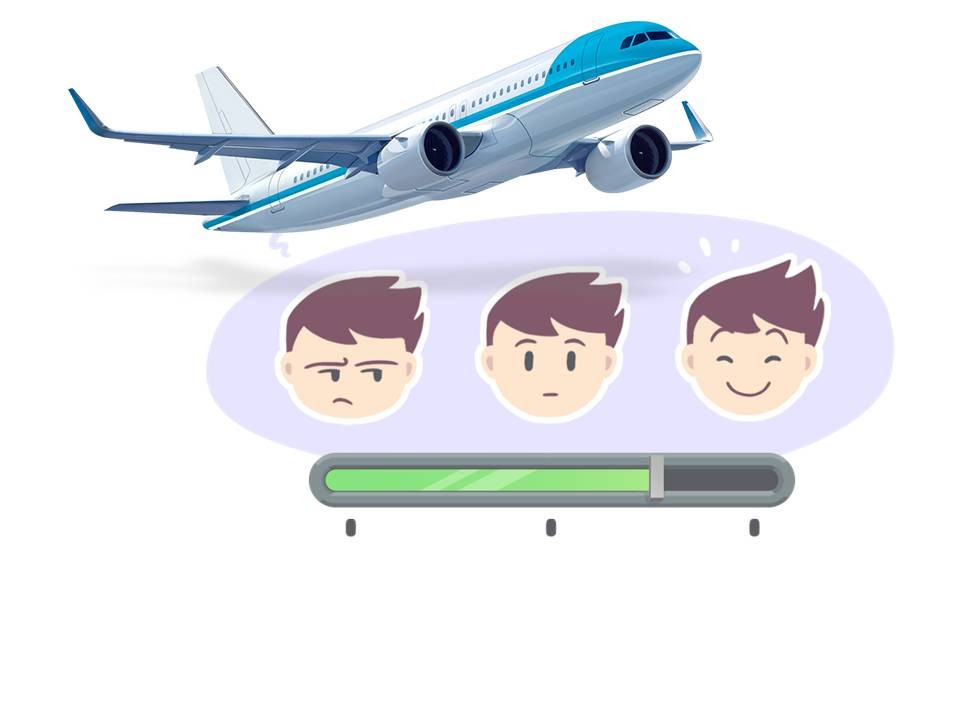
**פרויקט - חלק א'**



תוכן עניינים

[1. תיאור כללי- 3](#_Toc169696790)

[2. שאלת מחקר- 3](#_Toc169696791)

[1. תיעוד מקורות הנתונים ומשמעותם- 3](#_Toc169696792)

[1.1. Satisfaction (משתנה מוסבר, יח' מידה: NA, סוג: קטגוריאלי) 3](#_Toc169696793)

[1.2. gender (משתנה מסביר, יח' מידה: NA, סוג: קטגוריאלי) 3](#_Toc169696794)

[1.3. Age (משתנה מסביר, יח' מידה: שנים, סוג: רציף) 4](#_Toc169696795)

[1.4. Customer Type (משתנה מסביר, יח' מידה: NA, סוג: קטגוריאלי) 4](#_Toc169696796)

[1.5. Type of Travel (משתנה מסביר, יח' מידה: NA, סוג: קטגוריאלי) 4](#_Toc169696797)

[1.6. Class (משתנה מסביר, יח' מידה: NA, סוג: קטגוריאלי) 4](#_Toc169696798)

[1.7. Flight distance (משתנה מסביר, יח' מידה: קילומטר, סוג: רציף) 4](#_Toc169696799)

[1.8. Plane colors (משתנה מסביר, יח' מידה: NA, סוג: שלם, רציף) 4](#_Toc169696800)

[1.9. Inflight Wi-Fi service (משתנה מסביר, יח' מידה: NA, סוג: בדיד) 4](#_Toc169696801)

[1.10. Departure / Arrival convenient (משתנה מסביר, יח' מידה: NA, סוג: בדיד) 4](#_Toc169696802)

[1.11. Gate location (משתנה מסביר, יח' מידה: NA, סוג: בדיד) 4](#_Toc169696803)

[1.12. Food and drink (משתנה מסביר, יח' מידה: NA, סוג: בדיד) 4](#_Toc169696804)

[1.13. Seat comfort (משתנה מסביר, יח' מידה: NA, סוג: בדיד) 4](#_Toc169696805)

[1.14. On-board service (משתנה מסביר, יח' מידה: NA, סוג: בדיד) 4](#_Toc169696806)

[1.15. Ease of Online booking (משתנה מסביר, יח' מידה: NA, סוג: בדיד) 5](#_Toc169696807)

[1.16. Leg room service (משתנה מסביר, יח' מידה: NA, סוג: בדיד) 5](#_Toc169696808)

[1.17. Baggage handling (משתנה מסביר, יח' מידה: NA, סוג: בדיד) 5](#_Toc169696809)

[1.18. Inflight service (משתנה מסביר, יח' מידה: N/A, סוג: בדיד) 5](#_Toc169696810)

[1.19. Check-in service (משתנה מסביר, יח' מידה: N/A, סוג: בדיד) 5](#_Toc169696811)

[1.20. Cleanliness (משתנה מסביר, יח' מידה: N/A, סוג: בדיד) 5](#_Toc169696812)

[1.21. Departure Delay in Minutes (משתנה מסביר, יח' מידה: דקות, סוג: רציף) 5](#_Toc169696813)

[1.22. Arrival Delay in Minutes (משתנה מסביר, יח' מידה: דקות, סוג: רציף) 5](#_Toc169696814)

[**2.1.** הסתברויות אפריוריות 6](#_Toc169696815)

[2.2. איזון סט הנתונים 9](#_Toc169696816)

[2.3. קשרים מעניינים בין מאפיינים 9](#_Toc169696817)

[2.4. מאפיינים החשודים כבעלי השפעה על משתנה המטרה 10](#_Toc169696818)

[2.5. בדיקת מאפיינים החשודים כבעלי השפעה 10](#_Toc169696819)

[2.6. מאפיינים שניתן להסיר 11](#_Toc169696820)

[3. איכות הנתונים- 11](#_Toc169696821)

[נספחים 13](#_Toc169696822)

**הגדרת הבעיה**

# תיאור כללי-

מערך הנתונים שלנו מכיל סקר שביעות רצון של נוסעי חברת התעופה. נדרש לבדוק אילו גורמים נמצאים בקורלציה גבוהה לשביעות רצון הנוסע במגוון פרמטרים.

