# **פרויקט בקורס לימוד מכונה**

**חלק א'**

**דצמבר 2021**

**מגישים:**

ניתאי בן ברוך – 312473242

מירית זליצ'ונוק – 318381852

תוכן עניינים

[חלק א - Data collection and Sensing 2](#_Toc89372126)

[חלק ב - Dataset Creation 2](#_Toc89372127)

[1. Exploratory data analysis 2](#_Toc89372128)

[2. Dataset Creation Steps 5](#_Toc89372129)

[1. Pre-processing 5](#_Toc89372130)

[2. Segmentation 7](#_Toc89372131)

[3. Feature Extraction 7](#_Toc89372132)

[4. Feature Representation 8](#_Toc89372133)

[5. Feature Selection 9](#_Toc89372134)

[6. Dimensionality Reduction 10](#_Toc89372135)

[חלק ג - Model Training 10](#_Toc89372136)

# חלק א - Data collection and Sensing

ה-Data collection של הפרויקט מפרט פרטים שונים אודות אנשים הנמצאים בתוכנית הכשרה ל-Data Science בחברה מסוימת. ה-Sensing אשר בוצע על הדאטה הינו סטטי, משום שהינו מכיל פרטים קבועים יחסית אודות אותם אנשים - כמו מגדר, עיר מגורים, השכלה ופרטים אודות החברה בה עובדים כעת.

ניתן להוסיף סוג sensing חדש כמו תחושת שייכות לחברה בה המועמד עובד כעת (נניח טווח בין 1 ל-5, כאשר 1 מסמל את הדרגה הנמוכה ביותר (אי-שייכות) ו-5 את הגבוהה ביותר). סוג זה אמנם הינו דינאמי יותר מהסוגים האחרים הקיימים ב-data collection, אך יכול לתרום משמעותית לחיזוי של מי מבין האנשים ירצה להמשיך לעבוד בחברה לאחר התוכנית. כלומר, ייתכן כי אנשים אשר מרגישים תחושת שייכות נמוכה למקום עבודתם הנוכחי, ירצו להשתלב בחברה זו לאחר התוכנית.

קטגוריית משימת הלמידה היא Supervised learning כאשר סוג המשימה הינו binary classification. משימת הלמידה הינה supervised משום שה-labels שלנו ידועים (משתנה המטרה - target). כלומר, 1 - מחפש שינוי בעבודה, 0 - לא מחפש שינוי בעבודה.

יהיה ניתן להשתמש בנתונים אלה כדי לבצע משימות למידה נוספות, כמו למשל - האם לעובד יש ניסיון רלוונטי בעבודתו הנוכחית (כלומר לחזות את עמודת relevent\_experience).

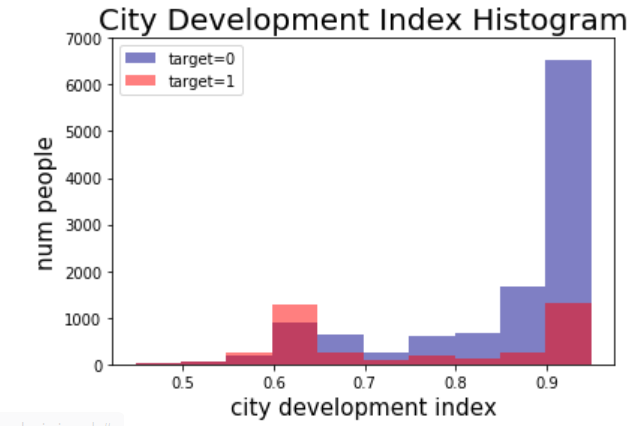
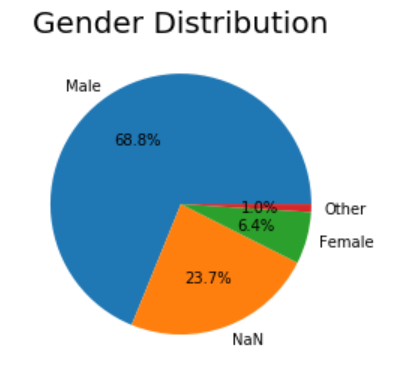
# חלק ב - Dataset Creation

## 1. Exploratory data analysis

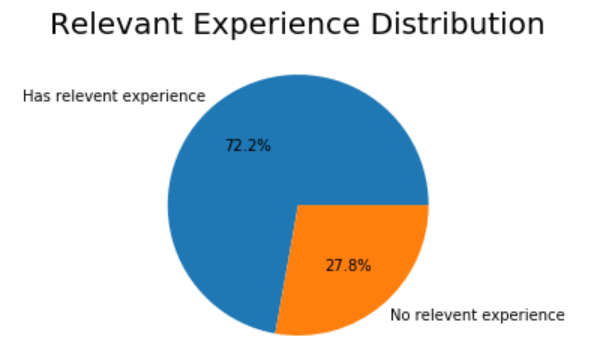
משתנה target - משתנה זה הינו משתנה המטרה של משימת הלימוד, וטווח הערכים שלו בינארי (1 - מחפש לעזוב, 0 - לא מחפש לעזוב). ניתן לראות כי סט הנתונים אינו מאוזן - קיימים פי 3 יותר אנשים שמחפשים להישאר מאשר לעזוב.

