

# פרויקט בלימוד מכונה

## חלק א'



### מגישים:

ניב פינטו, תעודת זהות: deleted

אופיר טופצ'י, תעודת זהות: deleted

עוז גיל - און, תעודת זהות: deleted

### מחברת עבודה:

[https://colab.research.google.com/drive/1SrEE0PVmV4cN\\_SGA6KAj9ef2iRcXeYJO?usp=sharing](https://colab.research.google.com/drive/1SrEE0PVmV4cN_SGA6KAj9ef2iRcXeYJO?usp=sharing)

ישנה גישה חופשית להרצת הבלוקים וכתובת הערות, נדרש להעלות את `Xy_train.csv` לסביבת העבודה.

## תוכן עניינים

3.....	הגדרת הבעיה
3.....	תיאור כללי של עולם התוכן הנחקר
3.....	הגדרת שאלת המחקר
3.....	הבנת הנתונים
3.....	תיעוד מקורות הנתונים ומשמעותם
4.....	הסתברויות אפרוריות וקשרים בין מאפיינים
11.....	איכות הנתונים
12.....	הכנת הנתונים
14.....	נספחים

## הגדרת הבעיה

### תיאור כללי של עולם התוכן הנחקר

הנתונים עבורם נבצע את הפרויקט עוסקים במידע אודות חווית לקוח בטיסה. כאשר המטרה היא לחזות האם הלקוחות מרוצים מהשירות שקיבלו בחברת התעופה ועל סמך זאת להחליט האם הם מרוצים (satisfied) או נטרלים/לא מרוצים (neutral or dissatisfied).

### הגדרת שאלת המחקר

אנו מצפים לבנות מערכת לומדת אשר תדע לזהות האם לקוח היה מרוצה מהשירות שקיבל בטיסה או לא, בהינתן נתונים מסוימים.

## הבנת הנתונים

### תיעוד מקורות הנתונים ומשמעותם

מקור הנתונים הוא מקורס למידת מכונה באוניברסיטת בן גוריון, אנו מניחים שהם נלקחו משאלונים שחולקו לנוסעים בטיסות מסוימות.

משתנה	סוג	טווח ערכים	הסבר
Gender	קטגוריאל	Male, Female	מין הנוסע
Customer type	קטגוריאל	Disloyal, loyal	סוג הנוסע
Age	רציף	7-85,156,157	גיל הנוסע
Type Of Travel	קטגוריאל	Business, Personal	סוג הנסיעה
Class	קטגוריאל – ניתן לסידור	eco, Eco plus, Business it is boring, unknown	מחלקה
Flight Distance	רציף	0-4963	מרחק הטיסה
Plane Color	בדיד	1,2,3	צבע המטוס
Inflight wifi services	בדיד – ניתן לסידור	0:Not Applicable;1-5	חווית שירות האינטרנט בטיסה
Departure/Arrival time convenient	בדיד – ניתן לסידור	0:Not Applicable;1-5	דירוג זמני הטיסה
Ease of Online booking	בדיד – ניתן לסידור	0:Not Applicable;1-5	דירוג נוחות הזמנה מקוונת

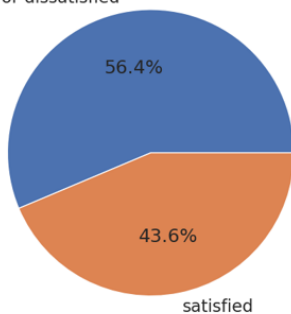
Gate location	בדיד – ניתן לסידור	0:Not Applicable;1-5	דירוג מיקום שער העלייה למטוס
Seat comfort	בדיד – ניתן לסידור	0:Not Applicable;1-5	דירוג נוחות הכיסא
Leg room service	בדיד – ניתן לסידור	0:Not Applicable;1-5	דירוג מקום ברגלים
Baggage handling	בדיד – ניתן לסידור	0:Not Applicable;1-5	דירוג מסירת המזוודה
Check-in service	בדיד – ניתן לסידור	0:Not Applicable;1-5	דירוג שירות הצ'ק אין
Inflight service	בדיד – ניתן לסידור	0:Not Applicable;1-5	דירוג חווית השירות של צוותי המטוס
Cleanliness	בדיד – ניתן לסידור	0:Not Applicable;1-5	דירוג הניקיון בטיסה
Departure Delay in Minutes	רציף	0-610	מס' דקות עיכוב בהמראה
Arrival Delay in Minutes	רציף	0-593	מס' דקות עיכוב בנחיתה
satisfaction	קטגוריאל	Satisfied, neutral or dissatisfied	שביעות הרצון מהטיסה

### הסתברויות אפרוריות וקשרים בין מאפיינים

#### משתנה מוסבר - Satisfaction:

Pie Chart of satisfaction

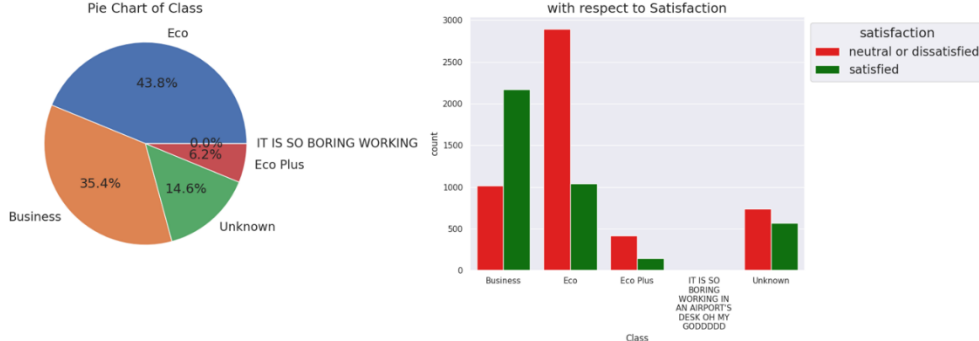
neutral or dissatisfied



ניתן לראות כי משתנה המטרה מאוזן וכ-56% מהנוסעים אינם מרוצים\ניטרליים ביחס לחוויה שלהם בחברת התעופה. בנוסף, סט הנתונים מאוזן ולדעתנו הוא תואם את המציאות.

