# ПРЕДСТАВЛЕНИЕ РЕЗУЛЬТАТОВ РЕКЛАМНЫХ АКЦИЙ ПО БОЛЬШИМ ДАННЫМ

**Владова А.Ю., Владов Ю.Р.**

*Институт проблем управления им. В.А. Трапезникова РАН,*
*Россия, г. Москва, ул. Профсоюзная, д.65*
*Финансовый университет при Правительстве Российской Федерации,*
*Россия, г. Москва, Ленинградский пр. 49*
avladova@ipu.ru,

**Якимов А.И.**

*Белорусско-Российский университет, Беларусь, г. Могилев, проспект Мира, 43*

*Аннотация: С 2012 года растет количество публикаций и программных средств в области анализа больших данных бизнес-процессов. Вклад данного исследования заключается в подтверждении гипотезы о возможности построения количественной модели бизнес-процесса в отсутствии обратной связи.*

Ключевые слова: EDA, большие данные, рекламная акция, статистический анализ.

## Введение

Для продвижения товаров и услуг дистрибьюторы управляют широкой сетью торговых представителей (ТП). Они представляют товары или услуги и демонстрируют презентации потенциальным клиентам до 10 раз в день. Презентации демонстрируются на планшетах, которые отправляют на центральный сервер дистрибьютора GPS-координаты, время начала и окончания презентации и название товара или услуги. Но современная технология геолокации не может различить, прокручивает ли ТП свою презентацию, находясь внутри или рядом с офисом клиента. Это означает, что ТП может имитировать свои действия. Существует возможность мошенничества и на этапе отправки на центральный сервер остальных параметров посещения клиента. Поскольку трудно отследить параметры посещения, распределенная база данных содержит некорректную информацию, и дистрибьюторская компания не может должным образом рассчитать KPI своих ТП. Уточнение бизнес-процессов дистрибьюторов и анализ больших данных, содержащихся в распределенной базе, позволили сформулировать предположение об возможности выявления значительной части ложной информации в случае надлежащей разметки данных с помощью алгоритмов машинного обучения.

## 1 Анализ публикаций

Согласно платформе dimensions.ai, которая обеспечивает доступ к грантам, публикациям, патентам и другим источникам, количество публикаций в области расширенного анализа данных (EDA) бизнес-процессов растет с 2012 года (рисунок 1). Экспоненциальное насыщение в 2020 году имеет место из-за одного-двухлетнего периода, который проходит от подачи статьи\патента к публикации\регистрации.

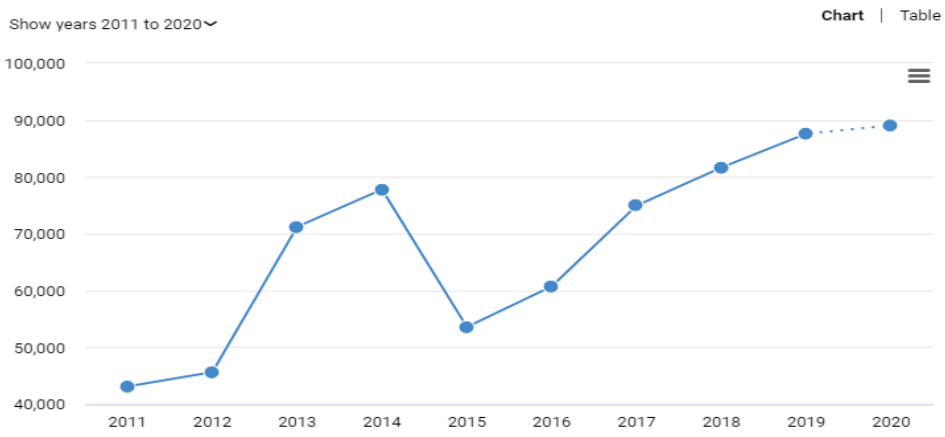



*Рис. 1. Рост количества публикаций по EDA бизнес-процессов*

EDA имеет важное значение для машинного обучения. Эксперты считают, что хорошо выполненный EDA помогает найти шаблоны в данных, которые в идеале будут играть роль гипотез

[2]. Авторы [3] изучали роль интеллектуальных платформ в исследовании и визуализации научных данных. Среди прочих они отметили Tableau, SAS, Splunk, Stata, Alteryx, Periscope; а в категории электронных таблиц и баз данных SQL, DbVisualizer, Alation, and Microsoft Excel. Описывая категорию языков программирования, они удивительным образом не упомянули, ни Python, ни R. Авторы поставили их в категорию, названную разметка данных. Но сегодня эти языки представляют самые мощные библиотеки для поддержки EDA (например, pandas profiling и sklearn) и визуализации (например, seaborn и mathplotlib).

