

Финальный проект DA: Анализ сайта "СберАвтоподписка"

Для анализа представлены две таблицы с данными:

1. **Sessions** – таблицами с данными о визитах на сайт
2. **Hits** - таблица с данными о событиях в рамках каждого визита

РАЗВЕДОЧНЫЙ АНАЛИЗ

В рамках исследования было проверено:

- наличие дубликатов – в обеих таблицах нет дубликатов
- типы данных: в признаке `visit_datetime` в таблице `sessions`, а также в признаке `hit_date` из таблицы `hits` был изменен типа данных на `datetime`

Были созданы новые признаки:

1. **month** в таблице `sessions`
2. **target_action** в таблице `hits` – по наличию целевого действия
3. **traffic** в таблице `sessions` – разделение органического и платного трафика
4. **region** в таблице `sessions` – разделение на «Мо и СПб» и другие
5. **brand** в таблице `hits` – выделение марки авто
6. **model** в таблице `hits` – выделение модели авто
7. **adv** в таблице `sessions` – разделение рекламы из социальных сетей и других

Были удалены признаки:

1. `'hit_number', 'hit_type', 'hit_page_path', 'event_category', 'event_action', 'event_label', 'event_value', 'hit_referer'` из таблицы **hits**
2. `'client_id', 'visit_time', 'visit_number', 'utm_medium', 'utm_adcontent', 'utm_keyword', 'device_os', 'device_model', 'device_screen_resolution', 'device_browser'` из таблицы **sessions**

Признак `target_action` из таблицы **hits** был добавлен к **sessions**, как новый признак `target`

В результате работы с таблицами получились две таблицы:

1. **hits_brand** с признаками: `'session_id', 'hit_date', 'target_action', 'brand', 'model'`
2. **sessions_df** признаками: `'session_id', 'target', 'visit_date', 'utm_source', 'utm_campaign', 'device_category', 'device_brand', 'geo_country', 'geo_city', 'traffic', 'region', 'month', 'adv'`

Пропуски были обработаны двумя способами:

1. Строки, где пропущено большинство значений, были удалены (менее 1%)
2. Остальные пропуски заполнены значением `other`

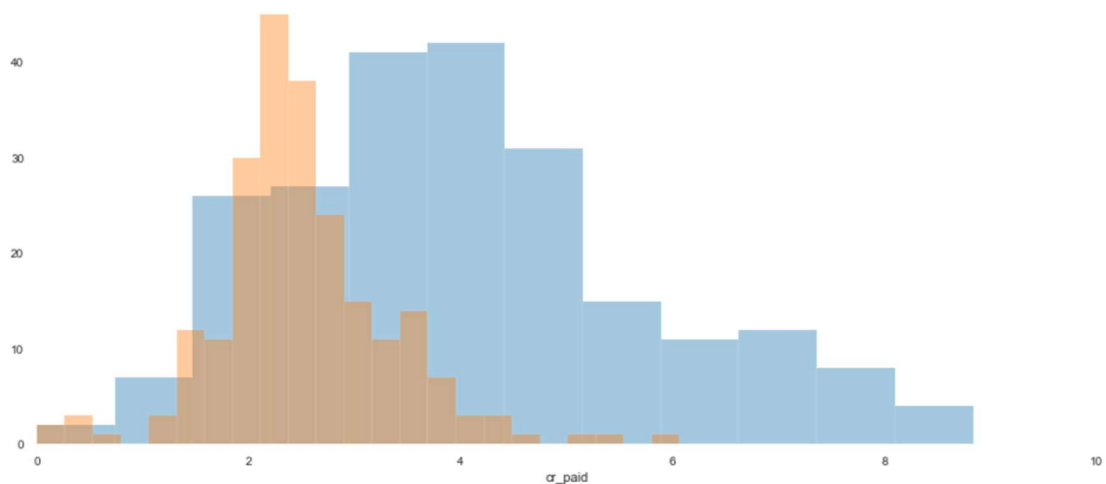
ПРОВЕРКА ГИПОТЕЗ

Органический трафик не отличается от платного с точки зрения CR в целевые события

Сформирован датасет для проверки гипотезы, где строки были сгруппированы по датам и высчитан CR в целевые действия по каждой дате

```
traff_all
```

	visit_date	cr_org	cr_paid
0	2021-05-19	7.173601	0.000000
1	2021-05-21	0.000000	0.000000
2	2021-05-22	1.443001	1.470588
3	2021-05-23	4.581901	2.542373
4	2021-05-24	6.499459	3.030303
...
221	2021-12-27	1.617710	2.214399
222	2021-12-28	1.223865	2.191194
223	2021-12-29	1.843003	1.763571
224	2021-12-30	1.626016	1.358234
225	2021-12-31	0.852273	0.730613



