



ЖИВІ ЛЮДИ

МТС-21, 2023



ПРОЩО ДАТАСЕТ

Датасет містить інформацію про працівників/-ниць однієї американської компанії.



ЧОМУ ВІН?

- містить багато ознак, що формують цілісне уявлення про фактори впливу на робочий процес та його ефективність
- дає змогу зрозуміти, яким чином можна краще взаємодіяти з різними демографічними сегментами
- показує, що найбільше впливає на роботу та ефективність роботи людей різного віку
- може слугувати базою для подальшого формування стратегій підвищення ефективності роботи, формування планів індивідуального розвитку та формування стратегій зниження кількості втрачених кадрів

- [D](int) Age - вік кожного працівника (роки)
- [D](str) Attrition - стан співробітника у компанії (покинув/не покинув)
- [D](str) BusinessTravel - частота подорожей (Travel_Frequently/Travel_rarely/non-travel)
- [D](str) Department - відділення, у якому працює співробітник (research & Development/Sales/human resources)
- [N](int) DistanceFromHome - відстань від дому робітника до його місця праці (милі)
- [D](int) Education - кількість здобутих рівнів навчання (1-5)
- [D](str) EducationField - освіта (life sciences/medical/marketing/technical degree/Human Resources/other)
- [D](int) EnvironmentSatisfaction - шкала, що показує на скільки працівник задоволений своїм місцем праці (1-4)
- [D](str) Gender - стать (Female/Male)
- [D](int) JobInvolvement - рівень залученості працівників у роботу(1-5)
- [D](int) JobLevel - рівень (посада) працівників (1-5)
- [D](str) JobRole - професія (Sales Executive/Research Scientist/Laboratory Technician/Manufacturing Director/Healthcare Representative/Maneger/Sales Representative/Research Director/Human Resources)
- [D](int) JobSatisfaction - шкала, що показує на скільки наш працівник задоволений своєю роботою (1-4)
- [D](str) MaritalStatus - сімейний статус (Single/Married/Divorced)
- [N](int) MonthlyIncome - зарплата працівника (у доларах)
- [D](int) PerformanceRating - рейтинг продуктивності (3-4)
- [D](int) RelationshipSatisfaction - шкала, що показує на скільки працівник задоволений власним колективом (1-4)
- [N](int) TotalWorkingYears - кількість років роботи загалом (роки)
- [D](int) WorkingLifeBalance - шкала, що показує баланс життя-робота (1-4)
- [N](int) YearsAtCompany - кількість років роботи у компанії (роки)
- [N](int) YearsInCurrentRole - скільки років людина працює на останній посаді (роки)

[D] - дискретне значення
[N] - не дискретне значення
(INT) - числове значення
(STR) - текстове значення

Цільова змінна
MonthlyIncome

ГОТУЄМО ДО РОБОТИ

- Видаляємо непотрібні стовпці для того, щоб не було перенавчання моделі, оскільки датасет містить значення, що повторюються.
- Підраховуємо кількість відсутніх значень (NaN або null значень).

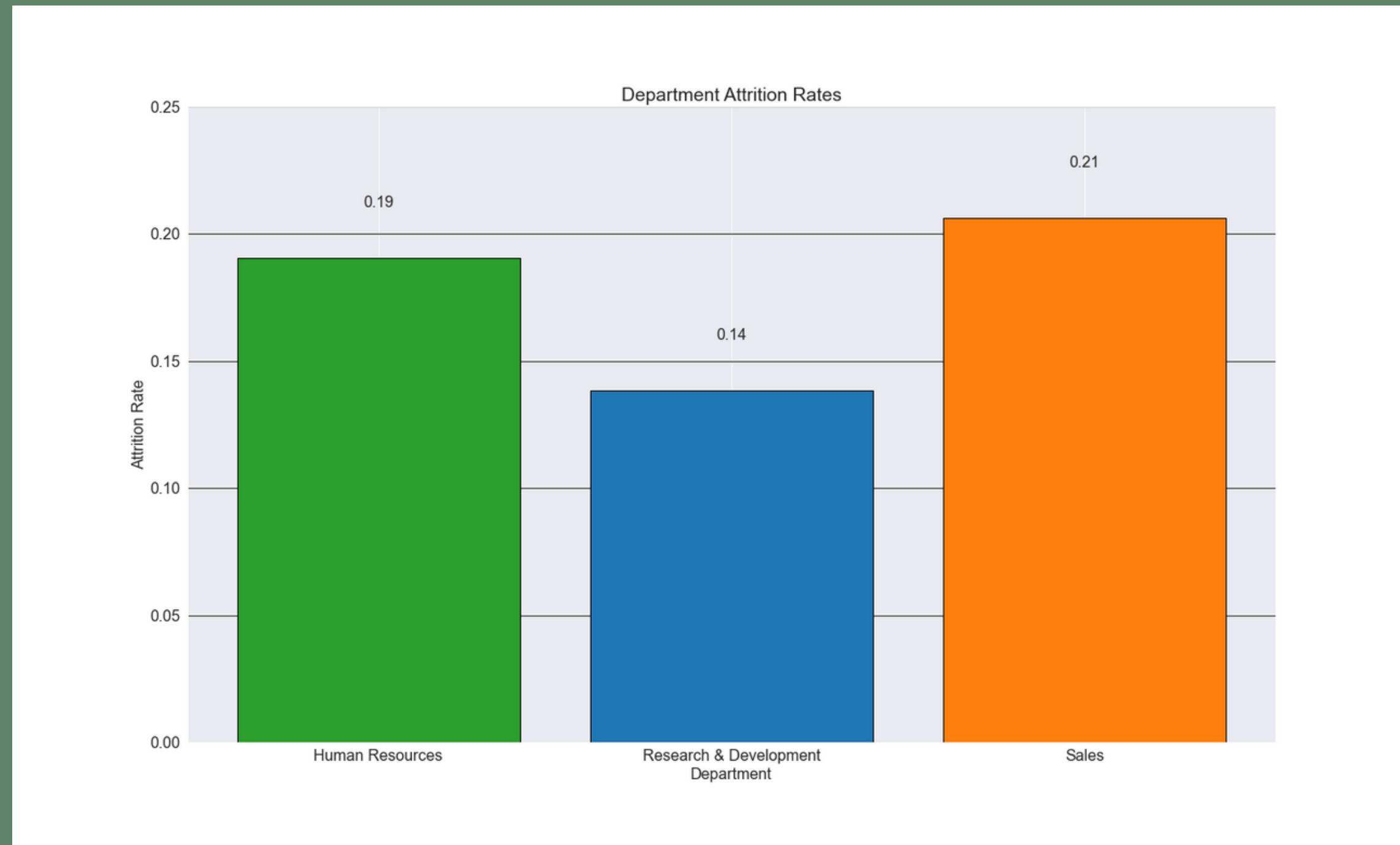
df.isna().sum()	
Age	0
Attrition	0
BusinessTravel	0
Department	0
DistanceFromHome	0
Education	0
EducationField	0
EnvironmentSatisfaction	0
Gender	0
JobInvolvement	0
JobLevel	0
JobRole	0
JobSatisfaction	0
MaritalStatus	0
MonthlyIncome	0
PerformanceRating	0
RelationshipSatisfaction	0
TotalWorkingYears	0
WorkLifeBalance	0
YearsAtCompany	0
YearsInCurrentRole	0
dtype: int64	



ВІЗУАЛІЗАЦІЯ ДАНИХ



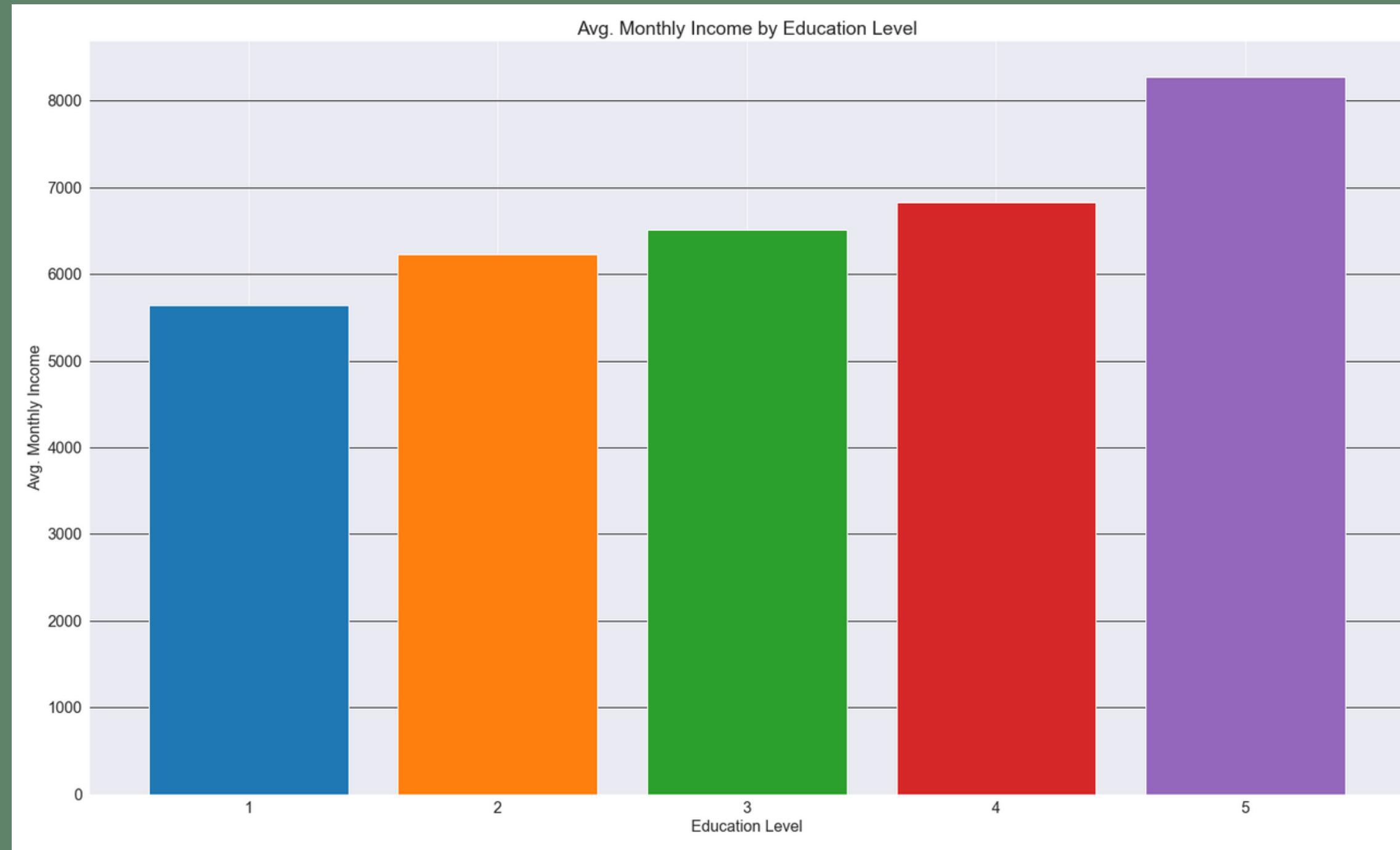
% людей, які покинули роботу з кожного відділення



Порівняльна гістограма, що інформує про кількість людей, які пішли з компанії у кожному з трьох відділів. Бачимо, що найбільше людей звільнилися та покинули відділ продажів (21%).

