כריית מידע וייצוג מידע* – פרויקט חלק א'

גילעד סבוראי ולירי בנזינו

* כל הגרפים והנתונים הסטטיסטיים נמצאים במחברת.

1) הצגת המידע – סוגי השדות והערכים

קיבלנו מסד נתונים טבלאי המכיל תכונות (מאפיינים) של חברת הזנק שונות ברחבי ארה"ב. מטרתנו מתוך נתונים אלו היא לחזות האם חברת ההזנק הבאה תצליח או תיכשל. בחלק זה של הפרויקט ננתח את מסד הנתונים בצורה כתובה ובצורה וויזואלית בעזרת תרשימים שונים, כדי להבין אותו לעומק, בנוסף להסקת קשרים ראשוניים בין הנתונים שקיבלנו.

נתחיל בטעינת מסד הנתונים והדפסתו בעזרת ספריית `pandas`. בעזרת מתודת `info` קיבלנו טבלה המתארת את השדות השונים של מסד הנתונים (מספר סידורי, קוד מדינה, מיקום גיאוגרפי, מיקוד, וכדומה). לכל שדה מתוארת כמות הפעמים בה הוא אינו ריק וסוג המידע (הממוחשב) ששמור בו.

לאחר מכן הצגנו טעימה ממסד הנתונים על ידי `head`, וכך קיבלנו תצוגה מהירה של חמש חברות ההזנק הראשונות ממסד הנתונים.

לבסוף על ידי `describe` הצגנו מידע שטחי נוסף לכל תכונה – כמו מספר הערכים הייחודים שאותה תכונה מקבלת, השכיחות המרבית שלה, הממוצע, השונות, הערך המזערי, הרבעון הראשון, שני (חציון) ושלישי שלה, והערך המרבי.

יש לציין שעבור ערכים מספריים (אורדינאליים/תחומים) קיבלנו הכי הרבה מידע בפקודה הזו.

2) הצגת המידע הסטטיסטי

בשלב זה הצגנו את ההתפלגויות של התכונות השונות ב-dataset שלנו, וחישבנו נתונים סטטיסטיים על כל אחד מהם – ממוצע, skewness, חציון, סטיית תקן וערך שכיח, באופן מעמיק.

תחילה, הצגנו את ההתפלגות של התכונות הנומינליות – state_code, category ו-Target, על ידי היסטוגרמות. בנוסף, על מנת לחשב את המידע הסטטיסטי של כל אחת מהתכונות קודדנו את הערכים על ידי שימוש במתודה map, בצורה זו יכלנו להתייחס לערכים הנומינליים כנומריים (המייצגים חלוקה לקטגוריות).

עבור התכונה city ראינו שישנם הרבה מאוד ערכים, ולכן החלטנו רק לחשב עבורה את הנתונים הסטטיסטיים ולא להציג עבורה היסטוגרמה.

לאחר מכן, הצגנו את ההתפלגויות של כל התכונות הנומריות ב-dataset שלנו, וגם כאן הצגנו את הנתונים הסטטיסטיים. את הערכים הבינאריים הצגנו על ידי דיאגרמות עוגה ואת הערכים האחרים הצגנו על ידי היסטוגרמות.

מתוך כל הגרפים למדנו על המיקומים הגיאוגרפיים בהן החברות נמצאות, וראינו כי קצת יותר ממחצית מחברות ההזנק ממוקמות ב-CA (קליפורניה), וגם רק פחות מרבע מהחברות נמצאות במדינות שאינן CA, NY, MA, TX.

בהמשך לכך, ראינו שסטיית התקן של הערים מאוד גדולה, ושישנם הרבה מאוד ערכים שונים, לכן הנחנו שאולי העמודה הזו תהיה פחות שימושית בשבילנו להמשך.

בנוסף לכך, ראינו כי רוב החברות הצליחו לבסוף (acquired), ואף יותר מ-80% מהן הגיעו ל-500 החברות המצליחות ביותר.

ראינו גם שקצת יותר מחצי מחברות ההזנק הגיעו לסבב הראשון (round A) מתוך סבבי המימון (ה-funding_rounds).

יתרה על כך, ראינו שרוב חברות ההזנק קיבלו את המימון הראשון שלהן כשהיו באוויר פחות מ-5 שנים.

