Проект НИС

Диана Лубневская, Парфенюк Елена

November 10, 2023

1 Введение

Данная работа посвящена анализу датасета "Customer shopping trends". Наш датасет позволяет получить ценную информацию о поведении покупателей и тенденциях покупок. Понимание предпочтений покупателей и трендов необходимо для ведения бизнеса в этой сфере и стратегий в маркетинге. Датасет охватывает широкий спектр характеристик покупателей включая возраст, пол, историю покупок, предпочитаемые способы оплаты, частота совершения покупок и многое другое. Анализ этих данных может помочь компаниям принимать решения, оптимизировать процесс продажи, тем самым удовлетворяя покупателя.

В ходе исследования нам предстоит предобработать данные, провести разведочный анализ данных и решить задачи классификации и линейной регрессии

2 Информация о датасете

Выбранный датасет представляет из себя набор сведений о покупках совершенных различными людьми. Каждая строка представляет из себя следующий набор данных: возраст покупателя, пол, выбранный товар, категория товара, цена, место покупки, размер, цвет, сезон, рейтинг товара на онлайн платформах, наличие/отсутствие подписки, способ доставки товара, наличие/отсутствие скидки на товар, использовался ли при покупке промокод, траты совершенные до данной покупки, способ оплаты и частота совершаемых покупок. Датасет составлен при помощи чата GPT и лишь моделирует реальную экономическую ситуацию. Всего 3900 строк