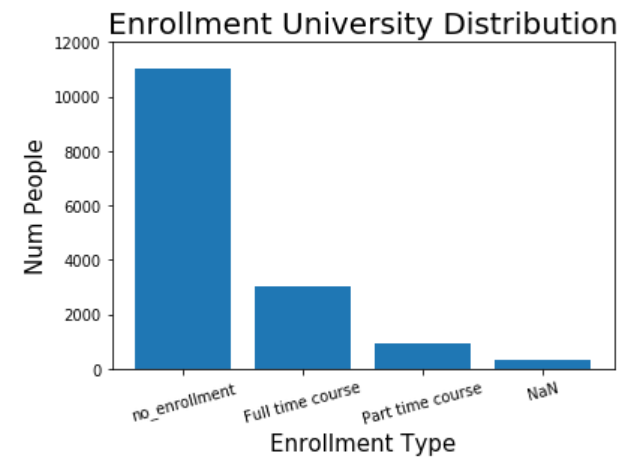
משתנה city - משתנה בעל ערכים שמיים, ואין בו ערכים חסרים. קיימים 122 ערים ייחודיות כאשר הפופולרית ביותר הינה city\_103. בחרנו לא להציג את התפלגות הערים, משום שקיימים ערכים רבים בשדה זה וכי הינו קשור באופן ישיר לשדה city\_development\_index אשר אינפורמטיבי במידה רבה יותר.

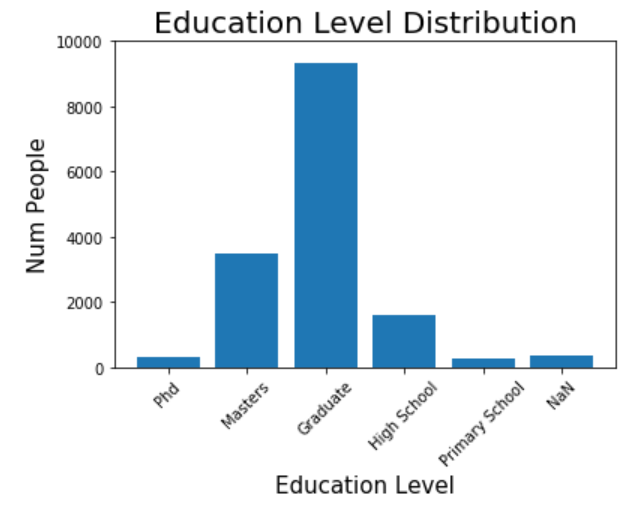
משתנה city\_development\_index - טווח הערכים הינו רציף ונע בין 0.448 ל-0.949.

משמעות המשתנה בהקשר של משימת הלימוד היא בחינת הקשר האם עזיבת מקום העבודה של העובד תלוי בדירוג פיתוח העיר ממנה מגיע. ערכנו בדיקה של ממוצע דירוגי פיתוח הערים בין העובדים שמחפשים לעזוב את החברה (הממוצע הינו 0.7558), אל מול עובדים שלא מחפשים לעזוב את החברה (הממוצע עומד על 0.8534). כלומר, ההשערה שלנו היא שככל שדירוג העיר גבוה יותר כך אחוז העזיבה נמוך יותר.

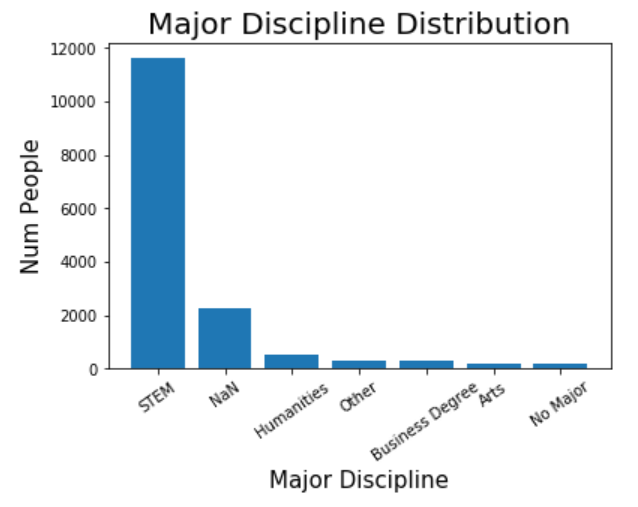
משתנה gender - טווח המשתנים: Male, Female, Other וקיימים גם ערכים חסרים (כרבע מהדאטה). משמעות המשתנה בהקשר של משימת הלימוד היא בחינת ההשפעה של המגדר על ההחלטה אם להישאר או לעזוב. כפי שניתן לראות בגרף, סט הנתונים אינו מאוזן.

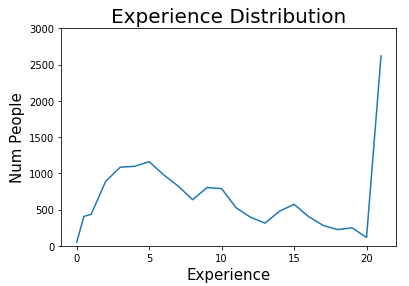
  
משתנה relevent\_experience - טווח המשתנים הינו  
בינארי - כלומר, יש ניסיון רלוונטי או אין ניסיון רלוונטי. משמעות המשתנה בהקשר של משימת הלימוד היא האם הניסיון רלוונטי להחלטה של עובד האם לעזוב או להשאר.

