МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра ИС

ОТЧЕТ

по лабораторной работе №1
по дисциплине «Машинное обучение»
Тема: Исследование набора данных

Студентка гр. 1373 _____ Новикова А.С.

Преподаватель _____ Татчина Я.А.

Санкт-Петербург
2023

Цель работы: исследовать алгоритмы классификации и кластеризации на выбранном наборе данных.

1. Краткое описание набора данных

Данные о ценах и объеме продаж авокадо на нескольких рынках США. Датасет был взят с сайта www.kaggle.com. Все данные реальны, они были загружены с веб-сайта The Hass Avocado Board в мае 2018 года.

Все данные в датасете - числовые. В датасете представлены следующие атрибуты:

- AveragePrice - средняя цена за авокадо
- Total Volume - сколько всего авокадо продано
- 4046 - продажи авокадо с кодом 4046
- 4225 - продажи авокадо с кодом 4225
- 4770 - продажи авокадо с кодом 4770
- Total Bags - сколько всего упаковок авокадо продано
- Small Bags - сколько маленьких упаковок авокадо продано
- Large Bags - сколько больших упаковок авокадо продано
- XLarge Bags - сколько очень больших упаковок авокадо продано

2. Определение параметров

а) Среднее значение, СКО

Среднее значение и СКО атрибутов были определены с помощью функций библиотеки numpy 'np.mean' и ‘np.std’ соответственно.

- AveragePrice: среднее = 1.40, СКО = 0.40
- Total Volume: среднее = 850644.01, СКО = 3453450.73
- 4046: среднее = 293008.42, СКО = 1264954.42
- 4225: среднее = 295154.57, СКО = 1204087.41
- 4770: среднее = 22839.74, СКО = 107461.12
- Total Bags: среднее = 239639.20, СКО = 986215.38
- Small Bags: среднее = 182194.69, СКО = 746158.07
- Large Bags: среднее = 54338.09, СКО = 243959.280.4
- XLarge Bags: среднее = 3106.43, СКО = 17692.41

b) Построить гистограммы распределения значений, определить есть ли выбросы

Поиск выбросов был осуществлен с помощью следующего алгоритма:

- Отсортировать данные и найти Q1 и Q3 (1-ый и 3-ый квартиль)
- Найти межквартильный размах IQR
- Проверить, какие наблюдения вышли за границы [Q1 – 1.5*IQR; Q3 + 1.5*IQR]

Наличие выбросов:

- AveragePrice: нет
- Total Volume: да
- 4046: да
- 4225: да
- 4770: да
- Total Bags: да
- Small Bags: да
- Large Bags: да
- XLarge Bags: да

