

SQL EDA: customers & marketing campaigns

[Вступление](#)

[Источник данных](#)

[Код схемы](#)

[Описание данных](#)

[Начальная работа с данными](#)

[Создание и редактирование полей](#)

[Исследование данных по категориям](#)

[Визуализация](#)

[Выводы](#)

[Ссылки](#)

[Контакты](#)

Вступление

Нам представлена таблица, содержащая данные о клиентах сети супермаркетов и их реакции на 5 маркетинговых кампаний. Данные сгенерированы искусственно.

С помощью SQL мы произведем разведывательный анализ данных и определим группы клиентов, наиболее и наименее активно реагирующих на маркетинговые кампании.

Источник данных

Датасет доступен на сайте Kaggle.com по [ссылке](#)

Код схемы


▼ Код для создания таблицы

```

CREATE TABLE campaigns
(
  user_id      INT PRIMARY KEY,
  year_birth   INT,
  education    VARCHAR(512),
  marital_status VARCHAR(512),
  income       FLOAT,
  kidhome     INT,
  teenhome     INT,
  dt_enrolled  DATE,
  recency      INT,
  amt_wine     INT,
  amt_fruit    INT,
  amt_meat     INT,
  amt_fish     INT,
  amt_sweet    INT,
  amt_gold     INT,
  discount_purchases INT,
  web_purchases INT,
  catalog_purchases INT,
  store_purchases INT,
  web_visits_mnth INT,
  cmp1_accepted INT,
  cmp2_accepted INT,
  cmp3_accepted INT,
  cmp4_accepted INT,
  cmp5_accepted INT,
  complain     INT,
  response     INT
);

```

▼ Cxema

campaigns	
user_id 	integer
year_birth	integer
education	varchar(512)
marital_status	varchar(512)
income	float
kidhome	integer
teenhome	integer
dt_enrolled	date
recency	integer
amt_wine	integer
amt_fruit	integer
amt_meat	integer
amt_fish	integer
amt_sweet	integer
amt_gold	integer
discount_purchases	integer
web_purchases	integer
catalog_purchases	integer
store_purchases	integer
web_visits_mnth	integer
cmp1_accepted	integer
cmp2_accepted	integer
cmp3_accepted	integer
cmp4_accepted	integer
cmp5_accepted	integer
complain	integer
response	integer

Описание данных

- **user_id:** уникальный идентификатор клиента
- **year_birth:** год рождения клиента
- **education:** уровень образования клиента
- **marital_status:** семейное положение клиента
- **income:** годовой доход клиента
- **kidhome:** количество детей младшего возраста у клиента
- **teenhome:** количество детей среднего и старшего возраста у клиента
- **dt_enrolled:** дата включения клиента в программу лояльности
- **recency:** количество дней с последней покупки клиента
- **complain:** если клиент подавал жалобу за последние 2 года - 1, иначе - 0
- **amt_wine:** сумма, потраченная клиентом на вино за последние 2 года
- **amt_fruit:** сумма, потраченная клиентом на фрукты за последние 2 года
- **amt_meat:** сумма, потраченная клиентом на мясные продукты за последние 2 года
- **amt_fish:** сумма, потраченная клиентом на рыбу за последние 2 года
- **amt_sweet:** сумма, потраченная клиентом на сладости за последние 2 года
- **amt_gold:** сумма, потраченная клиентом на продукты премиум категории за последние 2 года
- **discount_purchases:** количество покупок со скидкой
- **web_purchases:** количество покупок через вебсайт компании
- **catalog_purchases:** количество покупок с использованием каталога
- **store_purchases:** количество покупок непосредственно в магазине
- **web_visits_mnth:** количество посещений вебсайта за последний месяц

- **cmp1_accepted:** 1 - если клиент принял предложение в рамках кампании 1, 0 - если не принял
- **cmp2_accepted:** 1 - если клиент принял предложение в рамках кампании 2, 0 - если не принял
- **cmp3_accepted:** 1 - если клиент принял предложение в рамках кампании 3, 0 - если не принял
- **cmp4_accepted:** 1 - если клиент принял предложение в рамках кампании 4, 0 - если не принял
- **cmp5_accepted:** 1 - если клиент принял предложение в рамках кампании 5, 0 - если не принял
- **response:** 1 - если клиент принял хотя бы одно предложение в рамках одной из кампаний

Начальная работа с данными

Создание и редактирование полей

Первым делом, создадим новые колонки на основе существующих.

