**פרויקט במבוא למדע הנתונים**

**עוסק בקבצים: employeesו- clothes**



**שמות המגישות:**

שיר מוסאי – 212769152

חן מוסאי – 213062326

טל מוסאי – 213062318

דנה אליאס – 207895020

**תיאור הפרויקט:**

קריאת מקורות המידע של הקבצים: employees, clothes מאתר .Kaggle ניקוי הנתונים והכנתם לקראת ניתוחים שונים וביצוע ניתוחים בסיסיים על נתונים אלה.

**תובנות של 1.4**

**קובץ input1\_df (employees)**

* מספר רשומות: 1000, מספר עמודות:9
* מספר הערכים המלאים עבור כל משתנה (לא Null):
* ID: 1000, אין ערכים ריקים.
* First Name: 837
* Gender: 770
* Start Date: 504, כמעט חצי מהערכים ריקים.
* Last Login Time: 767
* Salary: 805
* Bonus %: 799
* Senior Management: 1000, אין ערכים ריקים.
* Team: 767

הוספנו לקובץ זה את העמודה ID ובנוסף שיבשנו ומחקנו חלק מהנתונים.

**תובנות של 1.4**

**קובץ input2\_df (clothes)**

* מספר רשומות: 1000, מספר עמודות: 7
* מספר הערכים המלאים עבור כל משתנה (לא Null):
* 1000 :ID, אין ערכים ריקים.
* Brand: 809
* Category: 805
* Color: 805
* Size: 840
* Material: 820
* 824 :Price

הוספנו לקובץ זה את העמודה ID ובנוסף שיבשנו ומחקנו חלק מהנתונים.

**תובנות 1.5**

**קובץ input1\_df**

אחוז הערכים החסרים הוא 19.45% (גדול מ 10%)

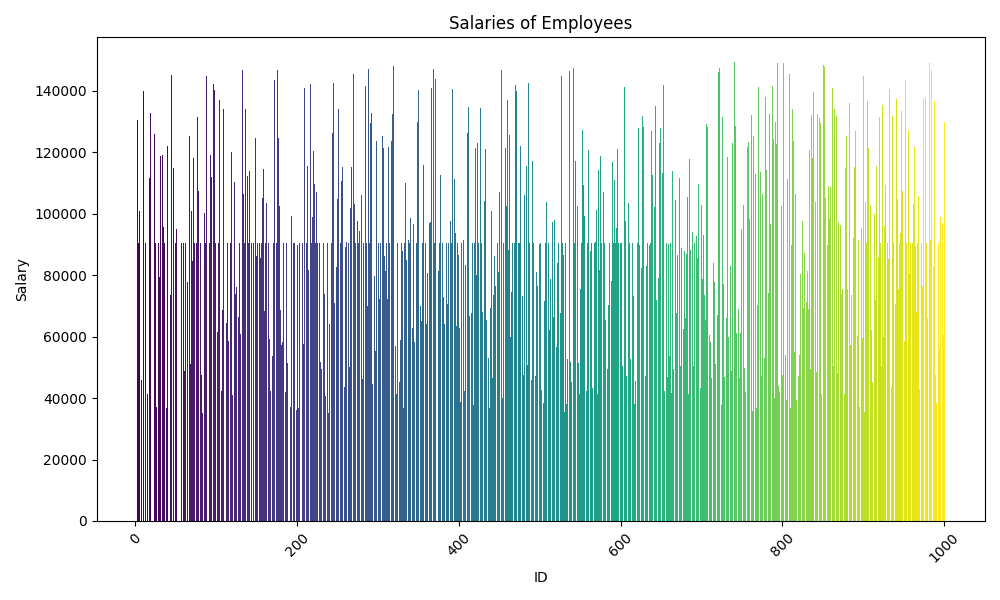
**קובץ input2\_df**

אחוז הערכים החסרים הוא 12.18% (גדול מ 10%)

