ויזואליזציה של מידע

נתוני הזמנות של טיסות ומלונות באתר KAYAK

זהר קרני | 205705262

מאי רגב | 205359706

עדי בולגרו |

תוכן עניינים

[תיאור של הנתונים והטיפול בהם 3](#_Toc93065801)

[דיון בעבודות קשורות ובוויזואליזציות קיימות לאותם נתונים 5](#_Toc93065802)

[ניתוח תיאורטי של הוויזואליזציות (מודל WWH), הסברים וצילומי מסך 7](#_Toc93065803)

[גרף 1 אינטראקטיבי – Revenue per Country and Trends by Quarterly 7](#_Toc93065804)

[גרף 2 אינטראקטיבי – Profit by World Areas and Countries 8](#_Toc93065805)

[גרף 3 אינטראקטיבי – Number of Returned Customers vs New Customers 9](#_Toc93065806)

[גרף 4 – Orders by Group Age and World Ares 10](#_Toc93065807)

[גרף 5 – Distribution of Temperature by Vacation Type 11](#_Toc93065808)

[גרף 6 – Box-plot of Cost per Traveler by Vacation Type and Gender 12](#_Toc93065809)

[קוד מלא ב-R 13](#_Toc93065810)

[איור 1- דוגמה ראשונה לגרף קיים 5](file:///C:\Users\karni\Documents\'שנה%20ה\ויזואליזציה%20של%20מידע\פרויקט\פרויקט.docx#_Toc93157105)

[איור 2- דוגמה שנייה לגרף קיים 5](file:///C:\Users\karni\Documents\'שנה%20ה\ויזואליזציה%20של%20מידע\פרויקט\פרויקט.docx#_Toc93157106)

[איור 3- דוגמה שלישית לגרף קיים 6](file:///C:\Users\karni\Documents\'שנה%20ה\ויזואליזציה%20של%20מידע\פרויקט\פרויקט.docx#_Toc93157107)

[איור 4- גרף 1 אינטראקטיבי - Revenue per Country and Trends by Quarterly 7](file:///C:\Users\karni\Documents\'שנה%20ה\ויזואליזציה%20של%20מידע\פרויקט\פרויקט.docx#_Toc93157108)

[איור 5- גרף 2 אינטראקטיבי - Profit by World Areas and Countries 8](file:///C:\Users\karni\Documents\'שנה%20ה\ויזואליזציה%20של%20מידע\פרויקט\פרויקט.docx#_Toc93157109)

[איור 6- גרף 3 אינטראקטיבי - Number of Returned Customers vs New Customers 9](file:///C:\Users\karni\Documents\'שנה%20ה\ויזואליזציה%20של%20מידע\פרויקט\פרויקט.docx#_Toc93157110)

[איור 7- גרף 4 - Orders by Group Age and World Ares 10](file:///C:\Users\karni\Documents\'שנה%20ה\ויזואליזציה%20של%20מידע\פרויקט\פרויקט.docx#_Toc93157111)

[איור 8- גרף 5 - Distribution of Temperature by Vacation Type 11](file:///C:\Users\karni\Documents\'שנה%20ה\ויזואליזציה%20של%20מידע\פרויקט\פרויקט.docx#_Toc93157112)

[איור 9- גרף 6 - Box-plot of Cost per Traveler by Vacation Type and Gender 12](file:///C:\Users\karni\Documents\'שנה%20ה\ויזואליזציה%20של%20מידע\פרויקט\פרויקט.docx#_Toc93157113)

# תיאור של הנתונים והטיפול בהם

אתר KAYAK הוא מנוע לחיפוש טיסות, מלונות, רכב וחבילות נופש המשווה מחירים ממאות חברות תעופה ואתרי תיירות או סוכנויות ברחבי העולם. האתר פעיל משנת 2004 וכיום הוא נגיש במחשבים, בטלפונים החכמים ובאפליקציה. הוא מקשר את לקוחותיו אל חברות תעופה, חברות נסיעות, מלונות וספקי רכב, ודרכם מקבל אחוזי מהעסקאות שמתבצעות.

עיקר הנתונים שלנו הם אלה שמגיעים מתוך מערכת המידע של אתר החברה ונוגעים להזמנות שלקוחות ביצעו של טיסות ומלונות. הנתונים עליהם התבססנו לקוחים מהשנים 2009-2018. הנתונים הופקו מתוך מחולל נתונים בהתבסס על השדות הרלוונטיים לכל טבלה, ויש לנו 4,091 רשומות.

הנתונים כוללים:

1. מידע על הלקוח: אימייל, מגדר, תאריך יום הולדת, פרטי כרטיס אשראי והאם הלקוח חוזר (כבר ביצע הזמנה באתר).
2. מידע על ההזמנה של הלקוח: מזהה של ההזמנה, תאריך ביצועה ומספר חבילה.
3. מידע על החופשה: תאריך התחלה, תאריך סיום, מזהה של המלון, מספר לילות, מספר נוסעים ועלות לנוסע.
4. מידע על הטיסות: מספר הטיסה בהלוך, חברת התעופה בהלוך, מספר טיסה בחזור וחברת התעופה בחזור.
5. מידע על המלון: שם המלון, העיר, המדינה והיבשת שבהן הוא נמצא, מספר כוכבים, דירוגו באתר, סגנונו והאם יש בו Wi-Fi חופשי.

לשקול להמיר לטבלה:

|  |  |  |
| --- | --- | --- |
|  | שדה | סוג |
| מידע על הלקוח | אימייל |  |
| מגדר |  |
| תאריך יום הולדת | תאריך |
| פרטי כרטיס אשראי |  |
| לקוח חוזר? | בוליאני |
| מידע על ההזמנה | מזהה של ההזמנה | מספר שלם |
| תאריך ביצוע | תאריך |
| מספר חבילה | מספר שלם |
| מידע על החופשה | תאריך התחלה | תאריך |
| תאריך סיום | תאריך |
| מזהה של המלון | מספר שלם |
| מספר לילות | מספר שלם |
| מספר נוסעים | מספר שלם |
| עלות לנוסע | מספר שלם |
| מידע על הטיסות | מספר טיסה בהלוך |  |
| חברת התעופה בהלוך |  |
| מספר טיסה בחזור |  |
| חברת התעופה בחזור |  |
| מידע על המלון | שם המלון |  |
| עיר |  |
| מדינה |  |
| יבשת |  |
| מספר כוכבים | מספר שלם |
| דירוג באתר | מספר עשרוני |
| סגנון המלון |  |
| יש Wi-Fi חופשי? | בוליאני |

כל תא ריק צריך להיות varchar- איך לרשום את זה?

