דו"ח פרויקט - חלק א' (תשפ"ד סמסטר ב')

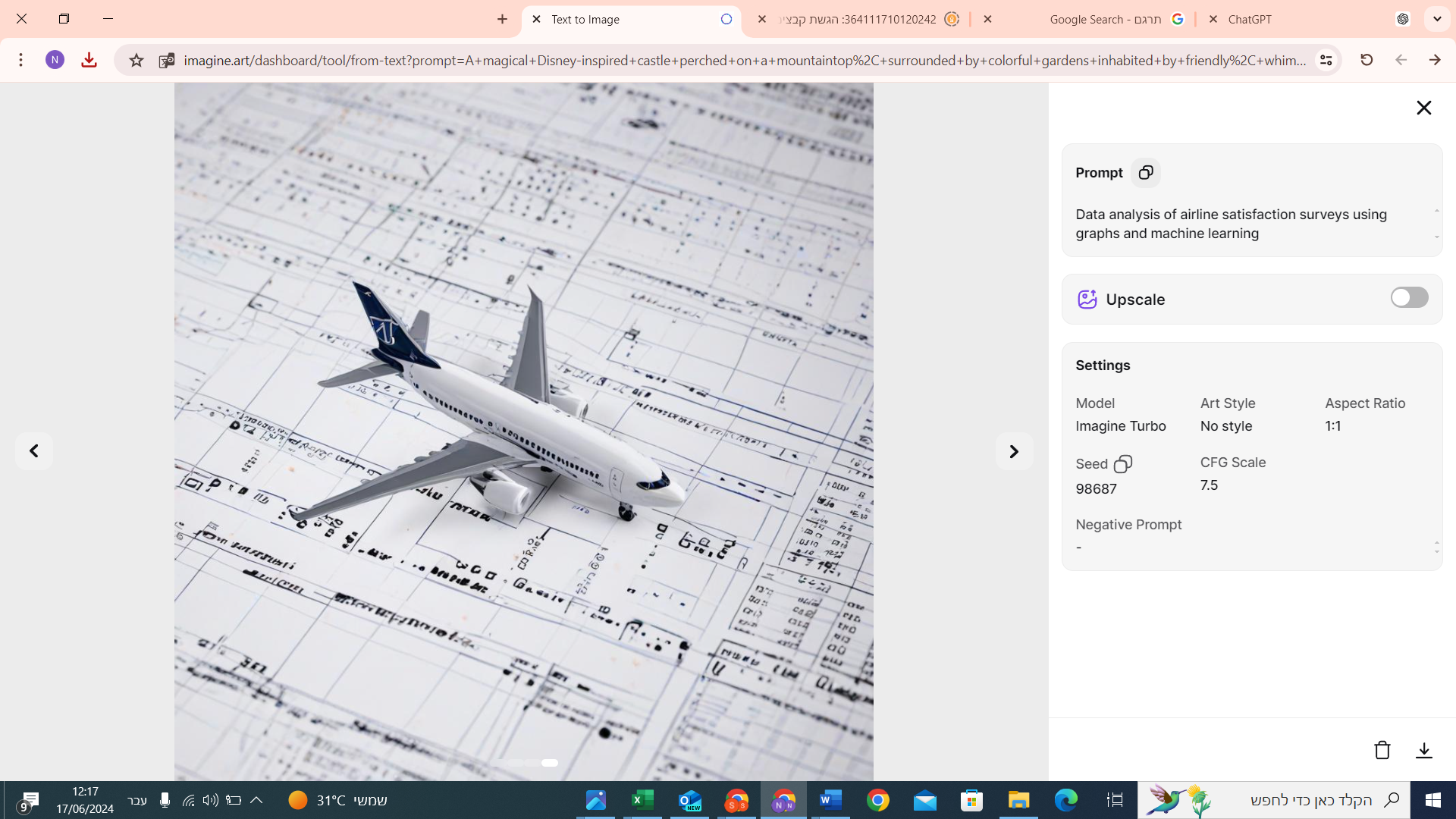
**קבוצה 6**

שחר לוי אטיאס 319042685

עדי נעמתי 209353838

תמיר לבנטר 206557456

יובל עמית 208956789



**תוכן עניינים**

[הגדרת הבעיה 3](#_Toc169640148)

[1. תיאור כללי של עולם התוכן הנחקר 3](#_Toc169640149)

[2. הגדרת שאלת המחקר 3](#_Toc169640150)

[הבנת הנתונים 3](#_Toc169640151)

[1. תיעוד מקורות הנתונים ומשמעותם 3](#_Toc169640152)

[2. הסתברויות אפריוריות וקשרים בין מאפיינים 4](#_Toc169640153)

[3. איכות הנתונים................................................................................................................................11](#_Toc169640154)

[הכנת הנתונים 11](#_Toc169640155)

[1. נימוק בחירות מאפיינים שביצענו 11](#_Toc169640156)

[2. טיפול פרטני במאפיינים 12](#_Toc169640157)

[נספחים 13](#_Toc169640158)

# **הגדרת הבעיה**

## **תיאור כללי של עולם התוכן הנחקר:**

קובץ הנתונים מכיל נתוני סקר שביעות רצון של נוסעים בחברת תעופה. עולם התוכן הנחקר הוא חוויית הטיסה והגורמים השונים המשפיעים על שביעות הרצון של הנוסעים. מטרת המחקר היא לבחון כיצד משתנים אלה משפיעים על שביעות הרצון ולנסות לזהות את הגורמים המשמעותיים ביותר המנבאים שביעות רצון גבוהה או נמוכה. הבנה זו עשויה לסייע לחברות התעופה לשפר את שרותיהן.

הבעיה המחקרית- חוסר הבנת גורמי ההשפעה על שביעות רצון הנוסעים בחברות תעופה.

מחקרים קודמים בתחום שביעות רצון לקוחות בענף התעופה וענפי שירות נוספים, השתמשו בשיטות הבאות:

1. סקרי שביעות רצון- איסוף נתונים באמצעות שאלונים וראיונות עם נוסעים לאחר טיסותיהם.
2. ניתוח רגרסיה- שימוש במודלים סטטיסטיים לזיהוי המשתנים המשפיעים ביותר על שביעות הרצון.
3. ניתוח גורמים- טכניקה סטטיסטית לזיהוי קבוצות של משתנים בעלי השפעה דומה.
4. ניתוח קבוצות- חלוקת הנתונים לקבוצות של נוסעים בעלי מאפיינים דומים לזיהוי דפוסים.
5. מודלים של למידת מכונה- שימוש באלגוריתמים ללמידה ממוקדת על מנת לחזות שביעות רצון לקוחות לפי מאפיינים שונים.

## **הגדרת שאלת המחקר:**

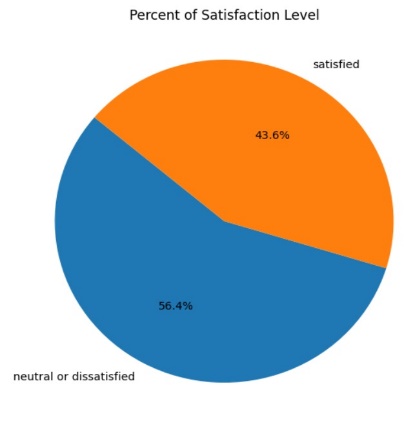
שאלת המחקר- מהם הגורמים המשפיעים ביותר על שביעות הרצון של נוסעים בחברות תעופה.