Вместе с предыдущими категориями авторы упомянули категорию под названием «доморощенная автоматизация». Она включает в себя скрипты, которые ученые создали для себя при выполнении повторяющихся задач. Платформа визуализации [4] является одним из примеров из этой категории. Пользователь платформы может наблюдать за эволюцией и производительностью внутренних структур модели. Согласно [5] в рамках рекламных акций используются следующие инструменты: RAW, Datawrapper, Timeline JS, and D3.js, визуализирующие карты, исторические данные и анимации.

## 2 Данные и методы обработки

Мы рассмотрели этапы анализа данных, предложенные в [6-8] и обобщили их на рисунке 2.

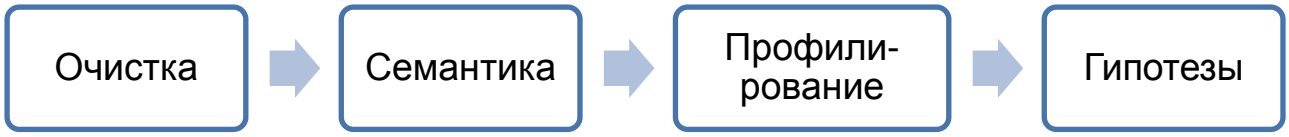



*Рис. 2. Процесс исследования.*

Образец данных для анализа, полученный из распределенной базы данных нашего клиента, содержит полмиллиона записей в 29 различных признаках. Для анализа пропусков информации применена библиотека Missingno. Выявленные пропуски по каждому признаку отражены рисунком 3.