3) הצגת הקורלציה

לאחר שהצגנו את התפלגות התכונות השונות, נחפש ונציג את רמות התאימות ביניהן על ידי חישוב קורלציות על כל התכונות המספריות בעזרת `corr(numeric_only=True)` וקיבלנו מטריצה ריבועית סימטרית כאשר התא ה-ij מתאר את רמת התאימות בין המשתנה ה-i לבין ה-j – על סקאלה של 1- עד 1. ערכים הקרובים למינוס 1 מתארים רמת תאימות (שלילית) גבוהה, 1 מתאר רמת תאימות חיובית גבוהה וערכים הקרובים לאפס מתארים היעדר תאימות. על מנת להבחין בקלות רבה יותר בערכי המטריצה, הצגנו אותם במפת חום על ידי `heatmap`. הצבעים הכחולים והאדומים החזקים מסמלים רמת תאימות (שלילית/חיובית) גבוהה.

ראינו קורלציה חיובית גבוהה בין הערכים הבאים –

1. גיל ההגעה לאבן הדרך הראשונה כנגד גיל ההגעה לאבן הדרך האחרונה – זה הגיוני כי ככל שמגיעים לאבן דרך הראשונה מאוחר יותר, כך גם נגיע לאבן דרך האחרונה מאוחר יותר, ואם נצליח להגיע להישגים ראשונים מהר, ככל הנראה שנצליח להישאר במגמה הזו לאחר מכן.
2. גיל המימון הראשון כנגד גיל המימון האחרון – ככל שגיל המימון הראשון גדל כך גם החברה תקבל את המימון האחרון בגיל מאוחר יותר. דבר זה ככל הנראה הגיוני, כי חברות הזנק לרוב לא נשארות פתוחות לאורך זמן ארוך, לכן ההפרש בין זמני המימון האחרון לראשון, ככל הנראה די קטן, ולכן הם יחסית יהיו קרובים. אך אם הם מקבלים את המימון הראשון קרוב לזמן הקמת החברה, המשקיעים ככל הנראה יתנו לחברה יותר זמן להגיע להישגים לפני שהם כבר לא יממנו אותה/החברה תיסגר.
3. גיל המימון הראשון כנגד גיל ההגעה לאבן הדרך הראשונה – כאשר חברה מקבלת מימון יהיה לה יותר קל להגיע לאבני דרך, כי לאחר המימון יהיו לה יותר משאבים, ולכן אנחנו רואים את הקורלציה בניהם.
4. כמות הקשרים שיש לחברה כנגד כמות אבני הדרך שהיא הגיעה אליה – אנחנו יכולים לראות שככל שיש יותר קשרים חיצוניים לחברה, בתוחלת החברה תגיע ליותר אבני דרך, וזה ככל הנראה בגלל העזרה מהקשרים.

מתוך נתונים אלו נגיע למסקנות על הקשרים בניהם ואת ההשפעות שלהם על ה-Target בסעיף הבא.

4) הצגת תובנות, יחסים בין features שונים והשפעתם על ה-target

מתוך הסעיף הקודם למדנו על כל מיני יחסים וקשרים ב-dataset שלנו.

ראשית, ראינו שישנה קורלציה חיובית גבוהה בין הערכים first_milestone_age ו-last_milestone_age (0.78), בין הערכים first_funding_age ו-last_funding_age (0.77), בין הערכים first_funding_age ו-first_milestone_age (0.59), ובין הערכים milestones ו-connections (0.53), והצגנו את הגרפים שלהם.

ראשית, עבור זוגות של תכונות עם קורלציה חיובית גבוהה הצגנו את הגרפים שמתארים את הקשר בניהם, ומהם למדנו –