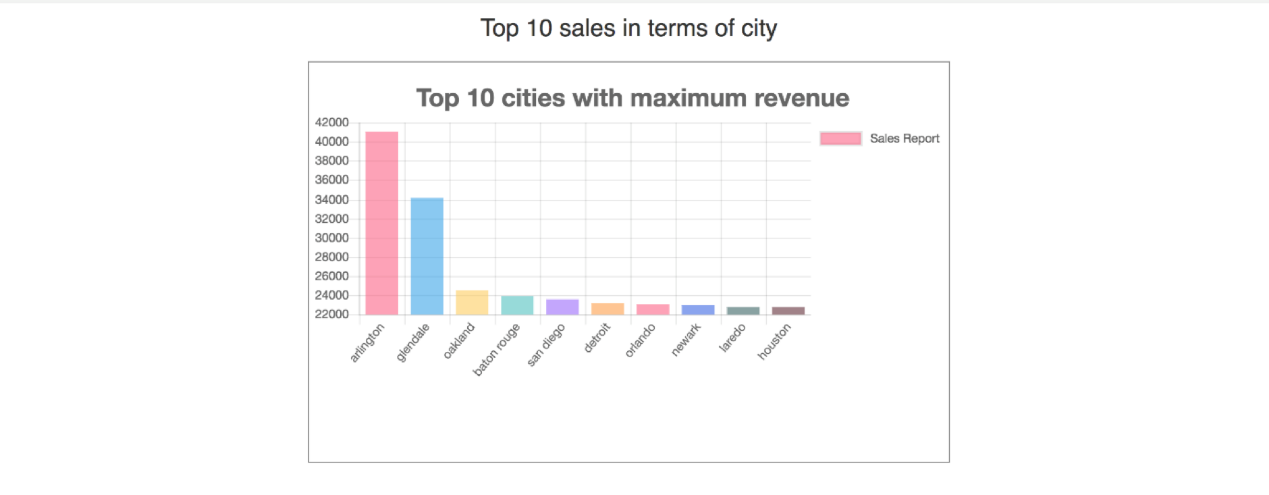
טיפול בנתונים:

1. הוספנו שדה של גיל: מחושב על ידי חיסור התאריך של היום מתאריך יום ההולדת של הלקוח.
2. הוספנו שדה קטגוריאלי של קבוצת גיל, על ידי חלוקת הגילים לשש קטגוריות: עד גיל 25, 25-35, 35-45, 45-55, 55-65, 65 ומעלה. הוספנו את השדה כדי שיהיה לנו יותר קל להציג ויזואליזציות שקשורות לגילי הלקוחות.
3. הוספנו שדה של הכנסה (revenue): מכפלה של עלות לנוסע במספר הנוסעים.
4. הוספנו שדה של הכנסה נטו (profit): אחוז מסוים מההכנסה מהעסקה שאותו החברה קיאק מרוויחה – האחוז משתנה כתלות בחודש בשנה שבו בוצעה העסקה. הוספנו את השדות של הכנסה והכנסה נטו כדי שנוכל להציג גרפים שנוגעים לרווחיות החברה ולכן מביאים לה ערך.
5. הוספנו שדה של טמפרטורה: מידע שנלקח ממקור חיצוני על הטמפרטורה הממוצעת בכל חודש בכל עיר מרכזית בעולם. מהמידע הזה יצרנו שדה על הטמפרטורה שצפויה להיות בחופשה שהלקוח הזמין. עניין אותנו להוסיף את שדה זה כדי לבחון את השפעת מזג האוויר על ההזמנות והרווחים.

# דיון בעבודות קשורות ובוויזואליזציות קיימות לאותם נתונים

מחיפוש באינטרנט מצאנו ב-GitHub אב טיפוס ללוח מחוונים לאתר KAYAK שיצר אדם פרטי. לוח המחוונים כולל: פלטפורמה לחיפוש והזמנת טיסות, מלונות ומכוניות, פאנל למנהל שבו הוא יכול להוסיף, לערוך ולמחוק טיסות, מלונות ומכוניות, גרף למעקב אחר פעילות משתמש וכן שני גרפים לניטור מכירות טיסות, מלונות ומכוניות.

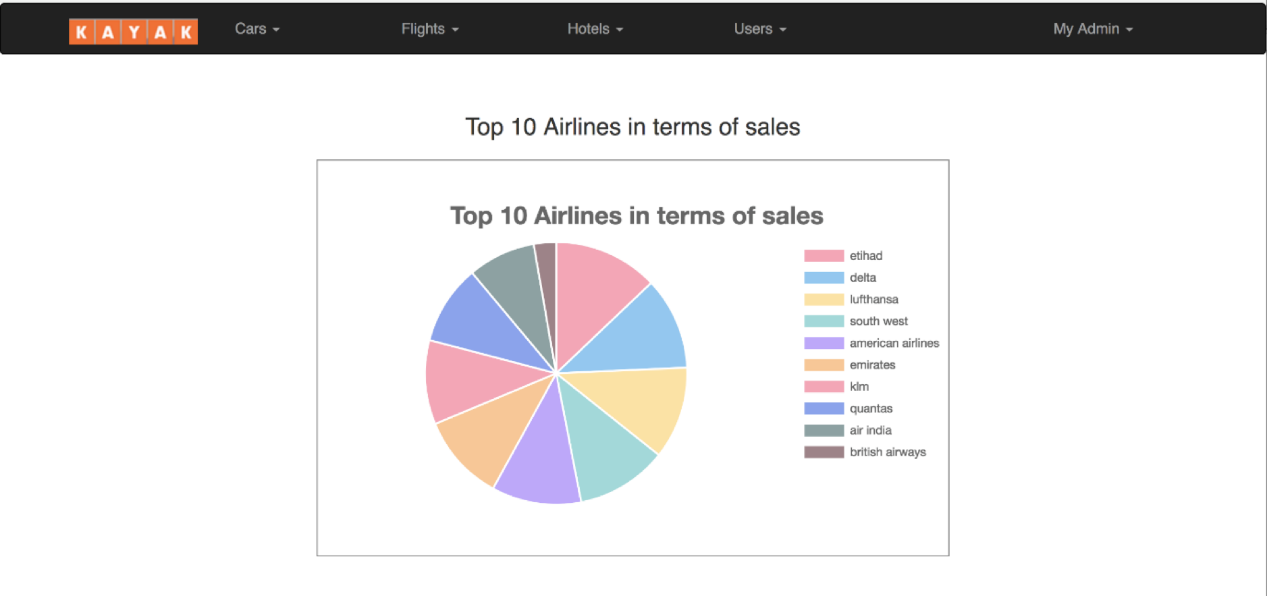
אנחנו נתמקד בגרפי הניטור. **הגרף הראשון** שהוא הציג היה "10 הערים עם ההכנסה הגבוהה ביותר", כאשר ההכנסה נמדדת לפי הזמנות למלונות שמשתמשי האתר ביצעו. לדעתנו הגרף הזה לא מספיק טוב. ראשית, לא ברורים המספרים בצד שמאל: הם באלפים? בדולרים? קנה המידה לא מתאים לכל העמודות – חוץ משתי העמודות השמאליות ביותר, ההכנסה בשאר העמודות היא נמוכה יחסית וקשה להעריך על כמה היא עומדת. לא ברור גם למה מופיע בצד ימין Sales Report.



