

# פרויקט ברגרסיה ליניארית



**פרטי המגישות:**

תום דמארי

מיה יערי



## תוכן עניינים

1.	בחירת מאגר נתונים	3
2.	יצירת טבלת משתנים	3
3.	תיאור המשתנים	4-5
4.	תיאור קשרים בין המשתנים	5-8
5.	ניתוח תיאורי של המשתנים	9-11
6.	ניתוח חריגים	11-13
7.	פונקציית צפיפות והתפלגות מצטברת	14-15
8.	ייצוג קשרים בעזרת תרשימים	15-16
9.	טבלאות שכיחות	16-17
10.	נספחים	18-20



חלק א' – בחירת בסיס נתונים וניתוח סטטיסטי של הנתונים

1. בחירת מאגר נתונים: Data Life

2. יצירת טבלת משתנים:

סוג המשתנה - מוסבר/מסביר	סימון במודל	יחידת מידה	סוג המשתנה – רצף / קטגוריאלי	הסבר קצר על המשתנה
משתנה מוסבר	$Y$	שנים	קטגוריאלי	תוחלת החיים במדינה
משתנה מסביר	$X_1$	%	רצף	מדד זיהום האוויר במדינה
	$X_2$	מספר	רצף	מספר האנשים החולים ב-HIV
	$X_3$	מספר	רצף	מספר האנשים החולים במלריה
	$X_4$	\$	רצף	הכנסה שנתית ממוצעת לאדם בדולרים
	$X_5$	ליטר לאדם פר שנה	רצף	צריכת אלכוהול שנתית לאדם בליטר
	$X_6$	$km^2$	רצף	צפיפות אוכלוסין ליחידת שטח
	$X_7$	%	רצף	אחוז צרכני הסיגריות
	$X_8$	-	קטגוריאלי	יבשת
	$X_9$	-	קטגוריאלי	שייכות לארגון ה-OECD



3. תיאור המשתנים:

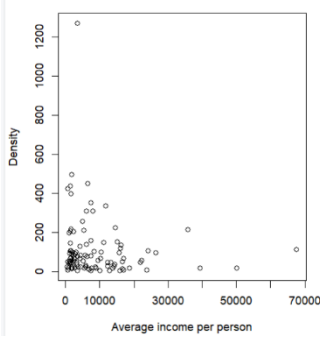
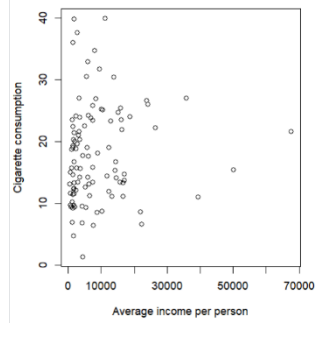
המשתנה המסביר	תיאור המשתנה
Outdoor air pollution (%) מדד זיהום האוויר במדינה	ידוע כי זיהום האוויר בעל השפעות מזיקות, בפרט בריאותיות בכך שיכול להיות גורם למחלות שונות. משתנה זה יכול להוסיף מידע לגבי מגמות השינוי במשתנה המוסבר על ידי מציאת קשר ישיר בין ירידה בתוחלת חיי האדם, המשתנה המוסבר שלנו, לבין עלייה במדד זיהום האוויר במדינה.
HIV - Estimated number of people that have been infected מספר האנשים החולים ב-HIV	ככל שכמות האנשים שנדבקו במדינה גבוהה יותר, כך נצפה לראות כי תוחלת החיים בה כנראה קטנה יותר. הרי ידוע כי מחלת ה-HIV ידועה כמסוכנת ועלולה אף להוביל למוות (מקצרת את תוחלת החיים).
Malaria - Estimated number of people that have been infected מספר האנשים החולים במלריה	ככל שכמות האנשים החולים במלריה במדינה גבוהה יותר, כך נצפה לראות כי תוחלת החיים בה תקטן. כידוע מחלת המלריה היא אחד מגורמי המוות העיקריים בעולם (מקצרת את תוחלת החיים).
Average income per person (\$) הכנסה שנתית ממוצעת לאדם בדולרים	ככל שסכום ההכנסה הממוצע לאדם קטן יותר נצפה שמצבו הבריאותי יהיה פחות טוב. סכום ההכנסה הממוצע של אדם משפיע על איכות המזון שיצרוך, על שירותי הרפואה בהם יבחר להשתמש, על צריכת התרופות ועל תנאי המחיה שלו. כאשר לאדם אין את האמצעים הנדרשים לדאוג לעצמו בריאותית נצפה שמצבו הבריאותי יתדרדר וכיוצא בכך תוחלת החיים שלו תתקצר.
Alcohol consumption per person (liters, year) צריכת אלכוהול שנתית לאדם בליטר	ככל שצריכת האלכוהול השנתית לאדם בליטר גדלה, כך נצפה לראות כי תוחלת החיים בה תקטן. מרבית המחקרים בנושא מראים כי אלכוהול הוא גורם סיכון התמותה והתחלואה המוביל ברחבי העולם, וקשור לבערך עשירית ממקרי המוות בגילאי 15 עד 49. נתון זה בהחלט יכול להיות גורם המסביר את הירידה בתוחלת החיים במדינה כי הנזק שלו הוא לא רק בריאותי, אלא נפשי.
density per square (km) צפיפות אוכלוסין ליחידת שטח	ככל שצפיפות האוכלוסין ליחידת שטח גדולה יותר משמע שיותר אנשים גרים ביחידת שטח אחת. כשאנשים חיים בצפיפות גדולה, מקדם ההידבקות במחלות לרוב גבוהה יותר ולכן מחלות מתפשטות בקצב גבוהה. מחלות אלו מסוכנות ועלולות להוביל לקיצור תוחלת החיים.
Cigarette consumption (%) אחוז צרכני הסיגריות	עישון ידוע כמזיק לבריאות. הוא גורם למחלות רבות במערכת הנשימה, התקפי לב, מחלות של מחזור הדם, פגיעה בריאות ועוד. לכן נצפה שככל שאחוז המעשנים במדינה גבוהה יותר המצב הבריאותי של התושב הממוצע יהיה פחות טוב ולכן תוחלת החיים תהיה קצרה יותר.
Continent יבשת	ביבשות שונות רמת איכות החיים הממוצעת שונה. ניתן לראות כי במדינות אפריקה תנאי המחיה פחות טובים ובהתאמה תוחלת החיים בהן נמוכה יותר. לכן, אצפה כי ביבשת אפריקה תוחלת החיים תהיה קצרה יותר מיבשות אחרות. בנוסף, ביבשות שונות ישנם תנאים שונים למחייה, בין אם מדובר בתנאי אקלים, אסונות טבע, דמוגרפיה, זמינות חיים או במשך שעות האור במהלך היום שכולם משתנים ביניהם ולכן עשויים להשפיע על איכות חיי התושבים ולהוביל לירידה או לעלייה בתוחלת החיים במדינות באותה יבשת.
Member of OECD שייכות לארגון OECD	ארגון ה-OECD, הארגון לשיתוף פעולה ולפיתוח כלכלי הוא ארגון בינלאומי של המדינות המפותחות הדוגלות בעקרונות הדמוקרטיה הליברלית והשוק החופשי. מטרתו לגבש עקרונות משותפים לביצוע מדיניות כלכלית וחברתית ברמה



הלאומית וליצור שיתופי פעולה אל מול האתגרים הכלכליים, החברתיים והסביבתיים בעולם. בארגון זה היינו מצפים לראות מדינות מפותחות ומתקדמות. מאחר והארגון שם דגש על צמיחה כלכלית, צמצום האבטלה ואיכות הסביבה נצפה שבמדינות שנמצאות בארגון יהיו פחות מחלות ולכן תוחלת החיים של האזרחים תהיה ארוכה יותר.

