Технически университет – София

Факултет по приложна математика и информатика

Курсова работа

по

Анализ на бизнес данни в социални мрежи

на тема

Анализ на бизнес данни в социалната мрежа Facebook

Изготвили: Савина Вълчанова, Денис Цолов, Цветомир Цветков

Фак. номера: 961324002, 961324006, 961324011

Група: 252

Съдържание

1. Въведение

Анализът на обяви за продажба е от съществено значение за разбиране на тенденциите на пазара и тяхната ефективност. Представените данни обхващат информация за различни обяви за продажба, като се фокусират върху показатели като местонахождение, тип, цена и други. Чрез детайлен анализ на тези данни можем да извлечем ценна информация за ценовите различия, предпочитанията на потребителите и пазарните тенденции в различните градове.

- Прегледът на данните разкрива различни аспекти на продажбите, като:
- Най-популярните категории продукти и техните ценови диапазони.
- Разликите в предпочитанията на потребителите в различни градове.
- Сезонни или времеви тенденции в търсенето.
- Влиянието на местоположението върху цените и обема на продажбите.

Чрез анализа на тези данни можем да идентифицираме най-новите тенденции и да оценим въздействието на различни фактори.

2. Цел

Целта на тази разработка е да анализира обяви за продажба в пазара на мрежата Facebook както и да постави оценка на най-търсените активи, техните цени и местата, където те биват предлагани. Целта също може да бъде разгледана и като комбинация от следните точки:

- **Регионален ценови анализ:**
    - **Колони:** City, Price
    - **Цел:** Анализ на цените и изучаване на ценовото разпределение и средните цени за всеки град

- **Анализ на типа:**
    - **Колони:** Title, Item Type
    - **Цел:** Събиране на информация относно най-продаваните активи и техния вид

- **Разпределителен анализ:**
    - **Колони:** City
    - **Цел:** Анализ на броя обяви според местонахождението на продавания артикул

3. Експериментална рамка

В тази секция ще опишем методите и техниките, използвани за анализа на данни от Facebook Marketplace.

3.1. Набори данни (Datasets):

- **Източник на данни:**
    - Данните са събрани с помощта на “Instant Data Scraper” Google Chrome Extension.
- **Обхванати колони:**
    - Name, City, Price, Item Type
- **Период на събиране:**
    - Данните обхващат периода от една седмица

3.2. Обработка на данни:

- **Почистване на данни:**
  - Справяне с липсващи стойности и коригиране на грешки в данните.
- **Преобразуване на данни:**
  - Форматиране на числови и категорийни стойности и модифициране на събрани данни с по-лесен анализ

3.3. Избор на метод и техника за анализ:

- **Регионален ценови анализ:**
  - **Метод:** Сравнение на средни цени по региони
  - **Техника:** Анализ и визуализация чрез хистограми и бар плотове

- **Типов анализ:**
  - **Метод:** Класификация и сравнение на категории
  - **Техника:** Анализ и визуализация чрез пай диаграми и хийт мапове

- **Синтактичен анализ:**
  - **Метод:** Извличане на отделни думи
  - **Техника:** Анализ и визуализация чрез бар плотове

3.4. Използвани библиотеки и софтуерни средства:

- **Instant Data Scraper**: За създаване на дейтасета
- **Pandas**: За манипулиране и анализ на данните
- **NumPy**: За математически изчисления и операции
- **Matplotlib**: За създаване на графики и визуализации на резултатите
- **Seaborn**: За по-сложни и изящни визуализации

3.5. Техника за интерпретиране на резултатите

- **Описание на резултатите:** Подробно описание на наблюдаваните тенденции и модели в данните. Определяне на ключовите открития спрямо резултатите.

4. Обработка и анализ на данните

4.1. Набор от данни

Out[377...