Распределение проверим тестом Шапиро – Уилка

```
stats.shapiro(traff_sh)
```

```
ShapiroResult(statistic=0.9192128777503967, pvalue=7.662228700643352e-15)
```

Выборки имеют ненормальное распределение - используем непараметрические критерии. Выборки независимы, поэтому используем критерий Манна Уитни

H0: Органический трафик не отличается от платного с точки зрения конверсии в целевые события

H1: Конверсия в целевые события от органического трафика выше, чем от платного

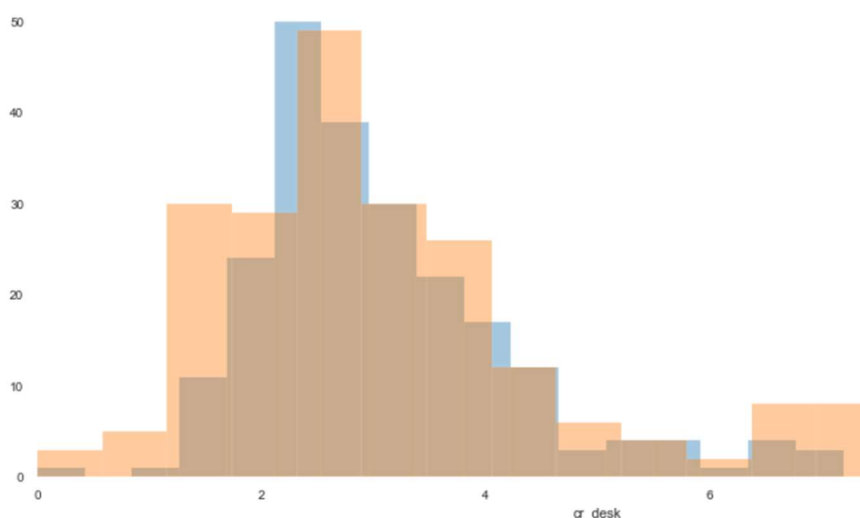
```
stats.mannwhitneyu(traff_all['cr_org'], traff_all['cr_paid'], alternative='greater')
```

```
MannwhitneyResult(statistic=39637.5, pvalue=1.5962465123037001e-24)
```

Нулевая гипотеза отвергнута - принимаем альтернативную: **Конверсия в целевые события от органического трафика выше, чем от платного**

Трафик с мобильных устройств не отличается от трафика с десктопных устройств с точки зрения конверсии в целевые события

Сформирован датасет для проверки гипотезы, где строки были сгруппированы по датам и высчитан CR в целевые действия по каждой дате



dev_all			
	visit_date	cr_mob	cr_desk
0	2021-05-19	6.918239	7.272727
1	2021-05-21	0.000000	0.000000
2	2021-05-22	1.369863	1.600000
3	2021-05-23	4.139715	4.166667
4	2021-05-24	6.441731	6.499578
...
221	2021-12-27	2.032141	2.450331
222	2021-12-28	2.134679	1.295160
223	2021-12-29	1.890130	0.992556
224	2021-12-30	1.590281	0.413983
225	2021-12-31	1.264299	0.032139

Распределение проверим тестом Шапиро – Уилка

```
stats.shapiro(dev_sh)
```

```
ShapiroResult(statistic=0.8704495429992676, pvalue=5.993145325748426e-19)
```

Выборки имеют ненормальное распределение - используем непараметрические критерии. Выборки независимы, поэтому используем критерий Манна Уитни

H0: Трафик с мобильных устройств не отличается от десктопного с точки зрения конверсии в целевые события

H1: Конверсия в целевые события от мобильного трафика отличается от платного

```
stats.mannwhitneyu(dev_all['cr_mob'], dev_all['cr_desk'])
```

```
MannwhitneyuResult(statistic=24634.0, pvalue=0.5152590369206732)
```

