**פרויקט**

**Machine Learning**

**חלק א'**



מגישים :

עומר ויצמן – 205505928

מיכאל אלפסי - 308307776

תוכן עניינים

[Data collection and Sensing 1](#_Toc69742402)

[Exploratory data analysis 2](#_Toc69742404)

[PreProcessing 5](#_Toc69742405)

[Segmentation 7](#_Toc69742406)

[Feature extraction 7](#_Toc69742407)

[Feature Representation 8](#_Toc69742408)

[Feature Selection 8](#_Toc69742409)

[Dimensionality Reduction 9](#_Toc69742410)

[Model Training 10](#_Toc69742411)

[נספח קורלציות פיצ'רים-משתנה מוסבר 10](#_Toc69742412)

# Data collection and Sensing

מטרת הפרויקט היא לחזות מי מבין האנשים ירצה להמשיך לעבוד בחברה DS ומי יצא לחפש עבודה במקום אחר בסיום התכנית ההכשרה עפ"י סט הנתוניםהקיים ונכון לרגע איסוף הדאטה.

Data Collection– הינו אוסף samples המגיעים מאותו ה.domain Data collection איכותי הינו אוסף samples המייצגים את המציאות, כלומר אוסף הsamples הינו מגוון ומציג את כלל המקרים הקיימים באוכלוסייה, אך ללא חזרתיות מיותרת, זהו למעשה הבסיס למערכת לומדת מוצלחת. ה-Data collection בפרויקט זה הינו אוסף נתונים שנאספו הודות האנשים שלוקחים חלק בהכשרה של חברת DS.

Sensing –שלב החישה בו אנו מבצעים חישה על הsmaples ומאפיינים אותם על פי פיצ'רים שונים. הSensing שבוצע על הדאטה בפרויקט הינו סטטי ודינמי. פיצ'ר סטטי הינו פיצ'ר המייצר אותם ערכים ללא תלות בזמן הדגימה ודינמי תלוי בזמן הדגימה.

פיצ'ר דינאמי – training\_hours בודק את מספר השעות שהעובד השלים מהקורס, זהו משתנה שיכול להשתנות בכל שעה ולכן החלטנו שהוא דינאמי.

פיצ'רים סטטיים – כל שער הפיצ'רים הינם סטטים ולא תלויים בזמן.

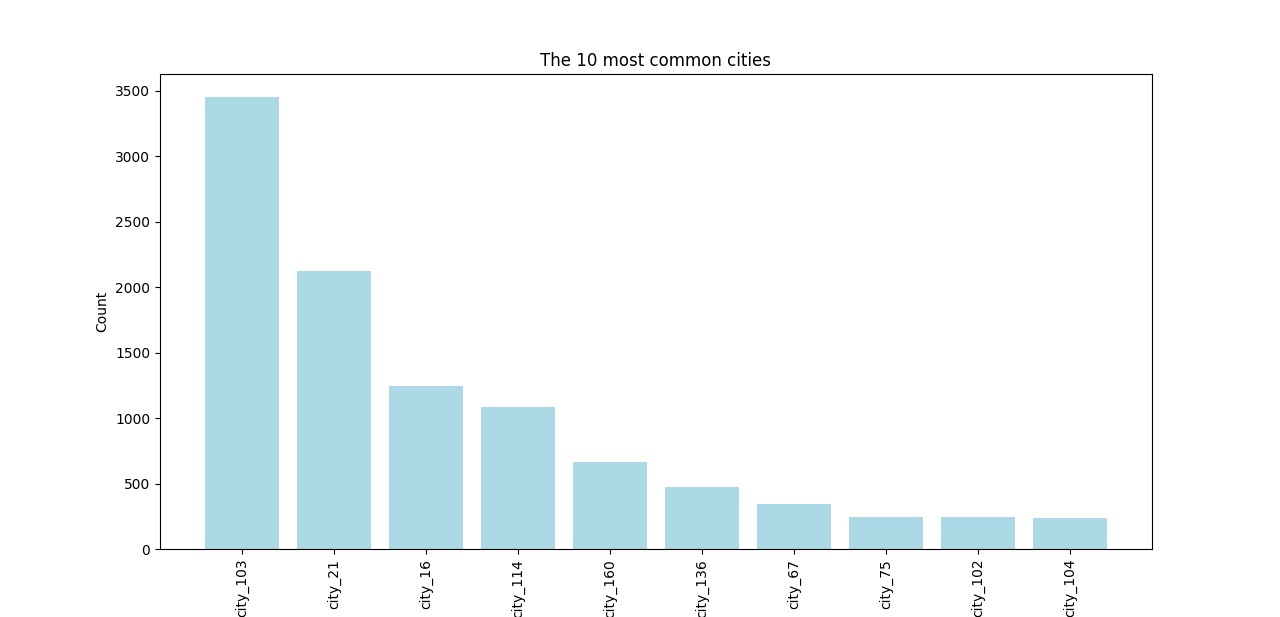
Add Sensing – מאחר ושני סוגי החישה שנלמדו בקורס מתקיימים בדאטה, בחרנו להציע פיצ'ר נוסף מסוג סטטי. הפיצ'ר הנוסף יקרא num\_interview ובו נשאל על מספר ראיונות העבודה שאדם ביצע ב-5 שנים האחרונות, בכך נוכל לבדוק האם העבוד נמצא בשלב חיפושי עבודה בחייו או שהוא מעוניין במציאת עבודה קבועה.

