**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра информационных систем**

отчет

**по лабораторной работе №1**

**по дисциплине «Машинное обучение»**

Тема: Исследование набора данных

|  |  |  |
| --- | --- | --- |
| Студент гр. 2372 |  | Михеева Э. Д. |
| Преподаватель |  | Татчина Я.А. |

Санкт-Петербург

2024

**Цель работы**

Выбрать набор данных, который будет использоваться в дальнейшем при исследовании алгоритмов кластеризации и классификации.

**Постановка задачи.**

1. Создать Jupyter Notebook
2. Выбрать набор данных
3. Описать выбранный набор:
   1. Указать предметную область, источник, характер данных;
   2. Описать представленные атрибуты, их тип, что они обозначают;
   3. Представить задачу анализа.
4. Для каждого атрибута:
   1. Определить среднее значение, СКО;
   2. Построить гистограмму распределения значений, определить наличие выбросов;
   3. Определить наличие пропусков и их количество;
   4. Предложить вариант обработки пропущенных значений.
5. Определить:
   1. Коррелирующие атрибуты, характер их корреляции;
   2. Атрибуты без корреляции;
   3. Построить графики или матрицу рассеивания.
6. Проанализировать полученные результаты.

**Выполнение работы**

Для анализа выбран набор данных о характеристиках мобильных телефонов и их ценах. Предметной областью работы стала классификация ценовой категории телефона по его характеристикам. Набор данных является синтетическим и содержит 21 атрибут, включая ценовую категорию.

Способом представления данных выбран pandas.DataFrame. После чтения данных из CSV-файла было проверено существование дубликатов и/или пустых значений – ни дубликатов, ни пустых полей не оказалось. С помощью метода describe() были определены основные параметры набора данных, включая среднее значение и СКО каждого признака.

Далее, с помощью инструментария seaborn были построены графики распределения значений. Для этого набор данных был разделен на логические и числовые составляющие. Так как данные синтезированы, распределения непрерывны и не содержат явных шумов или выбросов, большая часть атрибутов распределена равномерно.

Применив метод corr() была построена тепловая карта матрицы корреляции признаков, из которой выделены зависимые атрибуты:

1. Наблюдается средняя взаимосвязь разрешения основной и фронтальной камеры – как правило, в устройствах с качественной основной камерой, ставится более хорошая фронтальная;
2. Связаны параметры высоты и ширины экрана, что логично;
3. Поддержки стандартов 4G и 3G высоко коррелируют, ведь 3G является предшественником 4G;
4. Заметна слабая зависимость цены телефона от ресурса его батареи;
5. Цена устройства сильно (92%) зависит от объема оперативной памяти;
6. Разрешение экрана (параметры высоты и ширины) влияют на цену, корреляция составляет 16%.

Построив матрицы рассеяния для всех пар параметров с помощью seaborn.pairplot() были подтверждены указанные зависимости. Наиболее интересные для изучения зависимости были изучены более детально.

Разрешение камер действительно взаимосвязано, причём корреляции с ценой наблюдать не приходится (рис. 1).