**תובנות של 3**

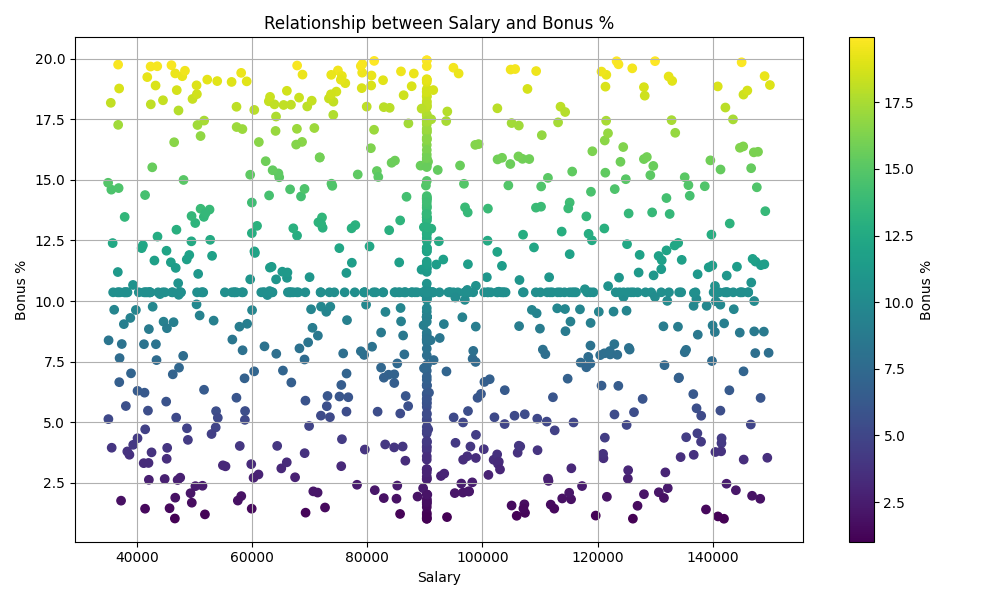
**הצגה ויזואלית של הנתונים**

גרף 1- :bar plot



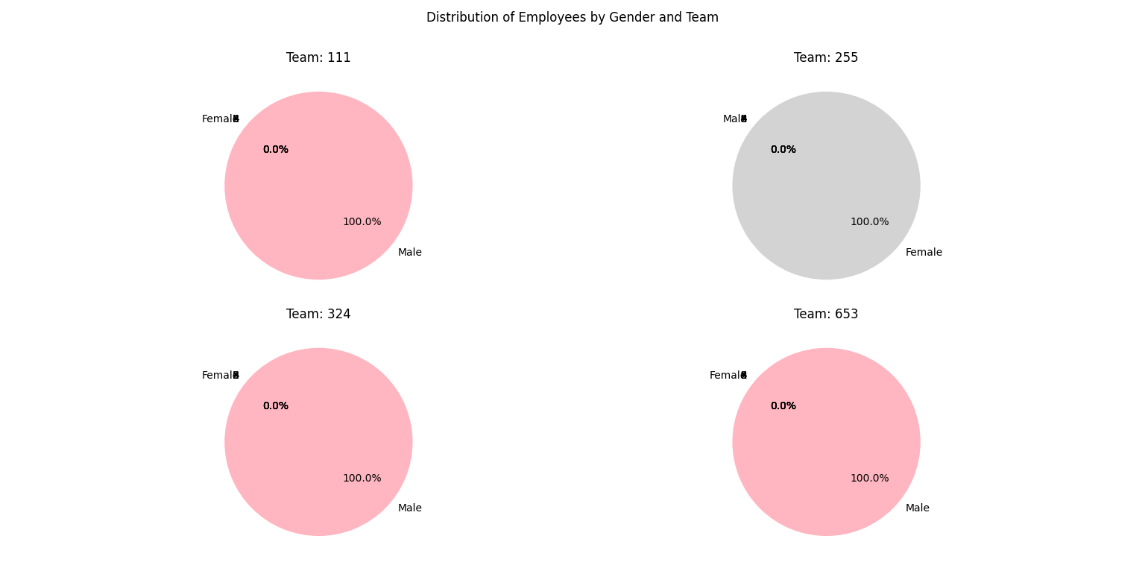
|  |  |
| --- | --- |
|  |  |
|  |

הגרף המוצג ממחיש את רמות השכר של העובדים באמצעות עמודות בצבעים שונים. ציר ה-X מציג את מספרי הזהות של העובדים, בעוד שציר ה-Y מייצג את רמות השכר המקבילות. ניתן לראות פיזור רחב של רמות שכר בין העובדים, כאשר חלקם משתכרים משכורות גבוהות בהרבה מאחרים. השונות ברמות השכר עשויה לנבוע ממגוון גורמים כגון תפקיד, ניסיון, הערכת ביצועים או מדיניות ארגונית. הצבעים הכהים מייצגים שכר נמוך יותר, והצבעים הבהירים מסמנים שכר גבוה יחסית. התרשים מאפשר זיהוי מהיר של פערי שכר ויכול לשמש כבסיס לניתוח נוסף של מבנה התגמול בארגון.