1. Добавим колонку **age** с возрастом клиентов, используя поле **year_birth**.

```
ALTER TABLE campaigns  
ADD COLUMN age INTEGER;
```

```
UPDATE campaigns  
SET age = DATE_PART('YEAR', NOW()) - year_birth;
```

2. Добавим колонку **total_kids** с общим количеством детей клиента, используя поля **kidhome** и **teenhome**.

```
ALTER TABLE campaigns
```

```
ADD COLUMN total_kids INTEGER;
```

```
UPDATE campaigns  
SET total_kids = kidhome + teenhome;
```

3. Колонка **total_spent** будет хранить общие траты клиента без разбивки на категорию продукта

```
ALTER TABLE campaigns  
ADD COLUMN total_spent INT;
```

```
UPDATE campaigns  
SET total_spent = amt_wine + amt_fruit + amt_meat + amt_fish + amt_others;
```

4. Колонка **total_purchases** будет отображать общее количество покупок клиента, без разбивки на способ покупки.

```
ALTER TABLE campaigns  
ADD COLUMN total_purchases INT;
```

```
UPDATE campaigns  
SET total_purchases = discount_purchases + web_purchases + store_purchases;
```

5. Колонка **marital_status** содержит ряд неоднозначных и/или дублирующих значений. Приведем их к единому значению.

```
UPDATE campaigns  
SET marital_status = 'Single'  
WHERE marital_status IN ('YOLO', 'Absurd', 'Alone');
```

7. Колонка с возрастом содержит подозрительные значения, которые мы удалим.

```
DELETE FROM campaigns
WHERE age > 100;
```

Исследование данных по категориям

Средние доходы и сумма расходов на покупки по уровню образования и семейному положению:

```
SELECT education, marital_status,
ROUND(AVG(income)::NUMERIC, 2) AS avg_income,
SUM(total_spent) AS total_spendings
FROM campaigns
GROUP BY education, marital_status
ORDER BY total_spendings DESC;
```

	education character varying (512) 🔒	marital_status character varying (512) 🔒	avg_income numeric 🔒	total_spendings bigint 🔒
1	Graduation	Married	50800.26	258030
2	Graduation	Together	55758.48	188468
3	Graduation	Single	51365.63	155075
4	PhD	Married	58138.03	137439
5	Master	Married	53286.03	78200
6	PhD	Together	55802.37	74146
7	Graduation	Divorced	54526.04	73353
8	PhD	Single	53039.67	61300
9	Master	Together	52109.01	59450
10	Master	Single	53787.14	57754

Группа клиентов с **наибольшими расходами** - женатые люди, получившее образование (Graduation)

Группа клиентов с **наибольшими средними доходами** - люди, получившее образование (Graduation), находящиеся в отношениях (но не в браке)

Средние доходы и сумма расходов на покупки по количеству детей:

```
SELECT total_kids,  
       ROUND(AVG(income)::NUMERIC, 2) AS avg_income,  
       SUM(total_spent) AS total_spendings  
FROM campaigns  
GROUP BY total_kids  
ORDER BY total_spendings DESC;
```

	total_kids integer	avg_income numeric	total_spendings bigint
1	0	65677.36	703794
2	1	47712.01	533156
3	2	44612.31	103544
4	3	46677.00	14554

Группа клиентов с **наибольшим средним доходом и наибольшими расходами на покупки** - люди без детей

Средние доходы и сумма расходов по возрастной группе:

```
SELECT CASE WHEN age BETWEEN 20 AND 30 THEN '20-30'  
           WHEN age BETWEEN 31 AND 40 THEN '30-40'
```



```

    WHEN age BETWEEN 41 AND 50 THEN '40-50'
    WHEN age BETWEEN 51 AND 60 THEN '50-60'
    WHEN age BETWEEN 61 AND 70 THEN '60-70'
    ELSE '70+' END AS age_group,
    ROUND(AVG(income)::NUMERIC, 2) AS avg_income,
    SUM(total_spent) AS total_spendings
FROM campaigns
GROUP BY age_group
ORDER BY total_spendings DESC;