1. הזוג first_milestone_age ו-last_milestone_age (corr=0.78) – ראינו מעין קשר לינארי, סוג של קו ישר. כלומר ראינו, שככל שה-milestone הראשון הגיע מאוחר יותר כך גם ה-milestone האחרון. דבר זה הגיוני כי ככל שהחברה מצליחה להגיע להישגים מהר יותר, היא ככל הנראה תמשיך במגמה הזו, ואנחנו רואים את זה אצלנו – ככל שהחברה מגיעה לאבן דרך ראשונה מהר יותר, היא מגיעה גם לאחרונה מהר יותר.
2. הזוג first_funding_age ו-last_funding_age (corr=0.77) – באופן דומה לזוג הקודם גם כאן מעין קשר לינארי, כלומר שככל שהמימון הראשון הגיע מאוחר יותר כך גם המימון האחרון.
3. הזוג first_funding_age ו-first_milestone_age (corr=0.59) – גם כאן ראינו קשר לינארי, אבל קצת פחות מובהק. כלומר, עדיין ככל שהמימון הראשון הגיע מאוחר יותר כך גם החברה הגיעה לאבן דרך מאוחר יותר, אבל בקשר "כמעט לינארי". זה הגיוני, כי לאחר שחברה מקבלת מימון, יש לה יותר אמצעים, ולכן תוכל להגיע לאבני דרך מהר יותר.
4. הזוג connections ו-milestones (corr=0.53) – מתוך הגרף אנחנו רואים שלרוב כאשר יש יותר קשרים לחברת ההזנק מגיעות ליותר אבני דרך, אבל כבר לא קיבלנו קשר לינארי מובהק כמו שראינו בזוגות הקודמים. כלומר, ישנן חברות שהגיעו להרבה אבני דרך, מבלי הרבה קשרים, ולהפך – חברות אשר יש להן הרבה קשרים וקצת אבני דרך, אבל סה"כ רואים מגמה הגיונית שככל שישנם יותר קשרים מגיעים ליותר אבני דרך.

מעבר לזוגות האלו, רצינו לבדוק את הקשר בין המדינה ממנה הגיעה החברה לתכונה שאומרת האם החברה נמצאת ב-500 החברות המצליחות ביותר, ואת ההשפעה של מדינה המוצא על התכונה target (הצלחה/כשלון החברה).

תחילה, ניזכר שראינו בסעיף 2 שפחות מרבע מהחברות נמצאות במדינות שאינן CA, NY, MA, TX החלטנו ליצור עמודה שלנו שתגיד באיזו מדינה החברה נמצאת, ואם היא לא באחת מארבע המדינות שציינו לעיל, כתבנו שהיא נמצאת ב-other-state. עשינו זאת בעזרת העמודות הבינאריות שיש לנו שאומרות מאיפה המדינה הגיעה.

לאחר מכן, יצרנו count-plot על המדינות והשוונו בין הכמויות שכן נמצאות ב-500 המצליחות ביותר, ובאופן דומה על ה-target (האם נרכשו בסוף או שנסגרו).

ראינו שרוב החברות בכל מדינה אכן הגיעו ל-500 החברות המצליחות ביותר, אבל לחברות מארבע המדינות שציינו לעיל – CA, NY, MA, TX יש סיכוי גבוה יותר להצליח מאשר לחברות שאינן מהמדינות האלו.

באופן דומה, חברות שנמצאות במדינות לעיל נרכשו בסיכוי גבוה יותר מאשר במדינות אחרות, שבהן הרוב המדינות בכלל נסגרו (על אף שהרוב הוא אכן רוב קטן), למרות המדינות שציינו שבהן רוב החברות (גם אם הרוב הוא קטן) אכן נרכשו ולא נסגרו.

כעת רצינו לראות את ההשפעה של להיות ב-500 המצליחות ביותר על ערך ה-target, ולכן יצרנו גרף דומה לקודם, וראינו כי לחברות שהגיעו ל-500 החברות המצליחות יש סיכוי גבוה בהרבה להירכש לעומת חברות שלא, ואף ראינו שרוב החברות שלא הגיעו ל-500 החברות המצליחות ביותר, נסגרו לבסוף.

לבסוף, בהמשך לקורלציה שראינו בסעיף הקודם (ובתחילת הסעיף הזה), רצינו לחקור את הקשר בין גיל החברה בזמן המימון הראשון לבין גיל החברה בהגעתה אל אבן הדרך הראשונה, ואת ההשפעה שלהם על ה-target. לכן ייצרנו גרף המתאר את הקשר בין גיל המימון הראשון לגיל אבן הדרך הראשונה, והפרדנו בין הרשומות של החברות שנרכשו (בירוק) לבין הרשמות של החברות שנסגרו (באדום). ראינו את אותו קשר לינארי חיובי שראינו בתחילת הסעיף הזה, ובנוסף לכך ראינו שעבור חברות שלבסוף נסגרו, המימון הראשון (ובהתאם גם אבן הדרך הראשונה) היה בזמן מאוחר יחסית לעומת חברות שנרכשו שבהן המימון הראשוני היה בגיל צעיר של החברה (ובהתאם גם אבן הדרך הראשונה).