# **הבנת הנתונים**

## **תיעוד מקורות הנתונים ומשמעותם:**

לפי התיאור והנתונים שקיבלנו, מקור הנתונים נראה כסקר שביעות רצון של נוסעים בחברות התעופה. אנו מסיקים כי הנתונים נוצרו באמצעות איסוף ידני של תשובות נוסעים לשאלון או ראיונות לאחר טיסותיהם.

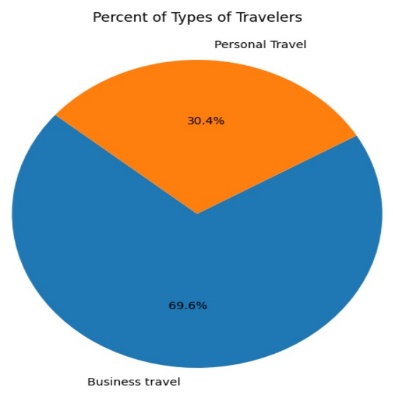
|  |  |  |
| --- | --- | --- |
| **משתנה** | **משמעות** | **סוג משתנה** |
| Satisfaction | רמת שביעות רצון כללית מהטיסה (שבע רצון, ניטראלי, לא שבע רצון) | קטגוריאלי ניתן לסידור |
| Gender | מגדר הנוסע (זכר/נקבה) | קטגוריאלי |
| Customer Type | סוג הלקוח (נאמן/לא נאמן) | קטגוריאלי |
| Age | גיל הנוסע | רציף |
| Type of Travel | מטרת הטיסה (נסיעה פרטית/נסיעת עסקים) | קטגוריאלי |
| Class | מחלקת הטיסה (עסקים, תיירים, תיירים פלוס) | קטגוריאלי |
| Flight distance | מרחק הטיסה | רציף |
| Plane colors | מספר צבעים על גוף המטוס | בדיד שנתייחס כרציף |
| Inflight wi-fi service | רמת שביעות רצון משירות Wi-Fi במהלך הטיסה (1-5, 0- לא היה Wi-Fi בטיסה) | בדיד שנתייחס כרציף |
| Departure/Arrival time convenient | רמת נוחות זמני ההמראה/נחיתה (0-5) | בדיד שנתייחס כרציף |
| Ease of Online booking | רמת קלות ההזמנה המקוונת (0-5) | בדיד שנתייחס כרציף |
| Gate location | רמת שביעות רצון ממיקום השער (0-5) | בדיד שנתייחס כרציף |
| Food and drink | רמת שביעות רצון ממזון ומשקאות (1-5, 0- לא היה אוכל ושתייה) | בדיד שנתייחס כרציף |
| Seat comfort | רמת נוחות המושבים (0-5) | בדיד שנתייחס כרציף |
| On-board service | רמת שביעות רצון משירות הצוות (0-5) | בדיד שנתייחס כרציף |
| Leg room service | רמת מרחב לרגליים (0-5) | בדיד שנתייחס כרציף |
| Baggage handling | רמת שביעות רצון מטיפול בכבודה (0-5) | בדיד שנתייחס כרציף |
| Check-in service | רמת שביעות רצון משירות בדלפק (0-5) | בדיד שנתייחס כרציף |
| Inflight service | רמת שביעות רצון משירות במהלך הטיסה (0-5) | בדיד שנתייחס כרציף |
| Cleanliness | רמת ניקיון (0-5) | בדיד שנתייחס כרציף |
| Departure Delay in Minutes | משך עיכוב בהמראה בדקות | רציף |
| Arrival Delay in Minutes | משך עיכוב בנחיתה בדקות | רציף |



## **הסתברויות אפריוריות וקשרים בין מאפיינים:**

**משתנה המטרה- Satisfaction**

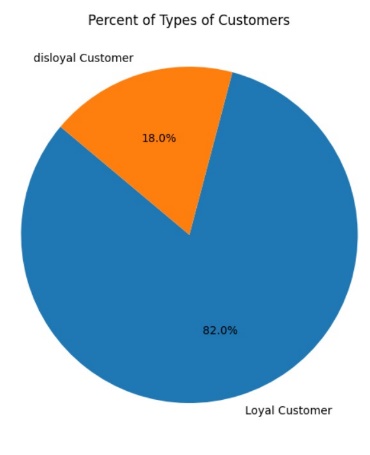
ניתן ללמוד מהסתברויות אלו כי רוב הנוסעים ניטרליים או אינם מרוצים מחוויית השירות של חברת התעופה. עם זאת, יש אחוז לא מבוטל של נוסעים שהיו מרוצים מחוויית השירות של החברה.



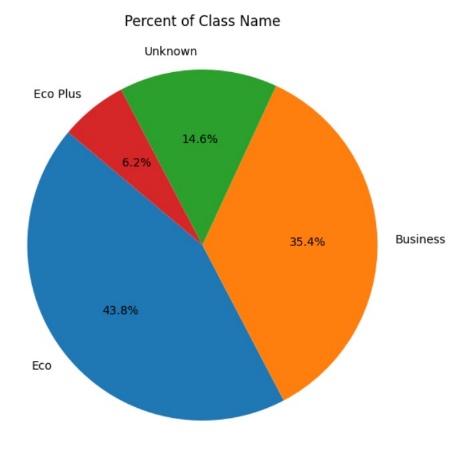
**משתנים קטגוריאליים-**

1. **מטרת הטיסה- Type of Travel**

מתוך ההסתברויות ניתן להסיק כי נסיעות עסקים מהוות חלק גדול יותר מהנסיעות. מהדבר הזה ניתן להבין שיותר אנשים שמילאו את הסקר טסו למטרות עסקיות מאשר למטרות פרטיות.

1. **סוג הנוסע- Customer Type**

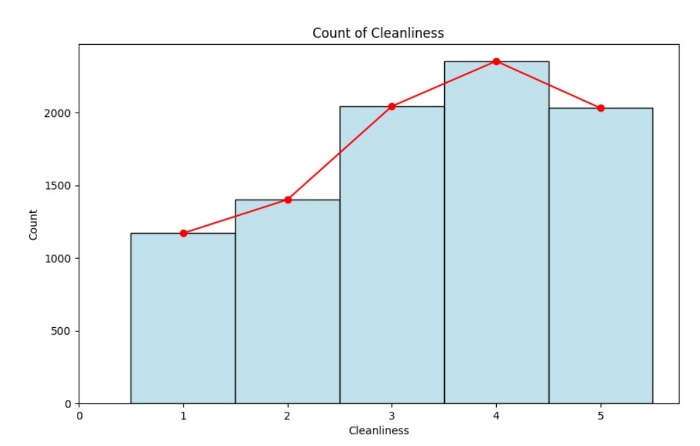
ניתן ללמוד מהסתברויות אלו כי רוב הנוסעים שמילאו את סקר חווית השירות הם לקוחות של חברת התעופה שנוסעים איתה באופן קבוע (לקוחות נאמנים), וכך הנתונים שמתקבלים מהם אמינים ומגובשים. עם זאת, איזון בין מספר הנוסעים הקבועים לנוסעים חד פעמיים עשוי לספק תמונה מלאה ומדויקת יותר של חוויית השירות ולאפשר שיפור מכוון, שכן לקוחות חד-פעמיים יכולים לספק תובנות על חוויה הראשונית.