```

	age_group text	avg_income numeric	total_spendings bigint
1	50-60	51441.32	372625
2	60-70	56431.57	327972
3	40-50	49776.88	306047
4	70+	59002.95	198960
5	30-40	44707.29	141128
6	20-30	58295.40	8316

Группа-лидер по **общим расходам** - люди в возрасте 50-60 лет, а по **среднему уровню доходов** лидируют люди 20-30 лет

Самые активные посетители веб-сайта

```

SELECT education, marital_status,
       SUM(web_visits_mnth) AS total_web_visits
FROM campaigns
GROUP BY education, marital_status
ORDER BY total_web_visits DESC;

```

	education character varying (512) 🔒	marital_status character varying (512) 🔒	total_web_visits bigint 🔒
1	Graduation	Married	2334
2	Graduation	Together	1481
3	Graduation	Single	1340
4	PhD	Married	1007
5	Master	Married	721
6	Graduation	Divorced	636
7	PhD	Together	620
8	Master	Together	540
9	PhD	Single	523
10	2n Cycle	Married	432

Наиболее активно за последний месяц веб-сайт посещали женатые люди, получившее образование

Самый популярный метод совершения покупки

```
SELECT CASE WHEN age BETWEEN 20 AND 30 THEN '20-30'
  WHEN age BETWEEN 31 AND 40 THEN '30-40'
  WHEN age BETWEEN 41 AND 50 THEN '40-50'
  WHEN age BETWEEN 51 AND 60 THEN '50-60'
  WHEN age BETWEEN 61 AND 70 THEN '60-70'
  ELSE '70+' END AS age_group,
SUM(discount_purchases) AS discount_pur_sum,
SUM(web_purchases) AS web_pur_sum,
SUM(catalog_purchases) AS catalog_pur_sum,
SUM(store_purchases) AS store_pur_sum
FROM campaigns
GROUP BY age_group;
```

	age_group text	discount_pur_sum bigint	web_pur_sum bigint	catalog_pur_sum bigint	store_pur_sum bigint
1	70+	561	1229	910	1747
2	30-40	429	824	587	1315
3	60-70	1205	2117	1492	3009
4	40-50	1327	2238	1336	3175
5	20-30	12	33	39	69
6	50-60	1671	2702	1592	3647

Наиболее **популярный метод покупки** - непосредственно в магазине, для всех возрастных групп

Оценка response rate

```
SELECT education,
       marital_status,
       total_kids,
       ROUND((SUM(response)::NUMERIC /COUNT(user_id)::NUMERIC), 2)
FROM campaigns
GROUP BY education, marital_status, total_kids
ORDER BY response_rate DESC;
```

	education character varying (512)	marital_status character varying (512)	total_kids integer	response_rate numeric
1	2n Cycle	Divorced	0	0.67
2	Master	Widow	0	0.60
3	PhD	Divorced	0	0.55
4	Master	Single	0	0.50
5	Master	Widow	1	0.50

Наиболее активно реагировали на кампании люди во втором цикле высшего образования (в зависимости от страны, это может быть бакалавриат или докторская степень) в разводе и без детей

Самая популярная категория продуктов по сумме расходов

```
SELECT *
FROM
  (SELECT 'amt_wine' AS product_group,
    SUM(amt_wine) AS amt
  FROM campaigns
  UNION ALL
  SELECT 'amt_fruit' AS product_group,
    SUM(amt_fruit) AS amt
  FROM campaigns
  UNION ALL
  SELECT 'amt_meat' AS product_group,
    SUM(amt_meat) AS amt
  FROM campaigns
  UNION ALL
  SELECT 'amt_fish' AS product_group,
    SUM(amt_fish) AS amt
  FROM campaigns
  UNION ALL
  SELECT 'amt_sweet' AS product_group,
    SUM(amt_sweet) AS amt
  FROM campaigns
  UNION ALL
  SELECT 'amt_gold' AS product_group,
    SUM(amt_gold) AS amt
  FROM campaigns)
ORDER BY amt DESC;
```

	product_group text	amt bigint
1	amt_wine	680038
2	amt_meat	373393
3	amt_gold	98358
4	amt_fish	83939
5	amt_sweet	60553
6	amt_fruit	58767

Самая популярная категория продуктов - вино, на нее потратили почти вдвое больше, чем на вторую по популярности категорию, мясо