בנוסף, ראינו שישנה גם קורלציה יחסית גבוהה בין המימון האחרון לאבן הדרך האחרונה (0.64), ולכן רצינו לבדוק את אותו הדבר כמו קודם, רק על הזוג הזה. ראינו דבר יחסית דומה, גם כאן יש לנו מעין קשר לינארי חיובי בין המימון האחרון לאבן הדרך האחרונה. יתרה על כן, ראינו שישנה הצטברות של נקודות סביב האפס עבור חברות שנסגרו, כלומר הרבה חברות שנסגרו קיבלו את המימון האחרון שלהן כאשר הן היו מאוד צעירות. דבר זה כנראה גורר את העובדה שהן נסגרו בגיל צעיר. בהמשך לכך שהמימון האחרון שלהן היה בגיל צעיר, בגלל הקשר הלינארי אנחנו רואים שגם אבן הדרך האחרונה שלהן הייתה בגיל צעיר.

5) ניקוי המידע – השלמת ערכים חסרים, תיקון חוסר עקביות

נתונים חסרים

שמנו לב שישנם מספר ערכים חסרים בנתונים, בתכונות של גיל ההגעה אל אבן הדרך הראשונה/האחרונה, והמדינה. את המידע החסר על המדינה השלמנו על פי ה-state code שנמצאת בנתונים. בעזרת הערך החסר הזה הבנו שיש חריג בשדות הבינאריים שמתארים את המדינה בה נמצאת החברה, כי את העמודה של המדינה country אנחנו יוצרים על פי השדות הבינאריים הללו – תיקנו את זה בהמשך.

את הגילים החסרים השלמנו בעזרת התאמה של רגרסיה לינארית בהתאם לנתונים ובהתאם לקורלציות שמצאנו בסעיף הקודם:

מתוך התאימות בין גיל המימון הראשון לגיל ההגעה אל אבן הדרך הראשונה, ביצענו התאמה לינארית והשלמנו את הנתון של גיל ההגעה לאבן הדרך הראשונה ברשומות החסרות.

תיקון דומה התבצע עבור גיל המימון האחרון כנגד גיל ההגעה אל אבן הדרך האחרונה.

חוסר עקביות

התחלנו בלבדוק חוסר עקביות בעזרת המידע הבינארי המהווה יתירות (המופיע בתכונות ה-category, state ביחד עם in/is בליווי של הקטגוריה/הארץ). יצרנו פונקציה אשר בוחנת את העמודות הבינאריות האלו ביחד עם העמודות הטקסטואליות ומחפשת אי תאימויות בין השניים: `check_binary_inconsistency`.

מהרצה של הפונקציה זיהינו בעיות בתכונת ה-state ותיקנו אותן. לא נמצאו בעיות בתכונת ה-category.

בשלב זה עברנו ידנית על הערים השונות של החברות וחיפשנו ערכים שלא הוקלדו נכון/התייחסו לאותה העיר ותיקנו ערכים אלו. בנוסף לכך עברנו על עמודות הגילים/תאריכים וחישבנו האם הפרשי התאריכים אכן מתאימים לגיל הנמצא בתכונה המתאימה (למשל תאריך המימון הראשון כנגד גיל המימון הראשון). לבסוף ווידאנו שהערכים הבינאריים המסמלים את מימון החברה בסבבים השונים לא סותר את כמות סבבי המימון בסך הכל.

מציאת ותיקון חריגים

חיפשנו ערכים אשר חורגים בעמודת הבאות מטווח של 3 סטיות תקן מסביב לממוצע: סך המימון, מספר הקישורים של החברה, מספר סבבי המימון, מספר אבני דרך וגודל קבוצה ממוצע.

אחרי בדיקה מצאנו כל מני עמודות עם ערכים חריגים, אבל כן רצינו להתחשב בהם, לכן שמרנו עליהם ועשינו על פיהם דיסקריטיזציה.