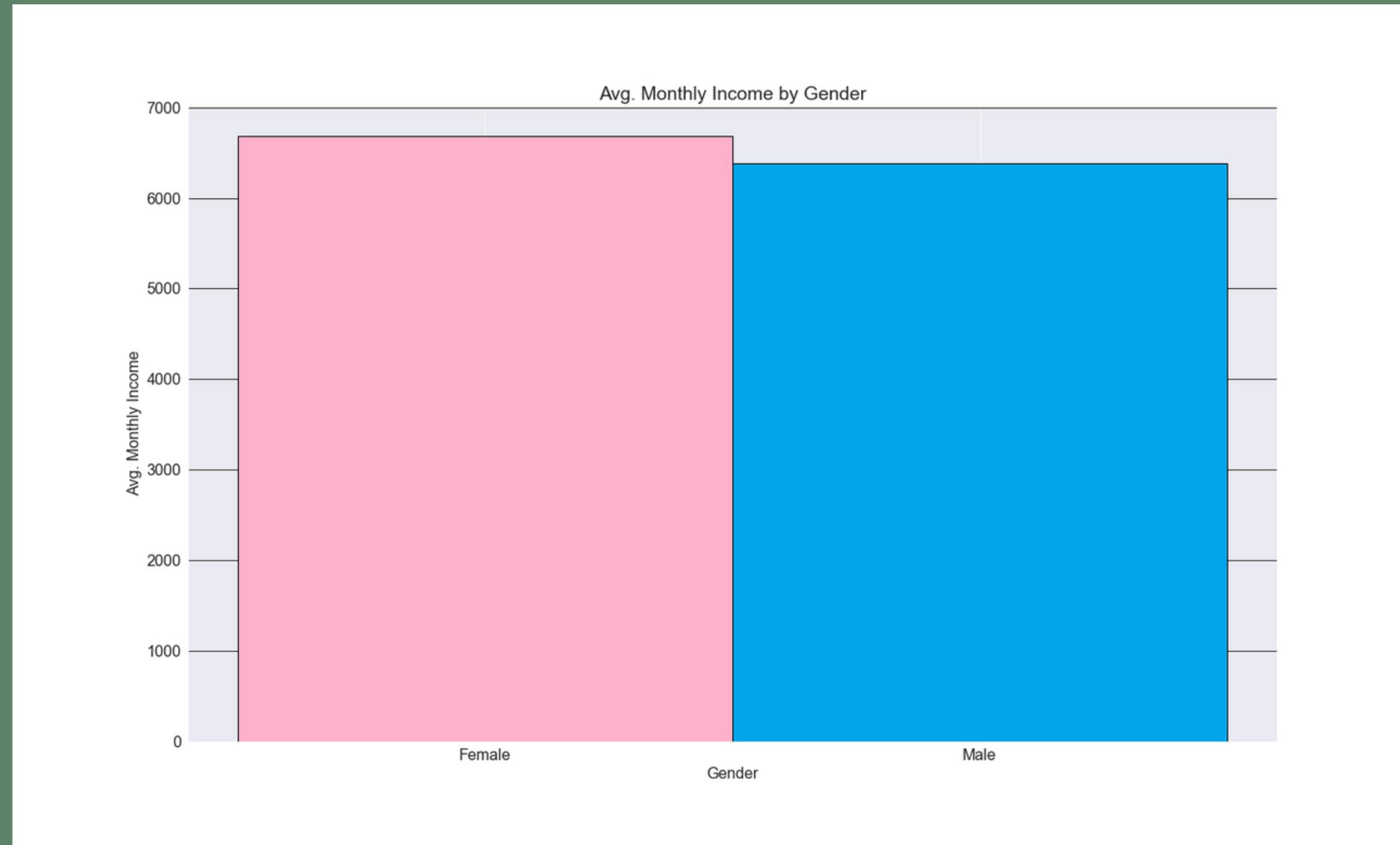
припускаємо, що це може бути пов'язано з стресами та вигоранням через комунікацію з великою кількістю людей

залежність сердньої зарплати від рівня освіти



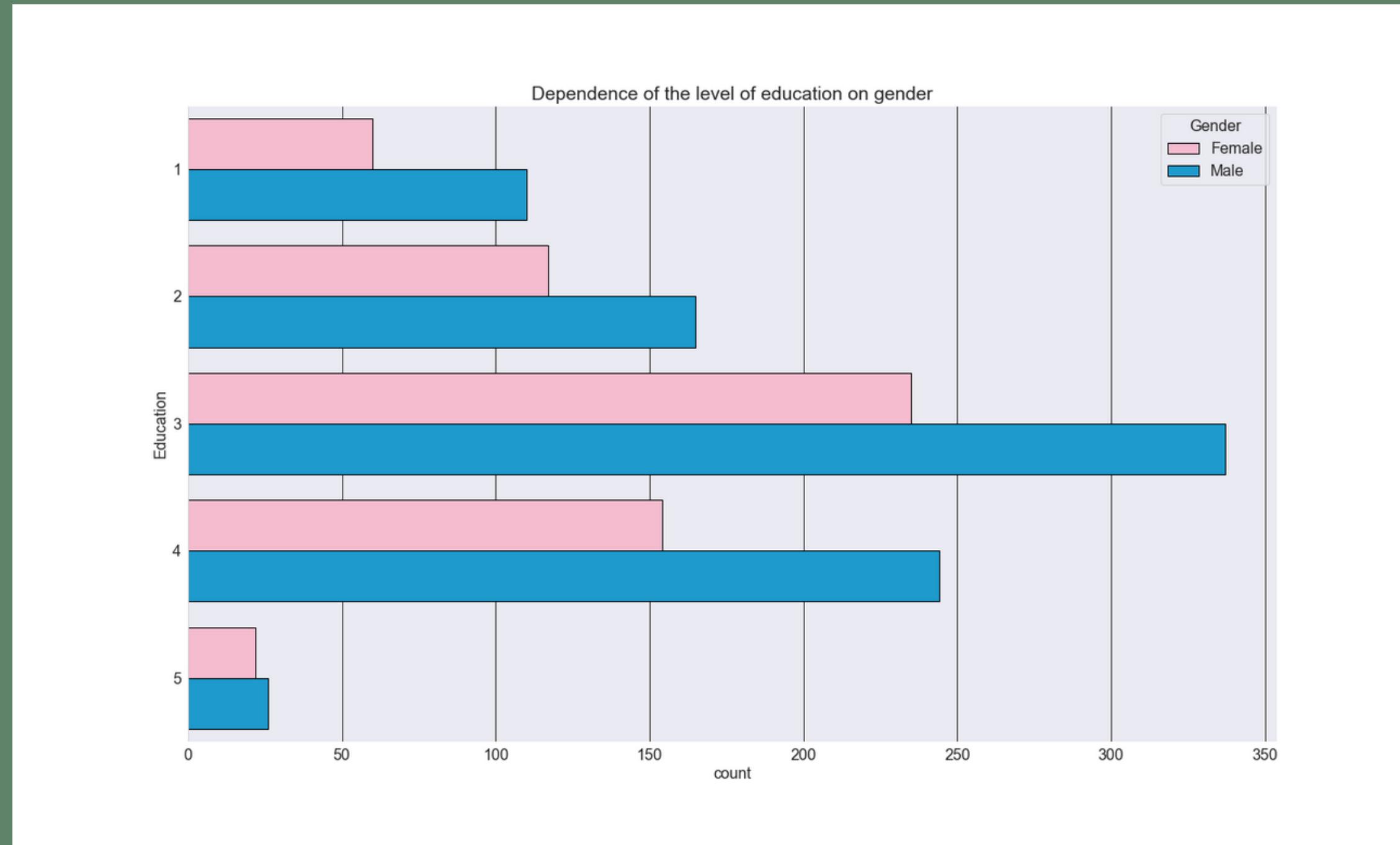
Графік, що показує залежність місячного доходу від рівня освіти. Бачимо, що разом з рівнем освіти росте і дохід. між другим, третім та четвертим рівнями спостерігаємо меншу різницю, аніж між першим та другим/четвертим та п'ятим.