Category & Type - קטגוריית משימת הלמידה היא Supervised Learning כיוון שהמאפיין העיקרי של קטגוריה זו היא זמינות ה labels שהרי מתקיימת בדאטה הקיים. סוג משימת הלמידה הינו Prediction משום שמשימתנו היא לפתח מודל חיזוי לאם עובד יבחר להישאר בחברה בתום ההכשרה. בהתבסס על המשתנים הקיימים. באמצעות אותו סט נתונים נוכל לבצע משימת למידה מסוג Multy Label Classification ולסווג על פי המידע על העובד האם הוא יבחר להישאר בחברה או לא.

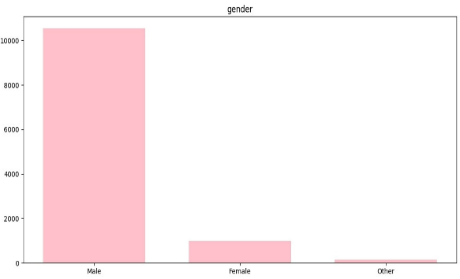
[**Exploratory data analysis**](https://en.wikipedia.org/wiki/Exploratory_data_analysis)

בשלב ראשוני זה נרצה לאפיין את המשתנים בסט הנתונים ולזהות אילו משתנים משפיעים על משימת הלמידה.

פיצ'רים קטגוריאליים:

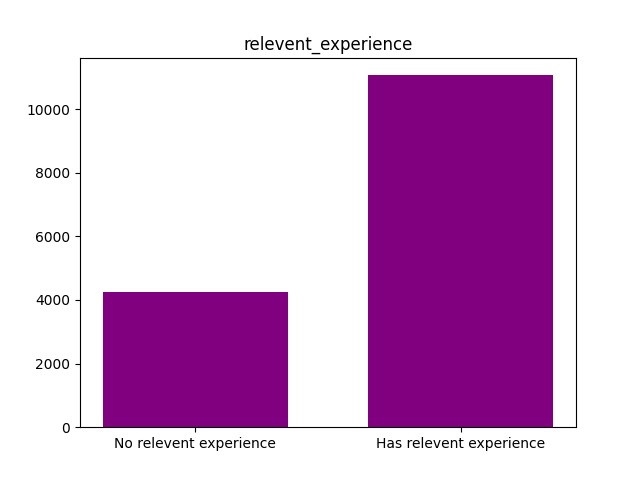


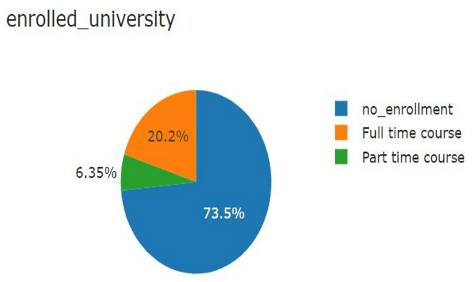
* City – פיצ'ר זה מציין באיזו עיר מתגורר כל אחד מהאנשים שלוקחים חלק בתוכנית ההכשרהDS .

בחרנו להציג את עשרת הערים בהם מתגוררים הכי הרבה אנשים מתוך כלל הערים. פיצ'ר זה יכול להוות חשיבות רק בהקשר לפיצ'ר מדד האיכות של כל עיר ובכך לתת מידע שאולי יהיה רלוונטי לחיזוי שאנו רוצים לעשות על כל עובד.

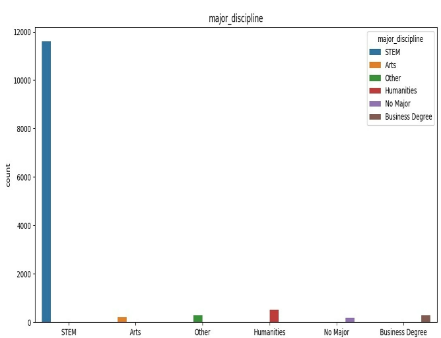
* gender – פי'צר המציין עבור כל אחד את מין העובד.

ניתן לראות כי מרבית העובדים הינם גברים וזה אולי מעיד על בעיה באספת הנתונים שהתמקדה בקבוצת הגברים (מעיד על פיזור לא נכון). בכל מקרה אנו מעריכים שאין משמעות למגדר העובד בהיבט החיזוי שאנו רוצים לעשות.

