МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Омский государственный технический университет»

Факультет информационных технологий и компьютерных систем

Кафедра «Прикладная математика и фундаментальная информатика»

Домашнее задание

по дисциплине Практикум по программированию

Студента(ки) Шохина Егора Павловича
фамилия, имя, отчество полностью

Курс 2 Группа ФИТ-221

Направление 02.03.02. Фундаментальная информатика и информационные технологии
код, наименование

Руководитель ст.преподаватель
должность, ученая степень, звание

Саматов А. П.
фамилия, инициалы, дата, подпись

Выполнил _____
дата, подпись студента(ки)

Итоговый рейтинг	

Омск 2023

ВВЕДЕНИЕ 3
1.Поиск и загрузка данных 4
2.1 Гистограмма распределения числового признака 5
2.2 Диаграмма «ящик с усами» числового признака 6
2.3 Круговая диаграмма номинативного признака 6
2.4 Тепловая карта 7
2.5 Диаграмма countplot с группировкой по двум номинативным признакам 8
3 Предварительная обработка данных 9
ЗАКЛЮЧЕНИЕ 11
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 12

ВВЕДЕНИЕ

Объемы накопленных данных в настоящее время настолько внушительны, что человеку просто не по силам проанализировать их самостоятельно, хотя необходимость проведения такого анализа вполне очевидна, ведь в этих "сырых данных" заключены знания, которые могут быть использованы при принятии решений, формировании статистических отчетов или составлении моделей машинного обучения. В ходе изучения курса были использованы следующие библиотеки для языка программирования Python:

1. NumPy — библиотека с открытым исходным кодом с поддержкой многомерных массивов (включая матрицы) и высокоуровневых математических функций, предназначенных для работы с многомерными массивами.
2. Matplotlib — это библиотека для визуализации данных. В ней можно построить двумерные (плоские) и трехмерные графики.
3. SymPy — это библиотека Python с открытым исходным кодом, используемая для символьных вычислений. Она предоставляет возможности компьютерной алгебры в виде отдельного приложения.
4. SciPy — библиотека с открытым исходным кодом, предназначенная для выполнения научных и инженерных расчётов.
5. Pandas — программная библиотека для обработки и анализа данных. Предоставляет специальные структуры данных и операции для манипулирования числовыми таблицами и временными рядами.
6. Seaborn — библиотека для создания статистических графиков на Python. Она построена на основе matplotlib и тесно интегрируется со структурами данных pandas.Эти библиотеки позволяют проводить обработку, анализ и визуализацию данных, строить статистику на их основе

1.Поиск и загрузка данных

Использован датасет Customer Shopping Trends Dataset (https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset). Kaggle.com, специализирующемся на исследовании данных и машинном обучении.

```
# Расчетно-графическая работа по дисциплине "Практикум по программированию"

Использован датасет Customer Shopping Trends Dataset
(https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset)

В нем представлены:
- Age-Возраст покупателя,
- Sex-Пол покупателя,
- Item Purchased-Приобретенная вещь,
- Category-Категория вещи,
- Purchase Amount (USD) - цена вещи в долларах,
- Location-Место покупки,
- Size-размер,
- Color-цвет вещи ,
- Season - время года
- Review Ranting-оценка вещи
- Subscription Status-статус подписки
- Shipping Type -тип доставки
- Discount Applied - применение скидки
- Promo Code Used - использование промокода
- Previous Purchases - предыдущие покупки
- Payment Method - способ оплаты
- Frequency of Purchases-частота покупок
```

Рисунок 1 – файл README.md

Датасет был загружен в ноутбук командой read_csv() библиотеки pandas.

```
Импорт датасета:

Ввод [108]: data=pd.read_csv('shopping_trends_updated.csv')
```

Рисунок 2 – загрузка датасета

Данный датасет выглядит как набор из 3900 строк и 17 столбцов, разделенных точкой с запятой, в которых описаны тренды покупок в зависимости от возраста, пола и других данных клиента и в зависимости от этих параметров заполнены его предпочтения.