איור 1- דוגמה ראשונה לגרף קיים

אחרי שנחשפנו לגרף זה החלטנו שכדאי להציג את המידע ברזולוציה נמוכה יותר, ועל כן אנחנו ביצענו את ההשוואה של ההכנסות לפי אזורים בעולם ב-treemap (). לכל אזור יש מלבן ובגלל שהגרף הוא אינטראקטיבי ניתן לעשות "drill down" ולקבל מידע על ההכנסות מהמדינות שבאזור מסוים.

**גרף הניטור השני** שנמצא באב הטיפוס של לוח המחוונים הוא "10 חברות התעופה המובילות במונחים של מכירות". לדעתנו גם גרף זה לא מספיק טוב. לא נכון להציג את החברות המובילות בגרף עוגה, בייחוד כשאין שדה נוסף לחברות האחרות שלא נמצאות בעשר המובילות. חסרים גם אחוזים על העוגה, ונראה שכמעט כל חברות התעופה תופסות נתח זהה מהמכירות.



איור 2- דוגמה שנייה לגרף קיים

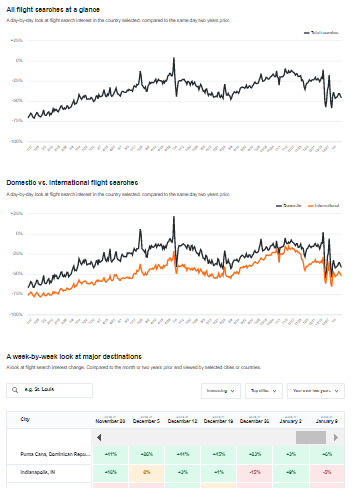
קישור לאתר: <https://github.com/akashgupta2903/Kayak-Prototype>

ויזואליזציה נוספת שמצאנו באינטרנט שנוגעת לנתונים מאתר KAYAK היא **מרכז מידע** שעוסק במגמות חיפוש טיסות גלובליות. את מרכז המידע הזה יצרה החברה בעצמה בניסיון לשפוך אור על ההשפעה של מגפת הקורונה על תעשיית הנסיעות. לפי אתר החברה, הם השתמשו בנתונים של יותר ממיליארד חיפושי טיסות, בחלוקה לפי מדינות המוצא.

במערכת שהם יצרו ניתן לבחור מדינה ומקבלים עבורה שלושה גרפים: קו מגמה שמציג את כל חיפושי הטיסות שבוצעו בה, קו מגמה נוסף שמציג את חיפושי הטיסות בחלוקה לטיסות פנימיות וטיסות בינלאומיות וטבלה שמציגה מבט של שבוע אחר שבוע על מגמות החיפוש לפי ערי יעד מרכזיות. הנתונים שבשלושת הגרפים מציגים את כמות החיפושים ביום מסוים בהשוואה לאותו יום שנתיים קודם לכן. נכון לשנת 2021, הנתונים מתעדכנים מדי יום.

אנחנו חושבות שהרעיון להקים מרכז מידע כזה הוא מעניין, מאחר שמסקרן לראות את השפעת הקורונה וללמוד עליה. אך איננו בטוחות שהדרך בה יצרו אותו היא מיטבית. תחילה, לא ברור האם נתוני החיפוש זה מה שמעניין אותנו בכלל – זה מעיד על משהו? אולי כשיש סגר אנשים נוטים לפנטז על היום שאחרי ונהנים לתכנן את החופשה שייצאו אליה כשהמצב יירגע. אולי דווקא בסגר אנשים מרגישים שמיותר לבדוק טיסות כשברור שזה לא ריאלי לצאת אליהן בזמן הקרוב. לכן אנחנו לא בטוחות אם הנתון של חיפושים הוא הנתון שבו נכון להתמקד כשרוצים לבחון את השפעת המגפה על עולם התעופה.

איור 3- דוגמה שלישית לגרף קיים



כמו כן, החברה מציגה את עצמה כ"מנוע חיפוש הנסיעות המוביל בעולם", אז כנראה שיש להם עוד המון נתונים שחבל שהם לא מציגים גם אותם. היה אפשר לנצל את כמות הנתונים הגדולה שבידי החברה כדי לייצר מבט מקיף על השפעת המגפה, שלא מתייחס רק לחיפושים שהלקוחות ביצעו.

מאגר הנתונים שאיתו אנחנו עבדנו לא כולל את תקופת הקורונה (תאריך ההזמנה האחרון הוא 13.12.2019) ולכן לא יצרנו גרפים שמתייחסים אליה.

קישור לאתר: <https://www.kayak.com/flight-trends>

# ניתוח תיאורטי של הוויזואליזציות (מודל WWH), הסברים וצילומי מסך

## גרף 1 אינטראקטיבי – Revenue per Country and Trends by Quarterly

בגרף זה אנו מציגות לוח מחוונים שבו בחירה של מדינה ממפת העולם משפיעה על שני גרפים רציפים (line plots) המציגים רווח (Revenue) ורווח נקי (Profit) בחלוקה לפי רבעונים ושנים. כל אחד מהגרפים מאופיין כך:

WHAT

* טבלה רב מימדית בעלת שדה אחד כמותי: רווחים בדולרים (Revenue או Profit).
* ושלושה שדות קטגוריאליים: שנה ורבעון (שנגזרים מתאריך ביצוע ההזמנה) ומדינת היעד של החופשה. השדה של השנים הוא משתנה בעל סדר.