מחקרים קודמים השתמשו במודלים של למידת מכונה כמו K-Nearest Neighbors, Decision Tree Classifier, Logistic Regression, Random Forest, Naive Bayes ו-AdaBoost כדי לחזות ולנתח רמות שביעות רצון של נוסעים בהתבסס על תכונות שונות של שירותי התעופה[[1]](#endnote-1). גורמים שנמצאו בקורלציה גבוהה לנוסעים מרוצים כללו היבטים כמו רישום מקוון לעליה למטוס (online boarding), בידור בטיסה, נוחות המושבים, שירות במהלך הטיסה, מרווח לרגליים, ניקיון, מרחק טיסה והימצאות Wi-Fi בטיסה[[2]](#endnote-2). בנוסף, נמצא כי תכונות אישיות של צוות המטוס (פתיחות, נעימות ומצפוניות) משפיעות על שביעות הרצון והנאמנות של הנוסעים לחברות התעופה, המחקר הדגיש את חשיבות איכות השירות בהשפעה על שביעות רצון הנוסעים[[3]](#endnote-3).יתר על כן, מחקרים ניתחו את שביעות רצון הלקוחות בתעשיית התעופה על ידי התחשבות בגורמים כמו גיל, עיכובים בטיסה, נוחות זמני יציאה והגעה, שירות צ'ק-אין, מזון ומשקאות וטיפול במטען, שכולם הראו קשרים חיוביים עם דירוג הנוסעים[[4]](#endnote-4).

# שאלת מחקר-

כיצד נתוני הטיסה ושירותים נלווים משפיעים על שביעות הרצון הכוללת של הטסים, ולאילו מהגורמים יש את ההשפעה המשמעותית ביותר על רמת שביעות הרצון של סוגי לקוחות במחלקות הנוסעים השונות?

**הבנת הנתונים**

# תיעוד מקורות הנתונים ומשמעותם-

## Satisfaction (משתנה מוסבר, יח' מידה: NA, סוג: קטגוריאלי)

המשתנה עונה על השאלה האם הנוסע מרוצה: כן- Satisfaction, לא-neutral or dissatisfaction. הנתון נאסף באמצעות סקר, שבו הנוסעים מציינים את רמת שביעות הרצון הכללית שלהם.

## gender (משתנה מסביר, יח' מידה: NA, סוג: קטגוריאלי)

המשתנה עונה על השאלה האם מדובר בזכר או נקבה. הנתון נאסף בד"כ באמצעות טופס סקר שבו נוסעים מציינים את מינם.

## Age (משתנה מסביר, יח' מידה: שנים, סוג: רציף)

גיל הנוסע בשנים, נאסף במהלך ההרשמה לטיסה או מפרטי חשבון המשתמש.

## Customer Type (משתנה מסביר, יח' מידה: NA, סוג: קטגוריאלי)

סוג לקוח: האם הוא לקוח נאמן או לא. הערך נקבע מתוך רישומי תוכנית הנאמנות של חברת התעופה ולפי דפוסי הרכישה שלו.

## Type of Travel (משתנה מסביר, יח' מידה: NA, סוג: קטגוריאלי)

סוג הנסיעה : האם מדובר בנסיעה למטרה אישית או עסקית. המידע נאסף באמצעות סקר בו מתבקש הנוסע לציין אם נסיעתו היא מסיבות אישיות או עסקיות.

## Class (משתנה מסביר, יח' מידה: NA, סוג: קטגוריאלי)

מחלקת נסיעות בטיסה. המידע נאסף מנתוני הזמנת הכרטיס, המציין אם הנוסע נסע במחלקת עסקים, אקונומי או אקונומי פלוס.

## Flight distance (משתנה מסביר, יח' מידה: קילומטר, סוג: רציף)

מרחק הטיסה, מחושב על סמך נקודות המוצא והיעד של הטיסה, שמקורן לרוב מנתוני הטיסה של חברת התעופה.

## Plane colors (משתנה מסביר, יח' מידה: NA, סוג: שלם, רציף)

מספר צבעים על גוף המטוס, נרשם בדרך כלל מרישומי חברת התעופה ומהנתונים על המטוס המשמש לטיסה.

## Inflight Wi-Fi service (משתנה מסביר, יח' מידה: NA, סוג: בדיד)

רמת שביעות רצון עבור שירות ה-Wi-Fi בטיסה תדורג מ 1 עד 5 ואפס אם לא רלוונטי, נאסף באמצעות סקר בו הנוסעים נשאלים על חווית השימוש ב-Wi-Fi.

## Departure / Arrival convenient (משתנה מסביר, יח' מידה: NA, סוג: בדיד)

רמת שביעות הרצון של זמן יציאה/הגעה. נאסף באמצעות סקר, כאשר הנוסעים מדרגים את נוחות זמני היציאה וההגעה.

## Gate location (משתנה מסביר, יח' מידה: NA, סוג: בדיד)

רמת שביעות הרצון ממיקום השער. המידע נאסף באמצעות סקר, כאשר הנוסעים מדורגים את שביעות רצונם ממיקום השער.

## Food and drink (משתנה מסביר, יח' מידה: NA, סוג: בדיד)

רמת שביעות הרצון מהמזון והשתייה בטיסה. המידע נאסף באמצעות סקר, שבו מדרגים הנוסעים את שביעות הרצון שלהם משירותי האוכל והשתייה בטיסה.

## Seat comfort (משתנה מסביר, יח' מידה: NA, סוג: בדיד)

רמת שביעות הרצון מנוחות המושב. המידע נאסף באמצעות סקר, כאשר הנוסעים מדורגים את נוחות המושב שלהם.

## On-board service (משתנה מסביר, יח' מידה: NA, סוג: בדיד)

רמת שביעות הרצון משירות אנשי הצוות במהלך טיסה. המידע נאסף באמצעות סקר, שבו הנוסעים מדרגים את השירות במהלך הטיסה.

## Ease of Online booking (משתנה מסביר, יח' מידה: NA, סוג: בדיד)

רמת שביעות הרצון של הלקוח מחוויית ההזמנה המקוונת. המידע נאסף באמצעות שאלון אינטרנטי קצר, שבו הנוסעים מדרגים את קלות ההזמנה באינטרנט.

## Leg room service (משתנה מסביר, יח' מידה: NA, סוג: בדיד)

רמת שביעות הרצון ממרווח הרגליים. מידע נאסף באמצעות סקר, כאשר הנוסעים מדרגים את מרווח הרגליים במושביהם.

## Baggage handling (משתנה מסביר, יח' מידה: NA, סוג: בדיד)

רמת שביעות הרצון מהטיפול בכבודה. המידע נאסף באמצעות סקר, בו הנוסעים מדרגים את שביעות רצונם מהטיפול בכבודה.

## Inflight service (משתנה מסביר, יח' מידה: N/A, סוג: בדיד)

רמת שביעות הרצון מהשירותים הניתנים במהלך בטיסה. המידע נאסף באמצעות סקר, בו הנוסעים מדרגים את רמת השירות בטיסה בהתבסס על איכות ומגוון השירותים הניתנים.

## Check-in service (משתנה מסביר, יח' מידה: N/A, סוג: בדיד)

רמת שביעות הרצון משירות הצ'ק אין. המידע נאסף באמצעות סקר, בו הנוסעים מדרגים את שירות הצ'ק-אין.