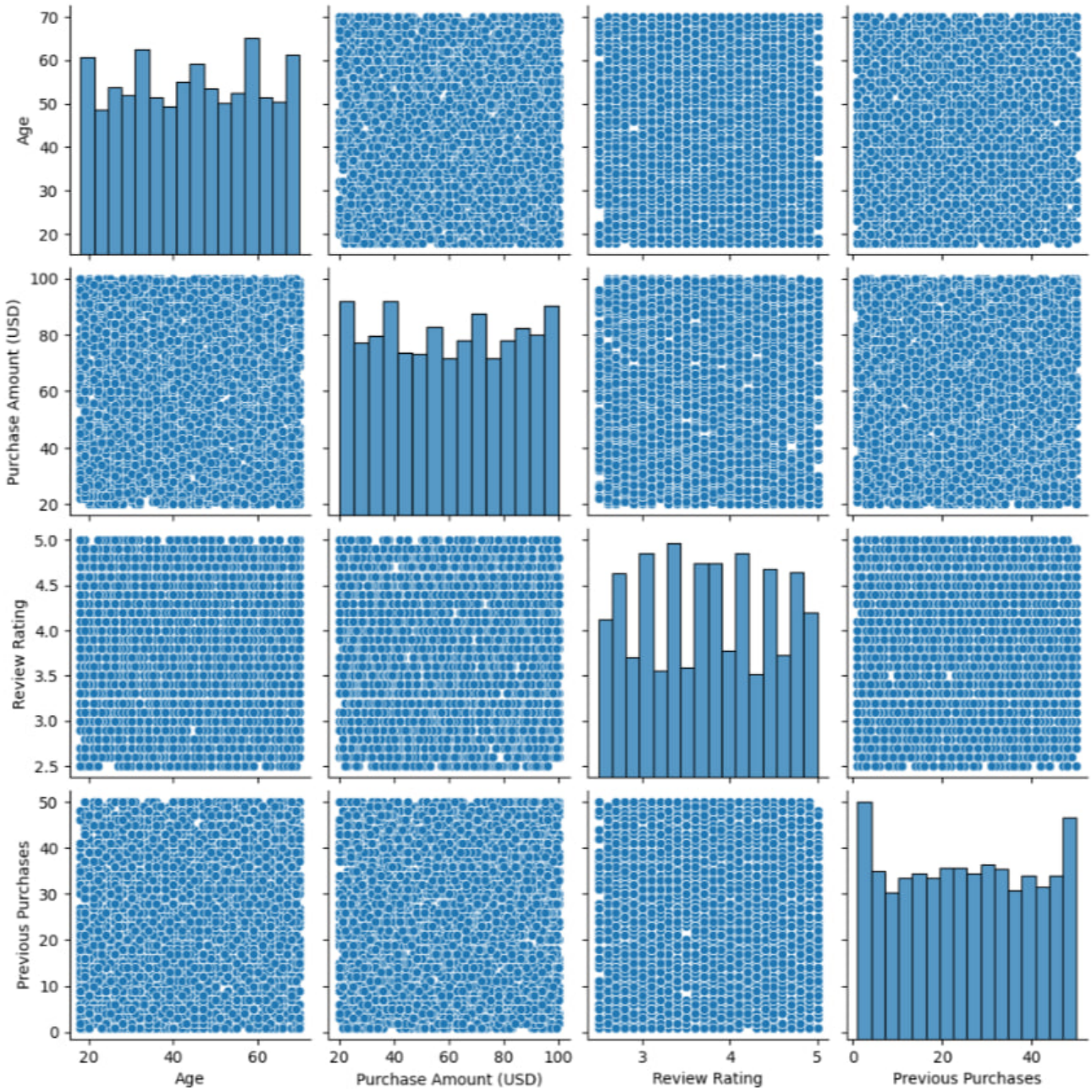
3 Цели проекта

Целью проекта является модель, определяющая по информации о покупке возраст покупателя. Это необходимо, для более точного определения целевой аудитории некоторого продукта в определенных условиях что помогает в выстраивании маркетинговой модели, а также определении оптимальных условий продаж.