	Link	Photo Url	Price	Name	City
0	https://www.facebook.com/marketplace/np/item/1...	https://scontent.fsof4-1.fna.fbcdn.net/v/t39.3...	160 000 лв.	Апартамент	Botevgrad
1	https://www.facebook.com/marketplace/np/item/5...	https://scontent.fsof4-1.fna.fbcdn.net/v/t45.5...	180 лв.	Lorelli бебешко легло 60/120см matrix new	София, България
2	https://www.facebook.com/marketplace/np/item/8...	https://scontent.fsof4-1.fna.fbcdn.net/v/t39.3...	125 000 лв.	Продавам къща в село Миланово област София.	Своге
3	https://www.facebook.com/marketplace/np/item/1...	https://scontent.fsof4-1.fna.fbcdn.net/v/t39.3...	139 500 лв.	Къща в с. Костенец	Kostenets
4	https://www.facebook.com/marketplace/np/item/1...	https://scontent.fsof4-1.fna.fbcdn.net/v/t39.3...	128 960 лв.	Двустаен апартамент в Надежда	София, България
...	...	...	...	...	...
207	https://www.facebook.com/marketplace/np/item/8...	https://scontent.fsof4-1.fna.fbcdn.net/v/t45.5...	150 лв.	Фритюрник 5 л. Bereket \nС чугунена горелка.	Kyustendil
208	https://www.facebook.com/marketplace/np/item/1...	https://scontent.fsof4-1.fna.fbcdn.net/v/t39.3...	106 000 лв.	Продава апаратамент в Малинова долина	София, България
209	https://www.facebook.com/marketplace/np/item/1...	https://scontent.fsof4-1.fna.fbcdn.net/v/t45.5...	95 лв.	Премиум ПВЦ мраморни панели	София, България
210	https://www.facebook.com/marketplace/np/item/2...	https://scontent.fsof4-1.fna.fbcdn.net/v/t39.3...	205 000 лв.	Двустаен апартамент в кв.Кръстова вада	София, България
211	https://www.facebook.com/marketplace/np/item/5...	https://scontent.fsof4-1.fna.fbcdn.net/v/t45.5...	17 лв.	Knauf Perlfix, Ceresit Thermo Universal	София, България

212 rows × 5 columns

Фигура 1.Използван набор от данни

4.2. Анализ на набора от данни

```
Out[400…   Link          object
           Photo Url     object
           Price         object
           Name          object
           City          object
           dtype: object
```

Фигура 2. Типове данни

Визуализирайки типовете (*Фигура 2*), забелязваме, че всички колони са от тип object и можем да пристъпим към предварителната обработка.

Преименуваме колоната *Name* на *Title* за по-акуратно репрезентиране на информацията и добавяме нова колона *Item Type*, която да съдържа информацията за типа на обявата (*Фигура 3*).

```
In [401…
data.rename(columns={'Name': 'Title'}, inplace=True)
```

```
In [402…
# Adding item type based on the title
def classify_item(name):
    if 'апартамент' in name.lower():
        return 'Apartment'
    elif 'къща' in name.lower():
        return 'House'
    elif 'легло' in name.lower():
        return 'Bed'
    else:
        return 'Others'

data['Item Type'] = data['Title'].apply(classify_item)
```

Фигура 3. Преименуване и създаване на колона за тип

Правим проверка за липсващи, безкрайни или нечислови стойности (*Фигура 4*) и виждаме, че в дейтасета такива няма.

In [403…
```
print("Check for missing values:")
print(data.isnull().sum())
```

```
Check for missing values:
Link         0
Photo Url    0
Price        0
Title        0
City         0
Item Type    0
dtype: int64
```

In [404…
```
print("Check for NaN values:")
print(data.isna().sum())
```

```
Check for NaN values:
Link         0
Photo Url    0
Price        0
Title        0
City         0
Item Type    0
dtype: int64
```

In [405…
```
print("Checking for infinite values:")
print((data == float('inf')).sum())
```

```
Checking for infinite values:
Link         0
Photo Url    0
Price        0
Title        0
City         0
Item Type    0
dtype: int64
```

Фигура 4. Проверка за липсващи, безкрайни или нечислови стойности

Правим проверка, дали успешно може да конвертираме цената към числен тип, като преди това почистваме редовете от ненужни символи (*Фигура 5*).

In [406…
```
data.dtypes
```

Out[406…
```
Link         object
Photo Url    object
Price        object
Title        object
City         object
Item Type    object
dtype: object
```

In [407…
```
# Price cleanup
def clean_price(price_str):
    # Remove "лв.", commas, etc.
    price_str = price_str.replace(' лв.', '').replace(' ', '').replace(',', '')

    try:
        return float(price_str)
    except ValueError:
        # In case we cannot convert, return NaN
        return np.nan
```

Фигура 5. Почистване на Price колоната

Уеднаквяваме стойностите на колоната *City* с цел събиране на обявите, където местоположението съвпада, но градът в обявата е написан на кирилица вместо на латиница. (*Фигура 6*).