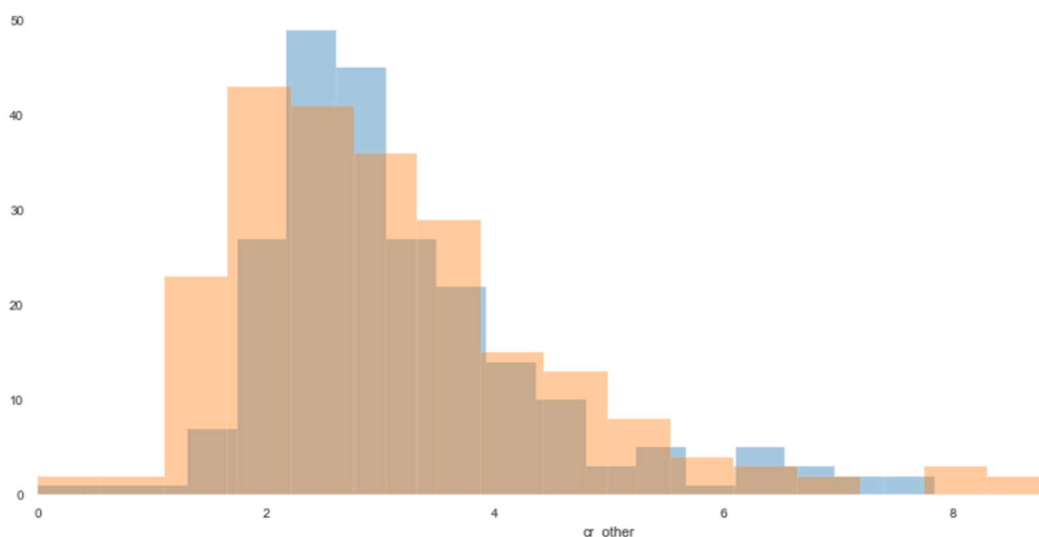
Нулевая гипотеза не может быть отвергнута: **Трафик с мобильных устройств не отличается от десктопного с точки зрения конверсии в целевые события**

Трафик из городов присутствия (Москва и область, Санкт-Петербург) не отличается от трафика из иных регионов с точки зрения CR в целевые действия

Сформирован датасет для проверки гипотезы, где строки были сгруппированы по датам и высчитан CR в целевые действия по каждой дате

```
reg_all
```

	visit_date	cr_mo_spb	cr_other
0	2021-05-19	11.764706	0.000000
1	2021-05-21	0.000000	0.000000
2	2021-05-22	1.717557	0.843882
3	2021-05-23	5.314685	2.030457
4	2021-05-24	6.456763	6.481256
...
221	2021-12-27	1.995835	2.188022
222	2021-12-28	1.979101	2.089060
223	2021-12-29	1.757619	1.793468
224	2021-12-30	1.165283	1.717715
225	2021-12-31	0.566305	1.195017



Распределение проверим тестом Шапиро – Уилка

```
stats.shapiro(reg_sh)
```

```
ShapiroResult(statistic=0.88506680727005, pvalue=7.421764722388121e-18)
```

Выборки имеют ненормальное распределение - используем непараметрические критерии. Выборки независимы, поэтому используем критерий Манна Уитни

H0: Трафик из городов присутствия не отличается от других регионов с точки зрения конверсии в целевые события

H1: Конверсия в целевые события от городов присутствия отличается от других регионов

```
stats.mannwhitneyu(reg_all['cr_mo_spb'], reg_all['cr_other'], alternative='greater')
```

```
MannwhitneyuResult(statistic=27303.5, pvalue=0.10184748890799017)
```

Нулевая гипотеза не может быть отвергнута - трафик из городов присутствия не отличается от других регионов с точки зрения конверсии в целевые события

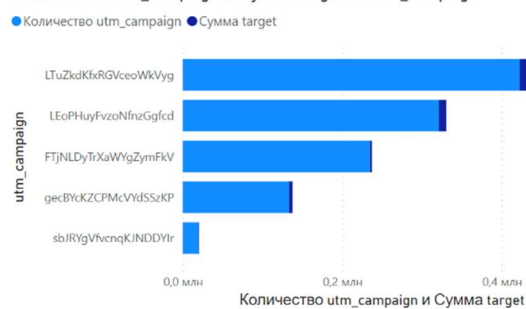
ОТВЕТЫ НА ВОПРОСЫ ПРОДУКТОВОЙ КОМАНДЫ:

1. Из каких источников / кампаний / устройств / локаций к нам идет самый целевой трафик?