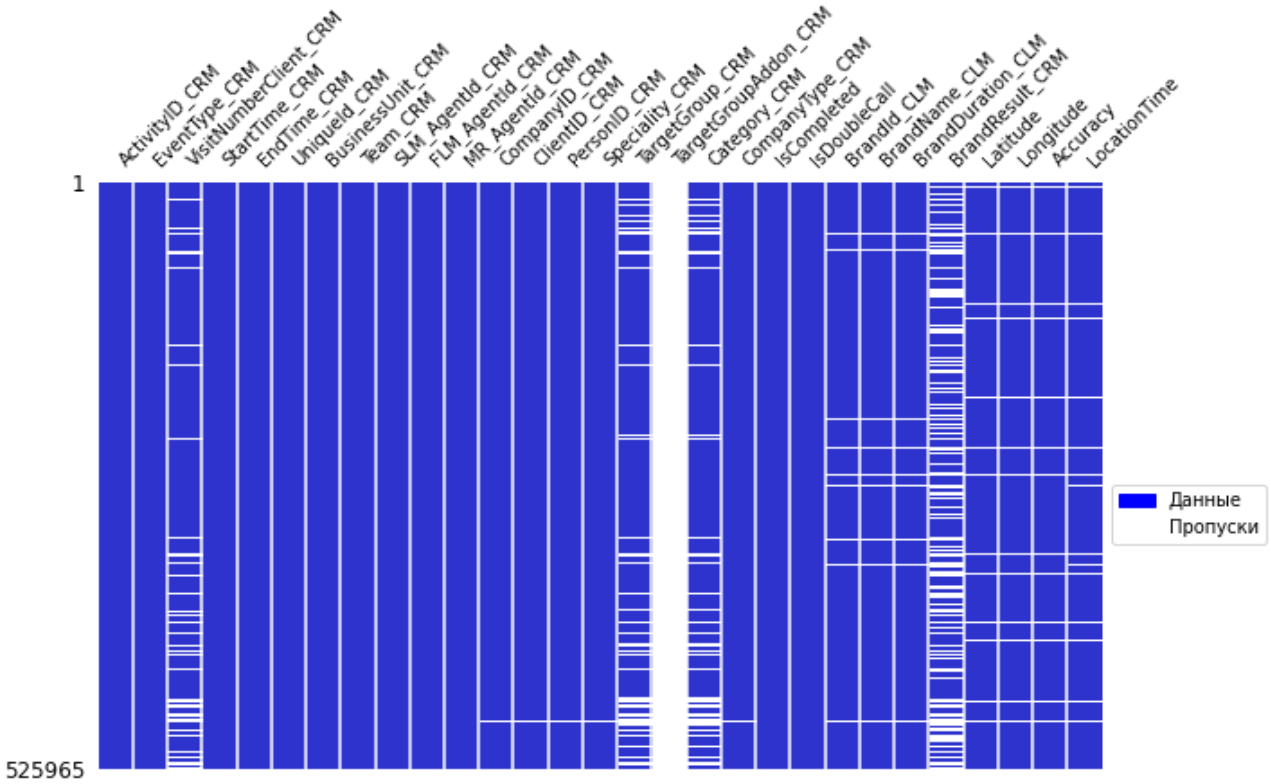



*Рис. 3. Матрица пропусков данных*

Очистка данных - важный шаг в процессе постановки статистических гипотез и особенно важный для уменьшения воздействия ошибок, допущенных при сборе информации [9]. Для улучшения качества входных данных используются три группы алгоритмов обнаружения объектов-выбросов:

статистические, метрические и эвристические. После обнаружения в выборке объектов-выбросов наиболее распространенной является стратегия фильтрации таких объектов [7].

Пытаясь понять семантику данных и то, что они представляют, а также логику генерации данных, мы изучили названия распределенных полей баз данных и как они соотносятся с данными в записях. Рисунок 4 отражает взвешенный список ключевых слов распределенной базы данных в виде облака тегов, которое мы получили с помощью библиотеки WordCloud.

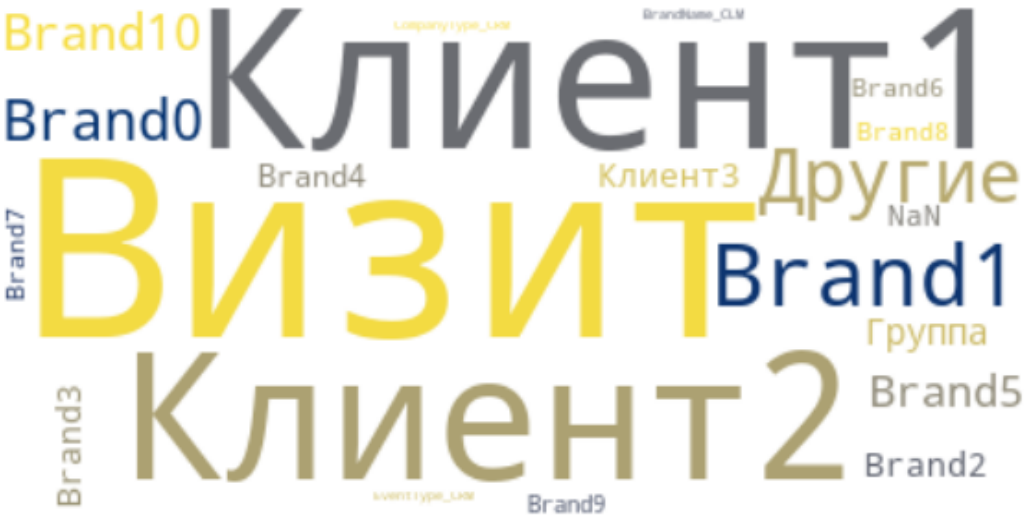



*Рис. 4. Облако тегов содержимого распределенной базы данных.*

Для выяснения закономерности между группами признаков выполнено их объединение по значению коэффициента линейной корреляции. Для этого построена дендрограмма связей признаков с помощью алгоритма иерархической кластеризации, заложенного в библиотеке Missingno (рис. 5).