## משתנים קטגוריאליים:

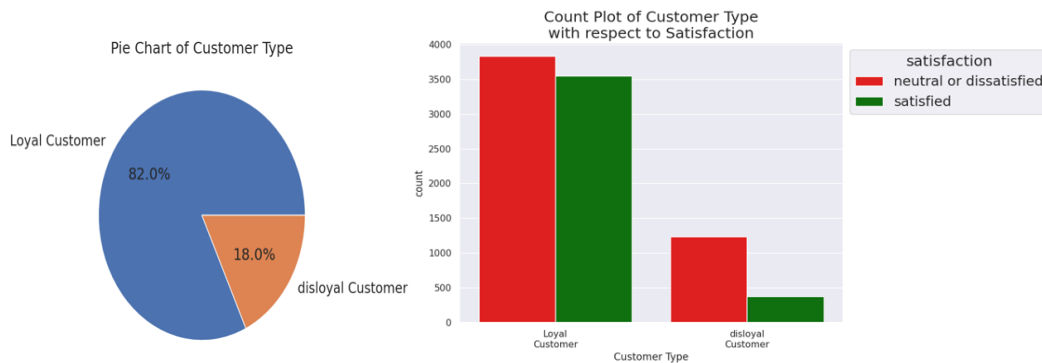
### Class



מתרשים העוגה עולה כי מכל הנוסעים שידוע על מחלקת הטיסה שלהם, עולה כי מרביתם טסים במחלקת Eco. אחוז דומה במחלקת עסקים ואחוז קטן שבחר לשדרג למחלקת EcoPlus. כאשר נפלח את הנוסעים לפי משתנה המטרה, ניתן לראות כי

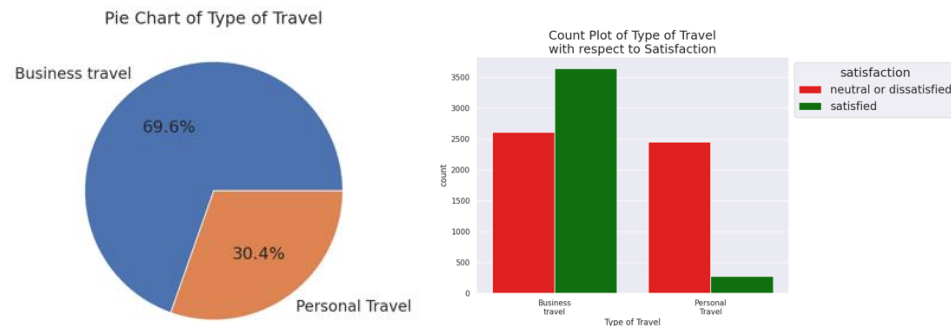
מרבית הנוסעים במחלקת עסקים מרוצים יותר. ניתן לומר כי שדרוג מינורי למחלקת EcoPlus לא בהכרח ישפר את חוויית הטיסה. לעומת זאת, שדרוג למחלקת עסקים בסיכוי גדול יותר עשוי לשפר את חוויית הטיסה של הנוסע.

### Customer Type



ניתן לראות באופן די בולט שרוב הנוסעים הם נוסעים נאמנים.

בחלוקה לפי משתנה המטרה קיים שוני בין הקבוצות. בקבוצת הלוקחות הנאמנים החלוקה מאוזנת כלומר הסתברות כמעט שווה בין שבעי הרצון לניטרליים\אינם שבעי רצון. לעומת זאת, רוב הלוקחות הלא נאמנים אינם שבעי רצון.



### Type Of Travel

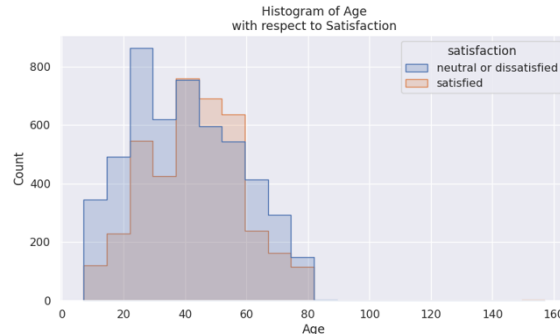
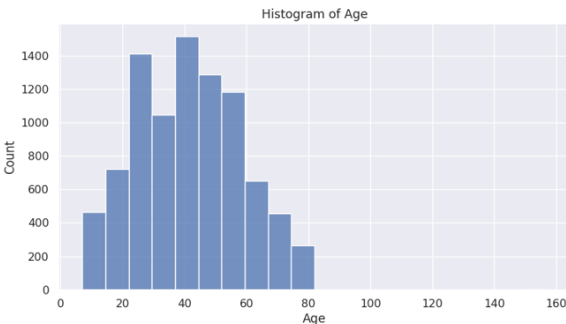
ניתן לראות שרוב הנוסעים הם נוסעים למטרות עסקיות. לפי משתנה המטרה ניתן לומר כי רוב הנוסעים בנסיעות העסקיות מרוצים מחוויית הטיסה שלהם לעומת מגזר הנסיעות הפרטיות שרובם המכריע אינו מרוצה\ניטרלי.

## משתנים רציפים:

### AGE

ניתן לראות שהתפלגות הגילאים היא יחסית נורמלית, ישנם מספר חריגים שמותחים את הגרף לערך 160. בהתייחס למשתנה

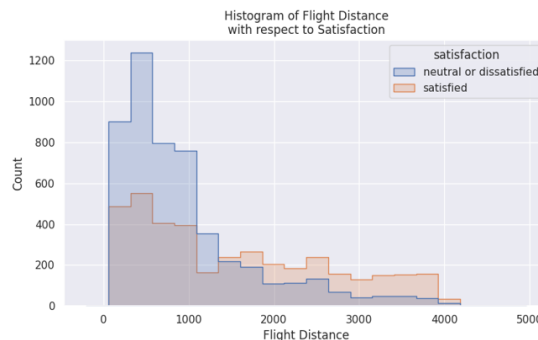
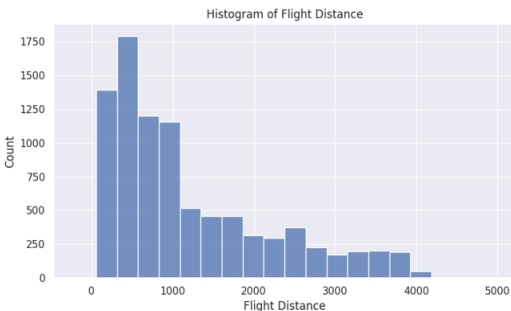
המוסבר, קשה לראות אינדיקציה חדה להבדלים משמעותיים ברמת שביעות הרצון בהבדלים בין הגילאים.



