 5.646061

```
> # X3
> summary(Dataset$X3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     0.0    33.5  3828.9  4044.5 81640.0
> sd(Dataset$X3)
[1] 9607.835
> skewness(Dataset$X3)
[1] 5.646061
```

מהנתונים עולה כי ישנן מדינות רבות עם מספר מצומצם של חולים במלריה או ללא חולים כלל, ולכן קיים פער בין החציון לממוצע. כלומר, ישנן מספר מדינות בעלות מספר חולים רבים המשפיעים על הממוצע (התפלגות א-סימטרית חיובית עם זנב ימני).

X4 - הכנסה ממוצעת לאדם:

ממוצע: 8,936
חציון: 5,685
תחום בין רבעוני: 1,895-12,550
סטיית תקן: 10,525.01
א-סימטריה: 2.806337

```
> # X4
> summary(Dataset$X4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    627    1895    5685    8936   12550   67700
> sd(Dataset$X4)
[1] 10525.01
> skewness(Dataset$X4)
[1] 2.806337
```

טווח ההכנסה הממוצעת לאדם הינו רחב ומשתנה בין המדינות. משום שהחציון נמוך מהממוצע, ניתן להסיק שברוב המדינות ההכנסה הממוצעת נמוכה בהכנסה הממוצעת במדגם. מכיוון והערך המייצג את הא-סימטריות הוא חיובי להתפלגות המשתנה קיים זנב ימני.

X5 - צריכת אלכוהול לאדם (ליטר בשנה):

```
> # X5
> summary(Dataset$X5)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.90    9.80   15.15   15.43   20.70   33.50
> sd(Dataset$X5)
[1] 7.730745
> skewness(Dataset$X5)
[1] 0.1485117
```

ממוצע: 15.43
חציון: 15.15
תחום בין רבעוני: 9.8-20.7
סטיית תקן: 7.730745
א-סימטריה: 0.1485117

על פי הנתונים ניתן לראות שערך הממוצע וערך החציון של כמות צריכת האלכוהול כמעט שווים, וכי מדד הא-סימטריות מזערי (זנב ימני לא משמעותי). כלומר, ההתפלגות דומה במאפיינה להתפלגות נורמלית כך שרוב התצפיות קרובות לממוצע.

X6 - צפיפות האוכלוסייה (אנשים לקמ"ר):

```
> summary(Dataset$X6)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.08   23.80   65.40  111.66  121.00 1270.00
> sd(Dataset$X6)
[1] 160.3229
> skewness(Dataset$X6)
[1] 4.257173
```

ממוצע: 111.66
חציון: 65.40
תחום בין רבעוני: 23.8-121
סטיית תקן: 160.3229
א-סימטריה: 4.257173

ברוב המדינות שנבדקו במדגם צפיפות האוכלוסייה הינה נמוכה מהממוצע, כלומר קיימות מדינות מעטות עם צפיפות גבוהה אשר מעלות את הממוצע כלפי מעלה ביחס לחציון (התפלגות א-סימטרית חיובית עם זנב ימני).

X7 - אחוז המעשנים:

```
> summary(Dataset$X7)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.30   12.01   16.68   18.04   23.45   39.90
> sd(Dataset$X7)
[1] 7.981384
> skewness(Dataset$X7)
[1] 0.6039282
```

ממוצע: 18.04
חציון: 16.68
תחום בין רבעוני: 12.01-23.45
סטיית תקן: 7.981384
א-סימטריה: 0.6039282

על פי הנתונים ניתן לראות שערך הממוצע וערך החציון של אחוז המעשנים קרובים, וכי מדד הא-סימטריות הינו נמוך (זנב ימני לא משמעותי). כלומר, ההתפלגות מזכירה במאפיינה להתפלגות נורמלית כך שרוב התצפיות קרובות לממוצע.

X8 - יבשת:

```
> summary(Dataset$X8)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    1.00    2.00    2.12    2.00    5.00
```

חציון: 2
תחום בין רבעוני: 1-2

משום שמדובר במשתנה קטגוריאלי כך שכל יבשת מיוצגת ע"י ספרה ייחודית, הערכים המתקבלים מייצגים את מיקום המדינות. לכן לא קיימת לערך הממוצע, לסטיית התקן, ולמדד הא-סימטריות. לעומת זאת, מניתוח ערכי הרבעונים והחציון ניתן להסיק את המסקנות הבאות:

1. לפחות רבע מהמדינות נכללות ביבשת אסיה (1).
2. לפחות מחצית מהמדינות נכללות ביבשות אסיה (1) או אפריקה והמפרץ הפרסי (2).
3. פחות מרבע מהמדינות נכללות תחת היבשות דרום אמריקה (3), אירופה (4), מרכז אמריקה (5).

X9 - חברות בארגון ה-OECD :

ממוצע : 0.02

חציון : 0

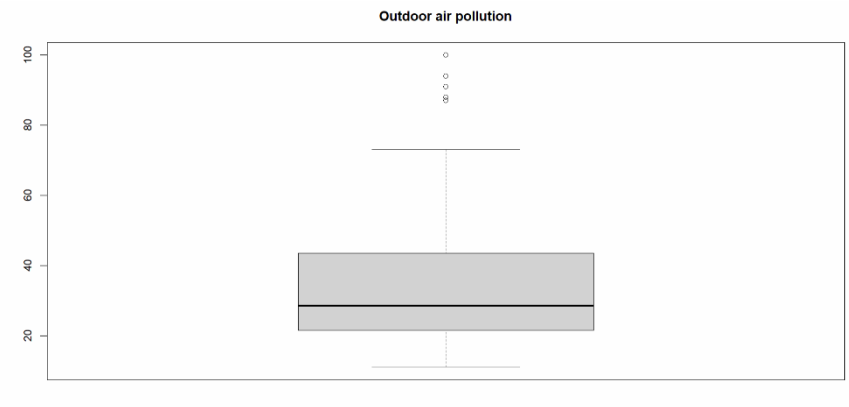
תחום בין רבעוני : 0-0

```
> # X9
> summary(Dataset$X9)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    0.00    0.02    0.00    1.00
```

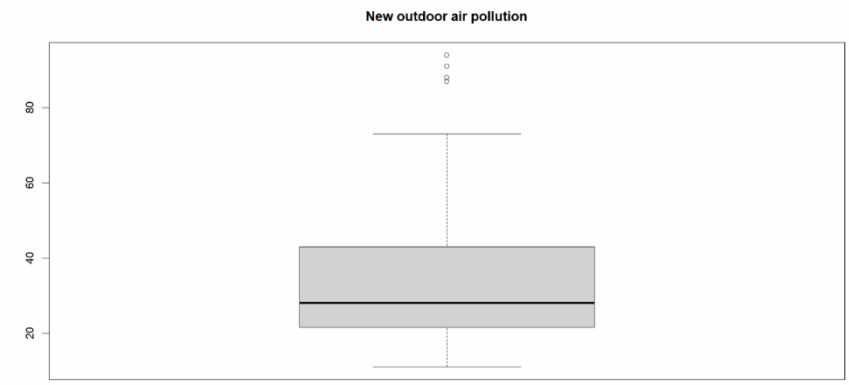
משתנה זה הינו קטגוריאלי מסוג בוליאני, ולכן ניתן לנתח את המשתנה בעזרת ערך הממוצע. מכיוון וערך הממוצע הוא 0.02, ניתן לומר שבדיוק מ-2% מהמדינות שנבדקו הינן חברות בארגון ה-OECD.