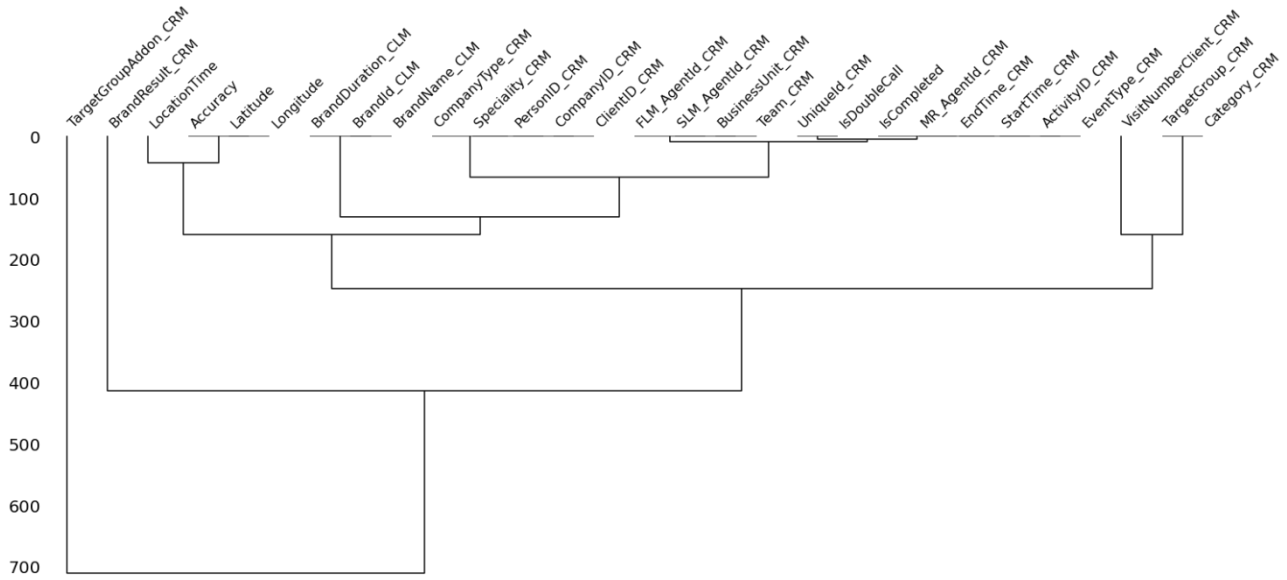
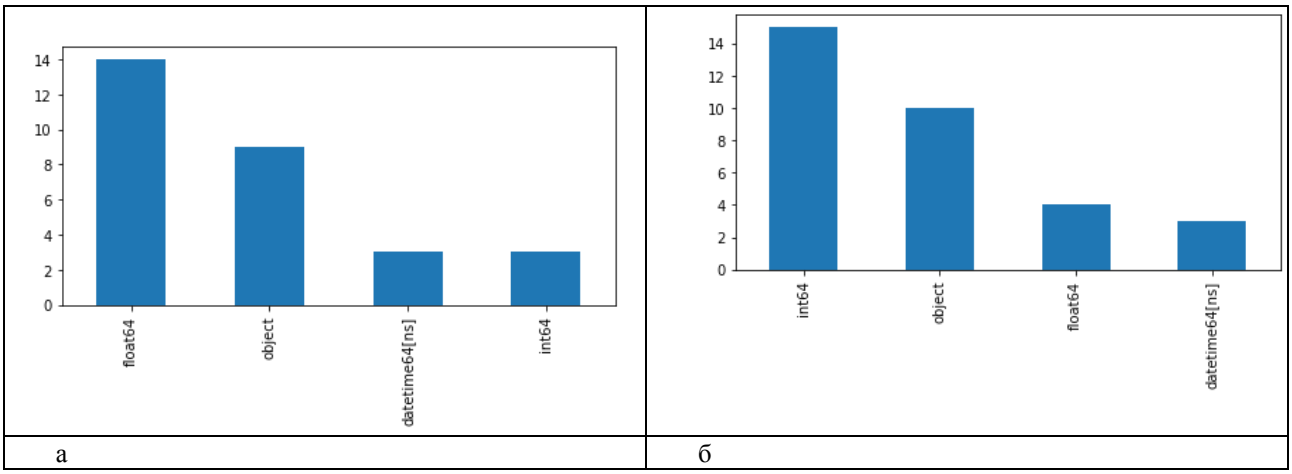



*Рис. 5. Дендрограмма с выделенными группами признаков*

Диаграмма иллюстрирует наличие семи групп близких по смыслу признаков, которые дальше от нуля формируют иерархические комбинации.

