Университет ИТМО

Практическая работа №1

по дисциплине «Визуализация и моделирование»

Автор: Мигулаева Татьяна Алексеевна
Поток: ВИМ 1.2
Группа: К3222
Факультет: ИКТ
Преподаватель: Чернышева А.В.

Санкт-Петербург, 2021 г.

Датасет:

В данном датасете представлены еженедельные данные сканирования розничной торговли за 2018 год по продажам разоичных сортов авокадо в различных городах. Данные сканирования розничной торговли поступают непосредственно из кассовых аппаратов розничных продавцов на основе фактических розничных продаж авокадо Хасс.

Данные датасета:

Date - Дата наблюдения - date - формат удобен для чтения

AveragePrice - средняя цена одного авокадо - float - в долларах, формат удобен для чтения

Type - обычные или органические - string - лучше перевести в числовой формат

Year - год - integer - формат удобен для чтения

Region - город или район наблюдения - string - перевести в числовой формат

Total Volume - Общее количество проданных авокадо - integer - округлить до целого

4046 - Общее количество проданных авокадо с PLU 4046 - integer - округлить до целого

4225 - Общее количество проданных авокадо с PLU 4225 - integer - округлить до целого

4770 - Общее количество проданных авокадо с PLU 4770 - integer - округлить до целого

Small bags - Количество проданных стандартных упаковок с авокадо - integer - округлить до целого

Large bags - Количество проданных больших упаковок с авокадо - integer - округлить до целого

Total bags - Общее количество проданных упаковок с авокадо - integer - округлить до целого

Задачи лабораторной работы:

Проанализировать данные и произвести их предобработку.

1. Дата в датасете хранится в удобном формате Date, нет никакой лишней или непонятной информации, с ней легко и удобно работать.

2. Средняя цена авокадо хранится в формате float, но только с двумя знаками после запятой, исчисление ведется в долларах. Формат удобен для обработки.

3. Тип авокадо - стандартный или органический. Так как здесь в каждой строке датасета прописано слово, то это затрудняет обработку. Переведем в числовой формат, обозначив традиционный тип - 1, а органический - 2.

4. Регион - различные регионы США. Намного удобнее будет работать с ними в числовом формате, так как названия порой написаны непонятно (какие-то с пробелом, какие-то нет), и это неудобно. Создадим массив со всеми названиями, и обозначим каждый его номером в массиве - таким образом регионы будут также отсортированны по алфавиту.

5. Год - поскольку данные собраны только за несколько лет, здесь варианты от 2015

до 2018, его формат удобен для обработки.

6. Общее количество проданных авокадо и их продажи, разделенные по разным видам. В каждой строке слишком много цифр после запятой, а учитывая смысл строк, их можно округлить до целого числа.

7. Количество проданных упаковок с авокадо и их деление по видам - ситуация аналогична с отдельными авокадо, зная целое число, работать было бы намного удобнее.

В данном датасете нет необходимости удалять пустые строки, так как поиск по датасету показал их отсутствие. Также нет необходимости добавлять новые столбцы, поскольку все нужные данные можно получить прямым путем, нет необходимости высчитывать что-либо. Таким образом, можно сказать, что на преобразовании столбцов предобработка датасета "Авокадо"закончена.