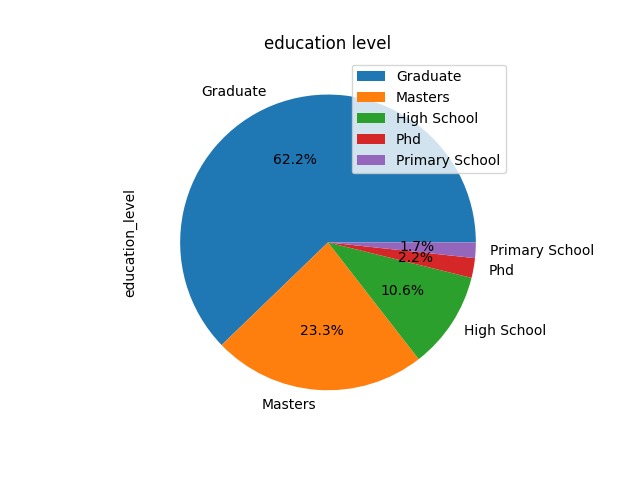


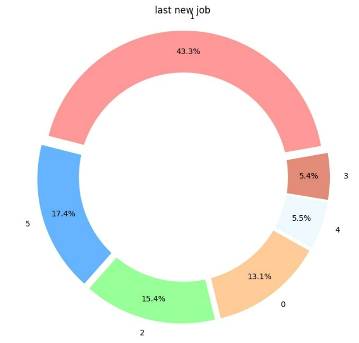
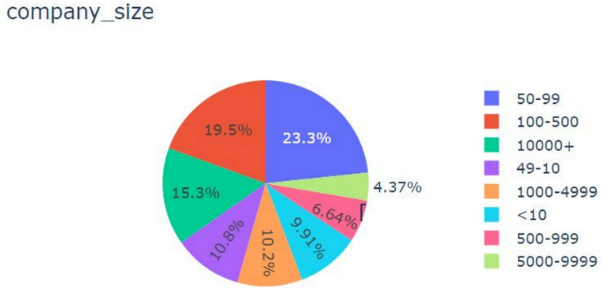
* relevent\_experience – פיצ'ר המציין עבור כל אחד האם יש לנו ניסיון בתחום הרלוונטי להתמחות הנעשית בחברה או לא (משתנה בינארי). אנו מניחים שהמידע הזה חשוב ויכול להשפיע על משתנה המוסבר שלנו, סביר להניח שאדם שאין לו ניסיון בתחום יעדיף להישאר בחברה לאחר ההכשרה, כי זו מציעה לו עבודה מובטחת בתחום לעומת אדם שכבר יש לו ניסיון ואולי יעדיף לחזור למקום העבודה הקודם שלו/ למצוא מקום עבודה חדש. כפי שניתן לראות מרבית העובדים מגיעים לקורס עם ניסיון רלוונטי.
* enrolled\_university – פיצ'ר המציין האם העובד רשום באופן מלא לקורס/ לא רשום כלל/ רשום באופן חלקי.

כפי שניתן לראות מעל 70% מהעובדים אינם רשומים בנוסף לקורס. המשתנה יכול להעיד אולי על כך שמרבית העובדים לא לוקחים קורסים נוספים במקביל ולכן זה יכול לעזור להעיד שהעובדים מחויבים לתוכנית ההכשרה.

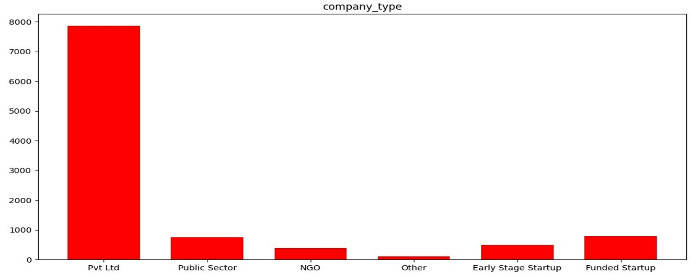
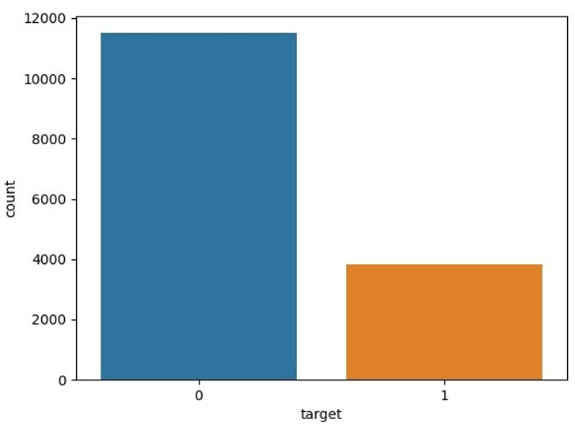
* major\_discipline- פיצ'ר זה מציג את העיסוק והתחום העיקרי של כל אדם בתוכנית. ניתן לראות שרוב מוחלט מהאנשים מגיעים מתחומי STEM הטכנולוגיים.

ניתן ללמוד מפיצ'ר זה על האפיון של כל תחום בהיבט תחלופת עובדים. לדגומא כידוע בתחומים הטכנולוגים ישנה נטייה גבוהה לתחלופת עבודה אל מול תחום האומנות למשל.

