

# פרוייקט

# בכריית מידע

מרצה: ד"ר אלי פקר

מגישים:

ברק 313584773

טליה נודל 311335665

אפי לדר 316012236

מור יואבי 205500606

## מטרת הפרוייקט: זיהוי קריטריונים למדד איכות חיים

ההנחה: ישנו קשר ישיר בין מצב סוציו-אקונומי לבין מדדי איכות החיים

### חתכים שנבדוק:

1. ממוצע ילודה במדינה למול מצב סוציו-אקונומי
2. השכלה למול מצב סוציו-אקונומי
3. מצב בריאותי למול מצב סוציו-אקונומי

### מחקר רקע:

- באתר EuroStat, של ארגון EuroStat העוסק בסטטיסטיקות ביבשת אירופה מטעם האיחוד האירופאי, הקריטריונים לאיכות חיים הינם: תנאי מגורים, סיפוק כללי מהחיים, תעסוקה, ניצול של הזמן, בריאות, השכלה, קשרים חברתיים, בטחון, ממשלה, סביבה ותנאים חומריים.

[https://ec.europa.eu/eurostat/cache/infographs/qol/index\\_en.html](https://ec.europa.eu/eurostat/cache/infographs/qol/index_en.html)

- במאמר "Quality Of Life" של מגזין U.S.News דובר על כך שמעבר לקריטריונים של מזון, תנאי מגורים, איכות ההשכלה, טיפול בריאותי, והמעסיק שמעסיק אותנו, איכות חיים כוללת גם משתנים פחות מדידים חומריים, כמו: בטחון תעסוקתי, יציבות פוליטית, חופש היחיד ואיכות סביבתית.

<https://www.usnews.com/news/best-countries/quality-of-life-rankings>

- במאמר "The Wealth Of Nations Revisited: Income and Quality Of Life" מוזכר ניסוי שבמהלכו מתוך קבוצה של 101 מדינות נדגמו 32 בעלות מדגם מייצג של ערכים אנושיים כלל עולמיים (שמחה, סדר חברתי, צדק חברתי). **עושר קושר משמעותית עם 26 מתוך ה-32, שבהן נצפו ציוני מדד איכות חיים גבוהים יותר.**

<https://www.jstor.org/stable/27522878>

### **איך אנחנו מגדירים איכות חיים -**

איכות חיים הוא מושג סובייקטיבי בעל חפיפה חלקית למושג רווחה. איכות חיים כוללת, בין היתר, את המידה בה אדם מעריך שהוא או סובביו מאושרים. אנשים שונים מעריכים איכות חיים על פי קריטריונים שונים או שנותנים משקל שונה לאותם קריטריונים בהערכתם זו.

בניתוח שלנו התייחסנו לאיכות חיים בקריטריונים הניתנים לחישוב סטטיסטי ולא על סמך סקרים אישיים. התייחסנו למונחים שאנו רואים בהם מדד לרמת איכות החיים והם תוחלת חיים, ילודה והשכלה גבוהה.

תוחלת חיים משקפת לנו את הקידמה הרפואית ומתן שירותים רפואיים לאוכלוסיה ואיכותם. בנוסף תוחלת חיים גם יכולה להצביע על איכות ומגוון המזון הנצרך באוכלוסייה, ותדירות הפעילות הגופנית. דברים שלא ספק מצביעים על איכות חיים.

השכלה גבוהה מצביע לנו על יכולת האדם לרכוש השכלה גבוהה, ויכולתו להתפרנס בזכות עצמו במקצוע אשר בחר ללמוד אותו, דבר המצביע גם על חופש בחירה, וגם הבטחת הכנסה.

ממוצע ילודה מצביע לנו על הרצון והיכולת של אוכלוסיה להקים משפחה, לדאוג ולטפל בילדהם ולרצות בעתידם הטוב.

**במחקר שלנו** החלטנו על מספר קריטריונים, הנגזרים מתוך קריטריונים שחזרו בשניים מהמקורות הנ"ל (אתר EuroStat והמאמר "Quality Of Life"). הקריטריון שחזר על עצמו, ושמעין במאמר השלישי עלתה חשיבותו, הינו עושר. לכן להבנתנו הכנסה כלכלית מהווה משקל משמעותי בקביעת איכות חיים במדינה. לכן בחרנו להשתמש בו, ולהשוות אותו ל-3 קריטריונים נוספים כדי להבין מי מהם, אם בכלל, מנבא מדד לאיכות חיים.

### נתוני ה-Data:

- **ארגון:** הלשכה המרכזית לסטטיסטיקה באיחוד האירופאי

### **הסבר על הארגון:**

ארגון EuroStat עוסק בסטטיסטיקות ביבשת אירופה מטעם האיחוד האירופאי. הארגון פועל בשיתוף עם מכוני סטטיסטיקה לאומיים ורשויות לרומיות אחרות במדינות החברות באיחוד האירופי. שותפות זו ידועה בשם "המערכת הסטטיסטית האירופית", והיא כוללת גם את הרשויות הסטטיסטיות של מדינות האזור הכלכלי האירופי ושווייץ.