1. תמונה שמכילה טקסט, צילום מסך, תרשים, מלבן

   התיאור נוצר באופן אוטומטי**מחלקת הטיסה- Class**

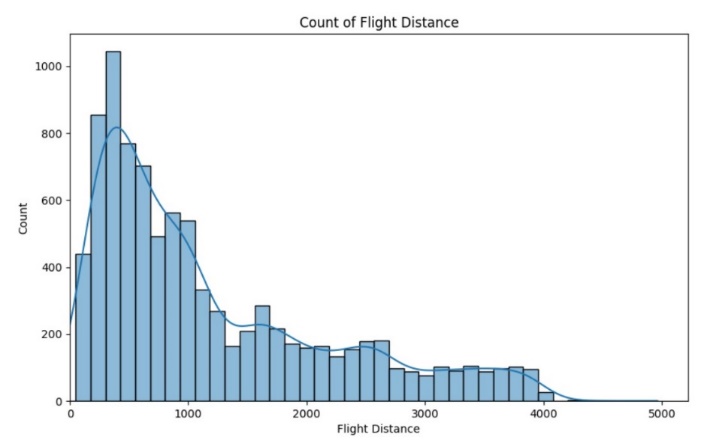
מההסתברויות ניתן להניח שרוב הנוסעים בחברת התעופה הזו הם במחלקה הסטנדרטית מחלקת התעופה הגבוהות והמשודרגות יותר מהוות רק שיעור קטן מהנוסעים. סט נתונים זה לדעתנו, מאוזן ומשקף את המציאות שכן רוב הנוסעים טסים במחלקות רגילות ולא הרבה יכולים להרשות לעצמם מחלקות אחרות.



**משתנים רציפים-**

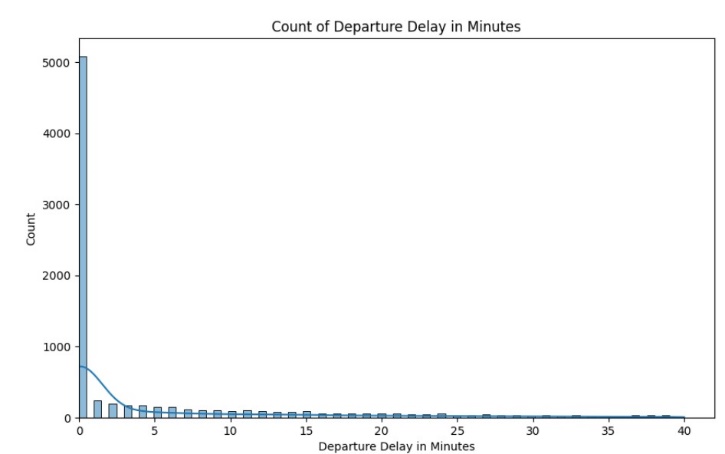
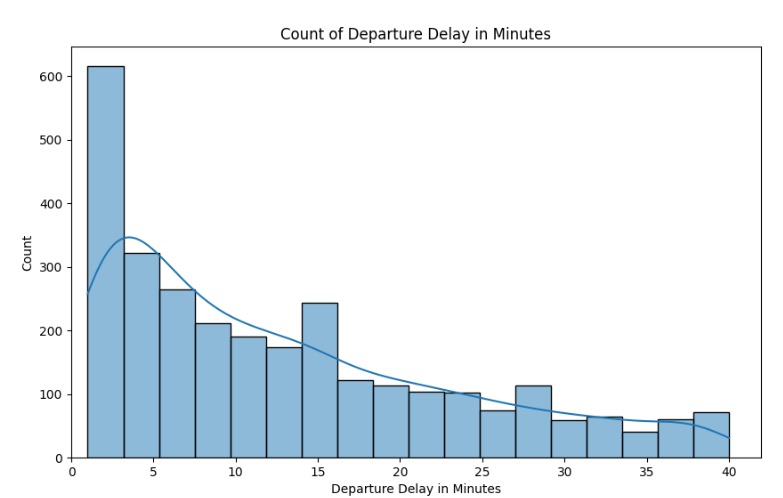
1. **ניקיון- Cleanliness**

מהגרף ניתן ללמוד כי רוב הנוסעים מדרגים את רמת הניקיון בציונים גבוהים, עם שיא של ניקיון ברמה 4. הציונים הנמוכים ביותר ניתנים לעיתים נדירות, והמספר הכולל של דירוגי ניקיון עולה באופן מתמשך מרמה 1 ועד רמה 4, ולאחר מכן יש ירידה קלה ברמה 5.



1. **מרחק טיסה- Flight Distance**

התפלגות הנ"ל מצביעה על כך שמרבית הטיסות הן קצרות טווח, כנראה בזכות הביקוש הגבוה לנסיעות מקומיות. טיסות ארוכות טווח למרחקים בינלאומיים הן נדירות יותר ומהוות חלק קטן מסך כל הטיסות ובנוסף, טיסות ארוכות יכולות להתחלק לטיסות קצרות עם קונקשנים וכך להעלות את מספר הטיסות הקצרות של חברת התעופה.

1. **משך עיכוב המראה בדקות- Departure Delay in Minutes**

מגרף ההתפלגות ניתן ללמוד כי חברת התעופה לרוב אינה מתעכבת בזמני ההמראה (מעל 5000 טיסות). ניתן לראות כי ככל שזמן עיכוב ההמראה עולה כך כמות הטיסות יורדת דבר שמלמד אותנו כי עיכוב גדול בהמראות הוא אינו נפוץ.

* הגרף הימני העתק של הגרף השמאלי אך ללא התצפיות בהן לא היה עיכוב ביציאה, על מנת שנוכל לראות את ההתפלגות בצורה ברורה יותר.

1. תמונה שמכילה טקסט, צילום מסך, מלבן, קו

   התיאור נוצר באופן אוטומטיתמונה שמכילה תרשים, טקסט, עלילה, קו

   התיאור נוצר באופן אוטומטי**משך עיכוב הנחיתה בדקות- Arrival delay in minutes**

מגרף ההתפלגות ניתן ללמוד כי חברת התעופה לרוב אינה מתעכבת בזמני הנחיתה (מעל 5000 טיסות). ניתן לראות כי ככל שזמן עיכוב בנחיתה עולה כך כמות הטיסות יורדת דבר שמלמד אותנו כי עיכוב גדול בנחיתות הוא אינו נפוץ.

* הגרף הימני העתק של הגרף השמאלי אך ללא התצפיות בהן לא היה עיכוב בנחיתה, על מנת שנוכל לראות את ההתפלגות בצורה ברורה יותר.