## Cleanliness (משתנה מסביר, יח' מידה: N/A, סוג: בדיד)

רמת שביעות רצון של מהניקיון בטיסה. המידע נאסף באמצעות סקר, שבו הנוסעים מדרגים את ניקיון המטוס.

## Departure Delay in Minutes (משתנה מסביר, יח' מידה: דקות, סוג: רציף)

דקות עיכוב בעת היציאה, נרשם מנתוני התפעול של הטיסה ממאגרי הנתונים בחברת התעופה המציינים את מספר דקות העיכוב בהמראה.

## Arrival Delay in Minutes (משתנה מסביר, יח' מידה: דקות, סוג: רציף)

דקות עיכוב בעת ההגעה, נרשם מנתוני התפעול של הטיסה ממאגרי הנתונים בחברת התעופה המציינים את מספר דקות העיכוב בהגעה.

1. **הסתברויות אפריוריות וקשרים בין מאפיינים –**

## הסתברויות אפריוריות

תמונה שמכילה טקסט, צילום מסך, תצוגה, תרשים

התיאור נוצר באופן אוטומטיSatisfaction:

סט הנתונים מאוזן, וכולל 43.6% נוסעים שמרוצים מטיסה ו-56.4% נוסעים שהם ניטרליים או שאינם מרוצים מהטיסה. מצב זה מאוזן מבחינת הנתונים, ואף אופייני להתפלגות באוכלוסייה הכללית[[5]](#endnote-5).

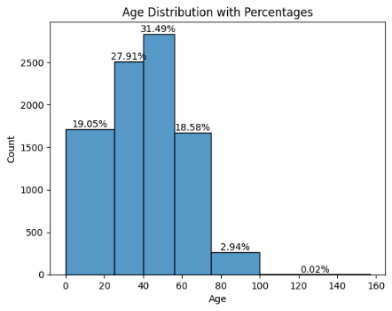
[1]

Age:

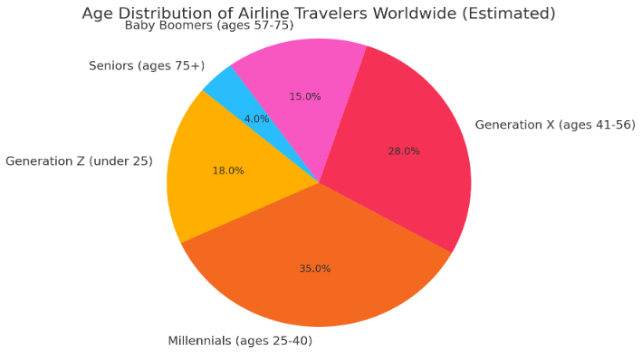
בסט הנתונים שלנו התפלגות הגיל תואמות את ההתפלגות הכללית באוכלוסיית הטסים[[6]](#endnote-6), בסט נתונים שלנו 19.05% מתחת לגיל 25 ובאוכלוסייה הכללית 18.0%, כמעט זהה. עבור קבוצת הגילאים 25-40 הייצוג בסט הנתונים שלנו קרוב לאוכלוסייה הכללית, אם כי לא מדויק, כך שבסט 27.91% נוסעים בגילאי 25-40 לעומת 35% באוכלוסייה הכללית. ביחס לקבוצת הגיל 41-56 בסט הנתונים יש 31.49% לעומת 28% באוכלוסייה הכללית, ובקבוצת הגיל 57-75 בסט הנתונים שלנו יש 18.58% לעומת 15% באוכלוסייה הכללית, ואלו נתונים דומים אם כי לא זהים. המצב דומה עבור קבוצת הגיל 80-100, בה לסט הנתונים שלנו לקבוצה זו פרופורציה של 2.94% לעומת 4% באוכלוסייה הכללית. אנו מעריכים כי הנתונים דומים דיו ומציגים התפלגות בעלת חלוקה דומה על מנת להוות בסיס מתאים להערכה במובן של משתנה הגיל.

[1]

סט הנתונים



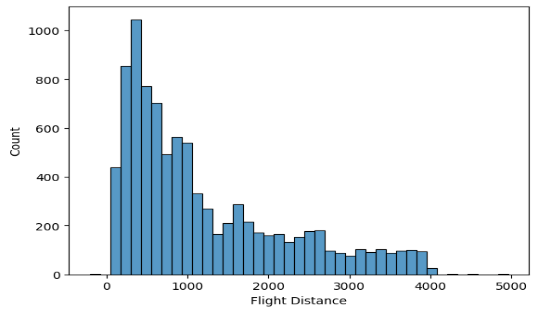
האוכלוסייה הכללית



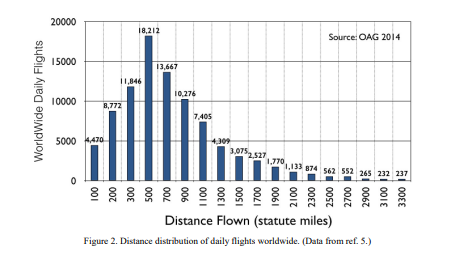
Flight Distance:

ההתפלגות בסט הנתונים שלנו דומה לאוכלוסייה הכללית[[7]](#endnote-7), כאשר שתיהן בעלות זנב ימני.

[1]



סט הנתונים

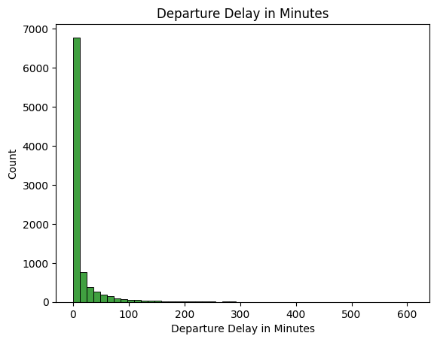
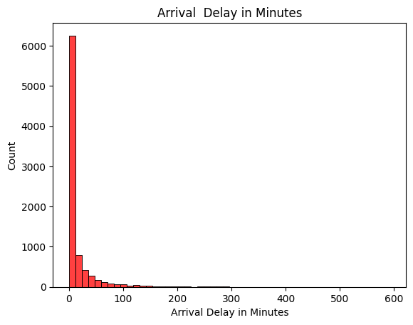


האוכלוסייה הכללית

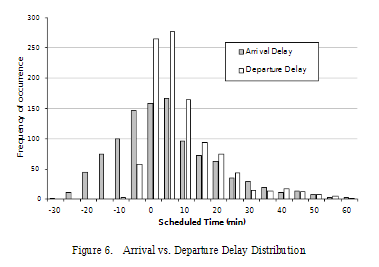
Delay Arrival & Delay Departure (התייחסות לשני המשתנים):

ניתן לראות בסט הנתונים ששני המשתנים בעלי התפלגות עם זנב ימני בדומה להתפלגות באוכלוסייה הכללית[[8]](#endnote-8). באוכלוסייה הכללית קיימים ערכים שליליים שמשמעותם הקדמה בזמן הטיסה, בניגוד לערכים הקיימים בסט הנתונים שלנו שהם כולם חיוביים ומייצגים איחורים בלבד, המשמעות היא שנוכל לעקוב אחר השפעת איחורי הטיסות על משתנה המטרה בצורה נאמנה למציאות, אך לא נדע להסביר את הקשר בין טיסות שהוקדמו למשתנה המטרה.