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Ship Type
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Expr
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Expr
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Ship
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Air
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Ship

Рисунок 3 – небольшая часть датасета, выведенного в виде таблицы

2.1 Гистограмма распределения числового признака

Гистограмма — способ представления табличных данных в графическом виде — в виде столбчатой диаграммы. Количественные соотношения некоторого показателя представлены в виде прямоугольников, площади которых пропорциональны. На гистограмме видно количество оценок и сопоставленный этому количеству рейтинг.

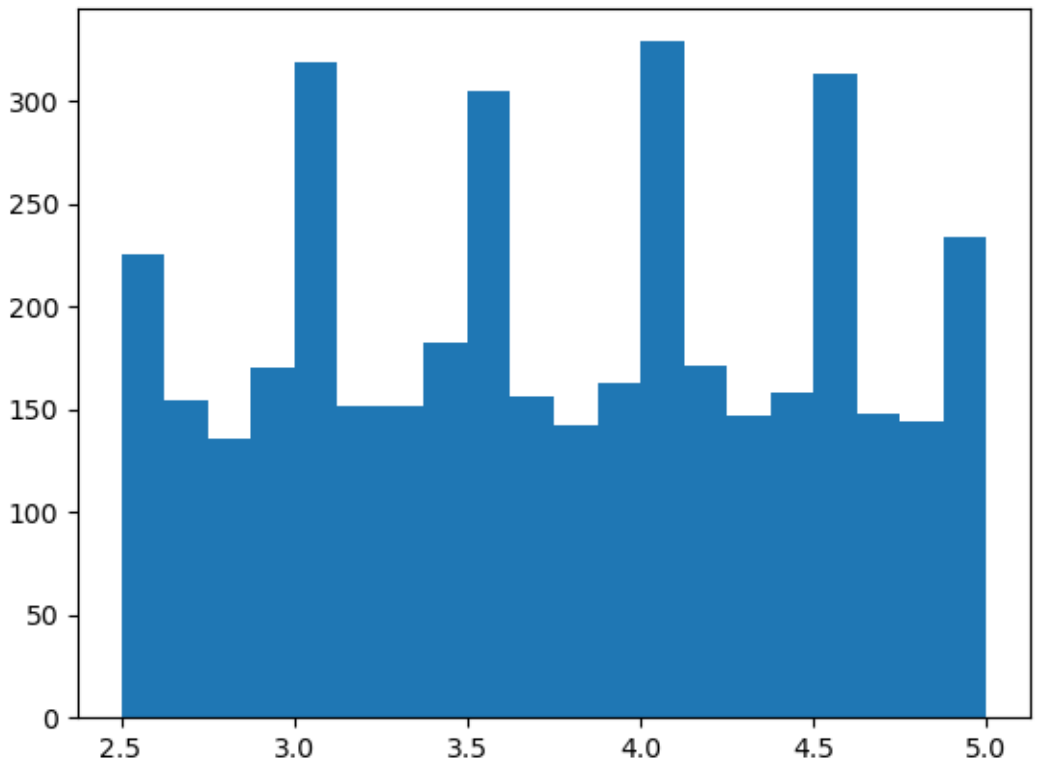

Рисунок 4 – гистограмма столбца Review Rating

2.2 Диаграмма «ящик с усами» числового признака

Диаграмма «ящик с усами» — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей. 7 Такой вид диаграммы в удобной форме показывает медиану (или, если нужно, среднее), нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы. Несколько таких ящиков можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим; их можно располагать как горизонтально, так и вертикально. Расстояния между различными частями ящика позволяют определить степень разброса (дисперсии) и асимметрии данных и выявить выбросы. На диаграмме видно что выбросов как таковых нет.

