**ויזואליזציה של מידע – פרויקט סיום**

המחלקה להנדסת תעשייה וניהול

**מרצה:**

**פרופסור גלעד רביד**

**מגישים:**

**מלודי עזרן (302311931)**

**אלמוג אליאס(312203243)**

**שחר שלי (307958512)**

**תאריך הגשה: 17/01/2022**

הסבר על מאגר הנתונים:

**מאגר הנתונים עליו התבססנו בפרויקט זה נלקח מאתר Data Gov המכיל מאגרי מידע ממשלתיים. הנתונים הינם נתוני תיירות פנים שנאספו לאורך שנת 2019 - שנה לפני פרוץ הקורונה שהייתה שנת "שיא" מבחינת כמות תיירים נכנסים לישראל והכנסות כתוצאה מכך.**

**המאגר מכיל 27,503 רשומות המתארות נתוני סקרים אשר מולאו ע"י התיירים הנכנסים לישראל. את הסקר מילאו התיירים בסוף שהותם בארץ. הסקר מכיל נתונים רבים ומגוונים – נתונים אישיים, דמוגרפיים, מהות החופשה,אופן הלנה, שביעות רצון, הוצאות כספיות במהלך החופשה ועוד.**

גרפים:

1. **גרף BoxPlot**

הנתונים שגזרנו לצורך הפקת גרף זה הינם:

* How many days prior to your tour did you purchase your flight? – כמה ימים מראש הוזמן הכרטיס.
* Country name – שם מדינת המוצא.
* TOT\_EXP $$ - כמה כסף מעריכים כי יבזבזו במהלך החופשה

המרנו את השדות בהתאם לערכים הבאים:

* country
* total\_cost
* prior\_days

העיבודים שנעשו:

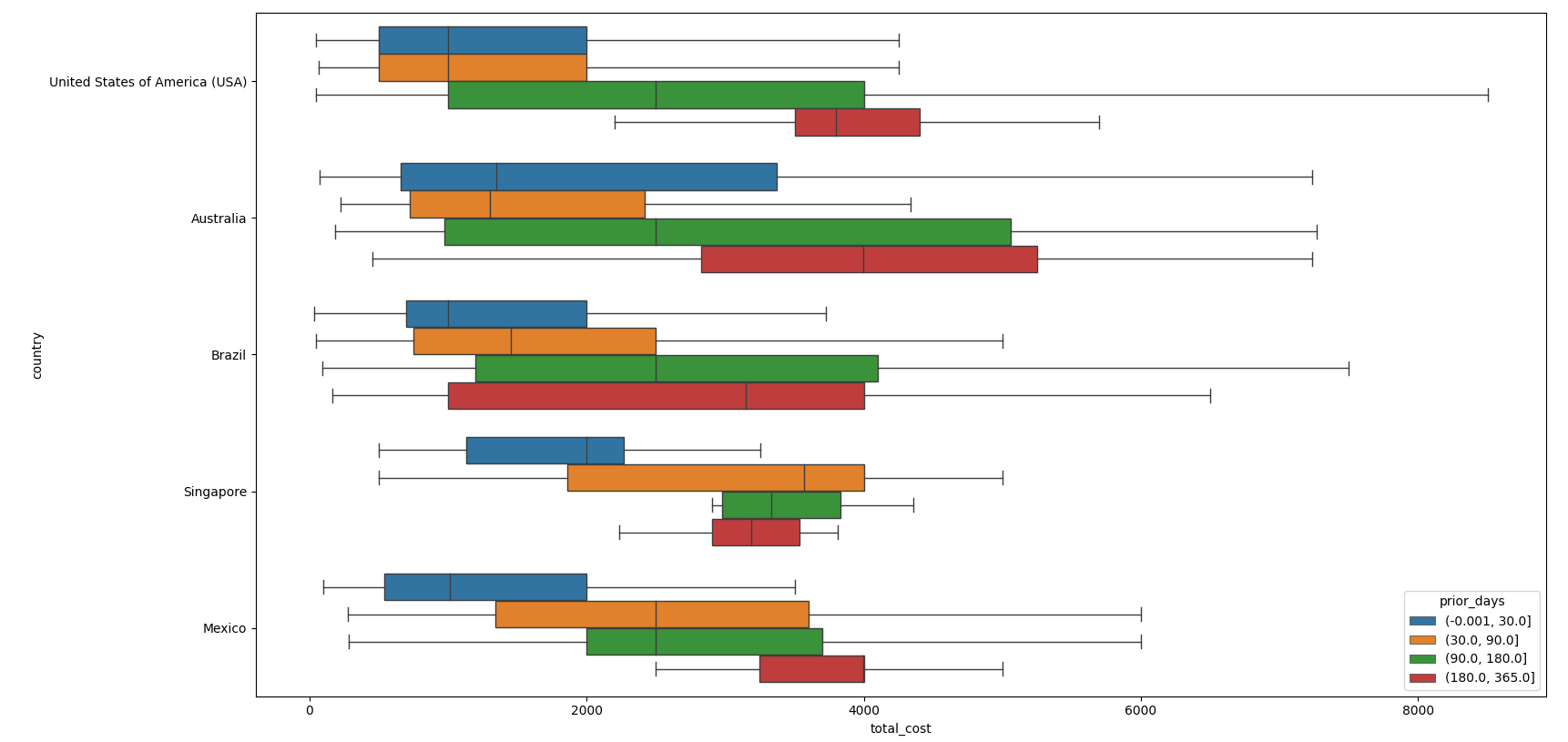
יצרנו טבלת pivot table לפי מפתח country והשתמשנו בפונקציה לחישוב ממוצע הערכים עבור שדה המחיר (total cost) לכל מדינה.

בחרנו את חמשת המדינות שמהן (בממוצע) כרטיסי הטיסה הוזמנו הכי הרבה זמן מראש לעומת שאר המדינות במאגר הנתונים.

ביצענו חלוקת BINS לשדה prior\_days ל - 4 בינים שונים באופן הבא:

1. כרטיס הטיסה הוזמן עד 30 יום לפני מועד הטיסה.
2. כרטיס הטיסה הוזמן 30-90 יום לפני מועד הטיסה.
3. כרטיס הטיסה הוזמן 90-180 יום לפני מועד הטיסה.
4. כרטיס הטיסה הוזמן מעל 180 יום לפני הטיסה (עד 365 יום במקסימום).

