****

**Universitatea Tehnică a Moldovei**

**Анализ данных.**

**Научная статья на тему  
«Предсказание пола покупателя по характеристикам товаров и транзакции»**

**Student: Casian Andrei**

**Coordonator: Munteanu Viorel**

**Chişinău, 2023**

СОДЕРЖАНИЕ

[АННОТАЦИЯ 3](#_Toc153836882)

[ВВЕДЕНИЕ 4](#_Toc153836883)

[1. ПРЕДСТАВЛЕНИЕ ДАННЫХ 5](#_Toc153836884)

[2. ВИЗУАЛИЗАЦИЯ ДАННЫХ 6](#_Toc153836885)

[3. ПРЕДСТАВЛЕНИЕ МОДЕЛИ 11](#_Toc153836886)

[ВЫВОДЫ 12](#_Toc153836887)

[БИБЛИОГРАФИЯ 13](#_Toc153836888)

АННОТАЦИЯ

Цель работы заключается в определении пола человека по характеристикам его покупок: итоговая сумма, цвет, тип, размер одежды или аксессуара. Полученные результаты могут помочь в определении потенциального потребителя или создании конкретных товаров для них.

Для решения поставленной задачи нужно визуализировать существующие данные и построить модель, прогнозирующую результат на их основе.

ВВЕДЕНИЕ

Современные технологии позволяют автоматизировать большинство вещей, недавно выполняемых людьми вручную. Маркетинг не стал исключением – сбор данных покупателей и анализ рынка может проводиться без участия человека, достаточно лишь правильно настроить систему и алгоритмы.

Целью работы является предсказывание пола потенциального покупателя для вещей и аксессуаров по их характеристикам, а также по информации самой транзакции.

Тема становится более актуальной из-за отхода людей от традиционного мужского и женского – все чаще можно увидеть людей в вещах, до недавнего времени считавшихся сугубо для конкретного пола.

Для начала в статье рассматривается набор данных, описываются его поля. Следующим шагом данные визуализируются и делаются выводы на основе полученных графиков. Создается модель, предсказывающая пол покупателя, и анализируются результаты.

1. ПРЕДСТАВЛЕНИЕ ДАННЫХ

Данные были взяты с сайта [www.kaggle.com](http://www.kaggle.com/).

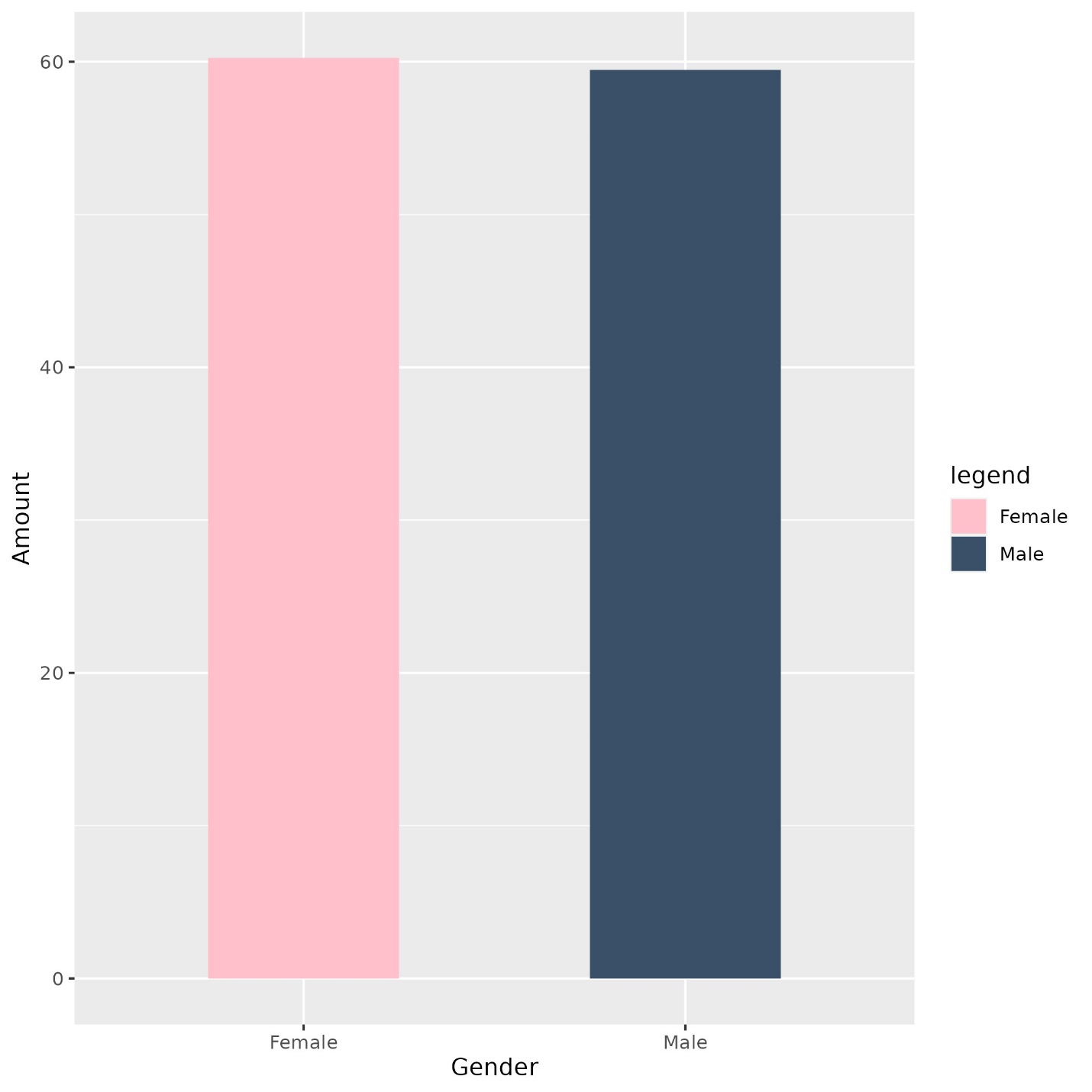
В наборе данных содержатся несколько полей, не использующихся в визуализации и классификации, так что они в описании не нуждаются.