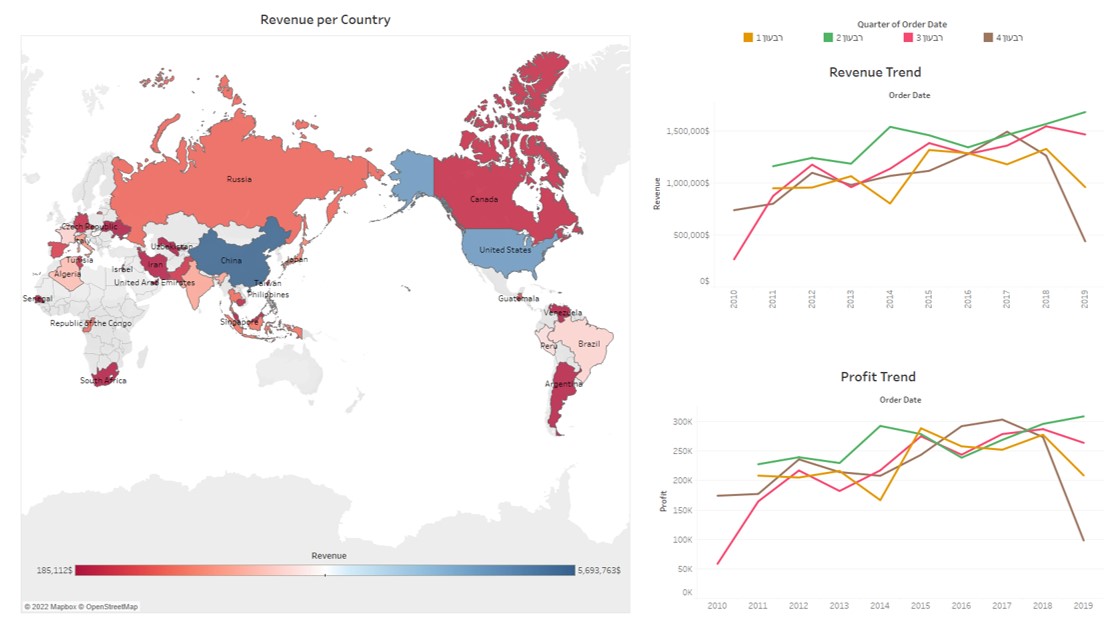
WHY

בעזרת גרפים אלו נוכל להראות את מגמת השתנות הרווחים במהלך השנים בחלוקה לרבעונים ובהסתכלות על מדינות שונות.

HOW

בכל גרף קיימים ארבעה קווים כך שכל אחד מהם מייצג רבעון. הקווים מחברים בין נקודות שמייצגות רווחים בכל שנה. השנים **מסודרות** בצורה כרונולוגית. בעזרת גרף המפה נוכל **לבחור** מדינה והבחירה **תשנה** את הנתונים שמוצגים ב-line charts.

לדעתנו, להצגת גרפים שקשורים לרווחיות החברה יש ערך משמעותי, ולכן יצרנו את לוח המחוונים הזה. צבע המדינה תלוי בערך ה-revenue שלה – בסך הרווחים מהזמנות ליעדים במדינה הזו. אפשר לסמן במקרא למעלה את הרבעון שמעניין אותנו וקווי המגמה שלו יובלטו על פני השאר. אפשר לבחור מדינה מסוימת ואז נראה רק את הנתונים שלה. ב- line plotהעליון רואים את ההכנסה על פני השנים ובתחתון את ההכנסה נטו, שהיא חלק יחסי מההכנסה שחברת קיאק מרוויחה.



איור 4- גרף 1 אינטראקטיבי - Revenue per Country and Trends by Quarterly

## גרף 2 אינטראקטיבי – Profit by World Areas and Countries

WHAT

* טבלה רב מימדית בעלת שדה אחד כמותי: רווחים בדולרים.
* ושני שדות קטגוריאליים: אזורים בעולם ומדינות של יעדי החופשות.

WHY

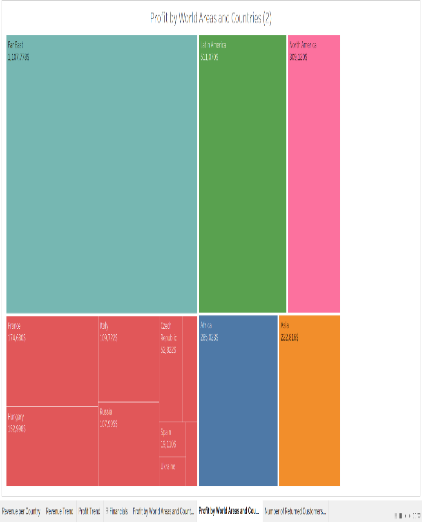
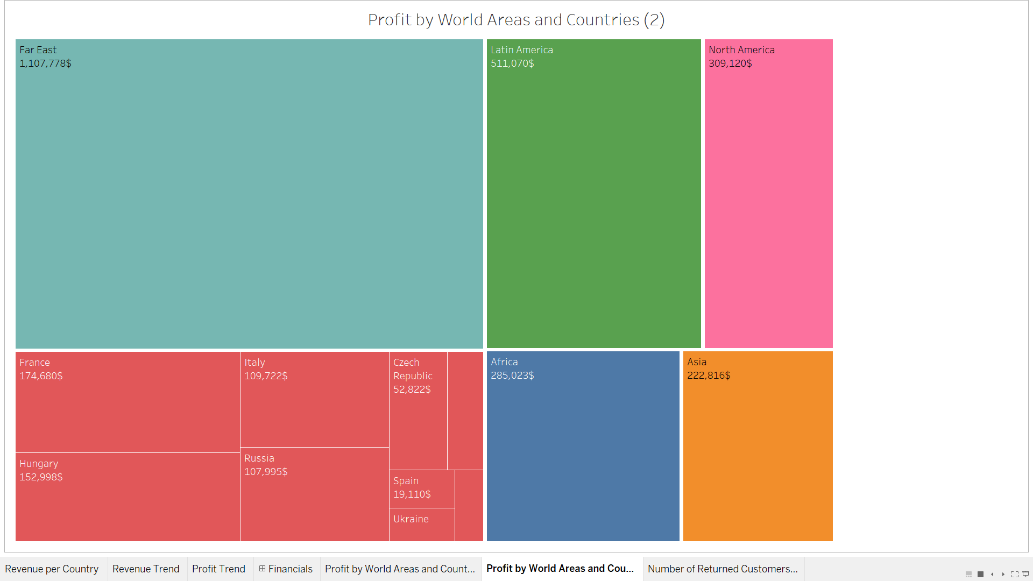
השוואה בין ערכים ובין פרופורציות בתוך היררכיה, בגרף המוצג נוכל להשוות תחילה בין הרווח הנקי בכל אזור בעולם ולאחר מכן נוכל לבצע drill down ולהציג את הרווח הנקי בכל מדינה לפי האזורים.

HOW

נשווה את הרווח הנקי בעזרת treemap – מלבנים גדולים של האזורים בעולם שמורכבים ממלבנים קטנים יותר של המדינות השונות. נאפשר לבצע מניפולציה של **בחירת** אזור מסוים בעולם כדי **לראות** את המדינות הקיימות בו ואת נתוני הרווח שמתאימים להן.

