DOI 10.31891/2307-5732-2023-319-1-300-307
УДК 004.043

ТКАЧИК ОЛЕКСАНДР
Національний університет "Львівська політехніка"
https://orcid.org/0000-0002-0728-4208
e-mail: oleksandr.a.tkachyk@lpnu.ua

ЗАСТОСУВАННЯ МЕТОДІВ КЛАСТЕРИЗАЦІЇ ДАНИХ ДЛЯ СТВОРЕННЯ ЦІЛЬОВИХ ГРУП КОРИСТУВАЧІВ НА РИНКУ НЕРУХОМОСТІ

У цій статті проведено неконтрольовану кластеризацію різнотипних даних щодо записів клієнтів із бази даних компанії з нерухомості. Сегментація клієнтів у групи — це практика розподілу клієнтів на певні групи, які відображають схожість між клієнтами в кожному кластері. Однією із задач поділу клієнтів на сегментовані групи є збільшення значущості кожного клієнта для бізнесу. У результаті поділу кожній групі можна буде запропонувати конкретні пропозиції, а також швидше знайти індивідуальний підхід для кожної одиниці певної групи. Це також дозволить допомогти бізнесу задовольнити потреби різних клієнтів та швидше скерувати їх у потрібному напрямку. Ключовим кроком є підготовка датасету для майбутньої кластеризації. Для роботи було взято зріз бази даних із 2000 користувачів, які зацікавлені ринком нерухомості. Після проведення аналізу даних, реалізовано підготовку та нормалізацію даних. Зменшено розмірність даних із допомогою методу PCA. Проведено кластеризацію даних і на їх основі створено та описано цільові групи користувачів.

Ключові слова: k-means, різнотипні дані, кластеризація даних, машинне навчання, ринок нерухомості, навчання без нагляду.

TKACHYK OLEXANDR
Lviv Polytechnic National University

APPLYING DATA CLUSTERING METHODS FOR CREATING TARGETING USER GROUPS FOR REAL ESTATE

In this paper applied unsupervised clustering to a dataset examines the application of k-means clustering to create target user groups for a real estate platform. The goal is to segment the user base into meaningful groups to better understand their preferences and behaviors, and tailor marketing campaigns and product features to the needs of each group. The key step in the application of k-means clustering to real estate data is data preparation. Real estate data can be particularly messy and incomplete, and thus requires careful cleaning and normalization before clustering can be applied. Data preparation includes several key steps, such as removing irrelevant or redundant features, creating new features as feature scaling is also an important step in data preparation. K-means clustering is sensitive to the scale of the data, so features may need to be normalized to ensure that they are on the same scale, handling missing or erroneous data, and scaling or transforming features to ensure they are on the same scale. Dataset of 2000 customers interested in real estate with the various types of data was taken as a basis. Then the data was observed, investigated and based on results it was prepared for clustering by doing data cleaning as irrelevant data or empty data points may include features that do not significantly contribute to the clustering process, data normalization as it is necessary to ensure that all features are on the same scale, feature selection to determine most relevant features for clustering, feature encoding and dimensionality reduction which was achieved through principal component analysis (PCA). By carefully cleaning, normalizing, and selecting relevant features, clustering algorithms such as k-means were applied more effectively and target user groups were identified.

Keywords: k-means, various types of data, data clustering, machine learning, real estate, unsupervised clustering

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Ринок нерухомості постійно розвивається, вдосконалюється та розширюється. З'являються нові та прогресивні рішення, кількість пропозицій, варіантів та опцій стає все більшою. Станом на сьогодні вже існує багато різноманітних інформаційних систем, Інтернет-ресурсів, порталів та площадок, які акумулюють пропозиції забудовників та дозволяють знайти ту чи іншу нерухомість, у відповідності до побажань. За останні декілька років ті ж Інтернет-ресурси та портали також удосконалились та пропонують клієнтам досить гнучкі системи пошуку нерухомості, реалізовані з допомогою різноманітних фільтрів. Крім цього, системи пропонують клієнтам підписуватись на їхні розсилки, які будуть періодично надсилати ті чи інші пропозиції, керуючись останніми пошуковими запитами клієнтів. Окрім цього, існують компанії які надають можливість клієнту спілкуватись із персональними ріелторами, котрі, на основі побажань, допомагають підібрати необхідну нерухомість і тим самим економлять час клієнта. Як правило, на початкових етапах складається портрет користувача - проводиться збір персональної інформації, вподобання, уточнюються вимоги до пошуку. В подальшому проводиться пошук пропозицій та їх демонстрація клієнту. Впродовж співпраці ріелтор постійно намагається максимально описати портрет клієнта, щоб у майбутньому підібрати оптимальну пропозицію. Цей крок також потрібний і для того, щоб, у разі необхідності, партнер ріелтора міг без проблем перейняти на себе обов'язки та продовжити роботу. Всі ці дані, які утворюють портрет користувача можна аналізувати у рамках деяких моделей, які можуть визначатись певними вхідними регресійними умовами. У свою чергу, регресійні моделі можуть використовуватись для того, щоб передбачити та зрозуміти зв'язки між змінними, спрогнозувати чи оцінити якусь тенденцію. Одним із прикладів може бути попереднє прогнозування портретів користувачів систем продажу та оренди нерухомості для кращого підбору пропозицій та розсилок.