Гистограммы распределений представлены на рис. 2.1 – 2.9

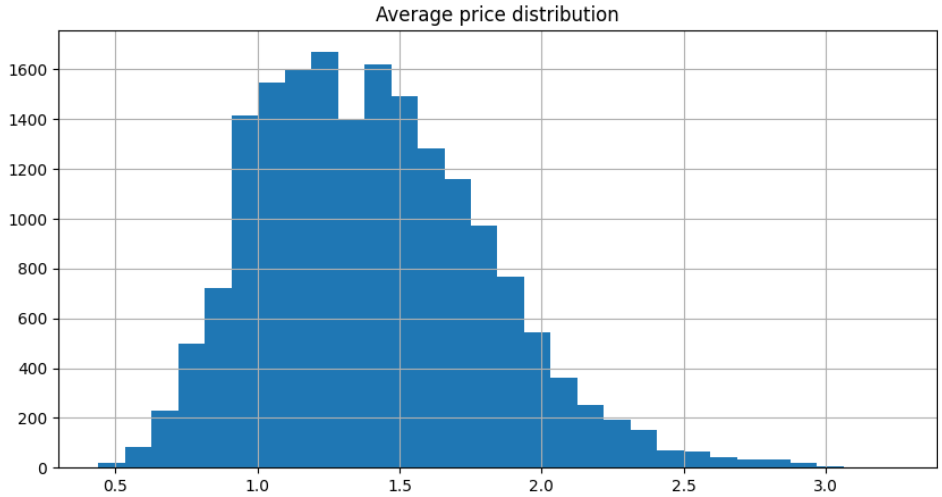

Рисунок 2.1 – Распределение средней цены

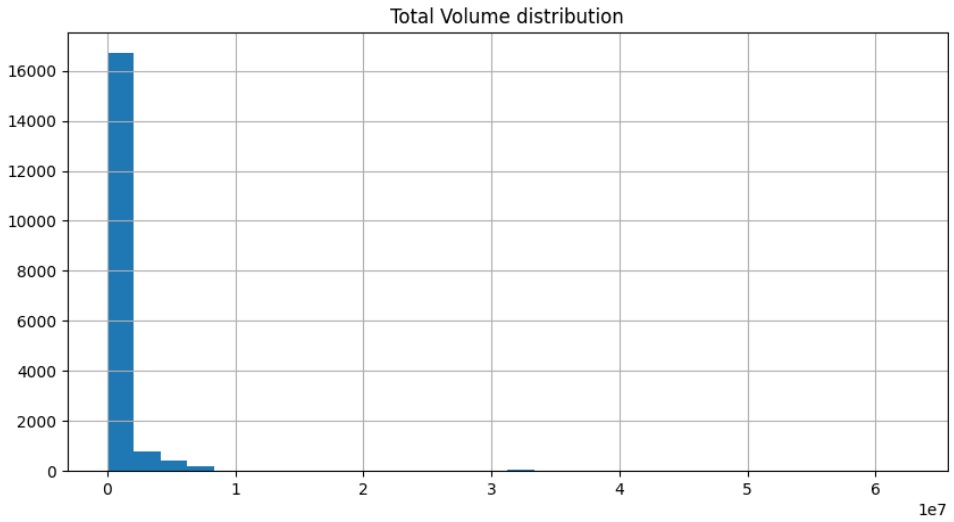

Рисунок 2.2 – Распределение общих продаж

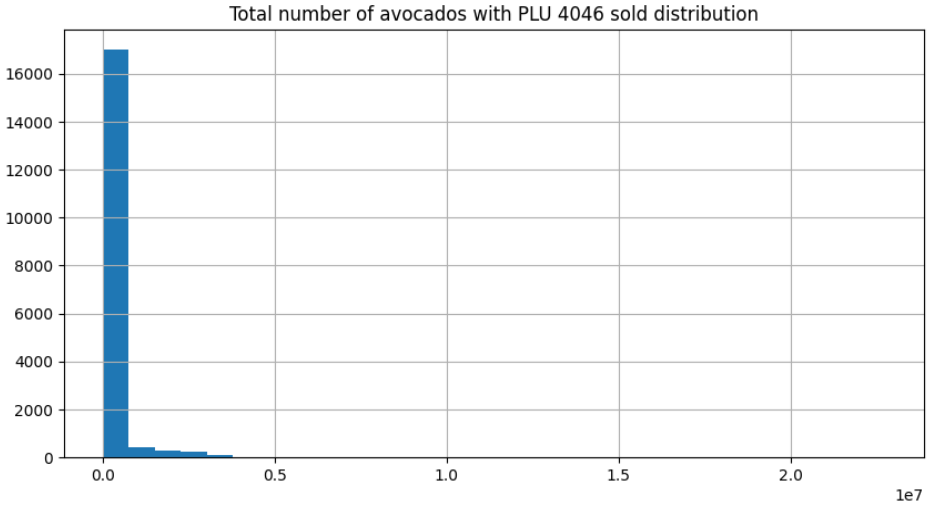

Рисунок 2.3 – Распределение продаж авокадо с кодом 4046

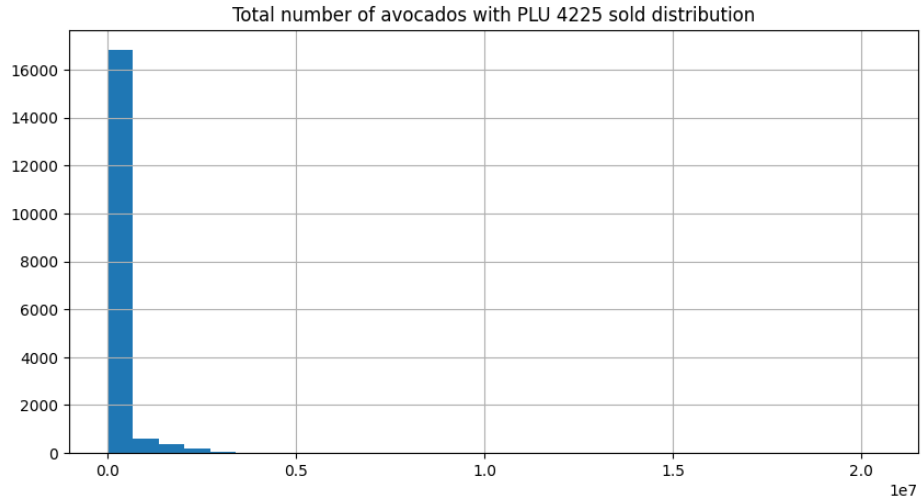

Рисунок 2.4 – Распределение продаж авокадо с кодом 4225

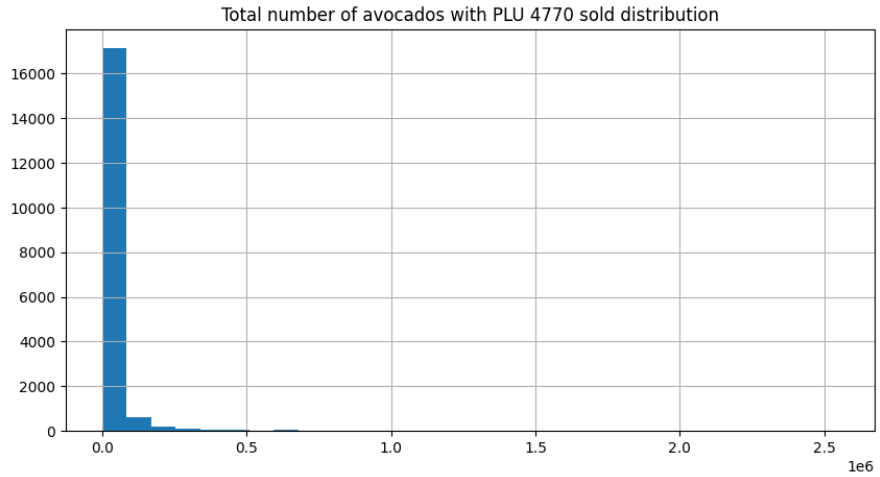

Рисунок 2.5 – Распределение продаж авокадо с кодом 4770

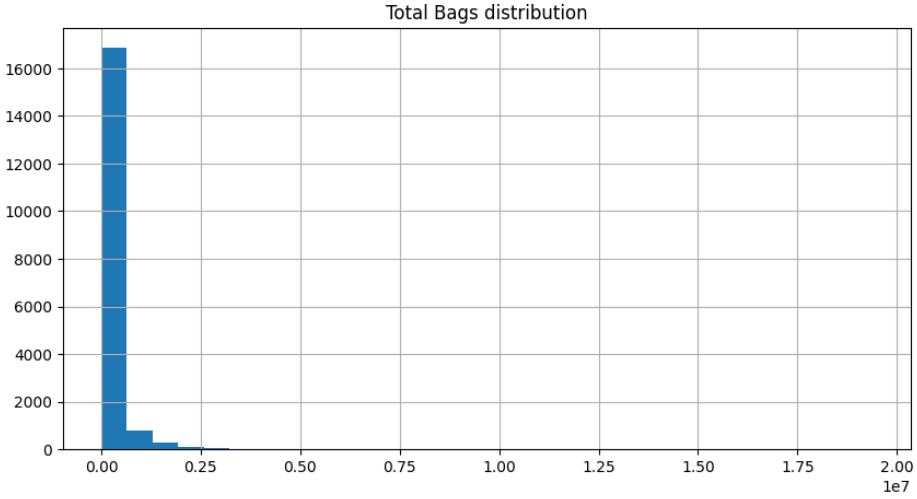