In [408…

```python
# Cities cleanup
cyrillic_to_latin = {
    'а': 'a', 'б': 'b', 'в': 'v', 'г': 'g', 'д': 'd',
    'е': 'e', 'ж': 'zh', 'з': 'z', 'и': 'i', 'й': 'y',
    'к': 'k', 'л': 'l', 'м': 'm', 'н': 'n', 'о': 'o',
    'п': 'p', 'р': 'r', 'с': 's', 'т': 't', 'у': 'u',
    'ф': 'f', 'х': 'h', 'ц': 'ts', 'ч': 'ch', 'ш': 'sh',
    'щ': 'sht', 'ъ': 'a', 'ь': 'y', 'ю': 'yu', 'я': 'ya'
}

data['City'] = data['City'].str.replace('България', '').str.strip()
data['City'] = data['City'].str.replace('Bulgaria', '').str.strip()
data['City'] = data['City'].str.replace(',', '')
data['City'] = data['City'].str.lower()

for key in cyrillic_to_latin.keys():
    data['City'] = data['City'].str.replace(key, cyrillic_to_latin[key])

data['City'] = data['City'].str.capitalize()

data['City']
```

Out[408…

```
0       Botevgrad
1          Sofiya
2           Svoge
3       Kostenets
4          Sofiya
           ...
207    Kyustendil
208        Sofiya
209        Sofiya
210        Sofiya
211        Sofiya
Name: City, Length: 212, dtype: object
```

Фигура 6. Уеднаквяване на градовете в таблицата

Правим последно почистване на дейтасета като премахваме редове, където липсват цена или град и официално конвертираме цената към числов тип (*Фигура 7*) като по този начин сме приключили с предварителната обработка на данните (*Фигура 8)*.

In [477…

```
# --- Data Cleanup ---

# Remove rows with missing Price or City data
data = data.dropna(subset=['Price', 'City'])

# Character cleanup
data['Price'] = data['Price'].replace(r'[^\d]', '', regex=True)

# Numeric convertion
data['Price'] = pd.to_numeric(data['Price'], errors='coerce')
```

In [478…

```
data
```

Out[478…

	Link	Photo Url	Price	Title	City	Item Type
0	https://www.facebook.com/marketplace/np/item/1...	https://scontent.fsof4-1.fna.fbcdn.net/v/t39.3...	160000	Апартамент	Botevgrad	Apartment
1	https://www.facebook.com/marketplace/np/item/5...	https://scontent.fsof4-1.fna.fbcdn.net/v/t45.5...	180	Lorelli бебешко легло 60/120см matrix new	Sofiya	Bed
2	https://www.facebook.com/marketplace/np/item/8...	https://scontent.fsof4-1.fna.fbcdn.net/v/t39.3...	125000	Продавам къща в село Миланово област София.	Svoge	House
3	https://www.facebook.com/marketplace/np/item/1...	https://scontent.fsof4-1.fna.fbcdn.net/v/t39.3...	139500	Къща в с. Костенец	Kostenets	House
4	https://www.facebook.com/marketplace/np/item/1...	https://scontent.fsof4-1.fna.fbcdn.net/v/t39.3...	128960	Двустаен апартамент в Надежда	Sofiya	Apartment
...	...	...	...	...	...	...
207	https://www.facebook.com/marketplace/np/item/8...	https://scontent.fsof4-1.fna.fbcdn.net/v/t45.5...	150	Фритюрник 5 л. Bereket \nC чугунена горелка.	Kyustendil	Others
208	https://www.facebook.com/marketplace/np/item/1...	https://scontent.fsof4-1.fna.fbcdn.net/v/t39.3...	106000	Продава апаратамент в Малинова долина	Sofiya	Others
209	https://www.facebook.com/marketplace/np/item/1...	https://scontent.fsof4-1.fna.fbcdn.net/v/t45.5...	95	Премиум ПВЦ мраморни панели	Sofiya	Others
210	https://www.facebook.com/marketplace/np/item/2...	https://scontent.fsof4-1.fna.fbcdn.net/v/t39.3...	205000	Двустаен апартамент в кв.Кръстова вада	Sofiya	Apartment
211	https://www.facebook.com/marketplace/np/item/5...	https://scontent.fsof4-1.fna.fbcdn.net/v/t45.5...	17	Knauf Perlfix, Ceresit Thermo Universal	Sofiya	Others