בגרף זה רצינו להראות את הרווחים (הכנסה נטו) לפי אזורים בעולם, כלומר את סך הרווחים מהזמנות ליעדים שבכל אזור. מכיוון שניתן להסתכל גם על המדינות בכל אזור בעולם ולהסיק נתונים נוספים רצינו להציג מידע גם עבור כל אזור בצורה שלא תכביד על הגרף וכתוצאה מכך על הקורא. לכן נעזרנו בהיררכיה ויצרנו drill down המציג חלוקה לפי מדינות בכל אזור בעולם. ברגע שלוחצים על מלבן של אזור מופיעים מלבנים קטנים יותר של המדינות שבאותו האזור, שמרכיבים את המלבן הגדול. עבור כל מדינה ואזור בעולם נוכל לראות את הרווחים מהזמנות שנעשו באתר. על ידי מעבר עם העכבר על המלבנים נוכל לראות את המידע הרלוונטי – את שם המדינה (גם אם בתחילה לא רואים אותו כי המלבן קטן מדי) ואת הרווחים בדולרים. בנתונים שלנו אין הרבה מאוד מדינות תחת כל אזור בעולם, אך אם היו היינו שוקלות להציג רק את חלקן ולהוסיף מלבן קטן של "מדינות נוספות".

אפשר לעשות גם שהמדינות ייצבעו בצבע לפי ההכנסה**:** [**https://help.tableau.com/current/pro/desktop/en-us/buildexamples\_treemap.htm**](https://help.tableau.com/current/pro/desktop/en-us/buildexamples_treemap.htm)



איור 5- גרף 2 אינטראקטיבי - Profit by World Areas and Countries

גרף 3 אינטראקטיבי –Number of Returned Customers vs New Customers

WHAT

* טבלה עם שני שדות קטגוריאליים: תקופת זמן (שנה/רבעון/חודש) שבה בוצעה ההזמנה ושדה בוליאני שמתאר האם הלקוח שביצע את ההזמנה הוא לקוח חוזר או לא (כלומר חדש).
* ושדה כמותי שסוכם לפי תקופת הזמן את כמות הלקוחות החוזרים ואת כמות הלקוחות החדשים.

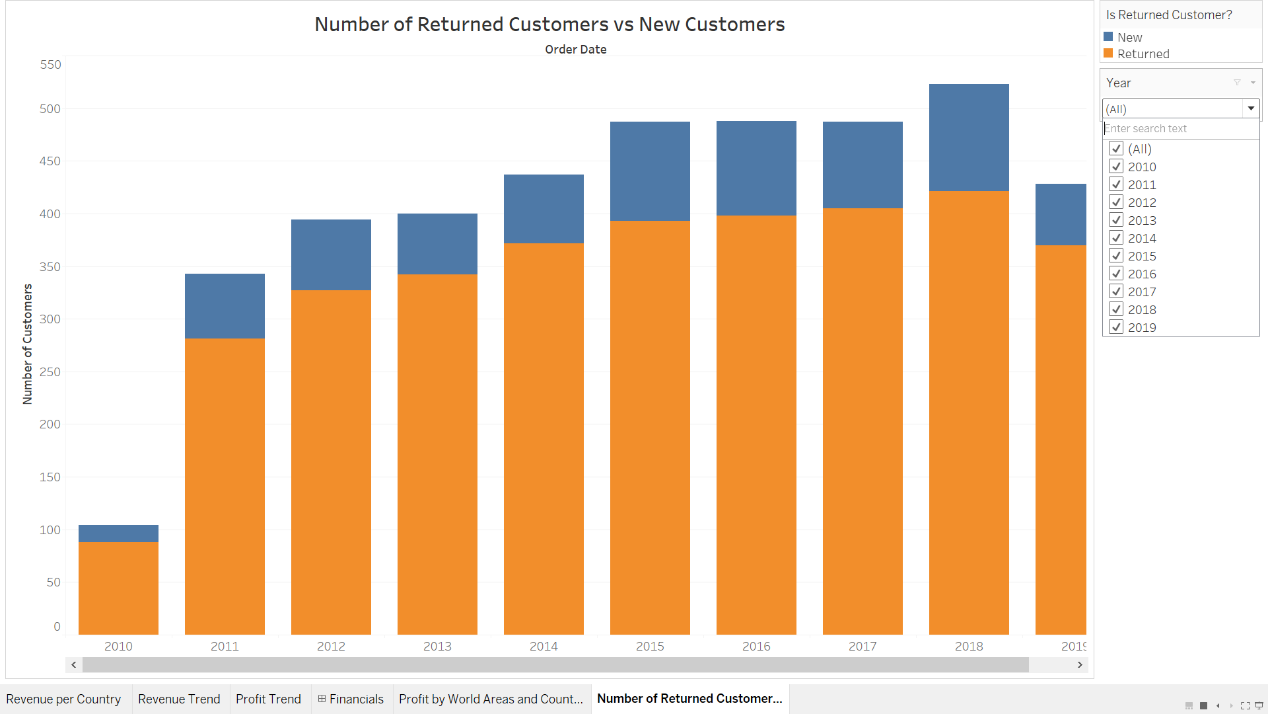
WHY

השוואה בין ערכים, בחינת מגמות במהלך תקופות הזמן השונות והסתכלות על חלק מהשלם.

HOW

בגרף זה נשווה את כמות הלקוחות החוזרים/החדשים בעזרת עמודות מאונכות ומיושרות המופרדות לפי תקופות זמן שונות. תקופות הזמן **מסודרות** בצורה כרונולוגית.

בגרף זה רצינו להראות את כמות הלקוחות החוזרים לעומת הלקוחות החדשים בחלוקה לשנים, רבעונים וחודשים. הגרף נעשה מכמה סיבות. ראשית, רצינו לראות את מגמת כמות הלקוחות הכוללת שמבצעים הזמנות באתר. השאיפה היא כמובן שמדי שנה (או רבעון/חודש) כמות הלקוחות תגדל. את מידע זה מספקים לנו גבהי העמודות. מידע נוסף שרצינו לקבל הוא על התפלגות הלקוחות – כמה מתוכם חוזרים וכמה חדשים, מדי כל תקופת זמן. כיוון שהעמודה היאstacked ומורכבת משתי קטגוריות, קל לראות את המידע הזה ולבצע השוואות. אפשר לבחון האם כמות הלקוחות החדשים גדלה מדי שנה (או רבעון/חודש) וכן האם חלקה לעומת הלקוחות החוזרים גדל מדי שנה. אפשר במקרא בצד ימין לסמן את הקבוצה הרצויה והיא תובלט בגרף. על ידי לחיצה על הפלוס שעולה בראשית הצירים ניתן לעשות drill down לרמות של רבעונים/חודשים, ולחזור ממנו על ידי לחיצה על מינוס שיעלה למעלה משמאל לתוויות תקופות הזמן.