залежність сердньої зарплати від статі



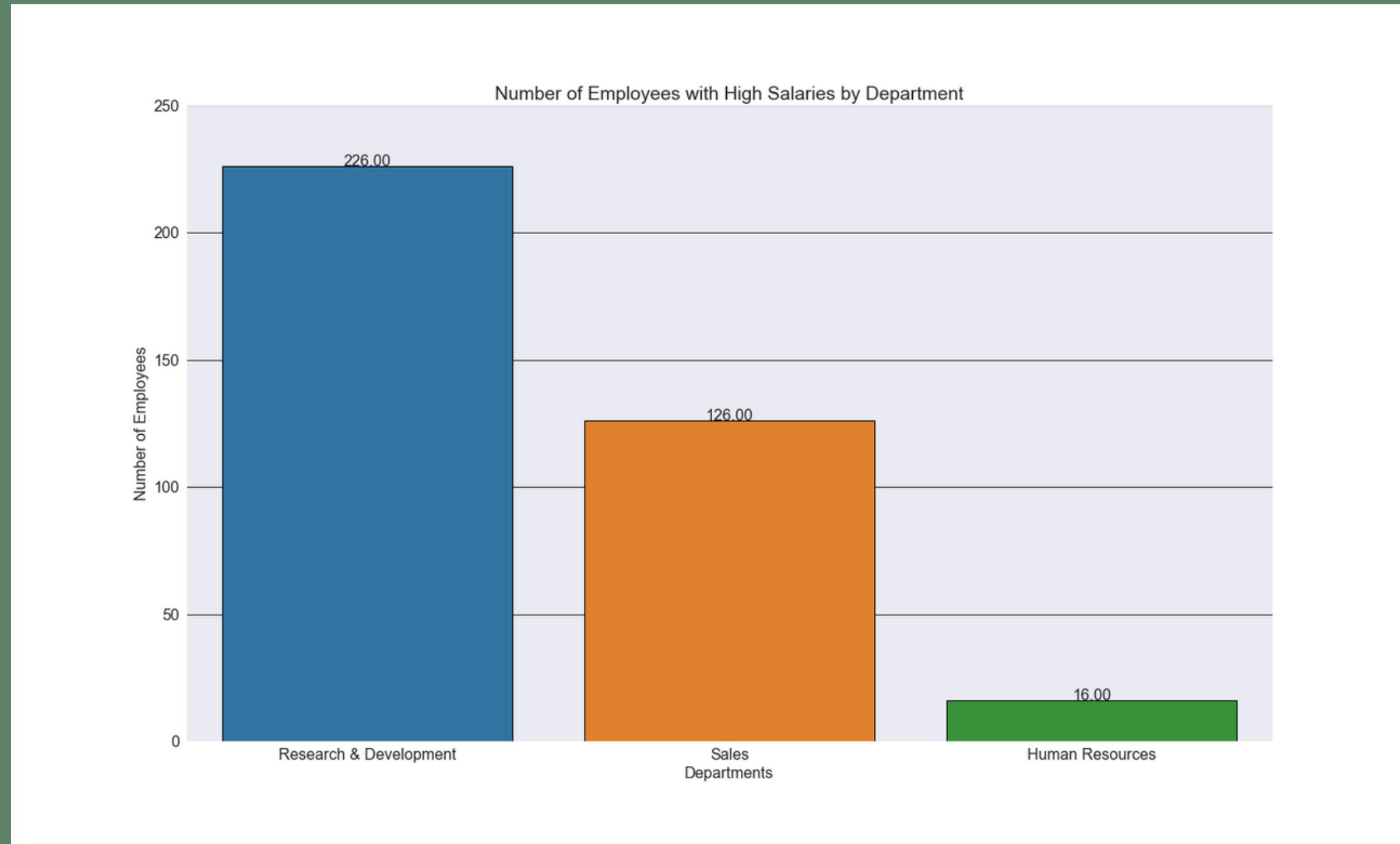
Графік, що показує залежність місячного доходу від статі. Бачимо, що середній дохід осіб жіночої статі вищий, ніж дохід чоловічої статі.

залежність рівня навчання від статі



Графік, що відображає залежність між рівнем навчання та статтю. Бачимо, що в середньому більшість осіб чоловічої та жіночої статі мають третій рівень навчання, найменше — п'ятий. Також спостерігаємо пропорційний розподіл осіб чоловічої та жіночої статі на кожному з рівнів.

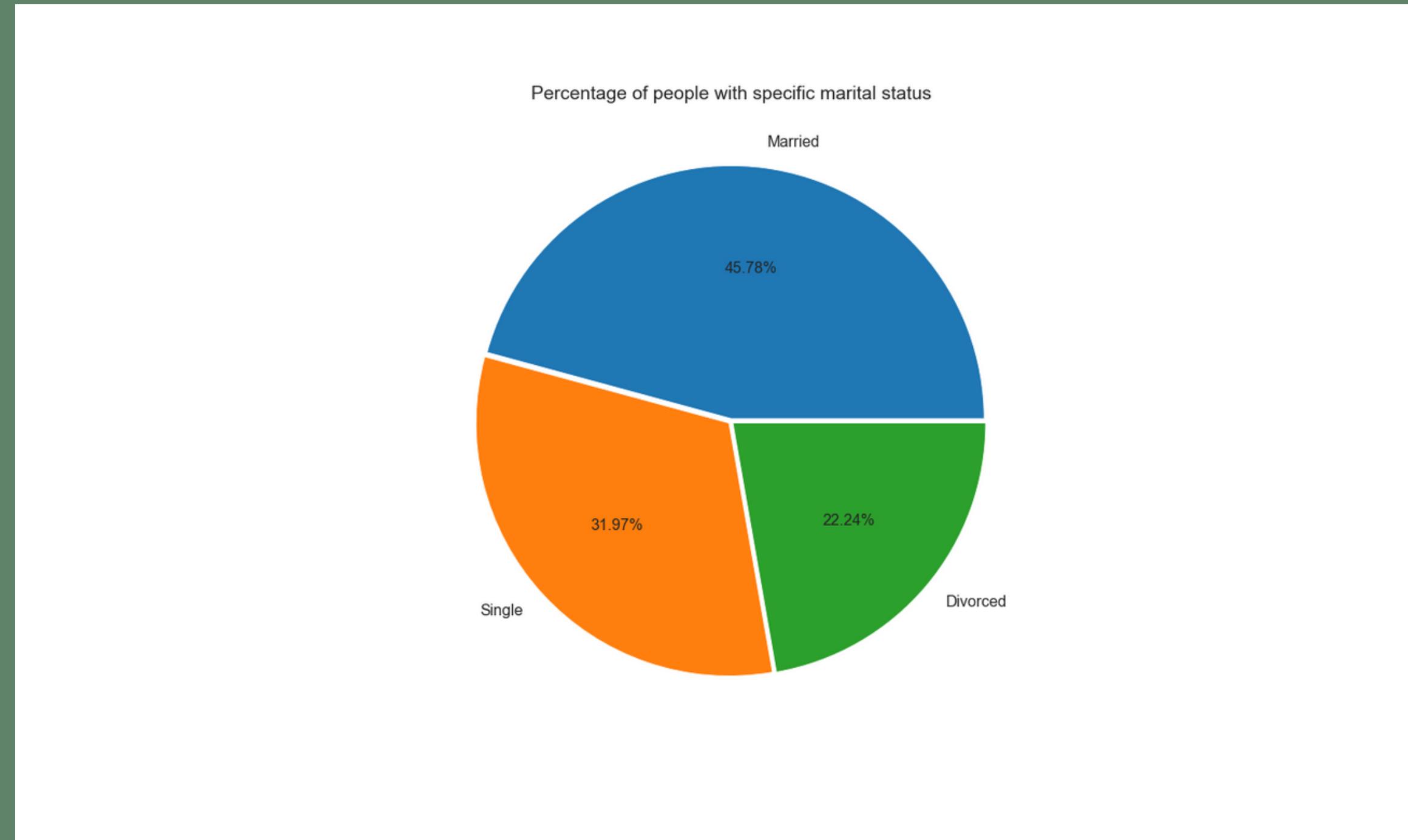
Кількість працівників з високою зарплатою у відділеннях



Порівняльна гістограма, що інформує про кількість людей з високим місячним доходом у трьох відділеннях з датасету (Research&Development, Sales Department, Human Resources). Бачимо, що серед вибірки з 250 людей 226 у RD, 126 у SD та 16 у HR мають високий місячний дохід.

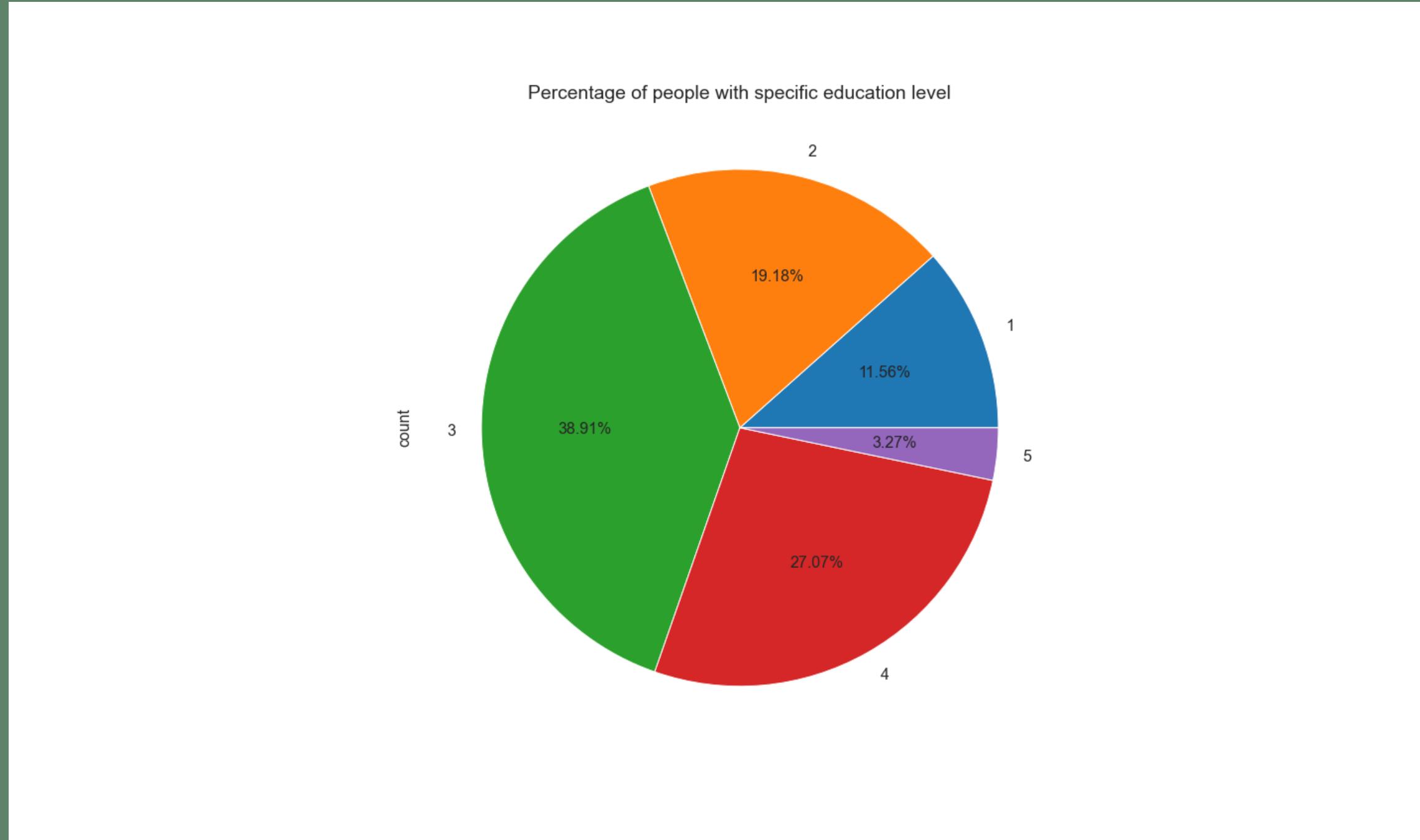
за високий місячний
дохід беремо
квантиль "75%" в
методі *describe()*.