הכנסנו את הטבלה שנוצרה לתוך boxplot כאשר ציר ה x הינו total\_cost וציר הy הם 5 המדינות שפילטרנו קודם לכן, והצבעים השונים מתייחסים לכל BIN לחוד.



על פי הגרף ניתן לראות כי ההוצאות נעות בין 1000$-4500$ בממוצע לתיירים מחמשת המדינות להלן אשר הגיעו לארץ ב2019.

בנוסף לכך, ניתן לראות בצורה מובהקת כי ככל שרכשו התיירים את כרטיסי הטיסה יותר זמן מראש כך הם ביזבזו יותר כסף כשטיילו בארץ.

המדינות בהן התיירים ביזבזו הכי הרבה כסף בממוצע כאשר רכשו כרטיסי הטיסה יותר זמן מראש הם אוסטרליה וארה"ב.

1. **מפת עולם ומפת חום לפי שביעות רצון – גרף אינטראקטיבי**
2. תיאור הנתונים:

הנתונים שגזרנו לצורך הפקת גרף זה הינם:

* Q date – תאריך כניסה לארץ.
* Country name – שם מדינת מוצא.
* Overall satisfaction – מדד רמת שביעות רצון.
* value for money – מדד ערך תמורת מחיר.

עיבודים שנעשו:

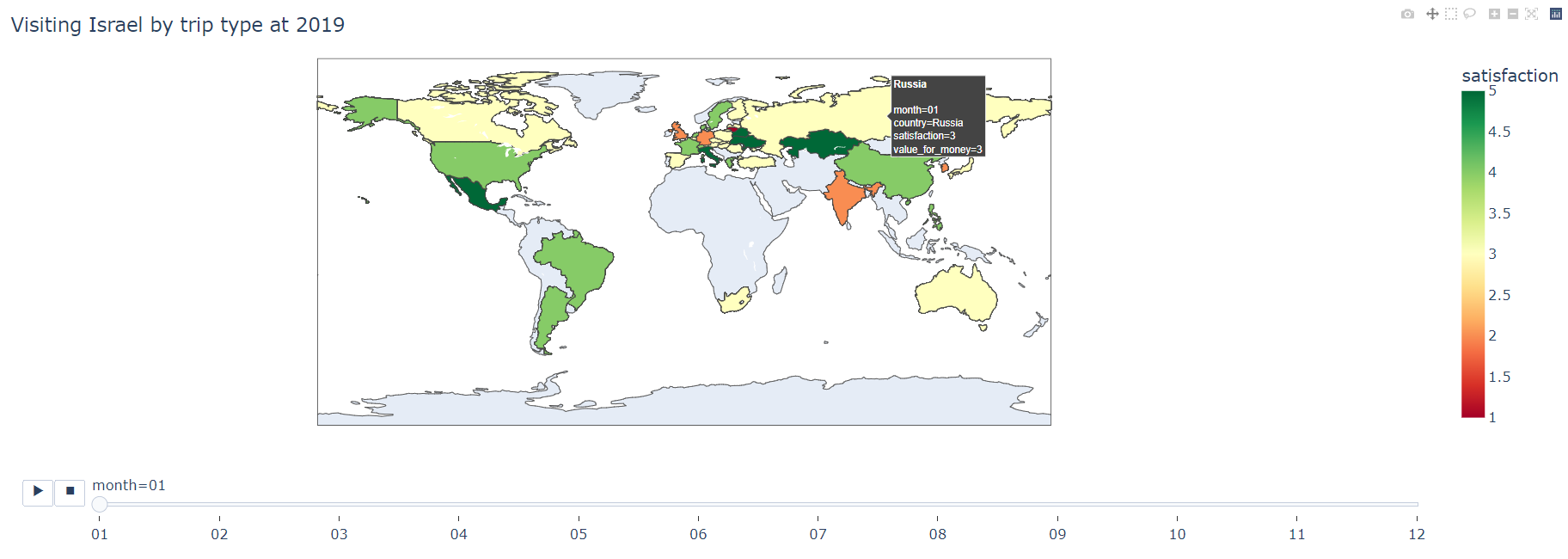
המרנו את השדות בהתאם לערכים הבאים:

* date
* country
* satisfaction
* value\_for\_money

הסרנו את כלל הרשומות הריקות בטבלה. המרנו את satisfaction ו- value\_for\_money לערכים מספריים (INT). שיננו את הערכים של המדינות הבאות:

* United States of America (USA) לערך – "USA"
* Russian Federation לערך – "Russia".

לסיום, סידרנו את הטבלה לפי תאריכים כך שתהיה התאמה עם מימד הזמן.