212 rows × 6 columns

Фигура 7. Обновен дейтасет

In [479…

```
data.dtypes
```

Out[479…

```
Link         object
Photo Url    object
Price         int64
Title        object
City         object
Item Type    object
dtype: object
```

Фигура 8. Изчистен и готов файл за анализ

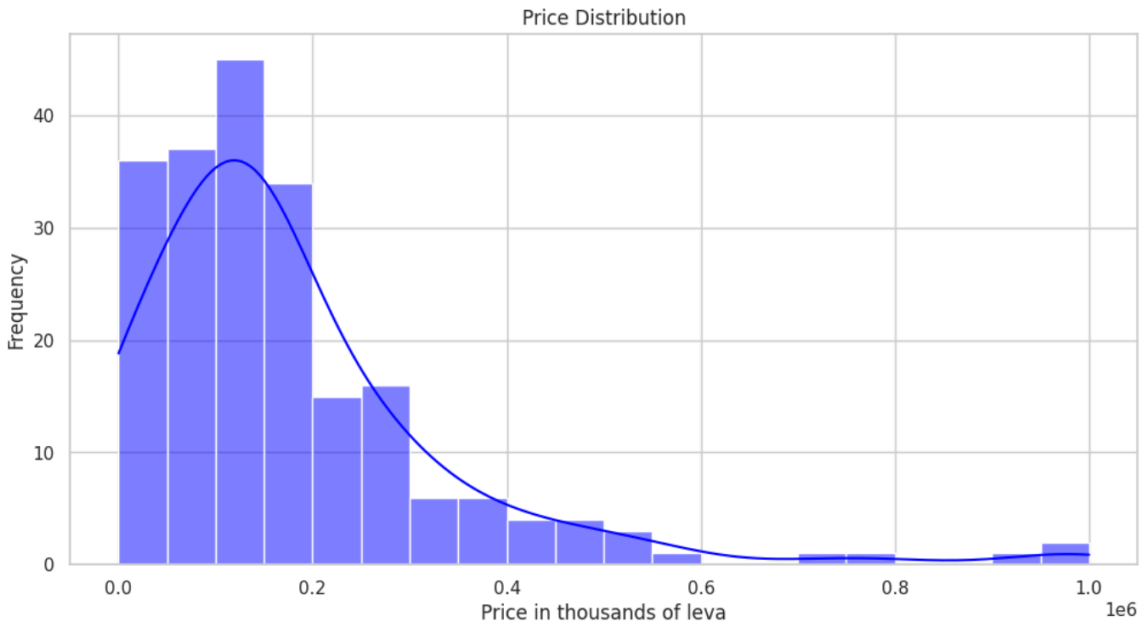
5. Представяне и визуализация на резултатите

5.1. Регионален ценови анализ

In [480…

```
# --- Data Analysis ---

# Price Distribution
plt.figure()
sns.histplot(data['Price'], bins=20, kde=True, color='blue')
plt.title('Price Distribution')
plt.xlabel('Price in thousands of leva')
plt.ylabel('Frequency')
plt.show()
```