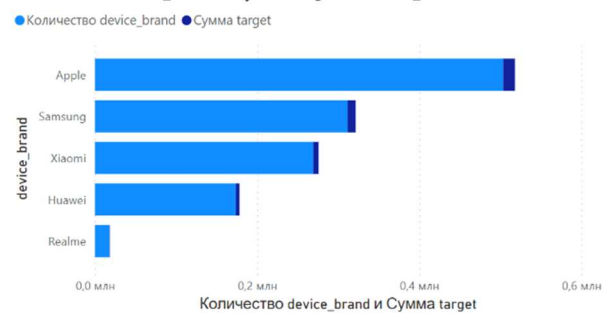
	С точки зрения CR	С точки зрения объема
<u>Источники</u>	<p>YpBKcihLLfJWuxOLfvW</p> <p>fJCYsujgSxIHfBOmgDdN</p> <p>XzfeEBYZWgSDtJNXOadn</p> <p>CqelpFwJscTsZoYXdHsP</p> <p>yxJKymISGVuKIPTxbysx</p>	<p>ZpYIoDJMcFzVoPFsHGJL</p> <p>fDLIAcSmythWSCVMvqvL</p> <p>kjsLglQLzykiRbcDiGcD</p> <p>bByPQxmDaMXgpHeypKSM</p> <p>BHcvLfOaCWwWTyKqHVe</p>
<u>Кампании</u>	<p>MHdHrBKQwbDaRalwnIJq</p> <p>JkhCpeDGctTwhwqWLywv</p> <p>IRKNegNgOUQLwudzMEIF</p> <p>SbYAsCvXapXBOIxEKBZs</p> <p>IndNlerCYECRQvBTyTye</p>	<p>LTuZkdKfxRGVceoWkVyg</p> <p>LEoPHuyFvzoNfnzGgfcD</p> <p>gecBYcKZCPMcVYdSSzKP</p> <p>FTjNLDyTrXaWYgZymFkV</p> <p>sbJRYgVfvcnqKJNDdYIr</p>
<u>Устройства</u>	<p>Motive</p> <p>Condor</p> <p>Land Rover</p> <p>Vertu</p> <p>Razer</p>	<p>Apple</p> <p>Samsung</p> <p>Xiaomi</p> <p>Huawei</p> <p>Realme</p>
<u>Локации</u>	<p>Brescia</p> <p>Qingdao</p> <p>Middletown</p> <p>Nybro</p> <p>Gravesend</p>	<p>Moscow</p> <p>Saint Petersburg</p> <p>Kazan</p> <p>Krasnodar</p> <p>Yekaterinburg</p>

Самые популярные источники / кампании / устройств / локации

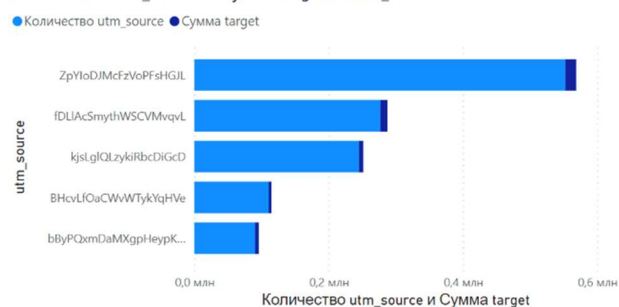
Количество utm_campaign и Сумма target по utm_campaign



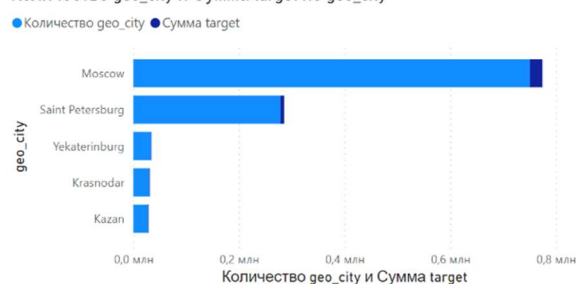
Количество device_brand и Сумма target по device_brand



Количество utm_source и Сумма target по utm_source

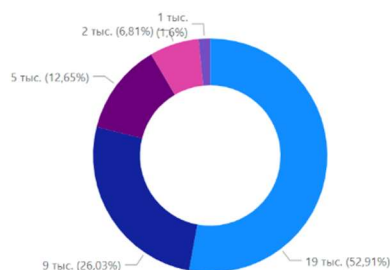


Количество geo_city и Сумма target по geo_city

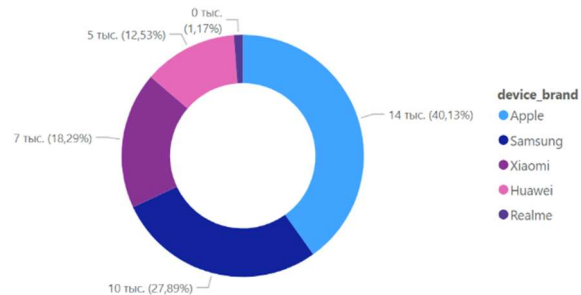


Распределение целевых действий между самыми популярными источниками, кампаниями, устройствами, локациями