איור 6- גרף 3 אינטראקטיבי - Number of Returned Customers vs New Customers

## גרף 4 – Orders by Group Age and World Ares

WHAT

* טבלה רב ממדית בעלת שדה אחד כמותי: כמות הזמנות.
* ושני שדות קטגוריאליים: קבוצת גיל ואזור בעולם.

WHY

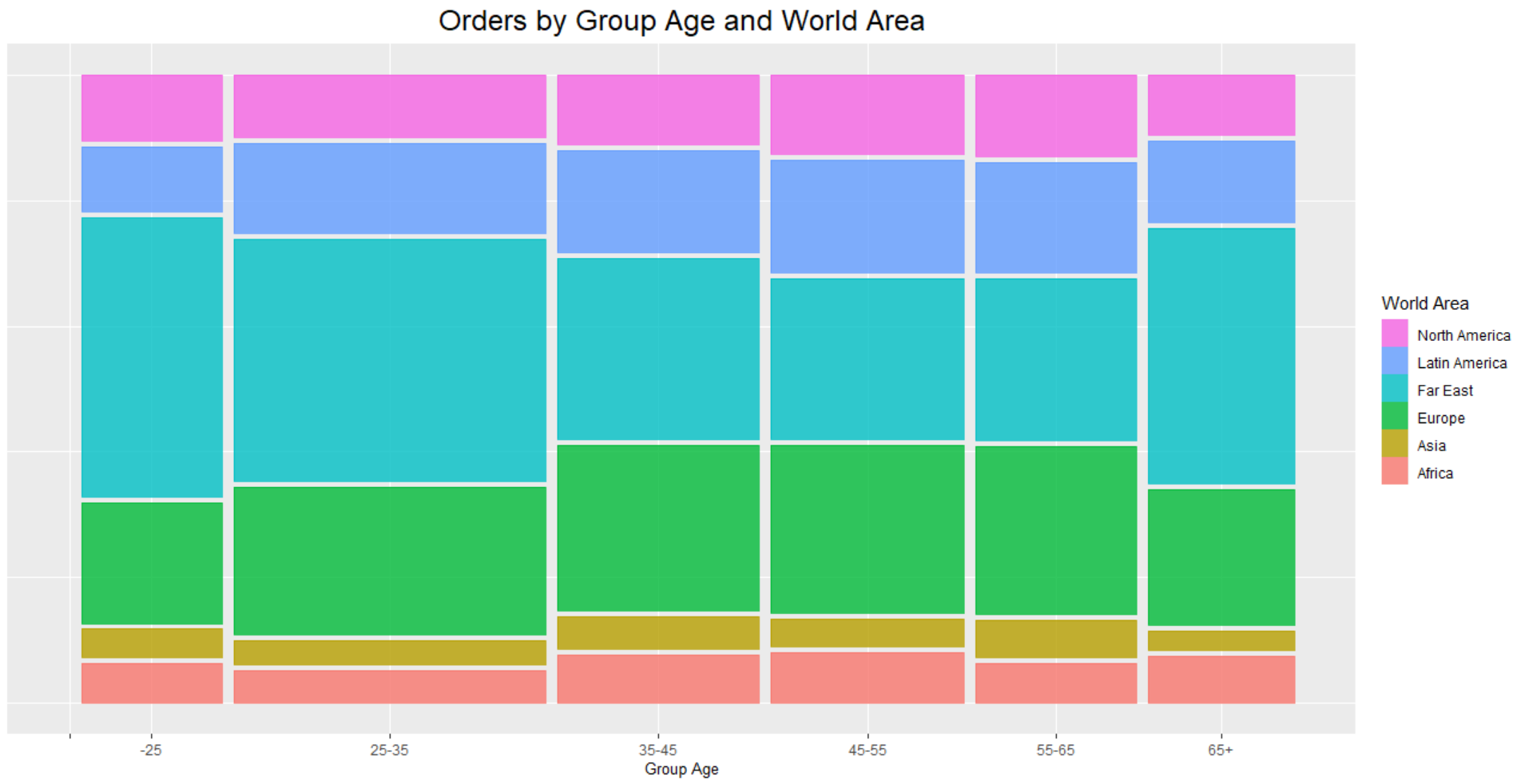
להראות את הקשר בין שני השדות הקטגוריאליים ולאפשר השוואה בין קבוצות האינטראקציה שנוצרו. בנוסף, מהגרף ניתן ללמוד על החלק מתוך השלם.

HOW

בגרף זה אנו מציגות תרשים פסיפס (מוזאיקה). כל עמודה בו מייצגת קבוצת גיל שונה. רוחב העמודות פרופורציונלי למספר התצפיות בכל רמה של משתנה זה שמונח על הציר האופקי. בציר האנכי מצוינים אזורים שונים בעולם. האורך האנכי (הגובה) של כל משבצת פרופורציונלי למספר התצפיות של האזור בעולם עבור כל קבוצת גיל, כלומר מציג את החלק היחסי מתוך השלם של כל אזור בעולם בתוך כל קבוצת גיל.

**את גרף זה יצרנו כי רצינו לקבל מושג לגבי פופולריות אזורי הטיול בעולם בקרב גילים שונים. שימוש אפשרי למידע כזה הוא למשל פרסום ממוקד. הגרף מאפשר לנו לראות גם את** פרופורציית כלל הלקוחות בחלוקה לקבוצת גיל (אפשר לראות למשל שהקבוצה של 25-35 היא הכי גדולה), וגם את פרופורציית ההזמנות לכל אזור בתוך קבוצת גיל מסוימת. בעמודה השמאלית של עד גיל 25 המלבנים של צפון אמריקה (ורוד) ואמריקה הלטינית (כחול) נראים כמעט זהים, כלומר בקרב הלקוחות עד גיל 25 אחוז כמעט זהה הזמין חופשות לכל אחד מהאזורים האלה. כל עמודה מייצגת את השלם – את כל ההזמנות בקרב קבוצת גיל מסוימת, וגבהי המלבנים מאפשרים לראות את החלק היחסי מההזמנות שמתאים לכל אזור בעולם.

ניסיתי לחפש דרך להוסיף בציר y אחוזים אבל לא מצאתי.. כי אין לנו בעצם עמודה שמתארת את הכמויות בכל אינטראקציה, הפונקציה של הגרף מחשבת את זה לבד על ידי כך שאנחנו נותנים לה את שני המשתנים הקטגוריאליים...