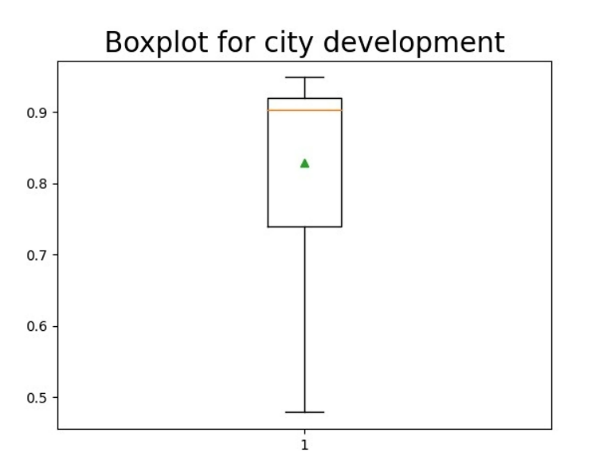


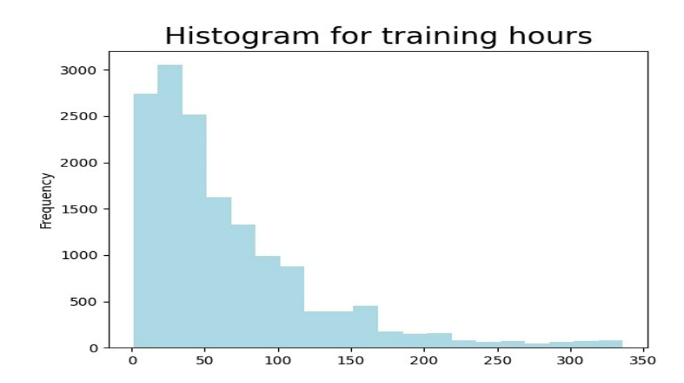
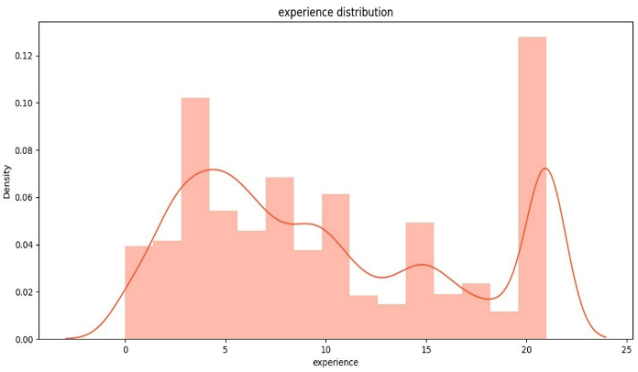
* education\_level – פיצ'ר המציין עבור כל אדם את רמת ההשכלה שלו. מידע זה רלוונטי מפני שאדם שאין לו השכלה ועושה קורס מקצועי סביר להניח שאין לו המון אפוציות לעבודה ואף סביר להניח שהוא מלכתחילה עושה את הקורס על מנת למצוא עבודה בתחום ובחברה לעומת אדם שיש לו השכלה גבוהה במיוחד ועושה קורס נוסף, אדם כזה סביר להניח שיש בפניו אופציות תעסוקה מגוונות, לכן משתנה זה עשוי להיות רלוונטי ולהשפיע על המשתנה המוסבר שלנו. כפי שניתן לראות מרבית העבודים בהכשרה שלנו בעלי תואר ראשון ותואר שני.
* last\_new\_job - פיצ'ר זה מציין כמה שנים עברו מאז העבודה האחרונה של כל עובד בהכשרה. ניתן לראות כי עבור כמעט מחצית מהעובדים חלפה שנה מאז העבודה האחרונה שלהם ושאר העובדים מתחלקים באופן יחסית אחיד בין שאר האופציות.
* company\_size - פיצ'ר זה מציין את גודל החברה ממנה הגיע כל אדם בהכשרה.

ניתן לראות כי קיים יחסית פיזור אחיד בין כלל האופציות הקיימות ולכן אנו מניחים שפיצ'ר זה לא יעזור לנו להסביר את המשתנה המוסבר בצורה טובה.

* company\_type – פיצ'ר זה מציין את סוג החברה ממנה על אדם בהכשרה הגיע ממנה. כפי שניתן לראות רוב מוחלט של העובדים מגיע מתחום השוק הפרטי. הנתונים של פיציר זה לא מאוזנים ולא יכולים לתרום לחיזוי מפני שרוב מוחלט מן הדאטה ישתייך לאותה סוג עבודה.
* Target- המשתנה המוסבר שלנו, משתנה בינארי המקבל 0 עבור עובד שלא מתכנן לעזוב את החברה בסוף ההכשרה ו1 עבור עובד שכן. כפי שניתן לראות בדרך מירב הדאטה שלנו הינה על עובדים שלא מתכננים לעזוב את החברה בסוף ההשכרה.

כלל הפיצ'רים הקיימים אמורים לעזור בחזוי המשתנה המוסבר.