Поля, использованные в работе:

* Age – возраст покупателя
* Item – вид одежды или аксессуара (hat, skirt, socks, pants)
* Category – тип, категория купленной вещи (clothing, accessories, footwear)
* Amount – сумма покупки в долларах США
* Size – размер вещи
* Color – цвет вещи
* Season – сезон вещи
* Shippng Type – тип доставки
* Discount Applied – куплена ли вещь по скидке
* Promo Code Used – был ли использован промокод
* Previous Purchases – сумма предыдущей покупки в долларах США
* Payment Method – тип оплаты
* Purchases – частота покупок (weekly, monthly, every 3 months, annually)
* Gender - пол

1. ВИЗУАЛИЗАЦИЯ ДАННЫХ

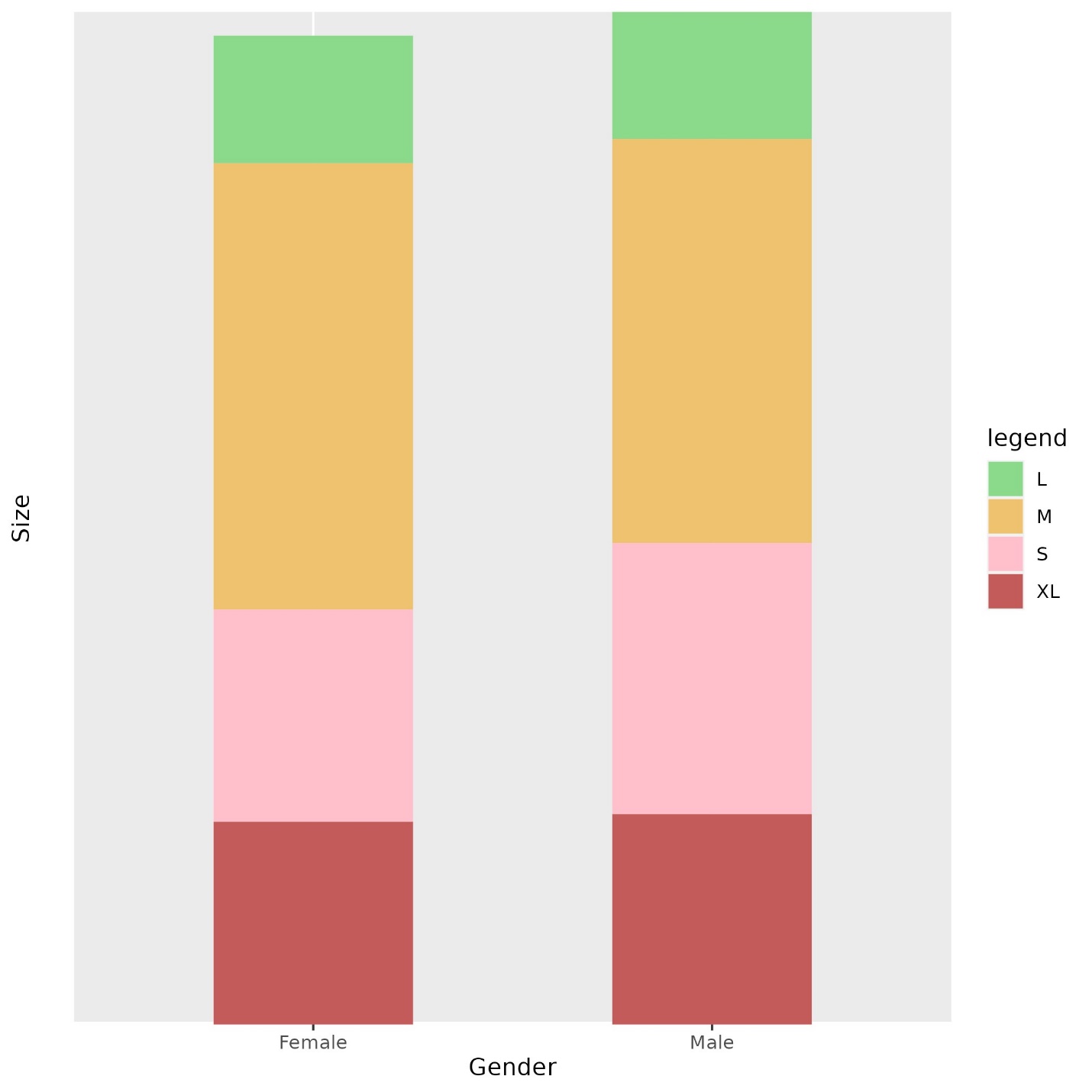
В первую очередь хочется узнать, кто в среднем тратит больше денег на поход в магазин.



