**Комментарии к данным**

**«Выручка по обучающей выборке»**

1. 2 ID неизвестны (были удалены для дальнейшего анализа)
2. Отсутствуют данные по выручке за некоторые месяцы (см. ноутбук)
3. Для ID 42383 и 42597 есть дупликаты по данным о выручке, которые не совпадают (были удалены для дальнейшего анализа из обоих файлов). Таким образом, в обучающей выборке 148 уникальных IDs.

**«Справочник точек»**

1. Для ID 42683 есть дупликат, для которого данные не совпадают. Обе строки были удалены.
2. В данных есть много пропущенных значений (см. ноутбук). Вопросы по данным:

* Дата закрытия пропущена в 233 позициях. Магазин работает по настоящий момент или информация потеряна?
* Конкуренты в ТЦ: 149 значений пропущено, причем есть значения 0. Заменить на 0 или нет данных по конкурентам?
* Необходимо уточнить данные по Широте и Долготе для магазинов с ID 42437, 42443, 42509, 42667
* Необходимо уточнить Дату открытия магазина c ID 42220
* Площадь ТЦ – пропущено 42 значения. Уточнить данные или удалить столбец?
* Необходимо уточнить Тип входа для магазинов с ID 42437, 42443 (пока замена на моду)
* Парковка – 22 значения пропущены для магазинов с ID 18010.0, 42001.0, 42056.0, 42059.0, 42072.0, 42100.0, 42118.0, 42132.0, 42189.0, 42191.0, 42288.0, 42345.0, 42437.0, 42443.0, 42499.0, 42505.0, 42780.0, 42800.0, 42810.0, 2633928.0 (пока замена на моду – бесплатная парковка)
* Необходимо уточнить Этаж для магазинов с ID 41914,42388, 42438,42598, 42775 (пока замена на моду – 1 этаж)
* Необходимо уточнить время работы для магазинов с ID 17987, 18527, 42319, 42389, 42406, 42419, 42420, 42437, 42443, 42509, 42667, 42745, 2633927, 2633930
* Заполненность ТЦ – 63 пропущенных значения. Удалить столбец?
* Необходимо уточнить Расположение в ТЦ для магазинов с ID 18010.0, 41842.0, 42001.0, 42051.0, 42056.0, 42072.0, 42100.0, 42118.0, 42132.0, 42189.0, 42191.0, 42288.0, 42345.0, 42466.0, 42481.0, 42499.0, 42505.0, 42730.0, 42780.0, 42792.0, 42810.0, 2633928.0

1. Обработка пропущенных и некорректных значений:

* Пропущенное значения для столбца Выборка (ID 42135) было отнесено к обучающей выборке, так как ID магазина был найден в данных по выручке
* Строка с пропущенным значением ID была удалена
* Этаж «цоколь» помечен как -1. Этажи 20 и 4,5 заменены на моду (1).
* «Рядом с эскалатором» и «Рядом с входом» – это номинальные переменные, где 0 – нет, 1 – да. Некорректные значения 2, 3 заменены на моду
* Некорректное значение времени в графе «Этаж» было удалено (можно заменить на моду при необходимости)
* Некорректное значения в столбце Выборка (-) было заменено на «Обучающая» для магазинов с ID 42192, 42204, 42156 и на «Тестовая» с ID 42340, 42122, 42546, 18931, 42388, 42706. Таким образом, в файле 155 значений обучающей выборки и 84 значений тестовой выборки.

1. Ошибки в разметке данных на обучающую и тестовую выборку:

* В справочнике магазинов отсутствует информация о 8 магазинах (IDs: 2633921, 45801, 42831, 2634000, 2633936, 2634008, 2634011, 426830). Эти значения были удалены из файла о выручке
* 63 магазина необходимо отнести к тестовой выборке, так как нет информации о выручке (ID см. в ноутбуке)
* 48 магазинов необходимо отнести к обучающей выборке, так как про них есть информация о выручке (ID см. в ноутбуке).
* Итого: в обучающей выборке 140 магазинов, в тестовой – 99 магазинов.

1. Ответы на вопросы:

* В тестовой выборке 8 магазинов формата Стрит
* В обеих выборках 157 магазинов с бесплатной парковкой (83 – в обучающей выборке, 74 – в тестовой выборке)
* Средняя выручка магазинов типа мини-ТЦ за 2016 год: вместо пропущенных значений (магазины 42720, 42743, 42803 ) присваиваю 0, так как нет данных о том, что магазины были закрыты до 2016 года. Значение среднего по выручке для каждого магазина представлено в таблице ниже (также см. ноутбук)

|  |  |
| --- | --- |
|  | **ID** |
| **0** | 18728 | 447705.269242 |
| **1** | 41824 | 329577.635033 |
| **2** | 41933 | 115160.841858 |
| **3** | 42067 | 652438.859274 |
| **4** | 42192 | 191355.248942 |
| **5** | 42219 | 180890.465756 |
| **6** | 42242 | 282498.755571 |
| **7** | 42355 | 134952.291671 |
| **8** | 42368 | 633120.934025 |
| **9** | 42422 | 224917.983557 |
| **10** | 42458 | 291624.632058 |
| **11** | 42496 | 297769.558321 |
| **12** | 42504 | 238500.484337 |
| **13** | 42548 | 316680.647989 |
| **14** | 42720 | 141401.151208 |
| **15** | 42743 | 209543.027145 |
| **16** | 42803 | 0.000000 |