#### 4. תיאור קשרים בין משתנים:

##### קשרים בין משתנים מסבירים שונים

הקשר בין המשתנים	PLOT	המנגנון שמקשר בין סיבה ותוצאה	המשתנים
ניתן לראות במדגם ריכוז גבוהה של אנשים עם הכנסה ממוצעת נמוכה שחיים בצפיפות אוכלוסין נמוכה יחסית. כלומר הכנסה נמוכה לא מעידה על צפיפות גבוהה. ערך הקורלציה הוא -0.118 ומעיד על כך שככל שההכנסה גדלה הצפיפות דווקא יורדת. זאת בסתירה להנחה שלנו ולכן נסיק כי מדובר בקשר מדגמי.		<b>הנחנו כי קיים קשר סיבתי בין הכנסה ממוצעת לאדם וצפיפות האוכלוסין בה חי. הנחנו כי אדם שההכנסה הממוצעת שלו יחסית נמוכה, יאלץ לחיות בתנאי צפיפות קשים יותר כי לא יוכל להרשות לעצמו כלכלית משהו אחר.</b>	Average income per person density per square
ניתן לראות כי רוב צורכי הסיגריות הם בעלי הכנסה ממוצעת נמוכה יחסית לאדם. עם זאת יש פיזור יחסית גבוהה של תצפיות באחוזי צריכה שונים כלומר הכנסה נמוכה לא בהכרח מעידה על אחוז צריכה גבוהה של סיגריות. יתכן שמדובר בצריכה נמוכה. בנוסף ערך הקורלציה עומד על 0.079 כלומר אין קשר בין המשתנים. לכן מהמדגם עולה כי הקשר הסיבתי אינו מתקיים. מאחר והממצאים סותרים את ההנחה שלנו נניח		<b>הנחנו כי קיים קשר סיבתי בין הכנסה ממוצעת לאדם ואחוז צריכת הסיגריות. הנחנו כי אדם שההכנסה הממוצעת שלו יחסית נמוכה ישתייך לאוכלוסייה מוחלשת, פחות ישמור על הבריאות שלו ויקפיד על הימנעות מחומרים ממכרים.</b>	Average income per person Cigarette consumption

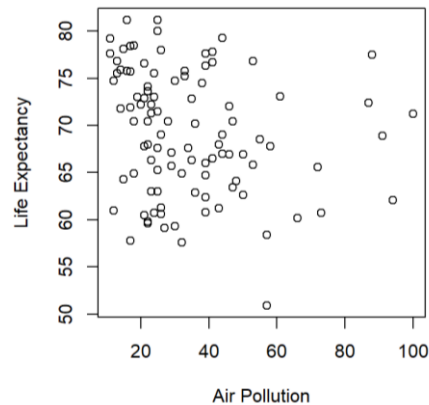


שהקורלציה מדגמית.			
<p>נראה כי ביבשת אפריקה (2) אכן ריכוז גבוהה של בעלי הכנסה ממוצעת נמוכה אך גם ביבשות אחרות המצב דומה. כלומר סוג היבשת לא מעיד בהכרח על הכנסה ממוצעת נמוכה. בנוסף ערך הקורלציה עומד על 0.045 ומעיד על כך שאין קשר סיבתי בין המשתנים. מאחר והממצאים סותרים את ההנחה שלנו נניח שהקשר מדגמי.</p>		<p><b>הנחנו כי קיים קשר סיבתי בין הכנסה ממוצעת ליבשת מאחר ויבשות מסוימות כמו אפריקה פחות מפותחות מיבשות אחרות. נצפה שרמת החיים בה תהיה נמוכה יותר ובהתאמה גם המצב הכלכלי של התושבים.</b></p>	<p>Average income per person</p>
			Continent
<p>נראה כי יש פיזור גדול בערכים. מהמדגם לא עולה כי צריכה גבוהה של אלכוהול מעידה על אחוז צריכה גבוהה של סיגריות. מהמדגם עולה שאין קשר סיבתי. בנוסף ערך הקורלציה עומד על 0.155, נמוך ומעיד על כך שאין קשר סיבתי מובהק.</p>		<p><b>הנחנו כי קיים קשר סיבתי בין צריכת אלכוהול ואחוז צריכה גבוהה של סיגריות מאחר והנחנו כי אנשים שנוטים לצרוך חומרים ממכרים הם אנשים שאינם מקפידים לשמור על אורח חיים בריא ולכן יותר סביר שיצרכו גם חומרים ממכרים אחרים.</b></p>	<p>Alcohol consumption per person</p>
			Cigarette consumption
<p>נראה כי היכן שיש צפיפות אוכלוסין נמוכה אנשים צורכים יותר סיגריות. עם זאת צפיפות אוכלוסין נמוכה לא מעידה בהכרח על אחוז גבוהה של צריכת סיגריות. בנוסף ערך הקורלציה עומד על -0.0004, כמעט אפס ומעיד כי אין קשר סיבתי מובהק. יש סתירה להנחה שלנו ולכן נניח כי מדובר בקורלציה מדגמית.</p>		<p><b>הנחנו כי קיים קשר סיבתי בין צפיפות אוכלוסין לאחוז צריכת הסיגריות מאחר והנחנו שהיכן שיש צפיפות אוכלוסין רמת החיים יותר נמוכה ולכן שם יתגוררו אוכלוסיות מוחלשות שפחות יקפידו על אורח חיים בריא.</b></p>	<p>Cigarette consumption</p>
			density per square

### הקשר בין משתנים מסבירים למשתנה המוסבר

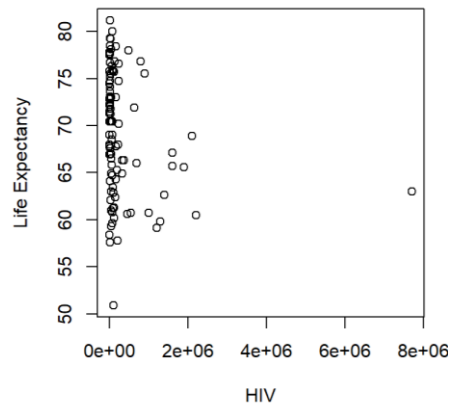
#### הקשר בין מדד זיהום האוויר לבין תוחלת החיים במדינה:

ניתן לראות ריכוז גבוהה יותר של תצפיות בהן אחוז זיהום האוויר נמוך ותוחלת החיים גבוהה. בנוסף ניתן לראות שהיכן שזיהום האוויר גבוהה תוחלת החיים נמוכה יותר. ניתן להניח קשר סיבתי כך שכלל שזיהום האוויר נמוך יותר תוחלת החיים ארוכה יותר. עם זאת, פיזור התוצאות יחסית גבוהה ולכן הקשר שנניח הינו קשר חלש (קשה להסיק באופן ברור מסקנות חותכות מהנתונים).



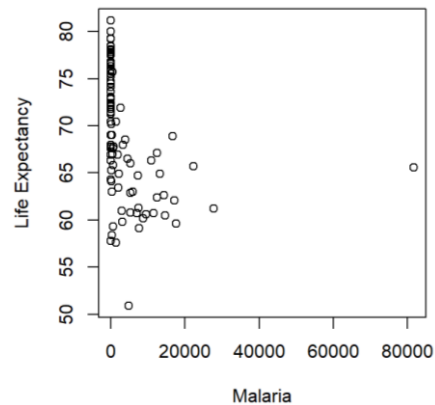
#### הקשר בין מספר האנשים החולים ב-HIV לבין תוחלת החיים במדינה:

ניתן לראות כי במדינות שאין בהן כמעט חולי HIV קיים פיזור בתוחלת החיים במדינות אלו. ניתן להסיק מכך כי אין גורם מסביר יחיד לשינוי בתוחלת החיים במדינה. אולם, לפי התוצאות שהתקבלו ניתן לראות כי במדינות בהן יש חולי HIV ניכר כי ככל שמספר החולים עולה במדינה כך תוחלת החיים בה יורדת.



#### הקשר בין מספר האנשים החולים במלריה לבין תוחלת החיים במדינה:

ניתן לראות כי במדינות שאין בהן חולי מלריה, ניתן להיווכח כי תוחלת החיים יחסית גבוהה ביחס לתוחלת החיים הנמוכה יותר של מדינות עם חולי מלריה. לעומת זאת, ניתן לראות כי אין קשר מובהק בין מדינות בעלות חולי מלריה בין מספר החולים לתוחלת החיים במדינה. ניתן להסיק מכך כי עבור מדינות בהן ישנם חולי מלריה איכות החיים פחותה ולכן נבדלות ממדינות ללא חולי מלריה כאשר במדינות אלה ישנם גורמים מסבירים נוספים לירידה בתוחלת החיים במדינה ולכן נניח קשר סיבתי יורד בין מספר החולים במלריה במדינה לבין תוחלת החיים בה.

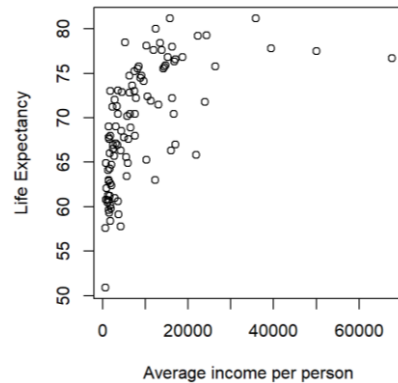






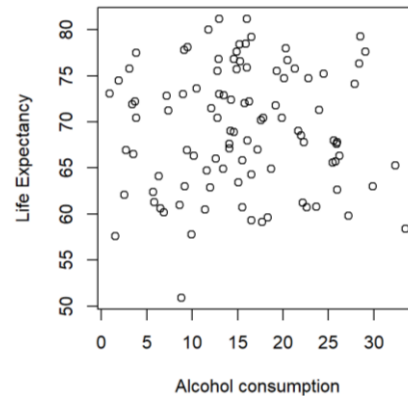
**הקשר בין הכנסה שנתית ממוצעת לאדם  
בדולרים לבין תוחלת החיים במדינה:**

ניתן לראות כי כאשר סכום ההכנסה לאדם נמוך מ-10000 אורך החיים שלו לרוב קצר יותר. בנוסף ככל שההכנסה קטנה נראה כי אורך החיים קטן עוד יותר בהתאמה. בנוסף ניתן לראות כי כאשר סכום ההכנסה גדול מ-10000 פיזור התוצאות יחסית גבוהה. ניתן להסיק כי עד רמת הכנסה מסוימת ישנו קשר סיבתי בין רמת ההכנסה לאורך החיים כך שככל שההכנסה יורדת כך גם תוחלת החיים. אולם מעל אותה רמת הכנסה לא ניתן להסיק קשר מובהק בין תוחלת החיים לסכום ההכנסה המסוים של אותו אדם.