Аналіз досліджень та публікацій

Із постійним розвитком ринку нерухомості, платформи, які пропонують пошук, підбір та пропозиції варіантів також повинні вдосконалюватись та підлаштовуватись під вимоги користувачів і, власне, самого ринку. Вчасне реагування на зміни дозволяє утримати та збільшити клієнтську базу, але саме виявлення тієї чи іншої зміни залишається складним питанням. Необхідно постійно моніторити ринок, слідкувати за зміною поведінки споживачів та враховувати нові побажання. Фактори можуть набувати різних форм і не завжди корелюються між собою. Це може бути, наприклад, еволюція чи зміна мислення користувачів, які з певним часом починають шукати той чи інший елемент інфраструктури (зелені зони, школи із наявністю певних параметрів тощо), або ж зміна підходів до будівництва нерухомості [1]. Зміна цінових пропозицій

також є дуже вагомим та впливовим фактором для платформ ринку нерухомості. В межах лише одного міста ціни можуть відносно швидко та значним чином змінюватись, у відповідності до чинників, які впливають на ці зміни. Цими чинниками можуть бути як макроекономічні параметри, так і локальні зміни, як-от розвиток інфраструктури [2]. Тому, зараз активно залучаються різноманітні моделі машинного навчання та підходи. Якість роботи цих моделей сильно залежить від повноти даних та правильності постановки завдання. Через це наразі не існує єдиного підходу для визначення портрету користувача та підбору пропозицій для нього [3].

Формулювання цілей статті

Метою роботи є процес підготовки та кластеризації масиву різнотипних даних клієнтів для поділу їх на цільові групи для подальшої систематизації подачі пропозицій та цільової розсилки пропозицій.

Виклад основного матеріалу

Датасет складається із 2000 одиниць даних та 11 атрибутів. Атрибути містять у собі різні набори даних, які необхідно структурувати та об'єднати у певні групи. Після завантаження датасету, було проведено поверхневий огляд загальної структури (рис. 1).

```
[4]: user_listing_data.head()
```

[4]:

id	date_of_birth	gender	family_status	kids_count	social_benefits	employment_status	employed_since	net_income	interested_in	search_performed_count
1	2000	male	Single	0	No	Employed	2021	2952.0	Rent	49
2	1962	male	Single	0	No	Unemployed	1982	9519.0	Rent	12
3	1973	male	Married	2	No	Employed	1993	9313.0	Buy	60
4	1959	male	Married	0	Yes	Employed	1980	5341.0	Rent	55
5	1979	female	Married	0	No	Employed	2002	9285.0	Rent	51

Рис. 1. Огляд структури даних

Для того, щоб отримати повне уявлення про те, які кроки потрібно вжити для очищення та підготовки набору даних, необхідно переглянути інформацію про дані (рис. 2).

```
user_listing_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2000 entries, 1 to 2000
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   date_of_birth          2000 non-null   int64
 1   gender                 2000 non-null   object
 2   family_status          2000 non-null   object
 3   kids_count             2000 non-null   int64
 4   social_benefits        2000 non-null   object
 5   employment_status      2000 non-null   object
 6   employed_since         2000 non-null   int64
 7   net_income             1942 non-null   float64
 8   interested_in          1876 non-null   object
 9   search_performed_count 2000 non-null   int64
dtypes: float64(1), int64(4), object(5)
memory usage: 171.9+ KB
```

Рис. 2. Інформація про дані

Із наведеного вище зведеного огляду про дані можна зробити наступний висновок:

- Існують пусті значення у деяких полях;
- Існують поля із набором різнотипних даних з типом object, які необхідно буде проаналізувати, згрупувати та декодувати у числові значення.