פיצ'רים רציפים:

* city\_development\_index – פיצ'ר זה מציין את מדד הפיתוח של כל עיר בה מתגוררים העובדים בהכשרה. תרשים box plot מציג כי החציון גבוה באופן משמעותי מהממוצע , דבר המעיד על הימצאותם של תצפיות בעלי כמות צפיות חריגה, אשר מטות את ממוצע הצפיות בdata כלפי מטה.
* training\_hours – פיצ'ר זה מציין עבור כל אדם בהכשרה כמו שעות התנסות הוא צבר עד כה בתוכנית ההכשרה, טווח הערכים עבור כמות צפיות הינו: [336- 1] ,טווח זה הינו רחב יחסית ולכן נוכל להגיד כי הנתונים אכן מגוונים, הממוצע הינו 65.337 , החציון הינו 47 וסטיית התקן הינה 60.007. היסטוגרמת השכיחויות מצביעה על פיזור הנתונים שאינו סימטרי עם זנב ימני, המעיד על הטרוגניות הנתונים.
* experience – פיצ'ר זה מציין כמה שנות ניסיון יש לכל עובד בעולם התעסוקה באופן ככלי. טווח הערכים נע בין [0-20] וניתן לראות כי ישנם הרבה עובדים עם ניסיון עד 10 שנים ו20 שנה ומעלה, אך מעט יחסית עובדים עם ניסיון בין 10-20, מכך ניתן ללמוד אולי כי הקורס ההכשרה פונה לעובדים חדשים יחסית או עובדים ותיקים שרוצים לשפר או להחליף את הכשרתם.

## **PreProcessing**

Redundancy – מבדיקה שערכנו נמצא כי לא קיימים ערכים כפולים בסט הנתונים.

Missing Values – מבדיקה שערכנו מצאנו כי בסט הנתונים קיימים ערכי null: פלט זה מתאר את כמות ואחוז הערכים החסרים עפ"י הפיצ'רים. נטפל בבעיה זו באופן פרטני עבור כל פיצ'ר.

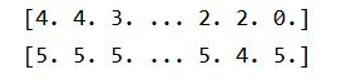
* עבור הפיצ'ר experience אשר חסרים בו 48 ערכים המהווים 0.31% מסך הנתונים לכן בחרנו להסיר את הרשומות מהדאטה כי זה כמעט ואינו משפיע על תוצאות המודל.
* עבור הפיצ'ר education\_level אשר חסרים בו 370 ערכים המהווים 2.41% מסך הנתונים למדנו כי לכל הרשומות הללו לא צויין גם ה- major\_discipline. מכך ניתן להסיק כי אלו אנשים אשר רמת הלימודים שלהם נמוכה/לא קיימת, כמו כן ראינו שישנם 1590 רשומות ברמת לימודים High School אשר לא ציינו את ה- major\_discipline שלהם ולכן בחרנו להציב ב-370 הערכים החסרים ב - education\_levelאת הערך High School.
* עבור הפיצ'ר major\_discipline אשר חסרים בו 2237 ערכים המהווים 14.6% מסך הנתונים עשינו התבוננות ביחס לפיצ'ר education\_level ולמדנו כי:
* ברמת לימודים High School ישנם 1590 רשומות שלא ציינו את ה- major\_discipline
* ברמת לימודים Primary School ישנם 254 שלא ציינו את ה- major\_discipline.
* ישנם 370 רשומות בהן לא מצויינות רמת הלימודים (כאמור הצבנו בהם את הערך High School) וה- .major\_discipline

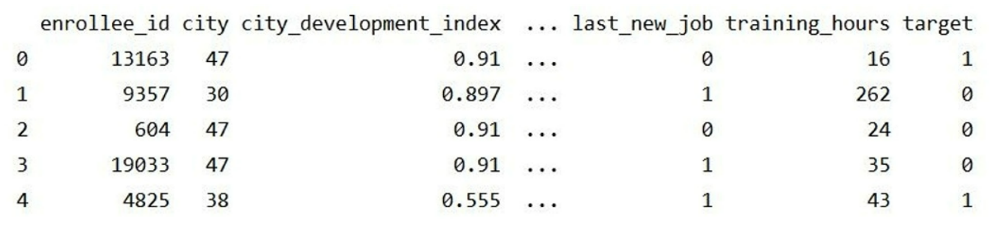
ניתן להסיק כי מרבית הרשומות החסרות ב - major\_discipline שייכים לרשומות של אנשים ברמת לימודים הנמוכה מ Graduate-( High School, Primary School ) ולכן אין להם מקצוע עיקרי עדיין, לכן נבחר לשים לכל אותם רשומות את הערך No Major ובכך טפלנו ב – 2214 רשומות.

בדקנו וראינו כי רוב מוחלט מרשומות אשר צויין בהם רמת לימודים Graduate ו- Masters צויין כי ה- major\_discipline שלהם הינו STEM ולכן בחרנו להציב ב18 הרשומות ברמת Graduate וב5 רשומות ברמת Masters אשר לא ציינו את ה- major\_discipline שלהם את הערך STEM.

* עבור הפיצ'ר last\_new\_job אשר חסרים בו 332 ערכים בחרנו להכניס ערך אשר מייצג את המספר הקרוב ביותר לממוצע של מי שציין כי עבד 4 ומטה שנים מהעבודה הקודמת שלו (ללא הערך 4<) מכיין שאנו מניחים שמי שלא ציין כמה שנים עברו מאז העבודה האחרונה שלו ככל הנראה שלא עבד כלל לפני.

הערך הממוצע של מספר השנים שעברו מהעבודה הקודמת (עם השמטה של כל מי שעברו מעל 4 שנים) הוא 1.357 ולכן בחרנו לשים את הערך 1 לכל 332 הרשומות הללו, זאת כי לדעתנו זה יפגע בצורה המזערית ביותר על אמינות הנתונים.

