בינה עסקית – פרויקט

חלק 1

מגישות

ורוניקה פרידמן

לי קישון

סמסטר ב, 2022

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:

```
newdic = {'Life Sciences':1,'Medical':2,'Marketing':3,'Technical Degree':4, 'Human Resources':5,'Other':6}
```

```
newdic = {'Single':1,'Married':2,'Divorced':3}
```

```
newdic = {'Female':1,'Male':2}
```

	EmployeeNumber	Age	DistanceFromHome	Education	EducationField	Gender	MaritalStatus	WorkLifeBalance
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
mean	1024.865306	36.923810	9.192517	2.912925	2.153741	1.600000	1.902721	2.761224
std	602.024335	9.135373	8.106864	1.024165	1.383865	0.490065	0.730121	0.706476
min	1.000000	18.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	491.250000	30.000000	2.000000	2.000000	1.000000	1.000000	1.000000	2.000000
50%	1020.500000	36.000000	7.000000	3.000000	2.000000	2.000000	2.000000	3.000000
75%	1555.750000	43.000000	14.000000	4.000000	3.000000	2.000000	2.000000	3.000000
max	2068.000000	60.000000	29.000000	5.000000	6.000000	2.000000	3.000000	4.000000

```
df.median()
```

```
EmployeeNumber       1020.5
Age                    36.0
DistanceFromHome        7.0
Education               3.0
EducationField          2.0
Gender                  2.0
MaritalStatus           2.0
WorkLifeBalance         3.0
dtype: float64
```

- טבלה 2: "WorkData"

```
newdic = {'Sales':1,'Research & Development':2,'Human Resources':3}
```

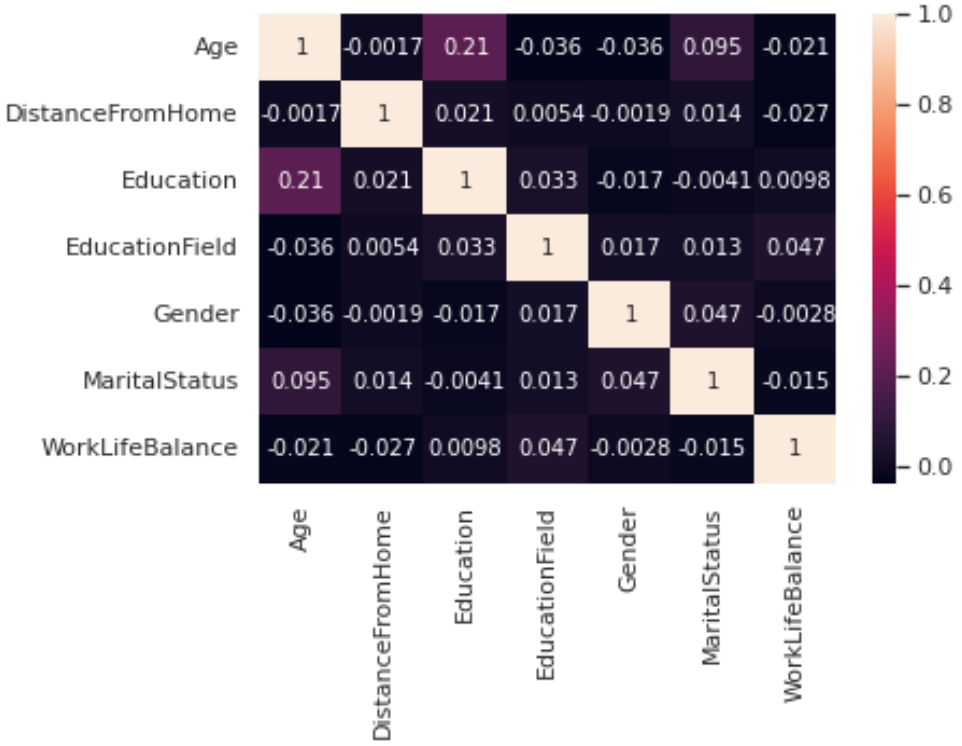
```
newdic = {'Sales Executive':1,'Research Scientist':2,'Laboratory Technician':3,'Manufacturing Director':4, 'Healthcare Representative':5,
'Manager':6, 'Sales Representative':7, 'Research Director':8,'Manager':9, 'Human Resources':10}
```