1. **תמונה שמכילה תרשים, עלילה, קו, צילום מסך

   התיאור נוצר באופן אוטומטיגיל- Age**

מהגרף ניתן ללמוד כי רוב הנוסעים הם בגילאים 30 עד 50, עם שיא סביב גיל 40. הגילאים מתפזרים באופן רחב בין 20 ל-80, כאשר ישנה ירידה משמעותית בכמות הנוסעים מעבר לגיל זה. כמו כן, ניתן לראות שיש נוכחות נמוכה של נוסעים בגילאים צעירים מאוד ומבוגרים מאוד, מה שמעיד על כך שהם פחות משתתפים בסקרים מסוג זה או פחות טסים.

1. תמונה שמכילה תרשים, עלילה, קו

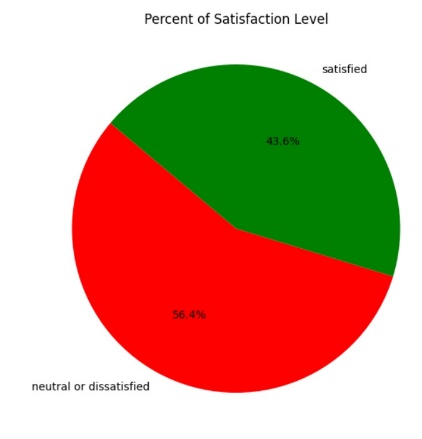
   התיאור נוצר באופן אוטומטי**רמת שביעות רצון משירות וויפי- Inflight wifi service**

הגרף של התפלגות דירוגי שירות ה- WiFi בטיסה ניתן ללמוד שמרבית הנוסעים נתנו דירוגים בינוניים, מה שמעיד על חוויה ממוצעת. יש פחות נוסעים שנתנו דירוגים קיצוניים. בנוסף, דירוג 0 בהתפלגות מראה כי לחברת התעופה אין הרבה טיסות ללא WiFi.המסקנה היא ששירות ה- WiFi נתפס כבסדר אך לא מצוין.

1. תמונה שמכילה טקסט, תרשים, עלילה, קו

   התיאור נוצר באופן אוטומטי**רמת נוחות זמני ההמראה/נחיתה- Departure/Arrival time convenient**

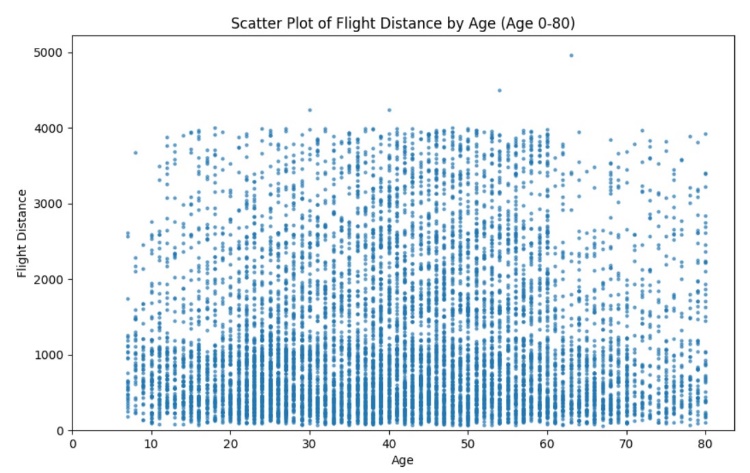
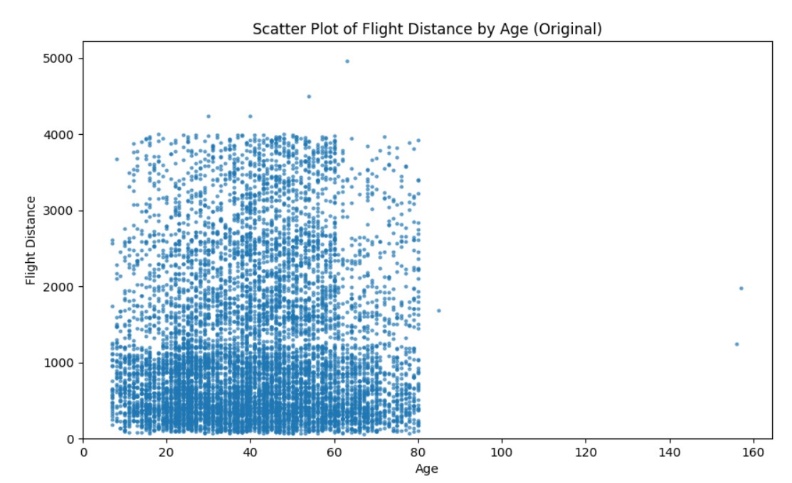
מהגרף של ההתפלגות של זמני ההמראה והנחיתה הנוחים ניתן ללמוד כי רוב הנוסעים מוצאים את זמני ההמראה/נחיתה נוחים, במיוחד כאשר הציון הוא 4, עם כמות גבוהה יותר של נוסעים שמוצאים את הזמנים נוחים. בסך הכל, זמני ההמראה והנחיתה נתפסים כנוחים בעיני רוב הנוסעים.



* בבדיקה האם סט הנתונים מאוזן מצאנו כי הוא לא מאוזן מכיוון ויש יותר אנשים שהיו ניטרליים או לא שבעי רצון מאשר אנשים שהיו שבעי רצון מחברת התעופה.

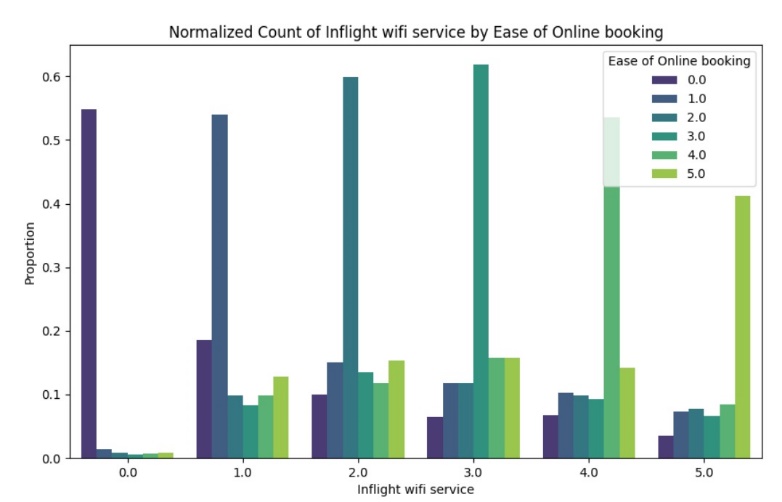
אנו חושבים כי סט הנתונים תואם את המציאות כי לרוב האנשים יש נטייה לבטא יותר חוסר שביעות רצון מטעמים פסיכולוגיים, חברתיים ותרבותיות. לכן גם בסט הנתונים הנ"ל ראינו כי אנשים נטו להתמקד בדברים הפחות טובים ולהביע יותר חוסר שביעות רצון.