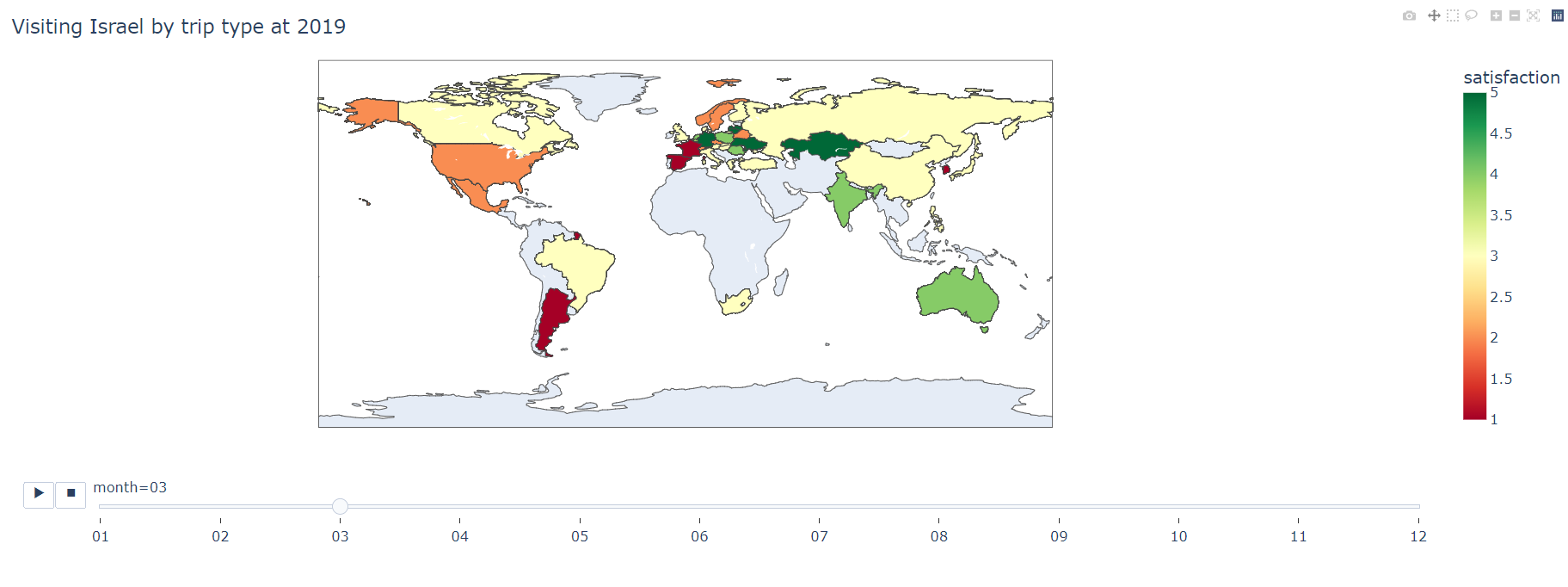


על פי הגרף ניתן לראות בצורה ברורה את ממוצע שביעות הרצון של התיירים שנכנסו לישראל לפי מדינות המוצא שלהם על ציר הזמן בשנה הנבחנת (2019).

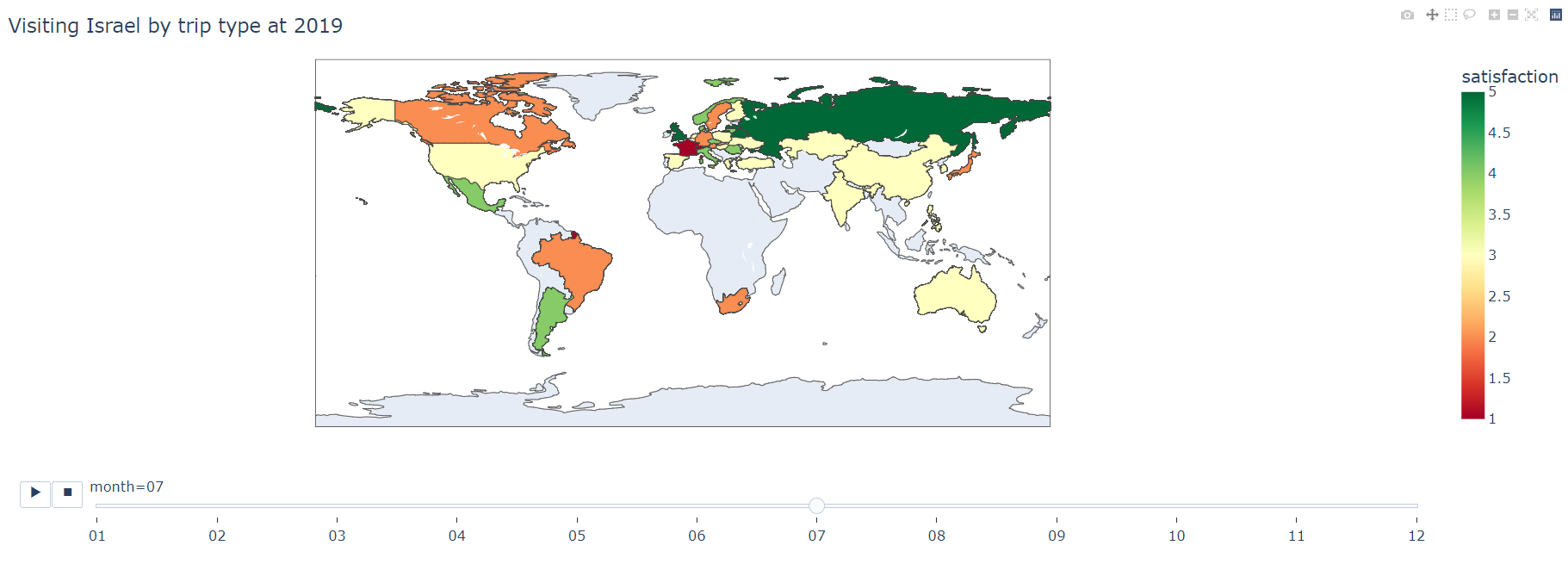
על ידי סימון המדינה ניתן לראות את מדד הערך תמורת המחיר שאותו דיווחו התיירים.

תובנות מעניינות שעולות מהגרף:

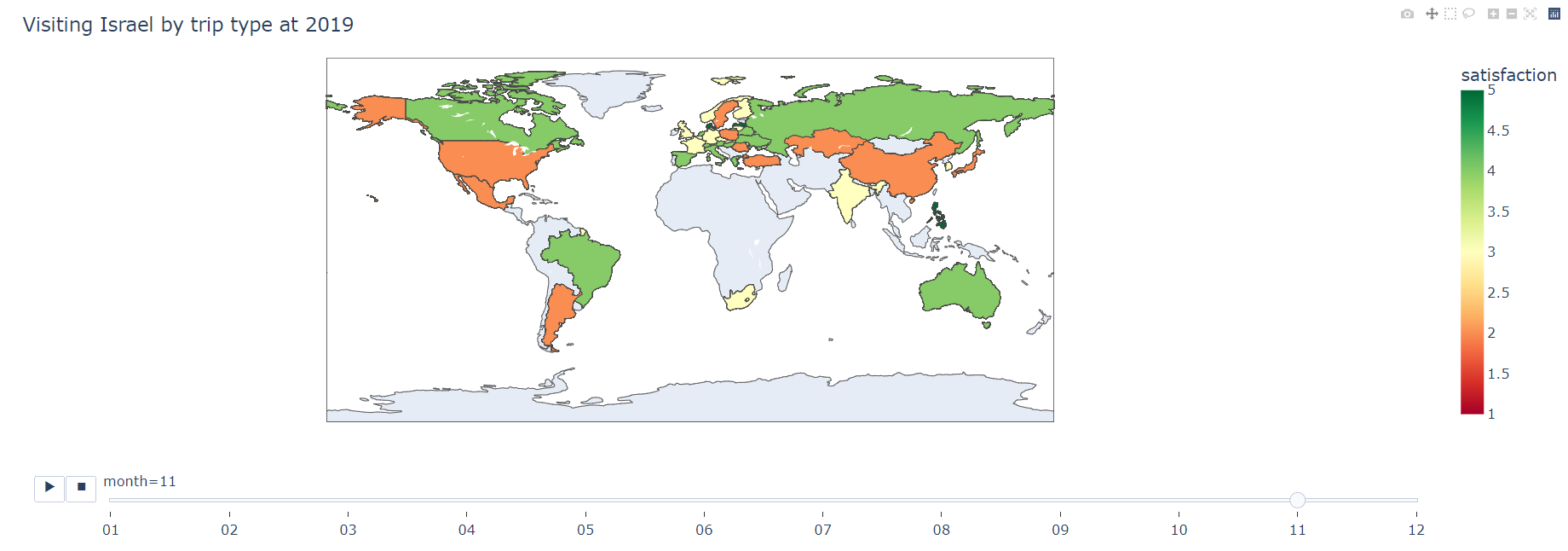
על פי הגרף נראה כי בחודש מרץ 2019, תיירים שהגיעו ממרבית מדינות העולם הביעו בממוצע שביעות רצון בינונית עד נמוכה מאוד:



בנוסף, נראה כי ביולי 2019 מרבית התיירים שהגיעו מיבשת אמריקה הביעו בממוצע גם שביעות רצון בינונית עד נמוכה:



לעומת זאת, בנובמבר 2019 תיירים שהגיעו מכמחצית ממדינות העולם הביעו שביעות רצון נמוכה:



1. מכיוון שישנו קושי להשוואת את שביעות הרצון של התיירים ממדינות שונות בחודשים השונים החלטנו ליצור מפת חום של שביעות רצון התיירים בממוצע בBins בין 0 ל-5 לפי מדינה בפילוח על פי חודשים:

תיאור הנתונים:

הנתונים שגזרנו לצורך הפקת גרף זה הינם:

* Q date – תאריך כניסה לארץ.
* Country name – שם מדינת מוצא.
* Overall satisfaction – מדד רמת שביעות רצון.

עיבודים שנעשו:

המרנו את השדות בהתאם לערכים הבאים:

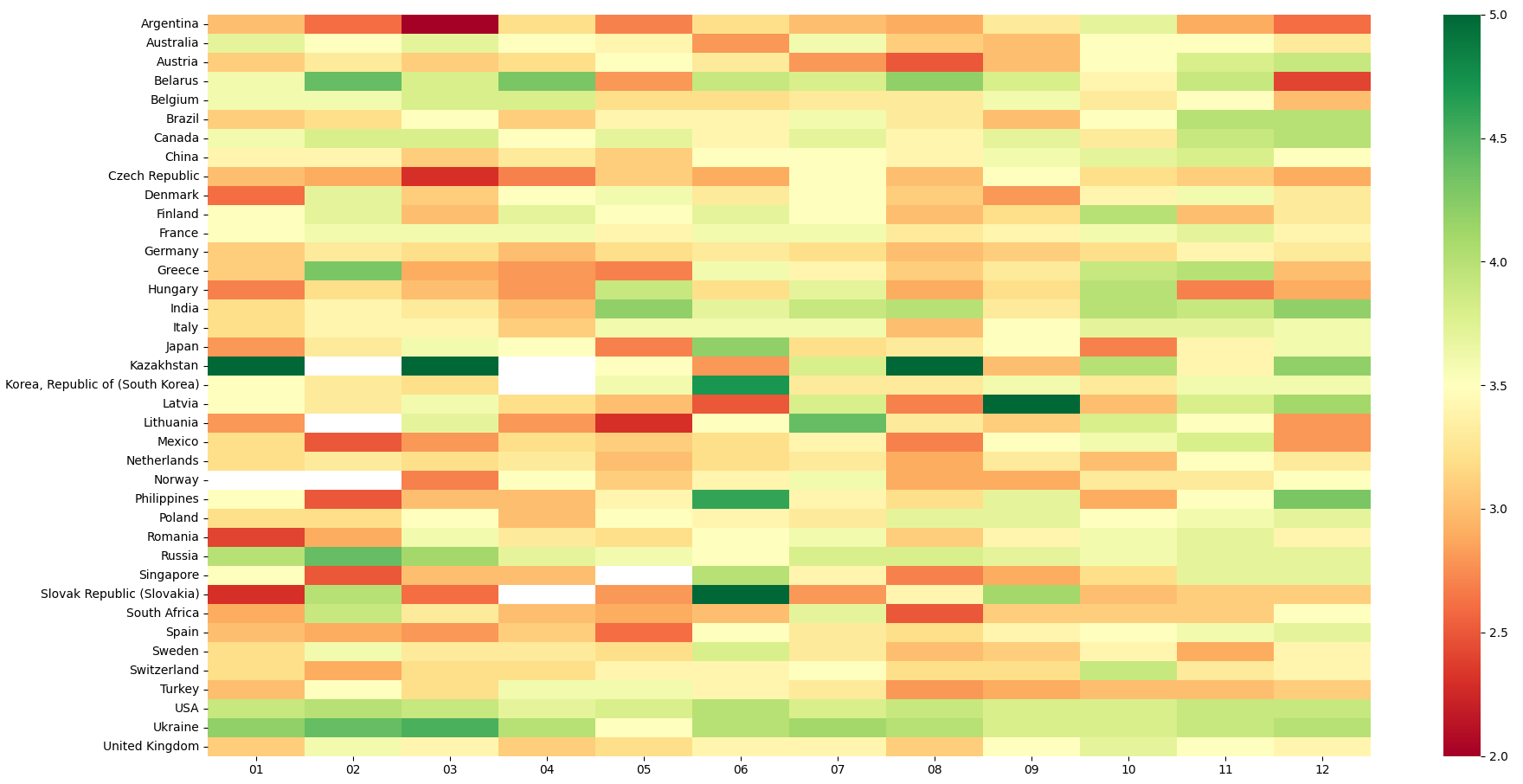
* date
* country
* satisfaction

הסרנו את כלל הרשומות הריקות בטבלה. המרנו את satisfaction לערכים מספריים (INT). שיננו את הערכים של המדינות הבאות:

* United States of America (USA) לערך – "USA"
* Russian Federation לערך – "Russia".

החלטנו להשתמש במיפוי הצבעים: " RdYlGn" מכיוון שבצורה כזאת ניתן יהיה לזהות את המדינות בהן התיירים הביעו שביעות רצון נמוכה מאשר מדינות אחרות ובכך לגזור תובנות נוספות.

לסיום, יצרנו את ה-Heatmap הבא:



ישנם 3 מימדים בגרף זה:

* ציר הx – הינו מספר החודש לשנת 2019.
* ציר הy- הינו קטגוריאלי שמכיל את מדינות העולם.
* מימד הצבע – מבטא את רמת שביעות הרצון שהביעו אותם התיירים שהגיעו למדינת ישראל ב2019.
  + צבע לבן – מופיע כאשר לא היו תיירים שמילאו את הסקרים באותו החודש.