Цель EDA на этапе профилирования состоит в том, чтобы описать основные особенности имеющейся численной и категориальной информации. Анализ показал значимое количество категориальных признаков, таких как типы посещений и клиентов, названия бизнес-подразделений, мест посещений, групп клиентов и ТП, флаги исполнения и контроля визита третьей стороной и т. д. В то же время остальные признаки являются уникальными идентификаторами посещений, торговых

представителей, клиентов, брендов и т. д. Соотношение количества типов признаков до и после явного приведения типов и обработки временных рядов приведено на гистограммах рисунка 6.




а б

*Рис. 6. Соотношение количества типов признаков: а – до обработки; б – после обработки*

На рисунке 7 отображены доли рекламных акций, приходящихся на типизированного клиента и рекламируемый бренд.

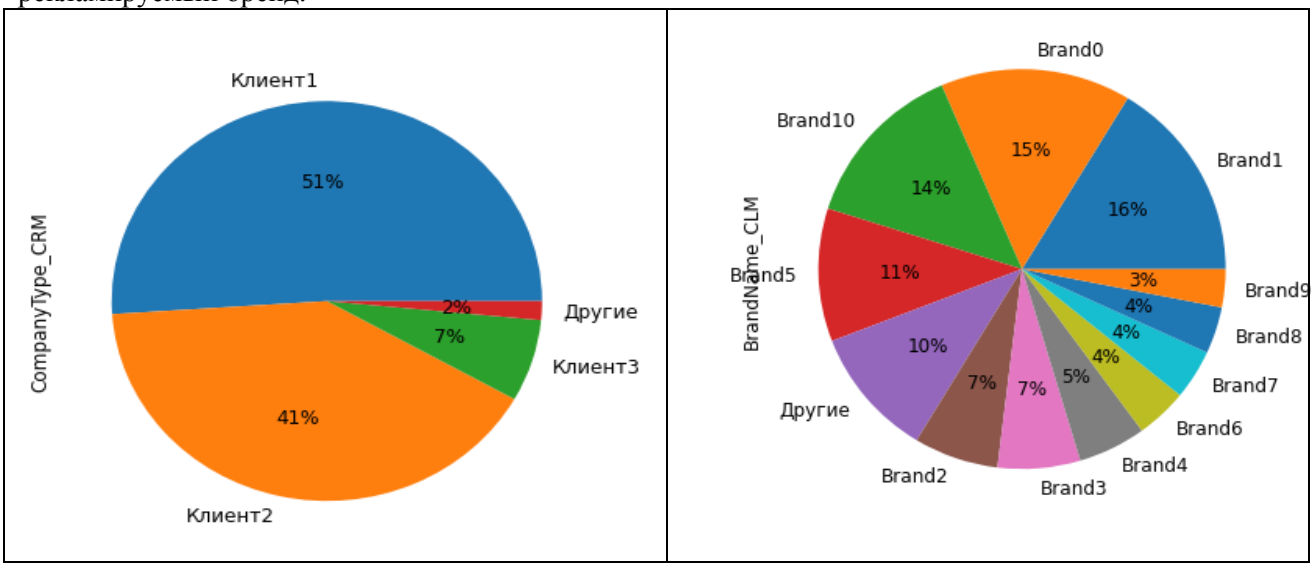



*Рис. 7. Доли рекламных акций, приходящихся на: а – клиентов; б – рекламируемые бренды.*

Полезно создать мозаичные диаграммы в качестве графического метода визуализации данных двух категориальных признаков. Для этого нам прежде нужно вычислить частотную таблицу признаков. Мозаичное представление двух категориальных признаков представлено на рисунке 8.

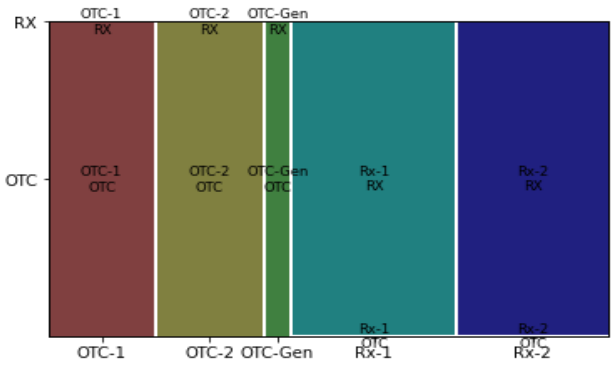



*Рис. 8. Диаграмма Маримекко*

Последней категорией исследовательской деятельности является генерация гипотез. Для этого построены графики распределений и выбросов (рис. 9).