**קשרים בין מאפיינים:** לאחר בחינת הנתונים, ניתן לזהות מספר קשרים מעניינים בין מאפיינים שונים:

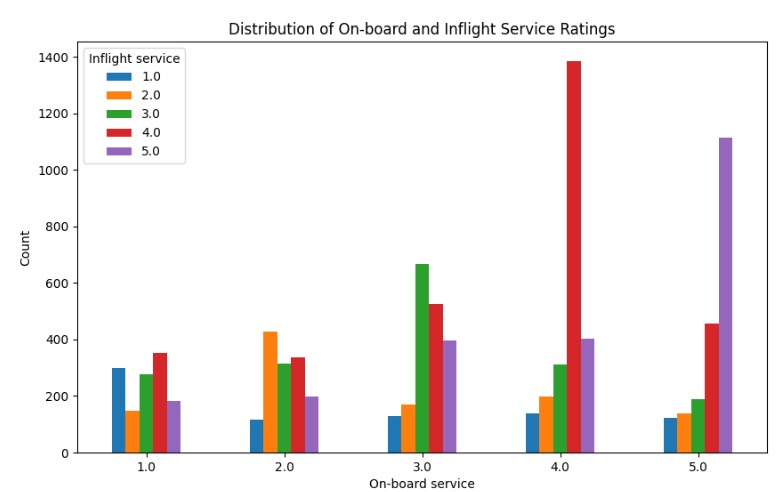
1. **קשר בין Age ו- Flight Distance:**

מהגרף ניתן ללמוד שאין קשר ברור בין גיל הנוסעים למרחק הטיסה. הנקודות מפוזרות באופן שווה יחסית לאורך כל טווחי הגילאים ומרחקי הטיסה, מה שמעיד על כך שמרחק הטיסה אינו תלוי בגיל הנוסע. הטווח הרחב של המרחקים קיים בכל קבוצות הגילאים, מראה שהגיל אינו גורם מכריע במרחקי הטיסות. כמו כן ניתן לראות שבגילאים 30-50 יש מספר תצפיות גדול יותר. תופעה זו יכולה להיות לאור העובדה שיש יותר תצפיות.

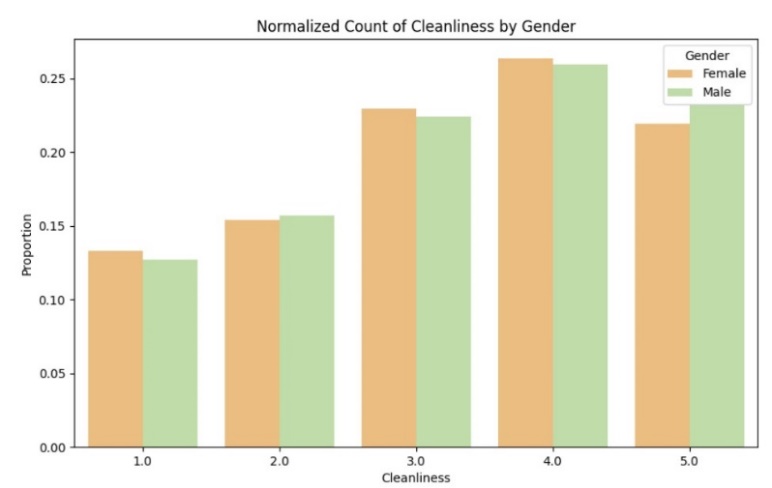
* צירפנו גרף לפני הוצאת חריגים וגרף לאחר הוצאת חריגים אשר רואים בו יותר בבירור.

1. **קשר בין Inflight Service ו- Ease of Online booking :**

הגרף מראה קשר ברור בין דירוג שירות ה-Wi-Fi במהלך הטיסה לבין קלות ההזמנה המקוונת. כאשר שירות ה-Wi-Fi מקבל דירוג נמוך, גם דירוג קלות ההזמנה המקוונת נוטה להיות נמוך. לעומת זאת, כאשר שירות ה-Wi-Fiמקבל דירוג גבוה, קלות ההזמנה המקוונת נוטה לקבל ציונים גבוהים. מעניין לראות שמי שלא הזמין באינטרנט כלל (דירוג 0 בקלות ההזמנה המקוונת) גם לא השתמש בשירותי ה-Wi-Fi בטיסה. באופן כללי, שביעות רצון גבוהה משירות ה-Wi-Fi קשורה לחוויית הזמנה מקוונת טובה יותר.

1. **קשר בין Inflight Service ו-On- board Service:**

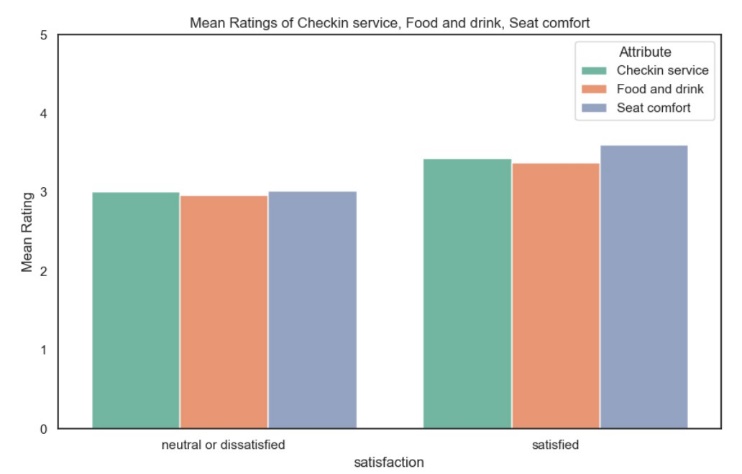
תרשים זה מציג את התפלגות הדירוגים עבור שירות על המטוס ושירות בטיסה. ניתן לראות שקיים קשר חיובי בין המשתנים, כאשר דירוג גבוה או נמוך באחד הפרמטרים נוטה להתאים לדירוג דומה בפרמטר השני.בנוסף, ניתן לראות שככל שהדירוג עולה כך גם כמות התצפיות דבר המעיד על שירות כללי טוב בחברת התעופה עבור סט נתונים זה.

1. **קשר ביןCleanliness ו-Gender:**

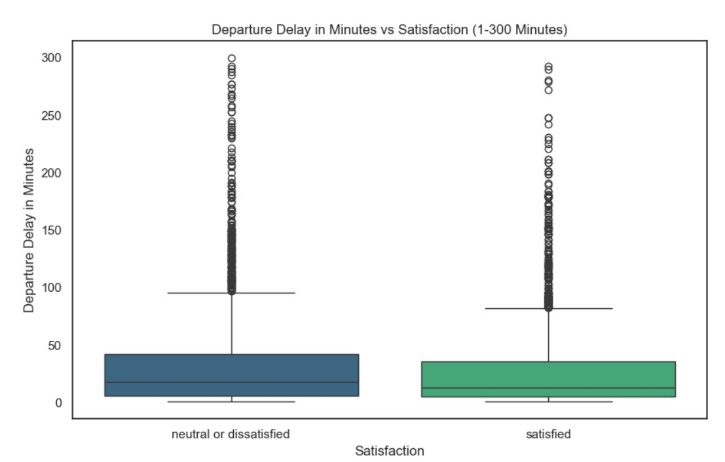
מהגרף ניתן ללמוד על הקשר בין דירוג הניקיון לבין מגדר הציפייה הייתה לראות הבדל משמעותי בין דירוגי הניקיון שניתנים על ידי נשים וגברים, שכן קיימת נטייה לחשוב שלנשים יותר אכפת מניקיון מאשר לגברים. עם זאת, מהגרף עולה כי אין הבדל משמעותי (קשר חיובי) בין הדירוגים שניתנו על ידי נשים וגברים. הדירוגים מתפלגים באופן דומה בשתי הקבוצות, כאשר רוב הדירוגים הם בין 3 ל-5, ללא העדפה מגדרית מובהקת.