### Flight Distance

ניתן לראות מגמה ברורה - רוב הטיסות הן טיסות במרחק 1000 ק"מ או פחות. בקנה אחד עולה שמספר חוות הדעת הן על טיסות קצרות. ניתן לראות כי רוב האוכלוסייה אינה

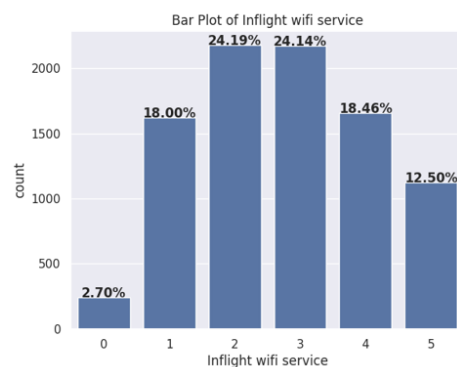
מרוצה/ניטרלית מאשר מרוצה בטיסות קצרות. לעומת זאת, כאשר מרחק הטיסה גדולה יותר, מספר המרוצים הוא גדול יותר מאשר הלא מרוצים/ניטרליים.



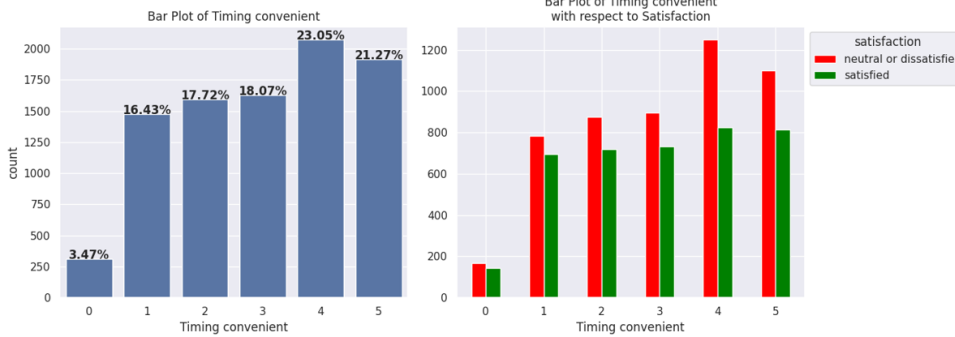
### Inflight Wifi service

ניתן לראות כי מחצית המדרגים דירגו את את רמת ה-wi-fi בציונים 2,3. רובם המכריע בסופו של דבר אינו שבע רצון או ניטרלי. לעומת זאת, רובם המכריע של מדרגי 5 ורוב

מדרגי 4 מרוצים בסופו של דבר מחוויית הטיסה. ניתן להסיק מכך שרמת הויי פי חשובה לרמת שביעות הרצון של הנוסעים ויש לשפרו.

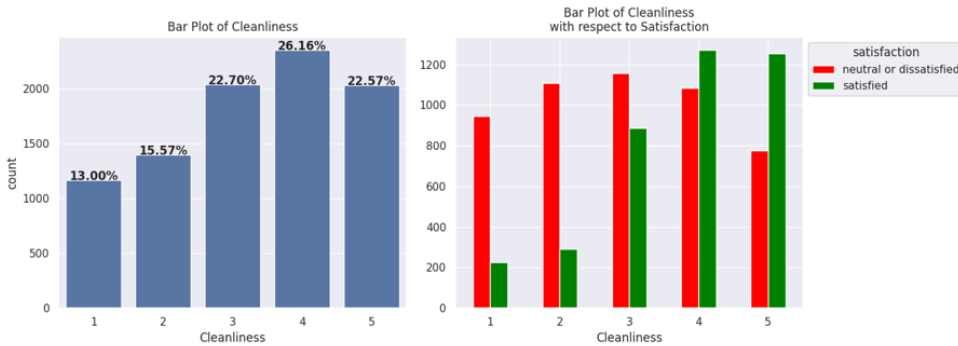


## Timing Convenient



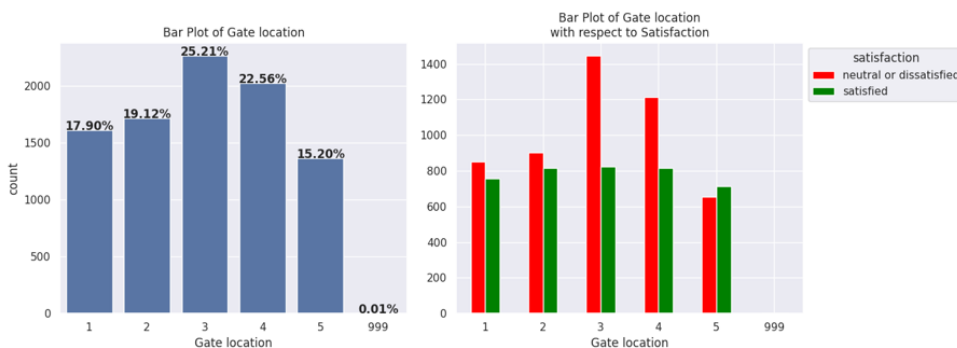
ניתן לראות כי כמות המצביעים לזמני המראה/נחיתה נוחים בציונים 4 ו-5 לעומת 1,2,3 היא גבוהה יותר. כאשר נתמקד בתרשים המחולק על פי ערכי המשתנה המוסבר - לכל הציונים יש כמות גבוהה יותר של לא מרוצים/ניטרליים מאשר מרוצים.