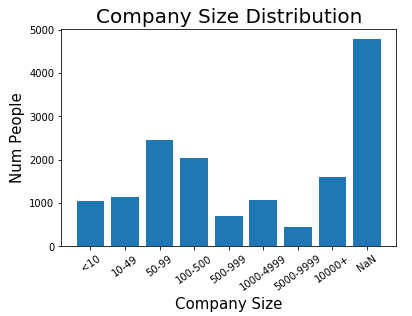
  
משתנה enrolled\_university - משתנה בעל ערכים שמיים (ללא, חלקית, ו-full time). משמעות המשתנה בהקשר של משימת הלימוד היא האם יש קיימת השפעה של הרשמה לקורס אוניבסיטאי, בין אם באופן חלקי או לא על ההחלטה.

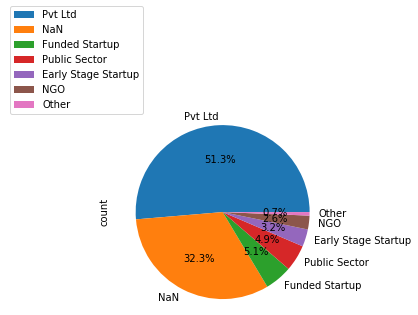


משתנה education\_level - משתנה מסוג קטגוריאלי בעל סדר. משמעות המשתנה בהקשר של משימת הלימוד היא האם לרמת ההשכלה קיימת חשיבות בבחירתו של העובד אם לעזוב או להישאר.

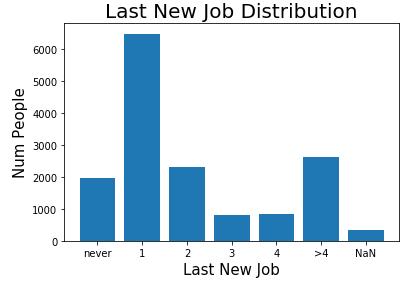
  
  
משתנה major\_discipline - טווח הערכים הינו קטגוריאלי לא אורדינלי. משמעות המשתנה בהקשר של משימת הלימוד היא האם לסוג ההשכלה קיימת חשיבות בבחירתו של העובד אם לעזוב או להישאר.

משתנה experience - טווח הערכים הינו מספרי אורדינלי עם שני ערכי קצה (1> , 20<). משמעות המשתנה בהקשר של משימת הלימוד היא האם לכמות הניסיון יש השפעה בבחירתו של העובד אם לעזוב או להישאר.

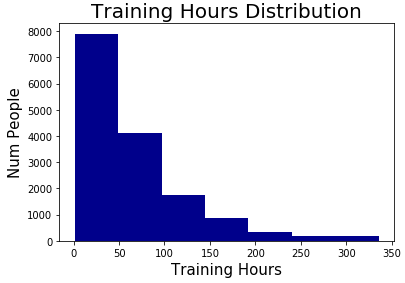
  
משתנה company\_size - טווח הערכים הינו קטגוריאלי אורדינאלי.

משמעות המשתנה בהקשר של משימת הלימוד היא האם לגודל החברה יש השפעה בבחירתו של העובד אם לעזוב או להישאר בעבודתו.

משתנה company\_type - טווח הערכים הינו קטגוריאלי לא אורדינלי. משמעות המשתנה בהקשר של משימת הלימוד היא האם לסוג החברה שבה עובד תהיה השפעה בבחירתו שלו אם לעזוב או להישאר בעבודתו.

  
משתנה last\_new\_job - טווח הערכים הינו מספרי אורדינלי עם שני ערכי קצה (Never=0 , >4).

משמעות המשתנה בהקשר של משימת הלימוד היא האם ההפרש (בשנים) בין העבודה הקודמת לנוכחית תהיה השפעה בבחירתו של העובד אם לעזוב או להישאר בעבודתו הנוכחית.



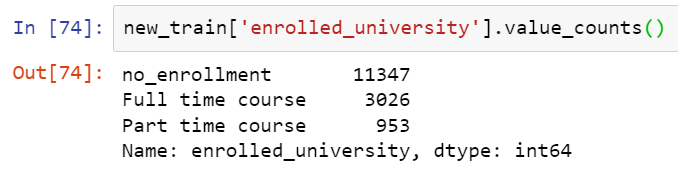
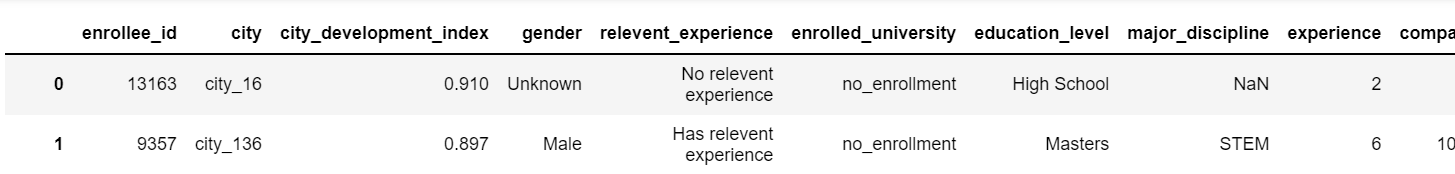
משתנה training\_hours - טווח הערכים הינו מספרי אורדינאלי (0-350). משמעות המשתנה בהקשר של משימת הלימוד היא האם לכמות השעות שהושלמו בקורס ישנה השפעה בבחירתו של העובד אם לעזוב או להישאר בעבודתו.