**מאפיינים חשודים**

המאפיינים שאנו חושדים בהם כבעלי השפעה על משתנה המטרה (שביעות הרצון מחברת התעופה) על סמך ידע אישי קודם ומחקרים הם:

* **איכות השירות בזמן הטיסה:** אחד הגורמים המשפיעים ביותר על שביעות רצון הנוסעים הוא איכות השירות שניתן בזמן הטיסה, מתחילתו ועד סופו. זה כולל את אדיבות הצוות והיכולת שלהם להתמודד עם בעיות, נוחות המושבים, וכן את איכות ומגוון האוכל והמשקאות ושירות הצ'ק אין. נוסעים מעריכים במיוחד שירות אישי ואדיב.

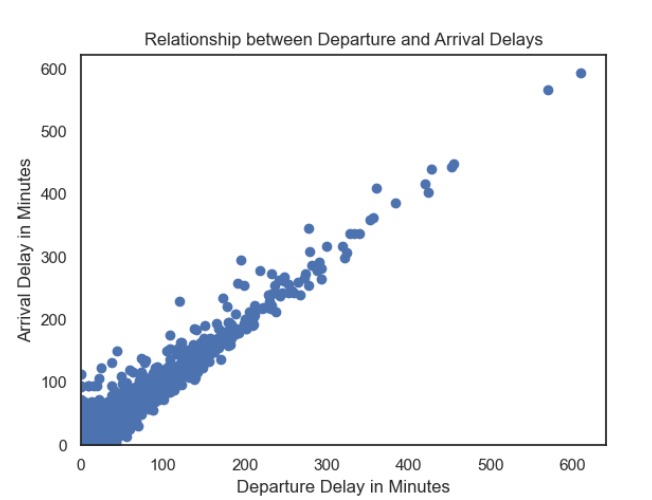
מהניתוח הזה ניתן להבין כי נוחות המושב ושירות הצ'ק אין הם שני הגורמים בעלי ההשפעה הגדולה ביותר על שביעות הרצון של הנוסעים. אוכל ושתייה גם משפיעים, אך באופן פחות ברור בהשוואה לשני הגורמים האחרים. במילים אחרות, כדי לשפר את שביעות הרצון של הנוסעים, חברת התעופה צריכה לשים דגש מיוחד על שיפור נוחות המושבים ושירות הצ'ק אין.

**דיוק בזמנים (בהמראה ובנחיתה):** גורם נוסף בעל השפעה קריטית הוא הדיוק בזמנים. נוסעים מצפים שהטיסות יצאו וינחתו בזמן, ואיחורים או ביטולים יכולים לגרום לתסכול ולפגוע בשביעות הרצון. דיוק בזמנים משדר אמינות ומקצועיות של חברת התעופה, ומבטיח לנוסעים תחושת ביטחון בתכנון זמנם. [מחקר](https://www.researchgate.net/publication/328468382_The_Impact_of_Delays_on_Customers'_Satisfaction_An_Empirical_Analysis_of_the_British_Airways_On-Time_Performance_at_Heathrow_Airport) בשם "The Impact of Delays on Customers' Satisfaction: An Empirical Analysis of the British Airways On-Time Performance at Heathrow Airport" מצא כי דיוק בזמנים הוא אחד הגורמים המשפיעים ביותר על שביעות רצון הנוסעים.  
בעקבות העובדה שהרוב המוחלט של הטיסות בסט הנתונים שלנו היו ללא עיכוב בחרנו להסתכל על זמן העיכוב של הטיסה, לפי גרף הBoxPlot ניתן לראות כי החציון של עיכובים בזמן ההמראה עבור הנוסעים המרוצים נמוכים יותר בהשוואה לנוסעים הניטרליים או הלא מרוצים. בנוסף, ישנה שונות גדולה יותר של עיכובים בקרב הנוסעים הלא מרוצים, מה שמעיד על כך שעיכובים ארוכים הם גורם מרכזי לאי שביעות רצון.

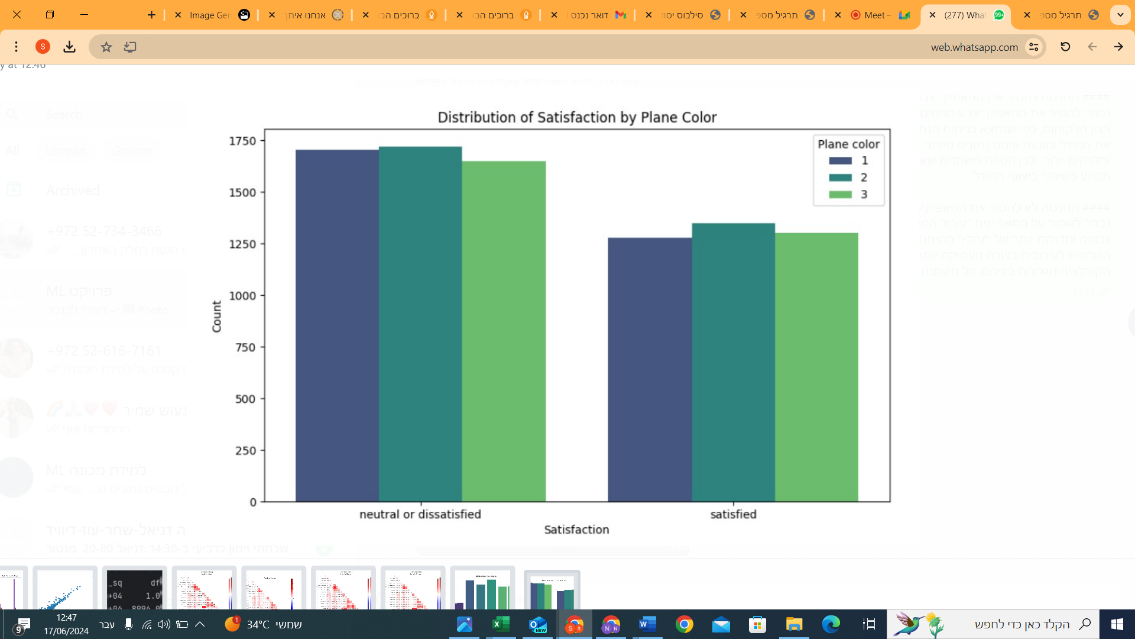
**בדיקת הסרת מאפיינים**

המשתנים שאנו מניחים שניתן להסירם על סמך ידע אישי: עיכוב הגעה בדקות ומספר צבעי המטוס.