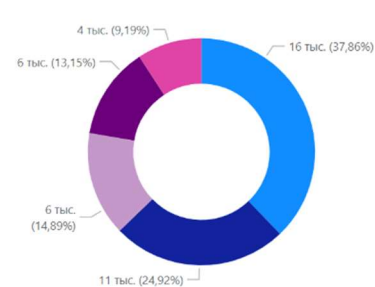
Сумма target по utm_campaign



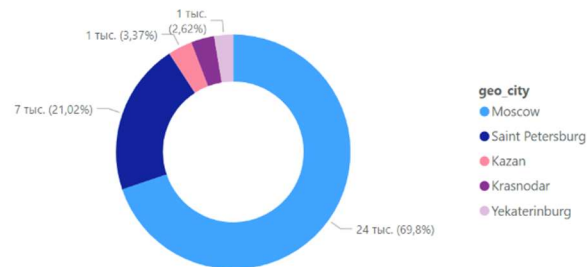
Сумма target по device_brand



Сумма target по utm_source



Сумма target по geo_city



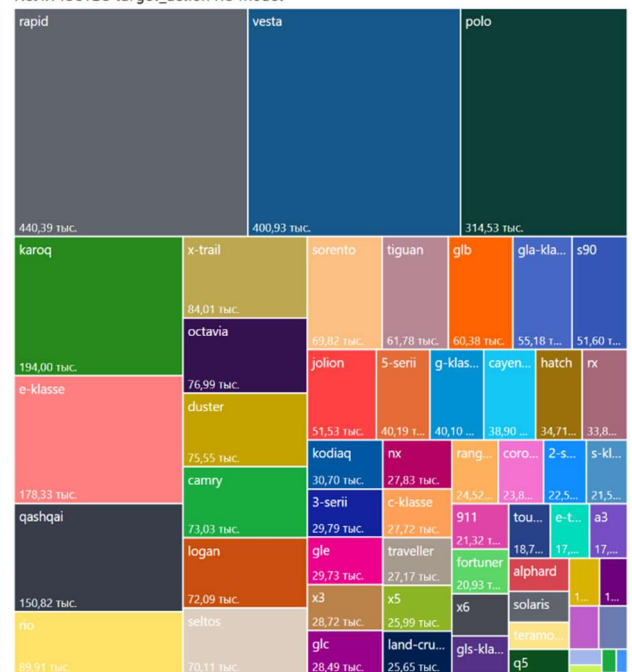
2. Какие авто пользуются наибольшим спросом?

С точки зрения CR	С точки зрения спроса
infiniti	skoda
hyundai	mercedes-benz
honda	volkswagen
lada-vaz	lada-vaz
volkswagen	nissan