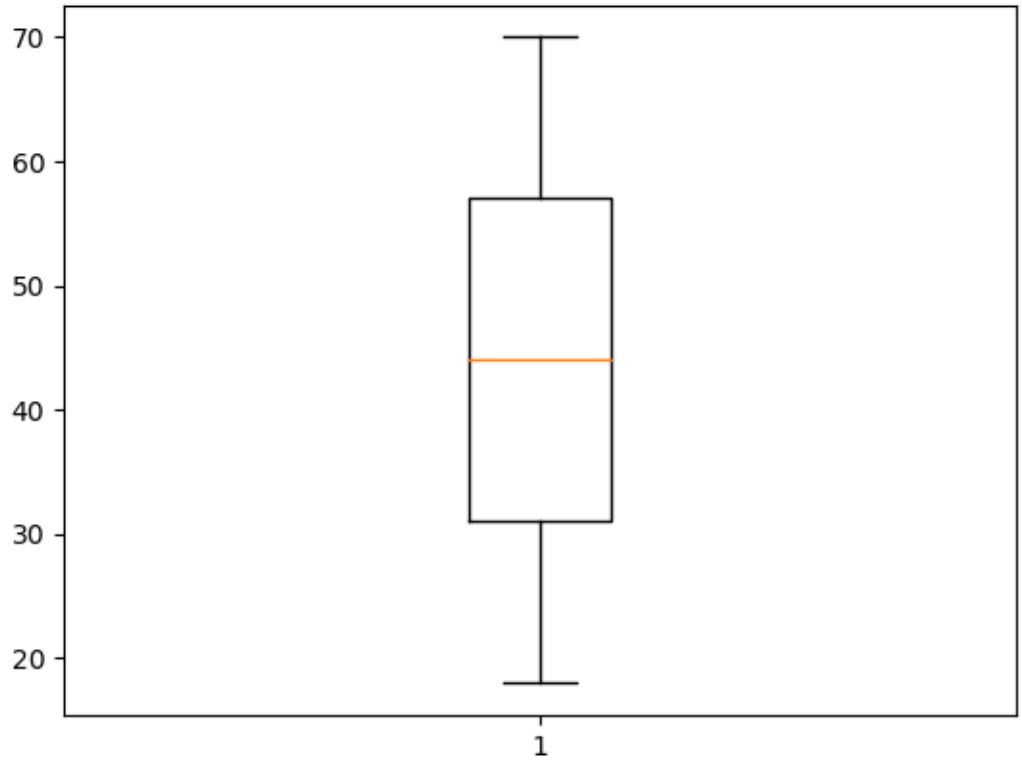

Рисунок 5 – Диаграмма «ящик с усами» столбца Age

2.3 Круговая диаграмма номинативного признака

Круговая диаграмма — это круговая статистическая диаграмма, которая разделена на срезы, чтобы проиллюстрировать числовую пропорцию. На круговой диаграмме длина дуги каждого среза пропорциональна величине, 8 которую он представляет. На данной круговой диаграмме видно, что

распределение клиентов по возрасту практически одинаковое.