הנחה ראשונה:

אנו יודעים שבדרך כלל זמן הטיסה הוא קבוע ומניחים על סמך ידע אישי, שם יש עיכוב בהמראה כנראה בוודאות יהיה עיכוב בנחיתה. נבחן *את ההנחה ע"י בחינת גרף הקורלציה בין שני המשתנים:*

*ניתן לראות כי ההשערה שלנו מתקיימת ואכן קיימת קורלציה גבוהה בין שני המשתנים. קורלציה זו מעידה על כך שמשתנה זה לא תורם לנו מידע נוסף ועל כן, ניתן להסיר את המשתנה "עיכוב הגעה בדקות".*

*הנחה השנייה:*

*אנו מניחים כי למספר צבעי המטוס אין השפעה על שביעות הרצון של הלקוחות מחברת התעופה. נבדוק את ההשערה באמצעות גרף עמודות: ניתן לראות כי אין השפעה ברורה וחד משמעית לצבע המטוס על שביעות הרצון של הלקוחות ועל כן, ההשערה שלנו מתקיימת וניתן להסיר מאפיין זה.*

## **איכות הנתונים**

נתונים חסרים לאחר ניתוח חסרים בקובץ הנתונים מצאנו כי ישנה רשומה שקיימים בה חוסרים ב- 16 מאפיינים ורשימה נוספת ובה יש כ- 10 מאפיינים חסרים. רשומה זו בעלת חוסרים כבדים ומביאה תועלת נמוכה לחקר הנתונים. כחלק מהתמודדות עם חוסרים, במידה וברשומה קיימים יותר מ-20% ערכים חסרים נבחר להסיר רשומה זו וזאת, על מנת לשמור על אמינות הנתונים ועל שיקוף המציאות. על כן, נסיר רשומות אלו ולא נשלים את הנתונים החסרים בהן. בנוסף, מצאנו כי במאפיין “Leg room service” יש כ-2700 ערכים חסרים ([נספח 1](#_נספח_1-_נתונים)) לפי גרף העוגה ([נספח 1](#_נספח_1-_נתונים)) המוצג ניתן לראות כי הערכים החסרים מהווים כ-30% מסך הרשומות. בנוסף, במאפיין “Arrival Delay in Minutes” יש כ-400 ערכים חסרים ([נספח 1](#_נספח_1-_נתונים)). על מנת להתמודד עם חוסרים אלו החלטנו לבדוק את הקורלציה בין המשתנים. לפי גרף Heatmap ([נספח 2](#_נספח_2-_גרף)) ניתן לראות כי יש קורלציה גבוהה בין עיכוב בזמני ההמראה לעיכוב בזמני הנחיתה. בעקבות הקורלציה הגבוהה נוכל להשלים את החוסרים בקלות על ידי שכפול הנתונים של משתנה Departure Delay in Minutes"" ולכן נבחר שלא להסיר את משתנה “Arrival Delay in Minutes”.   
נבחר להסיר מאפיינים בעלי 25% נתונים חסרים מהסיבה שלא נוכל להשלים אותם באופן איכותי ומדויק, למעט משתנה המטרה (Satisfaction).

נתונים בלתי הגיוניים

לאחר ניתוח נתונים חריגים או בלתי הגיוניים בקובץ ה-Data מצאנו כי במאפיין Age יש שני ערכים חריגים (156,157). במאפיין Class מצאנו כי יש ערך שונה מ- Eco, Eco plus, business ו-unknown. במאפיין flight distance מצאנו ערך שלילי (204-) ובמאפיין Gate location מצאנו ערך שלא בטווח [0-5].

נבחר להתמודד עם נתונים בלתי הגיוניים על ידי מחיקת הערך והשלמתו לפי חציון, ממוצע או לפי שכיחות הנתונים על מנת לשמור על היחס. מאחר וכל מאפיין יש ערכים שונים, נבחר להשלים באופן התואם כל מאפיין על מנת להבטיח כי ההתפלגות של הנתונים לא תיפגע באופן משמעותי וכי יהיה ניתן להפיק מסקנות אמיתיות מהנתונים.

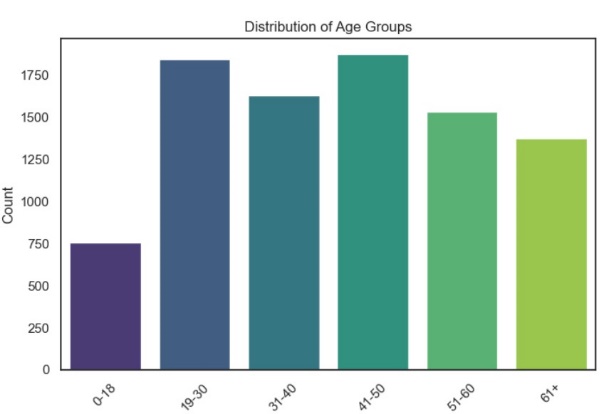
# **הכנת הנתונים**

## **נימוק בחירות מאפיינים שביצענו**

* השמטת מאפיינים "רועשים" או חסרי חשיבות- נבחר להסיר את המאפיין "Plane Colors" מכיוון שהראינו כי אין לו השפעה משמעותית על שביעות רצון הלקוחות. הסרת משתנים לא רלוונטיים עוזרת לפשט את המודל ומונעת עומס נתונים מיותר. בנוסף, מודלים פשוטים יותר נוטים להיות יעילים ומדויקים יותר, ולכן הסרת משתנים שאין להם תרומה משמעותית לשביעות רצון הלקוחות תסייע בשיפור ביצועי המודל.
* השמטת מאפיינים בעלי איכות נמוכה מדי- נבחר להסיר את המאפיין “Leg room service” מכיוון שלא ניתן להשלים את החוסרים בצורה איכותית עקב כמות הרשומות החסרות, שהן כ-30% מהנתונים. נוכחות כה רבה של נתונים חסרים עשויה לפגוע בדיוק ובאיכות המודל.
* השמטת תצפיות בעלות חוסר רב- נבחר להסיר את 2 התצפיות בעלות חוסר רב של מאפיינים (יותר מ-4 מאפיינים חסרים), מהסיבה שתצפיות אלו אינן מועילות והן עלולות לפגוע באיכות התוצאות ודיוק המודל. בנוסף, נבחר להסיר תצפיות בעלות חוסר במשתנה המטרה מכיוון שלא נוכל ללמוד עליהן ולא תורמות
* השלמה מושכלת של ערכים חסרים- הסבר על אופן השלמת הנתונים ([נספח 3](#_נספח_3-_אופן)).   
  בנוסף, שמירת שני המשתנים " Departure Delay in Minutes" ו "Arrival Delay in Minutes" מאפשרת לנו לנתח את הגורמים לעיכובים בצורה מעמיקה יותר ולהבין את השפעתם על חוויית הטיסה. למרות הקורלציה הגבוהה ביניהם, כל משתנה משקף היבטים שונים של העיכובים, והם יכולים לספק תובנות נוספות במודלים מתקדמים של למידת מכונה. בנוסף, שליפת הנתונים של מאפיין "Departure Delay in Minutes" יכולה לשמש ככלי יעיל להשלמת הנתונים החסרים במאפיין "Arrival Delay in Minutes" בשל הקורלציה הגבוהה ביניהם.