Сімейний статус працівників



Кругова діаграма, що ілюструє відсоткове співвідношення кількості працівників/-ць відповідно до сімейного статусу. Бачимо, що 45.78% працівників/-ць є одруженими, 31.97% — не перебувають у стосунках, 22.24% — розлучені.

рівень освіти працівників



Кругова діаграма, що ілюструє відсоткове співвідношення кількості працівників/-ць з одним із п'яти рівнів освіти.

Бачимо, що перший рівень освіти має 11.56%, другий — 19.18%, третій — 38.91% (найбільше), четвертий — 27.07%, п'ятий — 3.27% (найменше).

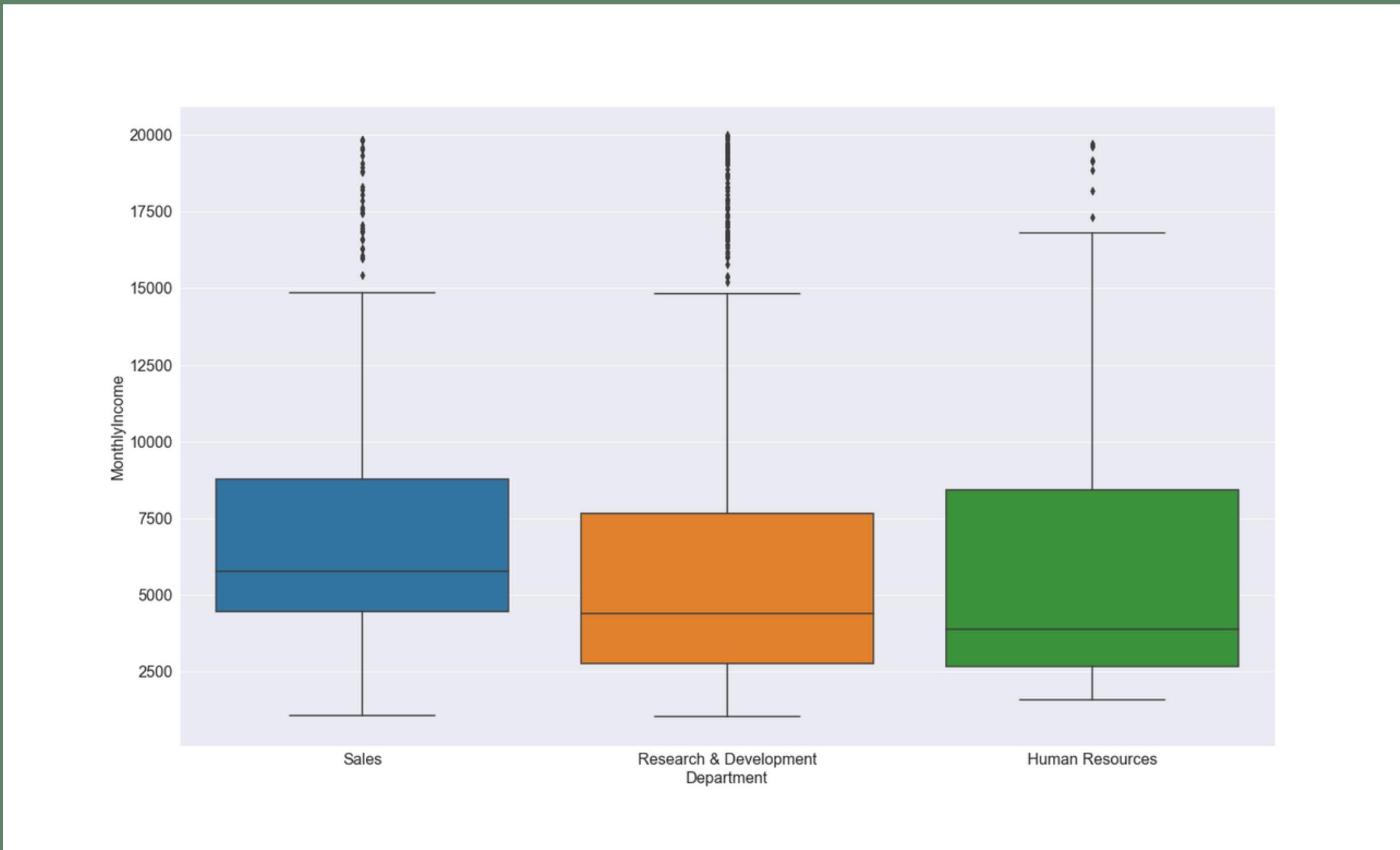
Вік і стать працівників



Діаграма розмаху, що ілюструє вік та стать працівників/-ць. Бачимо, що:

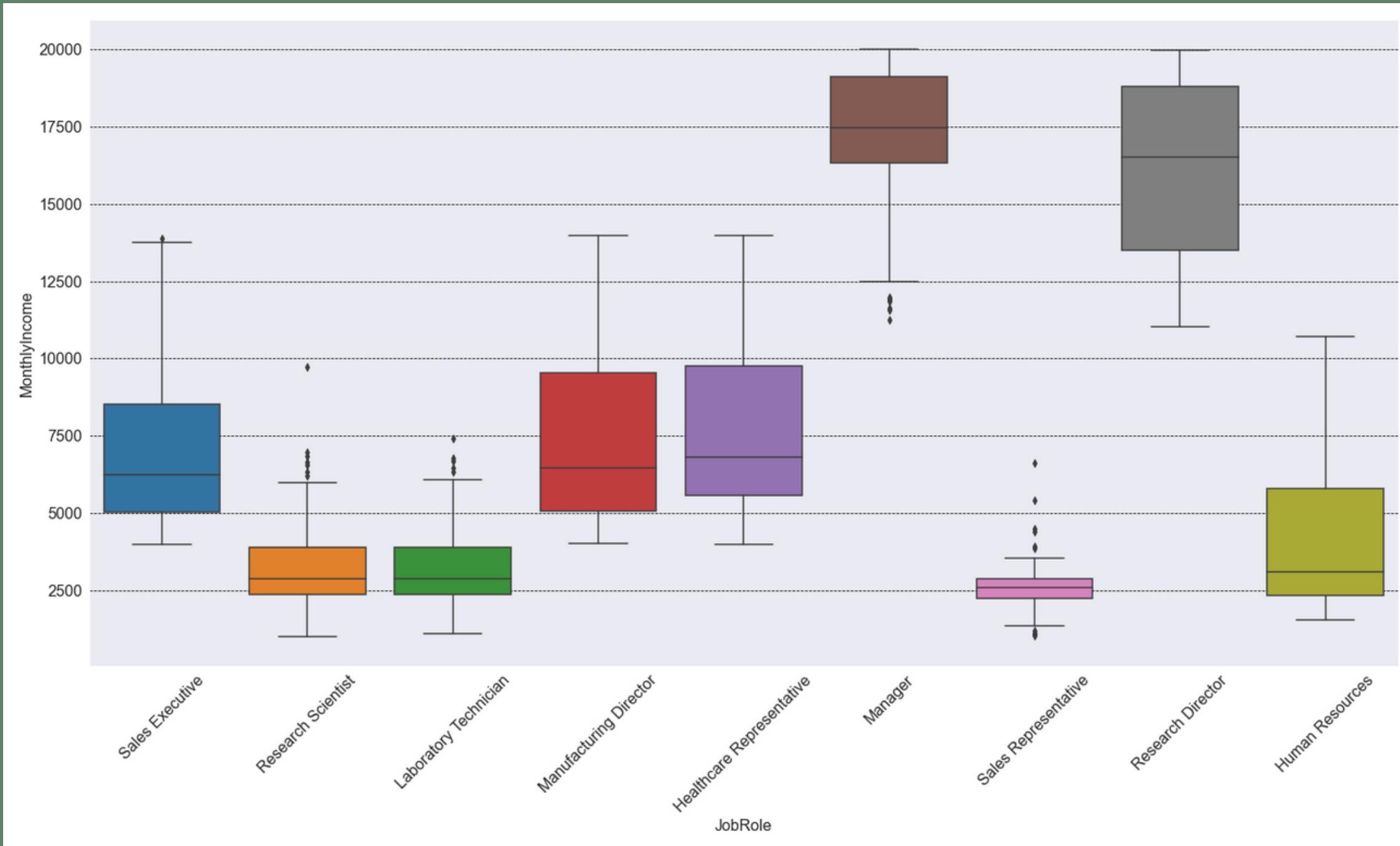
- загалом віковий діапазон у жінок більший за віковий діапазон у чоловіків.
- середній вік у чоловіків трохи нижчий за середній вік у жінок.
- у компанії мінімальний вік — вік переважно чоловіків, а максимальний — вік переважно жінок.