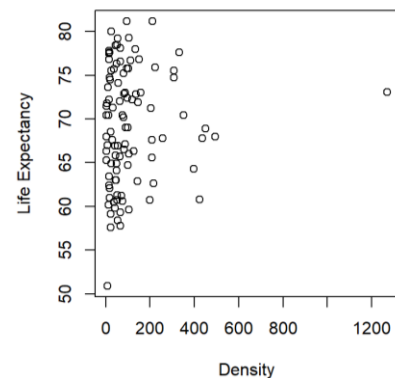
**הקשר בין צריכת אלכוהול שנתית לאדם בליטר לבין תוחלת החיים במדינה:**

ניתן לראות כי פיזור התוצאות בגרף יחסית גבוהה. כלומר שלא ניתן להניח מהתוצאות שום קשר מובהק בין צריכת האלכוהול לתוחלת החיים. מאחר ואלכוהול אכן מזיק לבריאות וצריכה מופרזת שלו אמורה להוריד את תוחלת החיים נניח כי התוצאות שקיבלנו מעידות קשר מדגמי בלבד שהרי תוצאות נכונות היו מעידות על קשר סיבתי.



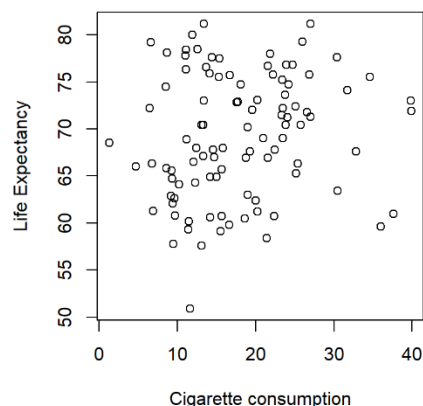
**הקשר בין צפיפות אוכלוסין ליחידת שטח לבין תוחלת החיים במדינה:**

ניתן לראות כי במדינות שאין בהן צפיפות האוכלוסין ליחידת שטח נמוכה מאוד קיים פיזור נרחב בתוחלת החיים. ניתן להסיק מכך כי אין גורם מסביר יחיד לשינוי בתוחלת החיים במדינה. אולם, לפי התוצאות שהתקבלו ניתן לראות כי מגמה של ירידה בתוחלת החיים ככל שהצפיפות אוכלוסין ליחידת שטח גדלה.



**הקשר בין אחוז צרכני הסיגריות לבין תוחלת החיים במדינה:**

ניתן לראות פיזור גבוהה בגרף. לא ניתן להניח מהתוצאות קשר מובהק בין צריכת סיגריות לתוחלת החיים של האדם. התוצאות מהוות סתירה להנחה שלנו כי צריכה גבוהה של סיגריות מהווה גורם לתוחלת חיים קצרה יותר. מאחר וידוע כי אלכוהול מזיק לבריאות ואכן מהווה גורם שכזה נניח כי התוצאות שקיבלנו מעידות על קשר מדגמי בלבד שהרי במציאות ישנו קשר סיבתי בין צריכת סיגריות לתחלואה ותוחלת חיים.







5. ניתוח תיאורי של המשתנים:

### משתנים רציפים

	Life Expectancy	Outdoor Air Pollution	HIV	Malaria	Average Income	Alcohol Consumption	Density	Cigarette Consumption
Min	50.90	11.00	200	0	627	0.90	3.08	1.30
1 <sup>st</sup> Qu.	64.25	21.75	14596	0	1895	9.80	23.80	12.01
Median	69.00	28.50	56000	33.5	5685	15.15	65.40	16.68
Mean	69.21	34.48	323178	3828.9	8936	15.43	111.66	18.04
3 <sup>rd</sup> Qu.	74.83	43.25	202500	4044.5	12550	20.70	121.00	23.45
Max.	81.20	100.00	7700000	81640	67700	33.50	1270.00	39.90
Sd	6.599589	19.38628	881656.9	9607.835	10525.01	7.730745	160.3229	7.981384
Skewness	-0.1858414	1.415481	6.274867	5.731823	2.848964	0.1507675	4.321838	0.6131017

טווחי התחום הבין רבעוני נמצאים בשורות הממוסגרות בסגול. הערך התחתון הוא הערך של השורה המסומנת בQ1 והערך העליון בQ3.

### פירוש תוצאות הניתוח:

**Life Expectancy** - נמצא כי למרות שהממוצע גדול מהחציון, מתקבלת אסימטריה שלילית נמוכה יחסית, כלומר נראה זנב שמאלי מתון יחסית. נסיק מכך שיש יותר מדינות בעולם תוחלת חיים כך שמספר השנים המוערך נמוך ממספר השנים הממוצע (69.21). מהתחום הבין רבעוני מתקבל כי עבור 50% מהמדינות תוחלת החיים בהן היא בין 64.25 ל-74.83 שנים.

**Outdoor Air Pollution** - נמצא כי הממוצע גבוה מהחציון, לכן בהתאם גם האסימטריה חיובית כלומר נראה זנב ימני. נסיק מכך שיש יותר מדינות שרמת זיהום האוויר בהן נמוכה מרמת זיהום האוויר הממוצעת (34.48%). מהתחום הבין רבעוני מתקבל כי עבור 50% מהמדינות רמת זיהום האוויר בהן היא בין 21.75% ל-43.25%.

**HIV** - נמצא כי הממוצע גבוה מהחציון, לכן בהתאם גם האסימטריה חיובית כלומר נראה זנב ימני. נסיק מכך שיש יותר מדינות שמספר האנשים החולים ב-HIV בהן נמוך ממספר האנשים החולים ב-HIV הממוצע (323178). מהתחום הבין רבעוני מתקבל כי עבור 50% מהמדינות מספר החולים ב-HIV בהן הוא בין 14596 ל-202500.

**Malaria** - נמצא כי הממוצע גבוה מהחציון, לכן בהתאם גם האסימטריה חיובית כלומר נראה זנב ימני. נסיק מכך שיש יותר מדינות שמספר האנשים החולים במלריה בהן נמוך ממספר האנשים החולים במלריה הממוצע (3828.9). מהתחום הבין רבעוני מתקבל כי עבור 50% מהמדינות מספר החולים במלריה בהן הוא בין 0 ל-4044.5.

**Average Income** - נמצא כי הממוצע גבוה מהחציון, לכן בהתאם גם האסימטריה חיובית כלומר נראה זנב ימני. נסיק מכך שיש יותר מדינות שממוצע ההכנסה השנתית לאדם שלהן נמוך מערך הממוצע של ממוצעי ההכנסה השנתית לאדם (8936). מהתחום הבין רבעוני מתקבל כי עבור 50% מהמדינות מספר החולים במלריה בהן הוא בין 1895 ל-12550.

**Alcohol Consumption** - נמצא כי הממוצע גבוה מהחציון, לכן בהתאם גם האסימטריה חיובית כלומר נראה זנב ימני. נסיק מכך שיש יותר מדינות שרמת צריכת האלכוהול בהן נמוכה מרמת צריכת האלכוהול הממוצעת (15.43%). מהתחום הבין רבעוני מתקבל כי עבור 50% מהמדינות רמת זיהום האוויר בהן היא בין 9.80% ל-20.70%.

**Density** - נמצא כי הממוצע גבוה מהחציון, לכן בהתאם גם האסימטריה חיובית כלומר נראה זנב ימני. נסיק מכך שיש יותר מדינות שצפיפות האוכלוסין ליחידת שטח בהן נמוך מצפיפות



האוכלוסין ליחידת שטח הממוצעת (111.66). מהתחום הבין רבעוני מתקבל כי עבור 50% מהמדינות מספר החולים HIV בהן הוא בין 23.80 ל-121.00.

Cigarette Consumption - נמצא כי הממוצע גבוה מהחציון, לכן בהתאם גם האסימטריה חיובית כלומר נראה זנב ימני. נסיק מכך שיש יותר מדינות שרמת צריכת הסיגריות בהן נמוכה מרמת צריכת הסיגריות הממוצעת (18.04%). מהתחום הבין רבעוני מתקבל כי עבור 50% מהמדינות רמת זיהום האוויר בהן היא בין 12.01% ל-23.45%.