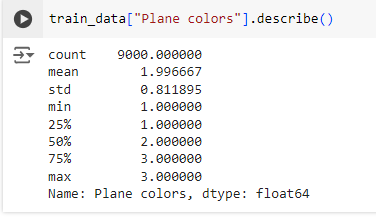
[1]



סט הנתונים



האוכלוסייה הכללית

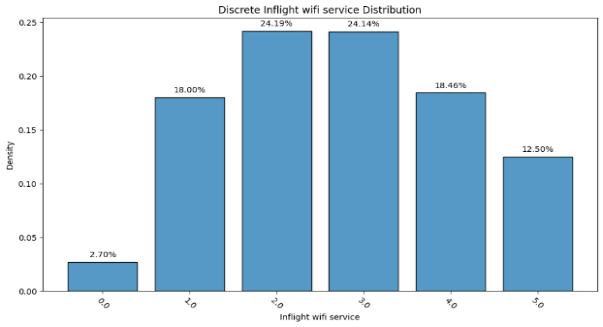
Plane Color:

לפי סט הנתונים שבידינו ממוצע הצבעים של מטוס הוא 2. מטוס נוסעים מסחרי מורכב בדרך כלל משניים עד שלושה צבעי יסוד בממוצע[[9]](#endnote-9).לפי סט הנתונים שלנו אנו רואים שהנתון מתיישב עם האוכלוסייה הכללית.

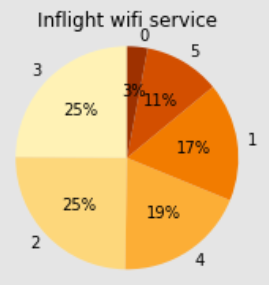
Inflight Wi-Fi service:

סט הנתונים מציג פרופורציות דומות לאלו שנמצאות באוכלוסייה בכללית[[10]](#endnote-10) עבור דירוג שירותי האינטרנט בטיסה (ממיפוי שנערך תחת ניסוי עבור אפיון שירותי הקישוריות בטיסות). ניתן לראות שפרופורציית הדירוגים דומה לזו שבאוכלוסייה הכללית בכל קבוצת הערכים.

[1]



סט הנתונים

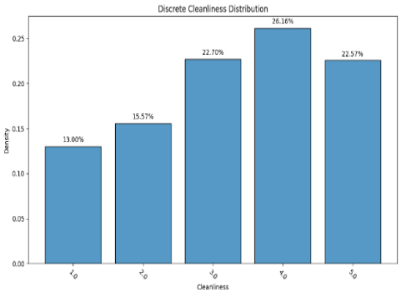


האוכלוסייה הכללית

Cleanliness:

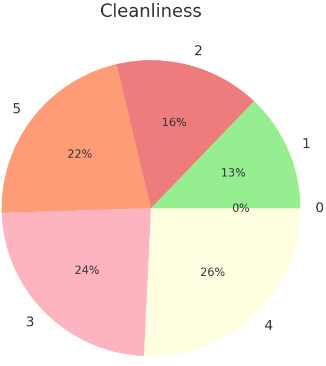
[1]

סט הנתונים מציג פרופורציות דומות לאלו שנמצאות באוכלוסייה הכללית[[11]](#endnote-11) עבור דירוג שירותי הניקיון בטיסה (ממיפוי של האגודה האמריקאית לשביעות רצון צרכנית). ניתן



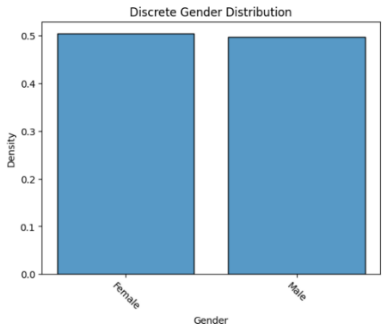
סט הנתונים

האוכלוסייה הכללית

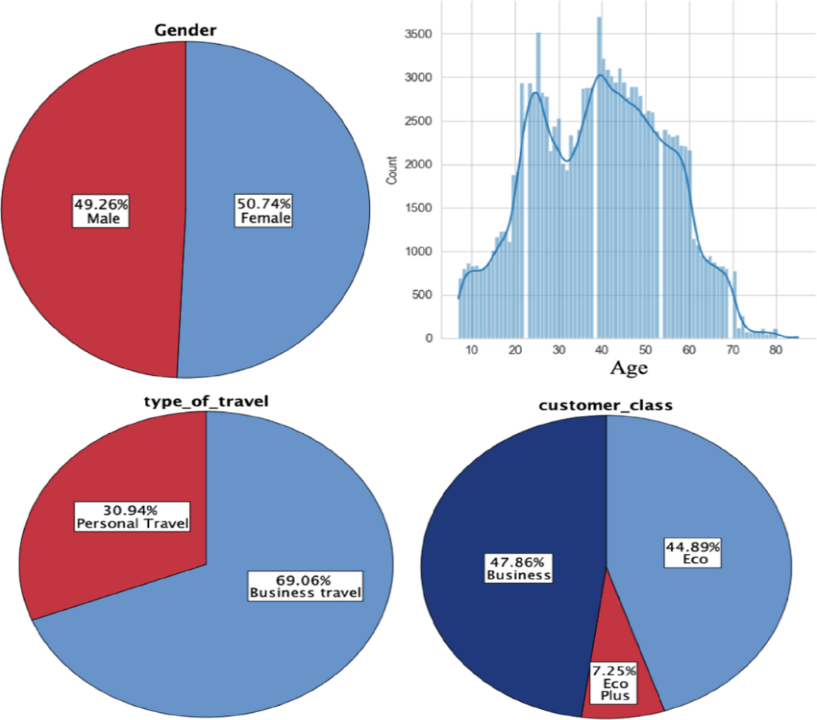


לראות שפרופורציית הדירוגים דומה לזו שבאוכלוסייה הכללית בכל קבוצת הערכים.

[1]



סט הנתונים



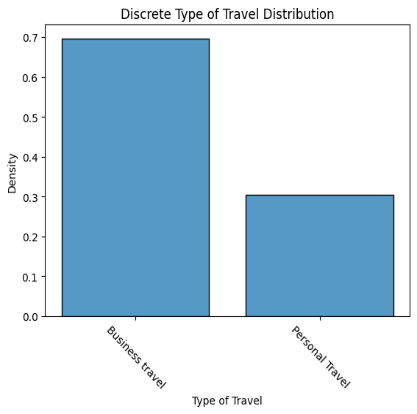
האוכלוסייה הכללית

Gender:

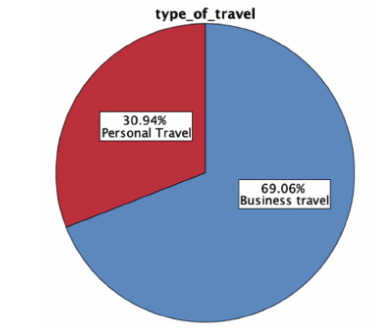
סט הנתונים  מייצג איזון בין גברים לנשים, ובכך מתאר את אותה ההתפלגות של אוכלוסיית הטסים הכללית[[12]](#endnote-12).