6. ניתוח חריגים:

- X1 אחוז זיהום האוויר :

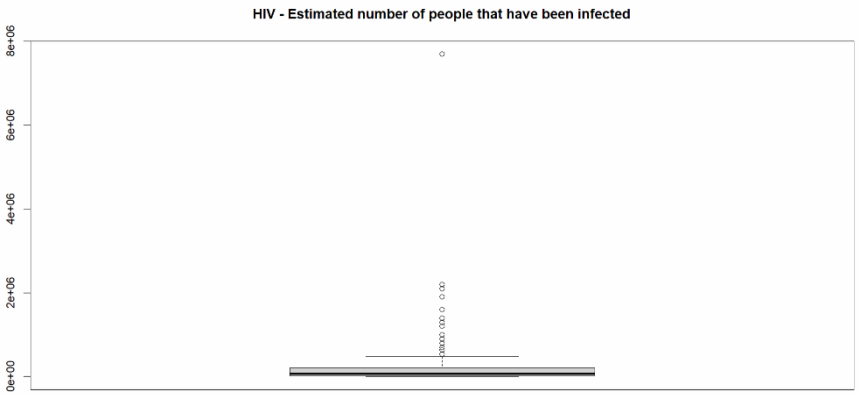


כפי שניתן לראות עבור משתנה זה ישנן חמש תצפיות חריגות ביחס למדינות שנבדקו. עם זאת, ייתכן ונתונים אלו הינם אמיתיים ונמדדו בשטח, ולכן לא ניתן להוריד את רשומות אלו על בסיס היותם חריגים בלבד. במקרה זה, אחוז זיהום הגבוה מ-98% מטיל ספק באמינות המדידה וייתכן והמכשיר אינו כויל כראוי. הנתון היחיד הגבוה מרף זה הוא 100% זיהום אוויר שנמדד במדינה נפאל, ולכן נבחר להוריד רשומה זו בהמשך ניתוח הנתונים.



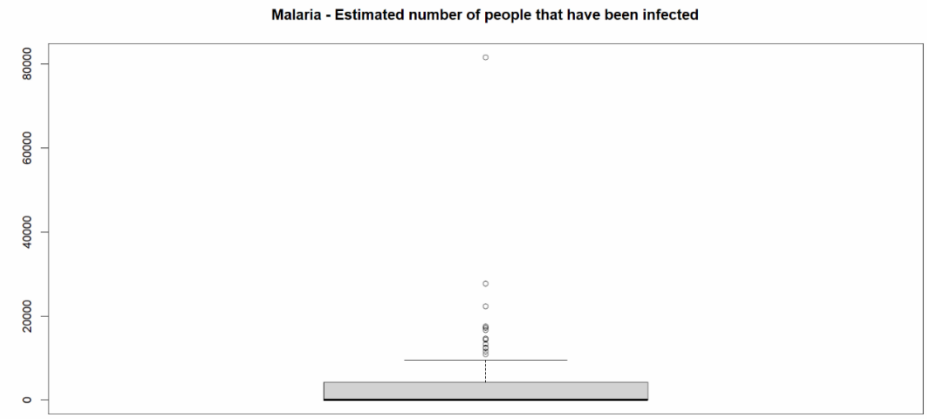
לאחר הוצאת הרשומה החריגה נקבל את התרשים הבא, ללא שינויים בשאר הנתונים החריגים :

- X2 - כמות הנדבקים ב-HIV :



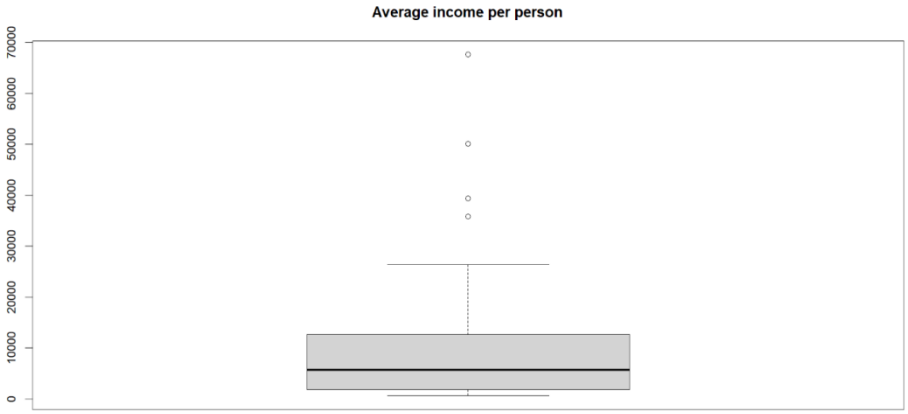
למרות שניתן לראות מספר רב של נתונים חריגים עבור משתנה זה, נבחר שלא להוריד את הרשומות החריגות. קיימות מספר רב של מדינות עם מספר מצומצם של נדבקים ב-HIV, ולכן מדינות בהן הנגיף נפוץ יותר יראו כחריגות. בהסתכלות על הנתונים, מרבית המדינות החריגות ממוקמות באפריקה ולכן לא נרצה להוריד רשומות אלה מכיוון שאין חשד במהימנות הנתון, וחשוב להתייחס בעת הניתוח גם למדינות מסוג זה.

- <u>X3 - כמות החולים במלריה :</u>



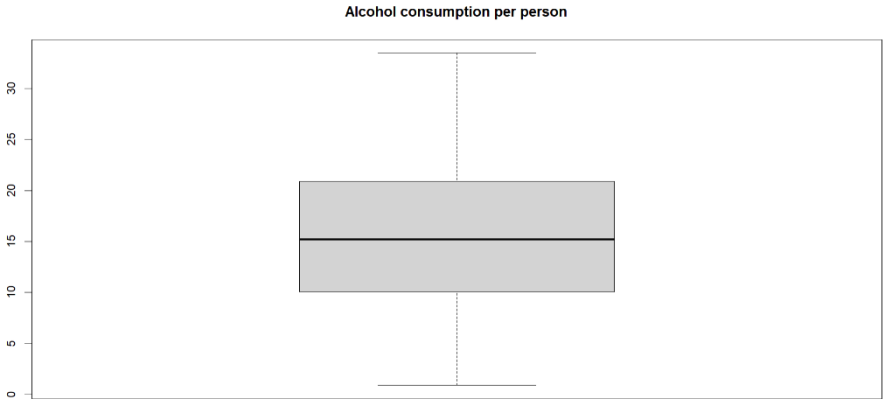
למרות שניתן לראות מספר רב של נתונים חריגים עבור משתנה זה, נבחר שלא להוריד את הרשומות החריגות. קיימות מספר רב של מדינות בהן אין חולים כלל במלריה, ולכן מדינות בהן המחלה נפוצה מסומנות כחריגות. בהסתכלות על הנתונים, מרבית המדינות החריגות ממוקמות באפריקה ולכן לא נרצה להוריד רשומות אלה מכיוון שאין חשד במהימנות הנתון, וחשוב להתייחס בעת הניתוח גם למדינות מסוג זה. מבדיקה על מחלת המלריה נמצא כי מוקדי ההתפשטות המרכזיים של המלריה הם מדרום למדבר סהרה. כמו כן, הנתון החריג ביותר מייצג את המדינה ניגריה הממוקמת באזור זה. לכן נתון זה רלוונטי, ונרצה להתייחס אליו בניתוח הנתונים.