איור 7- גרף 4 - Orders by Group Age and World Ares

## גרף 5 – Distribution of Temperature by Vacation Type

WHAT

* טבלה בעלת שדה רציף: **טמפרטורה** (במעלות צלזיוס).
* שדה כמותי: כמות הזמנות.
* ושדה קטגוריאלי: סוג חופשה.

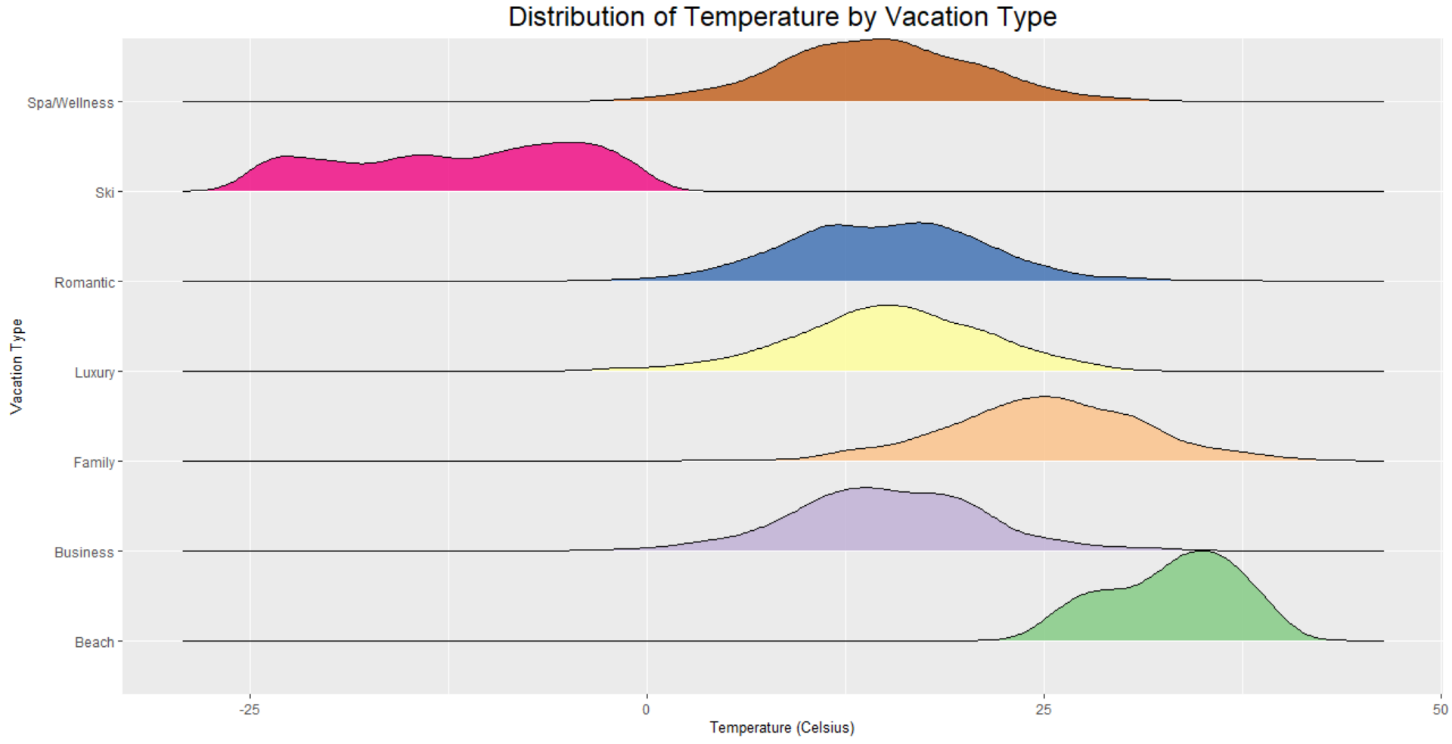
WHY

בעזרת גרף זה נוכל לראות את התפלגות ההזמנות על פני **הטמפרטורה** בחלוקה לסוגי חופשה.

HOW

בגרף זה מוצגת פונקציית התפלגות לכל סוג חופשה. פונקציות ההתפלגות מראות את ההזמנות שבוצעו על פני הטמפרטורה שצפויה במועד החופשה וביעד.

**גרף זה נעשה במטרה לבחון את ההשפעה של הטמפרטורה על סוג החופשה. נוכל לראות, למשל, שבחופשות סקי צפויות טמפרטורות נמוכות לעומת חופשות חוף, שבהן צפויות טמפרטורות גבוהות. לכל סוג חופשה יש פונקציית** התפלגות, עקומה אשר **נותנת מידע על כמות ההזמנות באופן רציף על פני הטמפרטורה. גרף זה מאפשר ללמוד על ההבדלים בביקוש לחופשה מסוימת כתלות בטמפרטורה הצפויה בה. הבחירה להציג את הנתונים האלה בפונקציות התפלגות עוזרת לנו לקבל את המידע בצורה קלה להבנה. ניתן לראות בבירור מבלי להתעמק בגרף יותר מדי את טווח הטמפרטורות הפופולריות בכל סוג חופשה, ואם מתמקדים בסוג מסוים אז רואים באיזה טמפרטורות נעשו יותר הזמנות בצורה ויזואלית שלא מצריכה להשוות מספרים.**



איור 8- גרף 5 - Distribution of Temperature by Vacation Type

## גרף 6 – Box-plot of Cost per Traveler by Vacation Type and Gender

WHAT

* טבלה רב ממדית בעלת שדה אחד כמותי: עלות חופשה לנוסע.
* ו-2 שדות קטגוריאליים: סוג החופשה ומגדר.