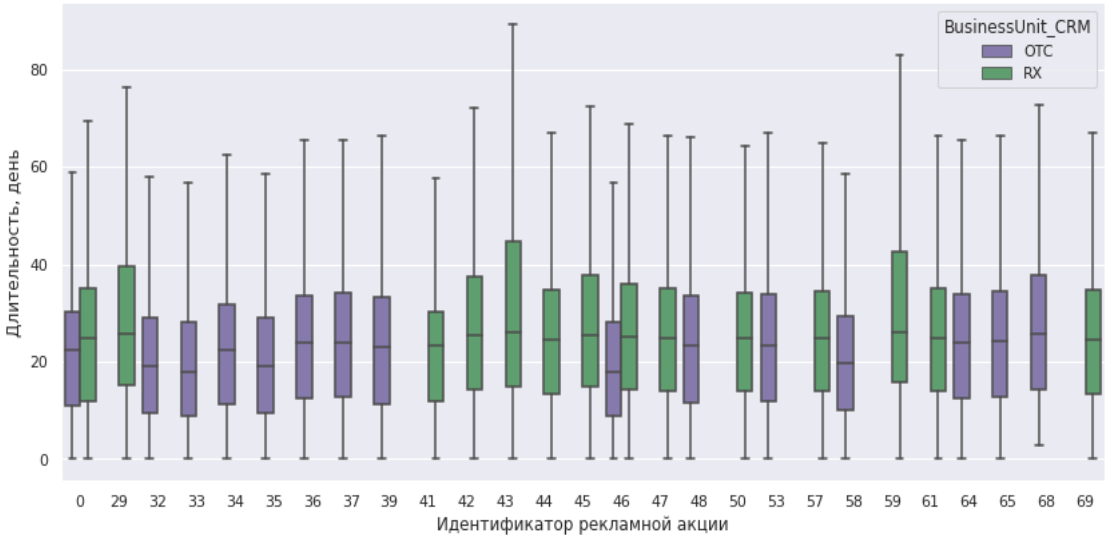



*Рис. 9. Диаграмма «ящик с усами»*

Наиболее длительными рекламными акциями были номера 42, 43, 45, 48, 59, 68. По ним наблюдаются и наиболее значительные выбросы.

Близкое к нормальному распределение признака Идентификатор активности ТП позволило сформулировать статистическую гипотезу об уровне искажения информации в распределенной базе данных.

## 3 Результаты

Для визуализации как численных, так и категориальных признаков мы использовали диаграмму рассеяния. Рисунок 10 показывает, что все бизнес-подразделения дистрибьютора заполняют выходные дни посещениями клиентов наравне с рабочими днями.