## 2. Dataset Creation Steps

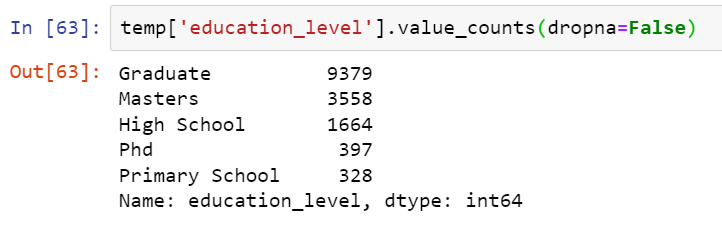
### Pre-processing

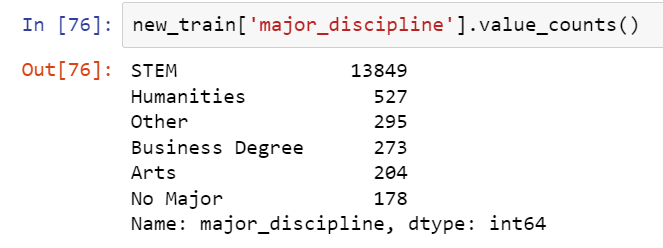
* ראשית, בדקנו האם קיימות רשומות כפולות עבור אותם האנשים, ונראה כי לא קיימים כאלו.
* בשלב זה נתמודד עם ה-missing values בעמודות הבאות:

**Gender**: בשדה זה יש 23.7% ערכים חסרים. בגלל שאחוז זה הינו יחסית גובה, ניצור קטגוריה נוספת מסוג Unknown.

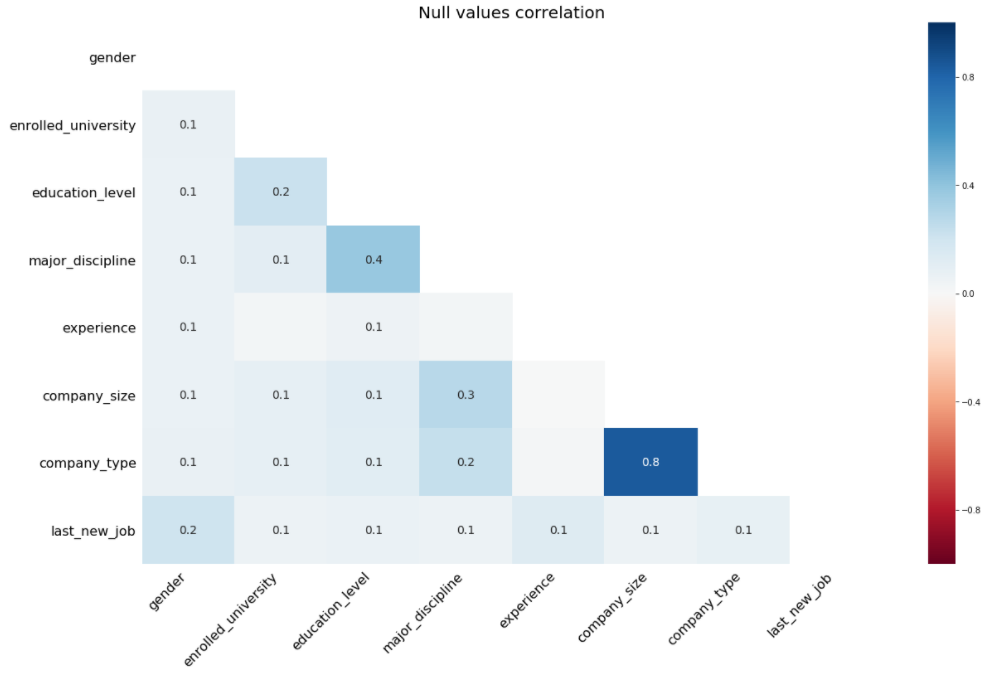


**Enrollment university**: בשדה זה יש 2.02% ערכים חסרים. מכיוון שאחוז הערכים החסרים הינו יחסית נמוך, החלטנו להחליף אותם בערך הפופולרי ביותר שהינו "no\_enrollment".

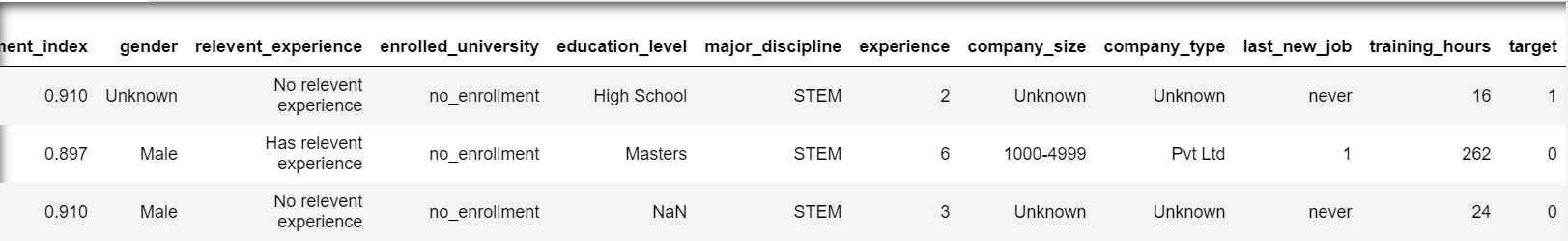
**Education level**: בשדה זה יש 2.41% ערכים חסרים ועל מנת לשמור התפלגות שווה נחלק את מספר הרשומות הלא ידועות שווה בין כל רמות הלימוד.