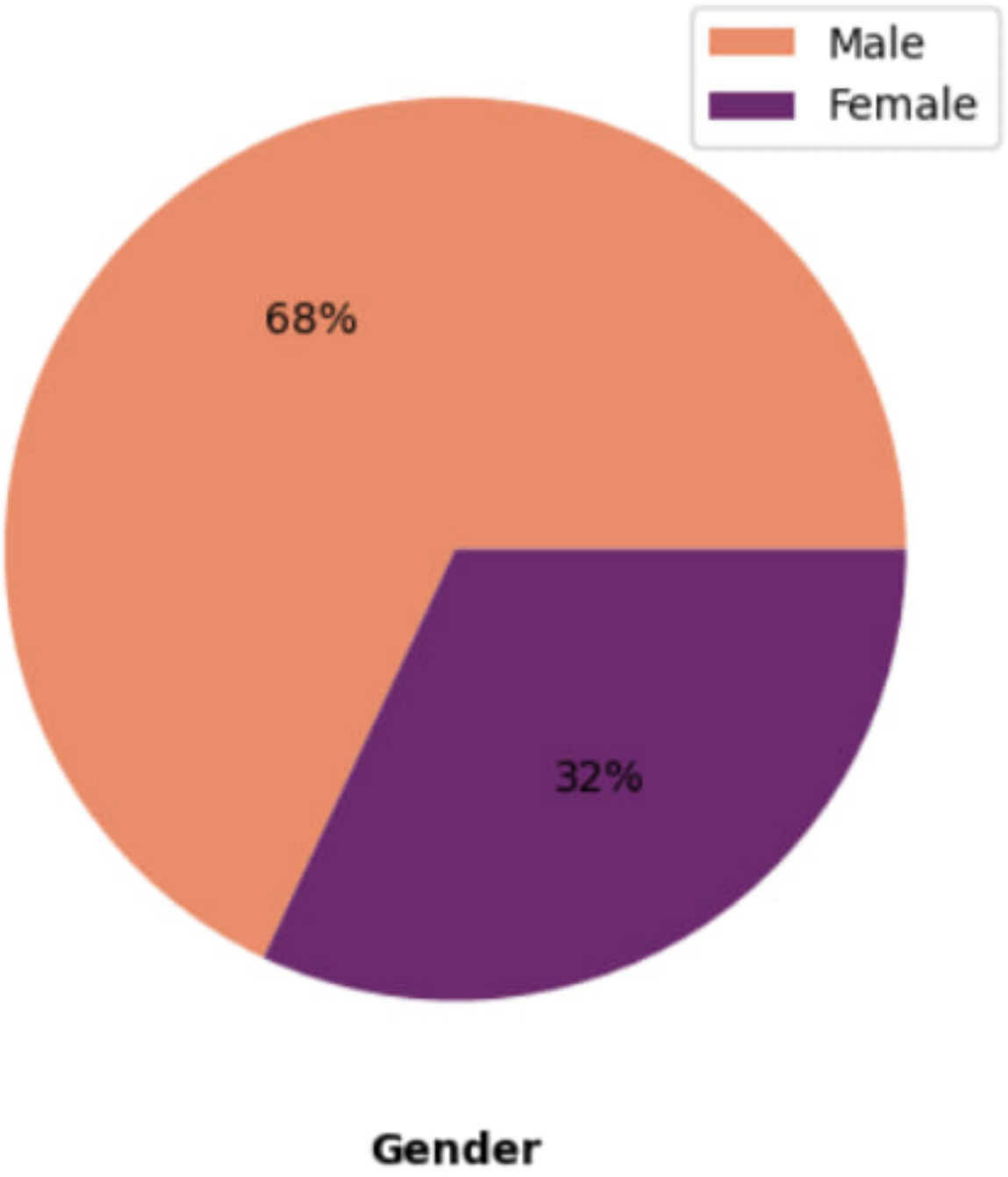
4 Описание процесса анализа

Перед построением модели мы проводим анализ имеющихся данных для того, чтобы понять, что вообще происходит в датасете и какую модель мы можем построить.
Начнем с построения тепловой карты. Это необходимо для того, чтобы определить, коррелируют ли какие-то признаки друг с другом. Если уровень корреляции высокий, то это значит, что от одного из признаков имеет смысл избавиться, чтобы не портить гипотетическую модель линейной регрессии.

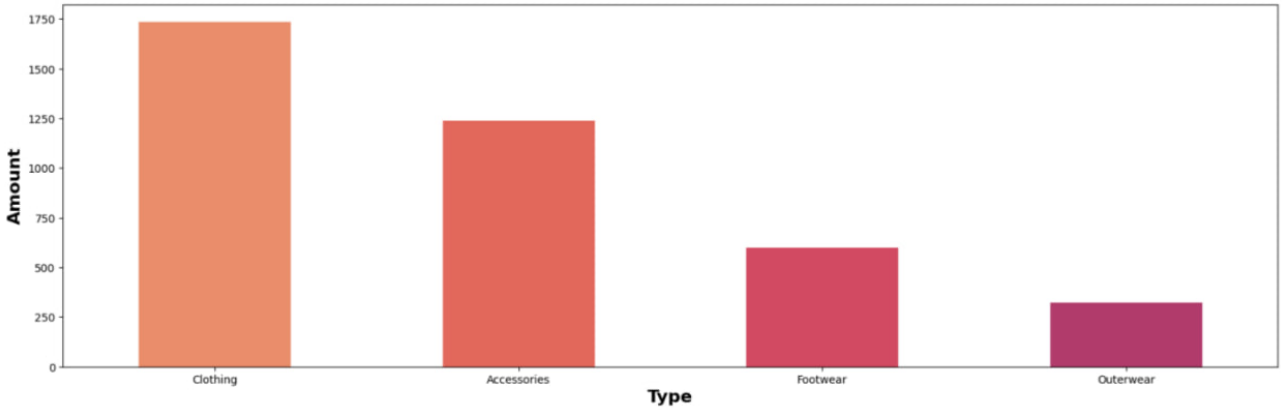
Нам повезло, все числовые признаки практически не коррелируют между собой. Воспользуемся еще одним распространенным способом визуализации данных – построением парных графиков для всех числовых признаков. Он демонстрирует связь между признаками. До этого мы убедились в отсутствии корреляции, но для наглядности решили также построить парные графики.