Для початку необхідно видалити рядки, які містять пусті значення (рис. 3).

```
[9]: user_listing_data = user_listing_data.dropna()
     print("The total number of data-points after removing the rows with missing values are:", len(user_listi
```

```
The total number of data-points after removing the rows with missing values are: 1819
```

Рис. 3. Видалення рядків із пустими значеннями

Далі проводиться створення ряду нових функцій, видалення зайвих атрибутів та конвертація різнотипних значень типу object у числовий для того, щоб можна було краще працювати із даними (рис. 4). Набір функцій наступний:

- Витягування інформації про те, скільки років користувач офіційно працевлаштований;
- Витягування віку користувача;
- Формування кількості людей у сім'ї;
- Формування статусу батьківства.

```
#Feature for deriving gender
user_listing_data["Gender"] = user_listing_data["gender"].replace({"male": 1, "female":2})
#Feature for deriving employment status
user_listing_data["Employment_Status"] = user_listing_data["employment_status"].replace({"Employed": 1, "Unemployed":2})
#Feature for deriving customer interests
user_listing_data["Interested_In"] = user_listing_data["interested_in"].replace({"Rent": 1, "Buy":2})
#Feature for deriving employed since status
user_listing_data["Employed_For"] = 2023 - user_listing_data["employed_since"]
user_listing_data["Employed_For"] = user_listing_data["Employed_For"][user_listing_data["Employed_For"] >= 0]
user_listing_data = user_listing_data.dropna()
#Feature for deriving customers age
user_listing_data["Age"] = 2023 - user_listing_data["date_of_birth"]
#Feature for deriving family size
data["family_status"]=data["family_status"].replace({
    "Married":"Married",
    "Together":"Married",
    "Widow":"Single",
    "Divorced":"Single",
    "Single":"Single",})
user_listing_data["Family_Size"] = user_listing_data["family_status"].replace({"Single": 1, "Married":2}) + user_listing_data["kids_count"]
#Feature pertaining parenthood
user_listing_data["Is_Parent"] = np.where(user_listing_data.kids_count> 0, 1, 0)
# For clarity
user_listing_data=user_listing_data.rename(columns={
    "net_income": "Income",
    "search_performed_count":"Search_Activity"
})
#Dropping some of the redundant features
to_drop = [
    "gender",
    "interested_in",
    "date_of_birth",
    "social_benefits",
    "employed_since",
    "employment_status",
]
user_listing_data = user_listing_data.drop(to_drop, axis=1)
user_listing_data.describe()
```

Рис. 4. Формування нових функцій

Після створення попередньо описаних функцій можна оглянути загальні статистичні характеристики даних (рис. 5).

	kids_count	Income	Search_Activity	Gender	Employment_Status	Interested_In	Employed_For	Age	Family_Size	Is_Parent
count	1824.000000	1824.000000	1824.000000	1824.00000	1824.000000	1824.000000	1824.000000	1824.000000	1824.000000	1824.000000
mean	0.355811	8440.408443	71.314693	1.50932	1.163925	1.428180	28.549890	49.885965	1.718750	0.262061
std	0.667430	3287.230877	49.385224	0.50005	0.370309	0.494951	15.619963	15.554873	0.983885	0.439876
min	0.000000	2330.000000	1.000000	1.00000	1.000000	1.000000	0.000000	21.000000	1.000000	0.000000
25%	0.000000	5574.000000	29.000000	1.00000	1.000000	1.000000	15.000000	36.000000	1.000000	0.000000
50%	0.000000	8157.500000	65.000000	2.00000	1.000000	1.000000	29.000000	50.000000	1.000000	0.000000
75%	1.000000	11329.750000	105.000000	2.00000	1.000000	2.000000	42.000000	63.000000	2.000000	1.000000
max	3.000000	14496.000000	307.000000	2.00000	2.000000	2.000000	58.000000	76.000000	5.000000	1.000000

Рис. 5. Статистичні характеристики моделі

Далі необхідно провести більш детальний огляд певної інформації. Для цього потрібно побудувати графік із обраних підмножин ознак (рис. 6). Графік дозволить візуально проаналізувати наявні дані та провести подальше коригування датасету.

Із графіка видно, що такі дані як стать та зайнятість (працевлаштований чи безробітний) не несуть корисної інформації, тому, при побудові кластерів, вони будуть видалені з загального обсягу інформації. Далі можна провести кореляційний аналіз між функціями (рис. 7).