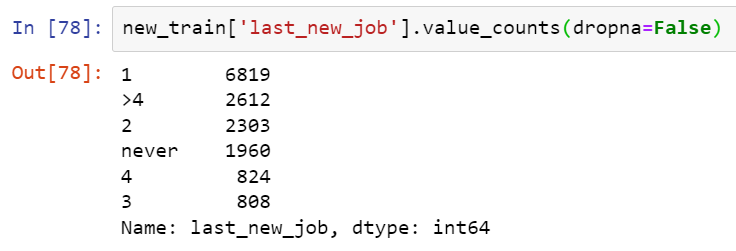
**Major discpline**: בשדה זה יש 14.59% ערכים חסרים, ועבורו החלטנו להחליף אותם בערך הפופולרי ביותר שהינו "STEM" ומהווה כ-76% מהדאטה כאשר שאר הערכים האחרים ביחד מהווים רק כ-9.41%.

**Company size ו-Company type**: בשדה הראשון יש 31.18% ערכים חסרים ובשני יש 32.25% ערכים חסרים. בגלל שמות השדות, בדקנו את הקשר ביניהם - וניתן להסיק מהגרף שב-80% מהמקרים בהם יש ערך חסר בשדה company\_type ישנו ערך חסר גם בשדה company\_size.



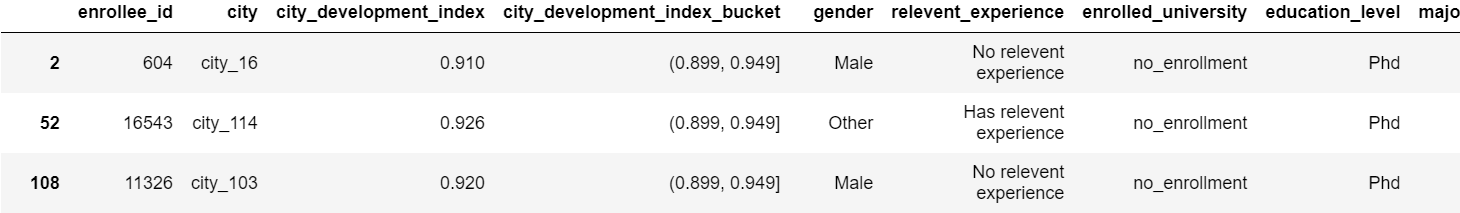
לכן, החלטנו להתמודד עם הערכים החסרים באופן הבא: עבור רשומות בהם compnay\_type קיים וcompany\_size לא, נשים את הערך הכי פופולרי עבור אותו company\_type, ולהפך. עבור רשומות בהן בשני השדות קיימים ערכים חסרים - ניצור קטגוריה נוספת של unknown עבור כל עמודה ונאכלס אותן בהן.



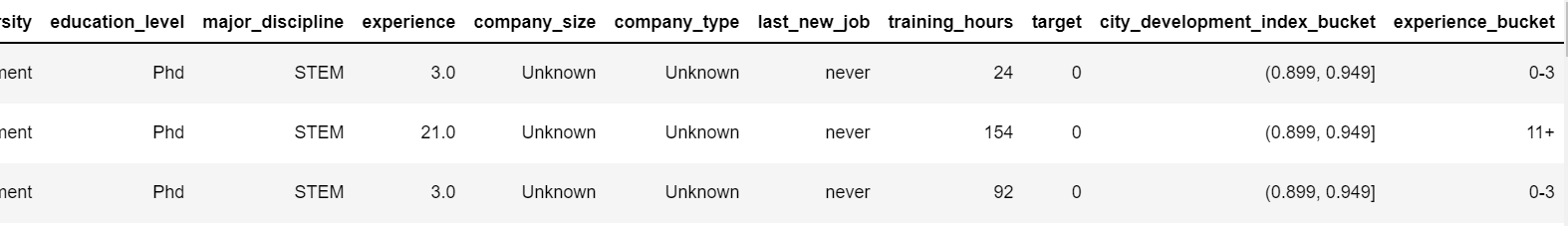
**Last new job**: בשדה זה יש 2.16% ערכים חסרים. מכיוון שאחוז הערכים החסרים הינו יחסית נמוך, החלטנו להחליף אותם בערך הפופולרי ביותר שהינו שנה אחת (1).

* בנוסף, בשלב זה נבצע שיוך ערכים רציפים לקטגוריות עבור השדות הבאים:

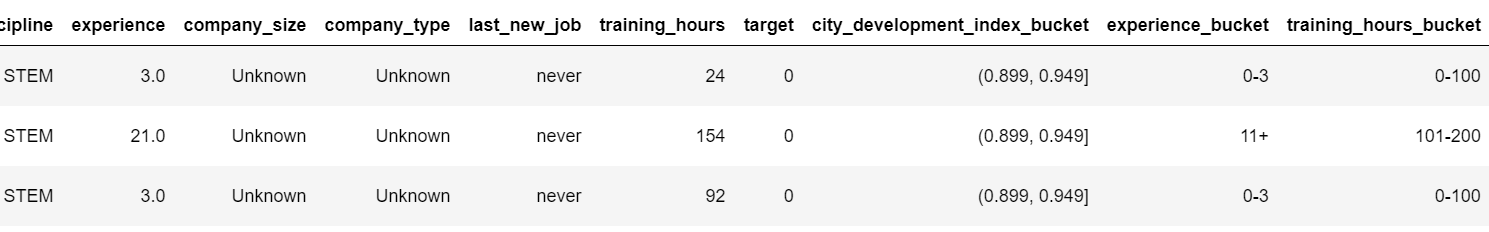
**city\_development\_index**: טווח הערכים של משתנה זה נע בין - 0.448 ל- 0.949 וניתן לחלק את הטווח לעשירונים לפי שיטת EWD. בחרנו לחלק את הטווח ל-10 ולא פחות, משום שניתן לראות בגרף שהעשירון הגבוה ביותר הינו גם בעל המשקל הגבוה ביותר, לכן אם נחלק את הטווח לפחות חלקים - ייתכן ונתעלם מערכים חשובים בטווח. ובעקבות כך הוספנו ל-DataFrame שלנו עמודה בשם city\_development\_index\_bucket.