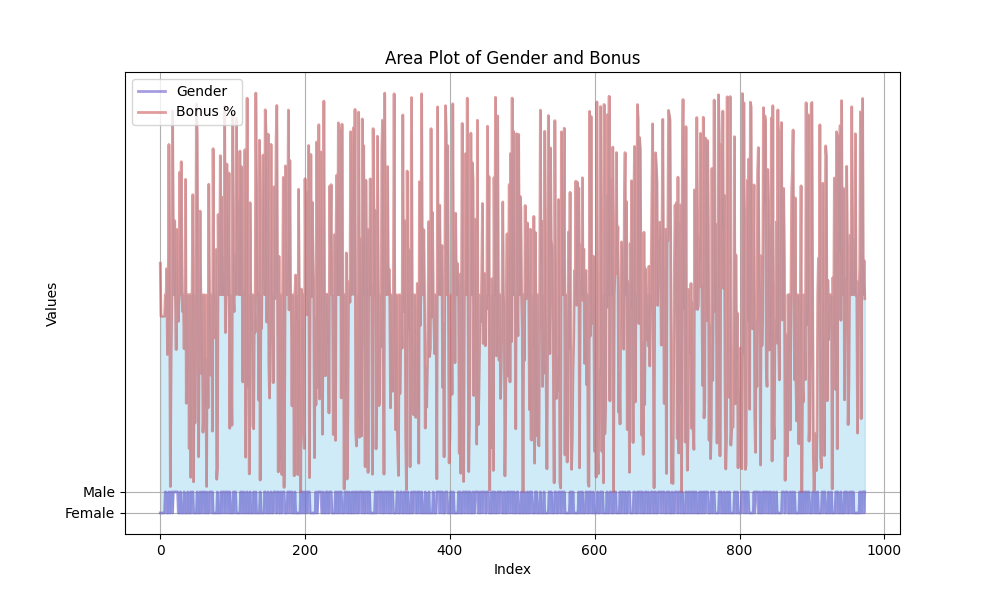
גרף 2 – :scatter plot

הגרף מציג את הקשר בין השכר לבין אחוז הבונוס. ניתן לראות כי אין קשר ברור או ישיר בין השכר לבין אחוז הבונוס שמקבלים העובדים. נראה כי אחוז הבונוס נע בטווח רחב לכל רוחב ציר השכר, אך יש מספר נקודות בולטות. לדוגמה, ניתן להבחין כי רבים מהעובדים מקבלים בונוס של 10%, ללא קשר לגובה שכרם. בנוסף, ישנה הצטברות של נקודות באזורים עם בונוסים נמוכים (2.5%-5%) ובונוסים גבוהים (17.5%-20%). ככל הנראה, ישנם מדדים או קריטריונים נוספים מעבר לשכר, המשפיעים על גובה אחוז הבונוס שמקבלים העובדים.

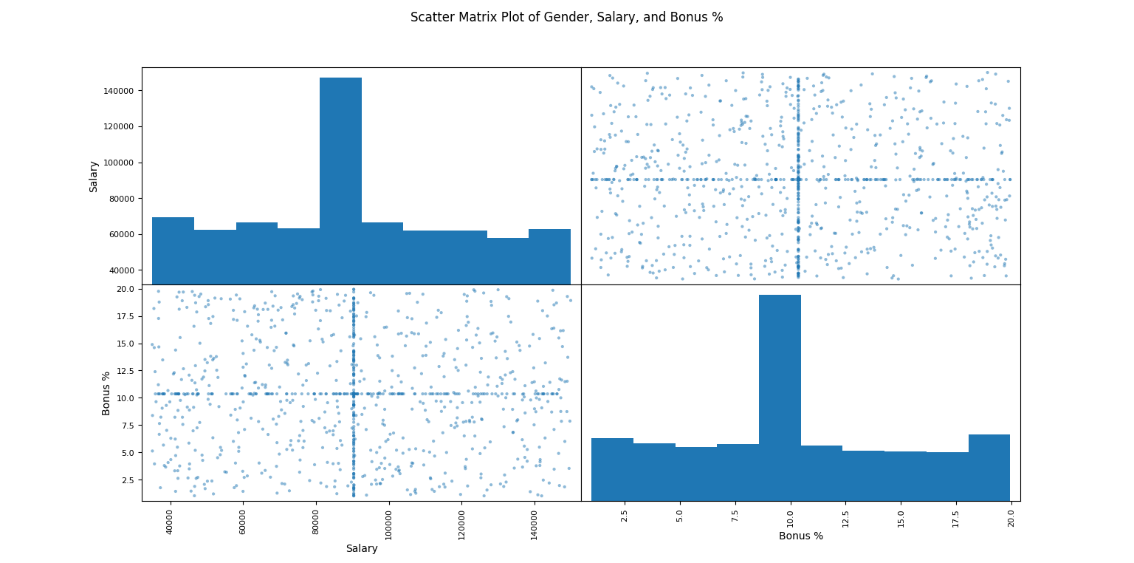
גרף 3 - Pie chart:



הגרף ממחיש את התפלגות העובדים לפי מגדר וצוותים שונים באמצעות עוגות עגולות. ניתן לראות כי בצוותים 111, 324 ו-653, כל העובדים הם גברים, כאשר בצוות 255 כל העובדים הן נשים. מכך ניתן להסיק שקיימת הפרדה מגדרית חדה בין הצוותים השונים, עם צוותים שמורכבים באופן בלעדי מגברים ואחרים שמורכבים רק מנשים. התופעה הזו יכולה להיות תוצאה של תרבות ארגונית או דפוסי גיוס והעסקה מסוימים. היעדר שילוב מגדרי בצוותים עלול להעיד על חוסר גיוון ועל סביבת עבודה לא משולבת, דבר שיכול להוביל לאתגרים בתפקוד הצוות ולהחמצת יתרונות הגיוון. חשוב לבחון את הגורמים לפערים אלו ולפעול לקידום גיוון והשתלבות טובים יותר של נשים וגברים בצוותים השונים.

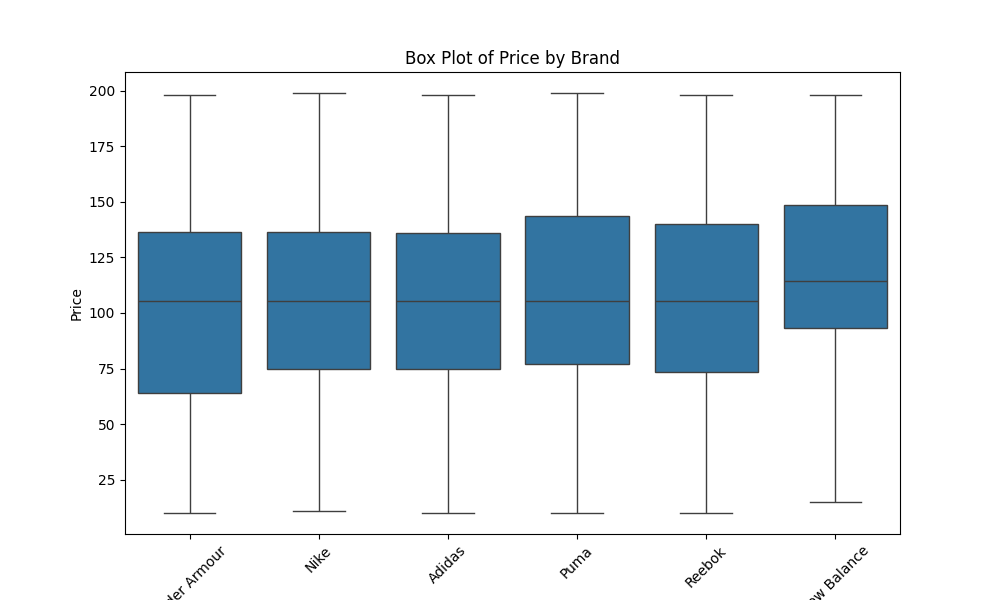
גרף 4 – Area plot :

הגרף מציג את ההתפלגות של מגדר ואחוז הבונוס במדגם נתון. ניתן לראות כי הנתונים המוצגים על ציר המגדר הם בינריים, כלומר הם מסווגים לזכר ונקבה, כאשר זכר מיוצג כ-1 ונקבה כ-0. רוב הנתונים נמצאים בתחתית הגרף ומתייחסים למגדר הנקבה, אך יש גם נוכחות ניכרת של זכר, המיוצג בקו העליון. אחוז הבונוס, המיוצג בצבע אדום, מראה שונות גבוהה מאוד עם תנודות משמעותיות לאורך ציר האינדקס, אך לא נראה כי יש קשר מובהק בין המגדר לאחוז הבונוס. השכיחות הגבוהה של קווים אדומים המשתרעים לכל אורך ציר האינדקס מרמזת כי אחוזי הבונוס מפוזרים בצורה רחבה ואינם תלויים במגדר.

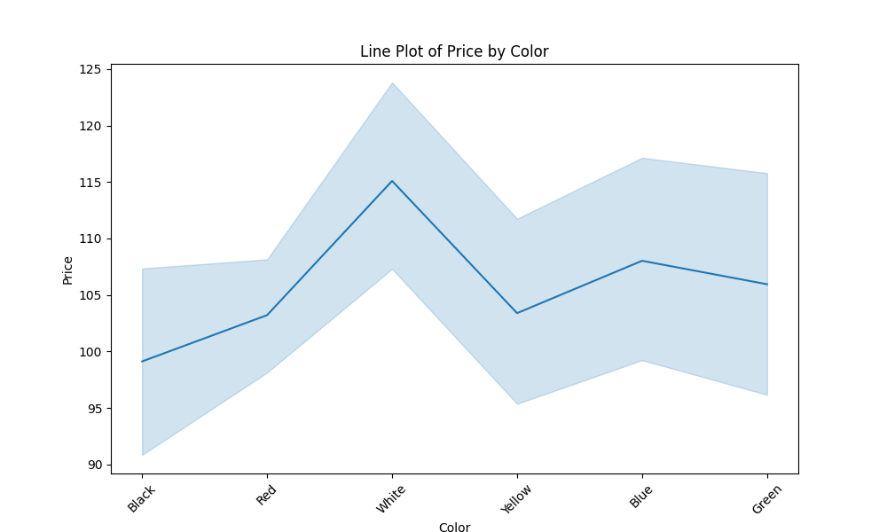
גרף 5 – Scatter matrix plot:

הגרף הנתון הוא מטריצת פיזור המציגה את הקשר בין המשתנים מגדר, שכר ואחוז בונוס. ניתן לראות התפלגות המקרים לפי מגדר בציר האלכסוני העליון – ישנם יותר גברים מנשים בארגון. הגרף התחתון משמאל מראה כי רוב הגברים משתכרים שכר גבוה יותר מנשים. עם זאת, בשני המגדרים קיימת שונות רבה בשכר. התפלגות אחוזי הבונוס, מוצג פיזור רחב יותר לגברים בהשוואה לנשים, ורובם מרוכזים סביב 15%-20% בונוס. סך הכל, התפלגות המקרים מצביעה על פערי שכר וגם פערים באחוזי הבונוס בין גברים לנשים, עם יתרון לגברים. פערי השכר נובעים כנראה גם מההבדלים בתפקידים ורמות הוותק של גברים לעומת נשים. חשוב לבחון את הגורמים לפערים המגדריים ולפעול למיגורם. בד בבד, ראוי לבדוק את מדיניות התגמול ולבחון דרכים לצמצום הפערים באופן הוגן ושוויוני יותר.

גרף 6 – :Box plot



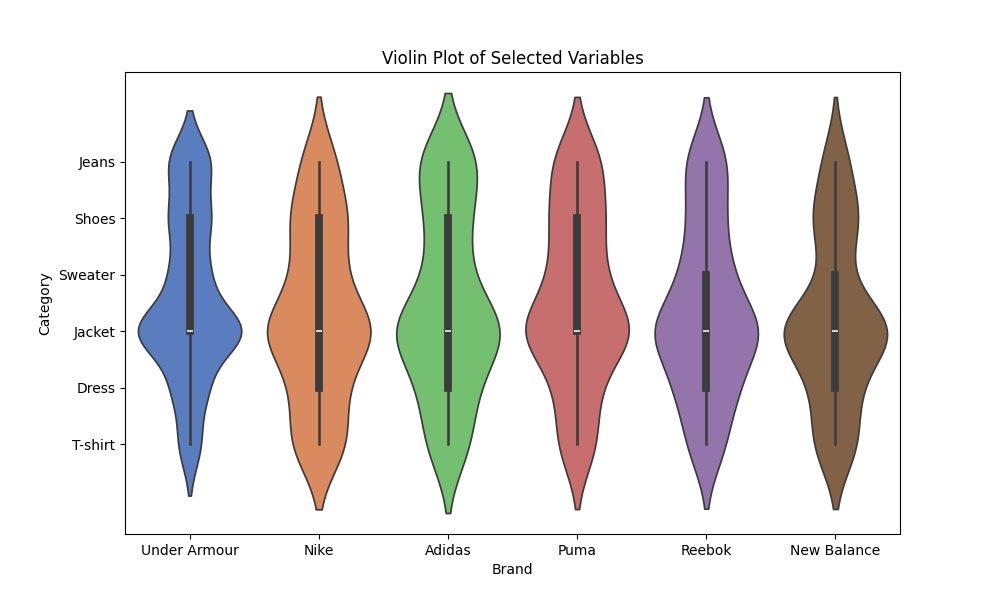
הגרף מציג את התפלגות המחירים לפי מותגי הנעלה שונים. ניתן לראות כי רוב המותגים, כמו Under Armour, Nike, Adidas, Puma, Reebok ו-New Balance, מציגים טווחי מחירים דומים. לכל מותג יש טווח מחירים שמתחיל בערך מ-25 דולר ומגיע עד ל-200 דולר. המותגים Under Armour ו-Puma מציגים את החציון הגבוה ביותר, כ-125 דולר, בעוד שהחציון של המותג New Balance הוא הנמוך ביותר, כ-100 דולר. ניתן לראות גם כי יש פיזור רחב של המחירים בתוך כל מותג, המוצג על ידי אורך ה"שפם" של התיבה, המייצג את ההתפלגות עד לנקודות הקצה. הדבר מצביע על כך שבכל מותג יש מוצרים בטווח רחב של מחירים.