לבסוף, הצגנו את חברות ההזנק השונות על מפת העולם, בעזרת עמודות המיקום שלהן. שמנו לב שרובן אכן ממוקמות בארה"ב, אך יש מספר מצומצם שלהן שמופיעות באירופה!

6) דיסקריטיזציה, הפחתת מידע ונרמול המידע

דיסקריטיזציה

בסעיף הקודם ראינו שעבור התכונות – connections, funding_rounds, milestones, avg_group_size יש ערכים חריגים שכן נרצה להתייחס אליהם, ובנוסף אם נהפוך אותם לדיסקרטיים נוכל ללמוד עליהם יותר מאשר הערכים הרגילים, כיוון שלכל אחד מהם יש משמעות לטווחים מסוימים. לדוגמה עבור גודל קבוצה ממוצע, נוכל להגדיר קבוצות קטנות, בינוניות, גדולות (וערכים חריגים), ובעזרת הקבוצות האלו אנחנו יכולים ללמוד יותר על ההשפעה של התכונה הזו מאשר הערכים הרציפים, כי כך נוכל ללמוד האם חברות עם קבוצות קטנות, מצליחות יותר או לא. באופן דומה עבור התכונות האחרות, האם הרבה קשרים משפיעים על ההצלחה של החברה וכו'.

על מנת לעשות דיסקריטיזציה לתכונות אלו, השתמשנו בשיטה שראינו בתרגול הנקראת reasoning partition, וזה כי רצינו להגדיר את הטווחים על פי ההיגיון וההתאמה לכל תכונה בנפרד. לכן כל הטווחים נבחרו כך שה-lower bound של ה-bin האחרון נקבע על פי הערך שממנו מתחילים הערכים החריגים, ושאר הטווחים נבחרו על פי בחירה הגיונית בהתאם לתכונה ולהתפלגויות שראינו בסעיף 2.

הטווחים שבחרנו הם –

1. connections –

```
connections = [0, 5, 10, 20, 30, data_frame['connections'].max()]
discretize(data_frame, 'connections', connections)
✓ 0.6s

0    376
1    193
2    121
3     35
4     13
```

2. funding_rounds –

```
funding_rounds = [0, 1, 2, 3, 4, 6, data_frame['funding_rounds'].max()]
discretize(data_frame, 'funding_rounds', funding_rounds)
✓ 0.5s

0    253
1    227
2    133
3     69
4     43
5     13
```

3. milestones –

```
milestones = [0, 1, 2, 3, 4, 5, data_frame['milestones'].max()]
discretize(data_frame, 'milestones', milestones)
✓ 0.6s

0    325
1    198
2    136
3     52
4     21
5      6
```

4. avg_group_size –

```
avg_group_size = [0, 1, 3, 5, 8, data_frame['avg_group_size'].max()]
discretize(data_frame, 'avg_group_size', avg_group_size)
✓ 0.6s

1    338
0    172
2    170
3     43
4     15
```

בהמשך לתכונות האלו, ולסעיפים קודמים, החלטנו לעשות דיסקריטיזציה גם לתכונה state_code כי ראינו שהמדינות שמופיעות הכי הרבה CA, NY, MA, TX, ולכן השתמשנו בעמודה country שהגדרנו על מנת לעשות את הדיסקריטיזציה – קידדנו את הערכים בה בעזרת המתודה map וקיבלנו למעשה ערכים דיסקרטיים עבור כל מדינה, בצורה הרבה –

CA =	0	376
other-state =	2	171
NY =	3	85
MA =	1	72
TX =	4	34

ראינו מגמה דומה גם בתכונה של ה-category, ולכן השתמשנו בעמודות הבינאריות האומרות איזו קטגוריה כל חברה היא, כיוון שהן מצביעות על הקטגוריות הנפוצות ביותר לעומת השאר, ויצרנו דיסקריטיזציה גם לקטגוריות –

other-category =	0	229
software =	3	129
web =	2	117
mobile =	5	64
enterprise =	1	59
advertising =	7	50
games-video =	4	40
ecommerce =	9	29
biotech =	6	19
consulting =	8	2

הורדת עמודות (הפחתת מידע)

בשלב הזה החלטנו להוריד את העמודות הבאות –

'id', 'state_code', 'latitude', 'longitude', 'zip_code', 'city', 'name', 'foundation_date', 'first_funding_date', 'connections', 'funding_rounds', 'milestones', 'category', 'avg_group_size', 'Target', 'state_code_codes', 'category_codes', 'city_codes', 'country', 'in_Top500_str', 'category_1'

ואת כל העמודות הבינאריות.

