**קלסטרינג – Clustering**

ניתוח אשכולות היא טכניקת קיבוץ של אובייקטים על בסיס הדמיון ביניהם. ניתוח אשכולות זה סוג של למידה לא מפוקחת, מה שאומר שהיא מוצאת דפוסים בנתונים מבלי להיות מונחית על ידי תוצאה או תיוג ידועים.

המטרה העיקרית של קלסטרינג היא לסווג אובייקטים לאשכולות בהתבסס על תכונותיהם כך שמידת השיוך תהיה חזקה בין חברי אותו אשכול וחלשה בין חברי אשכולות שונים.

**קלסטרינג מבוסס מרחק:** אלגוריתמים אלו משתמשים במרחקים בין נקודות הנתונים (למשל, מרחק אוקלידי) כדי להקצות נקודות נתונים לאשכולות.כל נקודת נתונים משויכת לאשכול על פי קרבתה למרכז אשכול (Centroid) או לפי קרבתה לנקודות נתונים אחרות. האלגוריתם משווה את המרחק בין הנקודות ובוחר את האשכול שבו המרחק הקטן ביותר.

דוגמאות:

* K-Means – משייך נקודות נתונים לאשכולות על פי הקרבה למרכזים (centroids).
* Hierarchical Clustering – בונה היררכיה של אשכולות על פי מרחק בין נקודות נתונים.

**קלסטרינג לא מבוסס מרחק:**  אלגוריתמים אלו אינם מתמקדים במרחקים, אלא מתחשבים בקשרים אחרים בין הנקודות, כמו צפיפות נתונים, קשרים היררכיים או מודלים מתקדמים יותר.

איך זה עובד: במקום להסתמך על מרחק, הם משתמשים במידע על הצפיפות של הנתונים (מה שמעיד על התכנסות לקבוצות) או משתמשים בטכניקות אחרות כמו גרפים או רשתות עצביות.

דוגמאות:

* DBSCAN (Density-Based Spatial Clustering of Applications with Noise) – מזהה אשכולות על פי צפיפות הנקודות.
* Gaussian Mixture Models (GMM) – משתמש בהשערות הסתברותיות ובמודלים סטטיסטיים כדי לזהות אשכולות.

**קריטריונים להערכת קלסטרינג:**

מדדים ללא קרקע אמת (Unsupervised) - מדדים אלו נועדו להעריך את איכות הקלסטרינג ללא תשובה ידועה מראש. הם בוחנים את המבנה של האשכולות ביחס לעצמם.

* Silhouette Score – כלי להערכת איכות הקלסטרינג על סמך המרחקים בין הנקודות בקבוצות השונות. המדד בוחן עד כמה כל נקודה שייכת היטב לקבוצה שלה, תוך השוואה למרחק שלה מקבוצות אחרות.

עקרונות בסיסיים:

* קרבה לקבוצה שלה - חישוב המרחק הממוצע של הנקודה מכל הנקודות האחרות באותה קבוצה (מסומן בa).
* קרבה לקבוצה הקרובה הבאה: חישוב המרחק הממוצע של הנקודה לנקודות בקבוצה השנייה הכי קרובה (מסומן בb).

ערכי המדד S: (נע בין -1 ל-1)

קרוב ל-1: הנקודה שויכה היטב לקבוצה שלה.

קרוב ל-0: הנקודה נמצאת על הגבול בין קבוצות (יש חפיפה בקבוצות)

קרוב ל--1: הנקודה שויכה לקבוצה הלא נכונה.

המדד פחות אפקטיבי עבור שיטות קלסטרינג שאינן מבוססות מרחקים, כמו DBSCAN (שמבוסס על צפיפות) והוא בעל רגישות למבנה הנתונים ולמדדי המרחק (למשל, אוקלידי).