### משתנים קטגוריאליים ללא חשיבות לסדר

#### היבשת

שכיחות קטגוריית המשתנה –

1	2	3	4	5
0.28	0.50	0.12	0.02	0.08

חלוקה פנימית –

מספר תצפיות מכל קטגוריה (100 מדינות במאגר): 28 מדינות מאסיה, 50 מדינות מאפריקה, 12 מדינות מאמריקה הדרומית, 2 מדינות מאירופה ו-8 מדינות מאמריקה הצפונית.

	אסיה (1)	אפריקה (2)	אמריקה הדרומית (3)	אירופה (4)	אמריקה הצפונית (5)
Mean	71.79643	65.1146	75.70833	76.7	74.1375
Sd	4.790846	5.597737	3.441315	3.676955	4.984243
Skewness	-0.7621529	0.5151854	-0.8808462	8.287509e-15	-0.6662611

Continent – במאגר המידע נתון מידע לגבי מדינות מ-5 יבשות שונות, כאשר מחצית מהמדינות הנתונות הן מיבשת אפריקה ולכן היבשת השכיחה הינה אפריקה. נמצא כי ביבשת השכיחה ביותר אפריקה, בעלת ממוצע תוחלת החיים הנמוך ביותר. בניגוד לכך, אירופה הינה היבשת הכי פחות שכיחה, כלומר בעלת מספר המדינות הנמוך ביותר מהמאגר. בעוד שבאפריקה ובאירופה קיימת אסימטריה חיובית, זנב חיובי ובהתאמה הממוצע גבוה מהחציון. נסיק כי יש יותר מדינות שתוחלת החיים בהן נמוכה מתוחלת החיים הממוצעת. בניגוד לכך, בשאר היבשות קיימת אסימטריה שלילית, זנב שמאלי ובהתאמה הממוצע נמוך מהחציון. נסיק מכך שיש יותר מדינות שתוחלת החיים בהן גבוהה מתוחלת החיים הממוצעת.

#### השכיחות לארגון ה-OECD

שכיחות קטגוריית המשתנה –

0	1
0.98	0.02

חלוקה פנימית –

מספר תצפיות מכל קטגוריה (מתוך 100 מדינות): 98 מדינות משתייכות לארגון OECD, 2 מדינות לא משתייכות לארגון OECD.

	לא שייך (0)	שייך (1)
Mean	69.03194	77.95
Sd	6.542781	1.909188
Skewness	-0.1653847	1.579345e-14

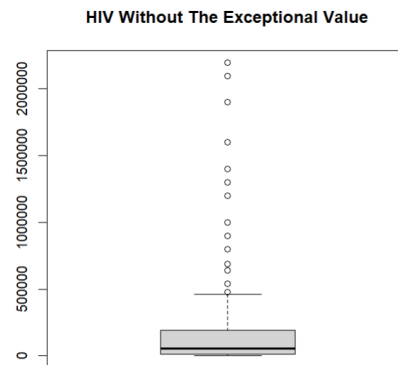
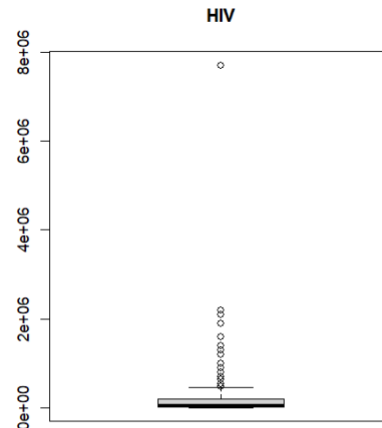
Member Of OECD – כמעט כל המדינות מתוך המאגר הנתון אינן משתייכות לארגון OECD, לכן הקטגוריה השכיחה הינה אי השתייכות לארגון OECD.

עבור המדינות אשר לא משתייכות לארגון OECD מתקבלת אסימטריה שלילית, זנה שמאלי ובהתאם לכך ניתן לומר כי התוחלת חיים בהן נמוכה מתוחלת החיים הממוצעת (69.21). לעומת זאת, עבור המדינות אשר משתייכות לארגון OECD מתקבלת אסימטריה חיובית, זנב ימני ולכן ניתן לומר כי תוחלת החיים של אותן המדינות גבוהה מתוחלת החיים הממוצעת הכוללת (69.21).

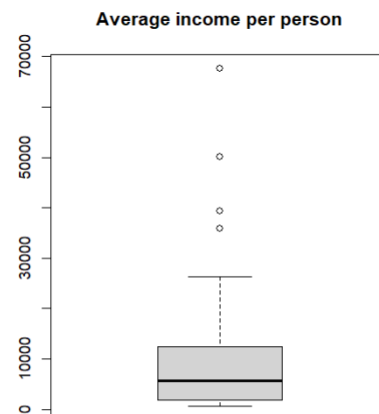
6. ניתוח חריגים:

#### ניתוח חריגים

ניתן לראות כי עבור המשתנה HIV קיבלנו כמות גדולה יחסית של חריגים, כלומר ישנן מדינות שכמות החולים בהן יצאו מחוץ לטווח הערכים. באחת המדינות החריגה בכמות החולים גדולה מאוד ומאחר והיא היחידה עם חריגה שכזו נרצה להוציא אותה מהתרשים. לעומת זאת שאר החריגות קרובות יחסית אחת לשנייה ולקופסא ולכן אותן נבחר להשאיר. בנוסף נרצה לתת ביטוי גם למדינות בהן כמות החולים גבוהה יותר על מנת לראות את ההשפעה שלהן על הסטטיסטיקות שלנו.

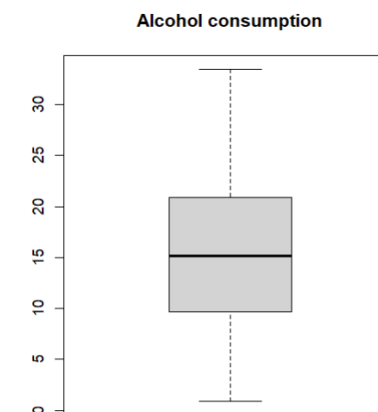
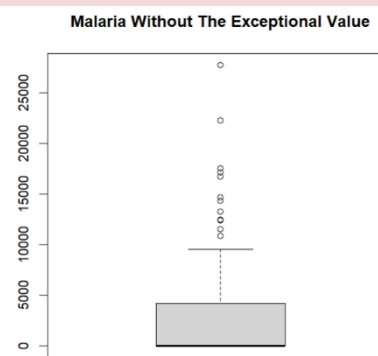
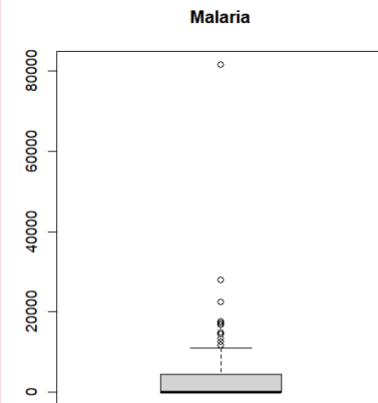


ניתן לראות כי עבור המשתנה הכנסה שנתית ממוצעת לאדם בדולרים קיבלנו 4 חריגים. נראה כי ברוב המדינות ההכנסה הממוצעת לאדם נשארת בטווח זהה אך יש 4 מדינות שבהן ההכנסה הממוצעת לאדם גבוהה יותר מהטווח. נרצה לתת ביטוי גם למדינות עם הכנסה ממוצעת גבוהה יותר מכיוון שיש לכך השפעה סטטיסטית על תוחלת החיים שאנחנו מחפשים לבדוק.



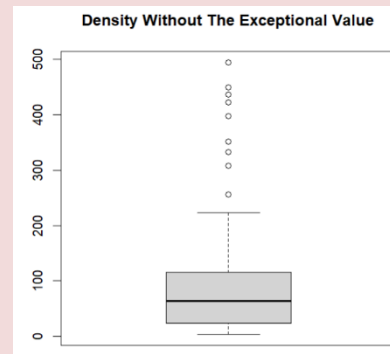
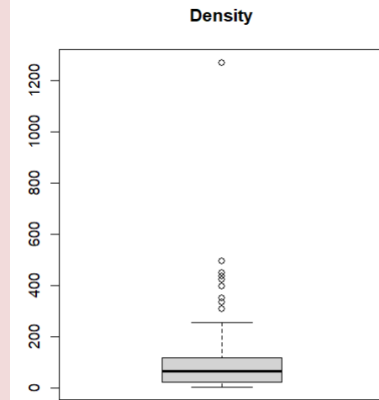
ניתן לראות כי עבור המשתנה Malaria קיבלנו מספר חריגים. כלומר ישנן מדינות שכמות החולים במחלה גדולה מטווח הערכים.

באחת המדינות החריגה גדולה במיוחד ומאחר והיא היחידה עם חריגה גדולה שכזו נרצה להוציא אותה מהתרשים. לעומת זאת שאר החריגות קרובות יחסית אחת לשנייה ולקופסא. לחריגות אלו יש משמעות סטטיסטית חשובה הרי נרצה לתת ביטוי גם למדינות בהן נתוני תחלואה גבוהים יותר. לכן נשאיר את אותן חריגות.