**experience**: במשתנה זה קיבלנו ערכים מספריים אך גם טווחים (כמו <1 ו- >20). לכן, החלטנו לחלק את משתנה זה לשלוש קטגוריות: בין 0 ל-3 שנות ניסיון (מתחיל), בין 4 ל-10 שנות ניסיון (בינוני), ו-11 שנות ניסיון ומעלה (מומחה).



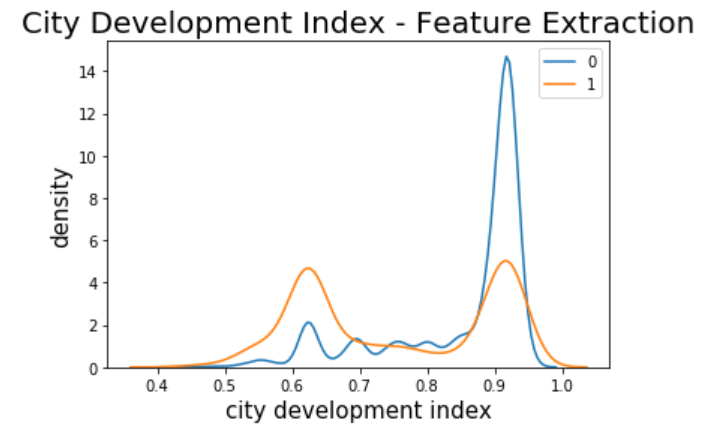
**training hours**: לפי ההיסטוגרמה של משתנה זה בחלק הקודם, החלטנו לחלק אותו לשלושה טווחים: 0-100 (מתחילים), 101-200 (אמצע קורס), 201+ (מתקדמים).



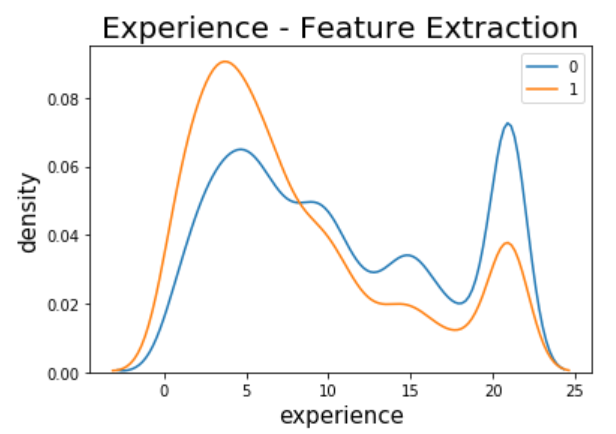
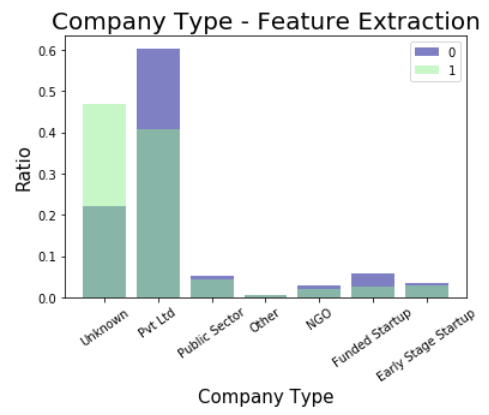
### Segmentation

ביצוע סגמנטציה לרוב מבוצע על Data שאינו טבלאי כמו תמונות, קבצי שמע וכו'. על שלב זה נוותר מכיוון שה-DataSet שלנו כבר במודל טבלאי, ואין צורך לבצע קיטוע בפיצרים מכיוון שהם ראשוניים.

### Feature Extraction

על פי ההבנה שלנו, המתודולגיה בטקסונומיה שמתאימה לסט הנתונים שלנו היא fixed number of features משום שאין כאן תלות בזמן, וסט קבוע יתאים יותר למשימת הלמידה שלנו (מי מהאנשים ימשיך ומי יעזוב) מאשר סט משתנה של פיצ'רים.

הפיצ'רים אשר בחרנו לחלץ:

1. **city\_development\_index** - רוב האנשים אשר רוצים להשאר בחברה מגיעים מעיר ברמת פיתוח גבוהה יחסית, לעומת אנשים אשר מתכננים לעזוב.
2. **experience** - מנקודת החיתוך של הגרפים, ניתן (בערך) לחלק את הגרף לשני חלקים - אחוז האנשים שרוצים להישאר בחברה הינו גבוה יחסית עבור אלו שיש להם כ-8 שנות ניסיון ומעלה, ואחוז האנשים בעלי רצון לעזוב את החברה הינו גבוה יחסית עבור אלו שיש להם פחות מ-8 שנות ניסיון.
3. **Company type** - עבור אלו שרוצים להשאר בחברה, אחוז האנשים אשר עובדים בPvt Ltd הינו גבוה ביחס לאלו שרוצים לעזוב. עבור אלו שרוצים לעזוב את החברה, אחוז האנשים אשר סוג החברה שלהם אינו ידוע הינו גבוה ביחס לאלו שרוצים להשאר. הבדיקה שלנו הייתה בהינתן שאנחנו יודעים את הtarget מראש, ולא בהינתן שאנחנו יודעים את הcompany\_type מראש (כפי שהמודל יקבל בסט ה-test).

