Напишите программу на Python согласно вашему варианту для датасета “brooklyn_sales_map.csv”. Напишите небольшой отчёт, который должен содержать титульный лист, вариант задания, код программы и полученные вами результаты. Для помощи в выполнении задания можно воспользоваться книгами «Изучаем Spark. Молниеносный анализ данных -2015», «PySpark SQL Recipes» или интернетом.

Вариант 1.

- Найдите среднюю стоимость жилья (sale_price) и выведите новую таблицу, содержащую две колонки – стоимость жилья и отклонение стоимости от среднего значения.
- Найдите среднюю стоимость жилья (sale_price) для каждого района.
- Выведите среднюю полную площадь жилья (gross_sqft) для всех сочетаний налоговых категорий (tax_class) и лет продажи (year_of_sale).
- Выведите таблицу, содержащую количество пустых (null) значений для каждой колонки.

Вариант 2.

- Найдите среднюю площадь жилья (gross_sqft) и выведите новую таблицу, содержащую стоимость жилья и отклонение стоимости от среднего значения.
- Найдите среднюю площадь жилья (gross_sqft) для каждого года, в котором оно было построено (year_built).
- Найдите среднюю стоимость жилья (sale_price) для всех сочетаний соседств (neighborhood) и категорий класса здания (building_class_category).
- В исходном датафрейме удалите все строчки с записями домов, которые были построены позже 2000 года (year_built), а также те, которые содержат только нулевые значения.

Вариант 3.

- Найдите среднюю стоимость жилья (sale_price) и выведите новую таблицу, содержащую две колонки – стоимость жилья и процент отклонения стоимости от среднего значения.
- Выведите таблицу, содержащую все категории класса зданий (building_class_category) и количество записей, которые к ним относятся.
- Выведите таблицу, содержащую средние значения по каждому столбцу в датафрейме.
- В исходном датафрейме заполните все нулевые значения средними по столбцу.

Вариант 4.

- Добавьте в датасет новую колонку, содержащую «возраст» жилья.
- Выведите таблицу, содержащую среднюю дату продажи для всех сочетаний индексов (zip_code) и налоговых категорий (tax_class) жилья.
- Выведите таблицу, содержащую суммарную стоимость жилья (sale_price) по всем сочетаниям налоговых категорий (tax_class) и индексов (zip_code).
- Создайте новую таблицу, в которой есть 10 колонок исходного датафрейма, в которых нулевые значения есть, но не преобладают. После этого удалите все строки, в которых содержатся исключительно нули.

Вариант 5.

- Найдите средний год постройки жилья (year_built) и выведите новую таблицу, содержащую год постройки жилья и отклонение года постройки от среднего значения.
- Отсортировать датасет по возрастанию цены продажи (sale_price) и убыванию индексов (zip_code) одновременно.
- Выведите таблицу с наибольшими ценами продажи (sale_price) и количеством зданий по каждому сочетанию соседства (neighborhood) и категории класса здания (building_class_category).