- <u>X4 - הכנסה ממוצעת לאדם :</u>



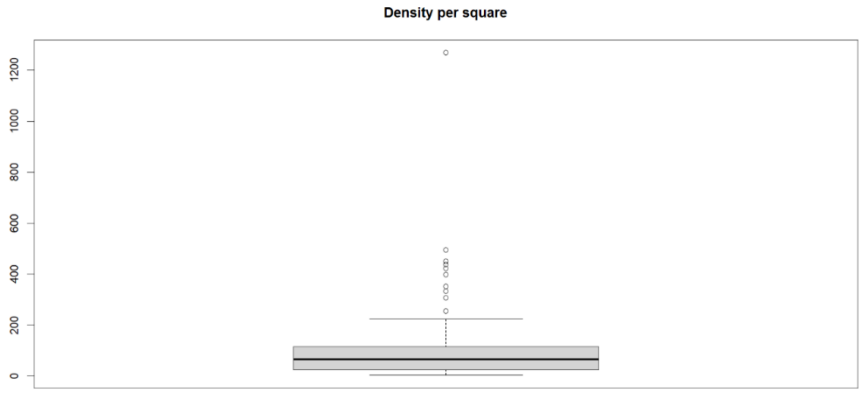
כפי שניתן לראות ישנן ארבע תצפיות חריגות, מכיוון ומשתנה זה מייצג הכנסה ממוצעת לאדם במדינה, הגיוני ויהיו פערים בין מדינות שונות בעולם. בהסתכלות על הנתונים שבידינו, ארבעת המדינות עם ההכנסה הממוצעת הגבוהה ביותר לאדם הן איחוד האמירויות, ערב הסעודית, עומאן ודרום קוריאה. מדינות אלו הינן מדינות מפותחות, עם כלכלה יציבה ומצב סוציואקונומי גבוה, ואכן מייצגות את האוכלוסייה במדינה. בעקבות זאת, נבחר שלא להוריד רשומות אלו.

- <u>X5 - צריכת אלכוהול לאדם (ליטר בשנה) :</u>



כפי שניתן לראות בתרשים, עבור משנה זה לא קיימות תוצאות חריגות.

- <u>X6 - צפיפות האוכלוסייה (אנשים לקמ"ר):</u>



כפי שניתן לראות בתרשים, קיימות מספר רב של מדינות המסומנות כחריגות. מדד זה מייצג את צפיפות האוכלוסייה במדינה ולכן הגיוני שפיזור הנתונים לא יהיה אחיד. לכן, נבחר שלא להוריד את הרשומות של מדינות אלה.

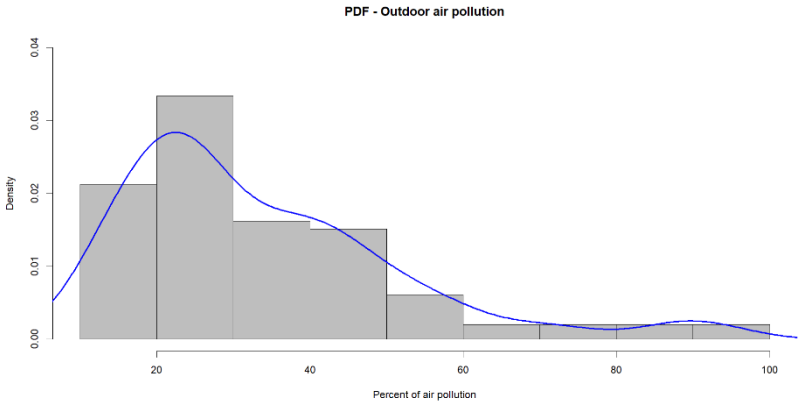
- <u>X7 - אחוז המעשנים:</u>

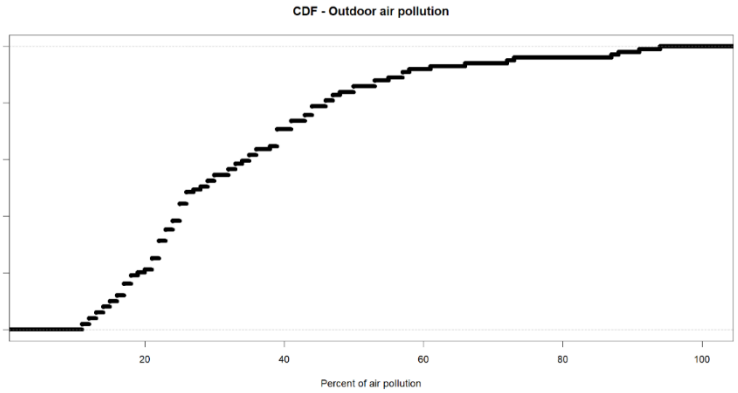
כפי שניתן לראות בתרשים, עבור משנה זה לא קיימות תוצאות חריגות.

- עבור משתנים קטגוריאליים אין משמעות לניתוח החריגים בעזרת תרשים מסוג זה, ולכן נצטרך לבצע בדיקה ידנית על מנת לוודא את נכונות הקטגוריה שנבחרה לכל רשומה. לאחר מעבר על שני משתנים אלו (יבשת, שייכות ל-OECD) לא קיימות תצפיות חריגות בבסיס הנתונים.

7. פונקציית צפיפות והתפלגות מצטברת:

- <u>משתנה מסביר X1 - אחוז זיהום האוויר</u>

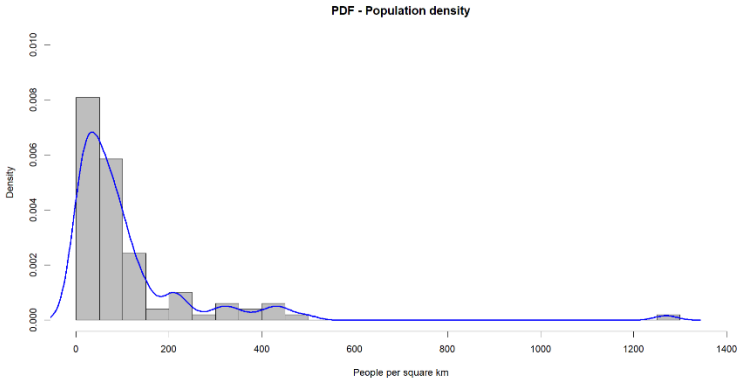


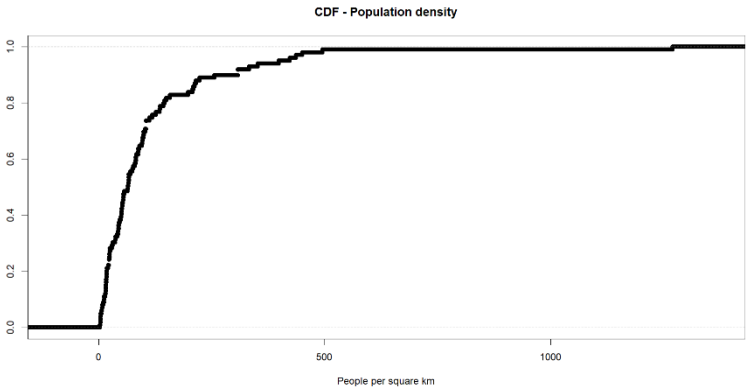


כפי שניתן לראות בפונקציית הצפיפות מרבית התצפיות נמצאות בטווח שבין 20-30, ומדובר בא-סימטריה חיובית עם זנב ימני. בנוסף, ניתן לראות בפונקציית ההתפלגות המצטברת כי כ-50% מהמדינות שנדגמו מדד זיהום האוויר שלהם הוא עד הערך 28%, בהתאמה לערך החציון של משתנה זה. כמו כן, בפונקציית הצפיפות מרבית הערכים של משתנה זה מתקבלים עד לערך 60%, זאת

בהתאמה לפונקציית ההתפלגות המצטברת שהינה קעורה עם שיפוע גדול יותר עד לערך זה, ולאחר מכן הפונקציה ממשיכה לעלות עם שיפוע קטן יותר.

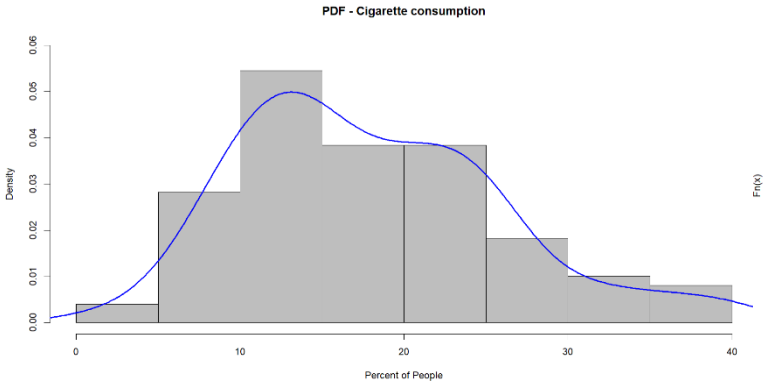
- <u>משתנה מסביר X6 - צפיפות אוכלוסין</u>

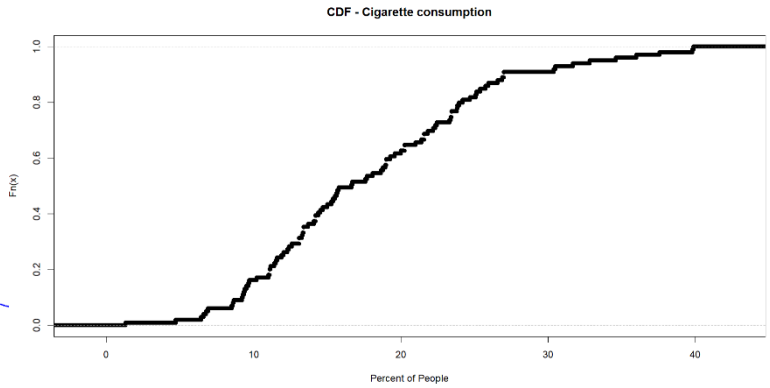




ברוב המדינות שנבדקו צפיפות האוכלוסין היא קטנה מ-150 אנשים לקילומטר, ובנוסף קיימות מספר מדינות בהן צפיפות האוכלוסין היא גבוהה יותר. לכן, ההתפלגות היא א-סימטרית עם זנב ימני. בנוסף, פונקציית ההתפלגות המצטברת היא קעורה עם שיפוע תלול עד הערך 150 (כ-75% מהתצפיות), ולאחר מכן היא ממשיכה לעלות בשיפוע קטן יותר.