* Elbow Method - שיטה פופולרית לקביעת מספר הקלסטרים האופטימלי (K) עבור אלגוריתם K-Means. השיטה מתמקדת בניתוח השונות בתוך הקבוצות (WCSS - Within-Cluster Sum of Squares) כדי לקבוע מתי הוספת קבוצות חדשות מפסיקה לשפר באופן משמעותי את איכות הקלסטרינג.
* עוזר למנוע בחירת מספר גבוה מדי של קבוצות (Overfitting).
* מספק שיטה אינטואיטיבית ומתמטית לזיהוי מספר הקבוצות.
* כלי פשוט ליישום בשיטות כמו K-Means, שבהן יש להגדיר מראש את מספר הקבוצות.

למה נקראת שיטת המרפק? הגרף נראה כמו זרוע מכופפת, והנקודה שבה הכיפוף (ה"מרפק") מייצגת את מספר הקלסטרים האופטימלי.

* Davies-Bouldin Index – מדד להערכת איכות הקלסטרינג שבוחן את היחס בין הקומפקטיות של כל קבוצה (Compactness) לבין ההפרדה בין הקבוצות (Separation), כאשר ערך נמוך יותר מעיד על קלסטרינג טוב יותר.
* מתאים לשיטות מבוססות מרחק: לדוגמה, יעיל עבור K-Means.
* לא מתאים לקבוצות שאינן כדוריות (Convex): כמו ב-DBSCAN, שבו הקבוצות אינן בהכרח כדוריות או כשיש רעשים (Outliers).

# "קבוצות כדוריות" מתארות קבוצות של נתונים שצורתן במרחב הדו-ממדי או התלת-ממדי דומה לכדור או לעיגול. הכוונה היא שכל נקודה בתוך הקבוצה קרובה יחסית למרכז הקבוצה (Centroid), והמרחק בין הנקודות במרכז לבין הקצוות הוא אחיד פחות או יותר.

* Calinski-Harabasz Index – מדד להערכת איכות הקלסטרינג שמבוסס על היחס בין:
* שונות בין הקבוצות (Between-cluster variance): כמה רחוקות הקבוצות זו מזו.
* שונות בתוך הקבוצות (Within-cluster variance): כמה קרובות הנקודות בכל קבוצה למרכז שלה.

כלומר - היחס בין הפיזור בין הקבוצות לפיזור בתוך הקבוצות.

מתאים לשיטות מבוססות מרחק ופחות מתאים לשיטות שמייצרות קבוצות לא כדוריות או מתמודדות עם רעשים.

ערך גבוה של CH(k) מעיד על קלסטרינג איכותי – קבוצות מובחנות וברורות, עם שונות גבוהה בין הקבוצות ושונות נמוכה בתוך הקבוצות.

מדדים עם קרקע אמת (Supervised Metrics) –

מדדים אלו משווים את התוצאה של אלגוריתם הקלסטרינג לחלוקה ידועה מראש של הנתונים. המטרה היא להעריך עד כמה הקבוצות שנוצרו על ידי האלגוריתם דומות לחלוקה "האמתית".

נשתמש כאשר יש בידינו תוויות אמיתיות (קרקע אמת) עבור כל נקודה במערך הנתונים.

* Accuracy – בודקים כמה נקודות סווגו נכון ביחס לתשובה הידועה מראש. מחושב ע"י חלוקת הנקודות שסווגו נכון בסך כל הנקודות.

נציין שהמדד פחות יעיל אם קבוצות הנתונים אינן מאוזנות בגודלן, כי המדד עלול להטות לטובת הקבוצה הגדולה יותר.

* Adjusted Rand Index (ARI) – בדיקה כמה הקבוצות שהתקבלו דומות לחלוקה הידועה מראש, תוך התחשבות בסיווג אקראי. מחשבים את הדמיון בין הקבוצות בפועל לקבוצות הידועות, תוך השוואה למקרה אקראי.

הערך נע בין 1- ל1 (כאשר - (1 = התאמה מושלמת, 0 = התאמה אקראית, 1- = התאמה גרועה).