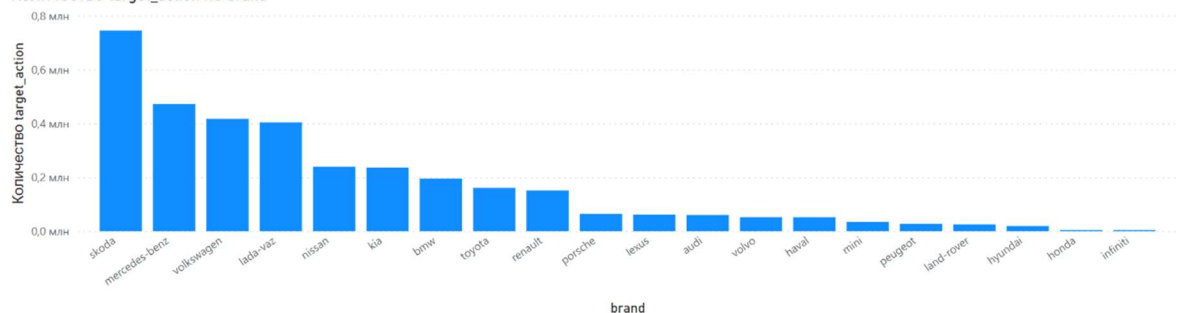
Количество target_action по brand



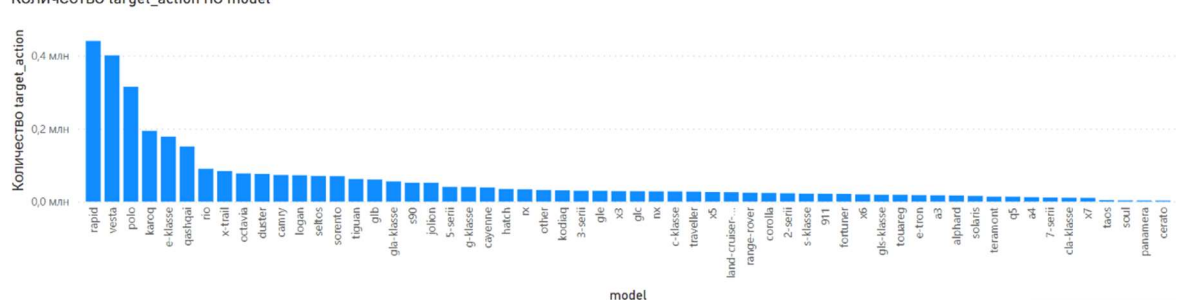
Количество target_action по model

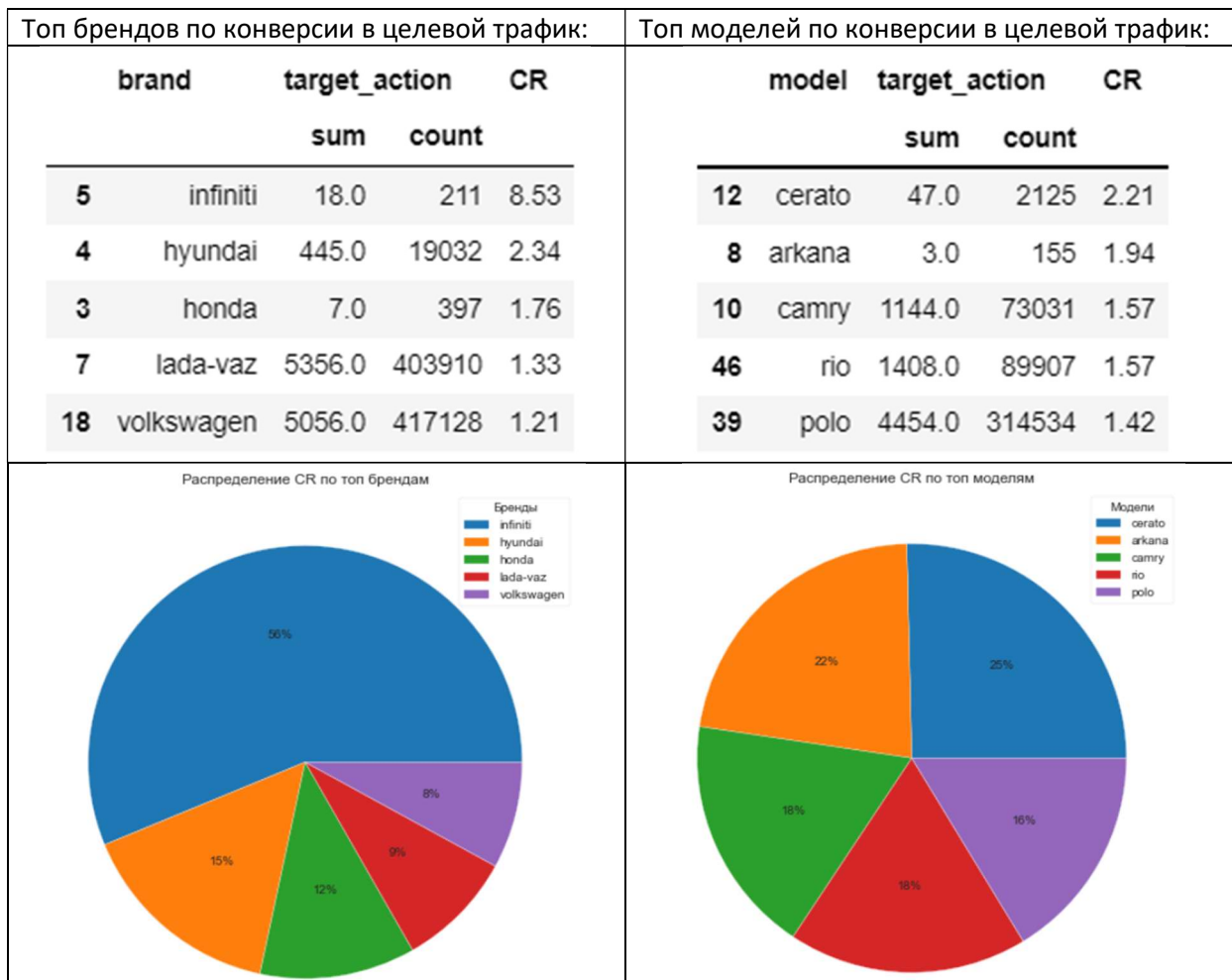


