

projects INSIGHTS

-затронутые темы в проектах:

исправить типы данных

кастомные столбцы

категоризация

pivot_table

чек корреляций

подробно пропуски, дубликаты и аномалии:

Пропуски

Задаться вопросом - в чем причина пропусков?

Причины пропусков

-нулевая категория

-больше некоторого порога

медиана медианы по группировке?

допускаем что группы пропущенные ведут себя как непропущенные

половину медианными - ок????

dropna кажд строки с хотя бы 1 na

дропаем так - теряем данные в др столбцах - поэтому заполнять важно

Важный столбец - сколько дней висело объявление - причина пропусков неясна -

последние пропуски скорее всего не *проданы vs еще висящие vs снятые и тд....*

пропуски не заполняем и не удаляем!

Дубликаты

Когда все пропущенные/аномалии/дубликаты исправлены - нужен (повторный) чек дубликатов

Для идентификации дубликатов см совпадения по нескольким столбцам

- тем где нет missin val-ues
 - в которых маловероятны совпадения
-

Аномалии

не в смысле как стат-термин, а любые ошибки/странности данных

Задаться вопросом - в чем причина аномалий?

мб можно нагуглить

Причины аномалий

опечатки (случайный минус у числа?)

единицы измерения

правила сервиса

человек не знает как заполнить параметр

аномалии/пропуски завязанные на время

влияние сроков

- нан of "проданы или нет" для последних в посл периоде длинной среднего времени продажи
 - правила сервиса (сроки публикации яндекс недвиж)
 - Значение или столбец отсутствовали, но в некоторую дату были добавлены

Номинативная

.value_counts()

Численная переменная

1. Чекаем max и min в .describe
2. Чекаем **hist** и/или **boxplot/violin**
(иногда достаточно
топы меньших и больших)
3. **hist без значений выше усов** или значения где выбросы становятся реже
(аномалией может быть мода/что угодно)
4. Удалить выбросы если есть
5. Если есть аномалия - смотрим др столбцы
6. Ищем причину аном/ чек долю аномалий - если мало - можно выкинуть
7. Чек .describe/hist/box еще раз
8. выводы - моды,
сегменты по кол-ву
симметрия столбца vs обраб столбца

Чек индикация категорий без ошибок

Идейное

Динамика цен может быть объяснена балансом спроса предложения

Чек корреляций/взаимосвязи

зависимая переменная - количественная

если **независимая - колич**:

1 кроме **scatter** можно применить **лайнplot медиан групп по независимой**

2 коэф корр-ции

если **независимая - номин**:

1 тот же **лайнplot**

2 процентные различие категорий

если **независимая - номин**:

1 кроме **scatter** можно применить **лайнplot медиан групп по независимой**

2 **pivot - median + count** или **процентные различия категорий**