**הערכת איכות הקלסטרינג:**

* הInertia הוא מדד שמחשב את הקומפקטיות של האשכולות על ידי סיכום ריבועי המרחקים בין כל נקודה במרחב לבין מרכז האשכול שלה. הוא מציין עד כמה הנקודות באשכול קרובות אחת לשנייה, אך אינו מתחשב בהפרדה בין האשכולות עצמם. כמדד פשוט ואפקטיבי, הוא יכול להצביע על איכות האשכולות, אך אינו מספק מידע על האופן שבו האשכולות מופרדים זה מזה.
* הSilhouette Score מודד עד כמה כל אשכול מבודד וברור, בכך שהוא משווה את המרחק בין כל נקודה בממוצע לנקודות שבאות מאשכולות אחרים. הערך נע בין -1 ל-1, כאשר ערך גבוה מציין שהאשכולות מופרדים היטב. מדד זה מספק תמונה טובה של איכות הקלסטרינג על ידי שקלול גם את הקומפקטיות וגם את ההפרדה בין האשכולות.
* הAdjusted Rand Index (ARI) משווה את התוצאה של הקלסטרינג לקרקע אמת (אם קיימת), ומודד את ההתאמה בין קבוצות שנוצרו על ידי האלגוריתם לבין קבוצות שמוגדרות מראש. המדד מנרמל את התוצאה כך שהיא נעה בין -1 ל-1, כאשר ערך של 1 מציין התאמה מושלמת ו-0 מציין שהקלסטרינג היה אקראי. הוא שימושי מאוד כשיש לנו תוויות מוכרות מראש להשוואה.
* הPurity מודד את האיכות של הקלסטרינג על ידי חישוב אחוז הנקודות באשכול שקשורות לאותה קטגוריה מתוך הקטגוריות הידועות מראש. המדד מראה עד כמה כל אשכול טהור, כלומר, כמה הוא מכיל נקודות שקשורות אחת לשנייה. הוא פשוט להבנה, אך עשוי להיות בעייתי אם יש אשכולות קטנים או כאשר יש חפיפות בין האשכולות.

**הכנות לביצוע קלסטרינג – אתגרים ומגבלות:**

הכנה לנתונים לפני ביצוע קלסטרינג היא שלב קריטי להבטחת תוצאות מדויקות ומשמעותיות. תהליך הקלסטרינג רגיש למבנה ולמאפייני הנתונים, כמו מספר הממדים, פיצ'רים לא רלוונטיים, ערכים חריגים והבדלים בטווחי הערכים בין המשתנים. הכנה מקדימה מסייעת להתמודד עם האתגרים הללו על ידי ניקוי, עיבוד והתאמת הנתונים כך שיתאימו טוב יותר לאלגוריתם הנבחר. הכנה נכונה לא רק משפרת את איכות החלוקה לקבוצות, אלא גם הופכת את התהליך למהיר ויעיל יותר.

בעת עבודה עם קלסטרינג, קיימים מספר אתגרים שעלולים להקשות על איכות התוצאות:

1. קללת הממדיות (Curse of Dimensionality): קללת הממדיות מתייחסת לאתגרים הנובעים מריבוי פיצ'רים (Features) בנתונים, שגורמים למורכבויות בתהליך הניתוח והמודלים. כאשר מספר הפיצ'רים גדל:

הנתונים נעשים פזורים מאוד במרחב, וקשה לזהות דפוסים ברורים.

המרחקים בין נקודות נעשים פחות משמעותיים, וקשה להבדיל בין נתונים קרובים ורחוקים.

מודלים מבוססי מרחקים (כמו K-Means) מתקשים לזהות אשכולות ברורים.

פתרון: הפחתת ממדים (Dimensionality Reduction) באמצעות שיטות כמו PCA (Principal Component Analysis), שמפחיתה את המימדים על ידי זיהוי רכיבים עיקריים שמרכזים את עיקר השונות בנתונים.

1. רגישות לחריגים (Outliers): ערכים חריגים (Outliers) הם נקודות בנתונים השונות באופן משמעותי משאר הנתונים, והם עלולים לשבש את תהליך הקלסטרינג, במיוחד בשיטות מבוססות מרחקים כמו K-Means. ערכים חריגים עלולים לגרום להסטה משמעותית של המרכזים (Centroids) ולהשפיע על חלוקה לא מדויקת.