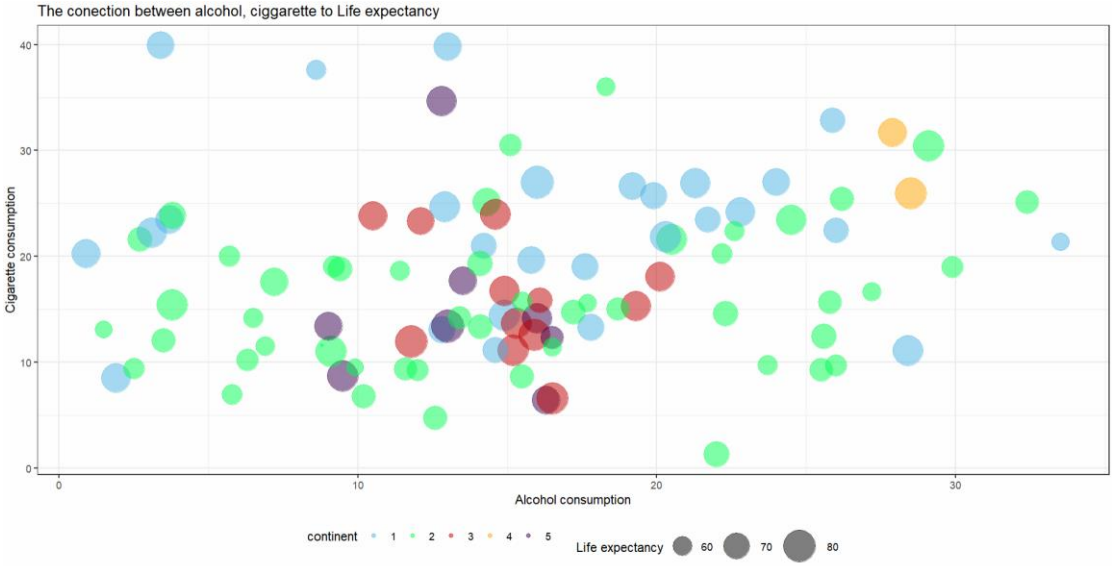
- <u>משנה מסביר X7 - אחוז המעשנים</u>



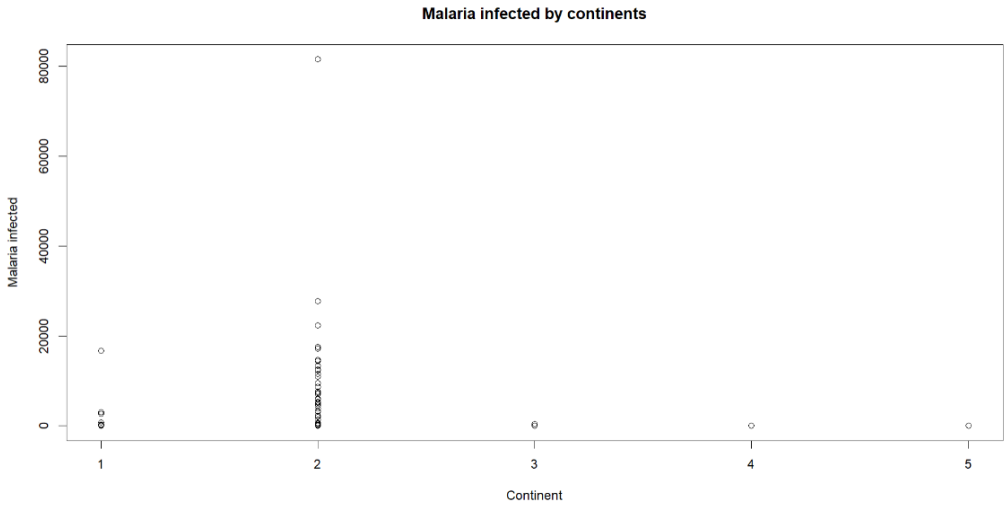


כפי שניתן לראות בפונקציית הצפיפות, ההתפלגות של משתנה זה מזכירה התפלגות נורמלית עם זנב ימני. זאת, משום שהערכים מתפזרים באופן יחסית סימטרי סביב התוחלת. בהתאם, בפונקציית ההתפלגות המצטברת השיפוע ברובו אחיד, ללא שינויים משמעותיים. כמו כן, כלל התצפיות שבהן עד 15% מהאוכלוסייה מעשנים, מהוות 50% מכלל המדינות שנבדקו, זאת בהתאמה לערך החציון של המשתנה.

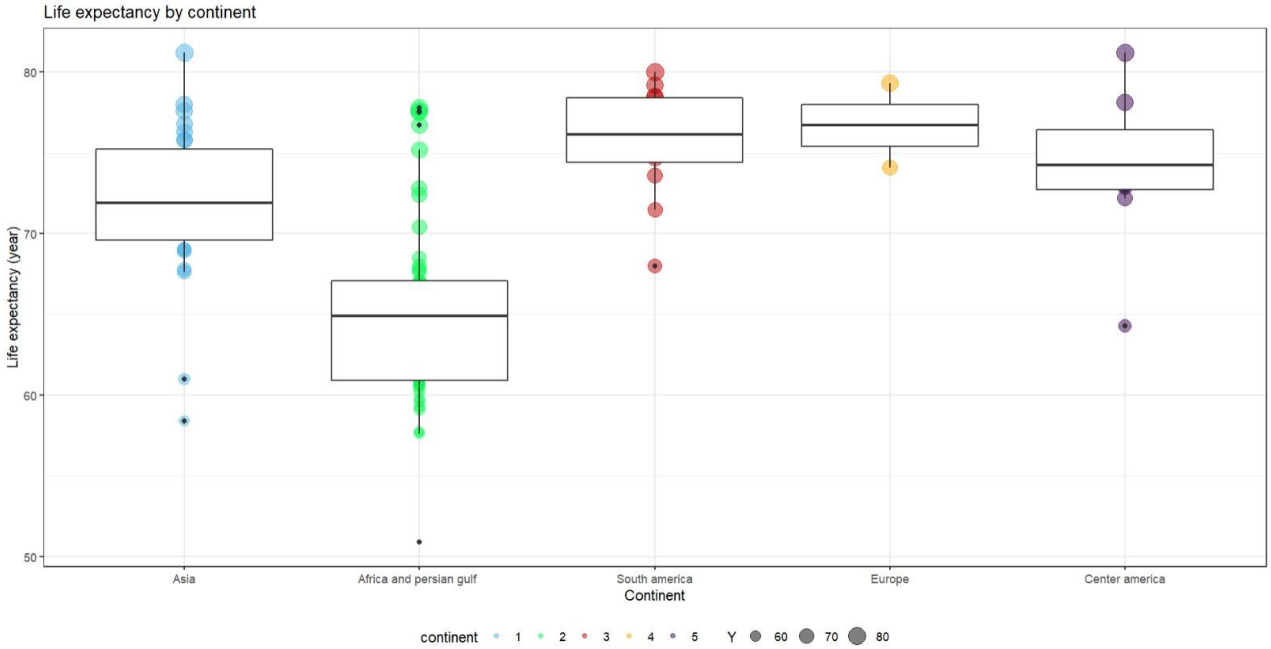
8. ייצוג קשרים בעזרת תרשימים:



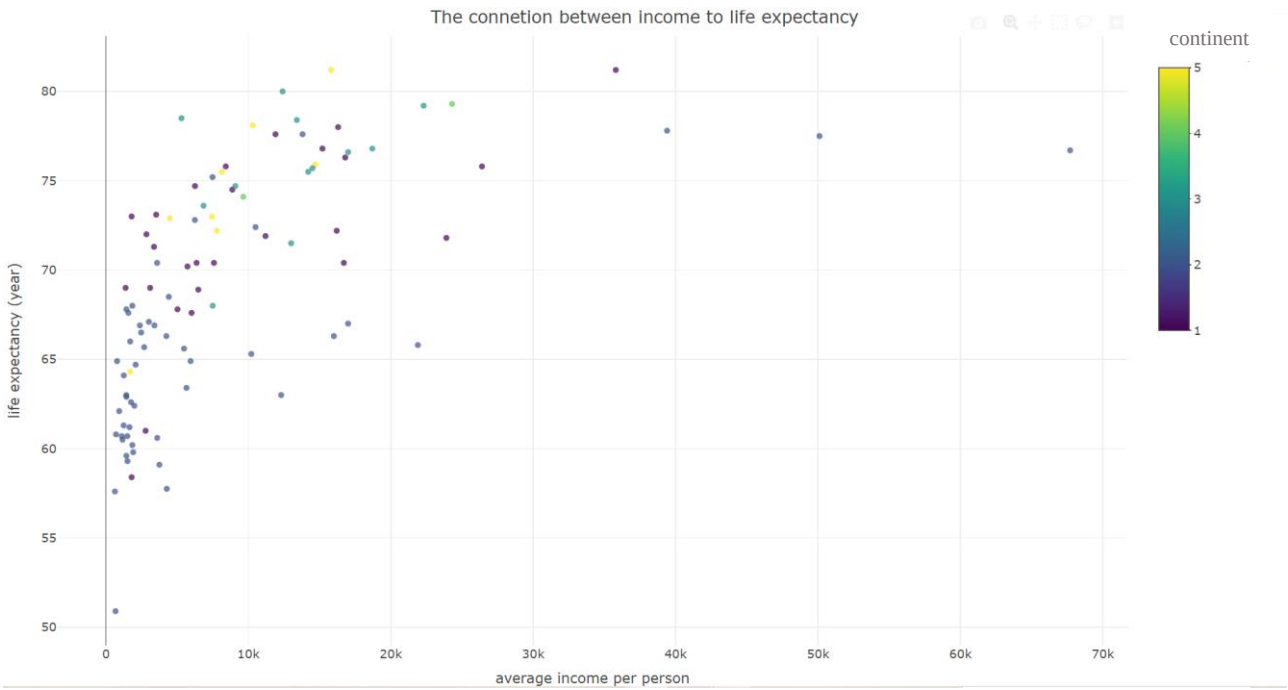
בגרף זה ניתן לראות את הקשר בין צריכת אלכוהול, צריכת סיגריות, לבין תוחלת החיים במדינה. כפי שכבר בדקנו, ישנה קורלציה חיובית מזערית בין המשתנים המסבירים (צריכת סיגריות וצריכת אלכוהול). היינו מצפים שבמידה וישנו קשר בין צריכת אלכוהול וסיגריות לתוחלת חיים נמוכה, נקבל בחלקו הימני העליון של הגרף עיגולים קטנים ביחס ליתר הגרף המייצגים מדינות עם תוחלת חיים נמוכה. על בסיס הנתונים הקיימים אין קשר בין משתנים אלו, ולכן ניתן להסיק שאינם משפיעים אחד על השני. ולכן, בניגוד למצופה לא ניתן לקבוע כי צריכת אלכוהול וסיגריות רבה מקטינה את תוחלת החיים במדינה.