### **קישור:**

- <https://ec.europa.eu/eurostat/databrowser/>

### השערות:

ההנחה הרווחת היא שישנו קשר בין מצב סוציו-אקונומי לבין מדדי איכות חיים. מדינות במצב סוציו-אקונומי גבוהה יותר גם איכות החיים תהיה גבוהה יותר, וההשערות שלנו הינן:

- במדינות בעלות הכנסה ממוצעת **גבוהה** יותר, יהיה ממוצע ילודה **נמוך** יותר, וזאת מתוך ההנחה הרווחת כי במדינות מפותחות יש ילודה פחותה יותר מאשר במדינות מתפתחות.
- במדינות בעלות הכנסה ממוצעת **גבוהה** יותר יהיה אחוז **גבוהה** יותר של בעלי השכלה גבוהה, כיוון שלימודים לתואר אקדמאי עולים כסף ברוב מדינות העולם. בנוסף, לימודים לתואר אקדמאי מצריכים לרוב בסיס כלכלי מסוים, כדי לאפשר שעות לימוד על חשבון שעות עבודה.
- במדינות בעלות הכנסה ממוצעת **גבוהה** יותר יהיה תוחלת חיים **ארוכה** יותר. לפי ההנחה שלנו, ככל לאנשים יש יותר כסף קח הם יכולים לאפשר לעצמם מזון בריא יותר וקבלת טיפולים רפואיים איכותיים יותר.

## תהליך העבודה:

- **Data cleaning:** לקחנו כל טבלה בנפרד והורדנו אותה לקובץ excel, קיבלנו טבלאות בצורה הבאה:

Data extracted on 29/12/2021 14:00:18		from [ESTAT]				
Dataset		Structure of earnings survey: monthly earnings [EARN_SES_MONTHLY\$DEFAULTVIEW]				
Last updated:		14/10/2021 23:00				
Time frequency		Annual				
Statistical classification of economic activities in International Standard Classification of Occupati		Industry, construction and services (except public administration, defense, compulsory social security)				
Working time		Total				
Age class		Total				
Sex		Total				
Structure of earnings indicator		Mean earnings in euro				
	TIME	2002	2006	2010	2014	2018
GEO (Labels)						
European Union - 27 countries (from	:		2,000	2,012	2,412.59	2,319
European Union - 28 countries (2013	:		2,010	2,097	2,278	.
European Union - 27 countries (2007	:			2,333.08	2,548.67	2,418
Euro area - 19 countries (from 2015)	:			2,228	2,387	2,521
Euro area - 18 countries (2014)	:		2,082	2,247	2,407	2,539
Euro area - 17 countries (2011-2013)	:			.	.	2,550
Belgium	:		2,438	2,749	2,956	3,360
Bulgaria	:	143	193	332	420	608
Czechia	:	472	714	922	909	1,225
Denmark	:		3,062	3,642	3,690	4,473
Germany (until 1990 former territory	:		2,457	2,520	2,620	3,349
Estonia	:		599	777	999	1,285
Ireland	:	2,549	3,142	3,274	3,340	3,780
Greece	:		1,563	1,744	1,511	1,356
Spain	:	1,475	1,597	1,775	1,829	2,043
France	:		2,342	2,400	2,603	2,895
Croatia	:		.	980	1,033	1,179
Italy	:		2,030	2,193	2,337	2,520
Cyprus	:		1,815	1,980	1,798	1,909
Latvia	:		431	573	692	1,158
Lithuania	:	299	441	525	640	970
Luxembourg	:		3,112	3,501	3,913	4,630
Hungary	:	451	608	729	774	996
Malta	:		1,206	1,375	1,566	1,939
Netherlands	:	2,035	1,917	2,281	2,964.09	3,076
Austria	:		2,124	2,288	2,494	3,066
Poland	:	573	648	793	948	1,105
Portugal	:		1,135	1,231	1,209	1,293
Romania	:	195	332	444	512	951
Slovenia	:	1,010	1,198	1,470	1,571	1,719
Slovakia	:	337	519	758	908	1,164
Finland	:		2,466	2,848	3,094	3,360
Sweden	:	2,449	2,510	2,809	3,286	3,488
Iceland	:		2,891	1,996	2,535	4,854
Norway	:		3,546	3,987	4,403	4,743
Switzerland	:			4,112	5,103	5,624
United Kingdom	:	2,681	2,893	2,504	2,795	3,258
Montenegro	:			.	717	.
North Macedonia	:			475	490	582
Albania	:			.	.	398
Serbia	:			.	567	626
Turkey	:		609	753	752	692
Special value						
:		not available				

מתוך הטבלה הנ"ל עשינו ניקוי של כל המידע הלא רלוונטי ביניהם:

- ניקינו את כל השורות שבעצם כבר עושות ממוצע לפי כמות מדינות ולפי חברות באיחוד האירופאי ולא חברות.
- ניקינו את המדינות אשר בשנת 2010 היה חסר להם DATA.
- ניקינו את כל השנים שהם לא 2010

• data integration :

לאחר שביצענו את הניקוי על ארבעת הטבלאות שהחלטנו לנתח, בנינו טבלה מאחדת בין הטבלאות ע"י פונקציית vlookup של אקסל. ובכך הצלחנו לשייך את ה-4 הנותנים מאותה שנה(2010) בעזרת גורם מקשר שהוא בעצם שם המדינה

החלטנו לנקות את המדינות שבהם לא היה נתונים של 4 הפרמטרים

התקבלה הטבלה הבאה:

TIME	2010			
מדינה	תחלית חיים	אחוז בעלי תואר	תבוסות בידור ממוצאן לאדם	ילודה במדינה
Switzerland	4112	37.4	64.5	1.52
Norway	3987	45	69.9	1.95
Luxembourg	3501	44.2	63.2	1.63
Finland	2848	39.2	58.2	1.87
Sweden	2809	42.3	66.1	1.98
United Kingdom	2504	41.6	65.3	1.92
France	2400	42.7	65.4	2.03
Austria	2288	20.7	65.1	1.44
Netherlands	2281	40.3	60	1.79
Iceland	1996	36.2	68.8	2.20
Cyprus	1980	48.1	59.3	1.44
Slovenia	1470	31.3	59	1.57
Malta	1375	24.3	59.4	1.36
Portugal	1231	25.5	48.1	1.39
Croatia	980	25.8	48.3	1.55
Poland	793	37.1	51	1.41
Slovakia	758	24	48.5	1.43
Hungary	729	26.1	51.3	1.25
Latvia	573	34.7	49	1.36
Lithuania	525	46.3	48.5	1.50
Romania	444	20.7	48	1.59

• data transformation:

מכיוון שקיימים לנו 4 מאגרים נותנים שונים (הכנסה ביורו, אחוז בעלי תואר ראשון, תוחלת חיים, ילודה ממוצעת) וכל אחד מהם נמדד בקנה מידה אחר החלטנו לעשות דירוג לכל סוג נתונים שבעצם מדרג ב-21 את המדינה בעלת ההכנסה ביורו הגבוהה ביותר, ודירוג 1 לבעלת ההכנסה הנמוכה ביותר. כנ"ל לגבי אחוז בעלי תואר ראשון, ניתן דירוג 21 למדינה עם האחוז הגבוהה ביותר ודירוג 1 לנמוך ביותר. וחוזר על עצמו בתוחלת החיים עם דירוג 21 למדינה עם תוחלת החיים הגבוהה ביותר ודירוג 1 לתוחלת החיים הנמוכה ביותר.

החלטנו שלא לדרג את הילודה כי לא ניתן באמת להגיד מה טוב יותר, הרבה ילדים או מעט ילדים, ולכן המבחן זה החרגנו את נתוני הילודה.

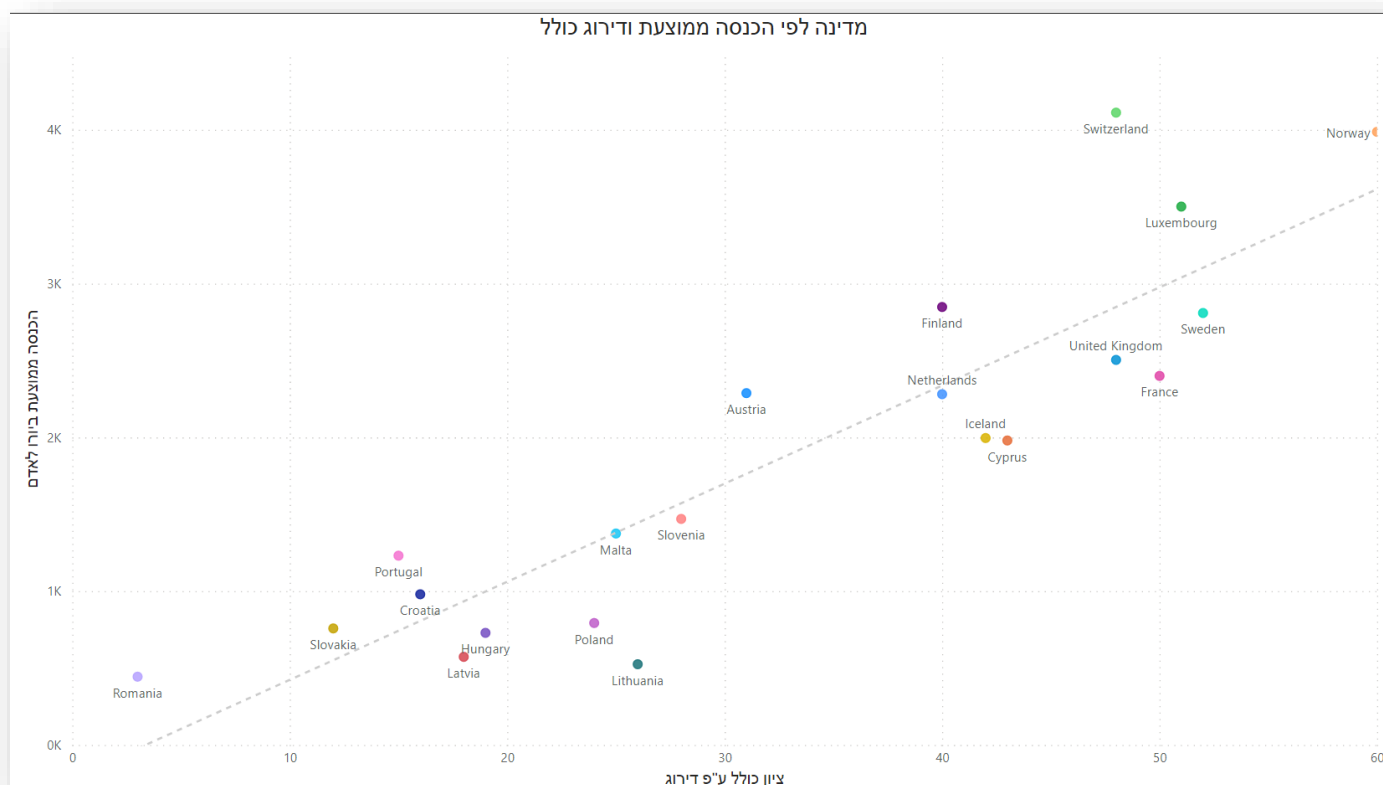
את הדירוג ביצענו בעזרת פונקציית RANK של אקסל שבעזרתה ניתן לומר את דירוג התא ביחס לשאר התאים.