Самая успешная кампания

```
SELECT *
FROM
(SELECT 'campaign1' AS campaign,
      SUM(cmp1_accepted) AS cmp_acceptance
FROM campaigns
UNION ALL
SELECT 'campaign2' AS campaign,
      SUM(cmp2_accepted) AS cmp_acceptance
FROM campaigns
UNION ALL
SELECT 'campaign3' AS campaign,
      SUM(cmp3_accepted) AS cmp_acceptance
FROM campaigns
UNION ALL
SELECT 'campaign4' AS campaign,
      SUM(cmp4_accepted) AS cmp_acceptance
FROM campaigns
UNION ALL
SELECT 'campaign5' AS campaign,
      SUM(cmp5_accepted) AS cmp_acceptance
```

```
FROM campaigns)
ORDER BY cmp_acceptance DESC;
```

	campaign text	cmp_acceptance bigint
1	campaign4	167
2	campaign3	163
3	campaign5	162
4	campaign1	144
5	campaign2	30

Самой успешной кампанией была четвертая, в рамках ее 167 клиентов приняли маркетинговые предложения

Визуализация

Для визуализации данных было создано подключение к базе данных PSQL в Jupyter Notebook, датасет был загружен в датафрейм.

1. Соотношение семейного положения, образования и доходов, расходов, количества покупок и реакций на кампании

▼ Код

```
metrics = ['income', 'total_spent', 'total_purchases', 'resp
features = ['education', 'marital_status']

fig = plt.figure(figsize=(15, 10))
gs = GridSpec(ncols = 2, nrows = 4, figure = fig)
c = 0
r = 0

for feature in features:
    for i, metric in enumerate(metrics):
```

```

grouped = df.groupby('education', as_index = False)[1]
plt.subplot(gs[c, r])
ax = sns.barplot(data = df, x = df[metric], y = df[fe
c+=1

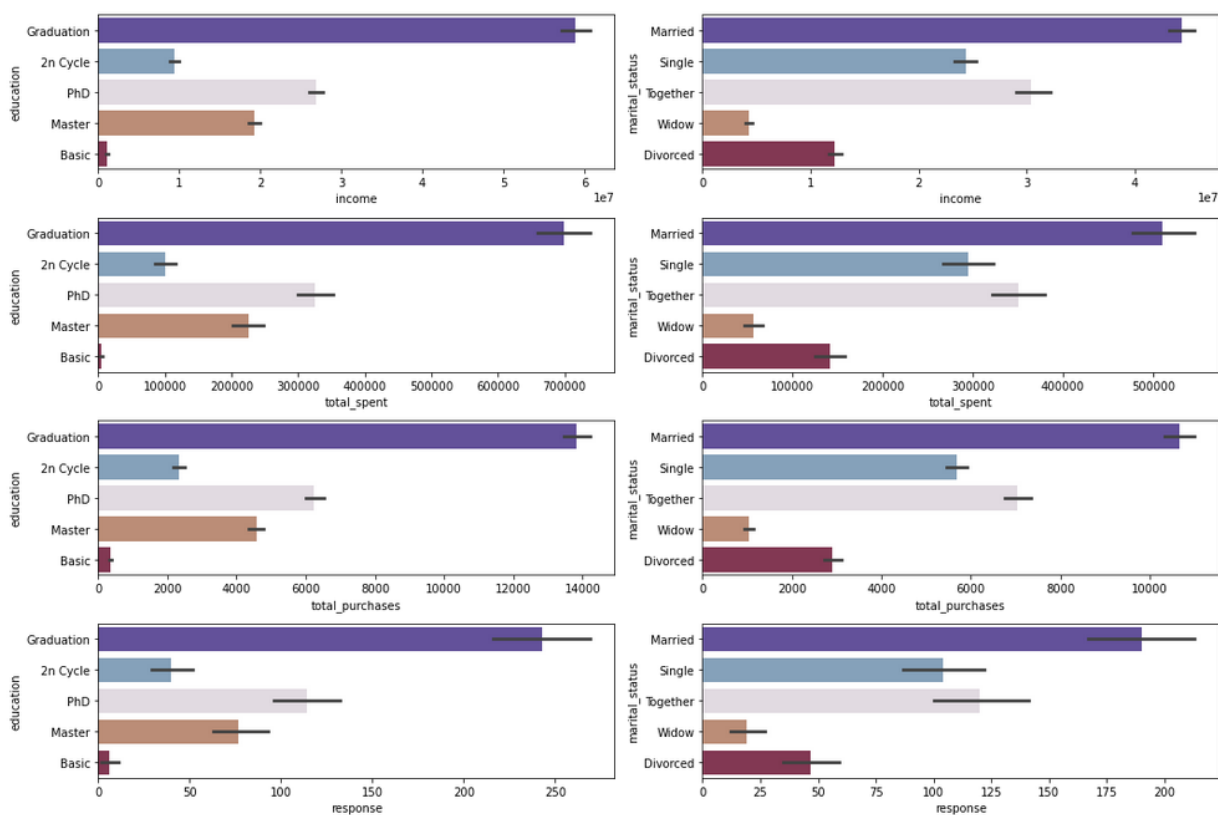
c=0
r+=1