залежність зарплати від відділення праці



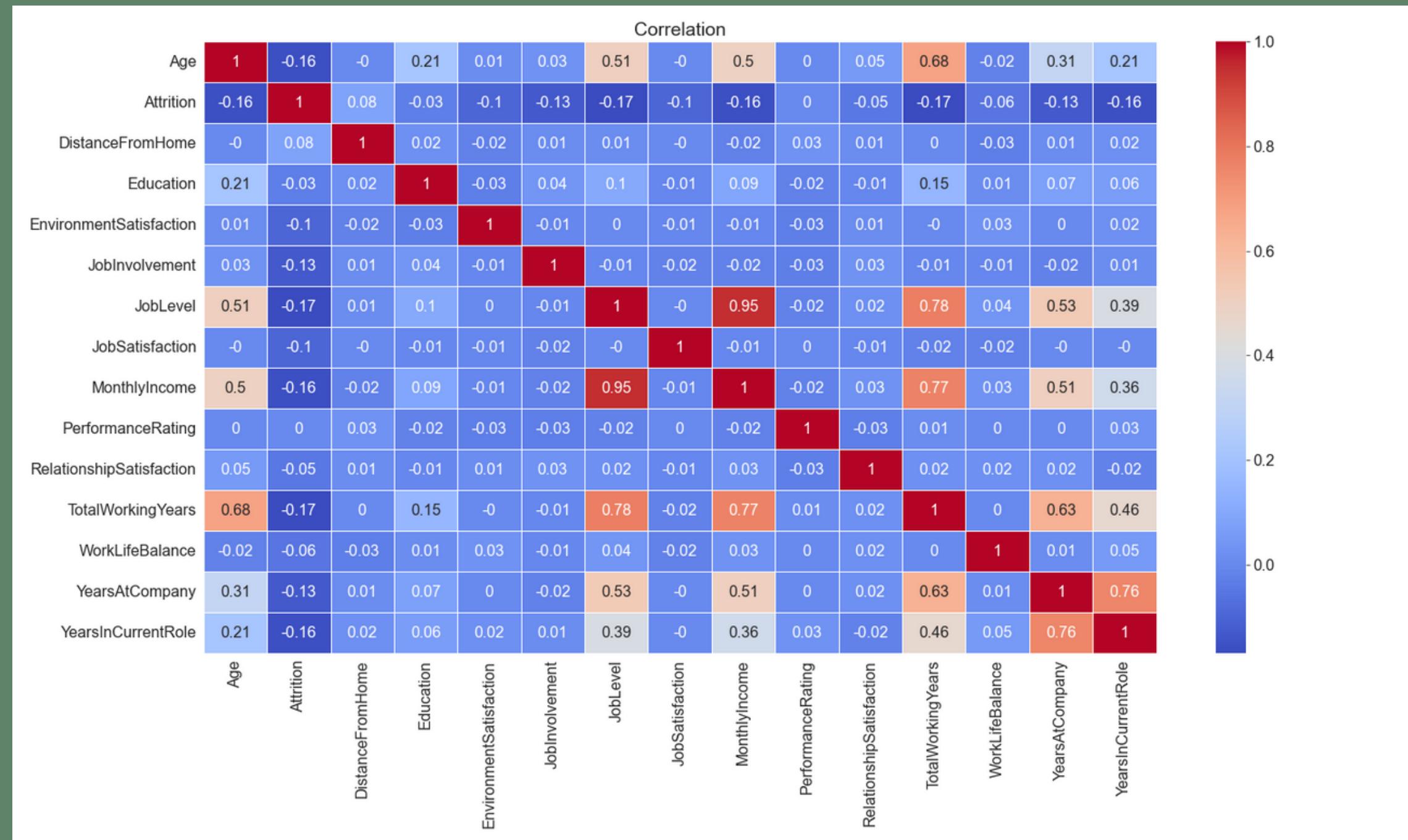
Діаграма розмаху, що ілюструє рівень місячного доходу працівників/-ць у різних відділеннях. Бачимо, що у Sales та Research&Development є досить багато викидів. Люди мають вищий за середній рівень місячного доходу. Бачимо, що у Human Resources теж є викиди, але їх менше.

зарплата робітників в залежності від виду діяльності



Діаграма розмаху, що ілюструє залежність посади робітників від їх зарплатні. Мінімальна зарплата у компанії є у таких спеціальностей як Research Scientist, Laboratory Techician, Human Resources та Sales Representative.