נאחד את סכום 3 הדירוגים הנותרים ונייצר גרף אשר בציר ה-X יהיה הכנסות ביורו וציר ה-Y יהיה סכום 3 המדדים. ונראה אם יש קשר ביניהם. נבדוק אם הקשר משמעי למבחן חי בריבוע

התקבלה הטבלה הבאה:

TIME	2010									
	תכנסות ביורו ממוצעת לאדם		אחוז בעלי תואר		תוחלת חיים		ילודה בג'יניני		ציון כולל	
	דירוג	דירוג	דירוג	דירוג	דירוג	דירוג	דירוג			
מדינה	Switzerland	4112	21	37.4	12	64.5	15	1.52		48.
	Norway	3987	20	45	19	69.9	21	1.95		60.
	Luxembourg	3501	19	44.2	18	63.2	14	1.63		51.
	Finland	2848	18	39.2	13	58.2	9	1.87		40.
	Sweden	2809	17	42.3	16	66.1	19	1.98		52.
	United Kingdom	2504	16	41.6	15	65.3	17	1.92		48.
	France	2400	15	42.7	17	65.4	18	2.03		50.
	Austria	2288	14	20.7	1	65.1	16	1.44		31.
	Netherlands	2281	13	40.3	14	60	13	1.79		40.
	Iceland	1996	12	36.2	10	68.8	20	2.20		42.
	Cyprus	1980	11	48.1	21	59.3	11	1.44		43.
	Slovenia	1470	10	31.3	8	59	10	1.57		28.
	Malta	1375	9	24.3	4	59.4	12	1.36		25.
	Portugal	1231	8	25.5	5	48.1	2	1.39		15.
	Croatia	980	7	25.8	6	48.3	3	1.55		16.
	Poland	793	6	37.1	11	51	7	1.41		24.
	Slovakia	758	5	24	3	48.5	4	1.43		12.
	Hungary	729	4	26.1	7	51.3	8	1.25		19.
	Latvia	573	3	34.7	9	49	6	1.36		18.
	Lithuania	525	2	46.3	20	48.5	4	1.50		26.
Romania	444	1	20.7	1	48	1	1.59		3.	

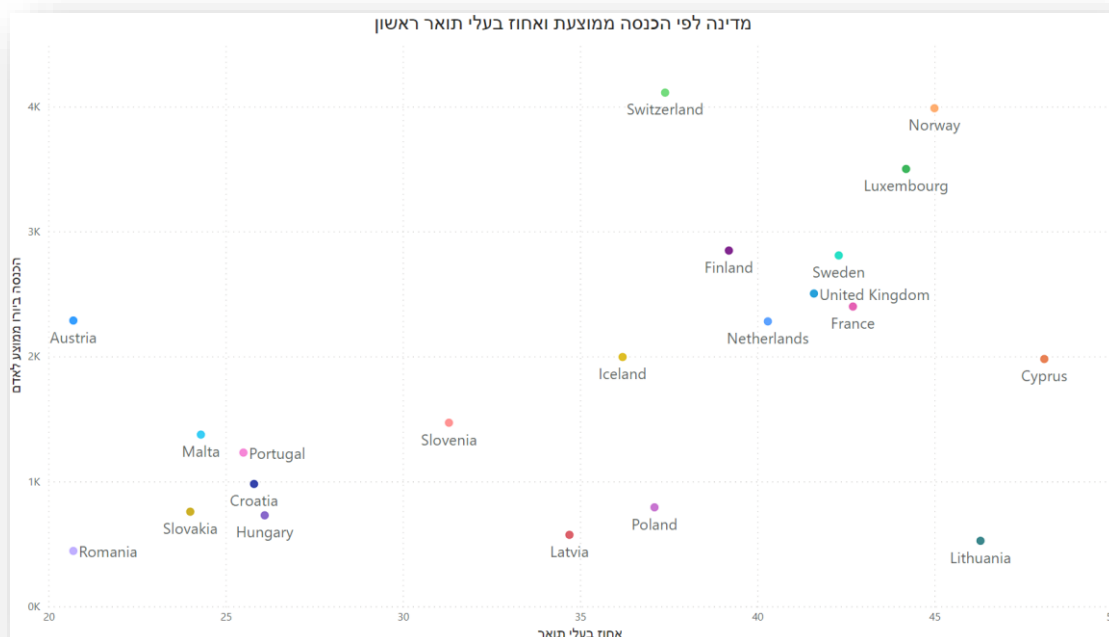
• **ניתוח הנתונים:**



לפי מבט ראשוני ניתן לחשוב כי בהחלט יש קשר בין ההכנסה לשאר הדירוגים שעשינו.