[1]

Travel Type:



סט הנתונים



האוכלוסייה הכללית

סט נתונים שלנו מתאר פרופורציה של כמעט 30% אחוז מהלקוחות כטסים למטרה אישית בהשוואה ל-70% בקירוב שהם לקוחות הטסים למטרה עסקית. מצב זה משקף באופן די מדויק את המצב באוכלוסייה הכללית בה 30.94% מהלקוחות טסים למטרה אישית ו-69.06% מהלקוחות למטרה עסקית[[13]](#endnote-13).

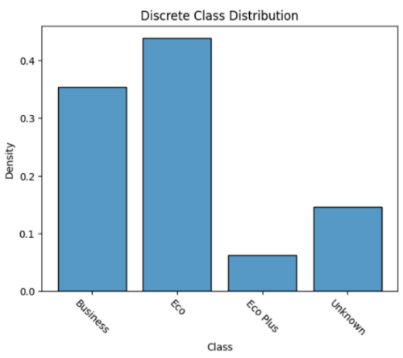
[1]

Class Distribution:

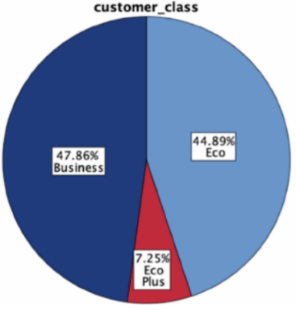
בסט הנתונים שלנו ניתן לזהות פרופורציה דומה של טסים במחלקת Eco לזו שבאוכלוסייה הכללית עם כ-45% מהאוכלוסייה. גם מחלקת Eco Plus מציגה פרופורציות זהות לאלו של האוכלוסייה הכללית. מחלקת Business מציגה שיעור נמוך באופן ניכר מזה של האוכלוסייה הכללית[[14]](#endnote-14), אך ייתכן וניתן להסביר את ההפרש למול הרשומות שאינן ידועות בסט הנתונים שלנו (Unknown). על פי התרשמות ויזואלית בלבד נבחר לקבוע שהחלוקה למחלקות בסט הנתונים שלנו אכן משקפת את האוכלוסייה הכללית (אל הרשומות החסרות נתייחס בסעיפים הבאים).

[1]

סט הנתונים



האוכלוסייה הכללית

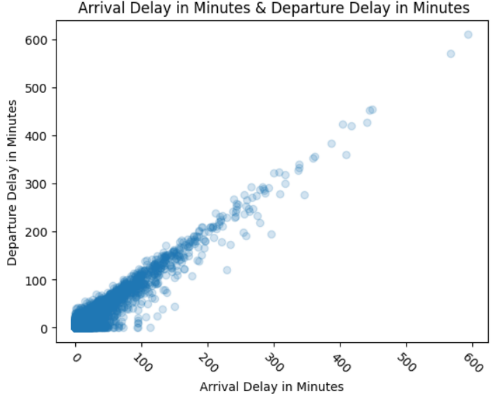


## איזון סט הנתונים

בהתבסס על בחינת ההסתברויות האפריוריות של המשתנים בסט הנתונים ועל בסיס הדמיון הרב בין פרופורציות הנתונים שהוצגו לבין האוכלוסייה הכללית אנו מניחים שסט הנתונים מאוזן ואכן מייצג נאמנה את המציאות.

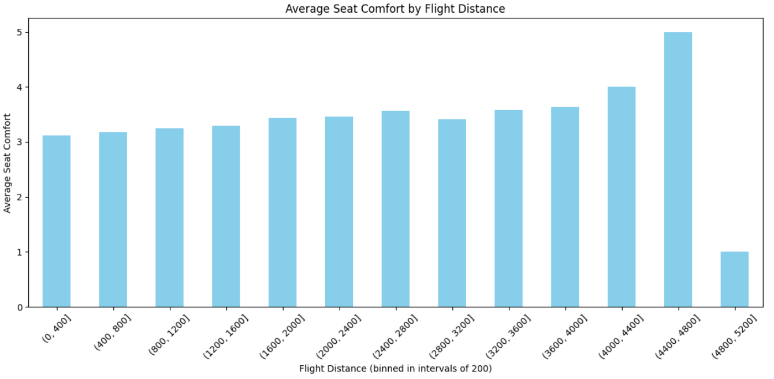
## קשרים מעניינים בין מאפיינים

Arrival and Departure Delay in Minuets:

בחרנו לבחון תחילה את הקשר בין זמן העיכוב בהמראה לבין זמן העיכוב בנחיתה. שיערנו שקיים קשר לינארי ישיר בין שני מאפיינים אלו עבור כל טיסה, ואכן ניתן לראות קשר לינארי חזק בין המאפיינים. המשמעות היא שטיסות שמתעכבות בהמראה לרוב טסות את משך הטיסה המתוכנן שלהן ולכן נוצר עיכוב זהה בנחיתה. סביר להניח שהקשר החזק בין המאפיינים מצביע על התאמת יתר וכפילות נתונים , ולכן נשקול את נחיצות שני המאפיינים בהמשך על מנת לבחון האם ניתן לפשט את המודל.

[2]

Average Seat Comfort by Flight Distance:

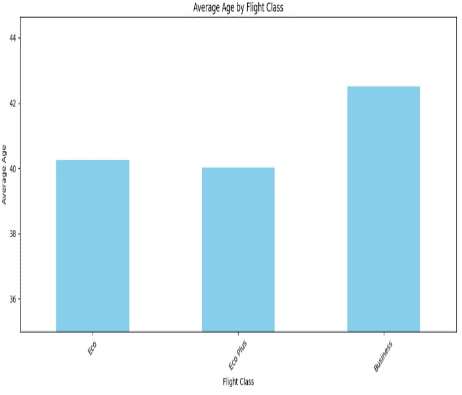
בחרנו לבחון את הקשר בין ציון נוחות הישיבה הממוצעת כתלות במרחק הטיסה:

היינו מצפים שככל שהטיסה ארוכה יותר הלקוחות ירגישו נינוחים פחות עקב ישיבה ממושכת, אך ניתן לראות שהגרף מציג מגמה הפוכה. ככל שאורך הטיסה עולה כך גם הדירוג הממוצע עולה. ניתן לשער שהסיבה לכך היא איכות מושבים ואיכות טיסה עדיפים במטוסים הנדרשים לבצע טיסות ממושכות. ניתן להניח אם כך, שנוחות המושבים אינה מאפיין שמשמש כמשתנה מתווך בין מרחק הטיסה לבין שביעות הרצון (משתנה המטרה) ועל כן לא נשקול לנקות אותו מהמודל.