## Cleanliness



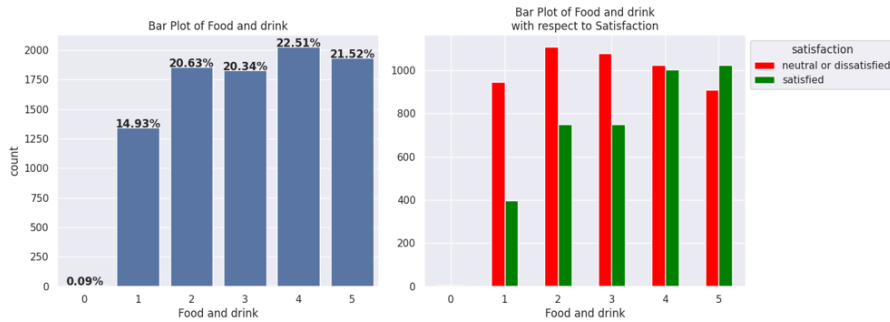
ניתן לראות כי ככל שרמת הניקיון של החברה גבוהה יותר – כך גם עולה כמות שבעי הרצון בקנה אחד. ניתן להבין שהציונים 1,2 ו-5 חוזים בצורה טובה את משתנה המטרה.

## Gate Location



ניתן לראות שציונים 3 ו-4 הם הציונים שדורגו הכי הרבה פעמים. ובהתאם גם כמות חוסר שבעי הרצון גבוהה יותר מאשר בשאר הציונים. ניתן לומר כי מיקום השער אינו משפיע על אחוז שבעי הרצון (גובה העמודות ירוקות) מאחר וכמות המצביעים שהם שבעי רצון בכל הציונים יחסית זהה.

## Food and Drink

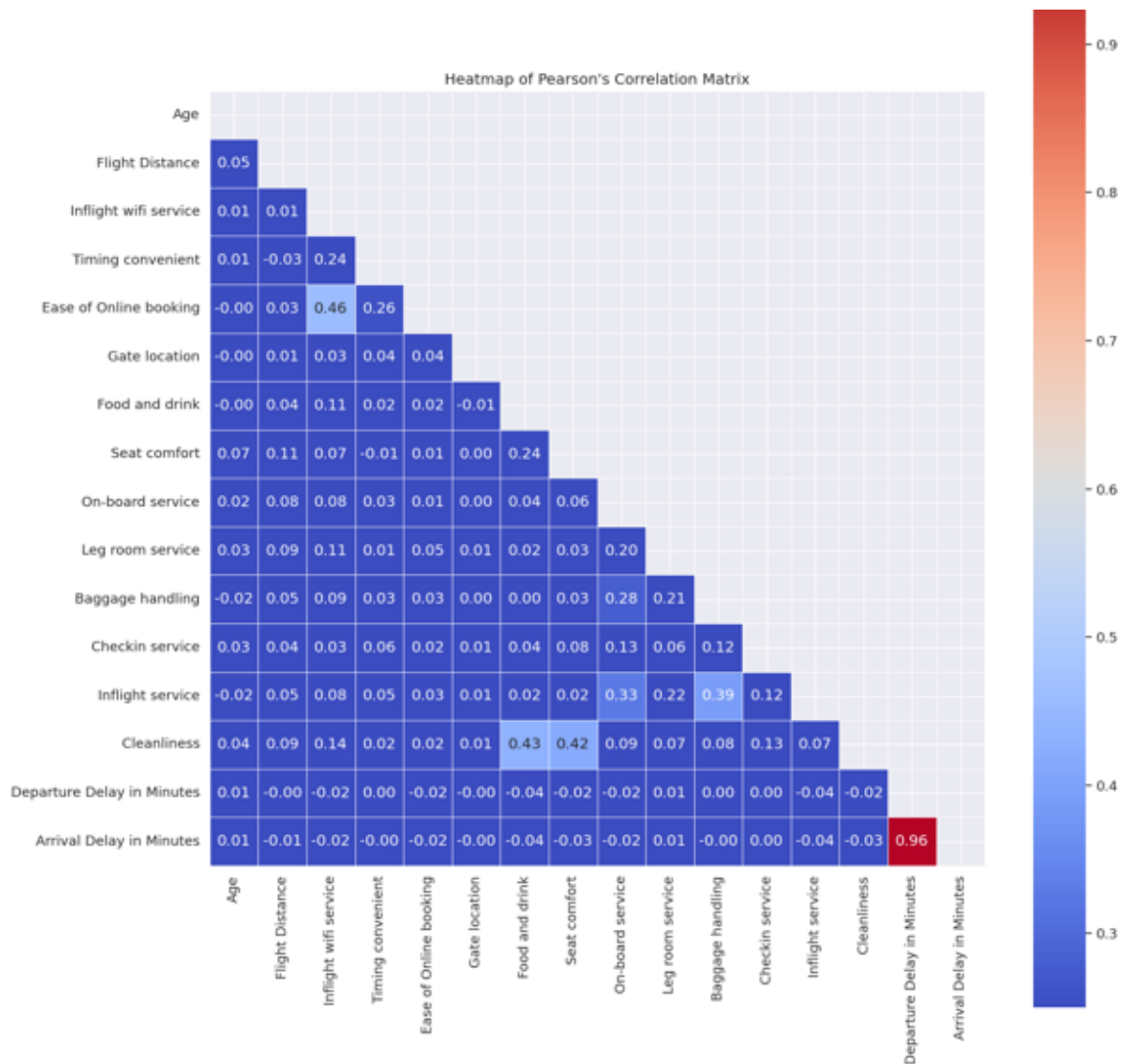


ניתן לראות כי התפלגות הציונים היא יחסית מאוזנת. כאשר התרשים מחולק על פי ערכי המשתנה המוסבר - ניתן לראות כי ישנו מתאם בין ציון האוכל והשתייה לבין כמות שבעי הרצון מהטיסה. לעומת זאת, התפלגות

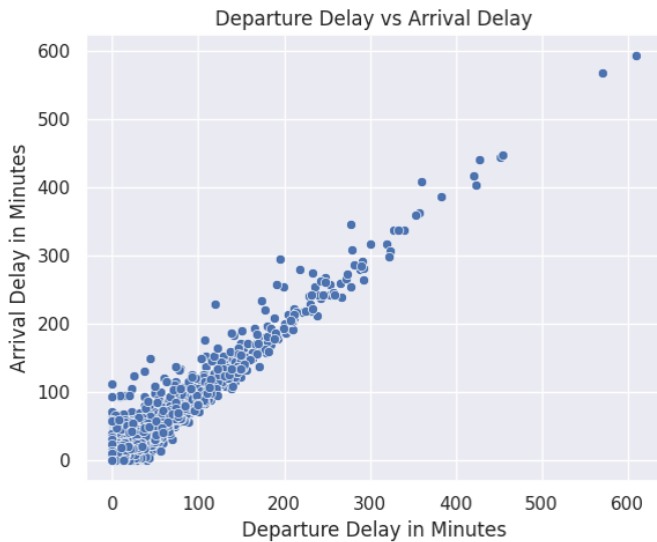
הניטרליים\לא שבעי רצון יחסית דומה בכל הציונים מה שעשוי להסביר שאין השפעה ברורה על המשתנה המסביר.