Рисунок 2.6 – Распределение общих продаж упаковок

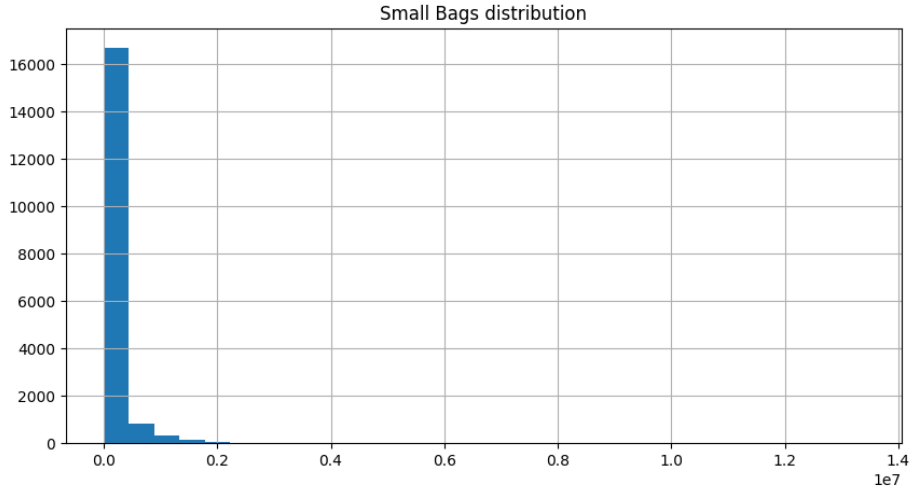

Рисунок 2.7 – Распределение продаж маленьких упаковок

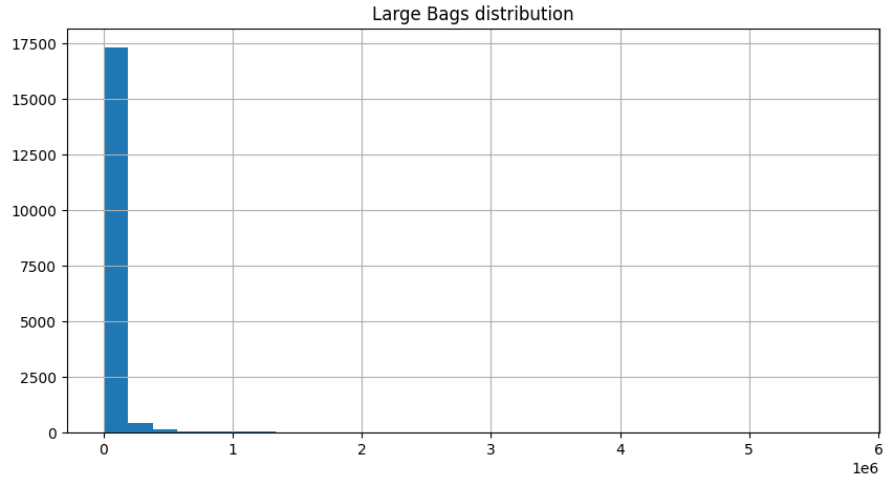

Рисунок 2.8 – Распределение продаж больших упаковок

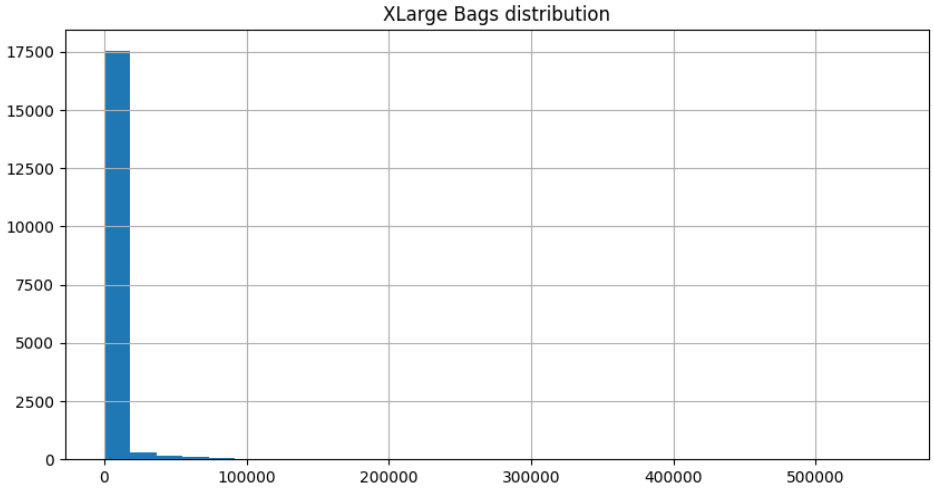

Рисунок 2.9 – Распределение продаж очень больших упаковок

c) Определить, есть ли пропущенные значения

Наличие пропущенные значений определялось с помощью функции df[‘Название’].isna().sum()

В ходе работы ни у одного атрибута не нашлось пропущенных значений

d) Предложить вариант обработки пропущенных значений

Так как пропущенных значений нет, то и обрабатывать их не нужно

3. Определение корреляции

В ходе работы я рассмотрела 5 зависимостей: средней цены от общих продаж, продаж авокадо с кодом 4046 от продаж авокадо с кодом 4770, продаж маленьких упаковок от продаж авокадо с кодом 4225, продаж очень больших упаковок от средней цены, продаж маленьких упаковок от продаж больших упаковок.

a) Определить, какие атрибуты высокоррелированы и характер корреляции

Для этого пункта я находила матрицу корреляции с помощью функции ‘np.corrcoef’.