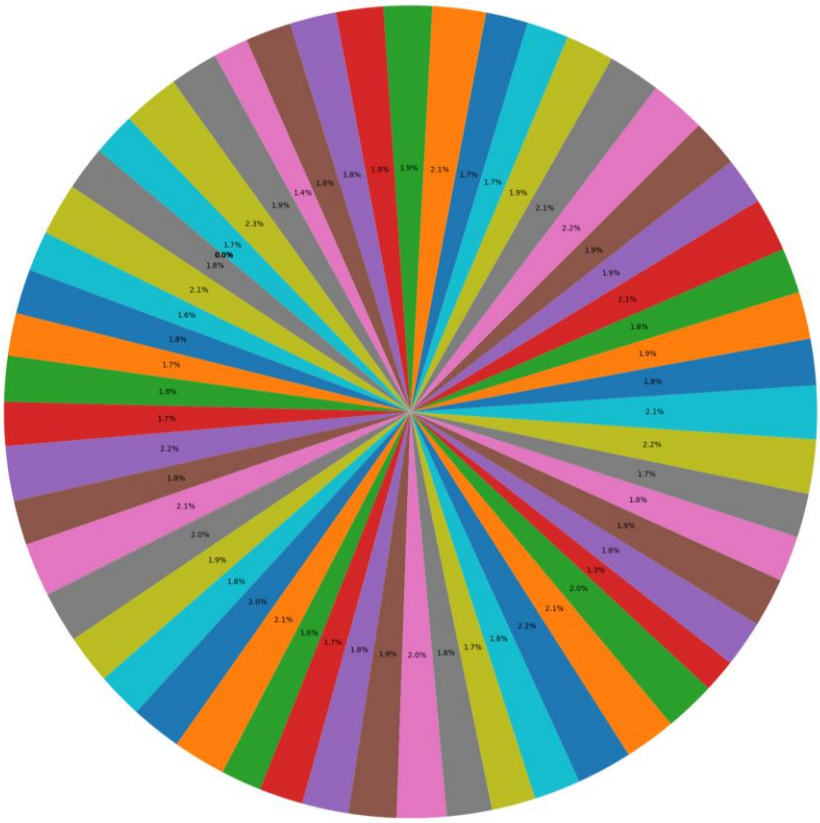

Рисунок 6 – Круговая диаграмма Age

2.4 Тепловая карта

Тепловая карта — графическое представление данных, где индивидуальные значения в таблице отображаются при помощи цвета. На тепловой карте данного датасета можно выявить несколько особенностей, что корреляции между данными как таковой нет.