* עבור הפיצ'ר gender אשר חסרים בו 3639 ערכים בחרנו להשלים את הנתונים על פי המשקולות היחסיים של כל אופציה (male, female, other ) ביחס לכל הדאטה שלנו (ללא 3639 הרשומות החסרות שלא ציינו את מגדרם), סך הכל ישנם 11687 רשומות כאלה. בפעולה זו נשמור על היחסים של כל אופציה ביחס לכל הדאטה וזאת תוך כדי ההנחה שאין חשיבות למגדר של האדם ביחס לשאלה שאנו רוצים לחזות.
* עבור הפיצ'ר enrolled\_university אשר חסרים בו 311 רשומות, ראינו כי 11036 רשומות משתייכות לאי רשום לקורס ולכן אנו מבינים שרוב המוחלט של הדאטה שלנו משתייכת לאנשים שעדיין לא נרשמו לקורס, לכן אנו מניחים שמי שלא רשם את סטטוס ההרשמה שלו ככל הנראה לא רשום לקורס עדיין והצבנו בערכים החברים האלה את הערך no\_ enrollment.
* עבור הפצ'ר company\_type אשר חסרים בו 4943 רשומות והפיצ'ר company\_size אשר חסרים בו 4779 רשומות מכיוון שלא מצאנו הסבר שסיפק אותנו ובשל מספר הרב של הדאטה החסרה בחרנו לעשות מודל למידה מסוג KNN אשר ינבע את הערך שיש להציב בכל אחד מהערכים החסרים בעמודות הללו ובכך לא נאבד את הדאטה. על מנת לבצע את המודל כעת ובכלל בשביל שנוכל לבצע בדיקות נדרשות בשלבים הבאים ביצענו Label encoding לפיצ'רים שהערכים שלהם לא היו נומרים. יש לקחת בחשבון כי המודל אינו נחשב מודל עם דיוק מירבי אבל הוא אמור לתת תוצאה מספקת לפיצ'רים הקטגוראלים הללו כי מרבית התצפיות משתייכות לקטגוריות מסויימות ולא קשה לחזות את הערך הנכון עבור הערכים החסרים. פלט של הערכים שנכנסו בעזרת KNN:
* בסוף תהליך מילוי הערכים החסרים ולאחר ביצוע בדיקה ניתן לראות כי אין יותר ערכים חסרים בדאטה וכי כלל הדאטה הומרה לערכים נומרים:



Outliers – אנו רוצים להציג סט נתונים המתאר את המציאות באופן המיטבי ביותר, לא נסיר תצפיות חריגות ונשתמש בכמות הdata הקיימת.

Data type conversions -Discretization – בחרנו להפוך את המשתנה הרציף city\_development\_index למשתנה קטגוריאלי עפ"י שיטת EFD (Equal Frequency Disc) לשלושה מקטעים –"Med developed" " High developed" ו ""Low developed ,מקטעים אלה מציינים את רמת האיכות של העיר ממנה כל עובד הגיע.

## **Segmentation**

שלב טכני בו מוציאים מהentity את המידע הרלוונטי למשימת הלמידה שלנו. ה Data בצורה מפורשת ולכן אין צורך לשלוף מידע.

## **Feature extraction**

זהו שלב הליבה של תהליך הdataset creation ,בשלב זה נחלץ אלמנטים שיתנו לנו אינדיקציה מהפיצ'רים הקיימים וכמו כן, נתאר וניצור פיצ'רים חדשים למטרת משימת הלמידה. נחלץ פיצ'רים חסרי חשיבות לזמן.

No Importance of features occurrence over time-Fixed number of Features

technical\_experience – בחרנו ליצור פיצ'ר בנארי המשלב את הדאטה הקיים בפצ'רים experience ו-major\_disciplin, פיצ'ר זה הינו ללא חשיבות לזמן כיוון שמספר שנות הניסיון או התחום ממנו העובד מגיע הינו סטטי. אנו רוצים להוסיף חשיבות לשילוב בין שנות הניסיון לתעסוקה בתחום ה-STEM. אנו מניחים כי אדם בעל ניסיון מעל 8 שנים אשר עבד בתחום טכנולוגי יהיה בעל סיכויים גבוהים לסיים את הקורס בהצלחה יתרה ולכן ככל הנראה יקבל אופציות תעסוקה רבות יותר ויגדל הסיכוי שיבחר לא לעבוד בחברה.

quality\_rate– בחרנו ליצור פיצ'ר שנותן ציון לכל עובד המעיד על "איכות העובד" הנקבעת לפי אינדקס הפיתוח של העיר בה הוא מתגורר ולרמת ההשכלה שלו. אדם התמגורר בעיר בפיתוח גבוה יקבל נקודה אחת, בעיר בפיתוח בינוני יקבל חצי נקודה ופיתוח נמוך לא יקבל נקודה כלל.