- Средняя цена от общих продаж: коэффициент = -0.19, если средняя цена растёт, то с небольшой вероятностью общие продажи низкие
- 4046 от 4770: коэффициент = 0.83, если продажи авокадо с кодом 4046 большие, то, скорее всего, и продажи авокадо с кодом 4770 тоже большие
- Маленькие упаковки от 4225: коэффициент = 0.92, если продажи маленьких упаковок большие, то, скорее всего, и продажи авокадо с кодом 4225 тоже большие
- Большие упаковки от средней цены: коэффициент = -0.12, если продажи больших упаковок большие, то с небольшой вероятностью цена ниже средней
- Маленькие упаковки от больших упаковок: коэффициент = 0.90 если продажи маленьких упаковок большие, то, скорее всего, и продажи больших упаковок авокадо тоже большие

b) Какие атрибуты не имеют корреляцию

Из исследованных атрибутов все имеют корреляцию. Самая близкая к 0 наблюдается у продаж больших упаковок авокадо и средней цены.

c) Построить графики рассеивания

Графики рассеивания представлены на рис. 3.1 – 3.5

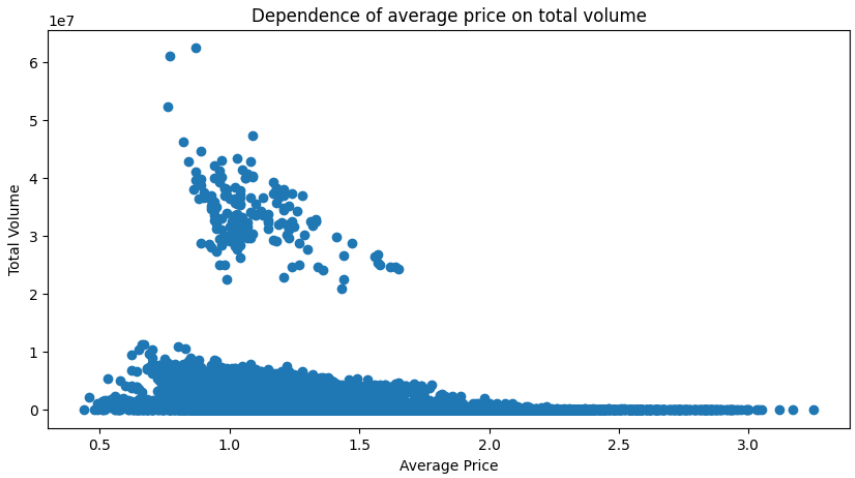

Рисунок 3.1 – Зависимость средней цены за авокадо от общих продаж

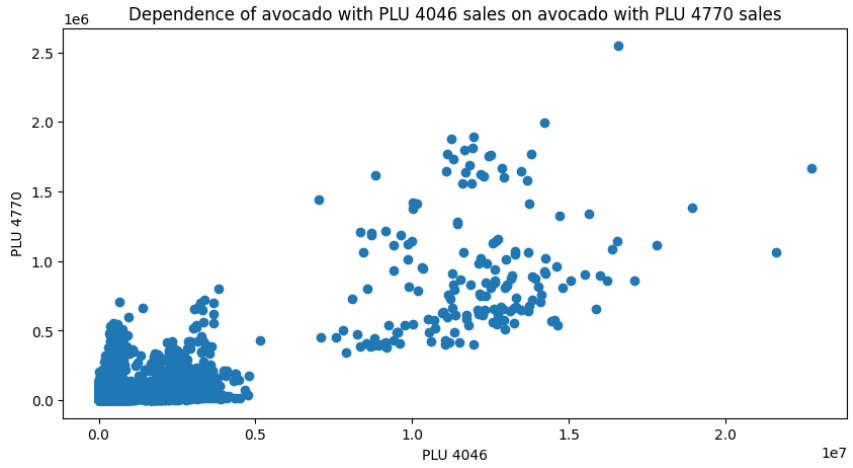

Рисунок 3.2 – Зависимость продаж авокадо с кодом 4046 от продаж авокадо с кодом 4770