כעת נבצע בדיקת רגרסיה לינארית מופרדת בין הקטגוריות הבאות ונדרוש כי R-square יהיה גדול שווה ל-80%. אם לא מתקיים התנאי נסיק כי אין קשר לינארי בין שתי הקטגוריות:

# 1. הכנסה ממוצעת ואחוז בעלי התואר במדינה:



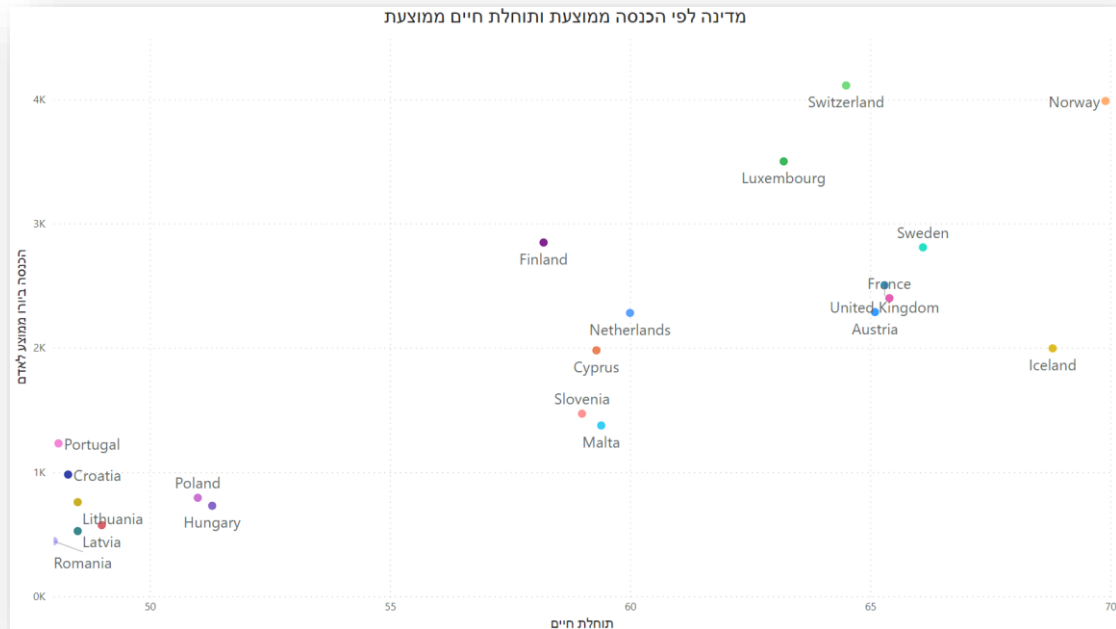
נבצע בדיקת רגרסיה לינארית:

הכנסה ביו"ר ממוצעת לאדם ואחוז בעלי התואר	
0.516071711	מקדם המתאם R
65.4434114	שיפוע קו ריגרסיה a
-400.8924886	חיתוך עם ציר b

ניתן לראות כי מקדם המתאם R קטן מ-0.8 ולכן לא מתקיים קשר לינארי מובהק



## 2. הכנסה ממוצעת ותוחלת חיים במדינה:



נבצע בדיקת רגרסיה לינארית:

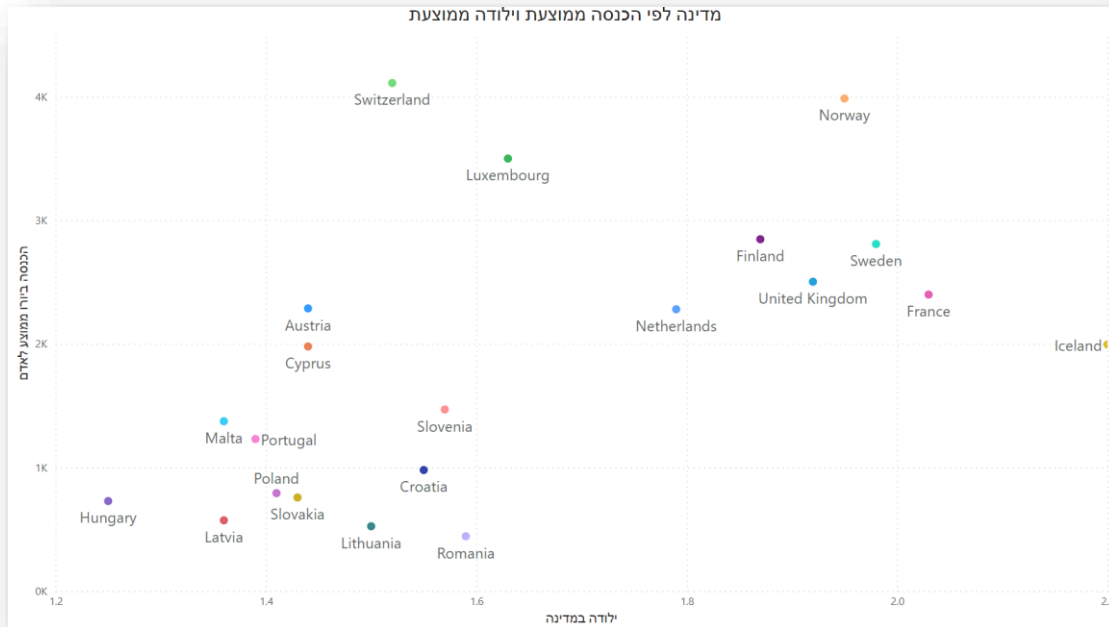
הכנסה ביורו ממוצעת לאדם ותוחלת החיים הממוצעת	
0.83836836	מקדם המתאם R
122.7362535	שיפוע קו רגרסיה a
-5227.321278	חיתוך עם ציר b