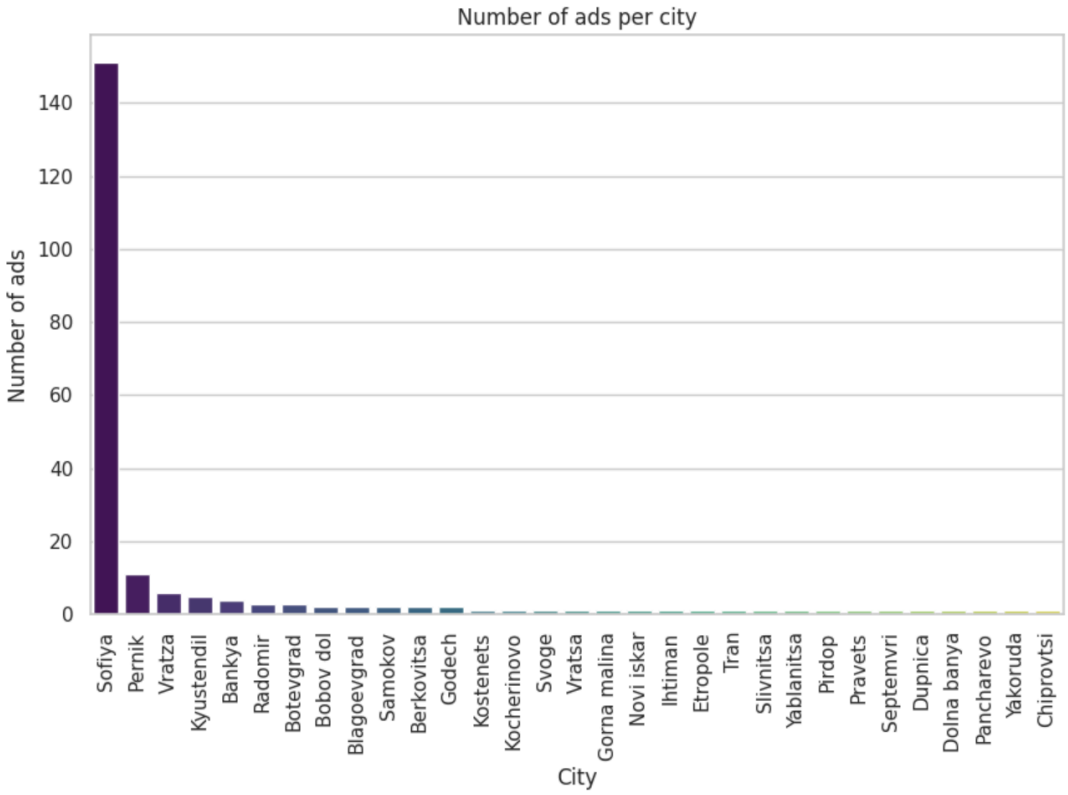
Фигура 9. Ценово разпределение

Най-висока е средната цена на обявите в град Чипровци, като това се дължи на единствената обява в размер на *721 721лв*. Изненадваща е позицията на град София, който се намира на едва пето място въпреки огромното превъзходство в броя на обявите с които разполага (*фигура 10)*. Град като Перник, например, постига по-добър резултат от София с над 10 пъти по-малък брой обяви (*Фигура 11*).

In [485…

```
city_counts = data['City'].value_counts()

plt.figure(figsize=(10, 6))
sns.barplot(x=city_counts.index, y=city_counts.values, hue=city_counts.index, dodge=False, palette='viridis', legend=False)
plt.title('Number of ads per city')
plt.xlabel('City')
plt.ylabel('Number of ads')
plt.xticks(rotation=90)
plt.show()
```