בתרשים מוצגת חלוקת המדינות (התצפיות) בין היבשות השונות עפ״י כמות החולים במלריה. ניתן לראות כי רוב המדינות בהן יש מספר רב של חולים במלריה נמצאות ביבשת אפריקה או המפרץ הפרסי (2), בעוד שביבשות דרום אמריקה (3), אירופה (4) ומרכז אמריקה (5) מספר מצומצם של מדינות עם חולים. לכן ניתן להסיק כי משתנה הקטגוריאלי רלוונטי ומשמעותי בעת הסתכלות על מספר החולים במלריה.



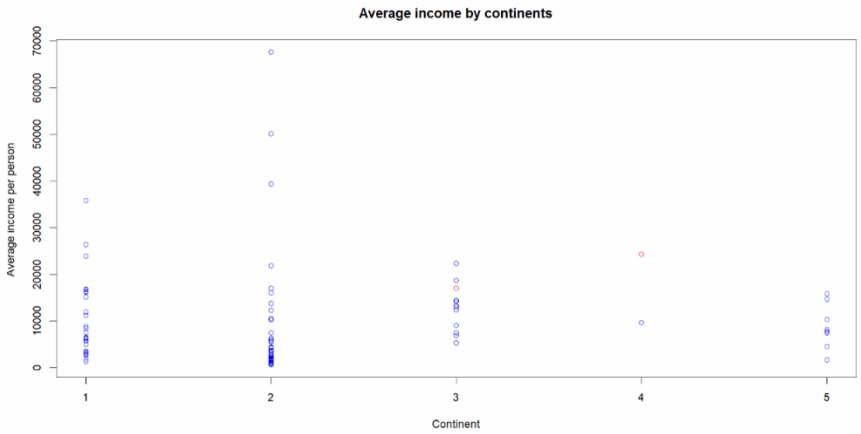
בתרשים ניתן לראות את תוחלת החיים עפ״י חלוקה ליבשות. בעזרת תצוגה מסוג זה ניתן להשוות בין תוחלות החיים בין כל אחת מהיבשות ולהבין את פיזור הנתונים בכל יבשת. כפי שניתן לראות תוחלת החיים הגבוהה ביותר היא ביבשת באירופה (כ-75) בעוד שתוחלת החיים הנמוכה ביותר הינה באפריקה והמפרץ הפרסי והיא (כ-65). ניתן לראות את פיזור הנתונים בכל יבשת עפ״י אורך הטווח הבין רבעוני, וכי ביבשת אפריקה הפיזור הוא הגדול ביותר, כלומר ישנם הבדלים משמעותיים בין המדינות השונות שביבשת.



התרשים מתאר את הקשר בין ההכנסה הממוצעת לבין תוחלת החיים. כל נקודה המייצגת מדינה צבועה בצבע על פי היבשת אליה היא שייכת. כפי ששיערנו ניתן לראות שככל ההכנסה הממוצעת לאדם עולה כך גם תוחלת החיים גדלה. בנוסף, ניתן לראות שבמרבית מדינות אפריקה ומפרץ הפרסי תוחלת החיים זהה וסביבות 60-65 שנים, והיא הנמוכה ביותר ביחס לשאר היבשות.

```
> V_Asia_income <- var(Asia_income)
> paste(V_Asia_income)
[1] "72433161.3333333"
> V_Africa_income <- var(Africa_income)
> paste(V_Africa_income)
[1] "163220528.577959"
> V_Samerica_income <- var(Samerica_income)
> paste(V_Samerica_income)
[1] "25614996.969697"
> V_Europe_income <- var(Europe_income)
> paste(V_Europe_income)
[1] "107457800"
> V_Camerica_income <- var(Camerica_income)
> paste(V_Camerica_income)
[1] "22663221.4285714"

> A_Asia_income <- mean(Asia_income$X4)
> paste(A_Asia_income)
[1] "10062.4444444444"
> A_Africa_income <- mean(Africa_income$X4)
> paste(A_Africa_income)
[1] "7220.56"
> A_Samerica_income <- mean(Samerica_income$X4)
> paste(A_Samerica_income)
[1] "12851.6666666667"
> A_Europe_income <- mean(Europe_income$X4)
> paste(A_Europe_income)
[1] "16970"
> A_Camerica_income <- mean(Camerica_income$X4)
> paste(A_Camerica_income)
[1] "8792.5"
```



בתרשים מוצגת החלוקה של המדינות ליבשות השונות עפ״י ההכנסה השנתית הממוצעת לאדם. ישנו שימוש בשני צבעים (בהתאם למשתנה X9): האדום - מדינות החברות בארגון ה-OECD, כחול - מדינות שאינן חברות בארגון. כך נוכל לזהות היכן ממוקמות מדינות החברות בארגון, ביחס לשאר המדינות. בתרשים ניתן לראות את הפיזור של הנתונים בכל אחת מהיבשות ולהסיק על פערי ההכנסות בין המדינות באותה היבשת, וכתוצאה מכך גם על פערים כלכליים-חברתיים ביניהן.

באפריקה והמפרץ הפרסי (2) השונות היא הגבוהה ביותר (163,220,528.577), כלומר ישנם הבדלים משמעותיים בין המדינות. ייתכן והשונות הגבוהה נובעת מאיחוד המפרץ הפרסי יחד עם אפריקה שהינם בעלי הבדלים כלכליים משמעותיים. בנוסף, היא יבשות זו היא הענייה ביותר (ממוצע ההכנסות ביבשת הוא 7220.56). מנגד, במרכז אמריקה (5) השונות היא הנמוכה ביותר (22,663,221.428), כלומר הפרשי ההכנסה הממוצעת בין המדינות הם הקטנים ביותר. קיימות שתי מדינות החברות בארגון ה- OECD וניתן לראות שכל אחת מהם מדורגת במיקום גבוה ביחס ליבשת שלה.