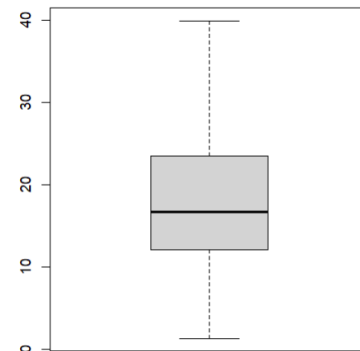
ניתן לראות כי עבור המשתנה של צריכת אלכוהול לא קיבלנו חריגים כלל. כלומר צריכת האלכוהול של כל המדינות יחסית דומה ונמצאת בטווח הבין רבעוני.

ניתן לראות כי עבור המשתנה Density קיבלנו מספר חריגים. כלומר ישנן מדינות שבהן צפיפות האוכלוסין ליחידת שטח גבוהה מהטווח. ניתן לראות כי במדינה אחת החריגה גדולה במיוחד. חריגה זו היא קיצונית ואינה מייצגת ולכן נעדיף להוציא אותה. שאר החריגות צמודות אחת לשנייה ומרוכזות קרוב יחסית לקופסא. נרצה לתת ביטוי למדינות בהן צפיפות האוכלוסין גבוהה מהטווח הרגיל מאחר ויש לכן משמעות בהסקה סטטיסטית שנסיק.



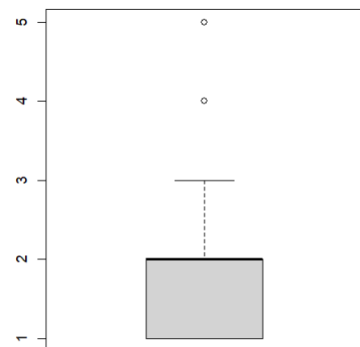
ניתן לראות כי עבור המשתנה של צריכת סיגריות לא קיבלנו חריגים כלל. כלומר צריכת הסיגריות של כל המדינות יחסית דומה ונמצאת בטווח הערכים.

Cigarette consumption



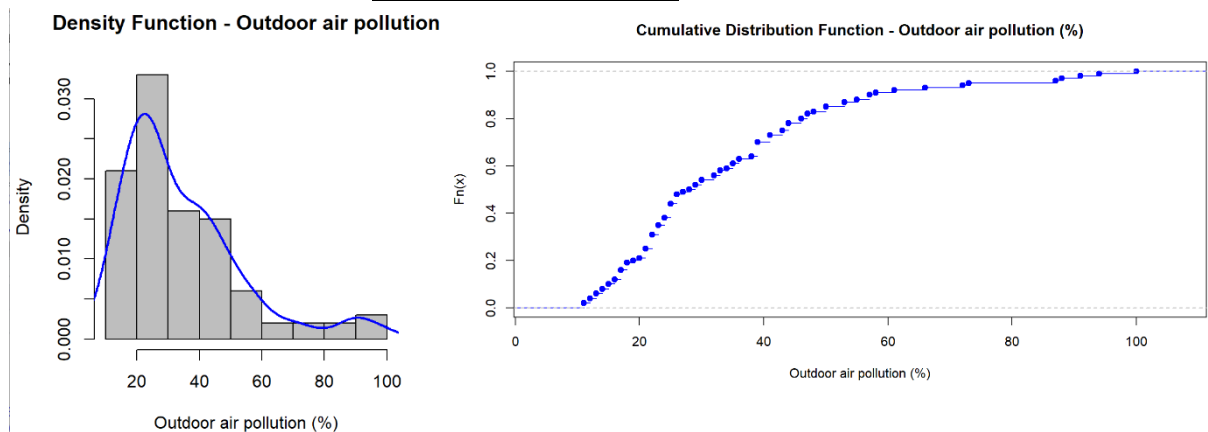
עבור המשתנה Continent ניתן לראות כי יש שתי חריגות. כלומר רוב המדינות משתייכות לאחת מיבשות 1,2,3. יש מדינה אחת בלבד שמשתייכת ליבשת 4 ומדינה אחת בלבד שמשתייכת ליבשת 5. נבחר להשאיר את החריגות כי נרצה מידע גם עבור יבשות 4 ו-5. חשוב לנו שנבצע הסקה שתסתמך על אוכלוסייה גדולה ומגוונת ככל האפשר. יש לכך משמעות על אמינות המידע שנקבל.

Continent



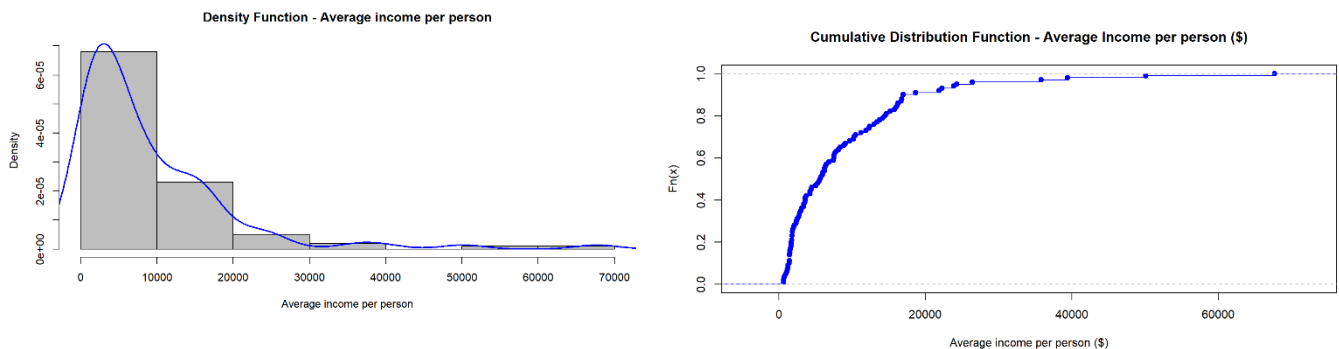
7. פונקציית צפיפות והתפלגות מצטברת:

**מדד זיהום האוויר במדינה**



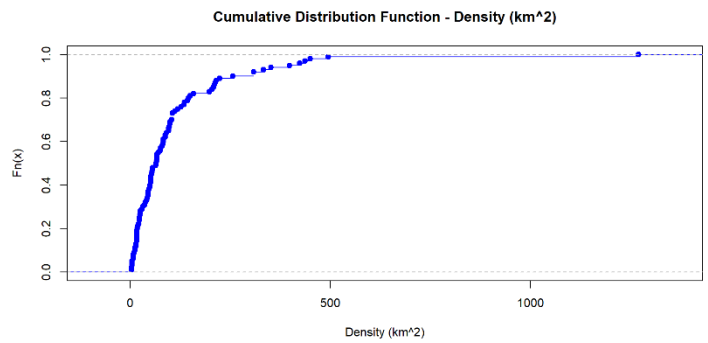
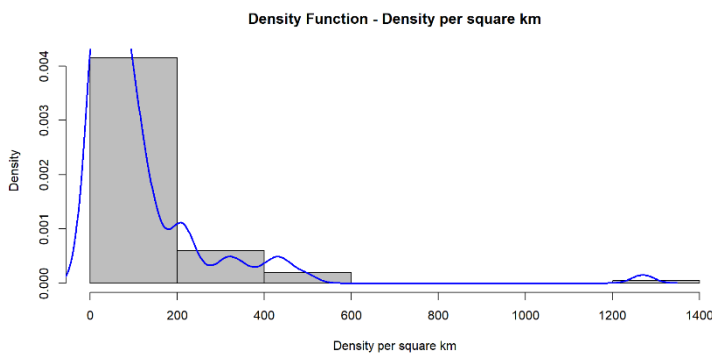
בפונקציית הצפיפות המצטברת ניתן לראות עלייה חדה יותר בטווח בין 0-20, בין 20-40 עלייה מתונה יותר והחל מ-40 העלייה עוד יותר מתונה. ניתן לראות שיש שאיפה להסתברות אפסית ככל שאחוז זיהום האוויר גדל. כלומר, הסיכוי למצוא מדינה מבין מאגר המדינות שקיבלנו, שיש לה יותר מ-90 אחוזים במדד זיהום האוויר הוא אפסי. דבר זה בא לידי ביטוי בזנב ימני בפונקציית הצפיפות שתואם את ערך האסימטריה החיובית הגבוהה יחסית שהתקבל מקודם ולעליות המתונות בפונקציה המצטברת. ישנה נקודת שיא בנקודה שבה אחוז מדד זיהום האוויר למדינה הוא עומד על בערך 25%. סטיית התקן היא קטנה (19.38628) ובהתאמה גם פיזור הנתונים הקטן יותר בגרף.

