



**חלק א': אוסף נתונים**

<https://www.kaggle.com/chirag9073/hr-analysis-prediction-and-visualisation/data>

אוסף הנתונים מתאר עובדי IBM, כפי שניתן לראות בהמשך, חילקנו את הנתונים ל -2 טבלאות כאשר הגורם המקשר הינו מספר העובד כאשר טבלה אחת בעלת הפרטים האישיים של העובד, והטבלה השנייה בעלת מאפיינים הקשורים לעבודה. כמו כן, רידדנו את העמודות שלא רלוונטיות לשאלות המחקר שלנו.

**חלק ב': שאלות המחקר**

שאלת מחקר 1 (supervised):

* האם התכונות הבאות: גיל, מרחק מהבית, השכלה, מגדר ומצב משפחתי -בעלות השפעה על איכות חיים-עבודה גבוה (מעל 2) ?

(השאלה ממוקדת ועוסקת מאפיינים ספציפיים, כמו כן התוצאה מדידה וברת השגה משום שמושתתת על נתונים קיימים).

* KPI’S:

מספר הגברים ביחס לנשים בחברה.

מצב משפחתי- השכיח ביותר (מסווג ל3)

מספר שנות השכלה לעובד: השכיח ביותר- מעל 3 נחשב גבוה (מסווג ל4)

מספר העובדים שגרים רחוק: התפלגות מגורי העובדים (כאשר רחוק מוגדר להיות מעל 15 ק"מ).

מספר העובדים בעלי WORK LIFE BALANCE גבוה (מדד נתון 1-4, מעל 2 יחשב גבוה).

WORK LIFE BALANCE השכיח ביותר בחלוקה למגדר (מדד נתון 1-4).

* SMART- כל המדדים שבחרנו ספציפים ומודדים פרמטר ברור (מספר עובדים, אחוז עובדים ביחס לפרמטר קיים אחר), כמו כן, התוצאות מדידות (מספר,אחוז) וברות השגה משום שמושתתות על נתונים קיימים, המדדים בוחנים תחומים בעלי השפעה ישירה על שאלת המחקר (מספר עובדים, נתוני מגדר מרחק וכו) וניתנים לביצוע ובדיקה במהלך תחום הזמן הקיים (עד סוף סמסטר ב').

שאלת מחקר 2 (unsupervised):

* מה ניתן ללמוד על תכונות נתוני ההעסקה - Department,JobLevel,JobRole,MonthlyIncome

של העובדים ביחס למגדר שלהם.

* KPI’S:

מספר הנשים בחברה.

מספר הגברים בחברה.

אחוז העובדים בעלי הכנסה מעל לממוצע.

אחוז העובדים בעלי JobLevel מעל לממוצע.

* SMART- כל המדדים שבחרנו ספציפים ומודדים פרמטר ברור (מספר עובדים, אחוז עובדים ביחס לפרמטר קיים אחר), כמו כן, התוצאות מדידות (מספר,אחוז) וברות השגה משום שמושתתות על נתונים קיימים, המדדים בוחנים תחומים בעלי השפעה ישירה על שאלת המחקר (מספר עבדים, נתוני מגדר ועבודה נוספים מתוך הנתונים) וניתנים לביצוע ובדיקה במהלך תחום הזמן הקיים (עד סוף סמסטר ב').

**חלק ג': הבנת הנתונים**

1. מדדי פיזור של התכונות

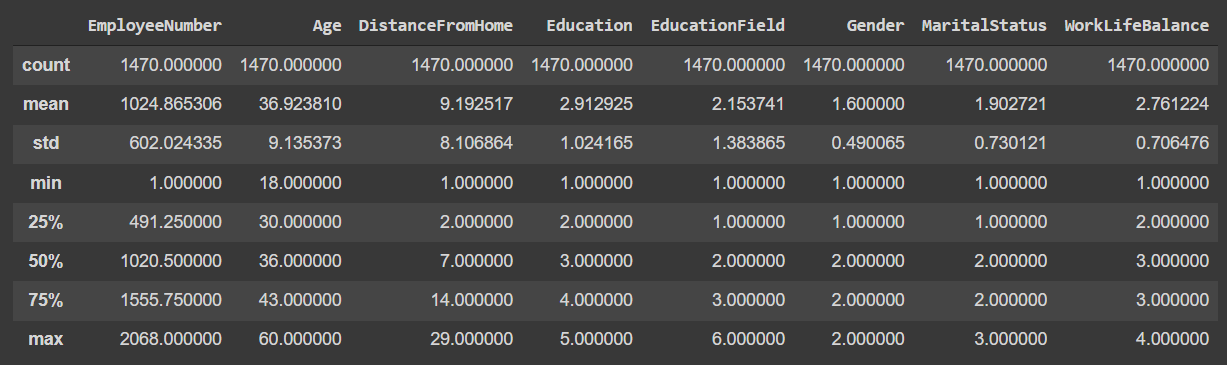
* טבלה 1: "PersonalData"