```

```

plt.tight_layout()
plt.show()

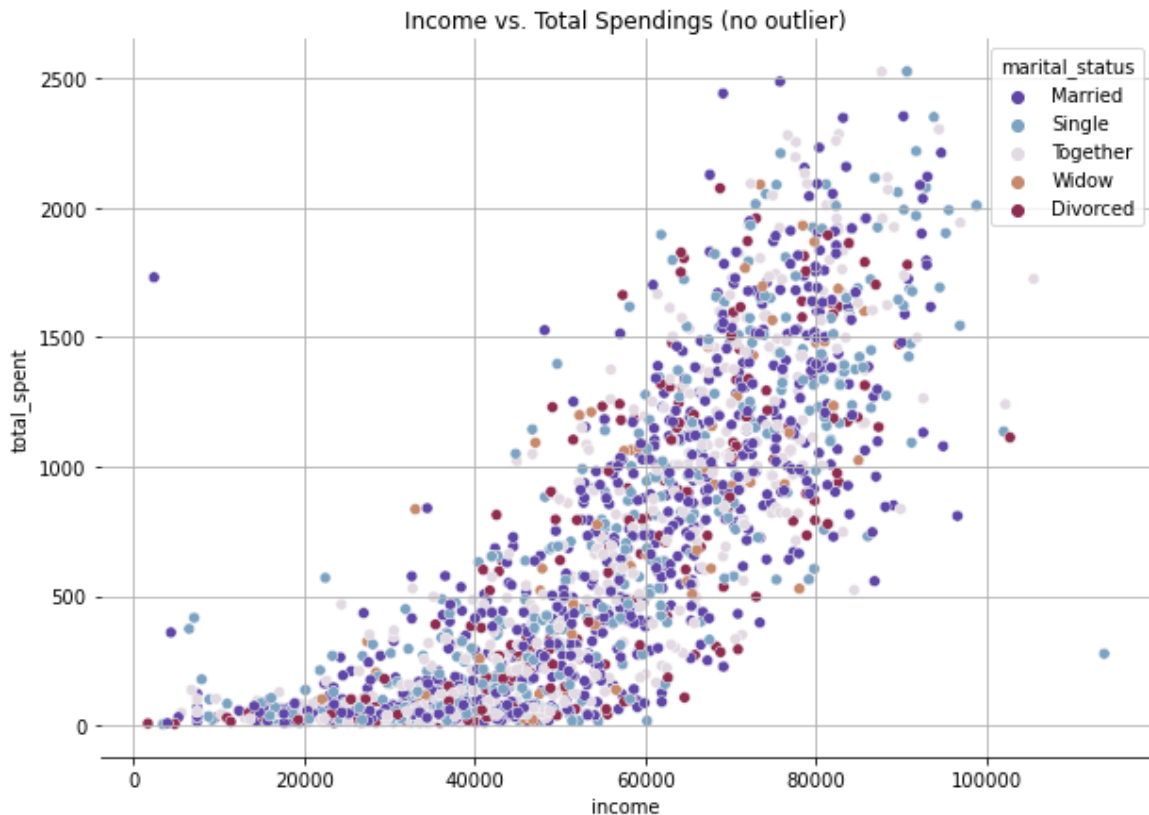
```



2. Соотношение доходов и расходов

▼ Код

```
plt.figure(figsize=(10,7))
sns.scatterplot(data = df[df['income'] < df['income'].mean()
plt.grid(True)
plt.title
sns.despine(left=True)
```