**הכנסה שנתית ממוצעת לאדם בדולרים**



בפונקציית הצפיפות המצטברת ניתן לראות עלייה חדה החל מ-0 עד בערך 20,000, ומשם ישנה עלייה מתונה יותר וצפיפות הנתונים דלה. ניתן לראות שיש שאיפה להסתברות אפסית ככל שממוצע השכר לאדם גדל. כלומר, הסיכוי למצוא מדינה מהמאגר שקיבלנו, שיש לה שכר ממוצע לאדם אשר גבוה מ-\$40,000 הוא אפסי. דבר זה בא לידי ביטוי בזנב ימני ארוך בפונקציית הצפיפות שתואם לאסימטריה החיובית הגבוהה שהתקבלה מקודם ולעלייה המתונה בפונקציה המצטברת. ישנה נקודת שיא שהיא מתארת את הנקודה שבה שיעור השכר הממוצע לאדם שווה בערך \$4,000. סטיית התקן הגבוהה (10525.01) באה לידי ביטוי בפיזור הערכים הגדול.

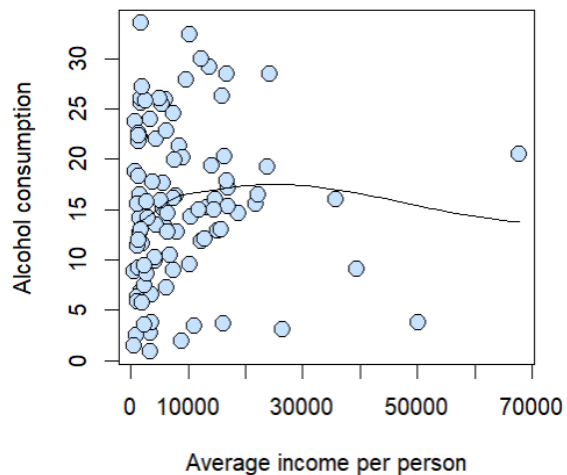
### צפיפות האוכלוסין ליחידת שטח



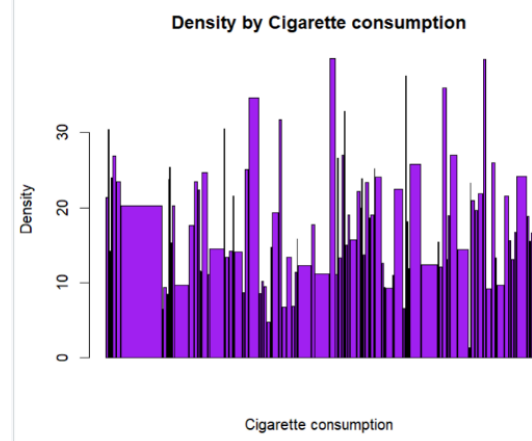
בפונקציית הצפיפות המצטברת ניתן לראות עלייה חדה החל מ-0 עד בערך 200, ומשם ישנה עלייה מתונה יותר וצפיפות הנתונים דלה. ניתן לראות שיש שאיפה להסתברות אפסית ככל שהצפיפות האוכלוסין ליחידת שטח גדלה. כלומר, הסיכוי למצוא מדינה מהמאגר שקיבלנו, אשר צפיפות האוכלוסין ליחידת שטח גבוהה מ-500 הוא אפסי. דבר זה בא לידי ביטוי בזנב ימני ארוך בפונקציית הצפיפות שתואם לאסימטריה החיובית הגבוהה שהתקבלה מקודם ולעלייה המתונה בפונקציה המצטברת. ישנה נקודת שיא שהיא מתארת את הנקודה שבה שיעור צפיפות האוכלוסין שווה בערך ל-50 קמ"ר. סטיית התקן הגבוהה יחסית (160.3229) באה לידי ביטוי בפיזור ערכים יחסית גדול.

התרשים הינו תרשים פיזור המתאר את הקשר בין ההכנסה הממוצעת של אדם לצריכת האלכוהול שלו. קודם כל ניתן לראות כי יש יותר קבוצות הנמצאות תחת רמות הכנסה נמוכות וממוצעות מאשר רמות הכנסה גבוהות. המשמעות היא שרוב האוכלוסייה לא במצב כלכלי מאוד גבוהה. אנחנו שיערנו כי רמות הכנסה נמוכות מובילות לצריכה גבוהה של אלכוהול בשל לחצים והתמודדויות. בפועל ניתן לראות כי עד סכום הכנסה מסוים (פחות או יותר 25000) ישנו קשר חיובי בין כמות ההכנסה לצריכת האלכוהול. נוכל להסיק כי אנשים ברמות הכנסה נמוכות יותר לא מוציאים את כספם על מותרות כמו אלכוהול גם אם נמצאים תחת לחצים כבדים. מעל אותו סכום הכנסה ניתן לראות מגמת ירידה. זה מתיישב עם ההשערה שלנו שברמות הכנסה גבוהות אנשים סובלים פחות מלחצים וצורכים פחות אלכוהול למרות שיש בידם את האמצעים הכלכליים לצרוך כמה שירצו.

### 8. ייצוג קשרים בעזרת תרשימים:

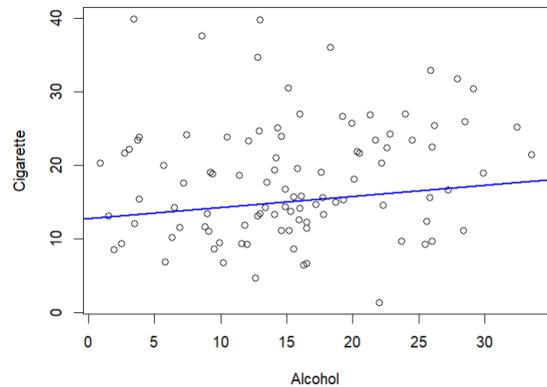


התרשים הינו תרשים בר המתאר את הקשר בין רמת הצפיפות לצריכת סיגריות. הוא מציג את צריכת הסיגריות עבור קבוצות שונות של רמות צפיפות שונות. קודם כל ניתן לראות כי קבוצות הנמצאות ברמת צפיפות מאוד גבוהה הן יותר נדירות וכי רוב הקבוצות נמצאות ברמת צפיפות ממוצעת. כפי שציפינו לראות אפשר לראות שאין קשר בין רמת הצפיפות לצריכת סיגריות שכן אפשר לראות צפיפות גבוהה בצריכת סיגריות נמוכה וההיפך.

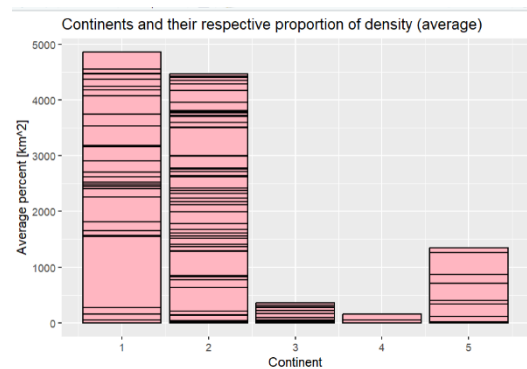




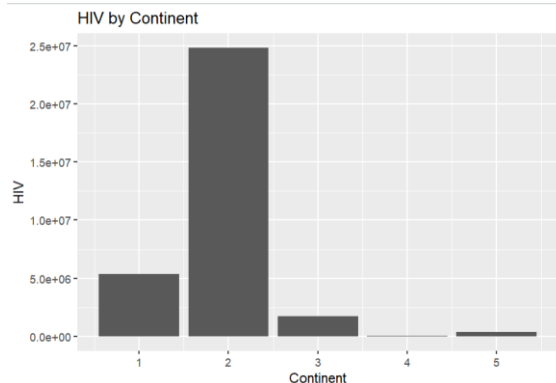
הגרף מתאר את הקשר בין צריכת אלכוהול לצריכת סיגריות. בגרף ניתן לראות כי רוב האוכלוסייה לא צורכת כמויות גבוהות מאוד של אלכוהול או סיגריות. בנוסף ניתן לראות כי כפי שחשדנו אכן יש קשר בין צריכת האלכוהול לצריכת הסיגריות. ככל שצריכת האלכוהול גבוהה יותר יש יותר סיכוי שאותו אדם יצרוך גם סיגריות. זה מתיישב עם ההשערה שלנו שאנשים עם התמכרות אחת הם אנשים שנוטים להתמכרויות בכללי. הם אינם שמים דגש על אורח חיים בריא ולכן בסבירות גבוהה יותר לסבול גם מהתמכרויות נוספות. ניתן לראות שעם זאת הקשר הליניארי אינו מאוד חזק.



הגרף מתאר את הקשר בין יבשת לרמת הצפיפות. בגרף ניתן לראות כי רמת הצפיפות גבוהה מאוד ביבשת אפריקה וביבשת אסיה ביחס לשאר היבשות. בנוסף ניתן לראות כי ביבשת אסיה הצפיפות גבוהה יותר מיבשת אפריקה. תוצאות אלו מפתיעות מאחר ולא צפינו לראות הבדל עד כי כך משמעותי בין יבשות אלו לשאר היבשות. בנוסף לא צפינו שרמת הצפיפות בין שתיהן תהיה עד כדי כך קרובה.



הגרף מתאר את הקשר בין יבשת לכמות החולים ב-HIV. ניתן לראות כי ביבשת אפריקה ישנה הכמות הגדולה ביותר של חולים ביחס לשאר היבשות. תוצאות אלה מתכנסות עם ההשערות שלנו שיבשת אפריקה הינו מקום פחות מפותח ומתקדם עם מודעות נמוכה יותר לגבי מחלות ונזקים בריאותיים ולכן גם שיעור גבוהה יותר של חולים במחלת ה-HIV.



9. טבלאות שכיחות:

בכל טבלה נציג מימין טבלת שכיחות הכוללת את מספר התצפיות (מדינות) בכל קטגוריה, משמאל נציג טבלת שכיחות יחסית עם האחוז היחסי של כל קטגוריה.

### טבלה ראשונה חז ממדית - ההכנסה השנתית הממוצעת לאדם

(0,1e+04]	(1e+04,2e+04]	(2e+04,3e+04]	(3e+04,4e+04]	(4e+04,5e+04]	(5e+04,6e+04]	(6e+04,7e+04]
68	23	5	2	0	1	1
0.68	0.23	0.05	0.02	0.00	0.01	0.01

בטבלת השכיחויות החד ממדית הזו הצגנו את ההכנסה השנתית הממוצעת לפי קפיצות של \$10,000. **תוצאות** – ניתן להסיק מן הטבלה כי ההכנסה השנתית הממוצעת לאדם במרבית המדינות (68%) נמוכה מ-\$10,000 לאדם. ניתן לשים לב כי אחוז המדינות שהכנסתן שנתית הממוצעת גבוהה מ-\$20,000 הינו נמוך מאוד (9%). ערך הכסף בכל מדינה שונה, ניתן להעריך כי ממוצע ההכנסה השנתי נמוך מ-\$10,000 במרבית המדינות בעולם.



### טבלה שנייה חד ממדית – רמת צריכת הסיגריות

(0,20]	(20,40]	(40,60]	(60,80]	(80,100]	(0,20]	(20,40]	(40,60]	(60,80]	(80,100]
0.62	0.38	0.00	0.00	0.00	62	38	0	0	0
(%)									

בטבלת השכיחות הזו ניתן לראות את ההתפלגות של המדינות בהתאם לצריכת הסיגריות שלהן, בקפיצות של 20% צריכה.

**תוצאות** – ניתן לראות כי מרבית המדינות (62%) עומדות על צריכה פחותה מ-20%, לעומת זאת 38% מהמדינות עומדות על צריכה בין 20% ל-40% של סיגריות. בעוד שאף מדינה לא עוברת צריכה של 40%. ניתן להסיק מכך כי לרוב האוכלוסייה בעולם יש כיום מודעות בריאותית בנושא נזקי העישון ולכן מרבית האוכלוסייה מקפידה לא לעשן.

### טבלה ראשונה דו ממדית - רמת זיהום האוויר ותוחלת חיי האדם

	(0,25]	(25,50]	(50,75]	(75,100]		(0,25]	(25,50]	(50,75]	(75,100]
(0,20]	0.00	0.00	0.10	0.11	(0,20]	0	0	10	11
(20,40]	0.00	0.00	0.40	0.09	(20,40]	0	0	40	9
(40,60]	0.00	0.00	0.17	0.04	(40,60]	0	0	17	4
(60,80]	0.00	0.00	0.04	0.00	(60,80]	0	0	4	0
(80,100]	0.00	0.00	0.04	0.01	(80,100]	0	0	4	1

בטבלת השכיחות הדו ממדית הראשונה ניתן לראות את רמת זיהום האוויר ותוחלת החיים בכל מדינה.

**תוצאות** – הנחנו הנחה כי ככל שרמת זיהום האוויר גדלה כך תוחלת החיים בה קטנה. ניתן לראות כי כאשר רמת זיהום האוויר עומדת על 0% ל-20%, תוחלת חיי האדם במדינה היא בערך 75 שנים. לעומת זאת, עבור רמת זיהום אוויר גבוהה יותר בטווח של 80%-100% תוחלת החיים היא בין 50-75 שנים. 40% מהמדינות רמת זיהום האוויר עומדת על 20%-40% ותוחלת החיים נמצאת בטווח של 50-75 שנים. ממצאים אלו מתכנסים עם השערתנו שאחוזים גבוהים של זיהום אוויר מובילים לירידה בתוחלת החיים.

### טבלה שנייה דו ממדית - יבשת ואחוזי צריכת

	(0,20]	(20,40]	(40,60]	(60,80]	(80,100]		(0,20]	(20,40]	(40,60]	(60,80]	(80,100]
1	0.19	0.09	0.00	0.00	0.00	1	19	9	0	0	0
2	0.34	0.16	0.00	0.00	0.00	2	34	16	0	0	0
3	0.11	0.01	0.00	0.00	0.00	3	11	1	0	0	0
4	0.00	0.02	0.00	0.00	0.00	4	0	2	0	0	0
5	0.08	0.00	0.00	0.00	0.00	5	8	0	0	0	0

בטבלת השכיחות הדו ממדית השנייה ניתן לראות עבור כל יבשת את ההתפלגות של צריכת האלכוהול.

**תוצאות** – אפשר להסיק על אופי צריכת האלכוהול ביבשות השונות מטבלאות השכיחות שלפנינו. ניתן לראות כי אין מדינות שעוברות את ה-40% צריכת אלכוהול. מרבית המדינות עומדות על אחוז צריכה בין 0% ל-20%. ניתן להיווכח כי באירופה (יבשת 4) אחוז צריכת האלכוהול גבוה יותר מהממוצע ועומד על 20%-40%. ניתן להסיק כי באירופה ישנם מודעות נמוכה יותר לנזקי האלכוהול ולכן גם הרגלי צריכה נפוצים יותר. לעומת זאת, באמריקה הצפונית (יבשת 5) אחוז צריכת האלכוהול נמוך יחסית ועומד על 0-20% ולכן ניתן להסיק כי קיימת מודעות בריאותית גבוהה ולכן צריכת אלכוהול נמוכה יותר. בשאר היבשות ישנה התפלגות מסוימת כאשר במרבית המדינות צריכת האלכוהול הינה נמוכה יחסית.

### אלכוהול



## נספחים – צילומי מסך של הקוד

```

1 # import data set from Excel
2 dataset <- read.csv(file.choose(), header = T)
3
4 # packages and libraries
5 install.packages("moments")
6 library(moments)
7 install.packages("car")
8 library("car")
9 install.packages("ggplot2")
10 library(ggplot2)
11 library(dplyr)
12 library(scales)
13 theme_set(theme_classic())
14 install.packages('vcd')
15 library('vcd')
16
17 # Creating data frame and excluding the NA row
18 Dataframe <- data.frame(
19   Life_expectancy = c(dataset$Life_expectancy), #1 Life_expectancy
20   outdoor.air.pollution = c(dataset$outdoor.air.pollution...), #2 air.pollution
21   HIV = c(dataset$HIV...Estimated.number.of.people.that.have.been.infected), #3 HIV
22   Malaria = c(dataset$Malaria...Estimated.number.of.people.that.have.been.infected), #4 Malaria
23   Average.income.per.person = c(dataset$Average.income.per.person...), #5 income
24   Alcohol.consumption = c(dataset$Alcohol.consumption.per.person.liters..year.), #6 Alcohol
25   density = c(dataset$density.per.square.km.), #7 density
26   cigarette.consumption = c(dataset$cigarette.consumption...), #8 cigarette
27   Continent = c(dataset$Continent), #9 Continent
28   Member.of.OECD = c(dataset$Member.of.OECD)) #10 OECD
29 Dataframe <- na.omit(Dataframe)
30
31 # x and y
32 plot(x = Dataframe[,2], y=Dataframe[,1], xlab = "Air Pollution", ylab = "Life Expectancy")
33 plot(x = Dataframe[,3], y=Dataframe[,1], xlab = "HIV", ylab = "Life Expectancy")
34 plot(x = Dataframe[,4], y=Dataframe[,1], xlab = "Malaria", ylab = "Life Expectancy")
35 plot(x = Dataframe[,5], y=Dataframe[,1], xlab = "Average income per person", ylab = "Life Expectancy")
36 plot(x = Dataframe[,6], y=Dataframe[,1], xlab = "Alcohol consumption", ylab = "Life Expectancy")
37 plot(x = Dataframe[,7], y=Dataframe[,1], xlab = "Density", ylab = "Life Expectancy")
38 plot(x = Dataframe[,8], y=Dataframe[,1], xlab = "Cigarette consumption", ylab = "Life Expectancy")
39
40 # x and x
41 plot(x = Dataframe[,5], y=Dataframe[,7], xlab = "Average income per person", ylab = "Density")
42 cor(Dataframe[,5],Dataframe[,7])
43
44 plot(x = Dataframe[,5], y=Dataframe[,8], xlab = "Average income per person", ylab = "Cigarette consumption")
45 cor(Dataframe[,5],Dataframe[,8])
46
47 plot(x = Dataframe[,5], y=Dataframe[,9], xlab = "Average income per person", ylab = "Continent")
48 cor(Dataframe[,5],Dataframe[,9])
49
50 plot(x = Dataframe[,6], y=Dataframe[,8], xlab = "Alcohol consumption", ylab = "Cigarette consumption")
51 cor(Dataframe[,6],Dataframe[,8])
52
53 plot(x = Dataframe[,7], y=Dataframe[,8], xlab = "Density", ylab = "Cigarette consumption")
54 cor(Dataframe[,7],Dataframe[,8])
55
56 # show the data about each variable
57 summary(Dataframe[1:8])
58
59 # show the Standard Deviation of each variable
60 sd(Dataframe[,1])
61 sd(Dataframe[,2])
62 sd(Dataframe[,3])
63 sd(Dataframe[,4])
64 sd(Dataframe[,5])
65 sd(Dataframe[,6])
66 sd(Dataframe[,7])
67 sd(Dataframe[,8])
68
69 # Show the skewness of the fluent variables
70 skewness(Dataframe[,1])
71 skewness(Dataframe[,2])
72 skewness(Dataframe[,3])
73 skewness(Dataframe[,4])
74 skewness(Dataframe[,5])
75 skewness(Dataframe[,6])
76 skewness(Dataframe[,7])
77 skewness(Dataframe[,8])
78