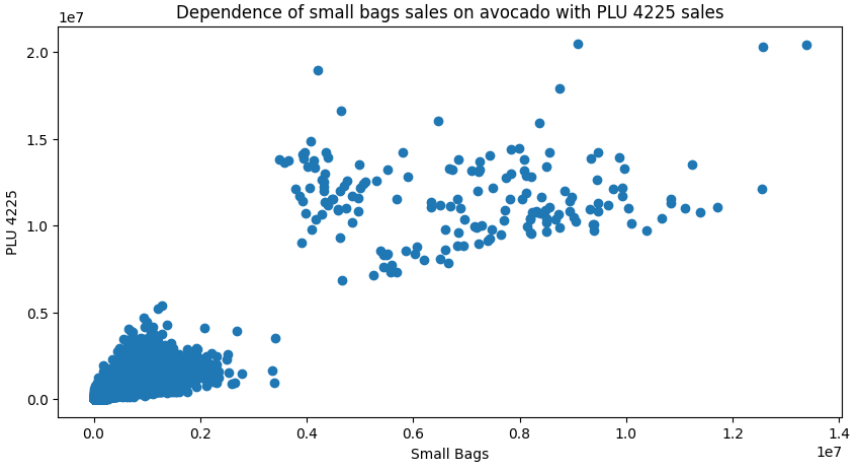

Рисунок 3.3 – Зависимость продаж маленьких упаковок от продаж авокадо с кодом 4225

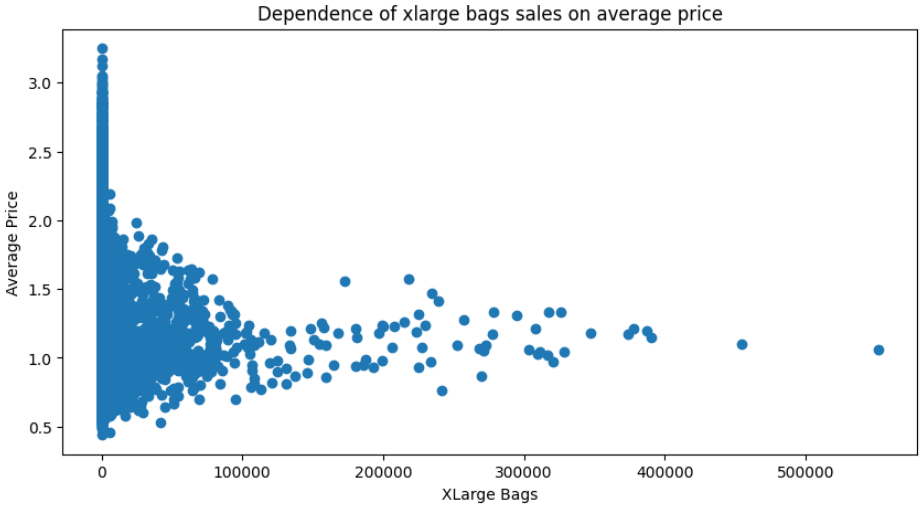

Рисунок 3.4 – Зависимость продаж очень больших упаковок от средней цены за авокадо

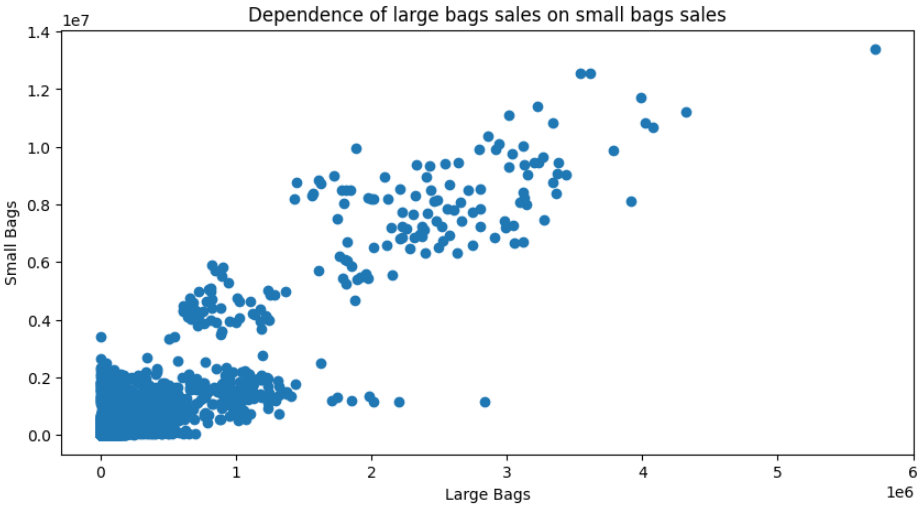

Рисунок 3.5 – Зависимость продаж маленьких упаковок от продаж больших упаковок

d) Проанализировать полученные результаты

Из полученных результатов мы можем сделать вывод, что в основном все виды продаж пропорционально зависят друг от друга, а также с увеличением стоимости за одно авокадо, продажи, скорее всего, будут низкими.