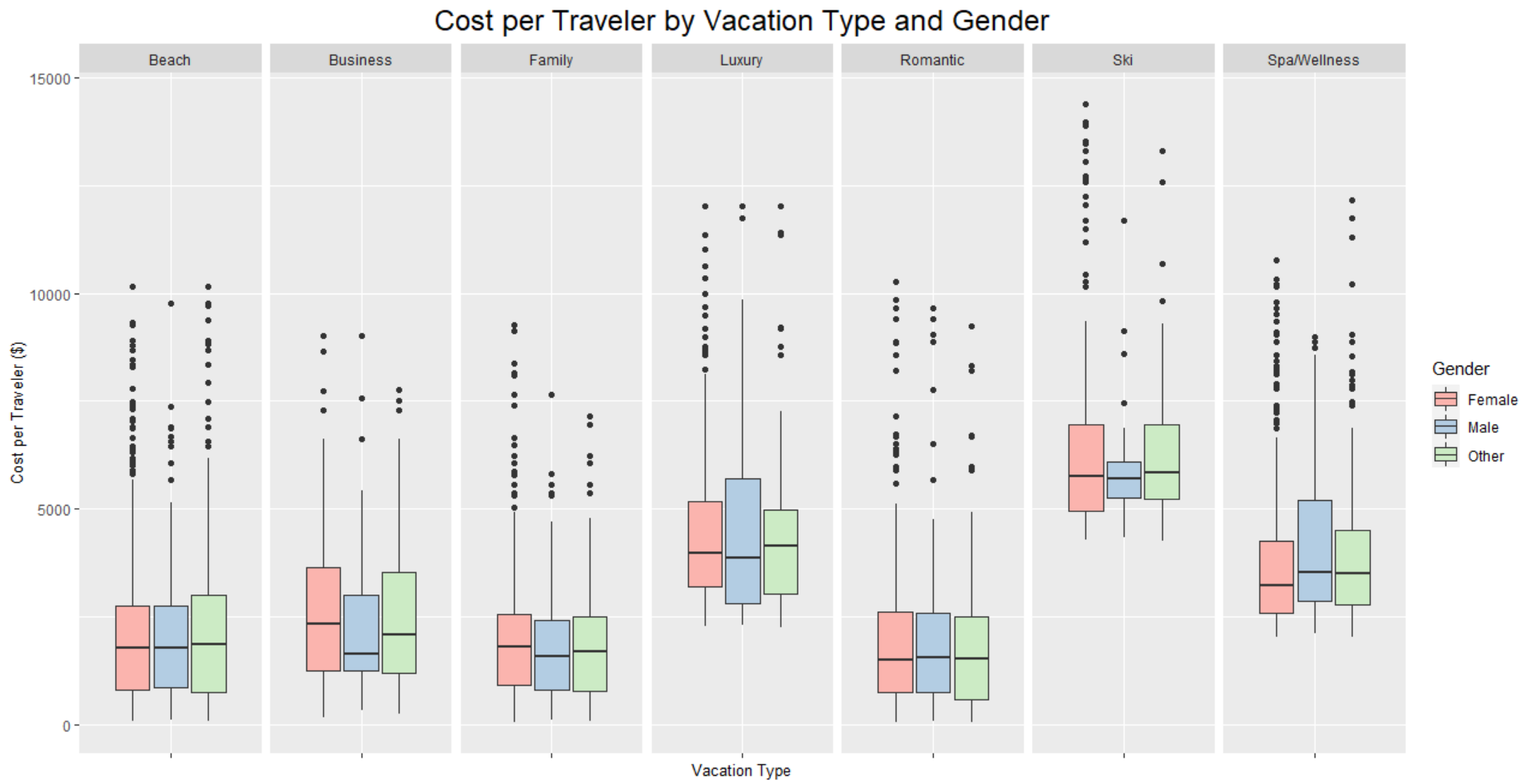
WHY

השוואת ההתפלגות והערכים הסטטיסטיים של משתנה רציף (עלות חופשה לנוסע) עבור סוגי חופשות שונים בחלוקה לפי מגדר המזמין.

HOW

בגרף זה כל קופסה מציגה את התפלגות עלות החופשה לנוסע. בכל קופסה כזאת נוכל לראות את החציון, את האחוזונים ואת התצפיות החריגות. ציר ה-x מציג את החלוקה לפי סוגי החופשות (ב- facets), והקופסאות השונות בכל סוג חופשה מייצגות את מגדר המזמין של החופשה.

גרף זה נעשה במטרה לבחון שני דברים: האם יש הבדלים בעלויות החופשה בין סוגי החופשות השונים והאם המגדר של מזמין החופשה (אישה/גבר/אחר) משפיע על העלות לנוסע. יש בנתונים שלנו לא מעט ערכי מגדר other, של משתמשים שלא רצו להזדהות בפנינו כאישה ולא כגבר. תרשימי הקופסה מאפשרים לנו השוואה מקיפה יותר של הנתונים הסטטיסטיים. ניתן לראות, למשל, שהעלות לנוסע של חופשות הסקי, הפאר והספא יקרות יותר משאר החופשות. בין המגדרים אין הבדלים מהותיים. יש הרבה חריגים שיצרו זנב ימני, אשר מצביע על הזמנות יחסית מעטות שבהן העלות לנוסע הייתה גבוהה משמעותית מערכי החציון של אותו סוג חופשה.



איור 9- גרף 6 - Box-plot of Cost per Traveler by Vacation Type and Gender

# קוד מלא ב-R

#libraries

library(ggplot2)

library(ggmosaic)

library(ggridges)

#install.packages("RColorBrewer")

library(RColorBrewer)

#import csv

filepath = choose.files()

dataset <- read.csv(filepath,header = TRUE)

##mosaic: Age Group & World Area

subset1 <- subset(dataset, GroupAge!="")

ggplot(data = subset1) +

geom\_mosaic(aes(x = product(WW.Erea, GroupAge), fill=WW.Erea)) +

labs(title="Orders by Group Age and World Area", x="Group Age", y="World Area", fill = "World Area") +

theme(plot.title = element\_text(size=18, hjust = 0.5),axis.title.x=element\_text(size=10,hjust = 0.5),axis.title.y=element\_blank(),axis.text.y=element\_blank(),axis.ticks.y=element\_blank()) +

guides(fill = guide\_legend(reverse=TRUE))

##boxplot: World Area & Revenues by Gender

subset2 <- subset(dataset, Gender!="NULL")

ggplot(subset2, aes(x=PropertyStyle, y=Price.per.traveler, fill=Gender)) +

geom\_boxplot()+

facet\_grid(. ~ PropertyStyle, scales='free')+

labs(title="Cost per Traveler by Vacation Type and Gender", x="Vacation Type", y="Cost per Traveler ($)")+

theme(plot.title = element\_text(size=18, hjust = 0.5),axis.title.x=element\_text(size=10,hjust = 0.5),axis.title.y=element\_text(size=10,hjust = 0.5),axis.text.x=element\_blank())+

scale\_fill\_brewer(palette="Pastel1",labels = c("Female", "Male", "Other"))

##density ridges: Temperature & Vacation Style

ggplot(dataset,

aes(x = temperature,

y = PropertyStyle,

fill = PropertyStyle)) +

geom\_density\_ridges(alpha = 0.8, scale = 1, show.legend = FALSE) +

theme\_ridges() +

theme\_gray() +

labs(title="Distribution of Temperature by Vacation Type", x="Temperature (Celsius)", y="Vacation Type")+

theme(plot.title = element\_text(size=18, hjust = 0.5),axis.title.x=element\_text(size=10,hjust = 0.5),axis.title.y=element\_text(size=10,hjust = 0.5),axis.text.y=element\_text(size=9))+

scale\_fill\_brewer(palette="Accent")