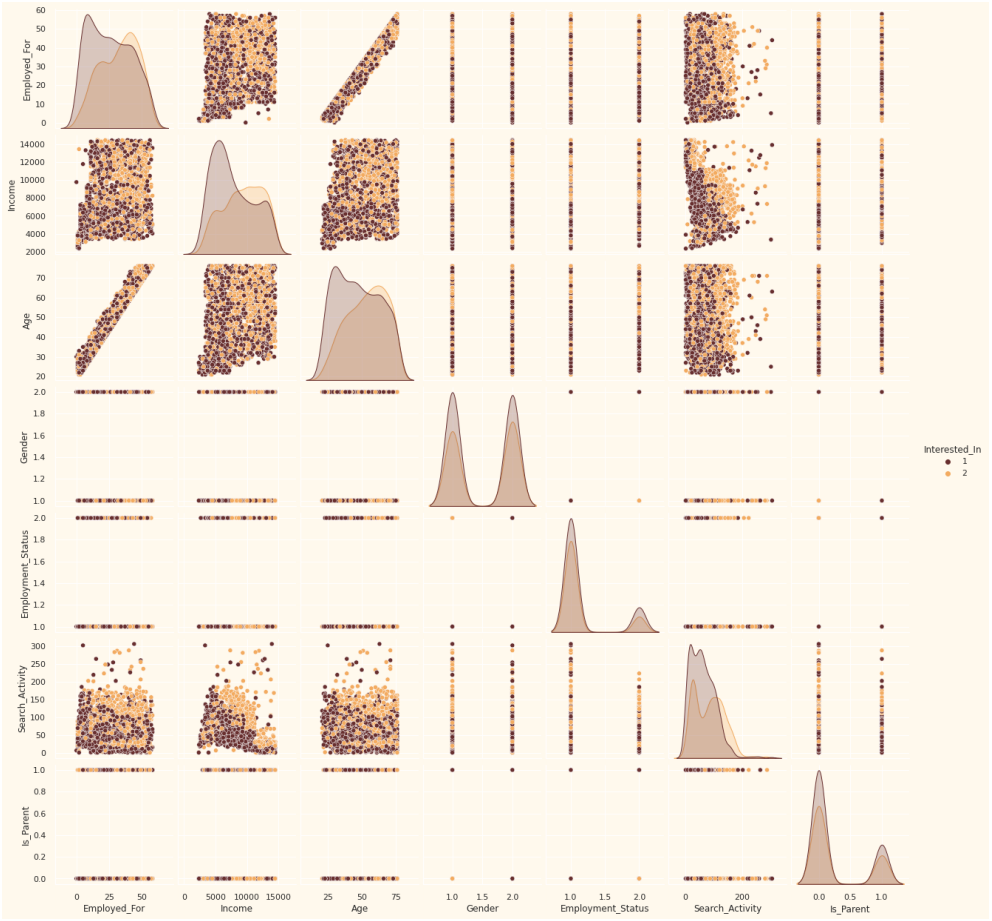

Рис. 6. Графічне відображення даних

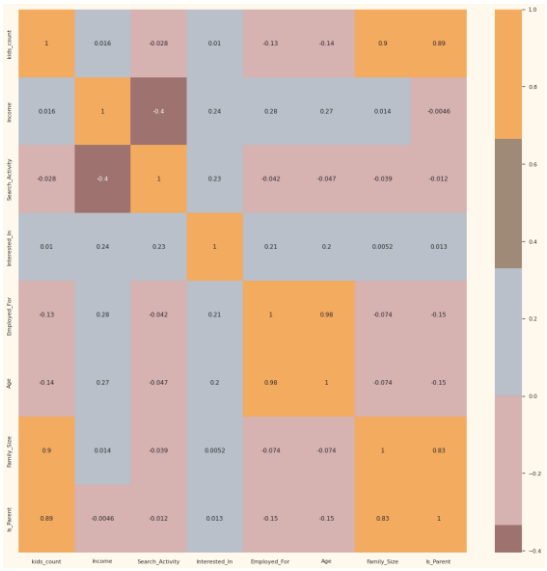

Рис. 7. Інфографіка кореляції між функціями

Кореляційний аналіз показує, що дані відносно чисті, тому ці функції можна додати та використовувати. Попри те, що кореляція між розміром сім'ї та доходом є невисокою, ця функція допоможе точніше описати майбутні кластери. Наступним кроком є препроцесинг даних. Для попередньої обробки застосовуються наступні кроки: мітка, що кодує категоріальні ознаки, масштабування функцій за допомогою стандартного масштабувальника та створення підмножини даних для зменшення розмірності. Після проведення обробки можна оглянути набір даних для подальшого моделювання (рис. 8).