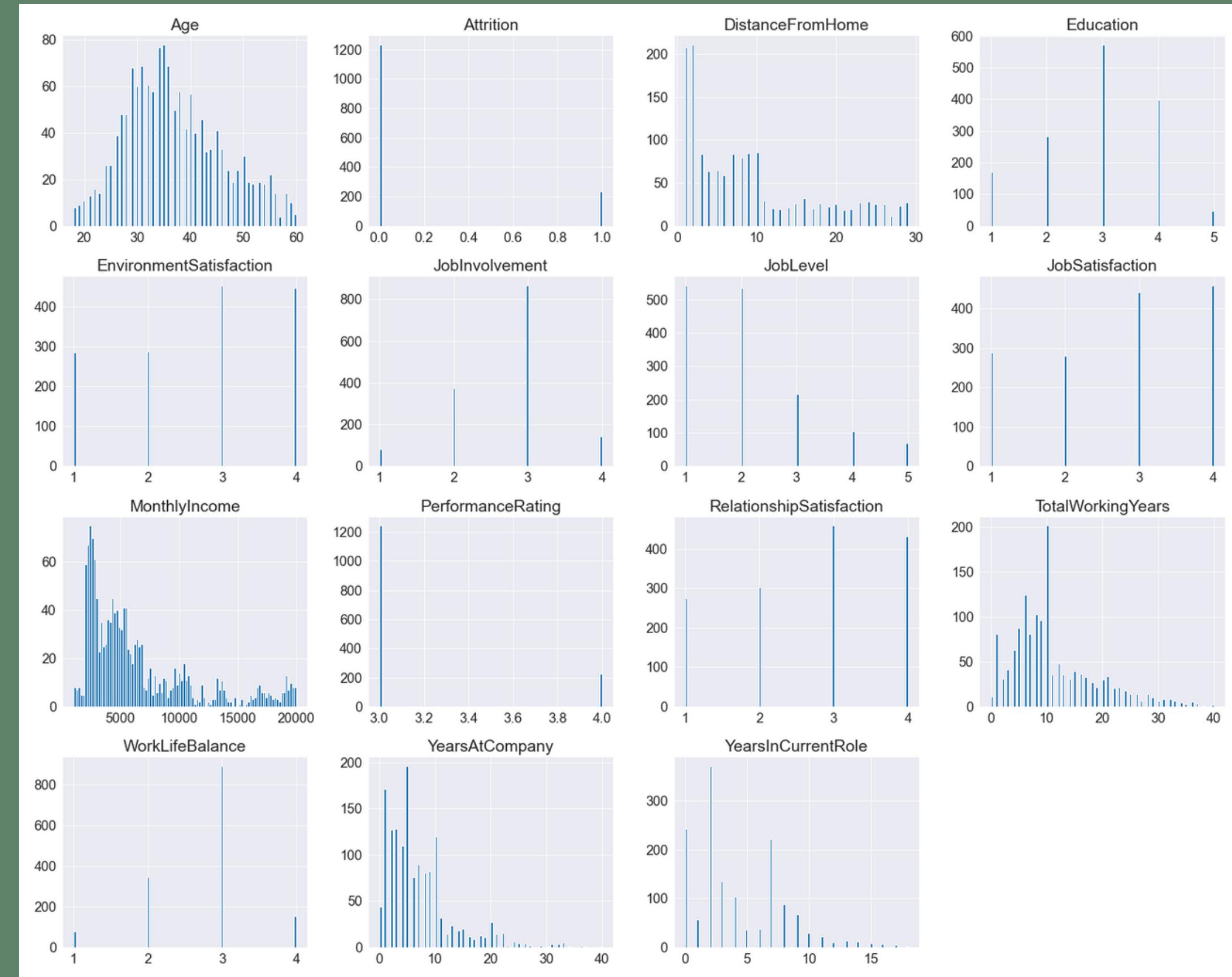
кореляція



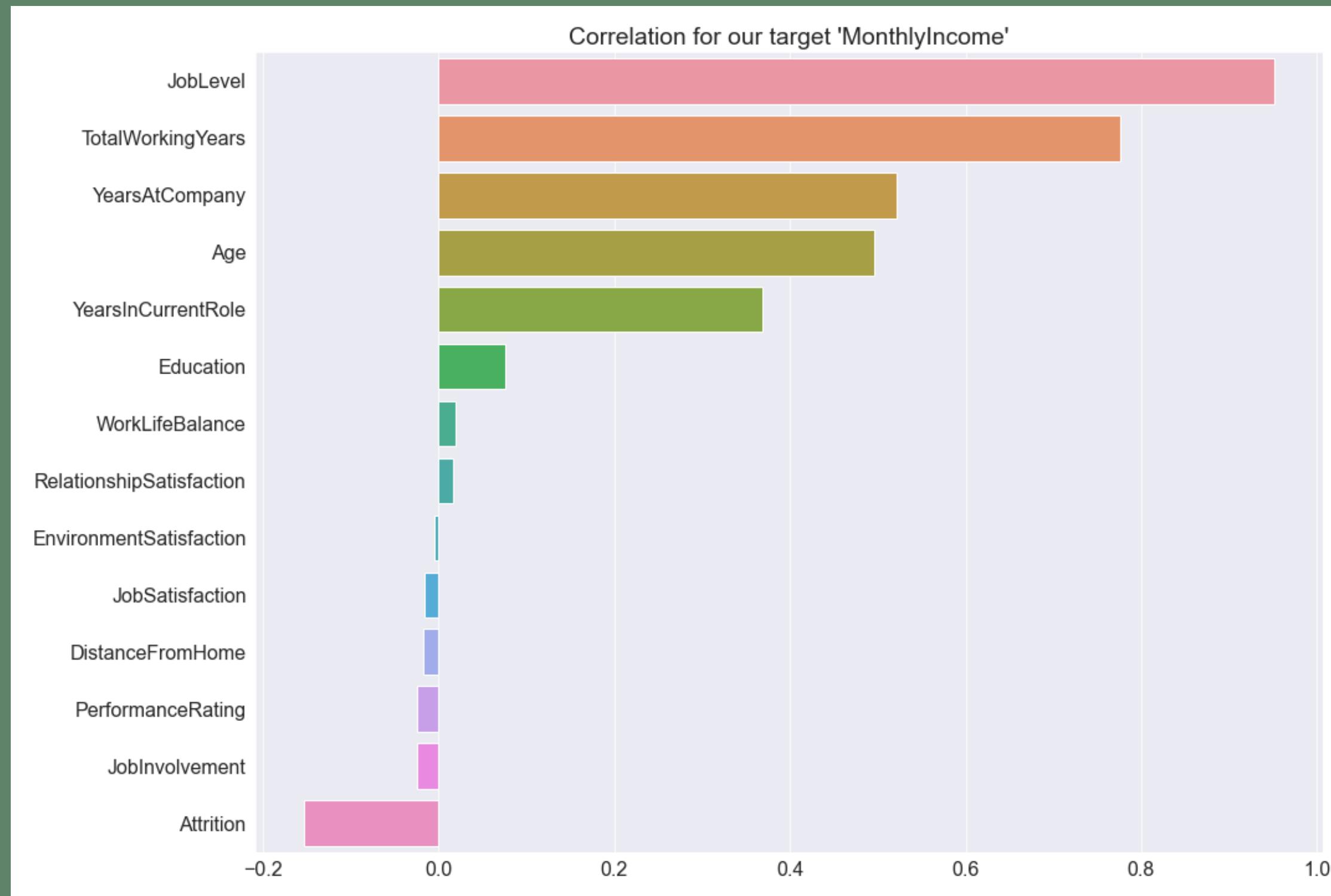
Із кореляції датасету бачимо, що найбільший позитивний вплив на цільову змінну мають такі ознаки:

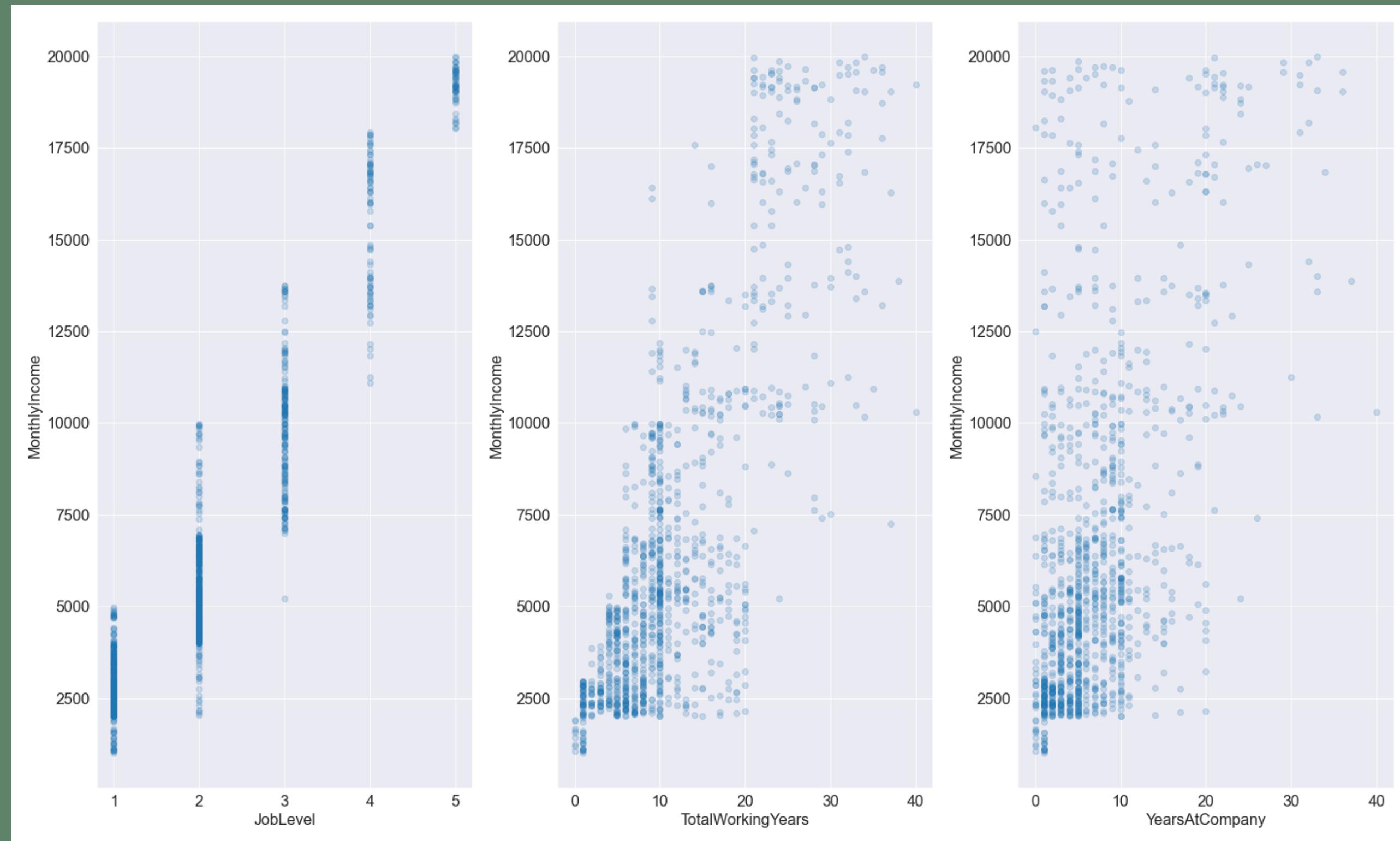
- JobLevel
- TotalWorkingYears
- YearsAtCompany
- Age

розподіл ознак



Залежність цільової змінної від інших числових ознак





`("JobLevel",`
`"TotalWorkingYears",`
`"YearsAtCompany")`

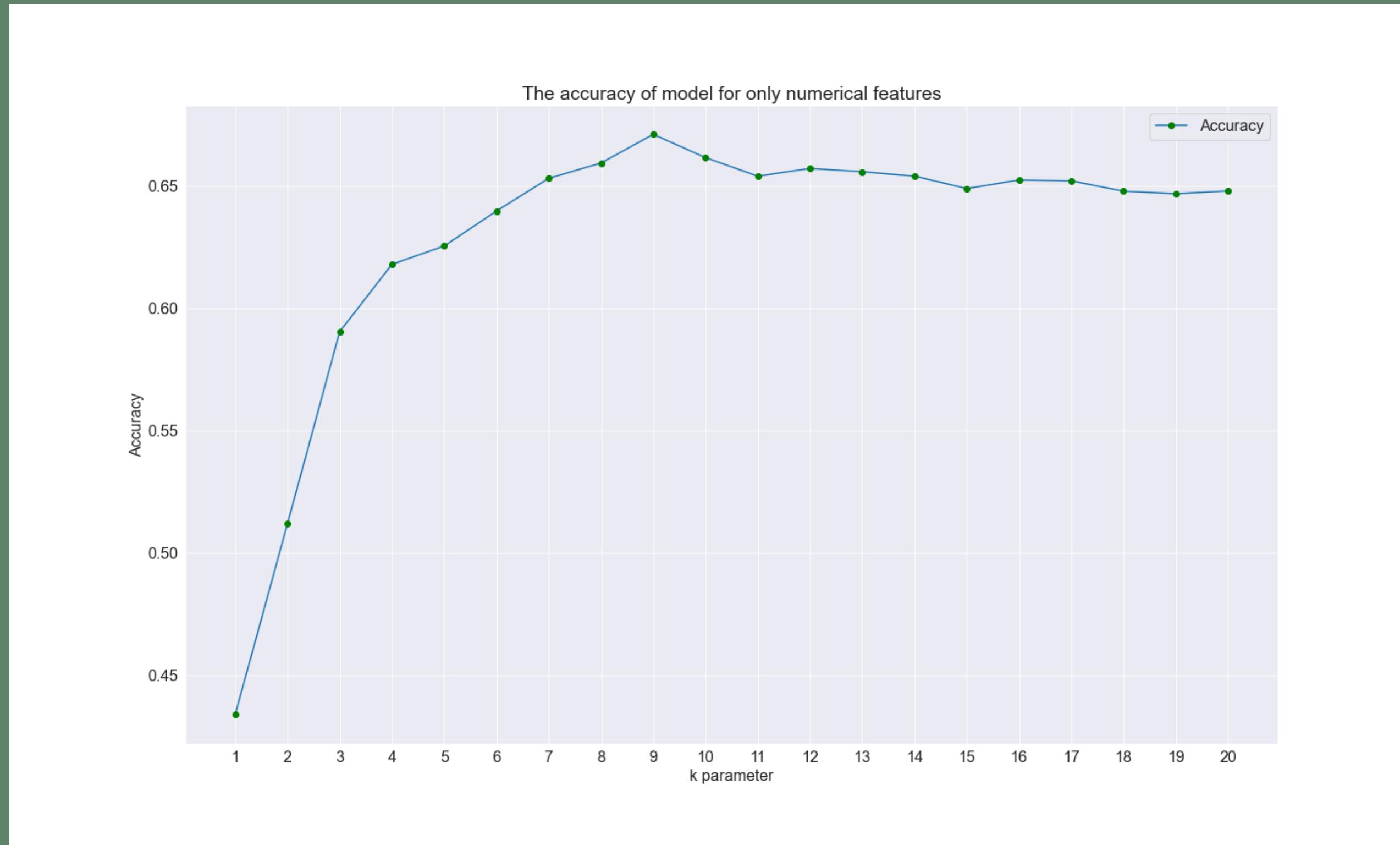
графіки розсіювання для трьох числових ознак
відносно нашої цільової змінної