[3]

[3]

Average Age by Flight Class:

בחרנו לבדוק את הקשר בין מחלקות הטיסה והגיל הממוצע של הנוסעים בהן. ההנחה שלנו היא שיש קשר בין גיל הנוסעים (ובעקיפין מצבם הכלכלי) לבין הסטנדרטים שהנוסעים דורשים בטיסה שלהם, ובהתאם בחורים מחלקות. ניתן לראות בגרף כי הגיל הממוצע במחלקות אקו ואקו פלוס כמעט זהה בעוד שהגיל הממוצע במחלקת העסקים גבוה יותר, ולכן הנחתנו נכונה ככל הנראה. למרות זאת, עדיין מוקדם לקבוע האם קיימים קשרי תיווך מובהקים בין העדפות הנוסעים הנובעת מגילם לבין המחלקות השונות, והאם צפויה השפעה מהותית על המודל.

Type of Travel and Class:

תמונה שמכילה טקסט, צילום מסך, תרשים, מלבן

התיאור נוצר באופן אוטומטיבחרנו לבחון את הפיזור של הנוסעים בטיסות אישיות ועסקיות למול מחלקות הטיסה השונות. הגרף מסמל את הקשרים בין המחלקה לסוג הנוסע. ניתן לראות כמצופה שרוב מוחלט מנוסעי העסקים טסים במחלקת עסקים ורוב הנוסעים הפרטיים נוסעים באקו. נתון אחד מפתיע הוא שיש יותר נוסעים עסקיים במחלקת אקו מאשר במחלקת אקו+. ניתן ללמוד מקשר זה בין המאפיינים שהנוסעים מתנהגים לרוב כפי שציפינו, ולא ניתן לזהות מגמות חריגות שעשויות לעוות את המודל.

[2]

## מאפיינים החשודים כבעלי השפעה על משתנה המטרה

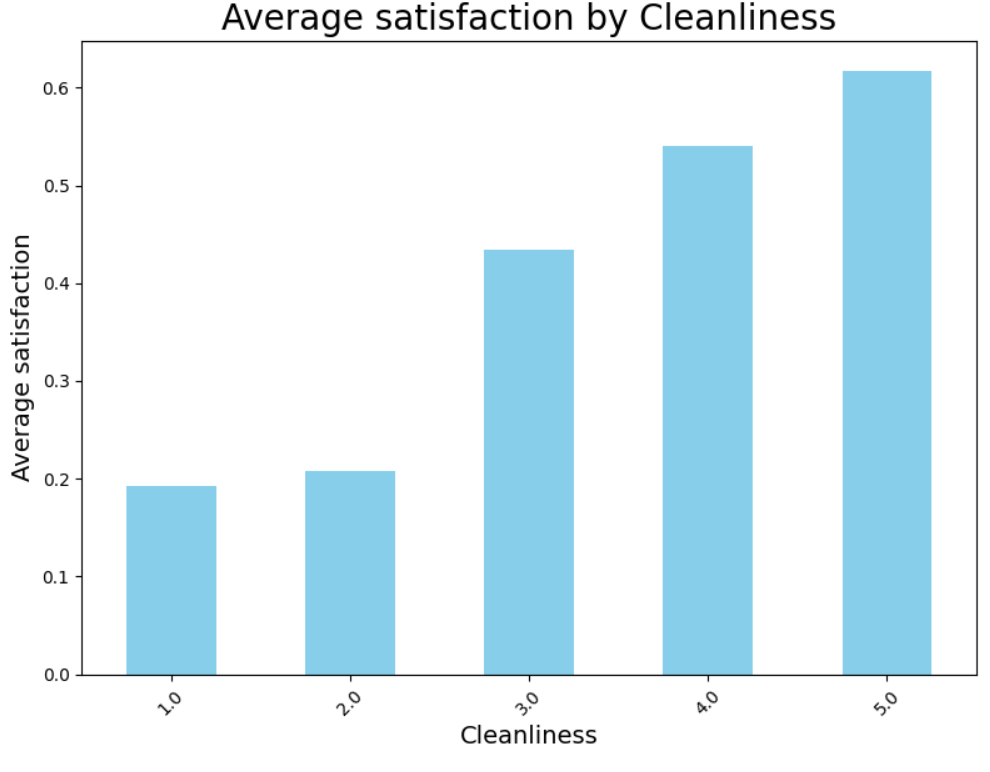
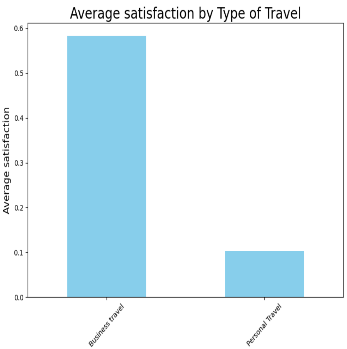
Cleanliness: חוות דעתם של הנוסעים על ניקיון הטיסה היא חיונית בהשפעתה על שביעות רצון הנוסעים מעצם השפעתה הישירה על התחושה הכללית שלהם במהלך הטיסה. תחושה של חוסר ניקיון עשויה להשפיע לרעה גם על תפיסת הנוסעים לגבי מאפיינם אחרים כמו מזון ושתייה או נוחות המושב. ניתן להסביר זאת באמצעות היגיון ולהניח שהנוסעים לא ישמחו לשבת על כיסא מלוכלך גם אם הוא נח, או לאכול ולשתות במקום לא נקי.

Type of Travel: הקשר בין סוג הנסיעה לבין שביעות רצון הנוסעים הוא קשר הגיוני שכדאי לבדוק. סביר להניח שנוסעים למטרות עסקיות לרוב אינם משלמים את עלות הטיסה מכיסם הפרטי, ולכן הציפיות שלהם מהתמורה עשויות להשתנות בהתאם. לכן ייתכן ונוסעים אלו יהיו פחות ביקורתיים במידת שביעות רצונם בהשוואה לנוסעים למטרות אישיות.

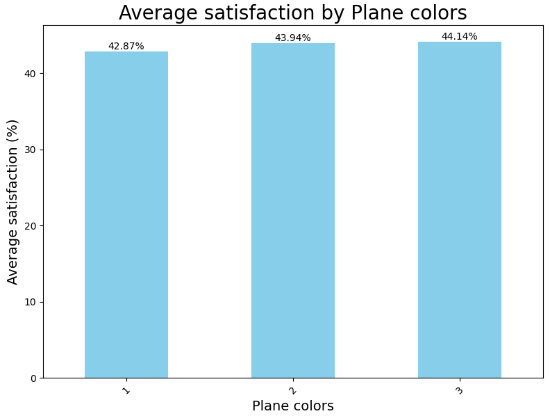
## בדיקת מאפיינים החשודים כבעלי השפעה

בבדיקה סטטיסטית של שני המשתנים שציינו אל מול משתנה המטרה ניתן לראות שאכן הקורלציה בין כל אחד מהם לבין משתנה המטרה היא גבוה ועומדת על 32% ו-44% בהתאמה. ניתן לראות בגרפים המצורפים בנספח שמידת שביעות הרצון של הנוסעים בנסיעות עסקיות אכן גבוהה משמעותית בהשוואה לאלו הנוסעים למטרות אישיות. ניתן גם לראות קשר מובהק בין מידת שביעות הרצון הממוצע לבין מידת הניקיון שדרגו הנוסעים.

