ETL: автоматизация подготовки данных

Урок 2. Введение в подготовку данных для аналитиков. Таблицы фактов и таблицы измерений

Задание .

1. Скачайте датасет fifаs2.сsv. Проанализируйте его и определите, какие данные являются

неполными. Удалите

ненужные колонки и недостающие значения.

2. Найдите в датафрейме полные дубликаты и удалите их. Значения могут быть одинаковыми, но

написаны по-разному. Например, может отличаться размер регистра (заглавные и строчные

буквы). Особое внимание уделить колонке с названиями команд.

3. Напишите функцию, которая добавит колонку с разбиением возраста по группам: до 20, от 20

до 30, от 30 до 36 и старше 36. Посчитайте количество футболистов в каждой категории.

ЗАДАНИЕ НУЖНО СДЕЛАТЬ С ПОМОЧЬЮ APACHE SPARK КАК ПОКАЗАНО НА СЕМИНАРЕ. Соберите

все в один pdf, код и скриншоты записанных таблиц с данными, скриншоты отработки лога

спарка.

















