Описание проекта

В проекте вам нужно обучить модель линейной регрессии на данных о жилье в Калифорнии в 1990 году. С этим датасетом вы уже работали в четвёртой теме курса.
В колонках датасета содержатся следующие данные:

- `longitude` — широта;
- `latitude` — долгота;
- `housing_median_age` — медианный возраст жителей жилого массива;
- `total_rooms` — общее количество комнат в домах жилого массива;
- `total_bedrooms` — общее количество спален в домах жилого массива;
- `population` — количество человек, которые проживают в жилом массиве;
- `households` — количество домовладений в жилом массиве;
- `median_income` — медианный доход жителей жилого массива;
- `median_house_value` — медианная стоимость дома в жилом массиве;
- `ocean_proximity` — близость к океану.

На основе данных нужно предсказать медианную стоимость дома в жилом массиве — `median_house_value`. Обучите модель и сделайте предсказания на тестовой выборке. Для оценки качества модели используйте метрики RMSE, MAE и R2.

Инструкция по выполнению проекта

1. Инициализируйте локальную Spark-сессию.
2. Прочитайте содержимое файла `/datasets/housing.csv`.
3. Выведите типы данных колонок датасета. Используйте методы pySpark.
4. Выполните предобработку данных:
    - Исследуйте данные на наличие пропусков и заполните их, выбрав значения по своему усмотрению.
    - Преобразуйте колонку с категориальными значениями техникой One hot encoding.
5. Постройте две модели линейной регрессии на разных наборах данных:
    - используя все данные из файла;
    - используя только числовые переменные, исключив категориальные.
      Для построения модели используйте оценщик `LinearRegression` из библиотеки MLlib.
6. Сравните результаты работы линейной регрессии на двух наборах данных по метрикам RMSE, MAE и R2. Сделайте выводы.