Количество target_action по brand



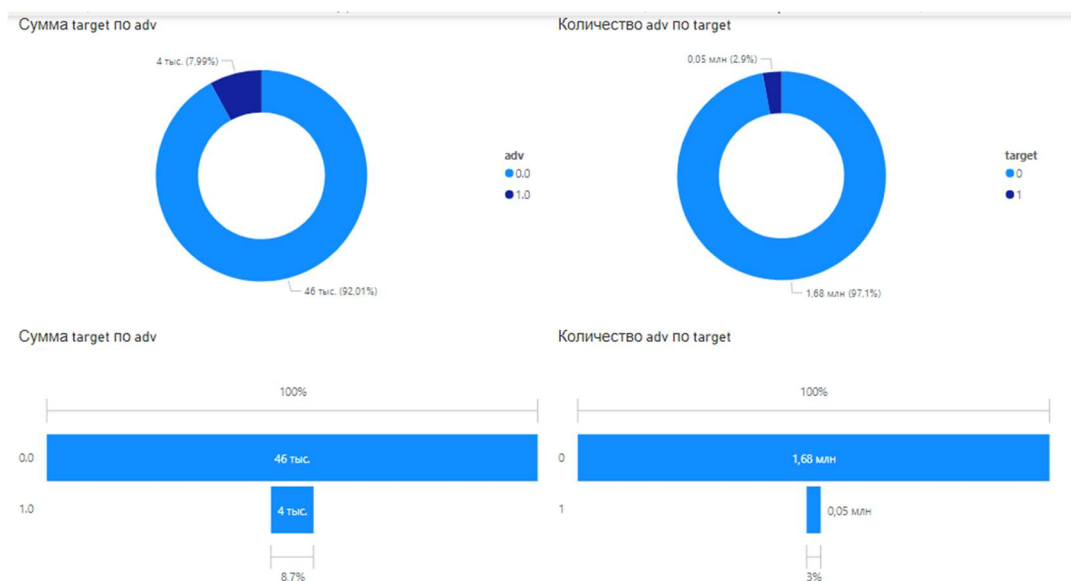
Количество target_action по model





3. Стоит ли увеличивать свое присутствие в соцсетях и давать там больше рекламы?