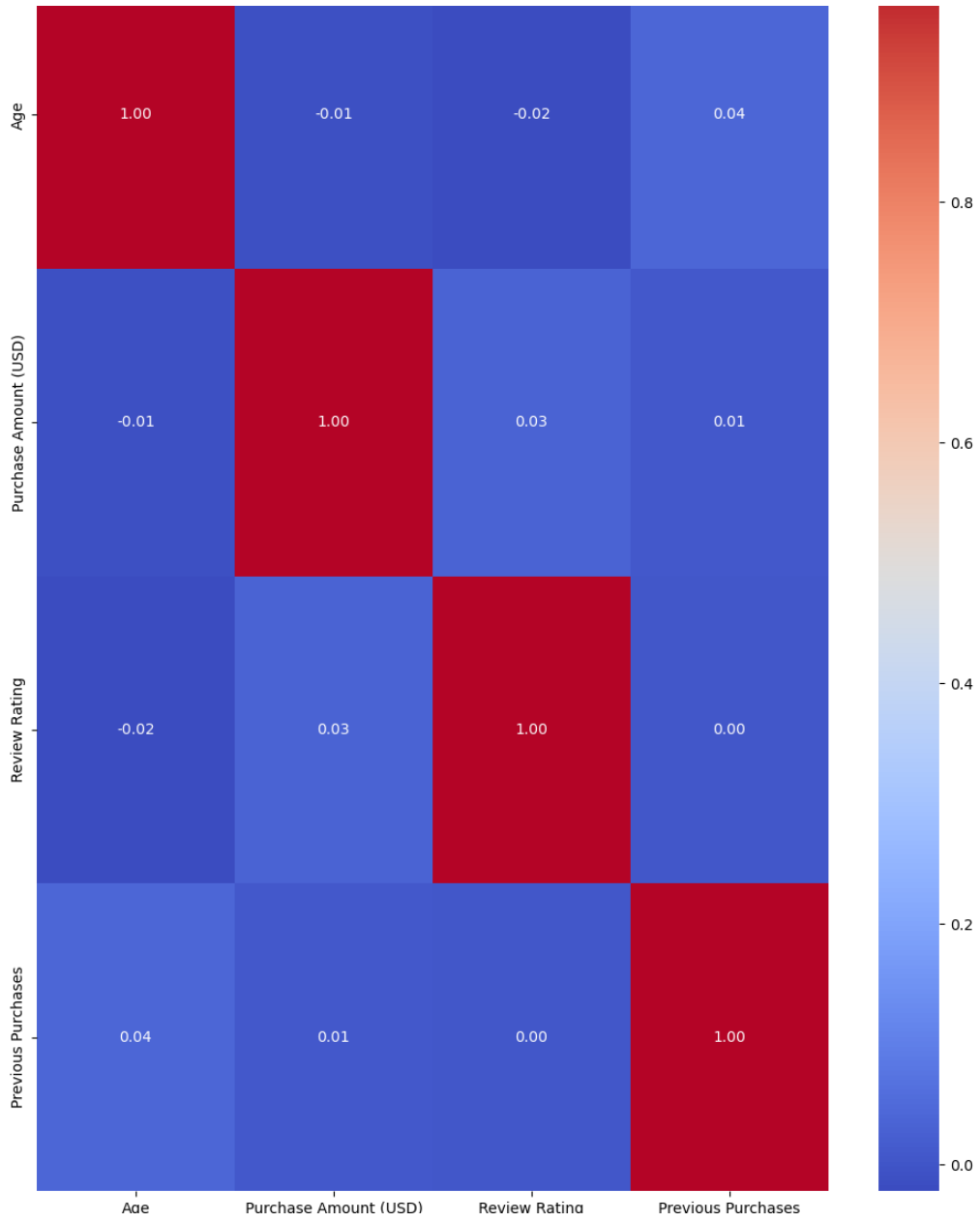

Рисунок 7 –фрагмент тепловой карты датасета

2.5 Диаграмма countplot с группировкой по двум номинативным признакам

CountPlot - столбчатая диаграмма, чаще всего используется для категориальных признаков в данных. Показывает, сколько строчек в датасете имеют каждое из выбранного значения категориального признака. Данная диаграма , что молодые оценивают чаще и выше , в отличие людей среднего возраста(от 40 до 60 лет) .

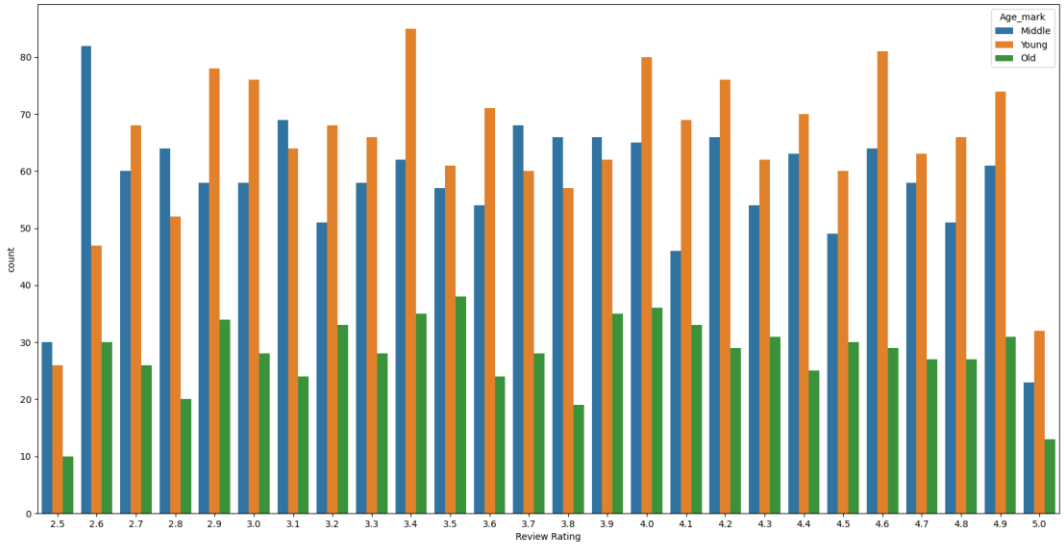

Рисунок 8 – Диаграмма countplot по столбцам Age и Review_Rating

3 Предварительная обработка данных

Данные заполнены без пропущенных значений ,значит заполнять как либо по моде и по среднему значению нам не надо.