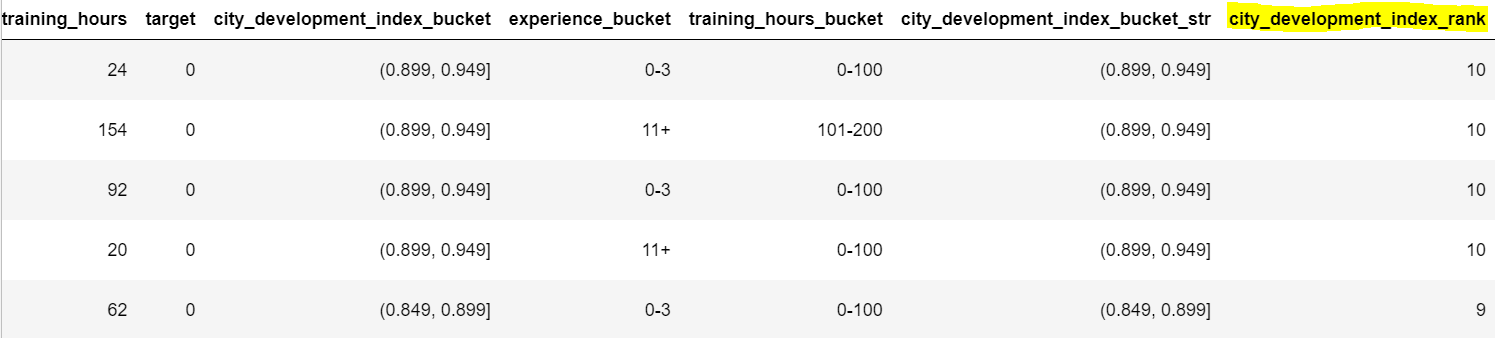


1. **Last New Job** - ניתן לראות כי גם במשתנה זה קיימים מעט הבדלים בין סמפלים שהם target=0 ל-target=1, אף על פי שקיימת חפיפה מאד גבוהה.

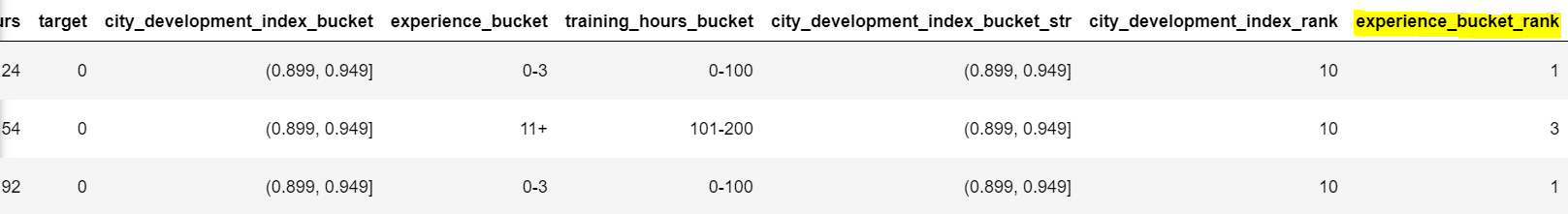
בשאר המשתנים אשר ניתחנו, ניתן היה לראות כי ההתפלגויות חופפות הן ב-target=0 והן ב-target=1, ולכן לא חילצנו אותם.

### Feature Representation

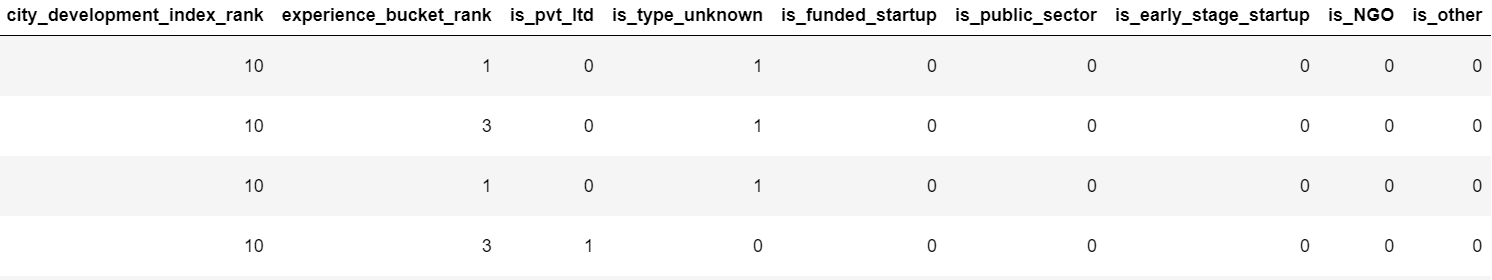
* **city\_development\_index -** את פיצ'ר זה חילקנו לעשר קבוצות שוות בין טווח הערכים 0.448-0.949, למשל את הקבוצה הראשונה (0.448-0.498) נדרג rank=1 וכך בהתאמה.