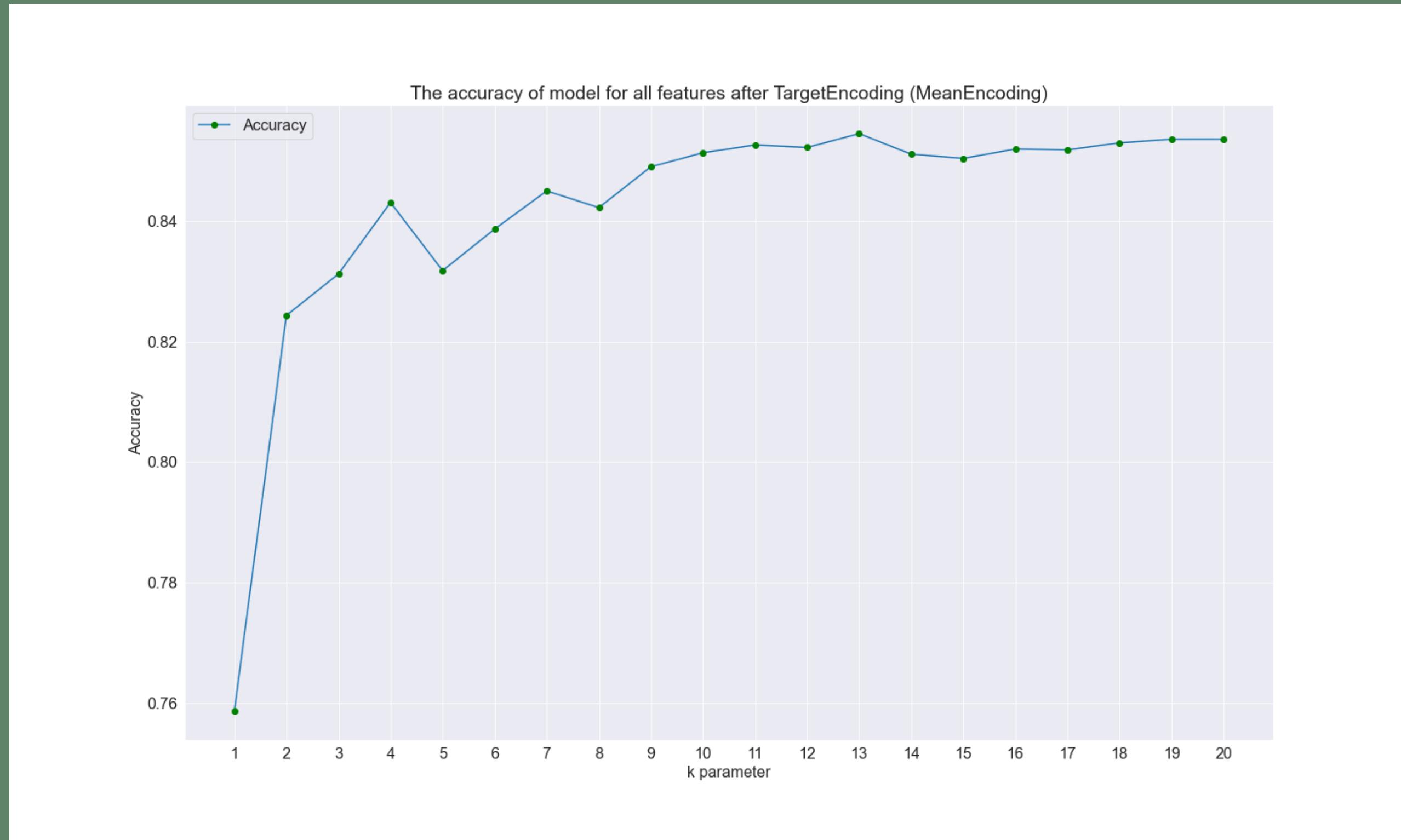
ML



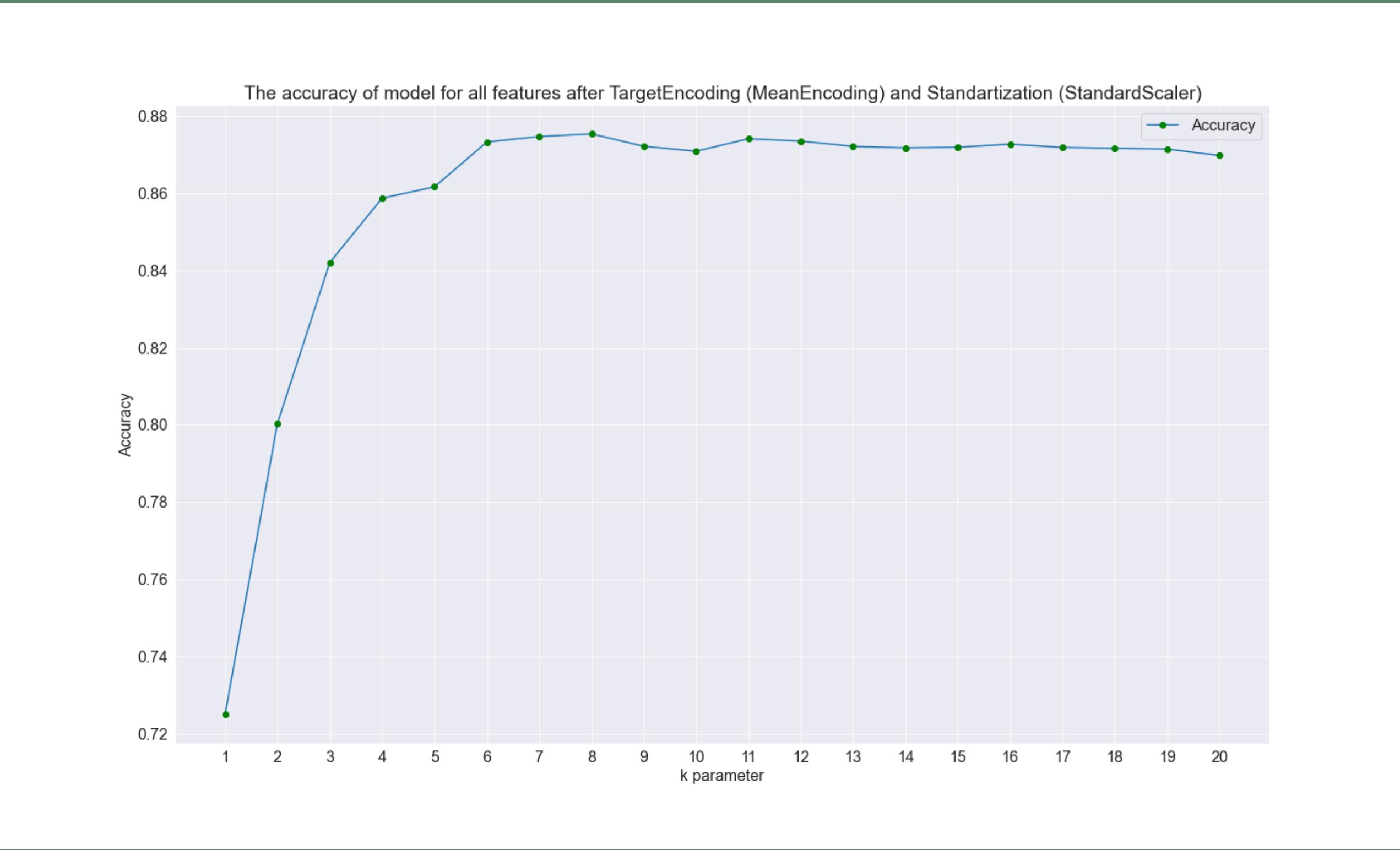
Модель KNNRegressor



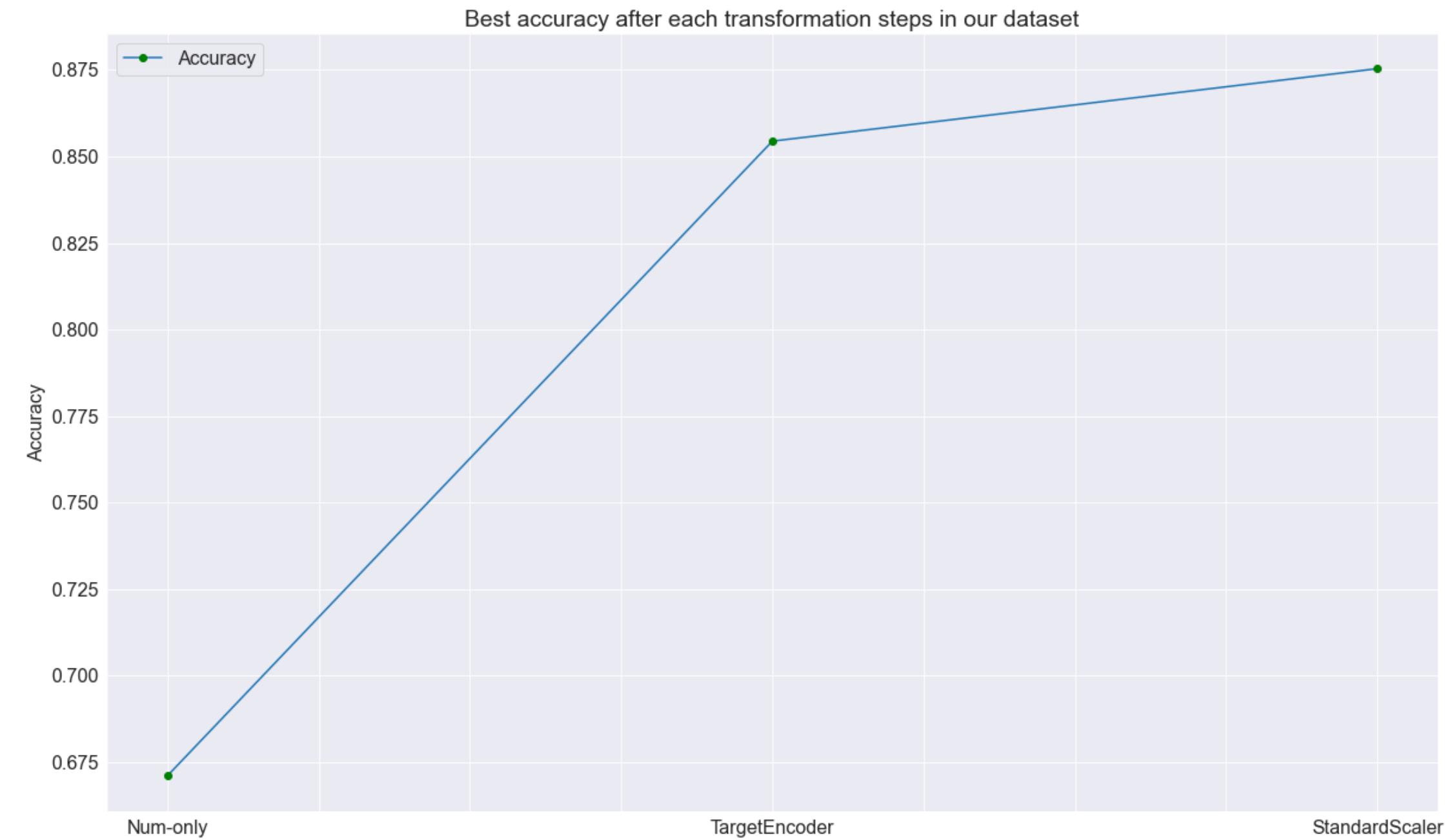
графік точності моделі лише для числових ознак