## קשרים בין מאפיינים:

נציג טבלת קורלציה בין משתנים רציפים בלבד מאחר ולמשתנים קטגוריאליים לא נכון לעשות קורלציית פירסון.

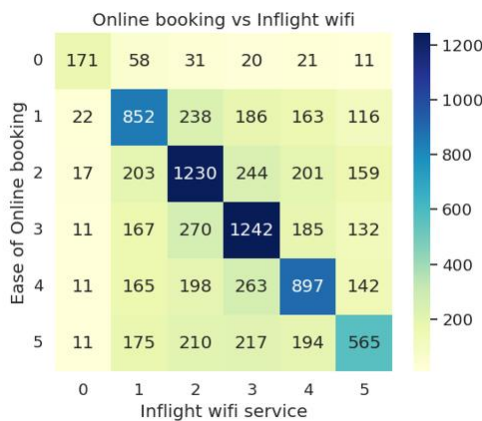






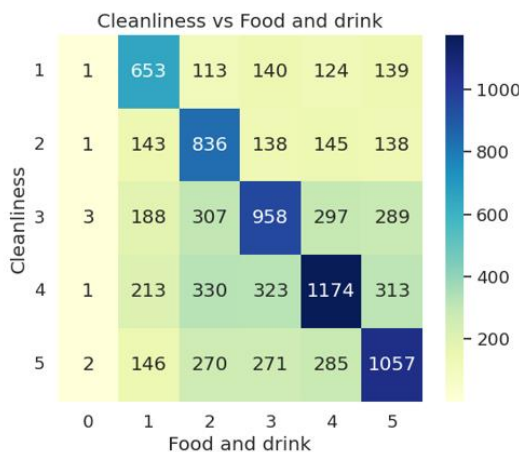
### Departure Delay vs. Arrival Delay

קשר ראשון שבחרנו להתייחס אליו הוא בחינת המתאם בין איחור המראת המטוס בדקות לאיחור נחיתת המטוס בדקות. לפי קורלצית פירסון מקדם המתאם הוא  $-0.96$  – מקדם גבוה מאוד ובחרנו להתעמק בו. באופן טבעי, אם המטוס מאחר בהמראתו במספר דקות סביר להניח שהוא יאחר בנחיתתו באותו מספר הדקות ומכאן רמת המתאם הגבוהה שנצפה לה. ניתן לראות קשר לינארי הדוק בגרף.



### Online Booking vs. Inflight Wifi

בחרנו להתייחס למתאם בין רכישת כרטיס נסיעה באופן מקוון לעומת ציון Wifin בטיסה. לפי קורלציית פירסון מקדם המתאם הוא  $-0.46$  – כלומר יש קורלציה מסוימת. לא ציפינו שיהיה קשר בין כמה נוח לבצע רכישה מקוונת לציון Wifin במטוס. לטעמנו, אין קשר בין השניים מאחר ומדובר בטכנולוגיות שונות שלא קשורות זו בזו. אתר מקוון הוא כלי שבאופן יחסי קל לתחזוק ותפעול לעומת רשת אלחוטית במטוס שהוא אתגר טכנולוגי ועלות התפעול והתחזוקה שלו גבוהה בהרבה.



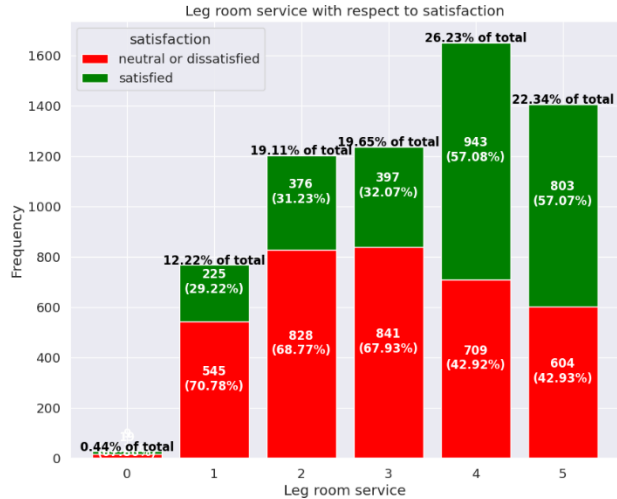
### Cleanliness vs. Food and Drink

בחרנו להתייחס למתאם בין ציון הנקיון של חברת התעופה לבין דירוג רמת האוכל והשתייה. לפי קורלציית פירסון מקדם המתאם הוא  $-0.43$  – כלומר יש קורלציה. ציפינו לקשר חזק יותר בין שני המשתנים הללו, כיוון שנקיון ואיכות האוכל והשתייה נתפסים לעתים קרובות כמאפיינים קשורים של חווית השירות הכוללת בחברת התעופה. ההסבר לכך הוא כנראה בסיסי יותר, הנקיון מתייחס בעיקר לניקיון המטוס וחלליו הציבוריים, בעוד איכות האוכל והשתייה קשורה יותר לתפעול המטבח ואספקת המזון והמשקאות.