את העמודות id, latitude, longitude, zip_code, city, name החלטנו להוריד, כי הן חח"ע, כלומר שונות לכל רשומה ורשומה, ולא נותנות הרבה מידע מעבר לעמודות האחרות שיש לנו.

את העמודות state_code, connections, funding_rounds, milestones, category, avg_group_size כי עשינו עליהן דיסקריטיזציה ושמרנו את התוצאה של הדיסקריטיזציה בעמודה נפרדת (לכן השארנו רק את העמודות אחרי הדיסקריטיזציה).

את העמודה Target הורדנו כי בתחילת הקוד יצרנו עמודה מקבילה שהיא קידוד של הערכים של עמודה זו, ואנחנו מעדיפים את הערך הנומרי.

את העמודות first_funding_date , last_funding_date הורדנו, כי בשלב הקודם וידאנו את ערכי הגילים בהן היה את המימון הראשון והאחרון על פי התאריכים ותאריך היצירה של חברת ההזנק, ולכן הם כבר לא רלוונטיים אלינו, ותמיד נוכל לחשבם על פי עמודות הגיל המתאימות.

את העמודות הבינאריות גם כבר אינן רלוונטיות אלינו כי עשינו דיסקריטיזציה למדינה וקטגוריה של רשומה על פי העמודות האלו, כלומר יש לנו עמודות שמייצגות את העמודות האלו, וניתן לחשב את העמודות האלו מתוכן, ולכן יכלנו להוריד אותן.

את שאר העמודות הורדנו כי הן עמודות שאנחנו יצרנו כדי לעזור לחשב ערכים סטטיסטיים, לייצר גרפים, וכדי לעשות את הדיסקריטיזציה, ועל כן הן לא רלוונטיות בשבילנו.

סה"כ העמודות איתן נשארנו הן –

```
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   foundation_date      738 non-null    datetime64[ns]
 1   first_funding_age    738 non-null    float64
 2   last_funding_age     738 non-null    float64
 3   first_milestone_age  738 non-null    float64
 4   last_milestone_age   738 non-null    float64
 5   total_funding        738 non-null    int64
 6   roundA               738 non-null    int64
 7   roundB               738 non-null    int64
 8   roundC               738 non-null    int64
 9   roundD               738 non-null    int64
 10  in_Top500            738 non-null    int64
 11  target_codes         738 non-null    int64
 12  connections_bins     738 non-null    category
 13  funding_rounds_bins  738 non-null    category
 14  milestones_bins      738 non-null    category
 15  avg_group_size_bins  738 non-null    category
 16  state_code_bins      738 non-null    int64
 17  category_bins        738 non-null    int64
```

נרמול העמודות

לאחר כל הפעולות האלו נותר לנו רק לעשות נרמול. את הנרמול אנחנו עושים על מנת להגדיר טווח זהה לשדות שונים, כדי שלא יהיה שדה שיקבל ערך משמעותי יותר משאר שדות אחרים, שאמורים להיות להם אותה השפעה.

לכן את הנרמול נעשה לכל העמודות הנומריות שלא עשינו להן דיסקריטיזציה, ושאינן בינאריות (וזאת כי הן כבר מנורמלות בין 0 ל-1), וגם לתאריך ההקמה של החברה. ביצענו עליהן את הנרמול, כי ראינו שעבור העמודות האלו לא היו ערכים חריגים בצורה קיצונית (כמו השדות שעליהם עשינו דיסקריטיזציה), וכן היה חשוב לנו לשמור על הערכים הרציפים שלהם. בנוסף, על מנת לעשות את הנרמול לתאריכים, הפכנו אותם ל-int שמיוצג על ידי כמות ה-ns מאז תאריך מסוים.