כעת, ניתן לגזור גם תובנות רוחביות כגון:

* תיירים שהגיעו ממדינות USA, אוקריאנה הביעו את שביעות רצון גבוה בממוצע משאר המדינות לשנת 2019.
* תיירים שהגיעו ממדינות כמו ארגנטינה, צ'כיה, וסלובקיה הביעו שביעות רצון נמוכה יותר ממוצע משאר מדינות העולם לשנת 2019.

1. Radar chart:

תיאור הנתונים:

הנתונים שגזרנו לצורך הפקת גרף זה הינם:

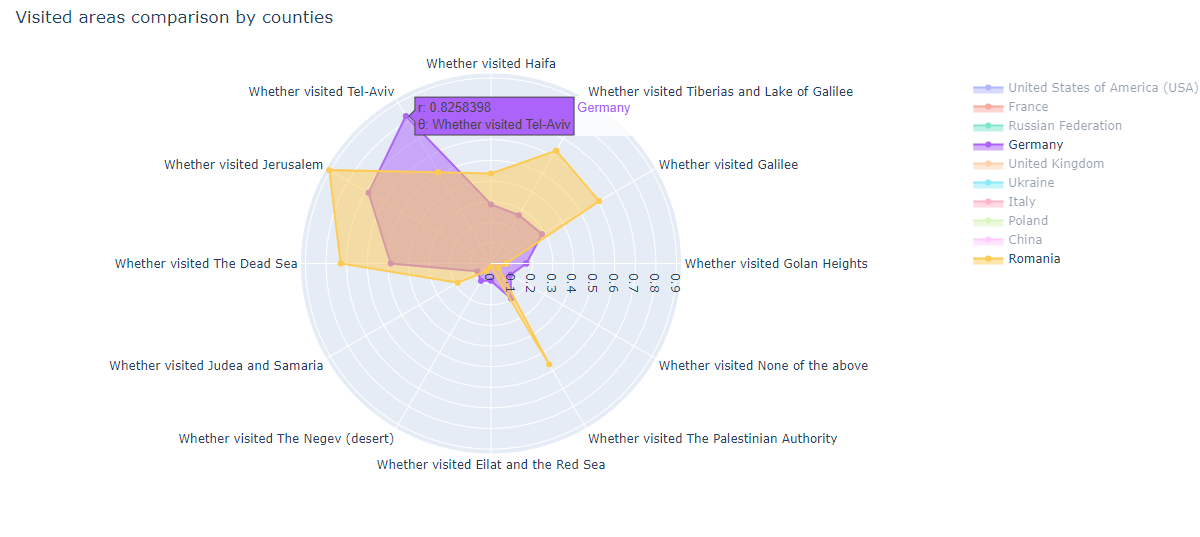
* Country name – שם מדינת מוצא.
* - Whether visited X סכ"ה כ12 אינדיקציות המתארות האם התייר ביקר (כן\לא) בשלל אזורים ברחבי הארץ:
  + Whether visited Golan Heights
  + Whether visited Galilee
  + Whether visited Tiberias and Lake of Galilee
  + Whether visited Haifa
  + Whether visited Tel-Aviv
  + Whether visited Jerusalem
  + Whether visited The Dead Sea
  + Whether visited Judea and Samaria
  + Whether visited The Negev (desert)
  + Whether visited Eilat and the Red Sea
  + Whether visited The Palestinian Authority
  + Whether visited None of the above

עיבודים שנעשו:

* בחירת 10 המדינות שמהן הגיעו הכי הרבה תיירים לישראל.
* חישוב ממוצע מענה על כל אחת מהשאלות (אינדיקציות) עבור כל אחת מהמדינות שנבחנו – טווח הערכים האפשרי שמתקבל הוא בין 0 ל1.

בחרנו להציג את הנתונים באמצעות גרף מסוג רדאר - מצד אחד על מנת לראות באופן ברור את ממוצע המענה של מדינה מסויימת עבור כל אחת מ12 השאלות, ומצד שני הגרף מאפשר לערוך השוואה בין המענה של המדינות השונות.

(שטח חופף או דומה מסמן שמדינה מסויימת מתנהגת בצורה דומה למדינה אחרת)



בגרף לעיל מוצגות 12 השאלות על פני הראדר שרדיוסו 1. ככל שצורה מסויימת תשאף להגיע למעטפת העיגול כך נוכל להסיק כי אחוז גבוה מהעונים על השאלה ממדינה מסויימת ענו "כן" – כלומר שביקרו באיזור מסויים בארץ.

בדוגמא לעיל, מוצגת השוואה בין התיירים שהגיעו מגרמניה (סגול) לעומת התיירים שהגיעו מרומניה (צהוב).

נוכל לראות כי 82.5% מהתיירים הגרמנים העידו כי ביקרו בתל אביב לעומת אחוז מופחת בקרב התיירים הרומנים (מסקנה הפוכה נקבל אם נבחן את השאלה עבור העיר ירושלים למשל).

אם יתקבלו צורות דומות (חופפות) נוכל להסיק מסקנה כללית יותר כי תיירים במדינה מסויימת מתעניינים ביעדים דומים בישראל.

**גרף 3:**

תיאור הנתונים:

הנתונים שגזרנו מתוך ה DATASET לצורך הפקת גרף זה הינם:

* אינדיקציות – האם ביצעתם את האטרציות הבאות בעת שהותכם בישראל:

• Culture (folklore, opera, concerts, festivals…) – אירועי תרבות.

• Sea sports (snorkeling, scuba diving, sailing…) – ספורט ימי.

• Desert tours – טיולי מדבר.

• Bicycle tours – טיולי אופניים.

• Birdwatching – מצפורים.

• Standard Sports (jogging, gym…) – ספורט סטנדרטי.

• Extreme sports (bungee, rapelling…) – ספורט אקסטרים.

• Daily guided tours – הדרכות יומיות (טיולי רחוב).

• Hiking – טיולים רגליים במסלולים.

• None of the above – אף אחת מהאטרקציות לעיל.