Проверка на пустые значения

```
Ввод [121] data.isnull().sum()
```

```
Age                       0
Sex                       0
Item Purchased            0
Category                  0
Purchase Amount (USD)     0
Location                  0
Size                      0
Color                     0
Season                    0
Review Rating             0
Subscription Status       0
Shipping Type             0
Discount Applied          0
Promo Code Used           0
Previous Purchases        0
Payment Method            0
Frequency of Purchases    0
Age_mark                  0
dtype: int64
```

Рисунок 9 – Проверка на наличие пропусков в таблице

Также было применено one-hot кодирование, то есть преобразование категориальных переменных в численные путем создания столбцов под каждую категорию и заполнения их значениями 0 и 1 в зависимости от категории каждой строчки.

Горячее кодирование:

Ввод [123]
```
data=pd.get_dummies(data)
```

Ввод [124]
```
data
```

	Age	Purchase Amount (USD)	Review Rating	Previous Purchases	Sex_Female	Sex_Male	Item Purchased_Backpack	Item Purchased_Belt	Item Purchased_Blouse	I
0	55	53	3.1	14	0	1	0	0	1	(
1	19	64	3.1	2	0	1	0	0	0	(
2	50	73	3.1	23	0	1	0	0	0	(
3	21	90	3.5	49	0	1	0	0	0	(
4	45	49	2.7	31	0	1	0	0	1	(
...	...	...	...	...	...	...	...	...	...	.
3895	40	28	4.2	32	1	0	0	0	0	(
3896	52	49	4.5	41	1	0	1	0	0	(
3897	46	33	2.9	24	1	0	0	1	0	(
3898	44	77	3.8	24	1	0	0	0	0	(
3899	52	81	3.1	33	1	0	0	0	0	(

3900 rows × 146 columns

Рисунок 10 – Горячее кодирование

Пред обработанные данные были сохранены в формате .csv в той же директории, что и изначальный датасет.

Экспорт датасета:

Ввод [125]
```
data.to_csv(r'C:/Users/djego/OneDrive/Рабочий стол/practicum/res.csv',sep=';',index=False)
```

Рисунок 11 –Экспорт датасета

ЗАКЛЮЧЕНИЕ

В ходе прохождения практики были изучены и использованы различные библиотеки Python, такие как matplotlib, seaborn, pandas и numpy. Эти библиотеки являются необходимыми инструментами для работы с данными и визуализации результатов.

Библиотека matplotlib позволяет создавать различные графики и диаграммы, которые помогают визуализировать данные и делать выводы. С помощью seaborn можно создавать более сложные графики, такие как тепловые карты и распределения. Библиотека pandas предоставляет возможность работать с данными в формате таблицы, что упрощает анализ и обработку данных. Библиотека numpy позволяет проводить математические операции с массивами данных.

В процессе работы с этими библиотеками были решены различные задачи, связанные с анализом данных и визуализацией результатов. Были созданы графики, диаграммы, тепловые карты и распределения, которые помогли понять структуру данных и выявить закономерности.

В целом, использование этих библиотек значительно ускоряет процесс анализа данных и позволяет делать более точные выводы. Они являются необходимым инструментом для работы с данными в Python и рекомендуются к изучению всем, кто занимается анализом данных.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. https://numpy.org/doc/stable/reference/generated/numpy.matrix.html (дата обращения: 30.10.23).
2. https://seaborn.pydata.org/installing.html(дата обращения: 30.10.23).
3. https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html (дата обращения: 30.10.23).
4. https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.tight_layout.html (дата обращения: 30.10.23).