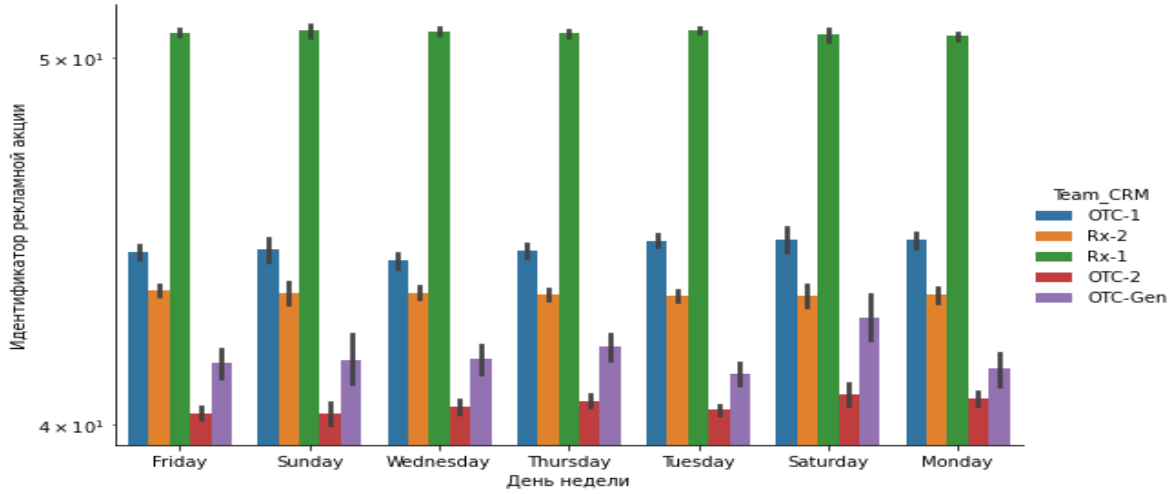



*Рис. 10. Интенсивность клиентских посещений различных бизнес-подразделений*

Таким образом, мы выполнили этап профилирования, имея в виду, что главной особенностью наших данных является отсутствие автоматически измеренной обратной связи.

## 4 Дискуссия

EDA анализ сформулировал несколько уточняющих вопросов для разметки данных: какова продолжительность рекламных акций; действительно ли ТП работают по выходным; сколько времени требуется, чтобы продемонстрировать презентации в быстром режиме; действительно ли бизнес-подразделение OTC-Gen работает с топ менеджментом; можно ли демонстрировать видеопрезентации по выходным и т.д.

Рисунок 11 отражает связь между продолжительностью рекламной акции и месяцем.

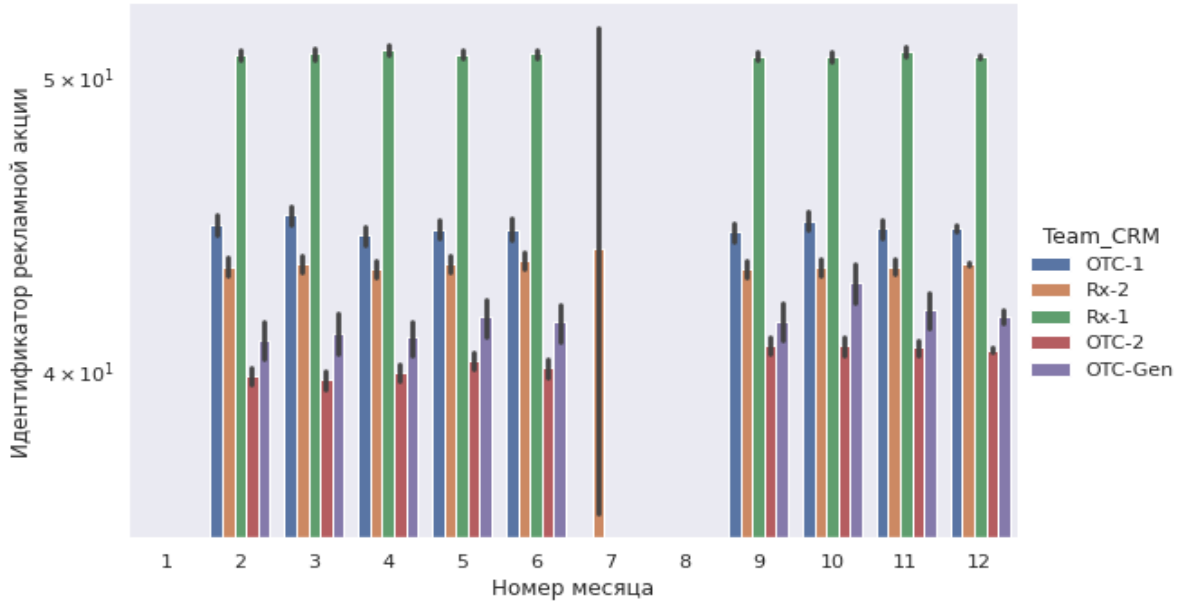



*Рис. 11. Соотношение продолжительности рекламных акций, в которых участвуют бизнес-подразделения*

За рассматриваемый год в летние месяцы и в январе рекламные акции не проводились. Наибольшая интенсивность работы у подразделений Rx-1 и OTC-1. Наименьшая интенсивность у подразделения OTC-2. В июле работало единственное подразделение Rx-2.