שיטות לזיהוי והסרה:

* בוקס-פלוט (Box Plot): לזיהוי נקודות קיצוניות באמצעות שיטת Tukey's Fences.
* Isolation Forest ו- Z-Score: כלים מתקדמים לזיהוי חריגים.

1. תלות בפרמטרים: אלגוריתמים רבים בקלסטרינג דורשים קביעת פרמטרים קריטיים מראש, המשפיעים על איכות התוצאות. בחירה לא נכונה של פרמטרים עלולה להוביל לחלוקה לא מדויקת של הנתונים ולהשפיע על מבנה האשכולות.

דוגמה בולטת היא אלגוריתם K-Means שדורש לקבוע מראש את מספר הקבוצות (K), מה שעלול להוביל לתוצאות לא מדויקות אם הערך לא נבחר בצורה נכונה.

פתרון: מציאת מספר הקבוצות האופטימלי באמצעות שיטת המרפק או Silhouette Score.

שיטות לשיפור איכות הקלסטרינג:

* הפחתת ממדים: כפי שצויין קודם, הפחתת ממדים כמו PCA מסייעת להסיר ממדים לא רלוונטיים שמכילים רעש ושיפור ביצועי המודל.
* נרמול הנתונים: נירמול חשוב בהכנת הנתונים לקלסטרינג, במיוחד כשמדובר במודלים מבוססי מרחקים. נרמול משווה את סקאלת הנתונים וימנע השפעה לא פרופורציונלית של נתונים עם ערכים גבוהים יותר על המרחקים.

שיטות לנרמול:

* Z-Score Normalization: ממיר את הנתונים כך שלכל משתנה יהיה ממוצע 0 וסטיית תקן 1.
* Min-Max Scaling: ממיר את הנתונים לטווח 0-1 תוך שמירה על יחסיות הערכים.
* בחירת פיצ'רים: השארת פיצ'רים מיותרים או בעלי מידע מועט עלולה לשבש את חישובי המרחקים ולפגוע באיכות הקלסטרינג. שיטות לבחירת פיצ'רים:
* Variance Threshold: הסרת פיצ'רים בעלי שונות נמוכה.
* Correlation Analysis: הסרת פיצ'רים עם קורלציה גבוהה.
* PCA: שימוש ב-PCA להמיר פיצ'רים למימדים חדשים שמייצגים את עיקר השונות.

**שלבי הקלסטרינג במחברת:**

שלב 1: סינון הטיסות על פי חברת ההלוך והחזור

בשלב זה אנחנו מסננים את הטיסות כך שמוצגות טיסות רק כאשר הטיסה הלוך והחזור מבוצעת ע"י אותה חברת תעופה. הסיבה היא כדי להבטיח שהתהליך יתמקד במדיניות תמחור אחידה של אותה חברת תעופה. אם נתחשב בטיסות שבהן חברת ההלוך שונה מחברת החזור, נוכל לקבל תמונה מעוותת או לא עקבית של מדיניות התמחור.

שלב 2: הבנת מדיניות התמחור

בשלב הזה, המטרה היא להבין את מדיניות התמחור של כל חברת תעופה על ידי זיהוי הגורמים שמשפיעים על המחיר. לצורך כך, יצרנו DataFrame בשם df\_cluster שמאגד את נתוני התמחור של כל חברה לפי מאפיינים שונים. כל שורה ב-DataFrame מייצגת חברת תעופה, וכל עמודה מכילה סטטיסטיקות שמתארות את מדיניות התמחור שלה, כמו משך זמן טיסה, כבודה יד, יום ושעה של טיסה, מחיר ממוצע וסטיית תקן של המחיר. נתונים אלו מאפשרים להשוות בין החברות ולהבין את השיקולים במדיניות התמחור שלהן.

שימוש בסטטיסטיקות כמו ממוצעים, חציון וסטיית תקן עוזרת להבין איך המחירים משתנים בין החברות, ומאפשרת זיהוי מגמות ודפוסים שמשפיעים על התמחור. ניתוח הקורלציות בין תכונות אלה למחיר מאפשר ליצור מדיניות תמחור מבוססת נתונים לכל חברת תעופה.