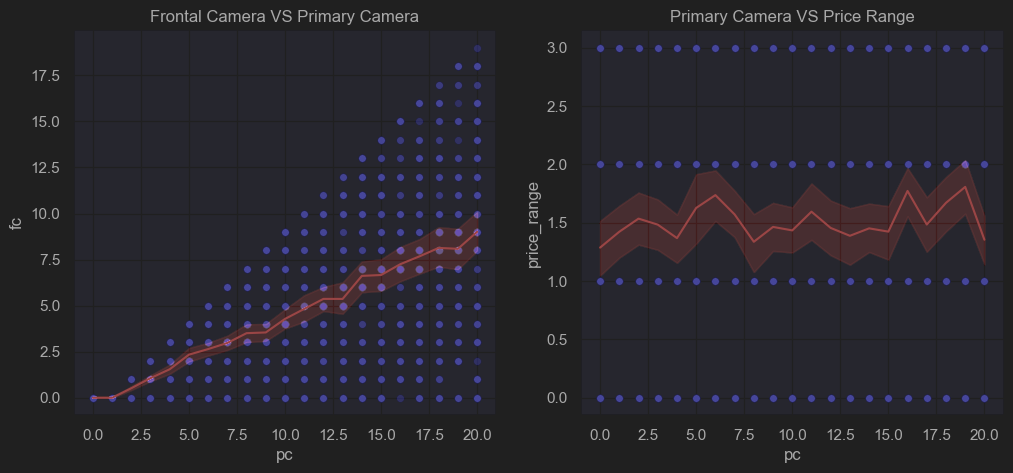


Рисунок . Зависимость разрешения камер, цены от разрешения

Действительно, можно наблюдать зависимость ценовой категории от емкости аккумулятора и объема оперативной памяти. При этом, первая зависимость выражена намного слабее (рис. 2, 3).

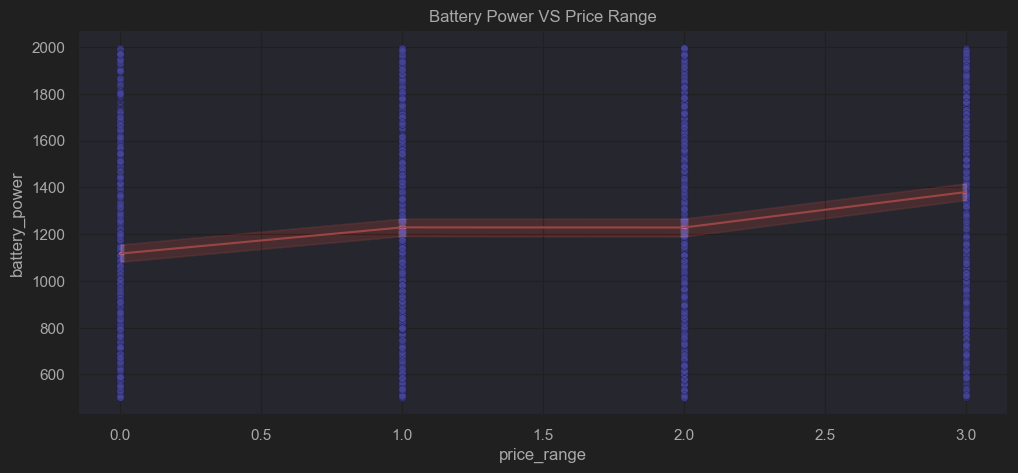


Рисунок . Зависимость емкости от цены

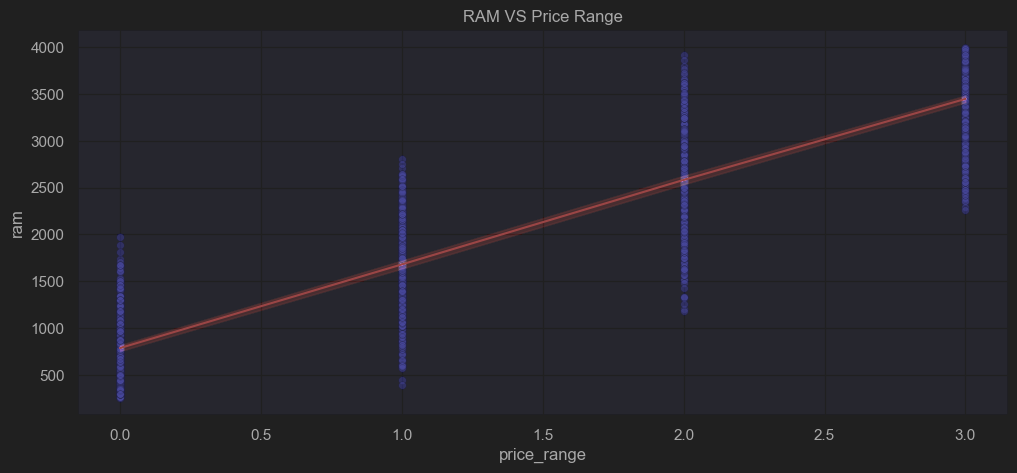


Рисунок . Зависимость объема ОЗУ от цены

**Выполнение работы**

ДВ ходе работы были получены навыки работы в Jupyter Notebook, анализа данных с помощью pandas, numpy и визуализации с применением matplotlib, seaborn. Были проанализированы распределения и корреляция атрибутов выбранного набора данных.