Gmail **Максим Иоган <maksiamiogan@gmail.com>**

(без темы)

Максим Иоган <maksiamiogan@gmail.com> 1 октября 2022 г., 03:50
Черновик

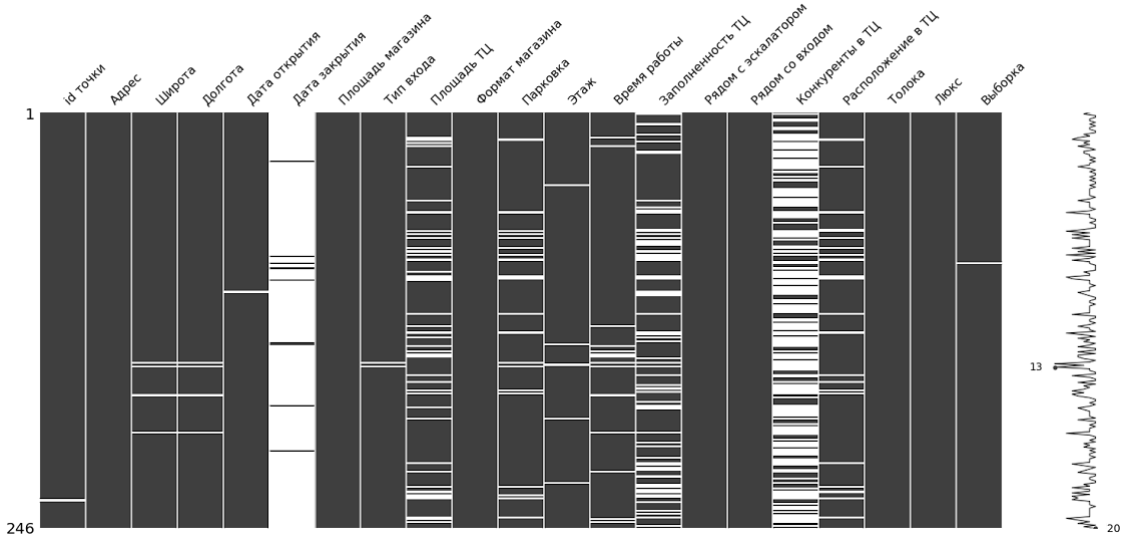
Здравствуйте, я проанализировал ваши данные и у меня есть несколько вопросов.

1. Заполненность данных

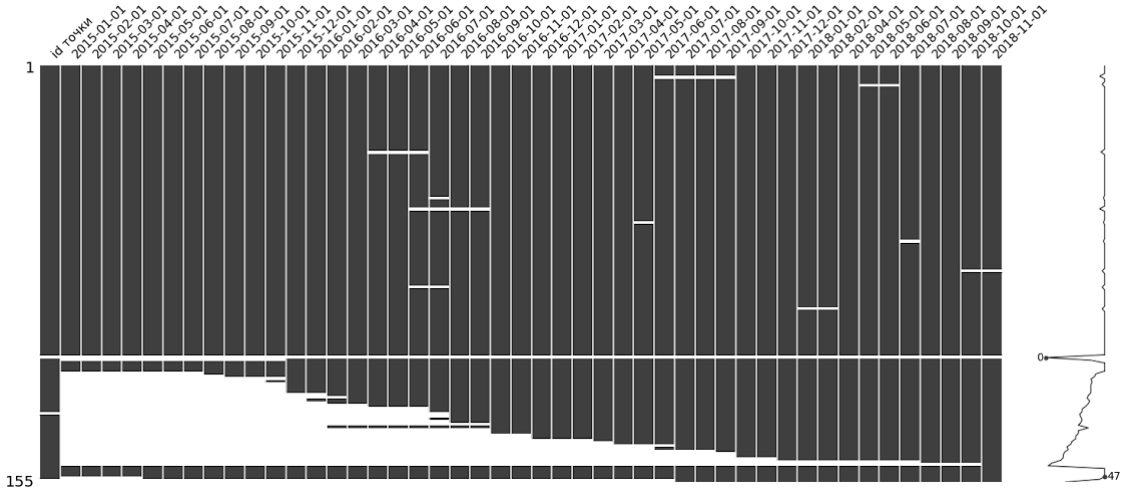
В данных есть большое кол-во пропусков, включая идентификатор точки магазина, широта, долгота, тип входа, площадь тц, парковка, этаж, время работы, заполненность тц, конкуренты в тц, расположение в тц, выборка

Данные пропуски означают отсутствие информации или отсутствие признака у магазина?