Источник трафика	CR
Из соц сетей	3,14
Из других источников	1,57

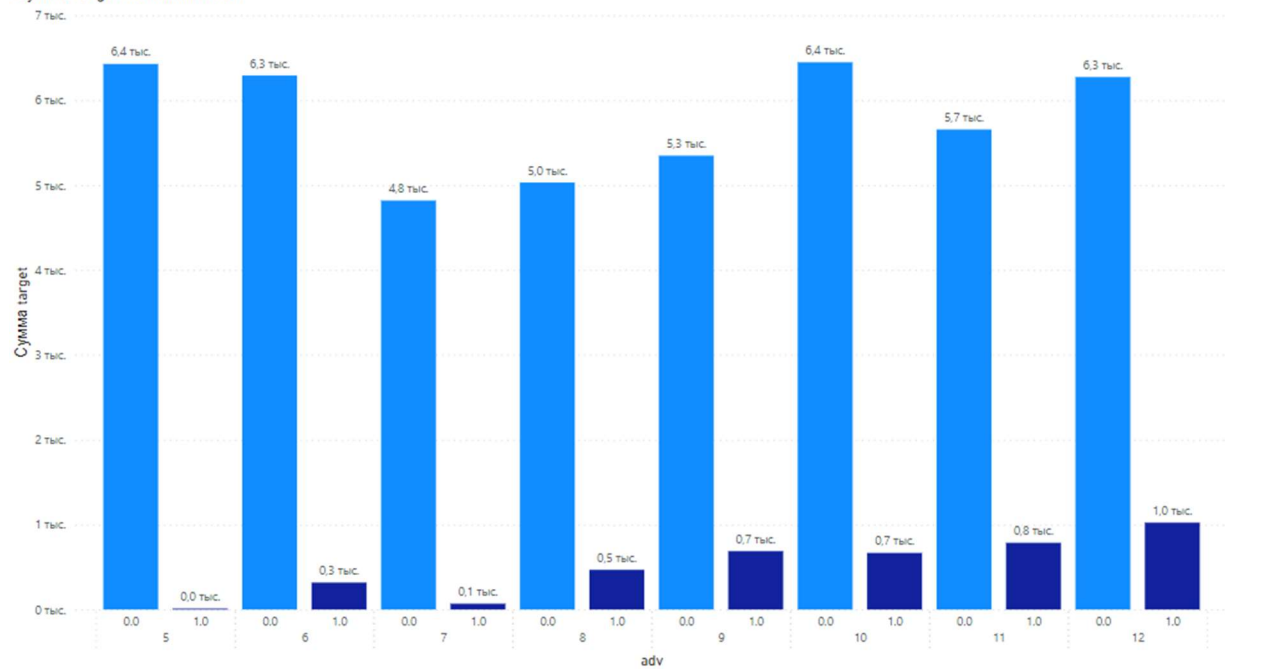


С первого взгляда видно, что целевых действий идет больше от трафика не из соц сетей.

Однако проверим, меняется ли что-то со временем

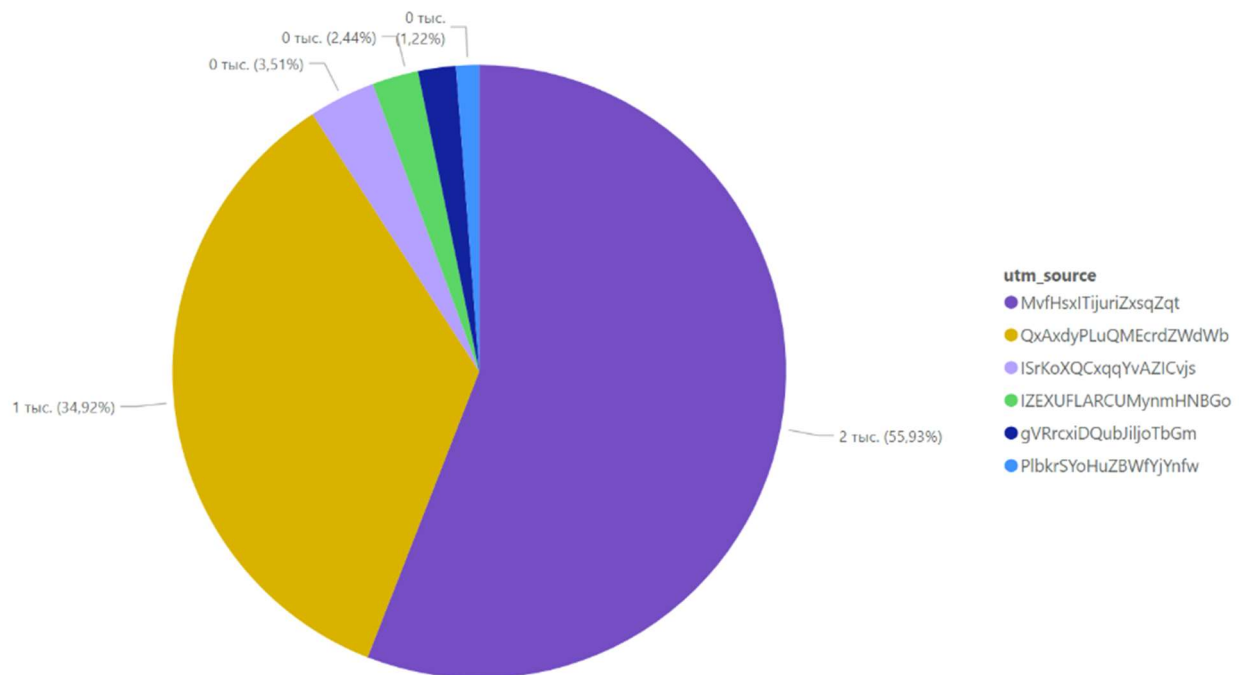
Распределение количества источников из соц сетей и других по целевым действиям по месяцам

Сумма target по month и adv



Количество целевых действий от соц сетей заметно увеличивается. Тенденция явно на рост влияния соц сетей. Можно предположить, что имеет смысл увеличить присутствие в соц сетях.

Распределение целевого трафика среди источников соц сетей



Наиболее качественно отрабатывает реклама в MvfHsxlTijuriZxsqZqt и QxAxdyPLuQMEcrdZWdWb. Можно выделять на них больше средств