גרף 7 - line plot:

|  |  |  |
| --- | --- | --- |
| |  | | --- | |  | |  |

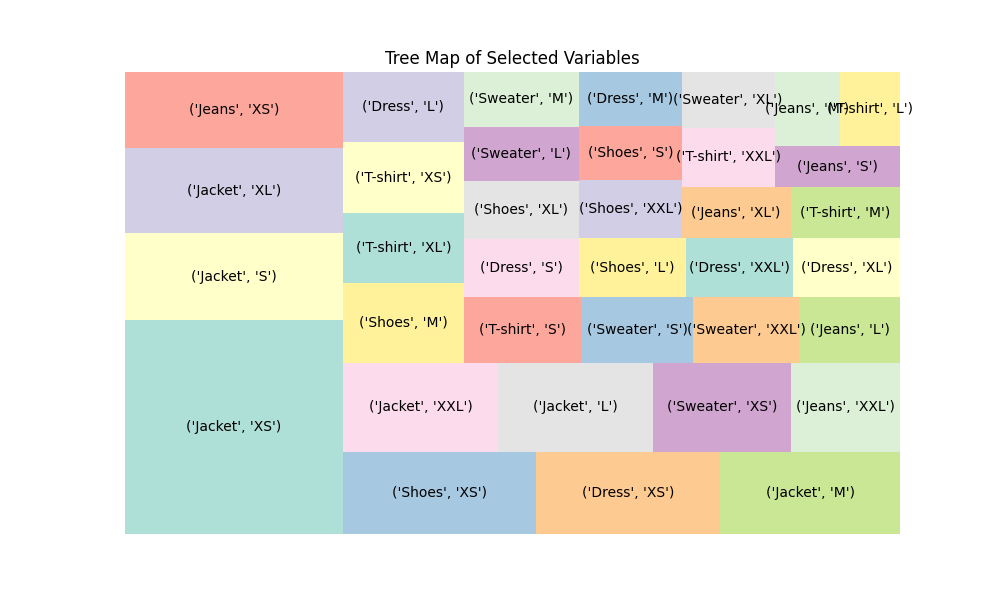
הגרף מציג את המחיר הממוצע לפי צבע של פריט כלשהו. הציר האופקי מציג את הצבעים השונים, ואילו הציר האנכי מייצג את המחיר. ניתן לראות שהמחירים הגבוהים ביותר הם עבור הפריטים בצבעים לבן ואדום, כאשר לבן הוא היקר ביותר. לאחר מכן, המחיר יורד בצורה תלולה עבור הצבעים צהוב, כחול וירוק, אשר עולים באופן דומה. הצבע השחור הוא הזול ביותר מבין כל הצבעים. ההבדלים המשמעותיים במחירים עשויים לנבוע ממגוון גורמים, כגון חומרי גלם שונים הנדרשים לכל צבע, תהליכי ייצור מורכבים יותר לצבעים מסוימים או היצע וביקוש שונים לפריטים בצבעים שונים. חשוב לבחון את הגורמים המדויקים המשפיעים על פערי המחירים כדי לקבל החלטות מושכלות ביחס למחירון המוצרים לפי צבעים.

גרף 8 – Violin plot:



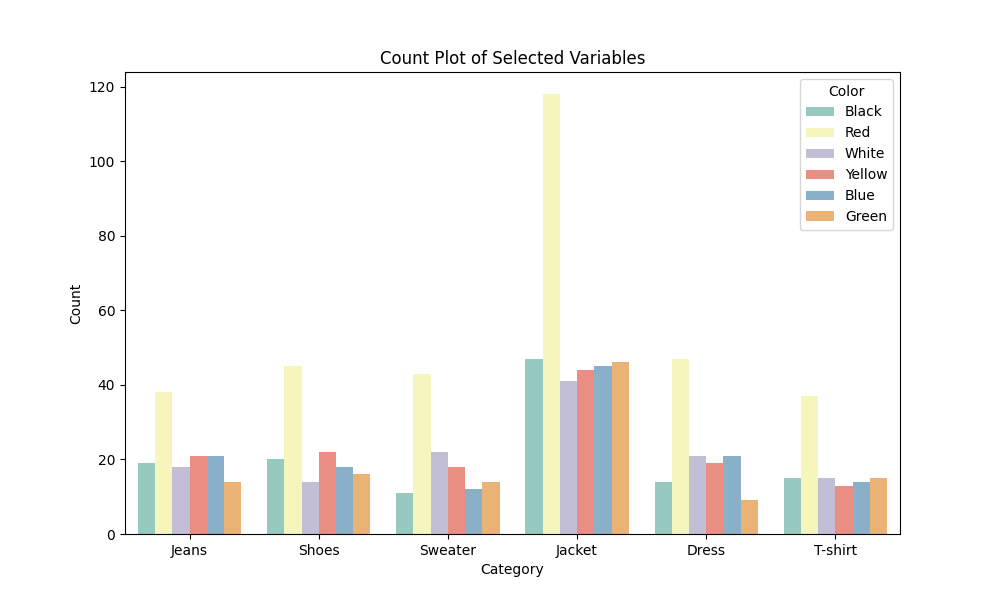
הגרף מציג את התפלגות הקטגוריות השונות של מוצרים עבור כל אחד מהמותגים: , Reebok Puma ,Under Armour, Nike, Adidas, ו-New Balance. ניתן לראות שהמותגים משווקים בעיקר מגוון דומה של קטגוריות מוצרים כגון ג'ינסים, נעליים, סווטשירטים, ג'קטים, שמלות וחולצות. עם זאת, יש הבדלים קלים בין המותגים במוצרים הספציפיים שהם מציעים תחת כל קטגוריה. הגרף מספק תמונה כוללת של הפריסה והמיקוד של כל מותג לפי קטגוריות המוצרים השונות.