Для наглядности прикреплю визуализацию пропусков, где темная область это заполненность, а белая - пропуск



Аналогичная ситуация с листом "Выручка по обучающей выборке". Данные пропуски из-за того, что в те времена учет выручки не велся, магазин не функционировал или данные пропали по ошибке?



2. Проблемы со значениями признаков

2.1) 2 элемента в идентификаторе точки имеют дубликаты, однако идентификатор должен быть уникален

2.2) Адреса не имеют единого шаблона, что может привести к трудностям расшифровки и разделения адреса на отдельные компоненты. Также есть опечатки, лишние кавычки, дер. - д, р-н - р-он, где-то литера написана отдельно от номера дома, где-то вместе, номер дома написан без опознавательных слов или со словами дом №.

2.3) В дате открытия есть две некорректные даты с 2098 года и с 1899 года

2.4) Хотелось бы узнать единицу измерение площади, однако есть подозрение на выброс, так как максимальная площадь составляет 1488, в то время как 75% квартиль - 430. Это касается признакова "Площадь магазина" "Площадь ТЦ", оба имеют выбросы

2.5) Нужна расшифровка типов входа, так как есть подозрение, что, например, сквозной проход и единственный вход могут классифицировать один и тот же магазин. Возможно стоит составить более уникальные категории

2.6) В признаке "Формат магазина" одни и те же категории имеют разное написание. Street - Стрит - Стрт, Мини ТЦ - Мини-ТЦ. Некоторые слова Street написаны с использованием буквы из кириллицы.
2.7) В признаке "Паркова" смешаны строчные категории с цифровыми и буква р в слова парковка в некоторых элементах перепутаны с английской буквой p, что воспринимается компьютером как две разные категории
2.8) Признак "Этаж". Что означает этаж 4.5? 20ый этаж сильно выделяется из общей выборки этажей, хотелось бы проверить на корректность. Среди этажей есть дата. Не очень хорошо, что среди числовых значений есть слово "цоколь", но это можно быстро исправить
2.9) 'Вс-Чт с 10:00 до 23:00 Пт-Сб с 10:00 до 24:00' - 24:00 это 00:00 ? Можно убрать лишние пробелы между словами. Добавить диапазон дней недель в категории, где их нет. 'с 10-00 до 22:00' 'с 10-00 до 21-00'- неверная форма времени.
Некоторые категории записаны с перемешкой кириллицы и латиницы в русских словах.
2.10) Признак "Заполненность ТЦ". Как вычислялись эти категории? Какая область допустимых значений?
2.11) Признак "Рядом с эскалатором". Как вычислялись эти категории? Логически подразумевается две категории "Да" "Нет", дано 3.
2.12) Признак "Рядом со входом". Как вычислялись эти категории? Логически подразумевается две категории "Да" "Нет", дано 4.
2.13) Признак "Расположение в ТЦ"". Как вычислялись эти категории? Что означают? Какая область допустимых значений?
2.14) Признаки "Толока" "Люкс". Как вычислялись? Что означают? Какая область допустимых значений?
2.15) Признак "Выборка". Что означает "-"? Слово "обучающая" иногда написано с латинскими буквами. "Тестовая" = "Тест" ?
2.16) Широта и Долгота некоторых магазинов расположена в океанах, морях и озерах, в местах, где физически не может находиться магазин. Прикрепляю карту с отмеченным магазинами

2.17) В обучающую выборку попали тестовые магазины (по id точки)
2.18) В обучающей выборке почти у всех большое стандартное отклонение и большой разрыв между 75 процентилем и максимумом. Есть года с нулевой выручкой. Слишком большие и слишком маленькие значения выручки выглядят аномально. Хотелось бы удостовериться в их корректности.

3. Ответы на вопросы
3.1) Сколько магазинов формата стрит в тестовой выборке?
Если учитывать, что тестовая выборка это значения "Тестовая" и "Тест" в столбце "Выборка", а магазин формата стрит это значения "Street",'Стрит', 'Street', 'Стрт', то данному условию будут удовлетворять 7

магазинов.
Если строго следовать условию задания, где значение выборки - "Тестовая", а формат магазина "Стрит", то данному условию будут удовлетворять 3 магазина.
3.2) Какова средняя выручка магазинов формата мини ТЦ за 2016 год?
Если учитывать что "Мини ТЦ" и "Мини-ТЦ" это одно и то же, то результат будет равен 346239.18395658414
Если строго следовать условию задания и брать только "Мини ТЦ", то результат будет равен 387582.70035861945
3.3) Сколько магазинов с бесплатной парковкой?
157 магазинов