Dataframe to be used for further modelling:

	family_status	kids_count	Income	Search_Activity	Interested_In	Employed_For	Age	Family_Size	Is_Parent
0	0.754790	0.965443	-1.670072	-0.451973	-0.865333	-1.700207	-1.728933	0.285935	1.678064
1	0.754790	-0.533253	0.328206	-1.201391	-0.865333	0.797283	0.714701	-0.730723	-0.595925
2	-1.324872	2.464138	0.265522	-0.229174	1.155624	0.092862	0.007333	2.319250	1.678064
3	-1.324872	-0.533253	-0.943122	-0.330446	-0.865333	0.925359	0.907619	0.285935	-0.595925
4	-1.324872	-0.533253	0.257001	-0.411464	-0.865333	-0.483481	-0.378504	0.285935	-0.595925

Рис. 8. Зразок оброблених даних

Наступним кроком буде зменшення розмірності даних. Ця задача містить у собі багато факторів, на основі яких буде зроблена остаточна класифікація. Всі ці фактори в основному є атрибутами або особливостями. В загальному, чим більше функцій існує в наборі даних, тим важче із ними працювати. Деякі з цих функцій є корельовані, а отже, зайві. Саме для цього зменшення розмірності для вибраних об'єктів є необхідним кроком. Процес зменшення розмірності – це зменшення кількості випадкових величин, що розглядаються, шляхом отримання набору головних змінних. Аналіз головних компонентів (PCA) – це техніка, яка використовується для зменшення розмірності таких наборів даних, підвищуючи інтерпретацію, але в той же час мінімізуючи втрати інформації [5]. Для цієї задачі розмірність буде зменшена до 3 вимірів (рис. 9).

```
#PCA method to reduce dimentions to 3
pca = PCA(n_components=3)
pca.fit(scaled_ds)
PCA_ds = pd.DataFrame(pca.transform(scaled_ds), columns=(["col1","col2", "col3"]))
PCA_ds.describe().T
```

	count	mean	std	min	25%	50%	75%	max
col1	1824.0	5.258951e-17	1.802498	-1.827483	-1.362546	-0.839866	1.171528	5.393486
col2	1824.0	2.337312e-17	1.488450	-3.437680	-1.152381	-0.105609	1.121915	3.574351
col3	1824.0	-1.947760e-17	1.153267	-2.471730	-0.831271	-0.013209	0.852033	4.504870

Рис. 9. Застосування методу PCA

Після зменшення атрибутів до трьох вимірів, можна провести кластеризацію за допомогою методу Elbow для визначення кількості кластерів, які мають бути сформовані (рис. 10). Це алгоритм неконтрольованого навчання, який використовується для вирішення проблем кластеризації в машинному навчанні [6].

```
print('Elbow Method to determine the number of clusters to be formed:')
Elbow_M = KElbowVisualizer(KMeans(), k=10)
Elbow_M.fit(PCA_ds)
Elbow_M.show()
```

Elbow Method to determine the number of clusters to be formed:

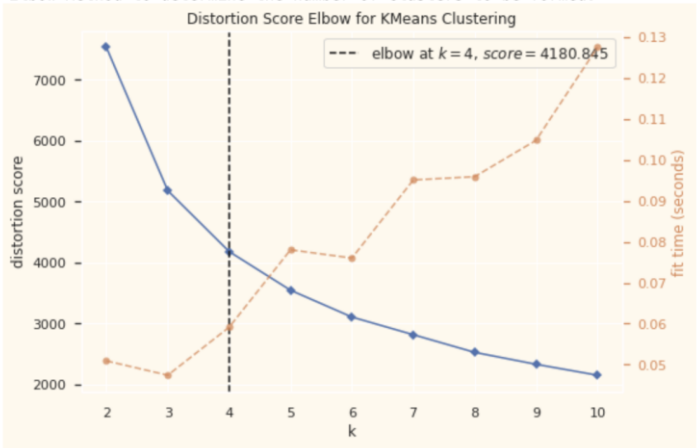

Рис. 10. Визначення кількості кластерів

Наведена вище інфографіка показує, що оптимальною кількістю кластерів для цих даних буде 4. Далі буде потрібно створити кластери та згрупувати їх за зазначеними вище характеристиками, використовуючи метод агломеративної кластеризації. Цей метод використовує принцип знизу-вгору,

об’єднуючи пари кластерів, які містять у собі найближчу пару елементів, які ще не належать до спільного кластера. Результат кластеризації зображений на рис. 11.