* AGE GROUP: yourself (age) – קבוצת גיל (מספר המהווה משתנה קטגוריאלי).
* religious affiliation – קבוצת דת (מספר המהווה משתנה קטגוריאלי)

העיבודים שנעשו:

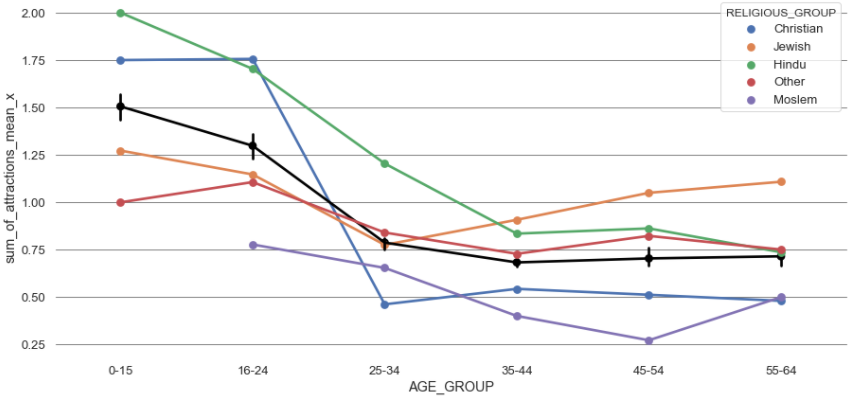
* AGE GROUP: yourself (age) – הומר ל AGE GROUP(המספרים הומרו לטקסט באמצעות טבלת מפתח)

|  |  |
| --- | --- |
| 1 | 0-15 |
| 2 | 16-24 |
| 3 | 25-34 |
| 4 | 35-44 |
| 5 | 45-54 |
| 6 | 55-64 |

* religious affiliation – הומר לRELIGIOUS\_GROUP- (המספרים הומרו לטקסט באמצעות טבלת מפתח)

|  |  |
| --- | --- |
| 1 | Christian |
| 2 | Jewish |
| 3 | Moslem |
| 4 | Hindu |
| 5 | Other |

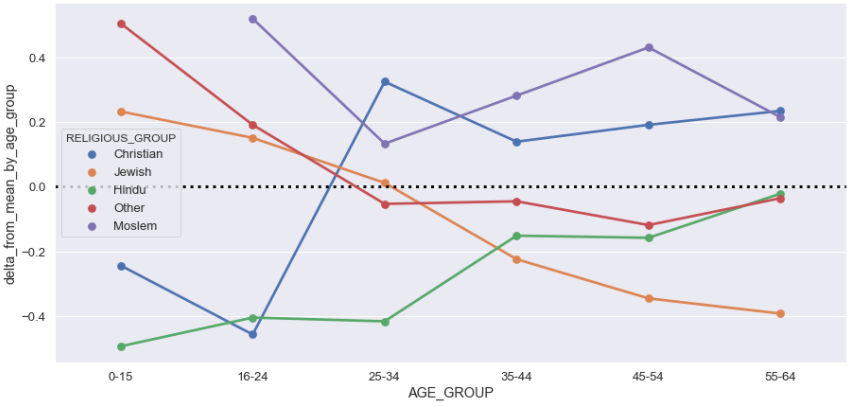
* סכמנו את כמות האינדיקציות עבור כל תייר (כלומר בכמה אטרקציות מהרשימה התייר השתתף בעת שהותו בישראל).
* עבור כל קומבינציה (group by) של השדות AGE GROUP, RELIGIOUS\_GROUP חישבנו את כמות התיירים באותו שילוב (num of people) ובנוסף את ממוצע כמות האטרקציות בהן שילוב מסויים השתתף (sum\_of\_attraction\_mean).
* חישבנו בנוסף את ממוצע כמות האטרציות לפי קבוצת גיל (ללא התייחסות לקבוצות הדת)



\*הקו השחור מציג את ממוצע כמות האטרקציות לפי קבוצת גיל בלבד.

* לאחר מכן ביצענו הפרש בין הממוצע (המיוצג ע"י הקו השחור בגרף לעיל) לבין כמות האטרציות הממוצעת לפי פילוח של קבוצות גיל וקבוצות דת.

הפרש זה מיוצג במשתנה דלתא (ציר הY) בגרף הבא:



הקו המקווקו ב Y=0 מייצג חוסר הבדל בין ממוצע קבוצת הגיל אל מול ממוצע קבוצת הגיל בשילוב עם קבוצת הדת.

מגרף זה אנו יכולים להסיק מסקנות מעניינות, לדוגמא:

* בקרב התיירים היהודים (קו כתום) בטווח הגיל שבין 25-34 – ממוצע כמות האטרקציות הייתה זהה לממוצע כמות האטרקציות של כלל קבוצת הגיל 25-34 (של כל הדתות יחד).
* כל נקודה על הגרף מעל Y=0 מסמלת עבורנו שאותה קבוצה (המורכבת מדת וקבוצת גיל) משתתפת ביותר אטרקציות מהממוצע של קבוצת הגיל התואמת, וההפך נכון עבור נקודה הנמצאת מתחת ל Y=0.
* הגרף מאפשר לדעת איזה שילוב "מתנהג" באופן חריג מקבוצת הגיל המתאימה לו – לדוגמא: בקרב היהודים בגילאים 55-64 ממוצע כמות האטרציות נמוך בהרבה מהממוצע שחושב עבור קובצת הגיל המבוגרת.
* בקרב דת האינדואיזם הפרש זה הולך ומצטמצם ככל שגיל התיירים עולה.

1. **lineplot graph**

תיאור הנתונים:

הנתונים שגזרנו לצורך הפקת גרף זה הינם:

* מטרת הביקור העיקרית - WHAT WAS THE MAIN PURPOSE OF YOUR VISIT TO ISRAEL? - משתנה קטגוריאלי.
* גובה ההוצאות במהלך הטיול בישראל - TOTAL EXPANSE - משתנה רציף.
* חודש הביקור בארץ.