גרף 9 – Tree map:



הגרף מציג קשרים בין קטגוריות של פריטי לבוש ובין מידות שונות באופן חזותי וניכר. בהתבסס על המפה, ניתן לראות את הקשרים בין הקטגוריות השונות ובין המידות המבוקשות, וכן לזהות טרנדים ודפוסי התנהגות ברורים. לדוגמה, ניתן לראות כי ישנם גדלים פופולריים יותר בקטגוריות מסוימות, כמו XS ו-S לג'ינסים, L לשמלות ולסוודרים, ו-XL ו-XXL לנעליים וחולצות T. המפה מאפשרת זיהוי שליטה יעילה על מידות המוצרים הנמכרים ביותר ועל רמת הביקוש או ההיצע ביחס למידות ספציפיות בכל קטגוריה. כמו כן, היא מאפשרת זיהוי שליטה במגוון רחב יותר של מידות בקטגוריות מסוימות מאשר באחרות, ואת זהות הטרנדים במדידת מידות מסוימות ביחס לקטגוריות שונות. הבנת דפוסי התנהגות זהים יכולה להוביל להחלטות תפעוליות ושיווקיות מותאמות היטב לצרכי הלקוחות.

גרף 10 - :Count plot



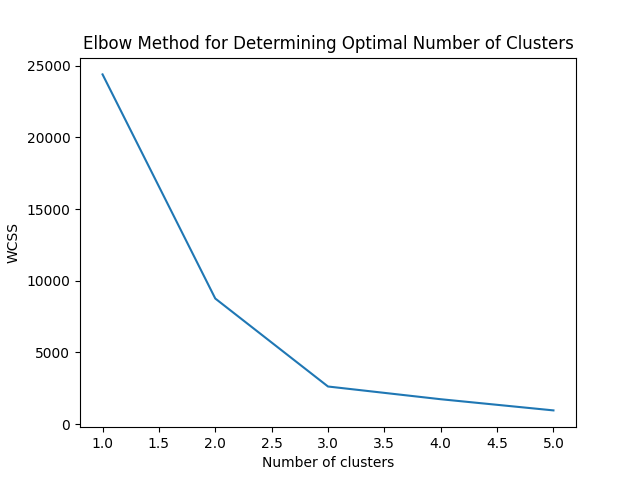
הגרף הזה מציג עמודות מונחות המייצגות את מספר הפריטים בכל קטגוריה, כאשר הצבעים השונים מייצגים את הצבעים של הפריטים בתוך כל קטגוריה. ניתן לראות שהקטגוריה הפופולרית ביותר היא ג'קט, עם כמות גדולה במיוחד של פריטים בצבעים לבן, שחור וצהוב. הקטגוריות הבאות בגודלן הן שמלות, נעליים וחולצות. בקטגוריות אלה, צבעים פופולריים כוללים שחור, לבן, כחול וירוק. הקטגוריות הקטנות יותר הן סווטשירט וג'ינס, עם כמות מוגבלת יותר של פריטים וגוון צבעים צר יותר. הגרף מספק תובנות על היחסים בין קטגוריות המוצרים השונות ועל צבעי המוצרים הפופולריים עבור כל קטגוריה.

**תובנות של 4**

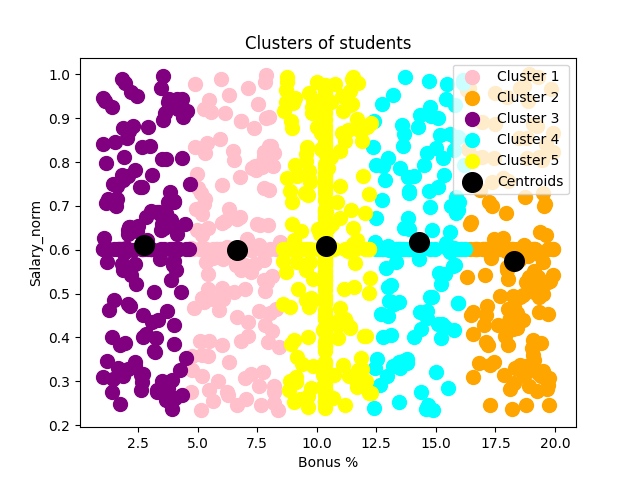
**ניתוח הנתונים**

4.1

Elbow Method- שיטת המרפק



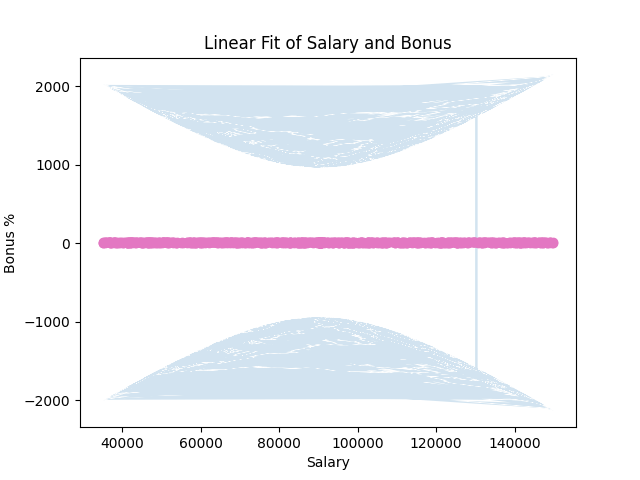
הגרף הזה מציג את שיטת "המרפק" לקביעת מספר האשכולות האופטימלי לצורכי ביצוע אלגוריתם ה k\_means למיקוד. הציר האנכי מייצג את ערך ה WCSS אשר מודד את "הוריאנס" הכוללת בתוך כל אשכול, בעוד הציר האופקי מייצג את מספר האשכולות. מתצורת הגרף ניתן לראות שככל שמספר האשכולות גדל, ערך ה WCSS יורד בצורה חדה, מכיוון שהדגימות מחולקות לקבוצות יותר צפופות. עם זאת, החל ממספר מסוים של אשכולות, הירידה בערך WCSS מתמתנת משמעותית. הנקודה שבה מתרחשת ההתמתנות הזו, המכונה "המרפק", מסמנת את מספר האשכולות האופטימלי, שכן תוספת אשכולות נוספים לא תשפר משמעותית את איכות המיקוד. על פי הגרף, נראה שהמרפק מתרחש סביב 3-4 אשכולות, מה שמצביע על כך שזה ככל הנראה המספר האופטימלי של אשכולות בעבור הנתונים הנוכחיים.

Clustering and Visualization- צבירה והצגת התוצאות

הגרף מציג חמישה אשכולות של סטודנטים המוצגים בצבעים שונים על ידי שילוב של אחוז הבונוס משכרם והשכר הנורמלי שלהם. ניתן לראות כי האשכול הסגול (אשכול 3) מכיל סטודנטים עם שכר נורמלי נמוך יחסית ואחוזי בונוס נמוכים. לעומת זאת, האשכול הכתום (אשכול 2) מכיל סטודנטים עם שכר נורמלי גבוה יחסית ואחוזי בונוס גבוהים. האשכולות הצהוב (5) והתכלת (4) ממוקמים באזור הביניים, עם שכר נורמלי בינוני ואחוזי בונוס מגוונים. האשכול הורוד (1) מראה סטודנטים עם שכר נורמלי נמוך מאוד אך עם אחוזי בונוס גבוהים יחסית. הנקודות השחורות מייצגות את מרכזי האשכולות השונים. חלוקה זו של הסטודנטים לאשכולות על בסיס השכר והבונוס יכולה לספק תובנות חשובות על סגנונות החיים והכנסות של קבוצות שונות בקרב אוכלוסיית הסטודנטים.

4.2

רגרסיה ליניארית



הגרף מציג את הקשר בין משכורת לאחוז הבונוס באמצעות התאמה ליניארית. הנקודות הוורודות מייצגות את הנתונים בפועל, בעוד הקו הכחול מייצג את המודל הליניארי המתאים ביותר. האזור הכחול הבהיר סביב הקו מייצג את טווח השגיאה, המצביע על רמת הוודאות של המודל. התמונה המתקבלת מראה שאין קשר ליניארי משמעותי בין המשכורת לאחוז הבונוס - הקו כמעט אופקי, מה שמעיד על כך שעלייה במשכורת אינה מובילה בהכרח לעלייה באחוז הבונוס. יתר על כן, אחוזי הבונוס מרוכזים סביב אותו טווח (קרוב לאפס) ללא קשר לגובה המשכורת, מה שמחזק את המסקנה שאחוז הבונוס נקבע על פי קריטריונים אחרים ולא על בסיס המשכורת.