## השפעת מאפיינים על משתנה המטרה:

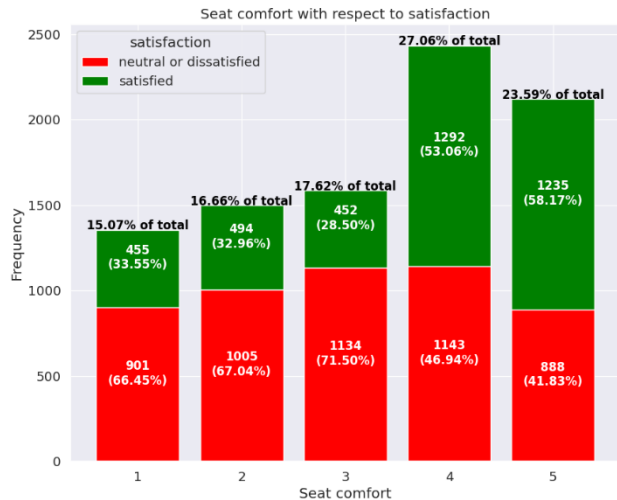
### Leg Room Service

על סמך ידע אישי, למיקום ברגליים יש השפעה גדולה על חווית הטיסה במיוחד אנשים שמתנשאים לגובה רב, כבדי משקל, ונכים. ניתן לראות מהתרשים שבציונים 1-3 יחס המרוצים נשאר פחות או יותר זהה (כ-30%). לעומת זאת, בציונים 4 ו-5 יש עלייה משמעותית ביחס המרוצים (כ-57%). כלומר, מהתרשים כי במידה והנוסעים אכן דירגו את רמת המקום ברגליים בציונים גבוהים (4,5) – בסבירות גבוהה יותר היו מרוצים. בהתאמה, אם דירגו נמוך (1-3) את רמת המקום ברגליים בסבירות גבוהה יותר היו לא מרוצים.



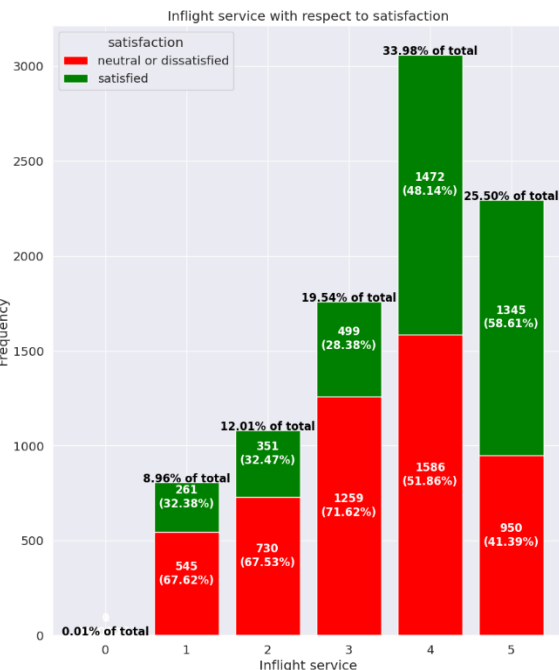
### Seat Comfort

לפי הניסיון שלנו, נוחות הכיסאות היא מרכיב משמעותי בחוויית הנוסע. הנוסע נמצא הרבה זמן בכיסאו והוא מצפה לנוחות מיטבית, בייחוד אוכלוסייה בעלת כאבי גב, ברכיים, וכו'. ניתן לראות מהתרשים שבציונים 1-3 יחס המרוצים נשאר פחות או יותר זהה (כ-31%). לעומת זאת, בציונים 4 ו-5 יש עלייה משמעותית ביחס המרוצים (כ-55%). כלומר, מהתרשים כי במידה והנוסעים אכן דירגו את רמת נוחות הכיסא בציונים גבוהים (4,5) – בסבירות גבוהה יותר היו מרוצים. בהתאמה, אם דירגו נמוך (1-3) את נוחות הכיסא בסבירות גבוהה יותר היו לא מרוצים.



### In Flight Service

להערכתנו, נוסע בטיסה חשוב לו להרגיש שהוא מקבל תמורה למחיר שהוא משלם ושנותני השירות במהלך הטיסה יסבירו פנים ויתנו לו את השירות הטוב ביותר שהם יכולים להציע. ניתן לראות מהתרשים שבציונים 1-2 יחס המרוצים נשאר פחות או יותר זהה (כ-32%). ואילו בציון 3 ישנה ירידה קטנה ביחס המרוצים (28%). לעומת זאת, בציונים 4 ו-5 יש עלייה משמעותית הן ביחס המרוצים והן בגידול ביחס המרוצים. כלומר, ישנה התאמה בין דירוג שירות הטיסה לבין ציון השירות: במידה והנוסעים דירגו את רמת השירות ברמה הגבוהה ביותר (5) בסבירות גבוהה שהיו מרוצים, אם דירגו 4 הסתברות שווה ואילו בציונים 1-3 בסבירות גבוהה ניתן להעריך שהנוסע היה לא מרוצה/ניטרלי. בניגוד לצפיותינו – בציון 3 יש קיטון ביחס המרוצים.

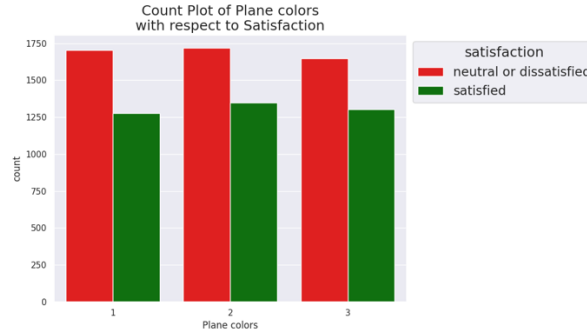
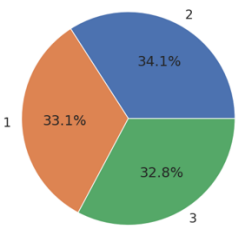


