



Направление Data Analyst, MTC

Аннотация

Добро пожаловать на виртуальную стажировку **Shift + Enter** от MTC — компании, где работа построена на сплоченных продуктовых командах, использующих agile¹-методологию. Big Data MTC — место, где телеком-данные превращаются в востребованные IT-продукты.

Стажировка от MTC в роли начинающего Data-аналитика позволит тебе:

- Приобрести продвинутые знания SQL и основных библиотек Python для работы с данными.
- Поработать над разными проектами в рамках твоего стажерского спринта.

Развиваемые компетенции

По результатам выполнения заданий ты прокачаешь такие навыки, как:

- Подготовка и визуализация данных.
- Анализ больших данных, хранящихся в разных таблицах.

А также умение самостоятельно искать информацию в интернете и структурировать ее, логику, аккуратность и внимательность к деталям.

Описание подзадач

Команда Big Data MTC создала и протестировала несколько десятков аналитических сервисов. Самые успешные из них уже стали элементами экосистемы MTC. А ты готов стать частью нашего коллектива и помочь с решением задач, связанных с анализом и визуализацией данных?

Выполнение всего блока заданий займет у тебя не более 80-100 минут.

Рекомендуемый тайминг

1. 20-30 минут на первое задание.
2. 30 минут на второе задание.
3. 30-40 минут на третье задание.

Информация о загрузке решения

Стажировка содержит несколько подзадач. Можно загрузить файл, содержащий решение части заданий, но по возможности постарайся сделать их все.

Желаем удачи!

¹ Agile — это группа методик для гибкого управления проектами в команде разработки. Рабочий процесс при таком подходе разбивается на небольшие временные промежутки, их еще называют спринтами или итерациями. Во время каждого спринта команда разработки создает часть продукта, которую можно протестировать и оценить.

Задание 1. Напиши SQL-запрос, помоги коллегам

Рады видеть тебя на стажировке от МТС! Илья², руководитель группы рекламной аналитики центра MTS Big Data и твой наставник, попросил помочь команде сервиса «[МТС Маркетолог](#)».

Привет!

Наверняка ты слышал о нашем умном сервисе «МТС Маркетолог», который позволяет нашим клиентам самостоятельно продвигать свои товары и услуги посредством таргетинга и ресурсов Big Data МТС с машинным обучением. Нашу платформу обычно используют фрилансеры, индивидуальные предприниматели, представители малого и среднего бизнеса.

Предположим, что нашу платформу будет использовать человек, только что открывший парикмахерскую по адресу г. Тверь, ул. Трехсвятская, д. 3, который хочет привлечь как можно больше новых клиентов.

Твоя задача – использовать данные, которые у нас имеются, чтобы:

- Описать алгоритм, который позволит ему собрать сегмент абонентов.
- Предложить наиболее подходящий способ привлечения новых клиентов.

Hints. Пока ты работаешь с фрагментами (сэмплами) данных. Но я рекомендую описать выгрузку сегмента в виде SQL-запроса, так как это позволит тебе применить уже готовый запрос позже для всего массива данных.

Спасибо!

Полезные материалы

- [Сэмплы данных](#) в формате .csv.
- [Руководство](#) по анализу данных с SQL.

Формат конечного результата

Файл в формате .docx, содержащий описание алгоритма, предложения по привлечению и SQL-запрос для выгрузки сегмента абонентов.

Форма загрузки результата

Пожалуйста, загрузи свой вариант ответа в формате zip-архива, используя инструмент «Загрузить решение». Необходимо сформировать единый zip-архив, содержащий решение одного или всех заданий по выбранной специальности.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания от эксперта после отправки собственной версии.

² Все имена и названия вымышленные, любые совпадения случайны. Данные заданий могут быть изменены в целях конфиденциальности

Вложение 1. Информация о таблицах, содержащих данные

- **cnum_d**: таблица со звонками всех абонентов МТС по всей России cnum_d с полями: msisdn (номер абонента), ts (дата и время звонка), called_party_number (номер, с которого или на который звонил абонент, в случае sms в этом поле может быть название компании-отправителя или короткий номер), call_type (тип вызова sms – 2 или звонок – 1), direction (направление вызова – входящий (I) или исходящий (O))

cnum_d

| msisdn | ts | called_party_number | call_type | direction |
|--------|-----|---------------------|-----------|-----------|
| ... | ... | ... | ... | ... |

- **host_d**: таблица с данными посещения хостов всех абонентов МТС по всей России host_id с полями: msisdn (номер абонента), ts (дата и время посещения), host (URL сайта типа yandex.ru)

host_d

| msisdn | ts | host |
|--------|-----|------|
| ... | ... | ... |

- **geo_d**: таблица с данными, обновляемыми раз в 10 минут о местонахождении абонента geo_d с полями: msisdn (номер абонента), ts (дата и время определения местонахождения), geo_area_key (идентификатор зоны местонахождения абонента)

geo_d

| msisdn | ts | geo_area_key |
|--------|-----|--------------|
| ... | ... | ... |