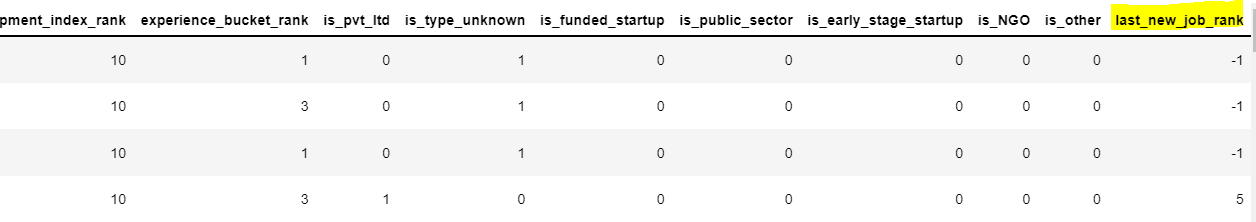
* **experience -** את פיצ'ר זה חילקנו לשלוש קבוצות, לא שוות. בהצגת פיצ'ר זה נציג 3 קבוצות על פי דירוג: בין 0 ל-3 שנים rank=1, בין 3 ל- 10 שנים rank=2, ו11 שנים ומעלה rank=3.



* **Company type -** פיצ'ר זה הינו קטגוריאלי, ולכן נציג אותו כ-7 פיצ'רים בוליאניים בכך שנהפוך כל קטגוריה לפיצ'ר למשתנה בוליאני ונוסיף ל-dataframe שלנו עמודות בהתאם (is\_type\_unknown, is\_pvt\_ltd וכדומה).

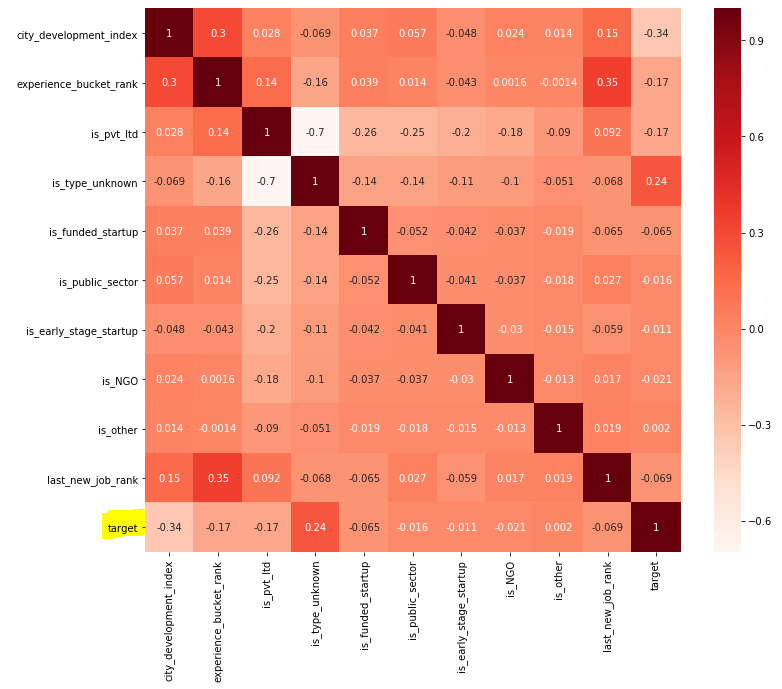


* **Last New Job -** פיצ'ר זה הינו קטגוריאלי אורדינלי, ומשום שקיימים בו רק 6 קטגוריות, בחרנו להשאיר אותו מספרי כפי שהוא אך רק לשנות את 4< ל-5 ואת never ל- 1-.



### Feature Selection

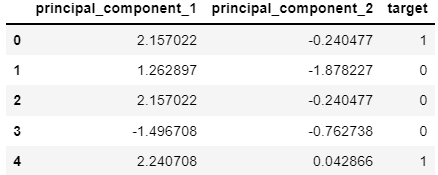
בשלב זה נשתמש ב-Correlation Feature Selection שהינה הערכה כמותנית מסוג Filter. במפת החום למטה, נראה כי הקורלציה של הפיצ'רים השונים עם משתנה הtarget (בשורה האחרונה) היא יחסית נמוכה. בכל זאת, בחרנו לקחת את כל הפיצ'רים שהקורלציה שלהם גדולה מ-10% (כלומר, city\_development\_index, experience, is\_pvt\_ltd ו-is\_unknown).



\*\*\* נציין כי מפת החום שלנו לא הצליחה לקלוט את השדה הקטגוריאלי שבנינו עבור דירוג העיר - city\_development\_index\_rank ולא הצלחנו לפתור זאת. לכן, נאלצנו להכניס את השדה המקורי.

### Dimensionality Reduction

בשלב זה נבצע PCA ונפחית את 4 המימדים שבחרנו בשלב הקודם ל-2 מימדים.



# חלק ג - Model Training

על סמך החומר אשר למדנו בהרצאה השלישית, נבחר בשיטת validation מסוג k-fold. בחרנו בה משום ששיטה זו הינה המדויקת ביותר מבין השיטות המתאימות לכמות נתונים גדולה. היתרון הבולט של שיטה זו הוא שלא נדלג על אף רשומה בזמן אימון המודל, לעומת שיטת holdout אשר בה ישנה סבירות גבוהה יותר שנאבד נתונים קריטים בזמן האימון.

בתהליך הולידציה שנבצע נחלק את סט האימון שלנו ל-10 foldים שווים, כאשר המודל ילמד על 9 מתוך 10 ה-foldים (ה-traning set) ויבצע validation על ה-fold הנותר. תהליך זה יימשך עד שהמודל יבצע validation בכל אחד מהfoldים כלומר, נבצע זאת 10 פעמים.