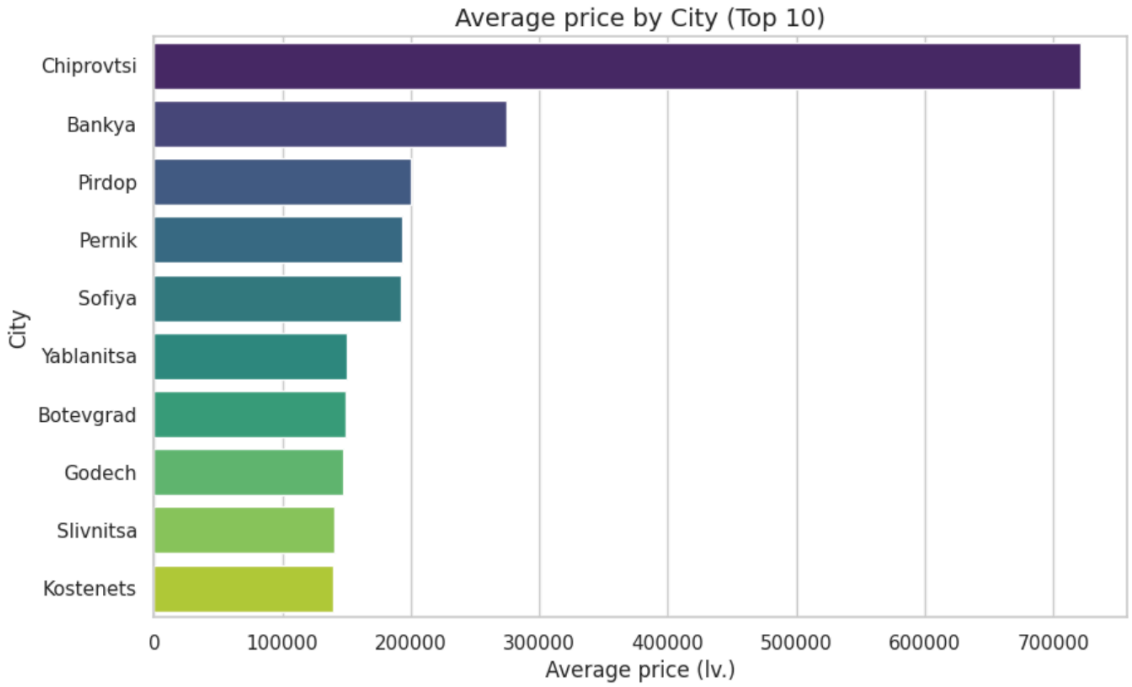


Фигура 10. Брой обяви за град

In [482…

```
# Average price by City (top 10 cities)
mean_prices_by_city = data.groupby('City')['Price'].mean().sort_values(ascending=False).head(10)

plt.figure(figsize=(10, 6))
sns.barplot(x=mean_prices_by_city.values, y=mean_prices_by_city.index, hue=mean_prices_by_city.index, palette='viridis')
plt.title('Average price by City (Top 10)', fontsize=14)
plt.xlabel('Average price (lv.)', fontsize=12)
plt.ylabel('City', fontsize=12)
plt.show()
```



Фигура 11. Средна цена на актив за град

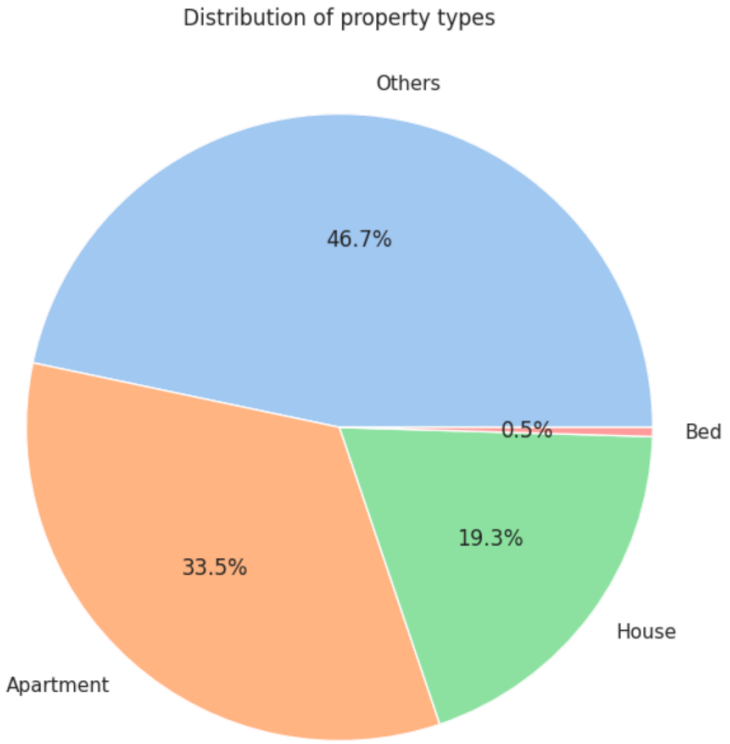
5.2 Типов анализ

С цел правилната обработка на събраната информация и определянето на типа на обявите, можем да използваме колоната *Item Type*, която е създадена за тази цел (*Фигура 3*). След като начертаем графиката (*Фигура 12*) виждаме, че след некласифицираните “други” в една от 3-те големи кофи обяви, на второ място се подреждат обявите за апартаменти, следвани от тези за къщи и накрая - тези за легла. Вземайки предвид сумата на броя обяви от тип апартамент и тип къща можем да заключим, че най-често срещани са обявите за продажба на жилища, които образуват *52.8%* от всички обяви.