	EmployeeNumber	Department	JobLevel	JobRole	MonthlyIncome
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
mean	1024.865306	1.739456	2.063946	3.796599	6502.931293
std	602.024335	0.527792	1.106940	2.721493	4707.956783
min	1.000000	1.000000	1.000000	1.000000	1009.000000
25%	491.250000	1.000000	1.000000	2.000000	2911.000000
50%	1020.500000	2.000000	2.000000	3.000000	4919.000000
75%	1555.750000	2.000000	3.000000	5.000000	8379.000000
max	2068.000000	3.000000	5.000000	10.000000	19999.000000

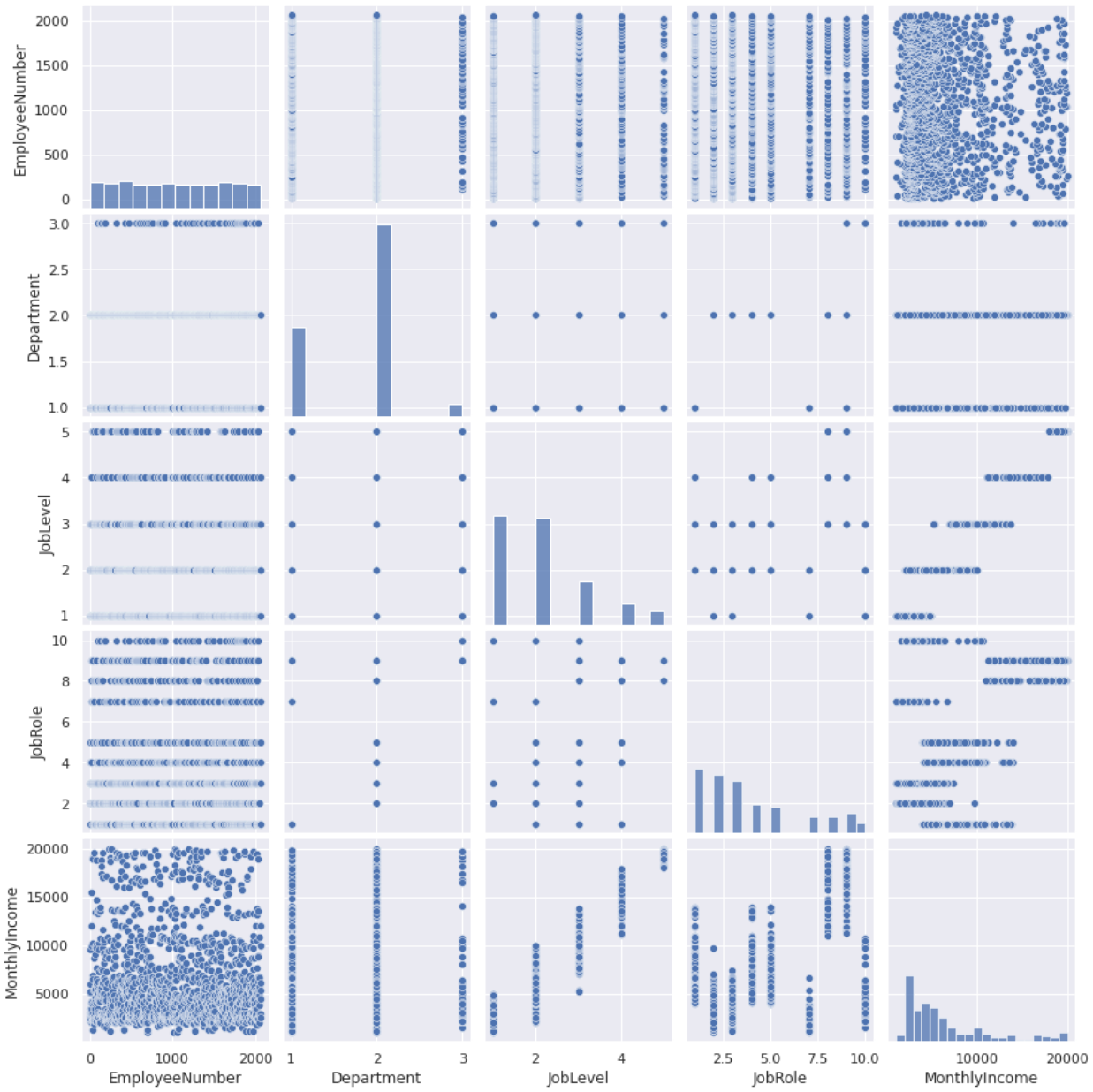
```
df.median()
```

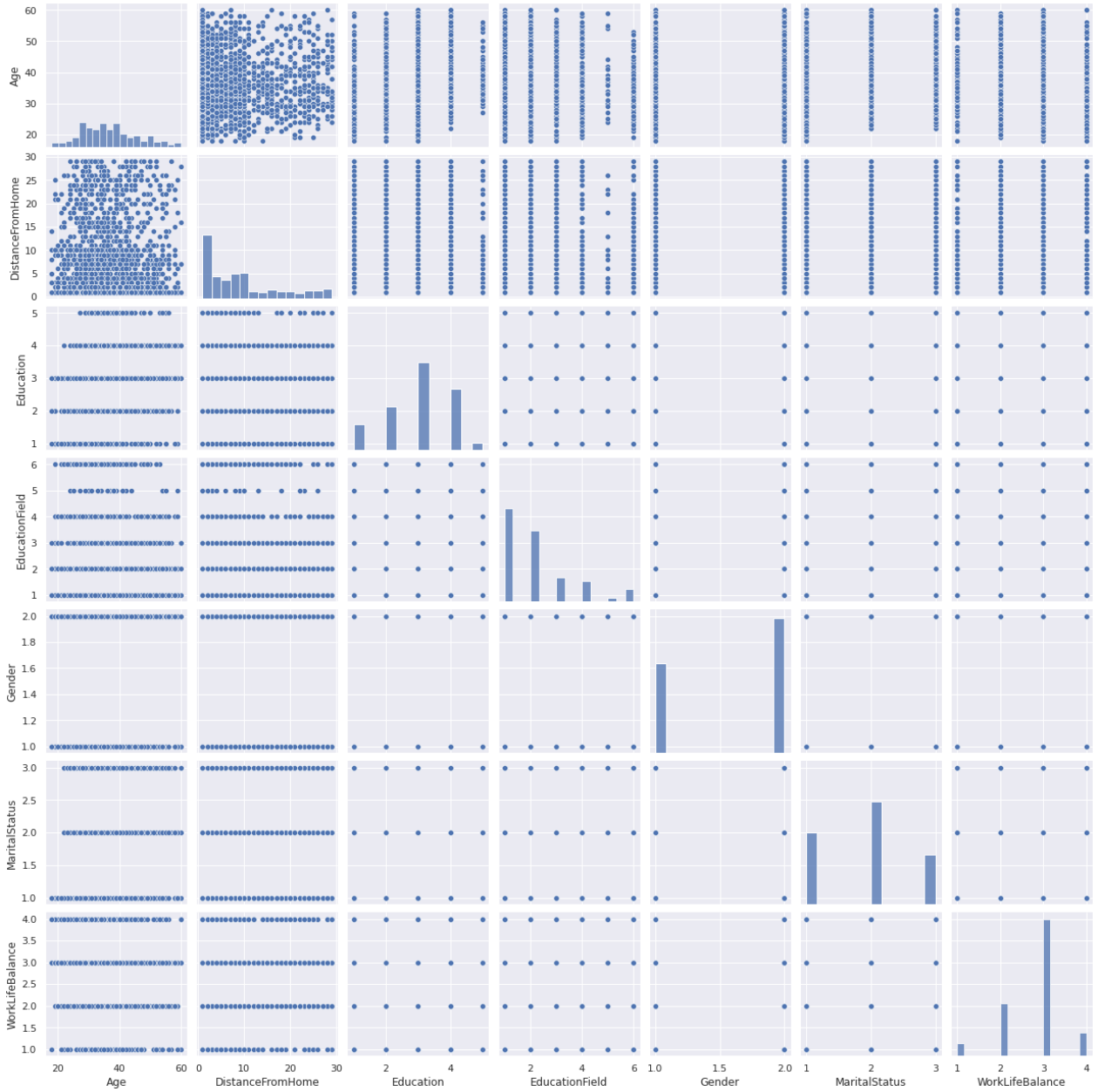
```
EmployeeNumber    1020.5
Department           2.0
JobLevel             2.0
JobRole              3.0
MonthlyIncome     4919.0
dtype: float64
```

2. <u>תלויות וקשרים (קורלציה)</u>



3. <u>כלי מדד סטטיסטי נוסף:</u>





4. <u>אנטרופיה עבור כל אחת מהתכונות לשאלת המחקר הראשונה (supervised):</u>

- 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

5. <u>עבור 2 התכונות בעלות האנטרופיה הנמוכה ביותר- GINI+INFO GAIN:</u>

- 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

6. <u>מסקנות מתחקור הנתונים:</u>

- 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.