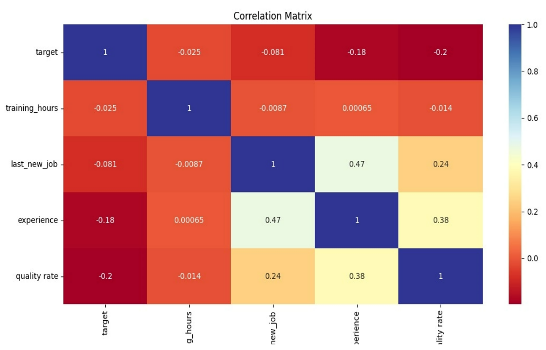
אדם בעל השכלה של עד 12 שנות לימוד יקבל חצי נקודה, אדם בעל תואר ראשון יקבל נקודה, אדם בעל תואר שני יקבל נקודה וחצי ואדם עם דוקטורט יקבל 2 נקודות.

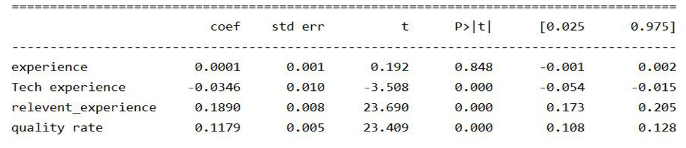
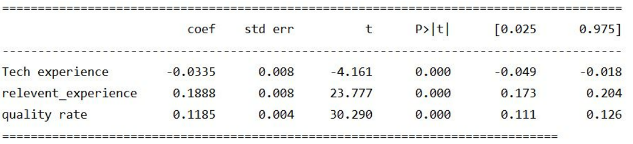
אנו מעריכים שככל שממד האיכות יהיה גובה יותר כך לעובד יפתחו אופציות תעסוקה רבות יותר ולכן גדלים הסיכויים שהוא יבחר שלא לעבוד בחברה בסוף ההכשרה.

## **Feature Representation**

בשלב זה נרצה לייצג את הפיצ'רים הקיימים לנו בדרך הטובה ביותר. על כן בחרנו לתקנן את המשתנה הרציף – training\_hours לסולם הערכים [0-1] לצורך מניעת הטיות בתהליך הלמידה של המודל ולקבל יחסיות עבור כל sample כאשר הערך הגבוה ביותר הינו 1 . השתמשנו בנרמול מינימום-מקסימום לפיצ'ר, על בסיס הנוסחה : .

## **Feature Selection**

בשלב זה נרצה להעניק לאלגוריתם הלמידה שלנו סט אינפורמטיבי וממצה של פיצ'רים, על כן נבצע בחירה של הפיצ'רים הרלוונטיים למודל. ראשית נסיר את הפיצ'רים – major\_discipline city\_development\_index, gender city, ו- enrollee\_id כיוון שאלו אינם רלוונטיים למשימת הלמידה והמידע הרלוונטי המופיע בהם חולץ בשלב הfeature extraction. השתמשנו בהנחה כי האינטראקציה בין המשתנים הרציפים למשתנים הקטגוריאליים היא זניחה ולכן בצענו פעולות נפרדות על המשתנים הרציפים ועל המשתנים הקטגוריאליים.

* ביצענו ויזואליזציה באמצעות תרשימים המציגים את הקשר של הפיצ'רים אל המשתנה המוסבר (ניתן לראות בסוף העבודה בנספח).
* Correlation Feature Selection- זו שיטה כמותית Filters Methods כלומר נדרג כל Feature בנפרד ללא קשר למודל הנבחר. נבחר את Features שהקורלציה שלהם עם משתנה המוסבר target יחסית גבוהה. עם זאת נרצה כי המתאם בין המשתנים יהיה נמוך כדי שלא נכניס משתנה למודל שלא תורם לנו מידע חדש. ניתן לראות כי הקורלציות עם המשתנים experience, quality\_rate יחסית גבוהה ולכן נבחר להכניס אותם למודל שלנו. לעומת זאת למשתנים last\_new\_job , training\_hours יש קורלציה יחסית נמוכה וקרובה ל-0 עם המשתנה המוסבר ולכן נבחר להוריד אותם מהמודל.
* על מנת לבחון את הערך של כל פיצ'ר בחרנו באסטרטגיית wrapper המוצאת את הקומבינציה הטובה ביותר של הפיצ'רים. הרצנו אלגוריתם **backward selection** כך שתחילה התחלנו עם כל הפיצ'רים הקיימים כעת(ללא קטגוריאלים)**.** בהרצה הראשונה גילינו כי experienceמשתנה לא מובהק ולכן הסרנו אותו מהמודל. בהרצה השנייה רואים כי כל המשתנים הנותרים מובהקים ולכן לא נסיר אותם.

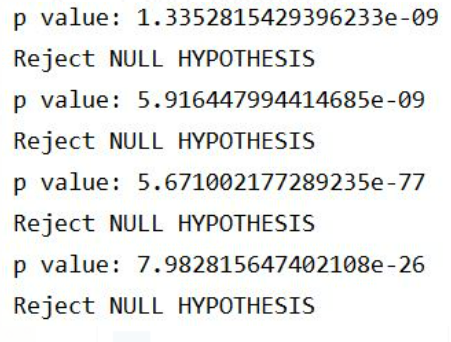
כדי לסנן בין ה Featuresהקטגוריים :

סדר ההרצה של המשתנים במבחן שעשינו היה: company\_size, company\_type, enrolled\_university, education\_level.