In [623…

```
item_type_counts = data['Item Type'].value_counts()

plt.figure(figsize=(8, 8))
plt.pie(item_type_counts, labels=item_type_counts.index, autopct='%1.1f%%', colors=sns.color_palette('pastel'))
plt.title('Distribution of property types')
plt.show()
```



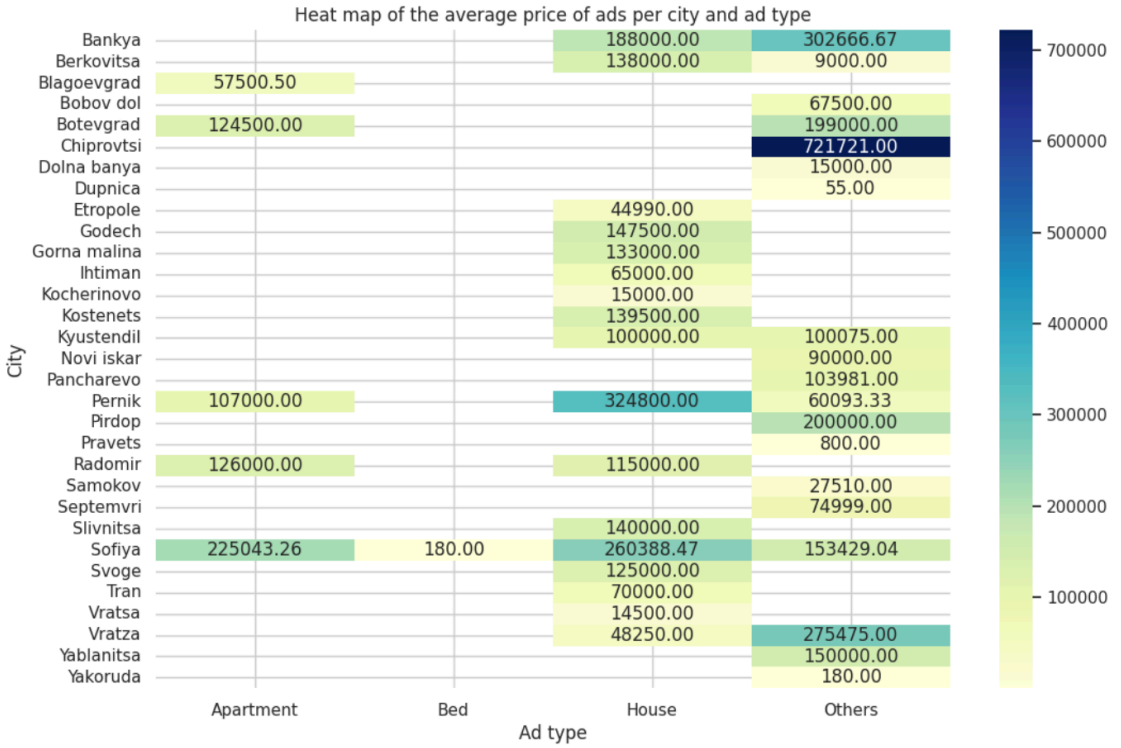
Фигура 12. Разпределение спрямо типа актив

Чрез помощта на т.н “топлинна карта” и можем да проследим най-скъпите активи в градовете и съответно към кой от 4-те главни типа спадат. (*Фигура 13*)

In [487…

```
# Pivot table for average price of ads per city and property type
pivot_table = data.pivot_table(values='Price', index='City', columns='Item Type', aggfunc='mean')

# Heat map
plt.figure(figsize=(12, 8))
sns.heatmap(pivot_table, annot=True, fmt=".2f", cmap='YlGnBu')
plt.title('Heat map of the average price of ads per city and ad type')
plt.xlabel('Ad type')
plt.ylabel('City')
plt.show()
```



Фигура 13. Средна цена спрямо населено място и тип на актива

5.3 Синтактичен анализ

След допълнително почистване на колоната *Title*, можем да създадем списък с всички думи, който да ни даде допълнителна информация за това кои са най-често срещаните (*Фигури 14, 15*). Очаквано предлогът “в” е на първо място, като се среща на цели *93* места. Изненадващо обаче е присъствието на думите “апартамент”, “къща” и “тристаен” съответно на второ, трето и четвърто място. Те разделят “в” от останалите два предлога “с” и “на”, които са на пета и шеста позиция.