С ростом доходов растут и расходы на покупки

3. Соотношение доходов и количества покупок

▼ Код

```
fig = plt.figure(figsize=(10,7))
ax = sns.scatterplot(data = df[df['income'] < df['income'].me
plt.grid(True)
```



```
sns.despine(left=True)
plt.title('Income vs. Total Purchases (no outlier)')
```

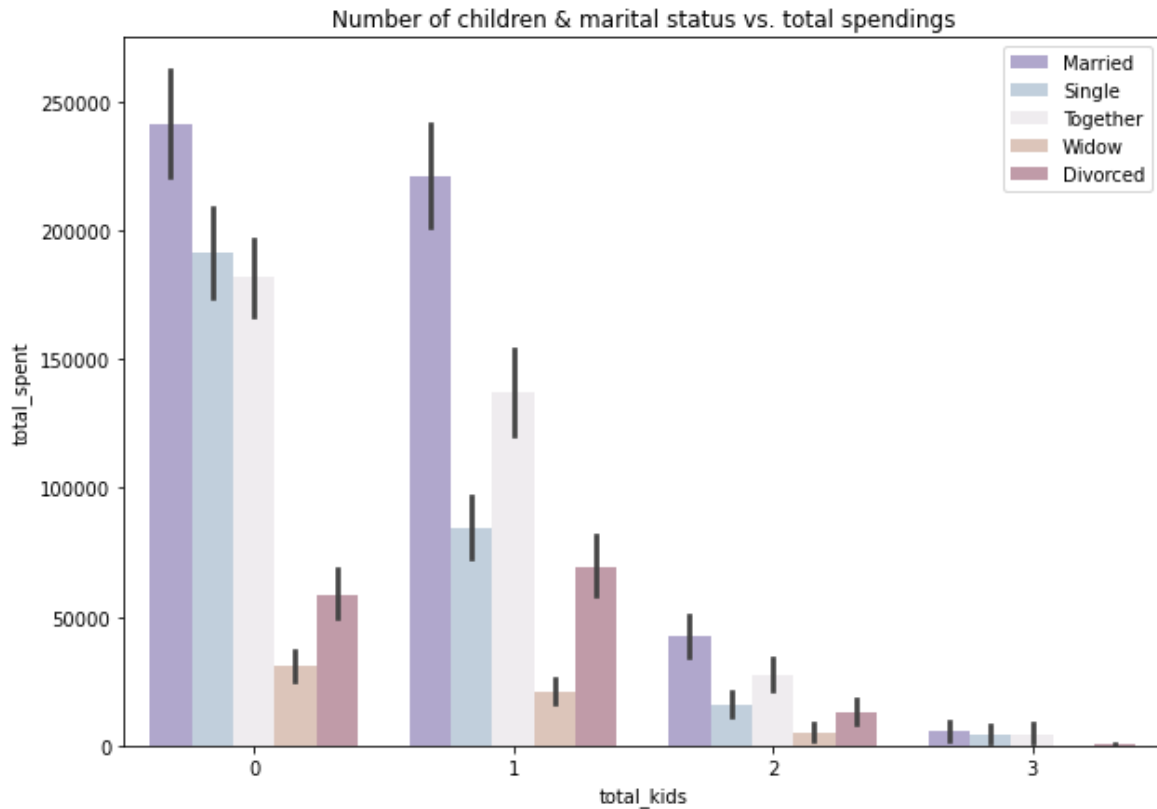


С ростом доходов растет и количество покупок

4. Количество детей + семейное положение и общие расходы

▼ Код

```
fig = plt.figure(figsize = (10,7))
ax = sns.barplot(data = df, x = 'total_kids', y = 'total_sper
ax.legend(loc='upper right')
plt.title('Number of children & marital status vs. total sper
```

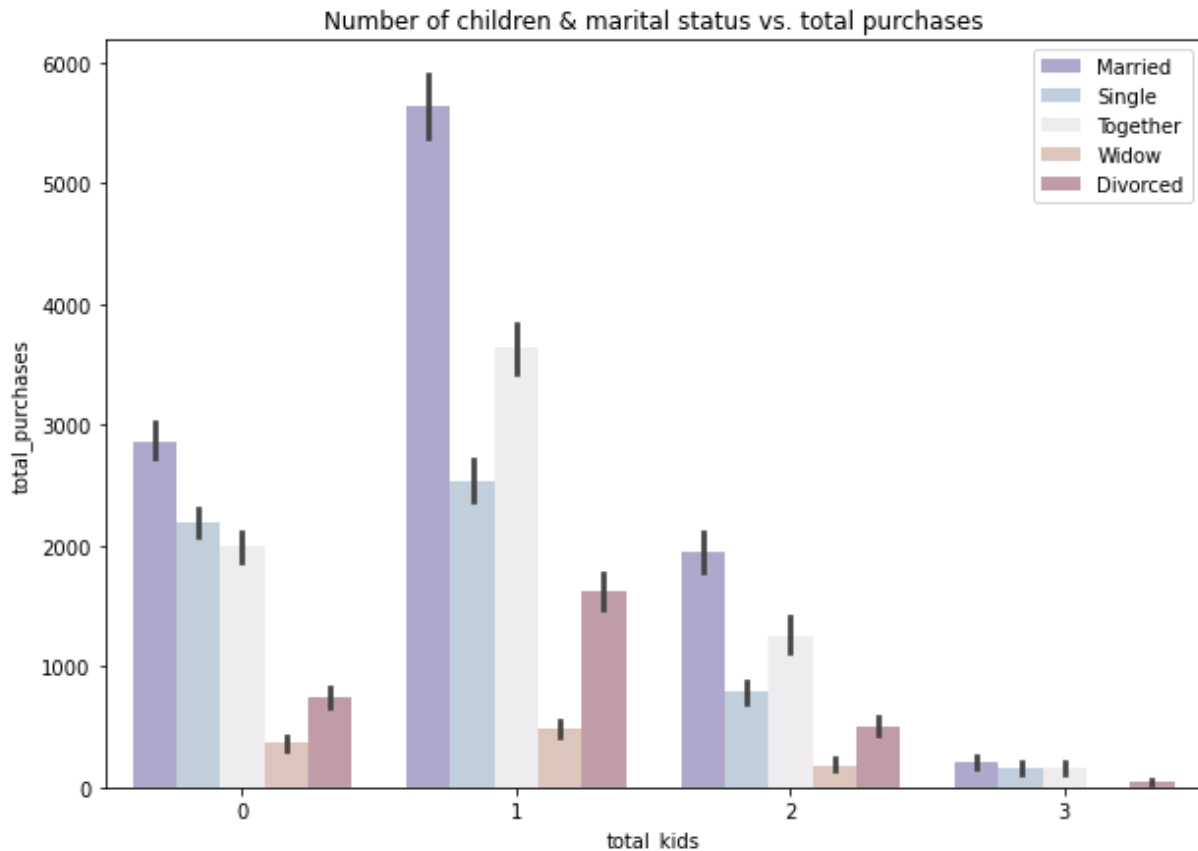


Самые высокие расходы на покупки - у женатых людей без детей

5. Количество детей + семейное положение и количество покупок

▼ Код

```
fig = plt.figure(figsize = (10,7))
ax = sns.barplot(data = df, x = 'total_kids', y = 'total_purch')
ax.legend(loc='upper right')
```



Больше всего покупок совершали женатые люди с 1 ребенком

Выводы

В рамках этого исследования данных мы проанализировали, насколько эффективными были 5 маркетинговых кампаний среди разных категорий пользователей, а также какие методы и категории покупок обладают наибольшей популярностью среди клиентов компании.

1. Самая эффективная кампания - кампания №5. На нее отклик был у 167 клиентов.
2. Самые активно откликнувшиеся на кампании клиенты - люди во втором цикле высшего образования (в зависимости от страны, это может быть бакалавриат или докторская степень) в разводе и без детей.
3. Группа клиентов с наибольшим средним доходом и наибольшими расходами на покупки - люди без детей.

4. Наиболее популярный метод покупок среди всех возрастных групп - покупка непосредственно в магазине.
5. Наиболее активно за последний месяц веб-сайт посещали женатые люди, получившее образование.
6. Категория продуктов, на которую за последние 2 года клиенты потратили больше всего, - это вино.

Было бы интересно оценить более полную картину, имея отдельные датасеты по клиентам, их покупкам, самим маркетинговым кампаниям.

Ссылки

[Ссылка](#) на полный блокнот Jupyter (EN)

[Ссылка](#) на полный блокнот Jupyter (RU)

Контакты

lianazaripovar@gmail.com

tg: @zaripova_liana