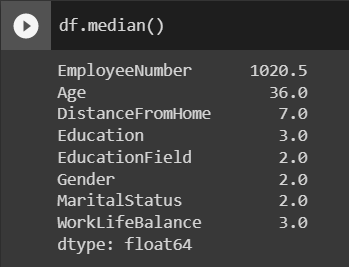
המרת משתנים לנומריים:











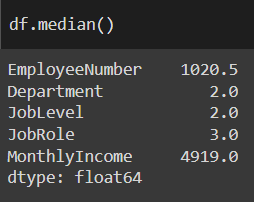
* טבלה 2: "WorkData"



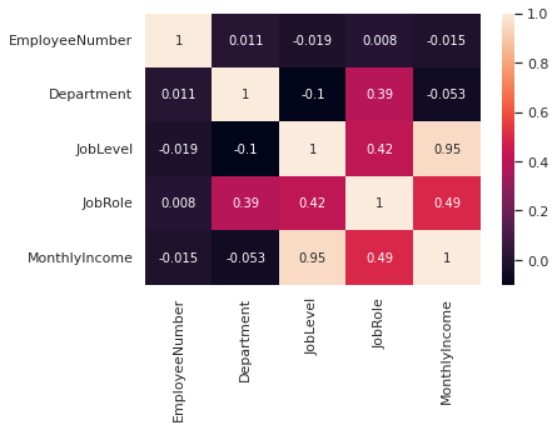


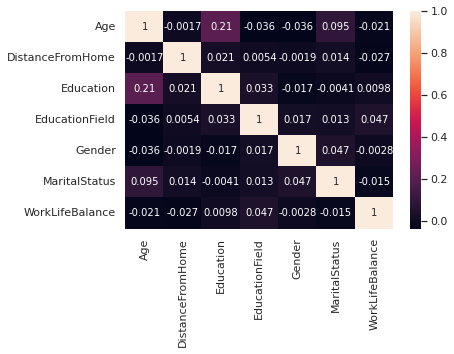




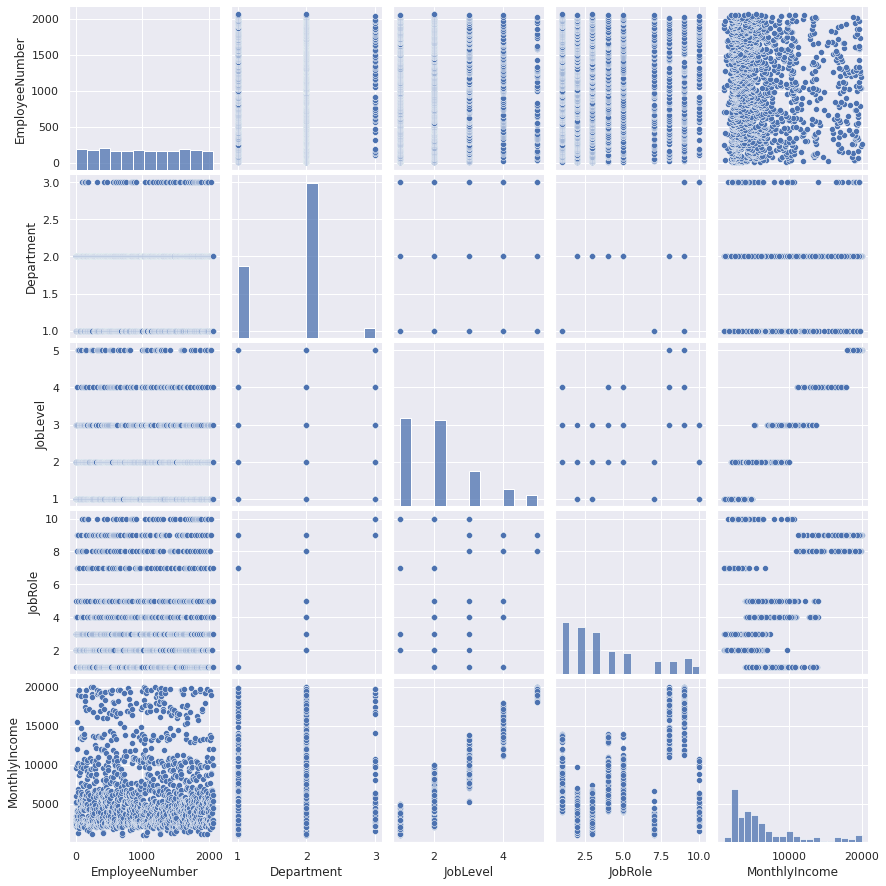


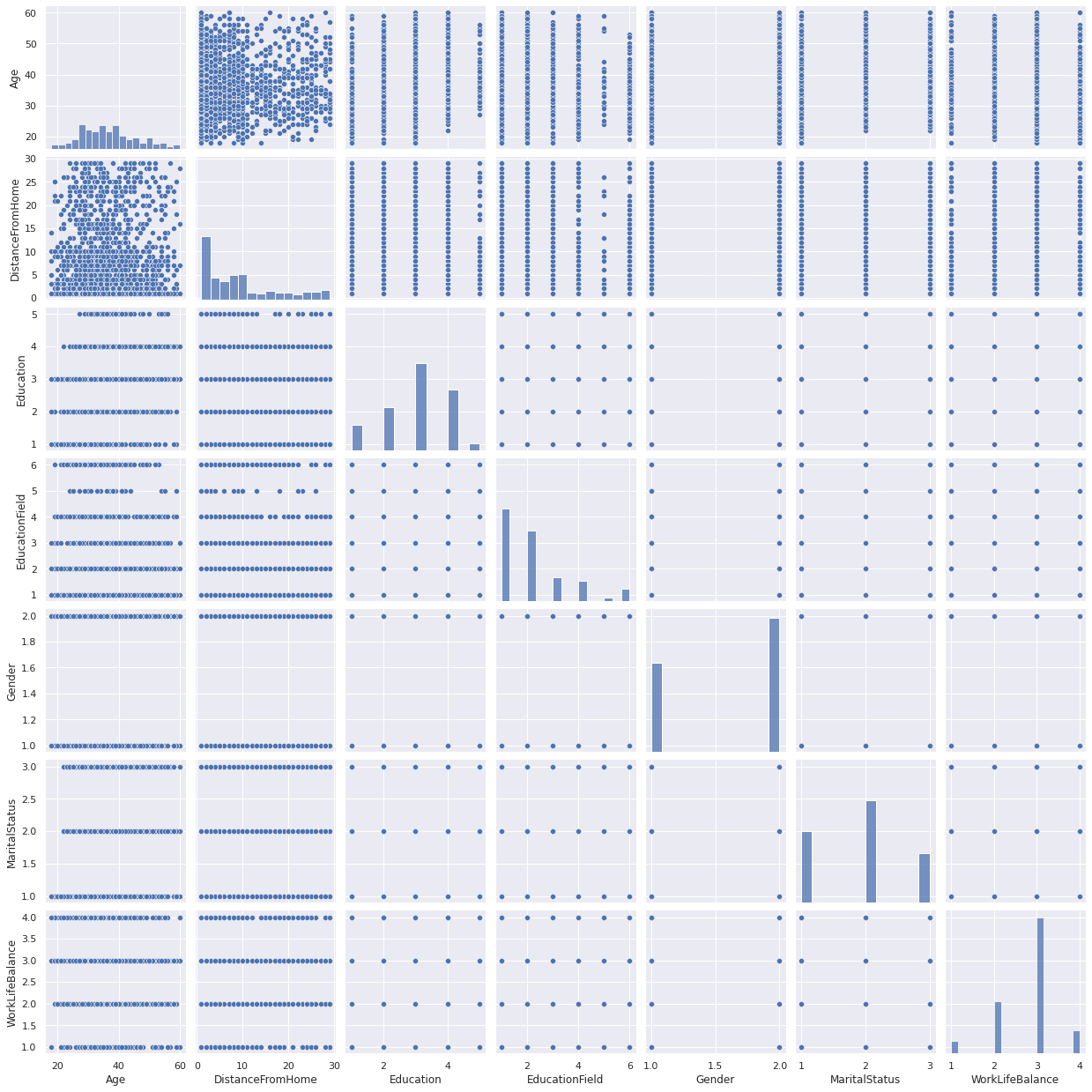
1. תלויות וקשרים (קורלציה)





1. כלי מדד סטטיסטי נוסף:





1. אנטרופיה עבור כל אחת מהתכונות לשאלת המחקר הראשונה (supervised):

* אנטרופיה של "WORKLIFE BALANCE" היא: 1.49
* אנטרופיה עבור מרחק מהעבודה: 4.356
* אנטרופיה עבור גיל: 5.14
* אנטרופיה של מגדר: 0.97
* אנטרופיה של מצב משפחתי: 1.524
* אנטרופיה של השכלה: 2.018

1. עבור 2 התכונות בעלות האנטרופיה הנמוכה ביותר- GINI+INFO GAIN:

* GINI של מגדר הוא:0.48
* GINI של מצב משפחתי: 0.63
* INFO GAIN של מגדר: 6.096691859069914e-05
* INFO GAIN של מצב משפחתי: 0.0017

1. מסקנות מתחקור הנתונים:

* הנתונים מצביעים על אי וודאות רבה.
* קיימת קורלציה נמוכה בין הכנסה חודשית למחלקה בה העובד מועסק וקורלציה גבוהה בין ההכנסה החודשית לתפקיד ולדרגה.
* ניתן לראות מהניתוח הסטטיסטי שיש יותר עובדים בעלי הכנסה חודשית נמוכה יחסית ומעט בעלי הכנסה חודשית גבוהה יחסית.
* רוב העובדים בעלי תואר במדעי החיים.
* רוב העובדים בעלי שלוש שנות השכלה גבוהה.
* רוב העובדים עובדים קרוב לבית.
* מרבית העובדים הם גברים.
* מרבית העובדים נשואים.