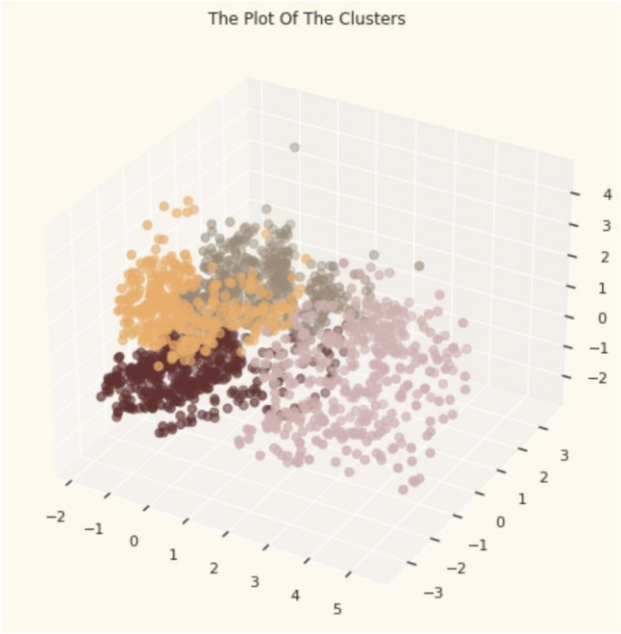

Рис. 11. Сформовані кластери

Оскільки проводиться неконтрольоване кластеризування, відповідно, немає конкретної функції з тегами, яка могла би допомогти оцінити або оцінити створену модель. Основною метою є вивчення закономірностей у сформованих кластерах та визначення характеру структур кластерів. Для цього дані можна оглянути за допомогою дослідницького аналізу даних і на їх основі зробити необхідні висновки. Із допомогою точкової діаграми (рис. 12) можна візуально оглянути профілі кластерів, сформованих на основі пошукової активності та доходів користувача.

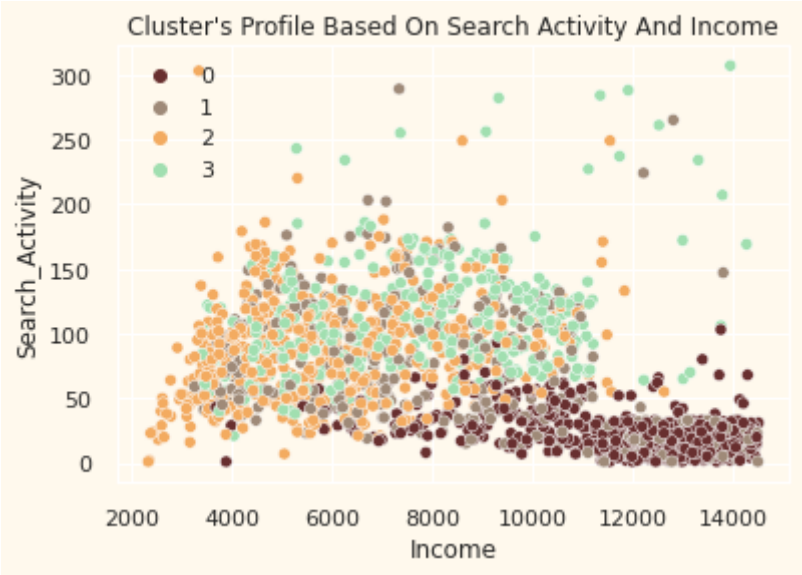

Рис. 12. Точкова діаграма на основі пошукової активності та доходів

Як результат, можна провести попередній опис чотирьох груп кластерів:

- Перша група (0-й кластер): високий дохід і низька пошукова активність;
- Друга група (1-й кластер): в основному високий дохід і дещо вища пошукова активність;
- Третя група (2-й кластер): низький та середній дохід і, відповідно, середня та висока пошукова активність;
- Четверта група (3-й кластер): середній дохід та висока пошукова активність.