- **twogis_phones**: справочник 2GIS, в котором есть рубрика, к которой относится номер телефона организаций и название самой организации, т. е. таблица twogis с полями: rubric, phone_num, organization_name

twogis_phones

| rubric | phone_num | organization_name |
|--------|-----------|-------------------|
| ... | ... | ... |

- **twogis_hosts**: справочник 2GIS, в котором есть рубрика, к которой относится сайт организации и название самой организации, т. е. таблица twogis с полями: rubric, host_name, organization_name

twogis_hosts

| rubric | host_name | organization_name |
|--------|-----------|-------------------|
| ... | ... | ... |

- **cities_regions**: справочник с идентификаторами зон, которые соответствуют региону и городу, району города, в которых они находятся, т. е. таблица cities_regions с полями: geo_area_key, city_name, region, city_area

cities_regions

| geo_area_key | city_name | region | city_area |
|--------------|-----------|--------|-----------|
| ... | ... | ... | ... |

- набор оборудования для рассылки SMS любому абоненту МТС с любым текстом, а также call-центр для обзвона абонентов

Задание 2. Напиши «красочный» SQL-запрос

Ты отлично справился с первым заданием и готов приступить ко второму. Оно будет связано с дополнением SQL-запроса, который поможет клиенту оптимизировать подачу пар разных цветов без их повторения. Чуть позже ты увидел письмо от Ильи с пояснениями к задаче.

Привет!

Спасибо за помощь с сегментацией абонентов, отличная работа!

Сегодня у нас был видеозвонок с фабрикой, выпускающей ткани различных цветов. Они обратились к нам за помощью в интеллектуальной настройке оборудования. В каждый автомат, который красит ткани, подается только два цвета, а на фабрике хотели бы, чтобы ткани выпускались с максимально разнообразной палитрой. Но нужно, чтобы каждый оттенок можно было получить смешиванием пары красок, которые подают в автомат.

Для решения этой задачи, пожалуйста, допиши SQL-запрос так, чтобы в результате были доступны пары разных цветов, не повторяющихся при их смешивании:

```
with a as (  
    select 'Красный' as color union all  
    select 'Зеленый' as color union all  
    select 'Синий' as color union all  
    select 'Желтый' as color union all  
    select 'Фиолетовый' as color union all  
    select 'Пурпурный' as color union all  
    select 'Белый' as color) ...
```

То есть нам подходят пары «красный + зеленый», «красный + синий», однако пары «красный + красный» или «зеленый + красный» недопустимы. В первом случае имеет место повтор одного и того же цвета и нет смысла заливать два одинаковых цвета в оба резервуара автомата. Во втором случае при смешивании мы получим такой же цвет, как и для пары «красный + зеленый», указанной выше.

Пожалуйста, пришли свое решение с пояснениями до конца рабочего дня. Спасибо!

Полезные материалы

[Статья](#) про оператор WITH в SQL.

Формат конечного результата

Файл в формате .docx, содержащий дополненный SQL запрос и пояснения к нему.

Форма загрузки результата

Пожалуйста, загрузи свой вариант ответа в формате zip-архива, используя инструмент «Загрузить решение». Необходимо сформировать единый zip-архив, содержащий решение одного или всех заданий по выбранной специальности.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания от эксперта после отправки собственной версии.

Задание 3. Построй диаграммы и графики для визуализации зависимостей

Настало время поработать над финальным проектом стажировки. Открыв почту, ты внимательно изучаешь детали задания, которые прислал Илья сегодня утром.

Привет!

У нас есть файл с данными о покупках, совершенных в трех ресторанах быстрого питания. Необходимо построить несколько визуализаций, которые помогут бы ответить на следующие вопросы:

1. Какова динамика трат посетителей в ресторанах? Есть ли дни наибольшего спроса? Можно ли проследить тренд? Какой ресторан находится на втором месте по популярности?
2. Каково распределение покупателей по возрасту в каждом из ресторанов?
3. Как можно визуализировать одновременно три параметра: сумму покупок, средний чек и количество покупателей?

При работе над этим заданием ты можешь использовать любой доступный инструмент анализа, например Excel, R/Python, но я рекомендую использовать Excel.

Спасибо!

Полезные материалы

[Файл с исходными данными](#) о покупках, совершенных в трех ресторанах быстрого питания.

Формат конечного результата

Файл с нужными диаграммами/матрицами и ответами на вопросы.

Форма загрузки результата

Пожалуйста, загрузи свой вариант ответа в формате zip-архива, используя инструмент «Загрузить решение». Необходимо сформировать единый zip-архив, содержащий решение одного или всех заданий по выбранной специальности.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания от эксперта после отправки собственной версии.