9. טבלאות שכיחות:

- <u>טבלאות שכיחות חד מימדיות:</u>

	continent	count	precent
1	1	27	0.27272727
2	2	50	0.50505051
3	3	12	0.12121212
4	4	2	0.02020202
5	5	8	0.08080808

- בטבלת השכיחות ניתן לראות את החלוקה של המדינות שבמדגם עפ"י יבשות. עפ"י התוצאות ניתן לראות הפרשים גדולים בין שכיחות היבשות, כלומר אין חלוקה אחידה של תצפיות המדגם עבור כל אחת מהיבשות. בקרב הנתונים היבשת השכיחה ביותר הינה אפריקה והמפרץ הפרסי (50.5%) ובעוד שהיבשת שהכי פחות שכיחה במדגם היא יבשת אירופה (2.02%).

	OECD	count	precent
1	0	97	0.97979798
2	1	2	0.02020202

- בטבלת השכיחות ניתן לראות את החלוקה של המדינות שבמדגם עפ"י שייכותם לארגון ה-OECD. עפ"י טבלת השכיחות רק 2.02% מהמדינות שנבדקו חברות בארגון זה, ובעקבות כך יהיה קשה לראות את ההשפעה של משתנה זה על משתנים אחרים.

- <u>טבלאות שכיחות דו מימדיות:</u>

- בטבלה שכיחות זו העמודות מייצגות את אחוז המעשנים והשורות מייצגות את היבשות. ניתן לראות מהטבלה כי התוצאה השכיחה ביותר במדגם הן מדינות ביבשת אפריקה והמפרץ הפרסי שאחוז המעשנים בהם הוא בין 10-20. כמו כן, טווחים רבים בטבלת השכיחות קיבלו את הערך אפס, ומכך ניתן להסיק שפיזור הנתונים הוא אינו רחב.

	(0,10]	(10,20]	(20,30]	(30,40]	(40,50]	(50,60]	(60,70]	(70,80]	(80,90]	(90,100]
(0,1]	0.01010101	0.07070707	0.15151515	0.04040404	0	0	0	0	0	0
(1,2]	0.12121212	0.26262626	0.09090909	0.03030303	0	0	0	0	0	0
(2,3]	0.01010101	0.08080808	0.03030303	0.00000000	0	0	0	0	0	0
(3,4]	0.00000000	0.00000000	0.01010101	0.01010101	0	0	0	0	0	0
(4,5]	0.02020202	0.05050505	0.00000000	0.01010101	0	0	0	0	0	0

- בטבלת שכיחות זו העמודות מייצגות את אחוז זיהום האוויר והשורות מייצגות את היבשות. ניתן לראות מהטבלה כי התוצאה השכיחה ביותר במדגם הן המדינות ביבשת אפריקה והמפרץ הפרסי בהן אחוז זיהום האויר הוא בין 20-30. עם זאת, קיימות שכיחויות נוספות הקרובות לשכיחות המקסימלית.

	(0,10]	(10,20]	(20,30]	(30,40]	(40,50]	(50,60]	(60,70]	(70,80]	(80,90]	(90,100]
(0,1]	0	0.08080808	0.08080808	0.04040404	0.02020202	0.03030303	0.01010101	0.00000000	0.00000000	0.01010101
(1,2]	0	0.02020202	0.15151515	0.12121212	0.12121212	0.03030303	0.01010101	0.02020202	0.02020202	0.01010101
(2,3]	0	0.07070707	0.05050505	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
(3,4]	0	0.00000000	0.01010101	0.00000000	0.01010101	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
(4,5]	0	0.04040404	0.04040404	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000

10. נספחים:

- טעינת ספריות:

```
library(rlang)
library(MASS)
library(fitdistrplus)
library(magrittr)
library(dplyr)
library(lazyeval)
library(parallel)
library(e1071)
library(plotly)
library(ggplot2)
library(triangle)
library(sqldf)
#library(readxl)
 #library(knitr)
 #library(rmarkdown)
library(simmer)
library(simmer.plot)
#install.packages("viridis")
library(viridis)
#install.packages("tidyverse")
library(tidyverse)
#install.packages("hrbrthemes")
library(hrbrthemes)
#install.packages("maps")
#library(maps)
```

- טעינת בסיס הנתונים:

```
filePath=choose.files ()
ImportData <-read.csv(filePath,header=TRUE)
Dataset <- na.omit(ImportData)
```

- קוד סעיף 4 - תיאור קשרים בין משתנים:

```
4 ##
subdata <- sqldf("select X1, X2, X3, X4, X5, X6, X7
            from Dataset(")
plot (subdata)

_4 #A
  plot (x = subdata$X2, y= subdata$X3,xlab = "HIV infected", ylab = "Malaria
infected" )
  fit <- lm(subdata$X3 ~ subdata$X2, data = subdata)
  abline(fit)
  cor(subdata$X2, subdata$X3) %>% print()
_4 #B
```

```
 plot (x = subdata$X6, y= subdata$X3,xlab = "Density population", ylab = "Malaria
infected" )
 fit <- lm(subdata$X3 ~ subdata$X6, data = subdata)
 abline(fit)
 cor(subdata$X6, subdata$X3) %>% print()
_4 #C
 plot (x = subdata$X5, y= subdata$X7,xlab = "Alcohol consumption", ylab =
"Cigarette consumption" )
 fit <- lm(subdata$X7 ~ subdata$X5, data = subdata)
 abline(fit)
 cor(subdata$X5, subdata$X7) %>% print()
_4 #D
 plot (x = subdata$X6, y= subdata$X1,xlab = "Density population", ylab = "Outdoor
air pollution" )
 fit <- lm(subdata$X1 ~ subdata$X6, data = subdata)
 abline(fit)
 cor(subdata$X6, subdata$X1) %>% print()
_4 #E
 plot (x = subdata$X4, y= subdata$X2,xlab = "Average income per person", ylab =
"HIV infected" )
 fit <- lm(subdata$X2 ~ subdata$X4, data = subdata)
 abline(fit)
 cor(subdata$X4, subdata$X2) %>% print()
```

- קוד סעיף 5 – ניתוח תיאורי של המשתנים:

```
5 ##

 #X1
summary(Dataset$X1)
sd(Dataset$X1)
skewness(Dataset$X1)

 #X2
summary(Dataset$X2)
sd(Dataset$X2)
skewness(Dataset$X2)

 #X3
summary(Dataset$X3)
sd(Dataset$X3)
skewness(Dataset$X3)

 #X4
summary(Dataset$X4)
sd(Dataset$X4)
skewness(Dataset$X4)

 #X5
summary(Dataset$X5)
sd(Dataset$X5)
skewness(Dataset$X5)
```

```
#X6
summary(Dataset$X6)
sd(Dataset$X6)
skewness(Dataset$X6)

#X7
summary(Dataset$X7)
sd(Dataset$X7)
skewness(Dataset$X7)

#X8
summary(Dataset$X8)

#X9
summary(Dataset$X9)
```

- קוד סעיף 6 - ניתוח חריגים:

```
6 ##

bp1 <- boxplot(Dataset$X1, main = "Outdoor air pollution")

DatasetNew <- sqldf("select*
          from Dataset
           where X1 < 98("
bp1New <- boxplot(DatasetNew$X1, main = "New outdoor air pollution")

bp2 <- boxplot(DatasetNew$X2, main = "HIV - Estimated number of people that
have been infected")
bp3 <- boxplot(DatasetNew$X3, main = "Malaria - Estimated number of people that
have been infected")
bp4 <- boxplot(DatasetNew$X4, main = "Average income per person")
bp5 <- boxplot(DatasetNew$X5, main = "Alcohol consumption per person")
bp6 <- boxplot(DatasetNew$X6, main = "Density per square")
bp7 <- boxplot(DatasetNew$X7, main = "Cigarette consumption")
```

- קוד סעיף 7 – פונקציית צפיפות והתפלגות מצטברת:

```
7 ##
hist(DatasetNew$X1,prob=TRUE,ylim = c(0,0.04) , main="PDF - Outdoor air
pollution",xlab = "Percent of air pollution" ,col="grey")
lines(density(DatasetNew$X1),col="blue",lwd=2)
plot.ecdf(DatasetNew$X1, main="CDF - Outdoor air pollution",xlab="Percent of air
pollution", lwd=7)

hist(DatasetNew$X6,xlim = c(0,1400),ylim = c(0,0.01), prob=TRUE, main="PDF -
Population density",xlab = "People per square km", breaks = 20,col="grey")
lines(density(DatasetNew$X6),col="blue",lwd=2)
plot.ecdf(DatasetNew$X6, main="CDF - Population density",xlab="People per
square km", lwd=7)
```

```
hist(DatasetNew$X7,prob=TRUE,ylim = c(0,0.06), main="PDF - Cigarette
consumption",xlab = "Percent of People" ,col="grey")
lines(density(DatasetNew$X7),col="blue",lwd=2)
plot.ecdf(DatasetNew$X7, main="CDF - Cigarette consumption",xlab="Percent of
People", lwd=7)
```

- קוד סעיף 8 - ניתוח קשרים בעזרת תרשימים:

```
8 ##
_8 #A
  plot(x=DatasetNew$X8,y=DatasetNew$X3,xlab
"=      Continent",ylab="Malaria infected", main ="Malaria infected by
continents( "
_#8B
  DatasetNew$X9 <-ifelse(DatasetNew$X9==0,"blue","red")
  plot(x=DatasetNew$X8,y=DatasetNew$X4,xlab
"=      Continent",ylab="Average income per person ", main ="Average income by
continents", col = DatasetNew$X9(

  DatasetNew$X9 <-ifelse(DatasetNew$X9=="blue","0","1")

  Asia_income  <- sqldf ("select X4
            from DatasetNew
            where X8 == 1( "
  Africa_income  <- sqldf ("select X4
            from DatasetNew
            where X8 == 2( "
  Samerica_income  <- sqldf ("select X4
            from DatasetNew
            where X8 == 3( "
  Europe_income  <- sqldf ("select X4
            from DatasetNew
            where X8 == 4( "
  Camerica_income  <- sqldf ("select X4
            from DatasetNew
            where X8 == 5( "

  V_Asia_income <- var(Asia_income)
  paste(V_Asia_income)
  V_Africa_income <- var(Africa_income)
  paste(V_Africa_income)
  V_Samerica_income <- var(Samerica_income)
  paste(V_Samerica_income)
  V_Europe_income <- var(Europe_income)
  paste(V_Europe_income)
  V_Camerica_income <- var(Camerica_income)
  paste(V_Camerica_income)


  A_Asia_income <- mean(Asia_income$X4)
  paste(A_Asia_income)
  A_Africa_income <- mean(Africa_income$X4)
```

```
  paste(A_Africa_income)
  A_Samerica_income <- mean(Samerica_income$X4)
  paste(A_Samerica_income)
  A_Europe_income <- mean(Europe_income$X4)
  paste(A_Europe_income)
  A_Camerica_income <- mean(Camerica_income$X4)
  paste(A_Camerica_income)
_#8C
  bubbleplot <- plot_ly(DatasetNew, x = ~DatasetNew$X4, y = ~DatasetNew$Y,
                  color =~DatasetNew$X8 ,
                  marker=
                   list(opacity = 0.7,
                       sizemode = "diameter(("
  bubbleplot <- bubbleplot%>%layout(title = 'The connetion between income to life
expectancy,'
                          xaxis = list(title = 'average income per person'),yaxis = list(title
= 'life expectancy (year)')(
  bubbleplot
  cor(DatasetNew$X4, DatasetNew$Y) %>% print()
_#8D
  ggplot(DatasetNew, aes(x = X5, y = X7))+
    ggtitle("The conection between alcohol, ciggarette to life expectancy")+
    xlab("Alcohol consumption")  + ylab("Cigarette consumption")+
#    scale_x_continuous(limits = c(0,60),breaks = seq(0,60,by=10))+
    geom_point(aes(color = as.factor(X8), size = Y), alpha = 0.5)+
    scale_color_manual(name = "continent" ,values = c("#4DB3E6", "#00FF50",
"#C00000", "#FFA500", "#37004D"))+
    scale_size(name = "Life expectancy", range = c(1, 13)) + # Adjust the range of
points size
    theme_set(theme_bw() +theme(legend.position = "bottom"))
_#8E
  ggplot(DatasetNew,   aes(x = factor(X8, levels = c ("1", "2", "3", "4", "5"), labels =
c("Asia", "Africa and persian gulf ", "South america", "Europe", "Center america")),
y = Y))+
     geom_point(aes(color = as.factor(X8), size = Y), alpha = 0.5)+
    scale_color_manual(name = "continent" ,values = c("#4DB3E6", "#00FF50",
"#C00000", "#FFA500", "#37004D"))+
    geom_boxplot+ ()
    labs(title = "Life expectancy by continent"  )+
    ylab("Life expectancy (year)") +
    xlab("Continent")

  data8 <- data.frame)
    continent=c(DatasetNew$X8),
    value=c(DatasetNew$Y)
(
```

- קוד סעיף 9 - טבלאות שכיחות:

```
9 ##
_#9A1
 table_continent<-sqldf("select X8 as continent, count(*) as count
```

```
        from DatasetNew
        group by X8("
table_continent["precent"]<- table_continent$count / sum(table_continent$count)
print (table_continent)
_#9A2
table_OECD<-sqldf("select X9 as OECD, count(*) as count
        from DatasetNew
        group by X9("
table_OECD["precent"]<- table_OECD$count / sum(table_OECD$count)
print (table_OECD)
_#9B1
PollutionByContinent <- cbind(Freq=table(cut(DatasetNew$X8,breaks =
            seq(0,5,1)),cut(DatasetNew$X1,breaks=seq(0,100,10((((
PollutionByContinent_Percent <- prop.table(PollutionByContinent)
_#9B2
CiggaretteByContinent <- cbind(Freq=table(cut(DatasetNew$X8,breaks =
                        seq(0,5,1)),cut(DatasetNew$X7,breaks=seq(0,100,10((((
CiggaretteByContinent_Percent <- prop.table(CiggaretteByContinent)
```