Наведені вище дані є досить інформативні, проте, потребують більшого аналізу. Потрібно детальніше зрозуміти, які саме групи користувачів є в цих кластерах. Для цього сформовані кластери будуть візуалізовані із допомогою спільної діаграми (рис. 13). Спільні діаграми корисні для дослідження взаємозв'язку між двома змінними, такими як їх кореляція, кластеризація або розподіл. Поєднуючи різні типи графіків, спільні діаграми можуть надати більш повне уявлення про дані, ніж окремі графіки. Вони також можуть бути настроєні для виокремлення певних особливостей або патернів в даних, таких як викиди, тенденції або кластери [7]. На основі спільних діаграм можна буде дійти висновку про те, яка група може бути цільовою, а також дізнатись, хто потребує більше уваги з боку маркетингової команди.

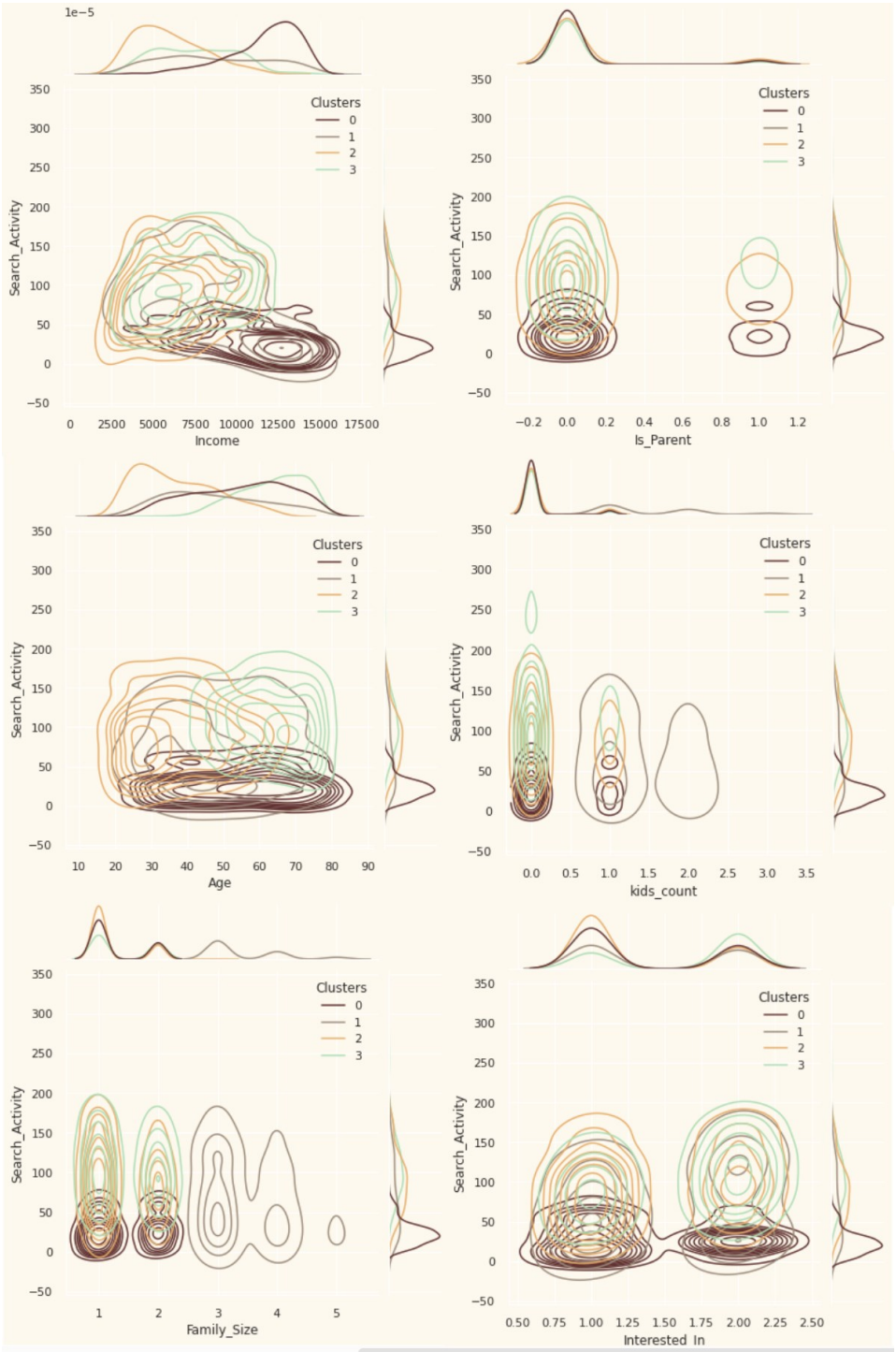

Рис. 13. Набір спільних діаграм на основі зазначених параметрів

Після огляду та аналізу наведених вище діаграм можна зробити загальне профілювання цільових груп:

- Перша група (0-й кластер) має високий дохід, в основному без дітей, вікова група переважно від 38 до 70 років. У більшості випадків користувачі є без пари і більше зацікавлені у купівлі чи інвестиції у нерухомість.
- Друга група (1-й кластер) має дохід вище середнього, в середньому має 1-2 дитини, вікова група переважно від 30 до 50 років. У більшості випадків користувачі мають пару та здебільшого зацікавлені у купівлі нерухомості.
- Третя група (2-й кластер) має невеликий дохід, більшість не має дітей, переважає у віці 20–30 років. В основному користувачі без пари та зацікавлені у оренді нерухомості.
- Четверта група (3-й кластер) має середній дохід, в загальному вже немає малих дітей. Вікова група 60–80 років. По розміру сім'ї можна бачити, що статистичні дані розподілені більш-менш рівномірно – можуть бути як із парою, так і без. Більшість зацікавлена в купівлі чи інвестиції нерухомості, хоча є група, яка зацікавлена в оренді.

На основі цих даних працівники агентства ринку нерухомості матимуть можливість краще оцінити які групи користувачів у них існують, а також сформувати якісні маркетингові кампанії. Також, враховуючи низьку активність першої групи користувачів, а також зважаючи на високий дохід цієї групи, можна певним чином стимулювати співпрацю.

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

У даній роботі проведено аналіз та підготовку даних, а саме видалення зайвих значень, нормалізацію, зменшення розмірності, а також створення нових функції, таких як розмір сім'ї, групування даних за певними ознаками та ін. На основі цих даних проведено кластеризацію і створено профільовані групи користувачів, із якими можна проводити маркетингові кампанії. В подальшому, до датасету можна буде додати більше неструктурованих даних, наприклад різноманітні метрики із соцмереж та створити на їх основі нові функції, які допоможуть виявити нові групи користувачів, або удосконалити інформацію про наявні групи.

Література

1. Zheng S., Hu W., Wang R. (2016). How much is a good school worth in Beijing? Identifying price premium with paired resale and rental data. Journal of Real Estate Finance and Economics, 53(2), 184–199.

2. Прогнозування цін на нерухомість в умовах фінансово-економічної кризи. Сейл Прайс Компані. URL: http://www.saleprice.com.ua/ua/publications/real_estate_price_forecasting.html (01.02.2023).

3. Interpretable machine learning for real estate market analysis. URL: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1540-6229.12397 (07.02.2023).

4. Seo W. (2018). Does neighborhood condition create a discount effect on house list prices? Evidence from physical disorder. Journal of Real Estate Research, 40(1), 69–88.

5. Principal component analysis (PCA) definition. URL: https://en.wikipedia.org/wiki/Principal_component_analysis (07.02.2023).

6. K-Means Clustering Algorithm in Machine Learning. URL: https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning (07.02.2023).

7. Seaborn jointplot Method. URL: https://www.educba.com/seaborn-jointplot/ (11.02.2023).

References

1. Zheng S., Hu W., Wang R. (2016). How much is a good school worth in Beijing? Identifying price premium with paired resale and rental data. Journal of Real Estate Finance and Economics, 53(2), 184–199.

2. Prohnozuvannia tsin na nerukhomist v umovakh finansovo-ekonomichnoi kryzy. Seil Prais Kompani. URL: http://www.saleprice.com.ua/ua/publications/real_estate_price_forecasting.html (01.02.2023).

3. Interpretable machine learning for real estate market analysis. URL: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1540-6229.12397 (07.02.2023).

4. Seo W. (2018). Does neighborhood condition create a discount effect on house list prices? Evidence from physical disorder. Journal of Real Estate Research, 40(1), 69–88.

5. Principal component analysis (PCA) definition. URL: https://en.wikipedia.org/wiki/Principal_component_analysis (07.02.2023).

6. K-Means Clustering Algorithm in Machine Learning. URL: https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning (07.02.2023).

7. Seaborn jointplot Method. URL: https://www.educba.com/seaborn-jointplot/ (11.02.2023).