סה"כ לאחר ההמרה ל-int, העמודות שעליהן היה נרמול הן –

#	Column	Non-Null Count	Dtype
0	foundation_date	738 non-null	int64
1	first_funding_age	738 non-null	float64
2	last_funding_age	738 non-null	float64
3	first_milestone_age	738 non-null	float64
4	last_milestone_age	738 non-null	float64
5	total_funding	738 non-null	int64
6	roundA	738 non-null	int64
7	roundB	738 non-null	int64
8	roundC	738 non-null	int64
9	roundD	738 non-null	int64
10	in_Top500	738 non-null	int64
11	target_codes	738 non-null	int64
12	connections_bins	738 non-null	category
13	funding_rounds_bins	738 non-null	category
14	milestones_bins	738 non-null	category
15	avg_group_size_bins	738 non-null	category
16	state_code_bins	738 non-null	int64
17	category_bins	738 non-null	int64

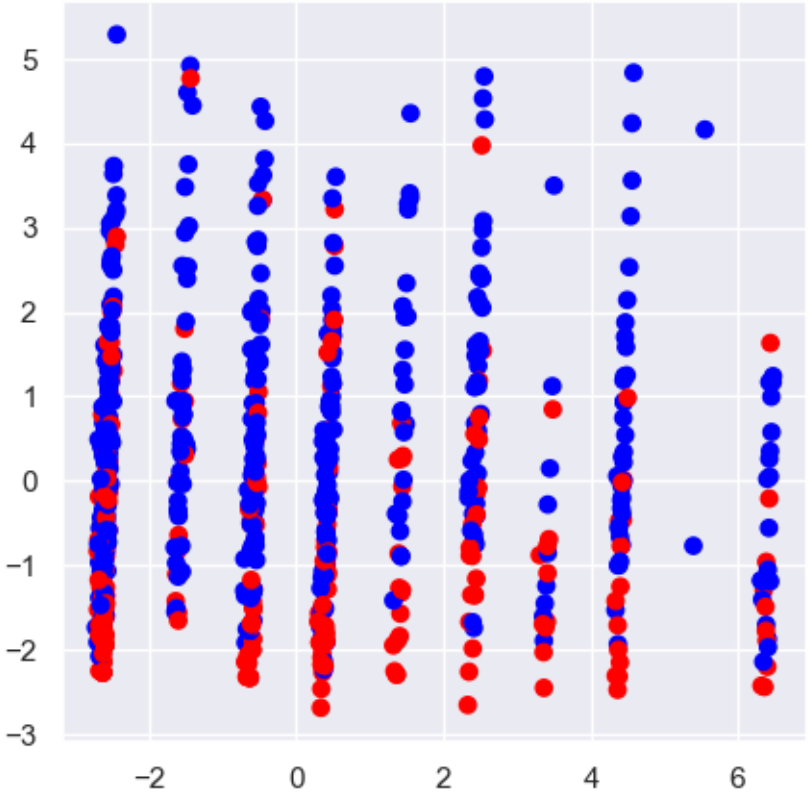
על מנת לעשות נרמול השתמשנו בשיטה `min-max-scalar`, ונרמלנו את הערכים להיות בטווח בין 0 ל-1.

7) Data Reduction – PCA

הפעלנו PCA על הנתונים, על העמודות שנשארנו עימן בסעיף הקודם, ששרדו את הניקוי, נרמול, זריקה ובידוד. יש לנו כעת רק ערכים מספריים, עליהם נפעיל ניתוח גורמים ראשיים (PCA), בעזרת מתודת PCA של ספריית sklearn. תוצאת פעולה זו הביאה data frame של המידע לאחר הורדת המימד, אל 13 תכונות על כל הרשומות (תוך כדי שמירה על מקסימום שונות).

מצאנו את השונויות של התכונות לאחר הורדת המימד (החדשות) בעזרת explained_variance_ratio_, חישבנו את סכומן, וראינו שהסכום שהתקבל הוא 0.999 כפי שכפינו על הספרייה.

ראינו ששתי התכונות הראשונות בעלות סכום שונויות של יותר מחצי לכן הן מכילות את רוב המידע, לכן הצגנו אותן בדיאגרמת פיזור scatter.



אכן ראינו שהמידע שמתקבל יחסית מסודר, כמו שציפינו.