העיבודים שנעשו:

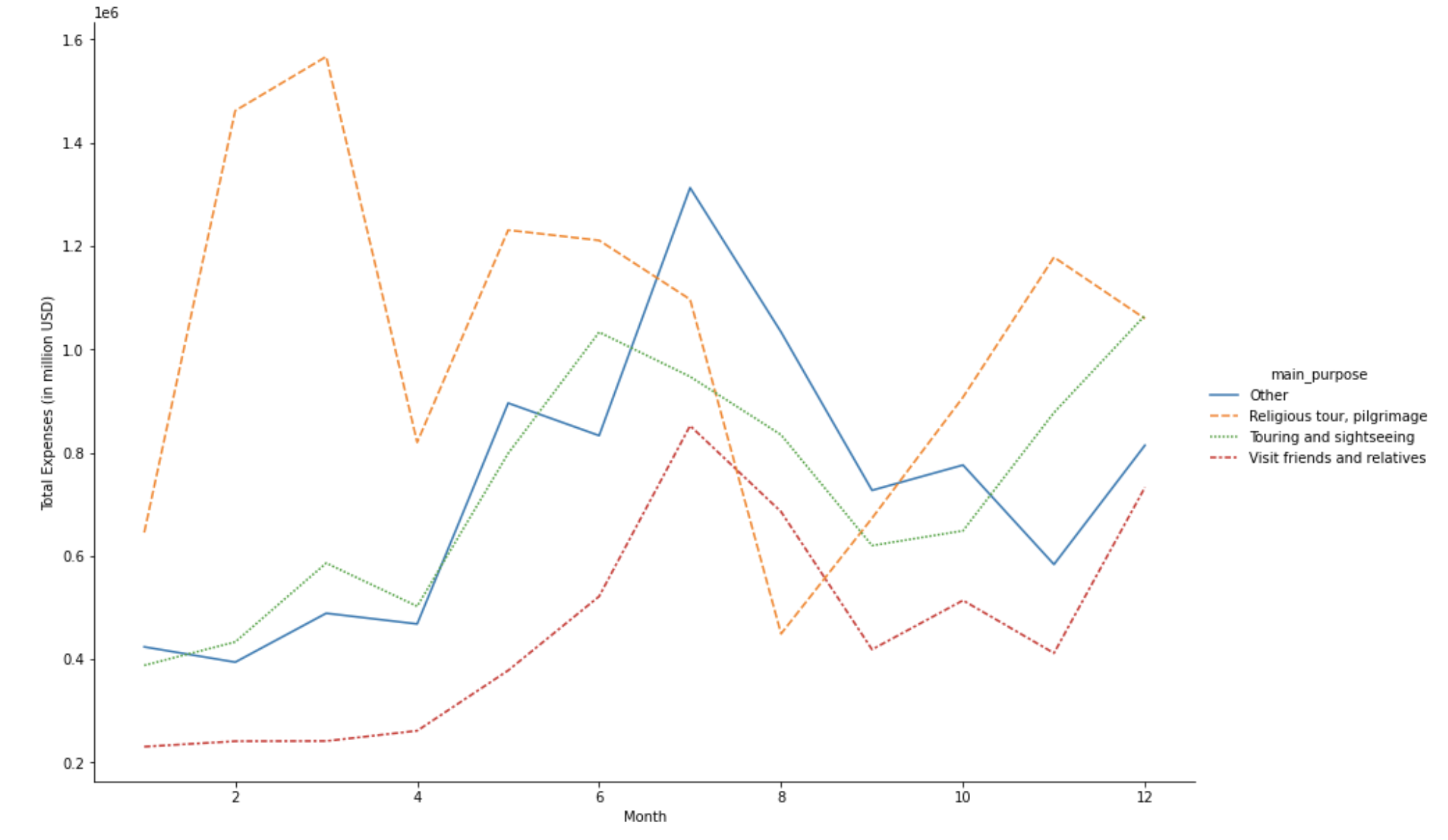
* מטרת הביקור הינה משתנה בעל 9 רמות אשר בחרנו באלו הבולטים ביותר לניתוח ולכן יצרנו איחוד של מספר קטגוריות לתוך קטגוריה ׳אחר׳

הערכים:

* Holiday, leisure, recreation
* Business
* Medical treatment
* Study and research

אוחדו יחד עם הקטגוריה other

* סכמנו את כמות ההוצאות החודשיות לפי מטרות הביקור השונות



בגרף לעיל מוצגות 3 המטרות הטיול הבולטות ביותר בניתוח חודשי של ההוצאות של כלל התיירים באותה קטגוריה.

מסקנות העולות מן הגרף:

* כמות ההוצאות הגדולה ביותר של תיירים בישראל היא באופן בולט מאוד ע״י קבוצת תיירות הדת.
* ההוצאות של התיירים ברבעון הראשון של השנה עולות בכמעט 30% על תקופת הקיץ שידועה כתקופה אטרקטיבית מבחינה תיירותית.
* בהסתכלות כוללת ניתן להבין כי ישנן שתי תקופות אטרקטיביות מבחינה תיירותית והן תקופת הקיץ וסוף השנה הקלנדרית.

1. **Chord Graph**

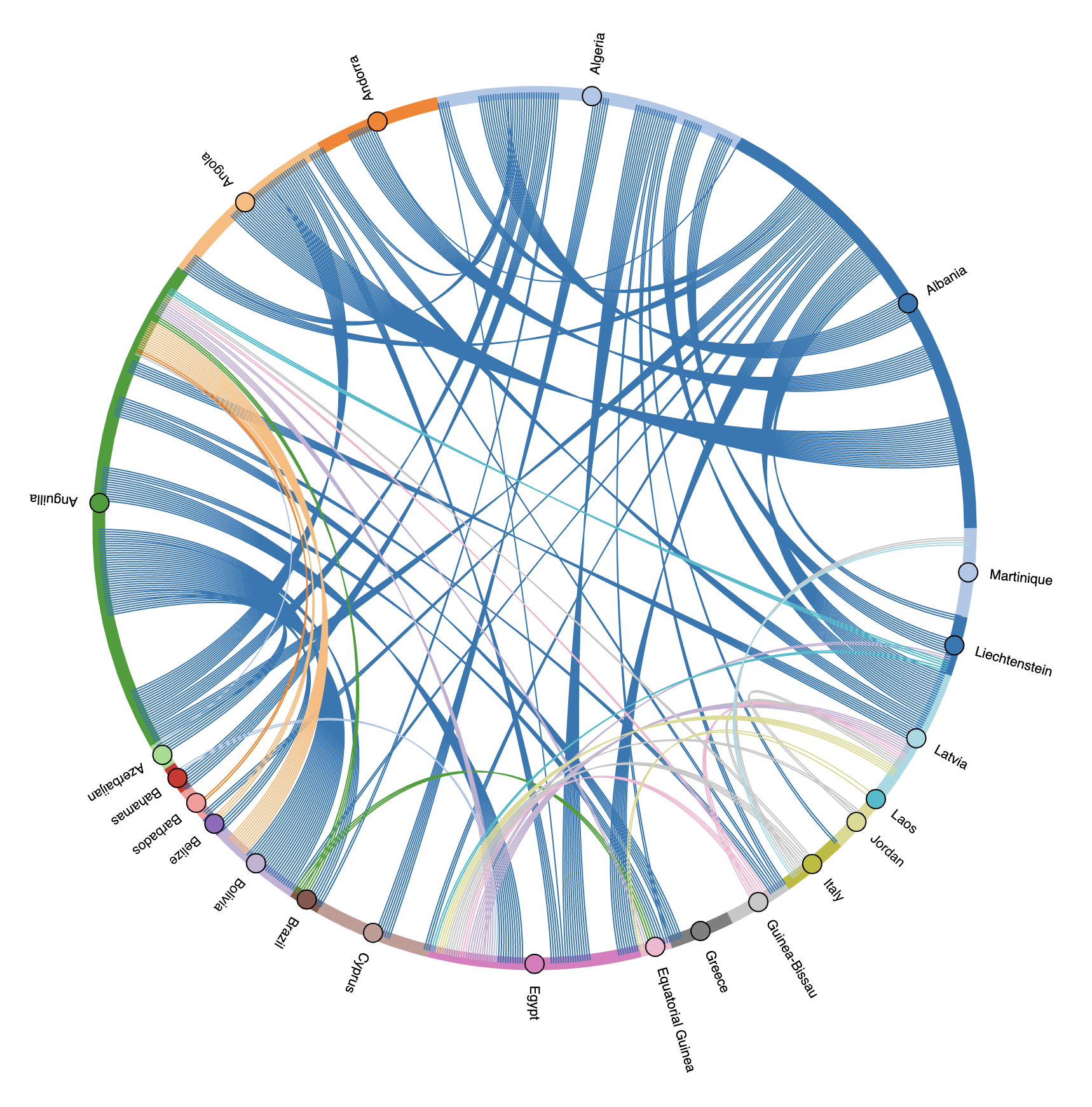
תיאור הנתונים:

הנתונים שגזרנו לצורך הפקת גרף זה הינם:

* ארץ המוצא של התייר.
* המדינה הראשונה אותה ביקר התייר כשיצא ממדינתו לראשונה.

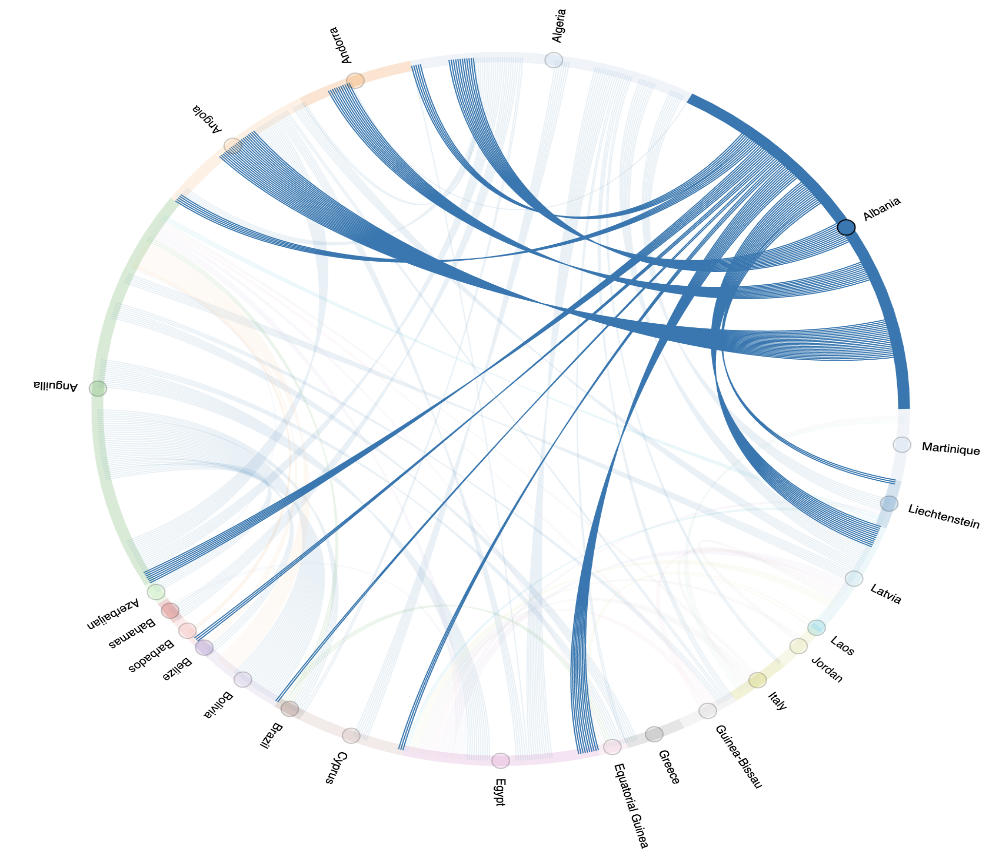
העיבודים שנעשו:

* סכימת הצירופים הקיימים ב-dataset של ארץ המוצא ומדינת היעד הראשונה
* סינון רק צירופים שמופיעים יותר מ-20 פעמים על מנת לאפשר תחקור



בגרף לעיל ניתן לראות על היקף הגרף את המדינות המופיעות לאחר העיבוד והקשתות העוברות ביניהם נותנות פרופורציה לכמה מאותה מדינה טיילו למדינת היעד וכמה הם תפסו מתוך התיירים לאותה מדינה מתוך המדינות המופיעות כאן.

גרף זה יכול לשמש את משרד התיירות לחיזוק קשרים עם מדינות אסטרטגיות כמו מדינות שלראשונה מגיעים לטייל במדינות האיזור כמו מצרים, יוון וקפריסין כדי לשכנע יותר מהם להגיע לישראל לראשונה או כחלק מטיול במדינות הללו.



מסקנות העולות מן הגרף:

* במשחק עם הגרף לעיל ניתן לתחקר עבור מדינה ספיציפית לאן האזרחים שלה לרוב מטיילים לראשונה, לדוגמא:

רוב האזרחים מאלבניה שטיילו בישראל טיילו לראשונה באנגולה ואלג׳יריה.

מסקנות נוספות מתחקורים אחרים:

* הרבה מאזרחי אלג׳יריה טיילו במצרים ובאלבניה לראשונה.
* מתוך התיירים שנשאלו רק שתי מדינות ענו שישראל הייתה המדינה הראשונה שלהם והם אזרבייג׳ן ואגווליה