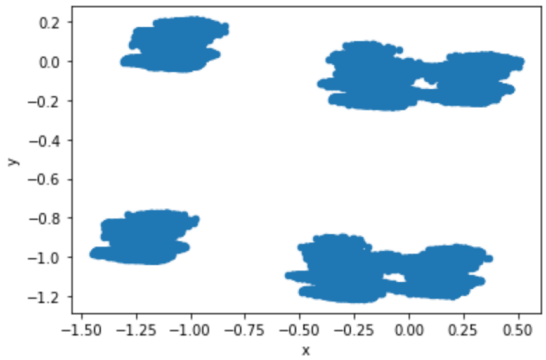
Vizual Analysis



אנחנו רואות 6 קלאסטרים בבירור.
3 קלאסטרים ממורכזים סביב ערך של 0.1 בציר ה-y ו-3 ממורכזים סביב ערך של 1.1− בציר ה-y. בנוסף, נשים לב שישנם שני זוגות קלאסטרים (אלה הקרובים יותר ל-0 בציר ה-o) אשר קרובים זה לזה באופן יחסי לשני הקלאסטרים הנותרים.

Cluster's Viewing Analysis

- לכל קלאסטר, ערך ה-diff_rank מוגדר ע"י ההפרש בין יחס הצפיות בתחנה מסויימת מתוך כלל הצפיות בקלאסטר, לבין יחס הצפיות בתחנה זו מתוך כלל הצפיות במסד המלא (כל הקלאסטרים).

יהיו קלאסטר c ותחנה s. אם ה-diff_rank של c ברשומה של s חיובי, זה אומר ש-households שנמצאות באותו קלאסטר c, שהן בעצם רשומות שדומות אחת לשנייה במאפיינים של ה-household שלהן, נוטות לבחור לצפות בתחנה s יותר מאשר כלל ה-households במסד באופן יחסי לנתוני הצפייה שלהן. כלומר אם ניקח ממוצע של מספר הצפיות של כל קלאסטר בתחנה s, אז כמות הצפיות של קלאסטר c בתחנה s תהיה גבוהה מהממוצע.

לעומת זאת, אם ה-diff_rank של c ברשומה של s שלילי, זה אומר ש-households שנמצאות באותו קלאסטר c, שהן בעצם רשומות שדומות אחת לשנייה במאפיינים של ה-household שלהן, נוטות לבחור לצפות בתחנה s פחות מאשר כלל ה-households במסד באופן יחסי לנתוני הצפייה שלהן. כלומר אם ניקח ממוצע של מספר הצפיות של כל קלאסטר בתחנה s, אז כמות הצפיות של קלאסטר c בתחנה s תהיה נמוכה מהממוצע.

- מהתבוננות בתוצאות חישוב ה-diff_rank של התחנות השונות, עבור כל subset נוכל להסיק את המסקנות הבאות:

1. לא מצאנו קשר בין subsets מאותו סוג בין קלאסטרים שונים, מפני שתחנות שמופיעות כחלק מ-7 התחנות בעלות ה-diff_rank הגבוה ביותר ב-subset של קלאסטר כלשהו לא בהכרח יופיעו כחלק מ-7 התחנות בעלות ה-diff_rank הגבוה ביותר ב-subset מאותו סוג של קלאסטר אחר.
2. מצאנו קשר בין subsets שונים באותו קלאסטר. התחנות שמופיעות כחלק מ-7 התחנות בעלות ה-diff_rank הגבוה ביותר ב-full data של קלאסטר מסויים יופיעו בסיכוי גבוה כחלק מ-7 התחנות

בעלות ה-diff_rank הגבוה ביותר ב-subset של ה-3 וה-17 של אותו הקלאסטר מפני שאלו דגימות מאותו הדאטא מאותו הקלאסטר. נשים לב כי קיים הבדל בין ה-subset של ה-3 וה-17 מפני שה-subset של ה-17 מייצג פחות רשומות מהדאטא ולכן הסיכוי ש-7 התחנות שלו יהיו תחנות בעלות diff_rank גבוה בדאטא המלא של אותו קלאסטר הוא נמוך יותר בהשוואה ל-subset של 3, שהוא בעל יותר רשומות ולכן מאפיין חלק גדול יותר של הדאטא בקלאסטר. כלומר, הסיכוי שיכיל מידע שמכליל את הדאטא לדאטא המלא בקלאסטר נמוך יותר (כי מייצג חלק קטן יותר של הדאטא והסיכוי שייצג את רובו קטן יותר).

Dynamic Data Analysis - Streaming

אנו שמות לב כי עבור כל קלאסטר, ככל שמספר ה-batch-ים גדל, כך גם ערכי ה-diff_rank נראים כמתכנסים לערכי ה-diff_rank הסטטיים שמצאנו בחלק הקודם. לדעתנו תופעה זו מתרחשת כיוון שמספר הרשומות שעליהן מתבצעים חישובי ה-diff_rank שלנו גדל, כי בכל פעם שאנחנו מקבלות מידע חדש ב-batch חדש, אנחנו מנתחות את המידע הזה לפי צרכינו ומאחדות אותו עם המידע שאספנו וחישבנו עד עכשיו מה-batch-ים הקודמים, ועושות זאת עבור כל קלאסטר. מצב זה יוצר ייצוג רחב יותר של הדאטא הכולל שלנו עם התווספות ה-batch-ים, ולכן בהסתברות גבוהה יותר גם מכליל אותו טוב יותר. כלומר, החישובים נעשים מדוייקים יותר, ועם עליית מספר ה-batches הם מתכנסים לערכים שיצאו לנו עבור כל קלאסטר עבור הדאטא הסטטי המלא.