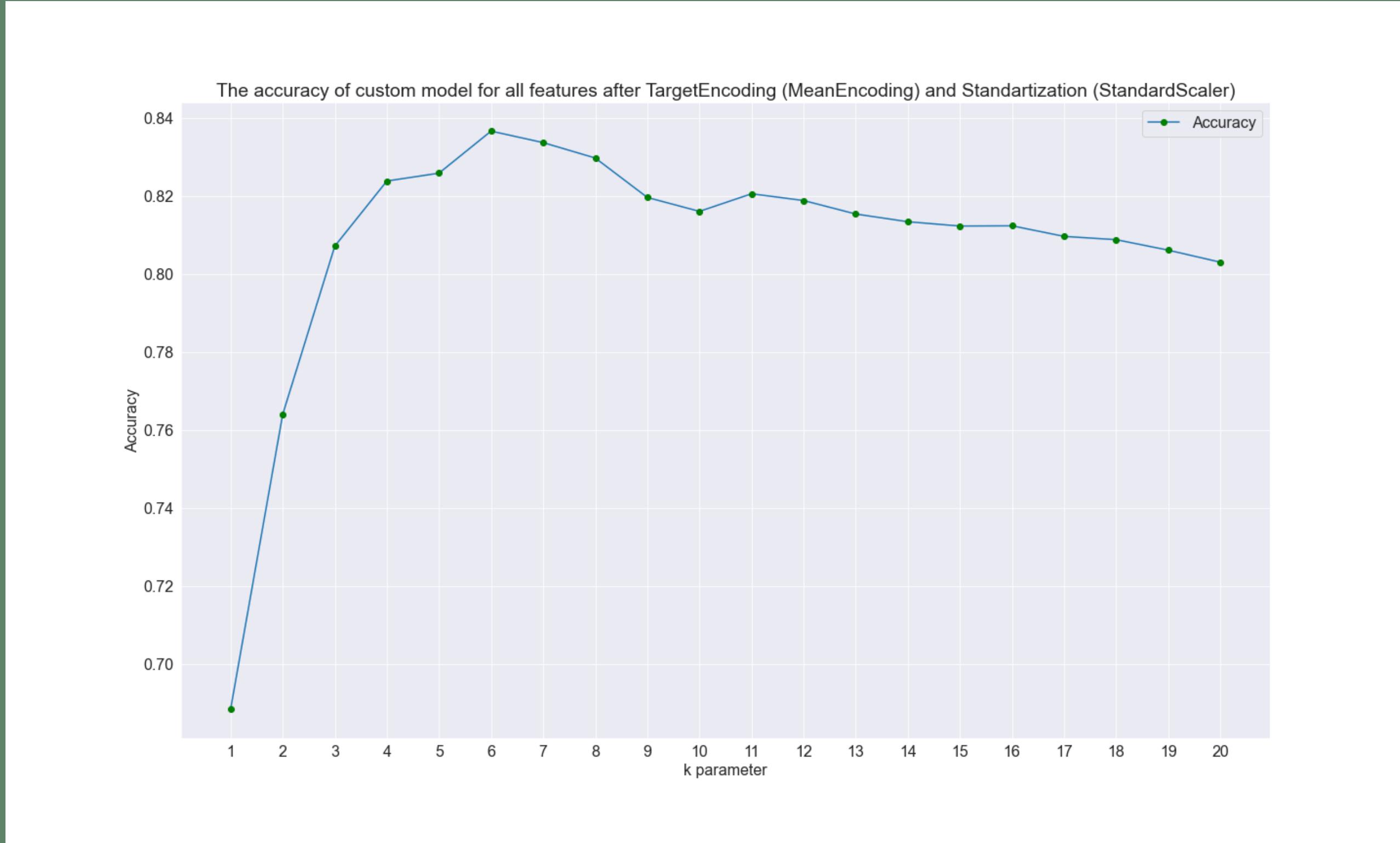
графік точності моделі для всього датасету
після TargetEncoding (MeanEncoding)



графік точності моделі для датасету після
TargetEncoding (MeanEncoding) та Standartization
(StandardScaler)



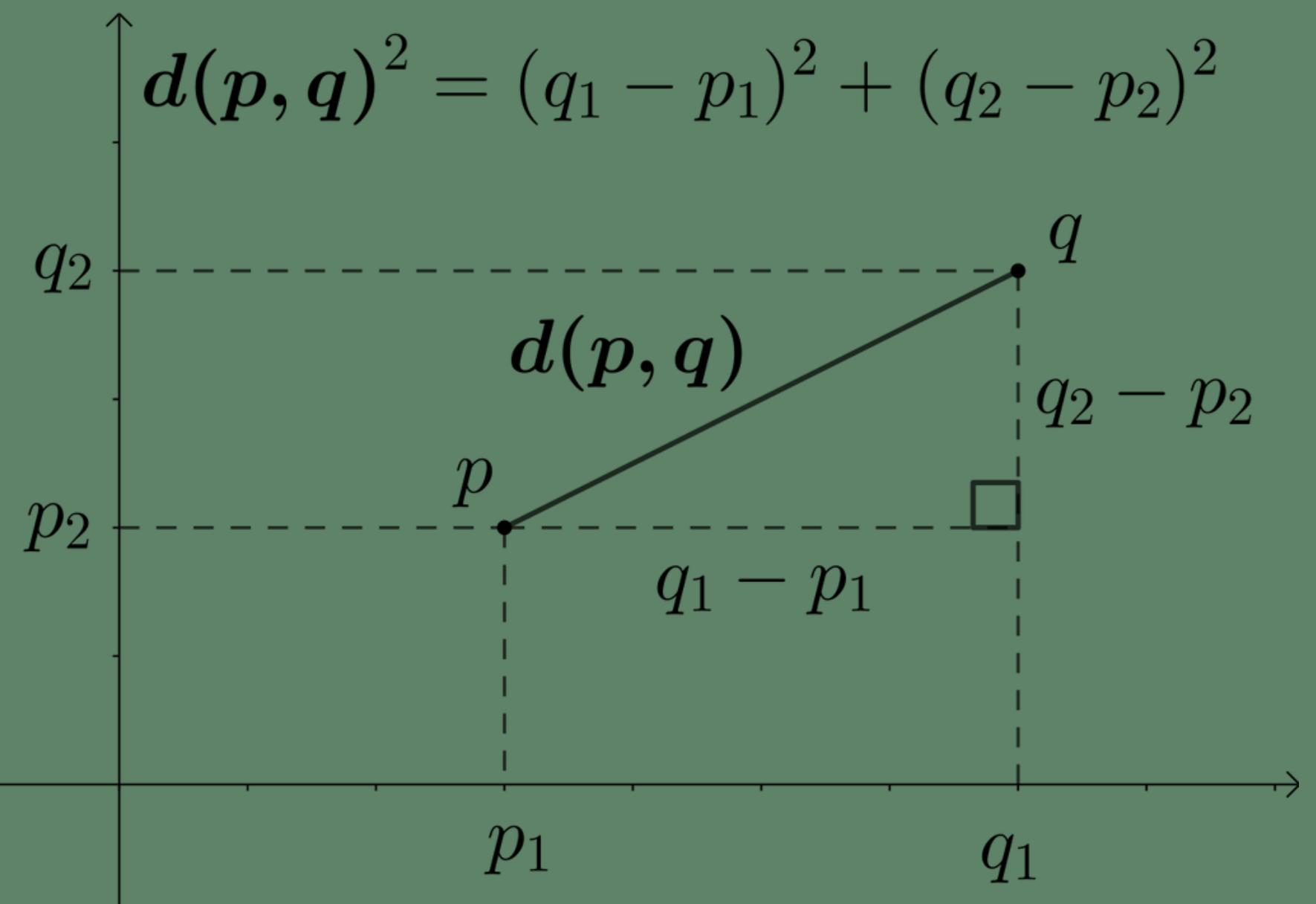
графік зростання точності для моделі залежно від
кроку трансформації датасету



графік точності (вручну написаної) моделі для всього датасету після TargetEncoding (MeanEncoding) та Standartization (StandardScaler)

формула Евклідової відстані

$$d(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$



вигляд при $n=2$

ЧАС ВИСНОВКУ

- Дивлячись на результати моделі KNeighborsRegressor, можна сказати що найкращим параметром n_neighbors(k) буде 8, а параметр metric – "manhattan". В такому випадку точність сягала **89.15%**.
- Використовуючи метрику відстані за замовчуванням ("minkowski"), і тільки перебиранням параметру n_neighbors(k) від 1 до 20, серед яких найкращим було значення 8, при якому максимальна точність була **87.53%**.
- Для вручно-написаного KNeighborsRegressor, використовуючи "euclidean", і перебираючи параметр n_neighbors(k) від 1 до 20, серед яких найкращим було значення 6, при якому максимальна точність була **83.66%**.