```



```

73 skewness(DataFrame[,4])
74 skewness(DataFrame[,5])
75 skewness(DataFrame[,6])
76 skewness(DataFrame[,7])
77 skewness(DataFrame[,8])
78
79 # analyze categorical variables
80 # Categorical variables table
81 prop.table(table(DataFrame$Continent))
82 prop.table(table(DataFrame$Member.of.OECD))
83 # Continent
84 mean(DataFrame[DataFrame$Continent == 1,1])
85 sd(DataFrame[DataFrame$Continent == 1,1])
86 skewness(DataFrame[DataFrame$Continent == 1,1])
87
88 mean(DataFrame[DataFrame$Continent == 2,1])
89 sd(DataFrame[DataFrame$Continent == 2,1])
90 skewness(DataFrame[DataFrame$Continent == 2,1])
91
92 mean(DataFrame[DataFrame$Continent == 3,1])
93 sd(DataFrame[DataFrame$Continent == 3,1])
94 skewness(DataFrame[DataFrame$Continent == 3,1])
95
96 mean(DataFrame[DataFrame$Continent == 4,1])
97 sd(DataFrame[DataFrame$Continent == 4,1])
98 skewness(DataFrame[DataFrame$Continent == 4,1])
99
100 mean(DataFrame[DataFrame$Continent == 5,1])
101 sd(DataFrame[DataFrame$Continent == 5,1])
102 skewness(DataFrame[DataFrame$Continent == 5,1])
103
104 # Member of OECD
105 mean(DataFrame[DataFrame$Member.of.OECD == 0,1])
106 sd(DataFrame[DataFrame$Member.of.OECD == 0,1])
107 skewness(DataFrame[DataFrame$Member.of.OECD == 0,1])
108
109 mean(DataFrame[DataFrame$Member.of.OECD == 1,1])
110 sd(DataFrame[DataFrame$Member.of.OECD == 1,1])
111 skewness(DataFrame[DataFrame$Member.of.OECD == 1,1])
112
113 # boxplot_Harigim
114 # boxplot_with_Harigim
115 bp<-boxplot(DataFrame[,3],main='HIV')
116 bp<-boxplot(DataFrame[,5],main='Average income per person')
117 bp<-boxplot(DataFrame[,4],main='Malaria')
118 bp<-boxplot(DataFrame[,6],main='Alcohol consumption')
119 bp<-boxplot(DataFrame[,7],main='Density')
120 bp<-boxplot(DataFrame[,8],main='Cigarette consumption')
121 bp<-boxplot(DataFrame[,9],main='Continent')
122

```

```

123 # boxplot_Without_Harigim
124 bp<-boxplot(DataFrame[,3],main='HIV',outline=FALSE)
125
126 # boxplot_Exclude_ChosenValues
127 bp<-boxplot(DataFrame[,3],main='HIV')
128 bp$out #show the exceptional values
129 dataset1<-subset(dataset,DataFrame[,3]<4e+06) #exclude the chosen exceptional value
130 bp<-boxplot(dataset1$HIV...Estimated.number.of.people.that.have.been.infected,main='HIV without The Exceptional Value')
131
132 bp<-boxplot(DataFrame[,4],main='Malaria')
133 bp$out #show the exceptional values
134 dataset1<-subset(dataset,DataFrame[,4]<80000) #exclude the chosen exceptional value
135 bp<-boxplot(dataset1$Malaria...Estimated.number.of.people.that.have.been.infected,main='Malaria without The Exceptional Value')
136
137 bp<-boxplot(DataFrame[,7],main='Density')
138 bp$out #show the exceptional values
139 dataset1<-subset(dataset,DataFrame[,7]<1200) #exclude the chosen exceptional value
140 bp<-boxplot(dataset1$Density.per.square.km.,main='Density without The Exceptional Value')
141
142 # create density function
143 hist(DataFrame[,2], prob = TRUE, main = 'Density Function - outdoor air pollution', xlab = 'outdoor air pollution (%)', col='grey')
144 lines(density(DataFrame[,2]), col="blue", lwd=2)
145 hist(DataFrame[,5], prob = TRUE, main = 'Density Function - Average income per person', xlab = 'Average income per person ($)', col='grey')
146 lines(density(DataFrame[,5]), col="blue", lwd=2)
147 hist(DataFrame[,7], prob = TRUE, main = 'Density Function - Density per square km', xlab = 'Density per square km', col='grey')
148 lines(density(DataFrame[,7]), col="blue", lwd=2)
149
150 # compute ecdf values and create ecdf plot
151 ecdf(DataFrame[,2])
152 plot(ecdf(DataFrame[,2]), main = "Cumulative Distribution Function - Outdoor air pollution (%)", xlab = "outdoor air pollution (%)", ylab = "Fn(x)", col='blue', pch=19)
153 ecdf(DataFrame[,5])
154 plot(ecdf(DataFrame[,5]), main = "Cumulative Distribution Function - Average Income per person ($)", xlab = "Average income per person ($)", ylab = "Fn(x)", col='blue', pch=19)
155 ecdf(DataFrame[,7])
156 plot(ecdf(DataFrame[,7]), main = "Cumulative Distribution Function - Density (km^2)", xlab = "Density (km^2)", ylab = "Fn(x)", col='blue', pch=19)
157

```



```

158 # representing connections with graphs
159 #plot1- scatter.smooth- Alcohol Consumption by Income Level- unexpected
160 scatter.smooth(x = DataFrame[,5], y = DataFrame[,6], xlab = "Average income per person", ylab = "Alcohol consumption" , pch=21, col = "black", bg = "slategray1", lwd=1, cex=1.5)
161 #plot2-continents and density-unexpected
162 ggplot(dataset,
163   aes(x = dataset$Continent,
164     y = dataset$density.per.square.km.)) +
165   geom_col(fill = "lightpink",
166     color = "black") +
167   ylab("Average percent [km^2]") +
168   xlab("Continent") +
169   ggtitle("Continents and their respective proportion of density (average)")
170 # plot 3- Density by Cigarette consumption- expected
171 b<-barplot(DataFrame[,8],
172   DataFrame[,7],
173   col = "purple",
174   border = "black",
175   ylab = "Density",
176   xlab = "Cigarette consumption",
177   main = "Density by cigarette consumption")
178 # plot 4- Alcohol and Cigarette- expected
179 plot(x=DataFrame[,6],y= DataFrame[,8],xlab = 'Alcohol',ylab='Cigarette')
180 abline(lm(DataFrame[,6] ~ DataFrame[,8], data=DataFrame), lwd=2,col="blue")
181 # plot 5- HIV by Continent- expected
182 ggplot(DataFrame,
183   aes(x = DataFrame[,9],
184     y = DataFrame[,3])) +
185   geom_bar(stat="identity")+
186   ylab("HIV") +
187   xlab("Continent") +
188   ggtitle("HIV by continent")
189
190 #IDM
191 #Average income per person:
192 #Cigarette consumption:
193 table1b <- table(cut(DataFrame$cigarette.consumption, breaks = seq(0,100,20)))
194 table1b
195 prop.table(table1b)
196 #ZDM
197 #outdoor air pollution - Life expectancy:
198 table2a <- table(cut(DataFrame$outdoor.air.pollution, breaks= seq(0,100,20)), cut(DataFrame$Life_expectancy, breaks= seq(0,100,25)))
199 table2a
200 prop.table(table2a)
201 #Continent - Alcohol Consumption:
202 table2b <- table(DataFrame$Continent ,cut(DataFrame$Alcohol.consumption, breaks= seq(0,100,20)))
203 table2b
204 prop.table(table2b)

```