* Test of Independence . במבחן זה בודקים האם התלות הקיימת בין המשתנה המוסבר למשתנה הקטגוריאלי היא זניחה או לא. במבחן זה בודקים שתי השערות:

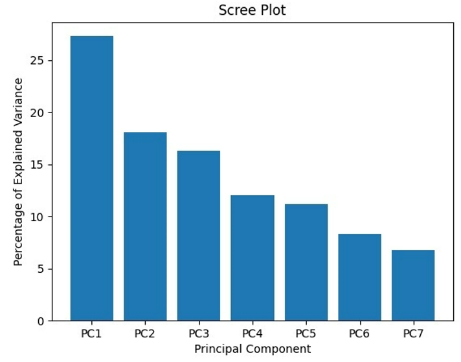
*Null Hypothesis (H0): There is no relationship between the variables*

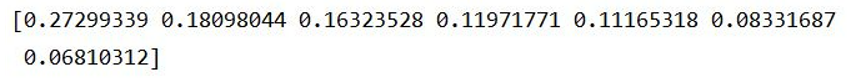
*Alternative Hypothesis (H1): There is a relationship between variables*



כאשר p-value=0.05. אם מקבלים את השערת האפס נשאיר את המשתנה במודל.

הבדיקה העלתה כי עבור כלל המשתנים הקטגוראלים שבדקנו דחינו את השערת ה-0 ולכן לא נוציא אף פיצ'ר נוסף מהמודל.

* **Dimensionality Reduction-** בשלב זה מפחיתים את המספר של ה Features כדי שנוכל ליצור Features set קטן יותר, פשוט יותר, שיחסוך זמן, ולהגדיל את ההבדלים בין ה class השונים. אחת השיטות לעשות זאת היא הPCA המורידה את הממד של ה Features אך שומרת את המידע החשוב והשונות. באמצעות שינוי מערכת הצירים תוך שמירה על השונות PCA בונה Features חדשים ומדרגת את אלו עם השונות הגבוהה ביותר, אותם נבחר למודל שלנו.

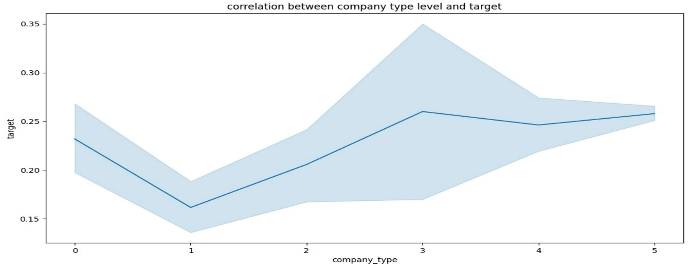
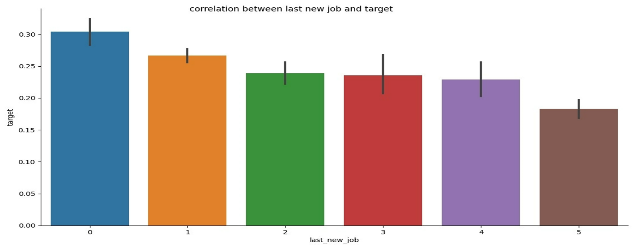
נבחר לקחת את pc1 - pc5 שלהם השונות הגבוהה ביותר, את קומפננטה 6 ו-7 נוציא מהמודל כי הן מהוות אחוז שולי משונות של הדאטה שלנו.ערכי השוניות של הPCA:

# **Model Training**

שיטת הוולידציה בה בחרנו היא שיטת **k-fold**, במסגרתה נחלק את סט הנתונים לK חלקים, על k-1 חלקים נבצע את תהליך אימון המודל ועל החלק הנותר נבצע תהליך הודליציה, נחזור על תהליך זה K פעמים (בכל פעם נוציא חלק אחר ונבצע את התהליך על כלל הקומבינציות האפשריות), לבסוף נחשב ממוצע עבור כל נתוני הולידציה. מטרות הוולידציה הן: הערכת שגיאת המודל, סיוע בבחירת מודל וביצוע אופטימיזציה של פרמטרי המודל וכן לאפשר אבחון מוקדם של שגיאות במודל.

שיטה זו עדיפה על שיטת hold out כיוון ששיטה זו מפחיתה את שונות ביצועי המודל ועדיפה על שיטת leave one out כיוון שסט הנתונים הינו גדול, מגוון וכולל מספר samples עבור אותה ישות.

נבחר לחלק את סט הנתונים ל100 חלקים כך שבכל מקטע יהיו כ-152 samples , באיטרציה הראשונה נקצה את המקטע הראשון לוולידציה, ואת יתר המקטעים (89 מקטעים) נקצה לאימון המודל. באיטרציה הבאה נקצה את המקטע השני לוולידציה והיתר לאימון המודל, וכן הלאה עד שנסיים להשתמש בכל המקטעים שיצרנו לוולידציה. לאחר מכן נחשב ממוצע עבור כלל מקטעי הולידציה , סה"כ 90 מקטעים.

נספח קורולציות- תרשימים המציגים את הקשר של הפיצ'רים אל המשתנה המוסבר:

