**הסבר על Exploration + Data Pre-Processing**

ספריות בשימוש:

* pandas – משמשת לעבודה עם נתונים בטבלאות, כולל טעינה, עיבוד, וסינון מידע.
* numpy – משמשת לחישובים נומריים יעילים, כולל עבודה עם מערכים רב-ממדיים.
* matplotlib.pyplot – משמשת ליצירת גרפים ותרשימים להצגת נתונים.
* seaborn – מספקת כלים מתקדמים ליצירת גרפים סטטיסטיים וויזואליזציות משופרות.
* warnings.simplefilter(action='ignore', category=FutureWarning) – מבטל אזהרות FutureWarning כדי למנוע הודעות על פונקציות שעלולות להשתנות בעתיד.

1. טעינת הנתונים

קריאה לקובץ (CSV, Excel, SQL וכו') באמצעות pandas (pd.read\_csv() וכו').

הצגת הנתונים (df.head(), df.info(), df.describe()) לקבלת מושג ראשוני.

2. בדיקת חוסרים (Missing Values) - (df.isnull().sum())

בדאטה שלנו רוב הערכים החסרים היו בשדות הקשורים לקונקשן נובעים מכך שהטיסות ברובן ישירות. שדות כמו Onward Flight Connection Airport ו-Return Flight Connection Airport רלוונטיים רק לטיסות עם עצירות ביניים, ולכן נותרים ריקים בטיסות ישירות.

3. הכנת הדאטה

* המרת תאריכים לפורמט של datetime64
* בזמני נחיתה זיהינו נתונים בעייתיים שבהם בנוסף לשעה מופיע +1 או +2, מה שמעיד על הוספת ימים לזמן ההגעה בפועל. הפתרון הוא לחלץ את הימים הנוספים, שמרנו רק את השעה, שילבנו אותה עם תאריך היציאה או החזרה, והמרנו את הכול לפורמט datetime תקני תוך הוספת הימים הנוספים לזמן ההגעה.
* חילצנו תכונות זמן מהעמודות Departure Time ו-Return Departure Time, כולל חודש, יום, שעה, ויום בשבוע (מותאם כך שיום ראשון יהיה 0 ושבת 6), כדי לאפשר ניתוח והשוואה מבוססי זמן.
* סיווגנו את זמני היציאה והחזרה לפי קטגוריות של חלקי היום והמרנו אותם לערכים מספריים (0 ללילה, 1 לבוקר, 2 לצהריים, 3 לערב).
* מיפינו את ימי השבוע כך שימי שני-חמישי יסומנו כלא סוף שבוע (False), וימי שישי-ראשון יסומנו כסוף שבוע (True), כדי לזהות אם טיסה יוצאת בסופ"ש.
* המרנו את שמות האתרים שמהם דגמנו לערכים מספריים. בהמשך נוכל להשתמש בזה כדי לנתח השפעה של כל אתר (לבדוק האם יש הבדל במחירים, זמינות טיסות או פרמטרים אחרים בהתאם לאתר שממנו הגיע המידע).
* הוספנו נתוני מרחק משדה התעופה למרכז העיר על בסיס קוד שדה התעופה, תחילה תכננו להשתמש בממוצע מרחקי השדות בכל עיר כשקוד השדה תעופה חסר אך בהמשך זיהינו שהרשומות בהן השדה חסר הן חברות אוטובוסים ורכבות במקום חברות תעופה – רשומות אלו הוסרו, מה שהשפיע רק על 0.101% מהנתונים.
* חישבנו את סך כל הקונקשנים בטיסה על ידי חיבור מספר עצירות הביניים בטיסת ההלוך (Onward Flight Connections) ומספר עצירות הביניים בטיסת החזור (Return Flight Connections), כדי לקבל משתנה חדש Total Connections שמשמש לניתוח והשוואת מסלולי טיסה.
* חישבנו את משך הטיסה הכולל בדקות על ידי חיבור משך כל טיסות ההלוך (Total Onward Flight Duration(min)) ומשך כל טיסות החזור (Total Return Flight Duration(min)), כדי לקבל את משך הטיסה הכולל (Total Flight Duration(min)) לניתוח והשוואה.

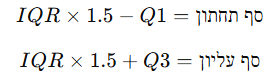
4. זיהינו והסרנו חריגות (Outliers) במחירי הטיסות באמצעות שיטת IQR של טוקי (Tukey's IQR Method)

שיטת Interquartile Range (IQR) של טוקי (Tukey's IQR Method)

אחת הדרכים הפופולריות לאיתור והסרת חריגים (Outliers) בנתונים. היא מבוססת על הרעיון שמרבית הנתונים נמצאים בטווח הרבעונים, ואילו ערכים שנמצאים רחוק מדי מהרבעונים העליון והתחתון נחשבים חריגים.

IQR (Interquartile Range) הוא המרחק בין הרבעון השלישי (Q3) לרבעון הראשון (Q1).

* Q1 (רבעון ראשון, 25%) – הערך שמתחתיו נמצאים 25% מהנתונים.
* Q3 (רבעון שלישי, 75%) – הערך שמתחתיו נמצאים 75% מהנתונים.
* IQR מודד את פיזור הנתונים בלב ההתפלגות ומונע השפעה של ערכים קיצוניים.

טוקי הגדיר טווח קביל של ערכים כך שחריגים יהיו הרחק מהרבעונים. ערך נחשב חריג אם הוא נמצא מחוץ לטווח הבא:

כל ערך שנמצא מתחת לסף התחתון או מעל לסף העליון – מסומן כחריג.

בעזרת שיטה זו הסרנו 2.27% מהנתונים שהם מחירי טיסות קיצוניים, כדי להבטיח שהניתוח הסטטיסטי יהיה מדויק יותר ולא יושפע מערכים חריגים. כעת הנתונים מאוזנים ואפשר לקבל תובנות ריאליות יותר לגבי מחירי הטיסות.

תמונה שמכילה עלילה, תרשים, צילום מסך, קו

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.

🔴 לפני הסרת החריגים → ההתפלגות מוטה מאוד ימינה עם זנב ארוך לטיסות יקרות במיוחד.

🔵 אחרי הסרת החריגים → ההתפלגות מאוזנת יותר, הרוב המוחלט של הטיסות מרוכז בטווח 300-800 ש"ח, בלי ערכים קיצוניים שמעוותים את הנתונים.

5. חקירה חזותית של הנתונים (Exploratory Data Analysis - EDA)

בשלב זה מבצעים חקירה חזותית של הנתונים כדי להבין תבניות, מגמות וחריגות בעזרת גרפים ותרשימים.

הסברים (לפי סדר הגרפים):

* הגרף מציג את מספר הטיסות לפי חברת תעופה. easyJet, Vueling, ו-Ryanair מובילות את השוק עם הכי הרבה טיסות, בעיקר בטיסות לואו-קוסט.

מסקנה: חברות לואו-קוסט שולטות בשוק, בעוד שחברות מסורתיות מפעילות פחות טיסות, כנראה באסטרטגיית שירות שונה.

* הגרף מציג את התפלגות מחירי הטיסות לפי חברת תעופה.  
  חברות לואו-קוסט (easyJet, Ryanair, Wizz Air) מציגות מחירים נמוכים ויציבים, בעוד שחברות מסורתיות (British Airways, Iberia) בעלות מחירים חציוניים גבוהים יותר וטווחי מחירים רחבים.  
  מסקנה: חברות מסורתיות מציעות מגוון מחירים רחב יותר, ככל הנראה בשל מחלקות שונות ושירותי פרימיום.

**הסבר על BOXPLOT:**

* הקו האמצעי בתוך התיבה – החציון (Median), שמראה את הערך המרכזי בנתונים.
* החלק התחתון והעליון של התיבה – הרבעון הראשון (Q1, 25%) והרבעון השלישי (Q3, 75%), שמייצגים את טווח הערכים המרכזיים.
* האורך של התיבה (IQR - Interquartile Range) – מראה את פיזור הנתונים בין הרבעונים.
* הזקיפים (Whiskers) (הקווים) – מייצגים את הגבולות של הנתונים הרגילים.
* הנקודות שמופיעות מחוץ לקווים – אלו הם חריגים (Outliers), כלומר ערכים שנמצאים מחוץ לטווח הצפוי של הנתונים.
* הגרף מציג את התפלגות מחירי הטיסות לפי יעד (פריז, רומא ולונדון). טיסות לפריז הן הזולות ביותר עם טווח מחירים קטן יותר, בעוד שללונדון ורומא יש טווח מחירים רחב יותר.

מסקנה: פריז היא יעד יותר חסכוני עם יותר טיסות במחירים נמוכים, בעוד שלונדון ורומא מציגות שונות גבוהה יותר במחירי הכרטיסים.

* הגרף מציג את התפלגות מחירי הטיסות לפי יעד באמצעות Box Plot.

לרומא יש את טווח המחירים הרחב ביותר וחריגים רבים, מה שמעיד על שונות גבוהה במחירי הטיסות.

מסקנה: טיסות לפריז וללונדון יציבות יותר, בעוד שרומא מציגה פיזור מחירים רחב עם טיסות זולות לצד יקרות מאוד.

* הגרפים הם Box Plots שמציגים את התפלגות מחירי הטיסות לפי חברת תעופה לכל יעד (פריז, רומא ולונדון).

פריז היא היעד הזול והיציב ביותר, עם טיסות לואו-קוסט רבות, במיוחד של easyJet.

רומא מציגה את טווח המחירים הרחב ביותר, עם שונות גבוהה וחריגים רבים, בעיקר בחברות פרימיום כמו British Airways ו-Iberia.

ללונדון יש פיזור מחירים רחב, אך החציון שלה נמוך יותר מרומא, עם טיסות יקרות במיוחד בחברות פרימיום כמו Air France.

באופן כללי, חברות לואו-קוסט שומרות על מחירים נמוכים יותר, אך עם שונות גבוהה, בעוד שחברות מסורתיות מציגות מחירים גבוהים ויציבים פחות.

* הגרף מציג את השפעת משך הקונקשן על מחירי הטיסות בשלושת היעדים: פריז, רומא ולונדון.

טיסות ישירות הן לרוב היקרות ביותר, בעוד שטיסות עם חניית ביניים עשויות להוזיל את המחיר אך לא תמיד באופן משמעותי.

קונקשנים קצרים לא בהכרח מובילים למחירים נמוכים יותר ועלולים אף לייקר את הטיסה.

ברומא יש את טווח המחירים הרחב ביותר, כולל טיסות לואו-קוסט לצד טיסות פרימיום יקרות.

חלק מהטיסות עם קונקשנים ארוכים כוללות מחירים חריגים, כנראה בשל טיסות פרימיום או ביקוש עונתי.

* הגרף מציג את השפעת מספר הקונקשנים על מחירי הטיסות.

טיסות ישירות או עם קונקשן אחד מציגות את טווח המחירים הרחב ביותר, עם שונות גבוהה.

טיסות עם שני קונקשנים נוטות להיות זולות יותר אך בעלות שונות גדולה במחירים.

טיסות עם שלושה קונקשנים מציגות את המחירים היציבים ביותר, אך אין מגמה ברורה שמראה שמספר הקונקשנים משפיע באופן ישיר על המחיר.

* הגרף מציג את הקשר בין משך הטיסה הכולל (בדקות) לבין המחיר הממוצע.

מחירי הטיסות אינם עולים באופן עקבי עם משך הטיסה, אלא משתנים בתנודות.

הטיסות היקרות ביותר הן בסביבות 400-600 דקות ו-3200-3400 דקות, ככל הנראה בשל מסלולי פרימיום או טיסות ארוכות טווח.

לאחר 600 דקות, המחירים נוטים לרדת לפני שהם עולים שוב בטיסות ארוכות במיוחד.

* הגרף מציג את השפעת משך הטיסה (הלוך וחזור) על המחיר.

טיסות הנמשכות 200-800 דקות מציגות שינויים חדים במחיר, עם עליות פתאומיות.

בטיסות ארוכות מ-1000 דקות המחירים יציבים יותר, אך חלקן עדיין יקרות מאוד.

הטיסות חזור מציגות דפוס דומה, אך עם יותר שינויים במחירים עבור טיסות קצרות.

* הגרפים מציגים את מספר ההמראות והנחיתות לפי שדה תעופה בכל אחד מהיעדים: לונדון, רומא ופריז.

בלונדון, Gatwick (LGW) הוא השדה העמוס ביותר, עם הרבה יותר טיסות משדות אחרים.

ברומא, Fiumicino (FCO) שולט בתנועת הטיסות, ומטפל בהרבה יותר טיסות מ-Ciampino (CIA).

בפריז, Orly (ORY) הוא השדה הפעיל ביותר, ואחריו Charles de Gaulle (CDG), בעוד ש-Beauvais (BVA) מטפל בפחות טיסות.

* הגרפים מציגים את מחירי הטיסות הממוצעים לפי שדה תעופה בלונדון, רומא ופריז.

בלונדון, City Airport (LCY) הוא היקר ביותר, כנראה בשל קירבתו לעיר ושירות לעסקים, בעוד ששדות רחוקים יותר כמו Stansted ו-Luton מציעים מחירים נמוכים יותר.

ברומא, המחירים ב-Fiumicino (FCO) וב-Ciampino (CIA) דומים, למרות שהמרחק שלהם מהעיר שונה, כנראה בגלל אופי הטיסות (בינלאומיות מול לואו-קוסט).

בפריז, Charles de Gaulle (CDG) מציע את המחירים הנמוכים ביותר בשל תחרות גבוהה, בעוד ש-Beauvais (BVA) מרוחק אך עדיין עם מחירים יחסית גבוהים בשל פחות אפשרויות תחבורה.

5. חקירה ממוקדת של השפעת הפיצרים של המחיר (בהמשך משתנה המטרה)

בשלב הראשון הצגנו מטריצת קורלציה (Heatmap) – שימוש ב-Seaborn (sns.heatmap()) כדי להמחיש את הקשרים בין המשתנים. ולבדוק לאיזה משתנים יש השפעה הכי גדולה על המחיר.

**קורלציה:**

קורלציה היא מדד סטטיסטי שבודק את מידת הקשר בין שני משתנים. הערך שלה נע בין -1 ל-1:

* קורלציה חיובית (0 עד 1) – כאשר ערך של משתנה אחד עולה, גם השני נוטה לעלות.
* קורלציה שלילית (-1 עד 0) – כאשר ערך של משתנה אחד עולה, השני נוטה לרדת.
* קורלציה אפסית (0) – אין קשר בין המשתנים.

חשוב לזכור שקורלציה אינה מעידה על סיבתיות – כלומר, קשר חזק בין שני משתנים לא בהכרח אומר שאחד גורם לשני.

ששת המשתנים עם הקורלציה הגבוהה ביותר (בערך מוחלט):

* Total Flight Duration (min) → 0.439 (ככל שהטיסה ארוכה יותר, המחיר גבוה יותר).
* Total Onward Flight Duration (min) → 0.410 (משך טיסת ההלוך קשור למחיר).
* Total Return Flight Duration (min) → 0.387 (משך טיסת החזור משפיע באופן דומה).
* Carry-ons → 0.199 (כמות הכבודה שניתן להעלות למטוס עשויה להשפיע על המחיר, במיוחד בטיסות לואו-קוסט).
* Departure Day → -0.169 (ככל שיום ההמראה גבוה יותר בחודש, המחיר נוטה לרדת).
* Onward Departure Airport Distance → -0.076 (ככל ששדה התעופה רחוק יותר מהעיר, המחיר נוטה להיות נמוך יותר).

PairGrid של ששת המשתנים עם הקורלציה הגבוהה:

PairGrid (הגרף הראשון)- יצרנו PairGrid המציג תרשימי פיזור (scatter plots) של ששת המשתנים שנבחרו אל מול מחיר הטיסה (Price(NIS)).

המטרה הייתה לבדוק אם קיימים דפוסים כלשהם בין כל משתנה למחיר הטיסה.

מטריצת Pairplot (הגרף השני) - יצרנו Pairplot Matrix, המציג את כל ששת המשתנים אחד מול השני, ולא רק מול מחיר הטיסה. מטריצה זו מאפשרת לבדוק קשרים בין כל המשתנים ולראות אם קיימת תלות או קורלציה ביניהם. האלכסון של המטריצה מציג גרפי KDE, הממחישים את התפלגות הנתונים בכל משתנה.

למה הגרפים נראים כך?

1. ערכים מספריים בדידים יוצרים אשכולות ברורים

* משתנים כמו Carry-ons ו-Departure Day מכילים מספר מוגבל של ערכים (למשל, 0, 1, 2 וכו').
* לכן, במקום פיזור אחיד של נקודות, ניתן לראות פסים אנכיים או אופקיים בגרפים.
* למרות שהם מוגדרים כמשתנים מספריים, ההתנהגות שלהם דומה למשתנים קטגוריים (למשל, משתנה עם רק 3 אפשרויות).

1. משתנים רציפים מראים מגמות ברורות יותר

* משתנים כמו Total Flight Duration(min) ו-Onward Departure Airport Distance הם בעלי ערכים רציפים, ולכן ניתן לראות דפוסים חלקים יותר בגרפים.
* הקשרים שלהם עם מחיר הטיסה נראים ברורים יותר בהשוואה למשתנים הבדידים.

1. ריכוזים גבוהים בטווחים מסוימים

* בגרפי ה-KDE שבאלכסון של מטריצת ה-Pairplot ניתן לראות שהנתונים אינם מתפלגים באופן אחיד.
* למשל, מחיר הטיסה מרוכז בטווחים מסוימים, מה שמעיד על כך שרוב הטיסות במחירים דומים.

מה זה KDE (Kernel Density Estimation)?

K כלי סטטיסטי שמטרתו להעריך את פונקציית הצפיפות של ההסתברות (PDF - Probability Density Function) של משתנה מסוים. במקום להציג היסטוגרמה (חלוקה של הנתונים לעמודות בדידות), KDE יוצר עקומה חלקה ורציפה שמתארת איך הנתונים מתפלגים. זה נעשה על ידי הצבת "גרעין" (Kernel) מסביב לכל נקודת נתונים וחיבור כל התרומות יחד כדי לקבל התפלגות משוערת.

בגרפי ה-Pairplot ניתן לראות את KDE באלכסון של מטריצת ה-Pairplot. שם הוא מציג כיצד הנתונים מתפלגים עבור כל משתנה, במקום להציג אותם כהיסטוגרמה ריבועית.

**הסברים על הגרפים (לפי סדר):**

* הגרפים מציגים את השפעת מספר תיקי היד והמזוודות על מחיר הטיסה.

טיסות הכוללות מזוודה נשיאה אחת (Carry-on) יקרות יותר בהשוואה לאלו ללא תיק נשיאה.

הוספת מזוודה שנשלחת לבטן המטוס (Checked Bag) מעלה את המחיר, אך ההשפעה קטנה יותר כאשר נשלחות שתי מזוודות.

מסקנה: דמי כבודה משפיעים על המחיר, במיוחד בטיסות לואו-קוסט שבהן חיובים נוספים על תיקים נפוצים.

* הגרפים מציגים את הקשר בין יום בחודש לבין מחיר הטיסה עבור טיסות הלוך וחזור.

מחירי הטיסות נוטים להיות גבוהים יותר במחצית הראשונה של החודש ולרדת בהמשך.

טיסות חזור מציגות דפוס דומה, עם שיאים סביב אמצע החודש.

מסקנה: ביקוש גבוה בתחילת החודש או זמינות מוגבלת עשויים לגרום לעליות מחירים, בעוד שבחלק מהימים יש עליות מחירים חדות, כנראה בשל אירועים או עומס עונתי.

* הגרף מציג את הקשר בין מספר הימים עד הטיסה (TTT - Time to Travel) לבין המחיר הממוצע.

מחירי הטיסות נוטים לרדת ככל שהתאריך רחוק יותר, כשהמחירים הנמוכים ביותר מופיעים באמצע תקופת ההזמנה.

לקראת מועד הטיסה, המחירים שוב עולים, ככל הנראה בשל ביקוש גבוה להזמנות של הרגע האחרון.

מסקנה: ההזמנה המשתלמת ביותר נעשית לא מוקדם מדי ולא מאוחר מדי, אלא סביב אמצע תקופת ההזמנה.

* הגרף מציג את הקשר בין מרחק שדה התעופה מהעיר לבין מחיר הטיסה (הלוך וחזור).

אין קשר לינארי ברור בין מרחק השדה למחיר – מחירים גבוהים מופיעים הן בשדות קרובים והן ברחוקים.

שדות תעופה מרוחקים מציגים שונות גבוהה יותר במחירים, בעוד ששדות קרובים לעיתים יקרים יותר בשל ביקוש ונוחות לנוסעים.

מסקנה: מרחק שדה התעופה אינו הגורם העיקרי בקביעת המחיר, וגורמים נוספים כמו ביקוש, זמינות ותחרות משפיעים עליו יותר.

* הגרף מציג את ההשפעה של טיסות בסוף השבוע על מחיר הכרטיס.

טיסות בסוף השבוע מציגות חציון מחיר מעט גבוה יותר בהשוואה לטיסות באמצע השבוע, אך ההתפלגות הכללית דומה.

קיום חריגים (outliers) בשתי הקטגוריות מצביע על כך שגורמים אחרים משפיעים יותר על המחיר מאשר עצם הטיסה בסוף השבוע.

מסקנה: טיסות סוף שבוע עשויות להיות מעט יקרות יותר, אך ההבדל אינו משמעותי ביחס לגורמים אחרים כמו משך הטיסה או מספר הקונקשנים.

* הגרף מציג את ההשפעה של חודש ההמראה והחזרה על מחיר הטיסה.  
  מחירי הטיסות באפריל מעט גבוהים יותר בהשוואה למרץ, אך ההבדל אינו משמעותי.  
  התפלגות המחירים בחודש החזרה דומה לזו של חודש ההמראה, מה שמעיד על ביקוש יציב יחסית.  
  מסקנה: נראה כי חודש ההמראה או החזרה אינם גורמים משמעותיים בקביעת מחיר הטיסה, וגורמים אחרים כמו משך הטיסה או מספר הקונקשנים משפיעים יותר.