[3]



## מאפיינים שניתן להסיר

 Plane colorהוא המאפיין הראשון שבחרנו להסיר. מבחינת ההשפעה הישירה שלו על משנה המטרה Satisfaction, ניתן לראות בגרף המצורף שאין שוני משמעותי בשביעות הרצון הממוצעת כתלות במספר הצבעים בהם המטוס צבוע. בבחינות נוספות שביצענו ניתן לראות שהקורלציה בין מאפיין זה לבין שאר המאפיינים נמוכה מאוד, בדגש על הקורלציה עם משתנה המטרה שהיא 1% בלבד. מדד P-value של Plane color בניתוח רגרסיה מרובה הוא אמנם נמוך בעל ערך של 2.87%, אך מרבית המאפיינים בעלי ערכים נמוכים ממנו. מדד ה- t-Valueשל Plane color בניתוח רגרסיה מרובה הינו נמוך בעל ערך של 2.87, מה שמצביע על השפעה נמוכה בין המאפיינים. בנוסף לכל האמור, לא מצאנו ספרות אקדמית שקושרת קשר מהותי בין צבעי המטוס לרמת שביעות הרצון של הנוסעים או למאפיינים אחרים במודל בסט הנתונים שלנו.

[3]

Arrival Delay in Minuets הוא המאפיין השני שבחרנו להסיר. מאפיין זה מציג קורלציה של 0.96 עם המשתנה Departure Delay in Minuets מה שעשוי להצביע על התאמת יתר בין המשתנים, בעוד שהקורלציה עם שאר המאפיינים נמוכה מאוד. הגרף שהצגנו בסעיף 2.3 מראה גם הוא קשר לינארי מובהק בין המשתנים. העבודות הללו בנוסף לכך שישנן כ-470 רשומות חסרות עבור מאפיין זה, מביאות אותנו להנחה שנכון להסיר אותו מסט הנתונים על מנת לפשט את המודל.

# איכות הנתונים-

המשתנה Leg room service הוא בעל מספר רב של ערכים ריקים (Nulls). לא נרצה לאבד את שאר הנתונים של נוסעים ברשומות החסרות, ולכן נזין במקום הרשמות החסרות את הערך השכיח. נבחר דווקא בהשלמה עם ערך שכיח בגלל שכך נוכל לשמור על מבנה ההתפלגות של המאפיין יחד עם שמירה על אדישות יחסית למול ערכים חריגים.[8]

עבור המשתנה Arrival Delay in Minutes שגם הוא בעל מספר גדול יחסית של ערכים ריקים, כדאי להחליף את הערכים הריקים על סמך עמודה בודדת אחרת של המאפיין Departure Delay in Minutes כדי לשמור על תצפיות אלו. בכך ניתן לשמור על נתוני המדד מבלי להשפיע על התנהגותו והתפלגותו הסטטיסטית. נציין כי בחרנו להסיר מאפיין זה מהמודל, אך בחרנו לציין כיצד כדאי להשלים את הרשומות החסרות במאפיין זה במידה ונדרש.

ישנם גם מספר מאפיינים בעלי מספר קטן מאוד של ערכים ריקים. תצפיות אלו כדאי להסיר מסט הנתונים שלנו כיוון שמדובר במספר זניח של תצפיות, ובך לא יאבד מידע רב. [8]

הסרת נתונים נוספת שניתן לבצע היא במאפיין הגיל, בו ישנן שתי תצפיות עם גיל של יותר מ-100 שנים. תצפיות אלו כדאי להסיר מסט הנתונים יחד עם כל הרשומה שלהן כיוון שאובדן הנתונים יהיה זניח.[8]

מאפיין class שהוא משתנה קטגוריאלי שמתאר באיזו מחלקה הלקוח טס. ישנן שלוש מחלקות, ובנוסף סימון עבור מחלקה לא ידועה בשם Unknown. הפרופורציה של השדה Unknown במשתנה זה קרובה ל 15%, ולכן כדאי להשאיר את הסימון הזה כמו שהוא על מנת שלא לייצר הטיה בנתונים עקב שינוי רשומות רבות כל כך. בנוסף ערכי Unknown עשויים לספק הבנה של דפוסים או מגבלות בנתונים עבור המודל, שאולי לא ניתן לזהות בשלב זה.[9]

**הכנת הנתונים [10]**

1. בחירת מאפיינים- השמטת מאפיינים "רועשים" או חסרי חשיבות:

* Gate Location – הסרת ערך חריג 999 עבור טווח הערכים 1-5.
* Class- הסרת תצפית עם ערך חסר משמעות לסוג כרטיס הטיסה.
* הסרת תצפיות עם ערכי null כאשר מספר המופעים מצומצם.

1. טיפול במאפיינים- דיסקרטיזציה של משתנים רציפים:

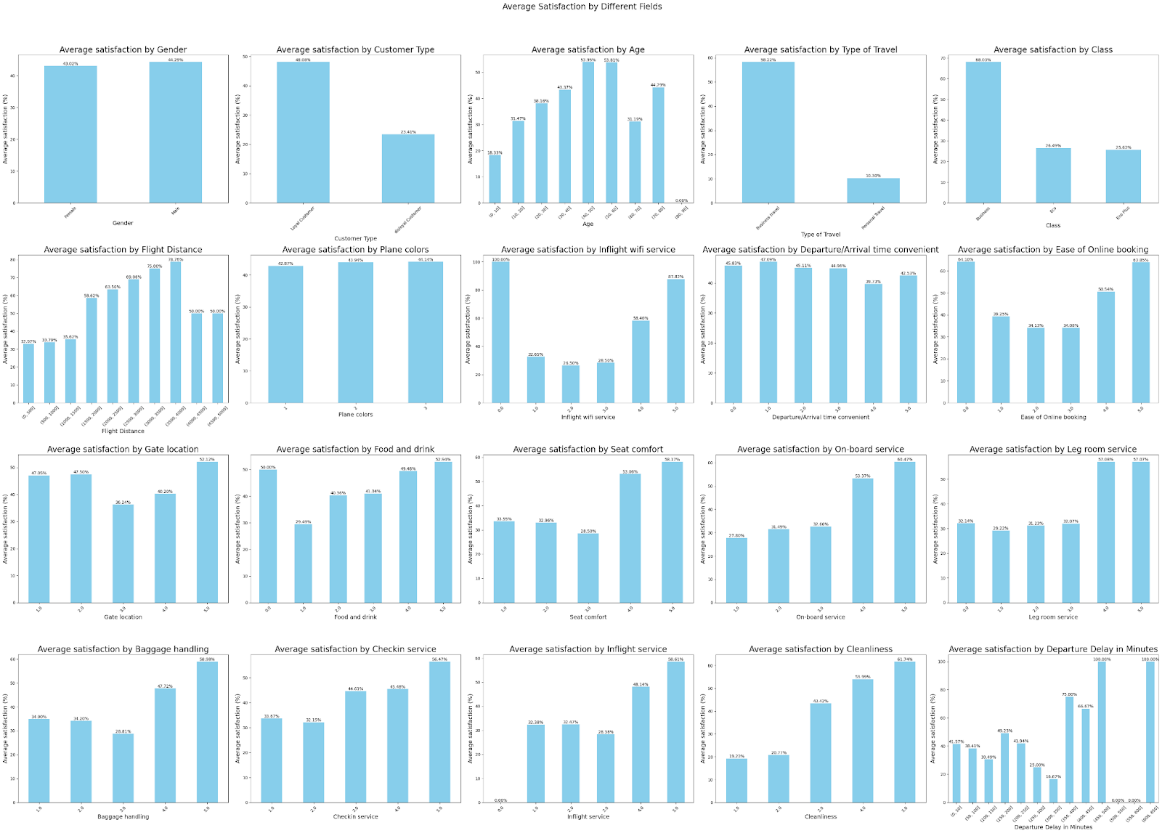
* Age - נחלק את המאפיין לקבוצות (Bins) באינטרוולים של 10.
* Departure Delay in Minutes- נחלק את המאפיין לקבוצות (Bins) באינטרוולים של 50.
* Arrival Delay in Minutes - נחלק את המאפיין לקבוצות (Bins) באינטרוולים של 50.
* Flight Distance - נחלק את המאפיין לקבוצות (Bins) באינטרוולים של 50.

# נספחים

תמונה שמכילה טקסט, צילום מסך, סגול, צבעוני

התיאור נוצר באופן אוטומטיניתוחי סטטיסטיקה תיאורית בהם עשינו שימוש להבנה מעמיקה יותר של המודל באמצעים ויזואליים שונים.

[4]



[3]

תמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

התיאור נוצר באופן אוטומטיתמונה שמכילה טקסט, תרשים, צילום מסך, מקביל

התיאור נוצר באופן אוטומטיתמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

התיאור נוצר באופן אוטומטי

[6]

[5]

[7]

**ביבליוגרפיה**

1. [R., Thandaiah, Prabu. (2023). Airline Passenger Satisfaction Prediction Using Novel Hybrid Random Forest Model Comparison with K-Nearest Neighbor Model. Doi: 10.1109/ICONSTEM56934.2023.10142807](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10142807) [↑](#endnote-ref-1)
2. [Wang, Y. (2023, July). Factors Affect Airline Customers’ Satisfaction: Data Mining. In 2023 4the International Conference on E-Commerce and Internet Technology (ECIT 2023) (pp. 344-362). Atlantis Press.‏](https://www.atlantis-press.com/proceedings/ecit-23/125989053)

   [↑](#endnote-ref-2)
3. [Steven, Matthew, Leon. (2023). Exploring personality moderation effects on satisfaction and loyalty in the airline industry. Journal of Airline and Airport Management, Doi: 10.3926/jairm.357](https://www.jairm.org/index.php/jairm/article/view/357/156) [↑](#endnote-ref-3)
4. [R., Thandaiah, Prabu. (2023). Airline Passenger Satisfaction Prediction Using Novel Hybrid Random Forest Model Comparison with K-Nearest Neighbor Model. Doi:](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10142807)

   [10.1109/ICONSTEM56934.2023.10142807](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10142807)

   [Steven, Matthew, Leon. (2023). Exploring personality moderation effects on satisfaction and loyalty in the airline industry. Journal of Airline and Airport Management, Doi: 10.3926/jairm.357](https://www.jairm.org/index.php/jairm/article/view/357/156) [↑](#endnote-ref-4)
5. [Wang, Y. (2023, July). Factors Affect Airline Customers’ Satisfaction: Data Mining. In 2023 4the International Conference on E-Commerce and Internet Technology (ECIT 2023) (pp. 344-362). Atlantis Press.‏](https://www.atlantis-press.com/proceedings/ecit-23/125989053)

   [↑](#endnote-ref-5)
6. [Air Travelers in America: Annual Survey](https://www.airlines.org/dataset/air-travelers-in-america-annual-survey/)

   [Demography, geography, and airport traffic](https://blog.aci.aero/demography-geography-and-airport-traffic/)

   [Travel Trends by Age Demographic](https://groupstoday.com/business/travel-trends-by-age-demographic) [↑](#endnote-ref-6)
7. [NTRS – NASA Technical Reports Server: Short-Haul Revitalization Study Final Report](https://ntrs.nasa.gov/citations/20180004393) [↑](#endnote-ref-7)
8. [Fricke, H., & Schultz, M. (2009, June). Delay impacts onto turnaround performance. In *ATM Seminar*.‏](https://www.researchgate.net/profile/Michael-Schultz-20/publication/262567633_Delay_Impacts_onto_Turnaround_Performance_-_Optimal_Time_Buffering_for_Minimizing_Delay_Propagation/links/0a85e53824955668c3000000/Delay-Impacts-onto-Turnaround-Performance-Optimal-Time-Buffering-for-Minimizing-Delay-Propagation.pd) [↑](#endnote-ref-8)
9. [Why do Airlines have Special Liveries?](https://www.brookfieldav.com/single-post/why-do-airlines-have-special-liveries)  
    [↑](#endnote-ref-9)
10. [Dinc, E., Vondra, M., Hofmann, S., Schupke, D., Prytz, M., Bovelli, S., ... & Cavdar, C. (2017). In-flight broadband connectivity: Architectures and business models for high capacity air-to-ground communications. *IEEE Communications Magazine*, *55*(9), 142-149.](https://typeset.io/pdf/in-flight-broadband-connectivity-architectures-and-business-1e63oolgrx.pdf)

    ‏ [↑](#endnote-ref-10)
11. [ACSI |The definitive measure of passenger satisfaction with cause-and-effect analysis](https://theacsi.org/industries/travel/airlines/)‏ [↑](#endnote-ref-11)
12. [↑](#endnote-ref-12)
13. [↑](#endnote-ref-13)
14. [Wang, Y. (2023, July). Factors Affect Airline Customers’ Satisfaction: Data Mining. In 2023 4the International Conference on E-Commerce and Internet Technology (ECIT 2023) (pp. 344-362). Atlantis Press.‏](https://www.atlantis-press.com/proceedings/ecit-23/125989053)

    [↑](#endnote-ref-14)