ניתן לראות כי מקדם המתאם R גדול מ-0.8 ולכן מתקיים קשר לינארי חזק בין העמודות

ולכן המשוואה הלינארית המתקיימת היא:

$$y = 122.73x - 5227.32$$

### 3. הכנסה ממוצעת וילודה במדינה:



נבצע בדיקת רגרסיה לינארית:

הכנסה ביו"ר ממוצעת לאדם	
וילודה ממוצעת	
0.540742561	מקדם המתאם R
2303.865009	שיפוע קו רגרסיה a
-1864.86219	חיתוך עם ציר b

ניתן לראות כי מקדם המתאם R קטן מ-0.8 ולכן לא מתקיים קשר לינארי מובהק

## מסקנות:

- מתוך הנתונים הנ"ל ניתן לראות כי רק בקשר בין הכנסה ממוצעת ביורו ותוחלת חיים מתקיים קשר לינארי חזק. לא נראה קשר בין אחוז בעלי תואר לבין הכנסה ממוצעת, וכן לא מזהה קשר בין ממוצע ילודה לבין אחוז בעלי תואר.
- מטרת הפרוייקט היתה זיהוי קריטריונים למדד איכות חיים, בהתבסס על כך שהכנסה כלכלית גבוהה מנבאת איכות חיים גבוהה. לכן, מהנתונים שקבלנו ניתן להסיק כי תוחלת חיים הינו קריטריון רלוונטי לניבוי איכות חיים במדינה.
- ניתן להעריך כי אחוז בעלי תואר במדינה וממוצע ילודה במדינה אינם מדדים רלוונטים לאיכות חיים, או לחילופין כי הניתוח בעבודה שלנו לא מספיק מעמיק, או מתבסס על data שאיננה רחבה מספיק.

## ניתוח נוסף:

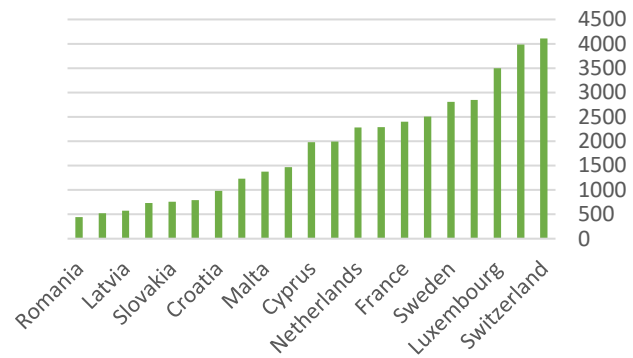
- לצורך ביצוע ניתוחים נוספים, בחרנו את העמודה של תוחלת חיים למול העמודה של הכנסה ביורו. כמו כן בחרנו לנתח את הנתונים לגבי שנת 2010. זאת בשל הסקנות הנ"ל המעידות על קשר אפשרי בין השתיים:

מדינה	הכנסה ביורו ממוצע לאדם	אחוז בעלי תואר
Switzerland	4112	37.4
Norway	3987	45
Luxembourg	3501	44.2
Finland	2848	39.2
Sweden	2809	42.3
United Kingdom	2504	41.6
France	2400	42.7
Austria	2288	20.7
Netherlands	2281	40.3
Iceland	1996	36.2
Cyprus	1980	48.1
Slovenia	1470	31.3
Malta	1375	24.3
Portugal	1231	25.5
Croatia	980	25.8
Poland	793	37.1
Slovakia	758	24
Hungary	729	26.1
Latvia	573	34.7
Lithuania	525	46.3
Romania	444	20.7

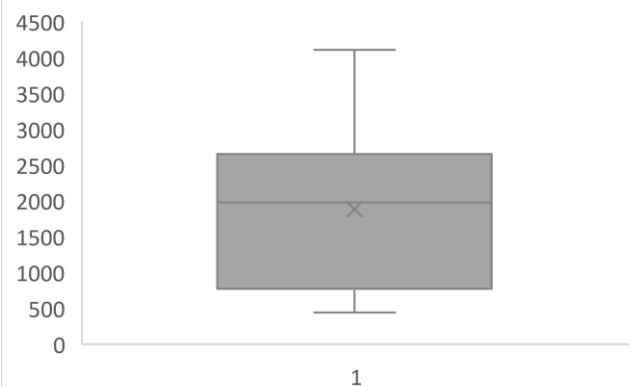
- הכנסה ביורו לאדם במוצק:

- ממוצע: 1884.93
- סטיית תקן: 1110.22
- חציון: 1980

הכנסה ביורו ממוצע לאדם - 2010



הכנסה ביורו ממוצעת לאדם - 2010

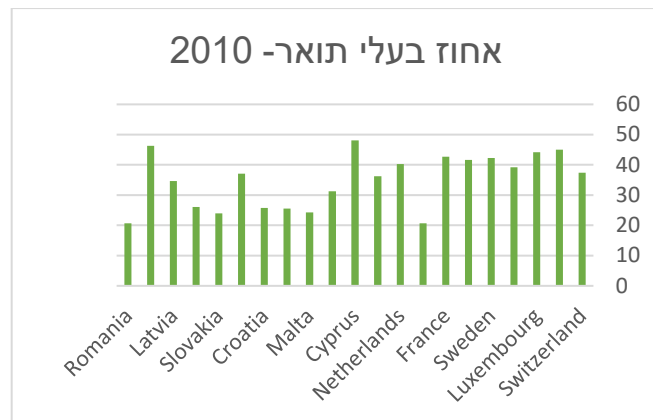


- אחוז בעלי תואר:

○ ממוצע: 34.93

○ סטיית תקן: 8.755

○ חציון: 37.1



• **ניתוח קורלציה – מבחן חי בריבוע**

ביצענו מבחן חי בריבוע הבודק האם קיימת תלות סטטיסטית בין 2 משתנים אשר נמדדו בסולם, כלומר האם קיים קשר סטטיסטי בין שני משתנים.

את המבחן ביצענו על הכנסה ממוצעת ותוחלת חיים בשנת 2010, בשנת 2014 ובשנת 2014. זאת בהמשך למסקנה אליה הגענו לפיה ישנו קשר ליניארי בין הכנסה ממוצעת ותוחלת חיים. במבחן חי בריבוע, ככל שהערך שנקבל יהיה קרוב יותר ל-0 כך הקורלציה בין שני המשתנים נמוכים יותר.

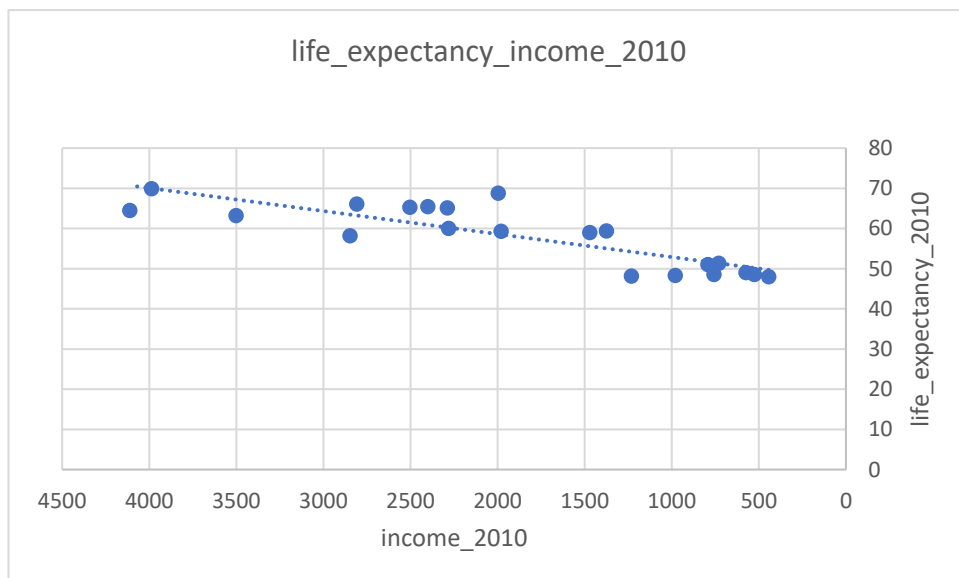
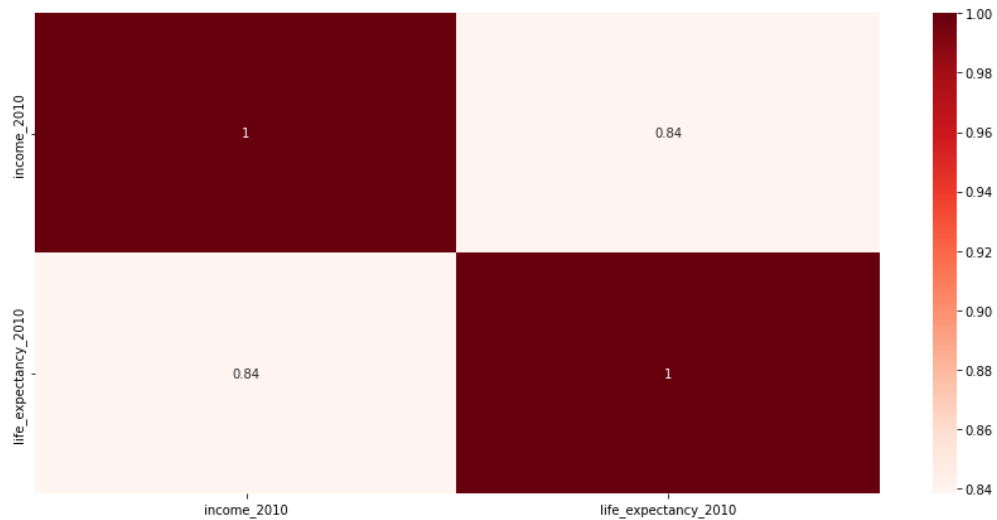
בחישובים הבאים ניתן לראות כי התוצאות שקיבלנו לגבי השנים 2010, 2014, 2018 הם בהתאמה: 0.84, 0.6, 0.45. המסקנה הנובעת מכך היא שבשנת 2010 הקורלציה בין הכנסה ותוחלת חיים הינה גבוהה.