## **טיפול פרטני במאפיינים**

* דיסקריטיזציה-

הגרף מראה התפלגות רחבה של גילאים בין 0 ל-80 ומעלה, ולכן מומלץ לבצע דיסקריטיזציה של המשתנה "גיל". חלוקה לקבוצות גיל מובחנות (0-18, 19-30, 31-40, 41-50, 51-60, 61+) תאפשר ניתוח נתונים פשוט יותר, תקטין רעש נתוני, ותשפר את ההבנה העסקית. בחרנו לבצע דיסקריטיזציה דווקא על "גיל" כיוון שההתפלגות הרחבה של גילאים מקשה על זיהוי מגמות ודפוסים כאשר הגיל נמדד כמשתנה רציף. דיסקריטיזציה תקל על השוואת נתונים בין קבוצות שונות, תאפשר זיהוי מגמות ודפוסים ברורים, ותסייע בביצוע ניתוחים סטטיסטיים אפקטיביים ובשיפור שירותים מותאמים לקבוצות גיל שונות.

בחרנו שלא לבצע דיסקריטיזציה על משתנים אחרים מכיוון שרבים מהם בדידים ובעלי ערכים מוגבלים (כמו Inflight Wi-Fi service, Seat comfort, Cleanliness), מה שמאפשר ניתוח ישיר ללא צורך בהמרה. בנוסף, בחרנו שלא לבצע דיסקריטיזציה על משתנים רציפים אחרים כמו Flight Distance מאפשרים ניתוחים סטטיסטיים מתקדמים ושמירה על רציפות הנתונים, מה שמספק תמונה מדויקת יותר של הנתונים ומאפשר הבנת השפעות ומתאמים טוב יותר.

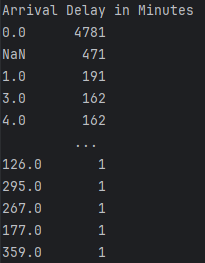
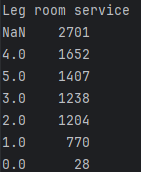
* גזירת מאפיינים חדשים-

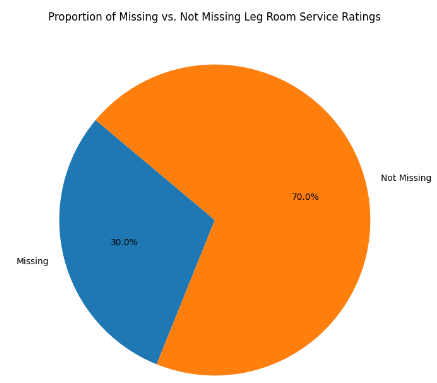
Average Service Rating:

* ממוצע של כל דירוגי השירות:  
   (Inflight service, On-board service ,Check-in service ,Cleanliness, Food and drink).
* משתנה נועד לספק מדד כולל על חוויית השירות של הנוסעים. בחרתי לחבר את הקריטריונים הללו למאפיין אחד מכיוון שהם מייצגים לדעתנו את ההיבטים המרכזיים והחשובים ביותר של חוויית הנוסע במהלך הטיסה. מדד זה מאפשר להבין כיצד מכלול משתני השירות הללו תורם לשביעות הרצון הכוללת, לזהות מגמות ודפוסים רחבים יותר, לפשט את ניתוח הנתונים.

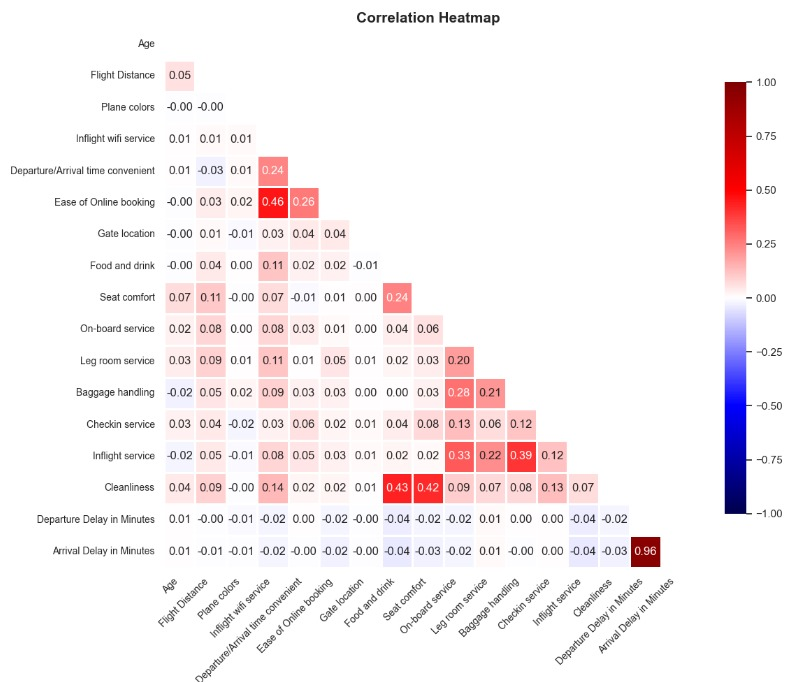
# **נספחים**

## **נספח 1- נתונים חסרים:**





## **נספח 2- גרף בדיקת קורלציה של המשתנים:**



## **נספח 3- אופן השלמת הנתונים:**

* Satisfaction לא נשלים מאחר וזה המשתנה עליו אנו מעוניינים ללמוד.
* Gender נוסיף לקטגוריית- לא ידוע, אנו מניחים שיש נוסעים שלא מעוניינים לציין את מינם ולכן החליטו לא לסמן.
* Inflight wi-fi service נשלים ל-0 (טיסה שלא היה בה Wi-fi).
* Class: נשלים ל- לא ידוע, מפני שכבר קיימת קטגוריה של לא ידוע, אז כל נוסע שלא מילא את המחלקה שלו, נכניס ל- לא ידוע.
* משתנים קטגוריאליים (Customer Type, Type of Travel) **נשלים לפי יחס על מנת לשמור על אופי ההתפלגות שקיים בנתונים.**
* משתנים רציפים Age), Flight distance, (Departure Delay in Minutes **נשלים לפי ממוצע מאחר ממוצע לוקח בחשבון את כל הערכים הקיימים במשתנה, ובכך מנצל את כל המידע כדי לספק מדד מרכזי מדויק.**
* משתנים בדידים (רציפים) (Departure/Arrival time convenient, Ease of Online booking, Gate location,Food and drink , Seat comfort, On-board service, Leg room service, Baggage handling, Check-in service, Inflight service, Cleanliness) **נשלים לפי חציון על מנת לשמור על מבנה ההתפלגות.**
* Arrival Delay in Minutes- עבור כל החוסרים במאפיין זה נשלים באמצעות המאפיין Departure Delay in Minutes"" מכיוון שהראינו כי יש קורלציה גבוהה בין שני המאפיינים נוכל לשכפל את הנתונים.