In [483…

```
from collections import Counter
import re

# Remove leading and following whitespace
data['Title'] = data['Title'].str.strip()

# Text conversion into lowercase and unnecessary symbols removal
data['Title'] = data['Title'].str.lower().apply(lambda x: re.sub(r'[^\w\s]', '', x))

# Splitting the text into words and creating a list
all_words = ' '.join(data['Title'].tolist()).split()

# Word frequency
word_counts = Counter(all_words)

# Sort by frequency
common_words = word_counts.most_common()

print("Top 10 most frequently used words in the 'Title' column:")
for word, count in common_words[:10]:
    print(f"{word}: {count}")
```

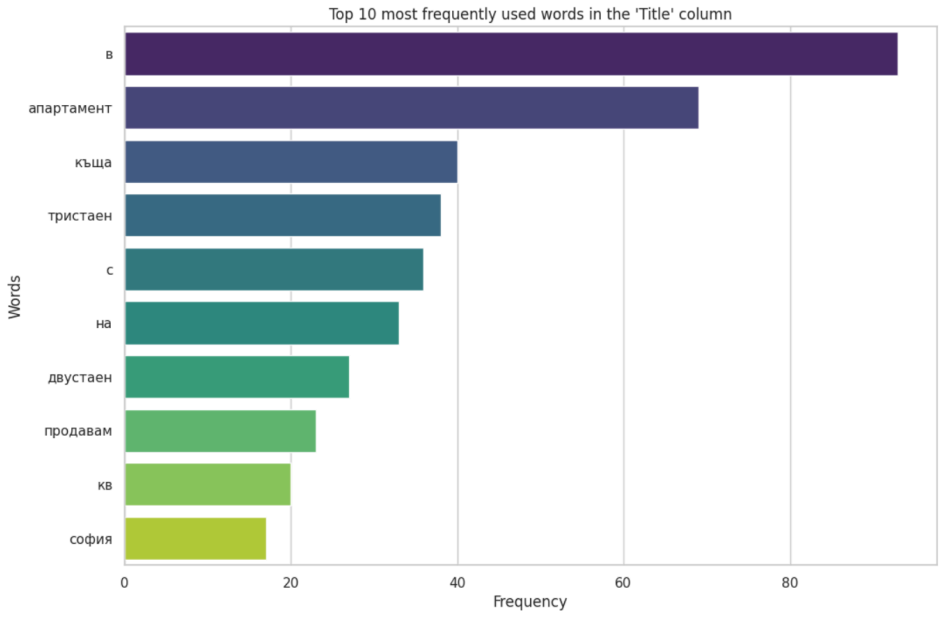
```
Top 10 most frequently used words in the 'Title' column:
в: 93
апартамент: 69
къща: 40
тристаен: 38
с: 36
на: 33
двустаен: 27
продавам: 23
кв: 20
София: 17
```

Фигура 14. Десетте най-често срещани думи

In [484…

```
# Zipping the words and their frequency
words, counts = zip(*common_words[:10])

# Bar plot creation
plt.figure(figsize=(12, 8))
sns.barplot(x=list(counts), y=list(words), hue=list(words), palette='viridis')
plt.xlabel('Frequency')
plt.ylabel('Words')
plt.title("Top 10 most frequently used words in the 'Title' column")
plt.show()
```



Фигура 15. Графично представяне на десетте най-често срещани думи

6. Заключение

Анализът на събраните данни от Фейсбук продажби разкрива, че поради ограничения обем и качество на данни, извличането на задълбочени изводи е трудоемка задача. Липсата на ключова информация като детайлни категории за продукти, демографски данни за купувачи или точни времеви периоди ограничава възможността за анализ на тенденции и потребителско поведение.

Въпреки това, наличните данни позволяват идентифицирането на основни модели, като най-продавани категории и средни ценови нива. За по-задълбочен анализ се препоръчва обогатяване на данните с допълнителна информация и повече на брой записи.

7. Сорс код

https://github.com/savina01/FB_Marketplace_Analysis/

8. Информационни източници

https://www.webfx.com/blog/social-media/social-media-analysis/