יש לציין שברגרסיה הליניארית ערכנו את ההשוואה בין הכנסה ממוצעת לבין 3 הקריטריונים הנוספים לאיכות חיים רק לגבי שנת 2010, לכן מראש היה ניתן להניח שנקבל במבחן חי בריבוע תוצאות שמראות על קורלציה בין הכנסה ותוחלת חיים בשנה זו. לגבי השנים 2014 ו-2018 לא היתה לנו הנחה.

מבחן חי בריבוע – הכנסה למול תוחלת חיים – 2010  
הקורלציה בין ההכנסה לבין תוחלת החיים בשנת 2010 הוא 0.84.

```
[15] feature_cols = ['income_2010', 'life_expectancy_2010']

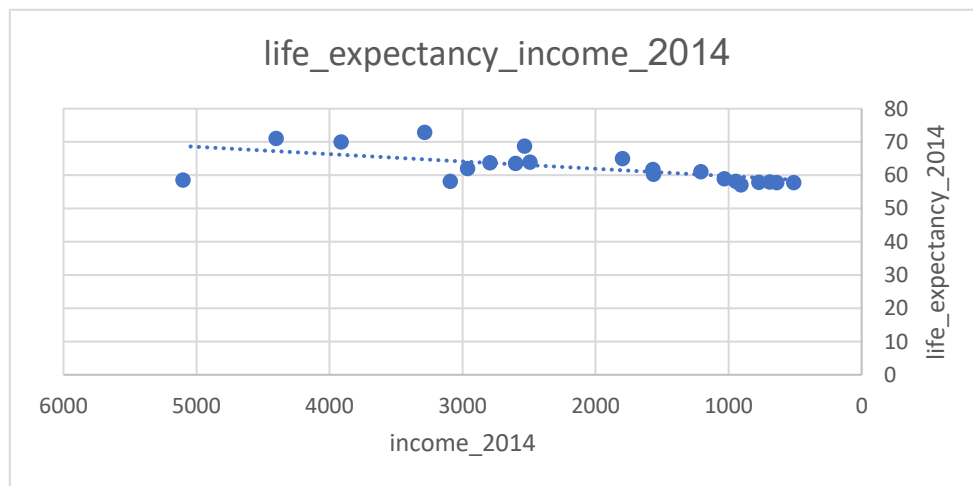
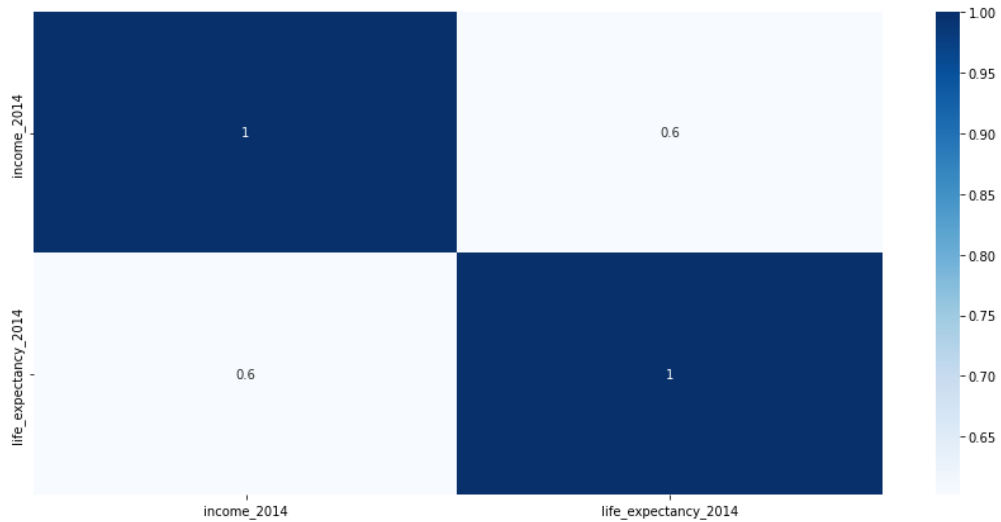
#correlation matrix
plt.figure(figsize=(15,7))
sns.heatmap(df[feature_cols].corr(),annot=True, cmap="Reds")
plt.show()
```



מבחן חי בריבוע – הכנסה למול תוחלת חיים – 2014  
הקורלציה בין ההכנסה לבין תוחלת החיים בשנת 2014 הוא 0.6.

```
[16] feature_cols = ['income_2014', 'life_expectancy_2014']

#correlation matrix
plt.figure(figsize=(15,7))
sns.heatmap(df[feature_cols].corr(),annot=True, cmap="Blues")
plt.show()
```





## מבחן חי בריבוע – הכנסה ותוחלת חיים – 2018

הקורלציה בין ההכנסה לבין תוחלת החיים לתואר בשנת 2018 הוא 0.45.

```
feature_cols = ['income_2018', 'life_expectancy_2018']

#correlation matrix
plt.figure(figsize=(15,7))
sns.heatmap(df[feature_cols].corr(),annot=True, cmap="Greens")
plt.show()
```