## משתנים שייתכן כי ניתן להסיר:

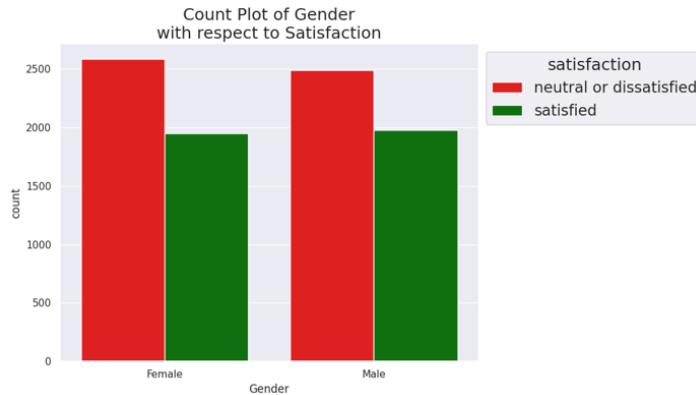
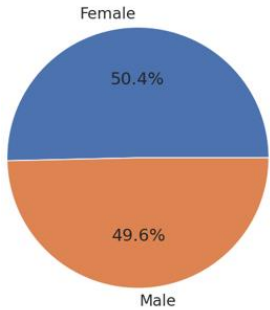
### Plane Color

חשדנו שמשתנה צבע המטוס ניתן להסרה היות ומניסיונו כנוסעים אין חשיבות כלל לצבע המטוס. יתר על כן, כאשר טסים בלילה או נכנסים דרך השרוול לא תמיד ניתן להבחין בצבע המטוס בבירור. הגרף שלפנינו מראה שאכן שביעות הרצון אינה מושפעת מאחר ובשלושת צבעי המטוסים יחס העמודות די זהה.

Pie Chart of Plane colors



Pie Chart of Gender



### Gender

להערכתנו, אין קשר בין חווית הטיסה של כל אדם באשר למגדר שלו. הגרף שלפנינו מראה שאכן שביעות הרצון אינה מושפעת מאחר ובשני המגדרים יחס העמודות די זהה.

## איכות הנתונים

### נתונים חסרים

תחילה, חילצנו מבסיס הנתונים שניתן לנו את כמות ערכי ה-NULL לכל משתנה. ניתן לזהות 2 עמודות "בעייתיות" בעלי כמות משמעותית של ערכים חסרים, ו-20 עמודות בעלי כמות זניחה של ערכים חסרים, אם קיימים (בין 0 ל-2).

סברנו כי תחילה יש להסיר את הרשומה הבודדת בה קיים ערך חסר למשתנה המטרה, להשלים בכל עמודה "לא בעייתית" את הערכים החסרים באמצעות המופע השכיח ביותר באותה עמודה, ולתת תשומת לב מיוחדת ל-2 העמודות הבעייתיות.

1. בהתייחס ל-**Arrival Delay in Minutes** – עקב מתאם גבוה מאוד וקשר ליניארי חזק עם עמודת **Departure Delay in Minutes**, נציע להשלים את הערכים החסרים ע"י גרסיה בה **Departure Delay in Minutes** חוזה את **Arrival Delay in Minutes**.
2. באשר ל-**Leg Room Service** – לדעתנו אין פתרון חד משמעי לבעיה זאת.

מאחר וכמות הערכים החסרים הוא כמעט שליש מסך הרשומות, רצינו תחילה להוסיף

קטגוריית ציון חדשה בשם "לא ידוע", על מנת לא לוותר על דפוס מסוים שעשוי להתקיים באותה קבוצת ערכים חסרים

Gender	0
Customer Type	1
Age	1
Type of Travel	1
Class	1
Flight Distance	2
Plane colors	0
Inflight wifi service	2
Timing convenient	2
Ease of Online booking	2
Gate location	2
Food and drink	2
Seat comfort	1
On-board service	1
Leg room service	2701
Baggage handling	1
Checkin service	1
Inflight service	1
Cleanliness	1
Departure Delay in Minutes	1
Arrival Delay in Minutes	471
satisfaction	1

גדולה. אך עם זאת, אפשרות זו כרוכה בהמרת המשתנה למשתנה קטגוריאלי, ומשמעות ההמרה הינה ויתור על המידע החשוב אודות האורדינליות הקיימת בין הערכים. בסופו של דבר, הצענו להשלים את 30% הערכים החסרים ע"י דגימה מהתפלגות המאפיינת את 70% הערכים הקיימים.

### נתונים חריגים

- Age – חיפשו ערכים שאינם בין 0 ל-122 (האישה המבוגרת בהיסטוריה).  
נמצאה רשומה אחת עם גיל 156 ורשומה אחת עם גיל 157.
  - Class – חיפשו ערכים שאינם בקבוצה הנתונה (Eco, Eco Plus, Business).  
נמצאה רשומה אחת עם "IT IS SO BORING..." ובנוסף כ-1300 רשומות עם "Unknown".
  - Flight Distance – חיפשו ערכים שאינם בין 0 ל-15,349 (הטיסה הארוכה בהיסטוריה).  
נמצא רשומה אחת עם ערך שלילי.
  - Gate Location – חיפשו ערכים שאינם בקבוצה הנתונה (0, 1, 2, 3, 4, 5).  
נמצאה רשומה אחת עם הערך 999.
- בנוגע לערך Unknown של Class – מאחר ומדובר בכמות ערכים חריגים גדולה (מופיע ב-1300 רשומות מתוך כ-9000), נציע להשאיר את הערכים כפי שהם ולא להחליפם ע"י שימוש באסטרטגיות כאלה ואחרות, לדעתנו פתרון זה סביר גם בגלל שמדובר במשתנה נומינלי.
- 5 הערכים החריגים האחרים מופיעים פעם אחת בלבד כל אחד, נציע להשלים אותם באמצעות הערך השכיח או החציון.