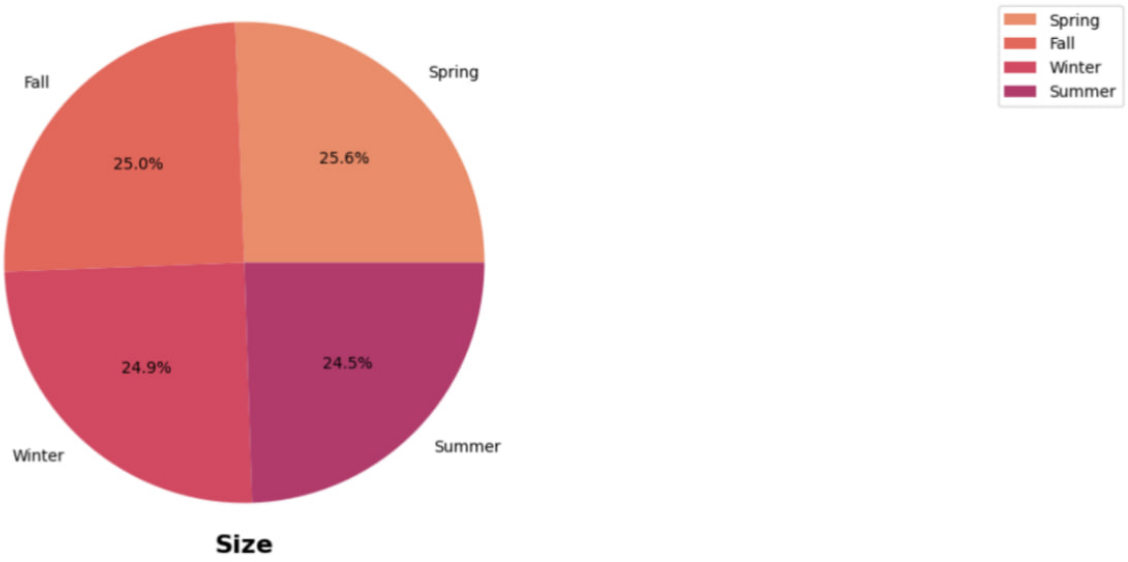


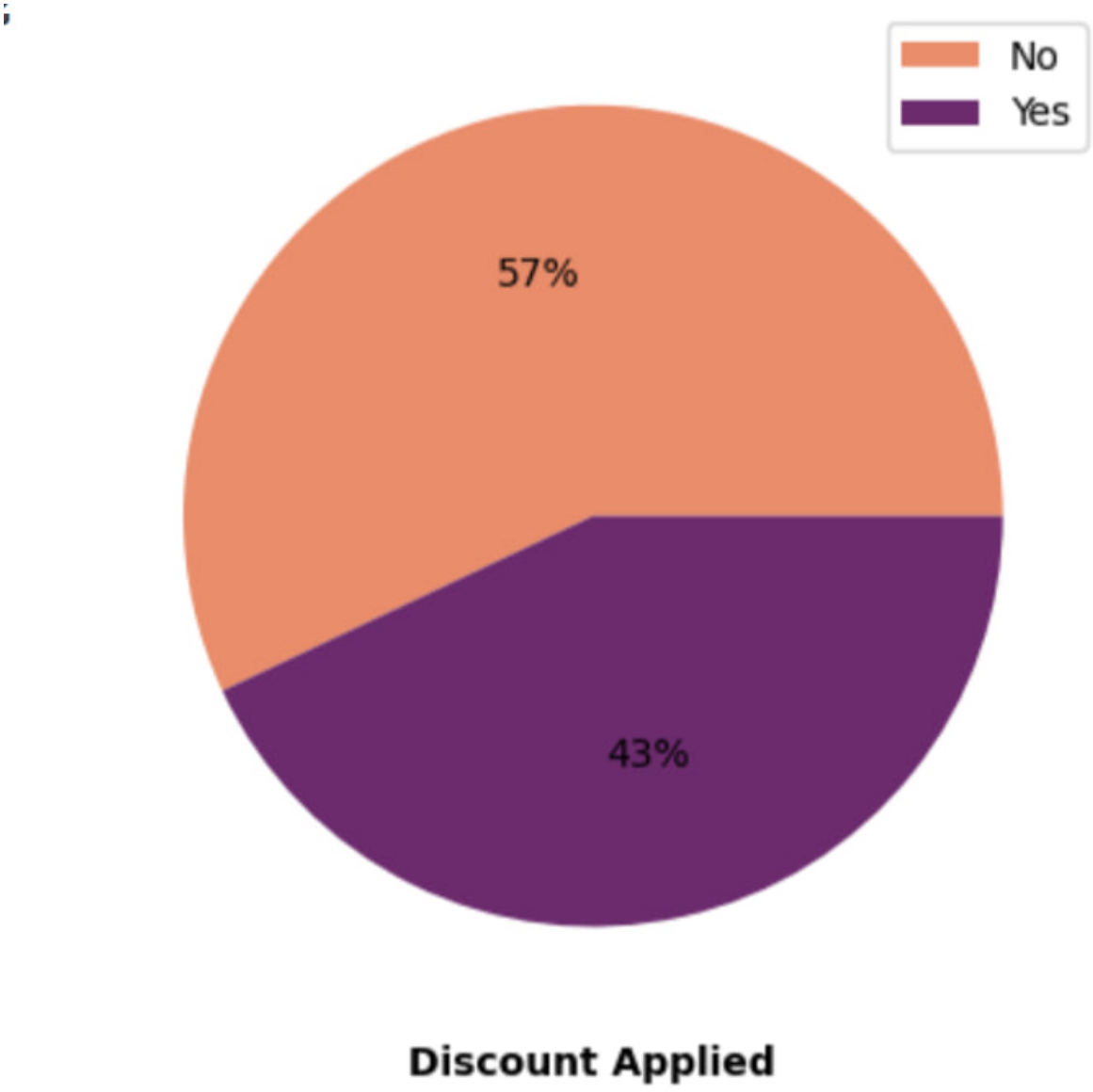
Как видно, никакой взаимосвязи между признаками не прослеживается.
Как вариант построения модели, мы рассматривали также логистическую регрессию для определения пола покупателя. Однако для этих целей соотношение между гендерами должно быть похоже на 50/50. Проверим, так ли это:



Такое соотношение показалось нам неудовлетворительным, поэтому мы отказались от этой идеи.
Далее мы построили несколько графиков с общей информацией о датасете для наглядности.







В конечном счете, мы подумали и решили делать линейную регрессию для определения возраста покупателя. Для этого нам пришлось избавиться от всех нечсиловых признаков, что очень плохо сказалось на качестве модели, однако для учебных целей мы посчитали такой вариант приемлемым. Разделив датасет на тренировочные и тестовые данные в соотношении

70/30, мы воспользовались библиотекой scikit-learn для построения уравнения линейной регрессии.

5 Результаты

В итоге мы получили модель линейной регрессии, однако на этих данных нельзя построить модель с высокой точностью, так как большинство признаков не являются числовыми. Однако, нам удалось построить графики, которые кластеризируют объекты по некоторым признакам.

Сделаем выводы из тех графиков, которые мы построили:

- Больше половины покупателей в нашем датасете (68%) - мужчины;
- Меньше всего покупают верхнюю одежду (объяснимо тем, что она необходима только в определенные сезоны);
- Больше всего покупают одежду весной и осенью;
- Больше половины покупателей применяют скидку (57%);

6 Заключение

В заключение, данная работа позволила нам ознакомиться с моделями анализа данных, работой с базами данных и построением графиков в Python. Это исследование пригодится нам в дальнейшем изучении анализа данных, когда мы сможем создавать более сложные модели.

Помимо этого, мы научились работать в Github, а именно загружать репозиторий и совершать коммиты.

Также, для описания нашего исследования, мы воспользовались сервисом Overleaf. Таким образом, мы научились работать в LaTex.