## Заключение

Согласно первичному статистическому исследованию, в среднем ТП выполняет около 1000 различных мероприятий в течение месяца, но среди наблюдений есть как значения 500 так и более чем 2500 мероприятий. Это может указывать на ТП, которые значительно завышают или уменьшают свое количество мероприятий. В то же время мы не нашли несоответствий во времени начала и окончания презентаций. Сравнительно короткие по продолжительности презентации (10-30 минут) были наиболее часто используемыми при самых длительных рекламных акциях. Самым успешным бизнес-подразделением является OTC-Gen (рис. 11).

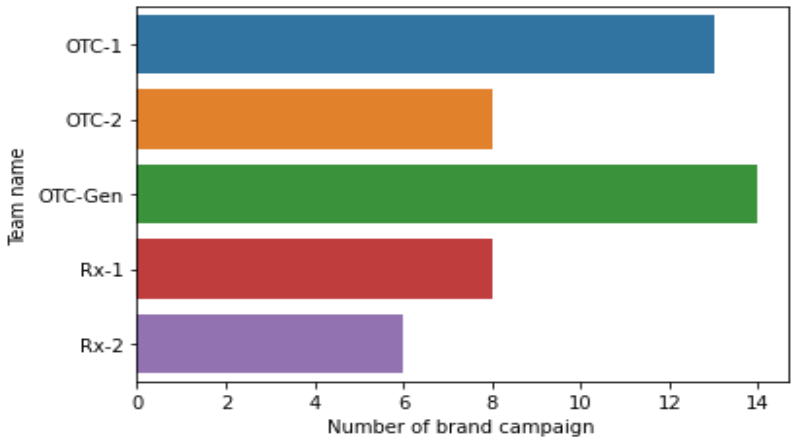



*Рис. 11. Соотношение количества рекламных акций, связанных с бизнес-подразделением*

Оно участвует в большинстве рекламных акций и тратит меньше времени на показ презентаций. В то же время бизнес-подразделение RX-2 имеет такое же количество участников при значительном количестве выбросов в продолжительности акций.

На этапе профилирования выполнен описательный анализ и сгенерированы гистограммы распределения данных, тенденции изменения, карты корреляционных значений.

Подводя итоги, наш трехэтапный анализ помог выделить большинство потенциальных выбросов.

**Литература**


1. https://app.dimensions.ai.
1. Skiena S.S. The data science design manual. - Springer, 2017. – 445 p.
2. *Alspaugh S., Zokaei N., Liu A., Jin C.,* Hearst M.A. Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices // IEEE Transactions on Visualization and Computer Graphics, 25(1).2019.– P.
3. *Li H., Fang S., Mukhopadhyay S., Saykin A.J., Shen L. I*nteractive Machine Learning by Visualization: A Small Data Solution // Procedure IEEE Int Conf Big Data 2018, P. 3513–3521.
4. *Krasser A.* How brands are using data visualization in social campaigns. Search engine watch. https://www.searchenginewatch.com/2016/06/17/how-brands-are-using-data-visualisation-in-social-campaigns/.
5. *Huxley K.* Data Cleaning // SAGE Research Methods Foundations. doi: 10.4135/9781526421036842861
6. *Kandel S., Paepcke A., Hellerstein J. M., Heer. J.* Enterprise data analysis and visualization: An interview study // IEEE Transactions on Visualization and Computer Graphics, 18(12), 2012. P. 2917–2926.
7. *Владова А. Ю.* Диджитализация маркетинговых кампаний / А. Ю. Владова, Ю. Р. Владов // Цифровая трансформация промышленности: тенденции, управление, стратегии - 2020 : Материалы ii международной научно-практической конференции, Екатеринбург, 27 ноября 2020 года. – Екатеринбург: Институт экономики УрО РАН, 2020. – С. 66-73.
8. *Vladova A.Yu., Vladov Y.R.* Machine Classification of Pore Space for Hydrocarbon Reservoir Characterization // Procedure IEEE 21st Conf on Business Informatics (CBI). 2019, P. 391-396.
9. *Борисова И. А.* Очистка данных от диагностических ошибок в признаковых пространствах большой размерности / И. А. Борисова, О. А. Кутненко // Математическая биология и биоинформатика. – 2019. – Т. 14. – № 2. – С. 464-476. – DOI 10.17537/2019.14.464.