### הכנת הנתונים

1. השמטת תצפיות בעלות חוסר רב – ספרנו את מספר הערכים החסרים בכל תצפית, וגילינו כי חסרים לכל היותר 2 ערכים בכל אחת מהן, למעט ב-2 רשומות בעייתיות בהן חסרות 10 או 16 ערכים. ראה [נספח 1](#).  
קודם לכן שיערנו שהסרה של רשומות בעלות 5 ערכים ומעלה (+25%) הינה פרקטיקה סבירה, אך הסרה של 2 תצפיות בעייתיות אלו בפועל עומדת בסף של 45% ומעלה, והינה מתבקשת והגיונית. לאחר ההסרה גילינו כי נותרו 2 עמודות בלבד בעלות ערכים חסרים (Leg room service, Arrival Delay in Minutes). ראה [נספח 2](#).
2. השמטת מאפיינים – מוקדם יותר בחנו את התפלגות **Plane Color** באמצעות תרשים עוגה וראינו שהוא מחולק כמעט שווה - כשליש לכל אחת מהקטגוריות. כמו כן, כאשר בחנו גרף המציג את הקשר בין "Plane Color" לבין משתנה המטרה של שביעות רצון הנוסע, לא הבחנו בהבדלי שביעות רצון בין הקטגוריות השונות של צבעי המטוס. לכן, החלטנו להשמיט את מאפיין זה שנחשד תחילה ולבסוף גם נראה כחסר כל חשיבות.
3. השלמת ערכים חסרים:
  - עבור המשתנה **Leg Room Service**, מאחר וישנם 2700 ערכים חסרים (30% מהעמודה) – לא רצינו למחוק את הרשומות כדי לא לאבד מידע רב. בחרנו להשלים את הערכים לפי מדגם אקראי (עם החזרות) של 2700 ערכים מתוך 6298 הקיימים. בדרך זו השלמנו ערכים ושמרנו בקירוב על התפלגות המשתנה. ראה [נספח 3](#).

- עבור המשתנה **Arrival Delay in Minutes**, למשתנה זה יש 470 ערכים חסרים (כ-5%). בהתבסס על תרשים הפיזור (Scatter Plot) בין המשתנים "Arrival Delay in Minutes" ו"Departure Delay in Minutes", ובהתבסס על מתאם פירסון-קיים מתאם גבוה מאוד (0.96) בין שני המשתנים. מתאם המעיד על קשר לינארי חזק בין השניים. בחרנו ברגרסיה לינארית כדי למלא את הערכים החסרים.

## **דיסקרטיזציה**

נרצה לפלג את כלל הערכים לפי הידע המוקדם שלנו, לפי טווחים מקובלים במחקרים דומים, לפי ההיסטוגרמה המתאימה לכל משתנה, ובהתאם לכמות התצפיות בכל טווח.

### **Age**

בחרנו לבצע דיסקרטיזציה למשתנה זה כך: [0-18, 19-40, 41-60, 61+]. נפלג את המשתנה ל-4 קטגוריות גיל. הראשונה היא ילדים ונערים, השנייה בוגרים עד גיל 40, השלישית בוגרים עד גיל 60, והקבוצה הרביעית גיל הפרישה.

### **Flight Distance**

בחרנו לבצע דיסקרטיזציה למשתנה זה ל-4 מרחקי טיסה שונים. טיסות קצרות (עד 500 ק"מ), בינוניות (501-1500 ק"מ), ארוכות (1501-2500 ק"מ), ארוכות מאוד (2501 + ק"מ).

### **Departure Delay in Minutes & Arrival Delay in Minutes**

נבצע דיסקרטיזציה למשתנים אלה לפי 5 קטגוריות שונות (ללא עיכוב, 1-15 דקות, 16-30 דקות, 31-60 דקות, 60 + דקות).

## **גזירת מאפיינים נוספים**

**Overall Service Rating** – ניצור משתנה חדש הסוכם את רמות שביעות הרצון של משתנים שונים הקשורים לרמה האנושית. משתנים כגון: on-board service, inflight service, baggage handling, check-in service. נסכום את כל המשתנים האלו לדירוג שירות כולל. משתנה זה עשוי לעזור לזהות דפוסים בין איכות הרמה האנושית של צוותי החברה לשביעות רצון הנוסעים.

**Convenience Rating** - ניצור משתנה חדש המשלב את רמות שביעות הרצון של שירותים פיזיים שונים שמציע המטוס. שירותים כגון: inflight WiFi service, food and drink, seat comfort, leg room service. בעזרת משתנה זה נוכל לנתח את ההשפעה של איכות השירותים שניתן לקבל במטוס על רמת שביעות הרצון של הנוסעים.

## נספחים

## נספח 1 – מספר רשומות עבור לפי מספר ערכים חסרים

```

Rows with 0 Nulls: 5970
Rows with 1 Nulls: 2886
Rows with 2 Nulls: 142
Rows with 3 Nulls: 0
Rows with 4 Nulls: 0
Rows with 5 Nulls: 0
Rows with 6 Nulls: 0
Rows with 7 Nulls: 0
Rows with 8 Nulls: 0
Rows with 9 Nulls: 0
Rows with 10 Nulls: 1
Rows with 11 Nulls: 0
Rows with 12 Nulls: 0
Rows with 13 Nulls: 0
Rows with 14 Nulls: 0
Rows with 15 Nulls: 0
Rows with 16 Nulls: 1
Rows with 17 Nulls: 0
Rows with 18 Nulls: 0
Rows with 19 Nulls: 0
Rows with 20 Nulls: 0
Rows with 21 Nulls: 0
Rows with 22 Nulls: 0

```

## נספח 2 – ערכים חסרים בכל עמודה לפני השלמה (משמאל) ואחרי השלמה (מימין)

	Null Count
Gender	0
Customer Type	1
Age	1
Type of Travel	1
Class	1
Flight Distance	2
Plane colors	0
Inflight wifi service	2
Timing convenient	2
Ease of Online booking	2
Gate location	2
Food and drink	2
Seat comfort	1
On-board service	1
Leg room service	2701
Baggage handling	1
Checkin service	1
Inflight service	1
Cleanliness	1
Departure Delay in Minutes	1
Arrival Delay in Minutes	471
satisfaction	1

DataFrame's shape, BEFORE rows drop:  
(9000, 22)

Rows  
Drop



	Null Count
Gender	0
Customer Type	0
Age	0
Type of Travel	0
Class	0
Flight Distance	0
Plane colors	0
Inflight wifi service	0
Timing convenient	0
Ease of Online booking	0
Gate location	0
Food and drink	0
Seat comfort	0
On-board service	0
Leg room service	2700
Baggage handling	0
Checkin service	0
Inflight service	0
Cleanliness	0
Departure Delay in Minutes	0
Arrival Delay in Minutes	470
satisfaction	0

Dataset's shape after drop: (8998, 22)

### נספח 3 – התפלגות המשתנה לפני השלמה (משמאל) ואחרי השלמה (מימין)

