Гид по безопасному Бостону

В этом задании предлагается собрать статистику по криминогенной обстановке в разных районах Бостона. В качестве исходных данных используется датасет https://www.kaggle.com/AnalyzeBoston/crimes-in-boston
С помощью Spark соберите агрегат по районам (поле district) со следующими метриками:

- crimes_total - общее количество преступлений в этом районе
- crimes_monthly - медиана числа преступлений в месяц в этом районе
- frequent_crime_types - три самых частых crime_type за всю историю наблюдений в этом районе, объединенных через запятую с одним пробелом “, ” , расположенных в порядке убывания частоты
    - crime_type - первая часть NAME из таблицы offense_codes, разбитого по разделителю “-” (например, если NAME “BURGLARY - COMMERICAL - ATTEMPT”, то crime_type “BURGLARY”)
- lat - широта координаты района, расчитанная как среднее по всем широтам инцидентов
- lng - долгота координаты района, расчитанная как среднее по всем долготам инцидентов

Программа должна упаковываться в uber-jar (с помощью sbt-assembly), и запускаться командой

```
spark-submit --master local[*] --class com.example.BostonCrimesMap
/path/to/jar {path/to/crime.csv} {path/to/offense_codes.csv}
{path/to/output_folder}
```

где {...} - аргументы, передаваемые пользователем.
Результатом её выполнения должен быть один файл в формате .parquet в папке path/to/output_folder.
Для джойна со справочником необходимо использовать broadcast.

Ссылку на репозиторий с кодом прислать в чат с преподавателем.

Методика оценки:

1. Программа выдает корректный файл на выходе - 3 балла
2. Здание сдано в срок до 00:00 04.10.2019 - 1 балл
3. Задание сдано с первой попытки (повторные попытки из-за неточностей в условиях задания не считаются) - 1 балл

Вам могут пригодиться следующие материалы:

1. Документация по Spark: https://spark.apache.org/docs/latest/
2. Документация по SQL функциям Spark: https://spark.apache.org/docs/latest/api/sql/index.html

Подсказка: в спарке есть функция percentile_approx, которая может посчитать медиану.