הטבלה מכילה את הנתונים הבאים:

* avgOnwardFlightDurationMin: ממוצע משך זמן טיסת ההלוך לכל חברת תעופה.
* avgReturnFlightDurationMin: ממוצע משך זמן טיסת החזור לכל חברת תעופה.
* medTTT: החציון של TTT (Total Travel Time) – הזמן הכולל של הנסיעה (ההפרש בין יציאה לחזרה).
* avgCarryOns: ממוצע כמות הכבודה שנלקחת ככבודה יד (Carry-ons) בכל טיסה של החברה.
* medDepDay: החציון של Departure Day – יום היציאה עבור כל חברת תעופה.
* medRetDay: החציון של Return Day – יום החזרה עבור כל חברת תעופה.
* medRetMonth: החציון של Return Month – חודש החזרה עבור כל חברת תעופה.
* avgPrice: ממוצע המחיר של הטיסות עבור כל חברת תעופה.
* priceStd: סטיית התקן של המחירים עבור כל חברת תעופה, שמציינת את פיזור המחירים.
* flightsCount: מספר הטיסות הכולל עבור כל חברת תעופה (מספר הרשומות של מחירי הטיסות).

שלב 3: סינון חברות עם פחות מ-1% מהנתונים

סיננו את החברות שמייצגות פחות מ-1% מהטיסות הרלוונטיות על מנת להימנע מהתעסקות עם קבוצות קטנות מדי שיכולות להשפיע על תוצאות הקלסטרינג באופן שגוי. נתונים מועטים יכולים להוות רעש במודל, להשפיע על יציבות הקלאסטרים ולפגוע בהבנה הכללית של מדיניות התמחור של החברות. בנוסף, נתונים כאלה לא תמיד מייצגים את המאפיינים הכלליים של השוק, ועשויים להוביל לקביעת קשרים לא נכונים בין הנתונים. סינון זה משפר את דיוק הקלסטרינג ומייעל את זמן החישוב.

שלב 4: נרמול הנתונים עם StandardScaler

בשלב הזה, בוצע נרמול של הנתונים באמצעות StandardScaler, שמסייע לאזן את המשקל של כל מאפיין (feature). הנרמול חשוב משום שכל מאפיין יכול להיות בסקאלה שונה, כלומר, יש לו ערכים בטווחים שונים (למשל, מחיר יכול להיות בטווח של 1,000-10,000, בעוד שמשך זמן טיסה יהיה בטווח של 30-600 דקות).

ללא נרמול, מאפיינים עם ערכים גבוהים יותר ישפיעו באופן לא פרופורציונלי על חישובי המרחקים (כמו במודלים מבוססי מרחקים כמו K-Means), מה שיכול לגרום לכך שהאלגוריתם יתמקד יותר במאפיינים בעלי ערכים גבוהים.

לאחר הנרמול, כל הפיצ'רים נשארים באותה סקאלה, ומאפשרים השוואה והשלכה מדויקת יותר על תוצאות הקלסטרינג.

**מה עושה StandardScaler?**

מבצע נרמול להתפלגות נורמלית כך ש כל מאפיין יהיה - ממוצע (Mean) של 0 וסטיית תקן (Standard Deviation) של 1. לאחר הנרמול, כל הפיצ'רים יהיו בקנה מידה אחיד וניתן יהיה להשוות ביניהם בצורה יותר הוגנת וניתן להבטיח שכל הפיצ'רים ישפיעו בצורה שווה על תהליך הקלסטרינג.

שלב 5: קביעת מספר הקבוצות האופטימלי לקלסטרינג

* Elbow Method: השתמשנו בשיטת המרפק כדי לקבוע את מספר הקבוצות האופטימלי. גרף ה- Elbow Method מראה את השינוי ב- Within-Cluster Sum of Squares (WSS) ככל שמספר הקבוצות עולה. הנקודה שבה הירידה ב-WSS מאיטה (ה"מרפק") היא הנקודה האופטימלית שבה כדאי לבחור את מספר הקבוצות. בנוסף, השתמשנו ב- KneeLocator כדי לאתר את הנקודה הזאת בצורה אוטומטית.
* דנדרוגרם: השתמשנו ב- Dendrogram כדי לראות איך הקבוצות מאוחדות בתהליך הקלסטרינג ההיררכי. הדנדרוגרם מראה את הדרך שבה הקבוצות מתמזגות, וזה עוזר להבין את הקשר בין הקבוצות השונות. כאשר מדובר באלגוריתם היררכי, כל צומת בגרף מייצג את האיחוד של שתי קבוצות.

מעבירים את הקו האופקי בנקודה בה המרחק בין הצמתים גבוה בצורה משמעותית, וזה קובע את מספר הקבוצות (K) האופטימלי. כל חיתוך של הקו ייצג קבוצה נפרדת.

* Silhouette Score: כדי לבדוק את איכות הקלסטרינג, חישבנו את Silhouette Score עבור מספרים שונים של קבוצות (K). ציון גבוה יותר מציין שקבוצות מופרדות בצורה ברורה יותר. עבור כל מספר K, חישבנו את הציון והדפסנו את התוצאה. הציון נע בין -1 ל-1, כאשר ערך קרוב ל-1 מציין שהקבוצות מופרדות בצורה טובה, וערך קרוב ל-0 מציין חפיפות בין הקבוצות.

בבדיקה שביצענו, השתמשנו ב-Silhouette Score כדי להעריך את איכות הקלסטרינג עבור מספר קבוצות שונות (2 עד 7). המסקנה היא ש-4 קבוצות הן האופטימליות, שכן ציון ה-Silhouette עבור 4 קבוצות היה הגבוה ביותר (0.52), מה שמעיד על חלוקה טובה וברורה יותר בין הקבוצות. מעבר לכך, הציון ירד, מה שמצביע על חפיפות גדולות יותר בין הקבוצות והפחתת איכות הקלסטרינג.

שלב 6: הגדרת המשתנים

X = df\_scaled – הגדרת X כנתונים המנורמלים מהדאטה.

labels = df\_cluster.index – הגדרת התוויות ((labels כשמות חברות התעופה.

שלב 7: הפעלה של אלגוריתם K-Means

**הסבר על האלגוריתם באופן כללי:**

אלגוריתם K-Means הוא אלגוריתם קלסטרינג (קיבוץ) לא מפוקח, שמטרתו לחלק את הנתונים ל-K קבוצות (clusters) כך שנקודות דומות יימצאו באותה קבוצה. זהו אלגוריתם מבוסס מרחק, כלומר הוא מתבסס על חישוב מרחקים בין נקודות נתונים ומרכזי הקבוצות (centroids) כדי לשייך את הנתונים לאשכולות.

המרחק המשמש לחישוב קרבה: האלגוריתם מתבסס על מרחק אוקלידי בין הנקודות למרכזי האשכולות, כך שלפני הפעלת האלגוריתם כדאי לנרמל את הנתונים. נרמול חשוב כדי למנוע ממאפיינים בעלי ערכים גבוהים יותר להטות את חישוב המרחק ולהשפיע באופן לא פרופורציונלי.

שלבי הפעולה של K-Means:

* אתחול: יש לקבוע את מספר הנקודות למספר הקבוצות (K) מראש (קבענו 4 לפי Silhouette Score). האלגוריתם בוחר K מרכזים אקראיים כנקודות התחלה (centroids). חשוב לציין שהשיטה הרגילה לבחירת המרכזים עלולה להוביל לתוצאות פחות טובות, ולכן השתמשנו ב- k-means++ לאתחול חכם של המרכזים. בשיטה זו, המרכזים נבחרים בצורה כזו שיבטיחו מרחקים טובים יותר בין המרכזים הראשוניים, מה שמפחית את הסיכון למיקומים לא אופטימליים של המרכזים הראשוניים ומסייע בהשגת תוצאות מדויקות יותר.
* שיוך נקודות: כל נקודת נתון משויכת לאשכול שהמרכז שלו הקרוב ביותר, בהסתמך על מרחק אוקלידי בין הנקודה לבין המרכזים. חשוב לציין שאלגוריתם רגיש לערכים חריגים שכן הם יכולים להשפיע על המיקום של המרכזים ולהשפיע על איכות הקיבוץ.
* עדכון המרכזים: לאחר שיוך כל הנקודות, האלגוריתם מחשב את המרכז החדש לכל אשכול, שהוא ממוצע כל הנקודות באותו אשכול.
* חזרה על התהליך: השלב השני והשלב השלישי חוזרים על עצמם עד שהמרכזים מתייצבים, כלומר אין שינוי משמעותי במיקומים שלהם או עד שהגיעו למספר איטרציות מקסימלי.

שלב 8: הפחתת מימדים

בשלב הזה, השתמשנו ב-PCA (Principal Component Analysis), שהיא שיטה להורדת ממדים (Dimensionality Reduction), שמטרתה להפוך את הנתונים המורכבים לפשוטים יותר. באמצעות PCA, צמצמנו את מספר הממדים (הפיצ'רים) בנתונים ל-2 רכיבים עיקריים, שמייצגים את עיקר השונות בנתונים. כך, ייצגנו את הנתונים ב-2 ממדים בלבד, תוך שמירה על המידע החשוב ביותר.

הפעלנו את ה-PCA על הנתונים המנורמלים (X), וכך הפכנו את הנתונים לשני ממדים בלבד. כל רכיב עיקרי שנוצר הוא שילוב לינארי של הפיצ'רים הישנים, והוא מכיל את רוב השונות וההבדלים בנתונים.

איך אפשר לדעת אילו פיצ'רים נשארו? ה-PCA לא שומר את הפיצ'רים המקוריים, אלא יוצר רכיבים חדשים שהם שילוב של הפיצ'רים. עם זאת, אפשר לדעת את המשמעות של כל רכיב ע"י הסתכלות על המשקלים של הפיצ'רים לכל רכיב עיקרי, שנמצאים ב-pca.components\_. כל שורה ב-pca.components\_ מייצגת רכיב עיקרי, וכל עמודה מייצגת פיצ'ר. הערכים הללו מציינים עד כמה כל פיצ'ר משפיע על כל רכיב עיקרי.

למה זה חשוב?

הפחתת ממדים בעזרת PCA מסייעת לצמצם את המורכבות של הנתונים, לשפר את ביצועי המודלים, ולהפוך את הנתונים לברורים יותר, תוך שמירה על עיקר המידע. בנוסף, מאפשרת לנו להבין איזה פיצ'ר תורם ביותר לכל רכיב עיקרי, ולבצע ניתוחים יעילים יותר.

**בשלב זה ניתן לראות מהגרף שיצרנו שבאופן כללי חברות תעופה עם מדיניות תמחור דומה מקובצות יחד. עם זאת, קשה להבין את כל הקשרים בגלל שהנתונים צומצמו ל-2 ממדים בלבד. יש צורך לפרק את הנתונים לקטגוריות שונות כדי להבין את הדפוסים והקשרים בצורה יותר ברורה.**

בשלבים אלו בדקנו את הקיבוץ (קלסטרינג) של הנתונים על פי פיצ'רים שונים, תוך שימוש בשיטות זהות – שימוש באותו אלגוריתם, קביעת מספר הקבוצות באמצעות שיטת המרפק ושימוש במדד Silhouette Score להערכת איכות הקיבוץ.

1. קלסטרינג על פי מחיר ממוצע ומספר טיסות

"Good Value" – חברות תעופה עם מחירים סבירים.

"High Quantity, Low Quality" – חברות עם מחירים נמוכים ומספר טיסות גבוה.

"Low Quantity, High Quality" – חברות עם מחירים גבוהים ומספר טיסות נמוך.

1. קלסטרינג על פי פרמטרים נוספים (זמן טיסה (Onward ו-Return Duration), כבודה (Carry-Ons), ימים, ועוד). השתמשנו בהפחתת ממדים באמצעות PCA לצמצום הנתונים לשני ממדים בלבד, ולאחר מכן ביצוע קלסטרינג על הנתונים המנורמלים.

חברות תעופה ממקמות את עצמן לפי קריטריונים שונים כמו מחירים (זולים/יוקרה) ומאפיינים של הטיסות.

השימוש ב-PCA עזר להציג את הנתונים בצורה ברורה וממוקדת.

1. קלסטרינג לפי מדדי מחיר וסטיית תקן של המחיר

ניתוח כיצד סטיית התקן משפיעה על הקבוצות השונות של מחיר.

הצגת המחירים הממוצעים לכל קבוצה בגרפים כמו box plots ו-violin plots.

הקבוצות שהתקבלו מייצגות מדיניות תמחור שונה, למשל, חברות עם מחירים יציבים מול חברות עם מחירים משתנים מאוד.

4. קלסטרינג לפי חודש חזרה וסטיית תקן של המחיר

**מסקנות עסקיות עבור מחלקות שיווק וניתוח נתונים בתחום התעופה:**

1. מיקוד באסטרטגיות תמחור

* מחירים נמוכים ומספר טיסות גבוה (High Quantity, Low Quality): חברות תעופה בקבוצה זו מציעות טיסות זולות עם פחות דגש על איכות השירות. חברות כמו easyJet ו-Ryanair מתאימות לאסטרטגיה הזו, שבה המחירים אטרקטיביים, אך לא תמיד רמת השירות היא הגבוהה ביותר.
* מחירים גבוהים ומספר טיסות נמוך (Low Quantity, High Quality): חברות כמו British Airways ו-Air France משתייכות לקבוצה זו. הן מתמקדות בטיסות יוקרה עם דגש על חווית לקוח גבוהה, ותפנה ללקוחות שמוכנים לשלם יותר עבור שירותים יוקרתיים וייחודיים.
* Good Value: חברות שמספקות מחירים סבירים עם תמורה טובה. גישה זו פונה לרוב השוק, במיוחד ללקוחות שמחפשים איזון בין מחיר ושירות.

1. אסטרטגיות שיווק מותאמות

* שיווק ללקוחות תקציביים: עבור חברות בקטגוריית "High Quantity, Low Quality", ניתן לפתח אסטרטגיות שיווק שמיועדות ללקוחות שמחפשים טיסות זולות ובכמויות גדולות, לדוגמה חבילות למשפחות או צעירים שמחפשים אטרקציות במחירים נמוכים.
* שיווק ללקוחות יוקרה: חברות כמו British Airways ו-Air France יכולות לפתח קמפיינים ממומנים שממוקדים לקהלים שמחפשים שירותים יוקרתיים, טיסות למטרות עסקים או חופשות יוקרתיות. השיווק יתמקד בלקוחות המעוניינים בטיסות ייחודיות או במעמד גבוה.
* הצעת עסקאות טובות: עבור הקבוצה בקטגוריית "Good Value", ניתן להדגיש את היתרון של המחיר יחד עם איכות השירות. אפשר להציע טיסות במחירים תחרותיים לקהל שמחפש יחס טוב יותר בין מחיר לשירות.

1. תכנון מסלולים ושירותים

* הבדלים בין טיסות ישירות לארוכות: חברות שמתמקדות בטיסות ישירות כמו easyJet ו-Ryanair צריכות להציע יותר טיסות במחירים נמוכים. חברות יוקרה כמו British Airways ו-Air France צריכות לפתח מסלולים ייחודיים לאזורים מיוחדים או אירועים יוקרתיים.
* זמינות טיסות: חברות כמו Transavia France ו-Wizz Air שמציעות טיסות במחירים סבירים אך תכופות, צריכות להציע יותר גמישות בשעות הטיסה, כדי למשוך לקוחות שמחפשים טיסות זולות וגמישות.

4. שיפור חווית הלקוח

* כבוד ושירות: חברות עם מחירים גבוהים, כמו British Airways, צריכות להדגיש את שירותי הכבוד (כמו ביטול ושינוי טיסות) וליצור חווית לקוח ייחודית.
* שירות עצמי ותקציב: חברות כמו Ryanair ו-easyJet יכולות לשפר את חווית הלקוח על ידי יצירת אפליקציות ושירותים שמייעלים את החוויה מבלי להוסיף עלויות נוספות.
* תכנון מבצעי לקראת תקופות מסוימות