**Рисунок 1. Средняя сумма покупки (USD)**

Вывод: в среднем девушки тратят на покупки одежды и аксессуаров больше, чем мужчины.

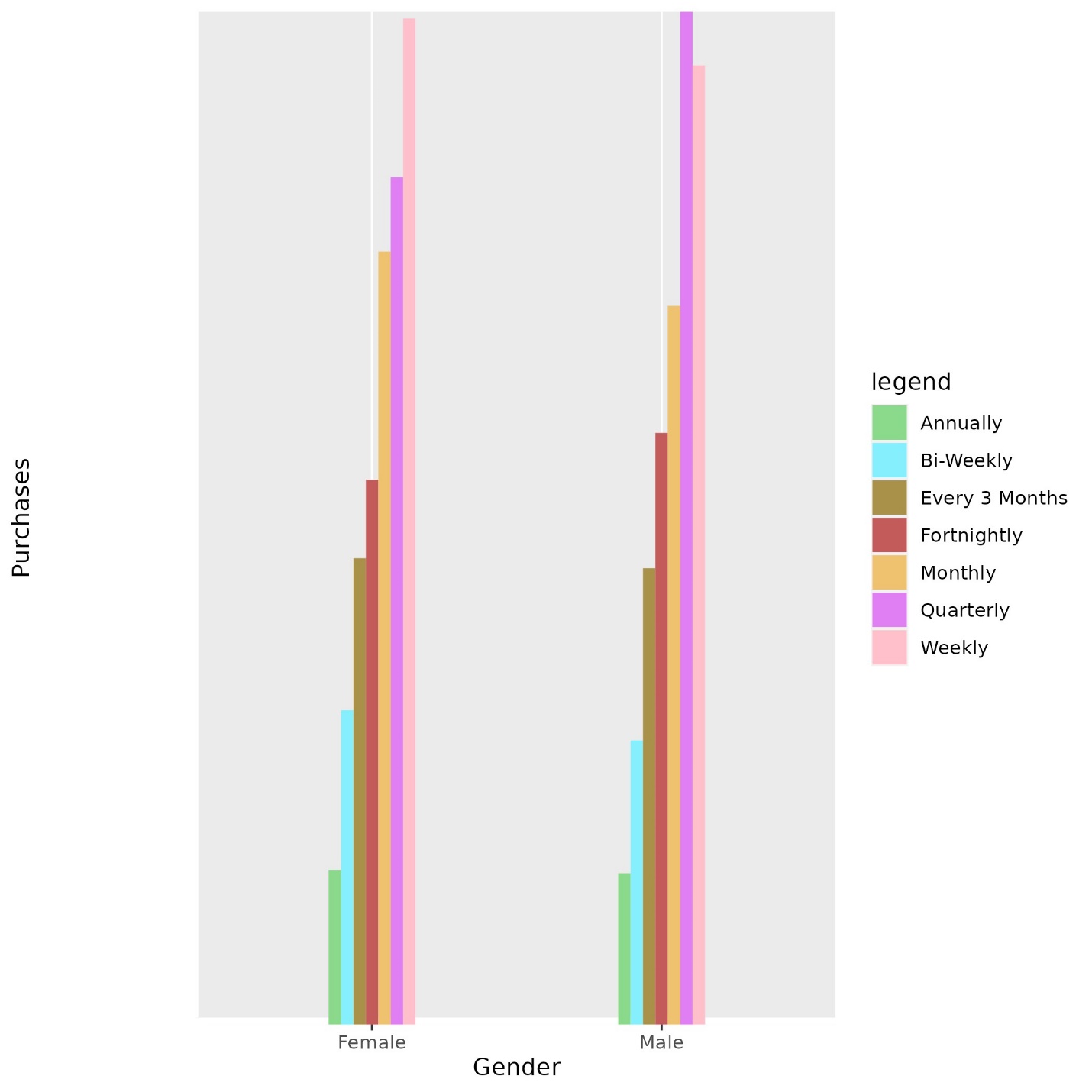
Также важно знать распределение размеров.



**Рисунок 2. Распределение размеров**

Вывод: преобладание M размера у женщин и S у мужчин.

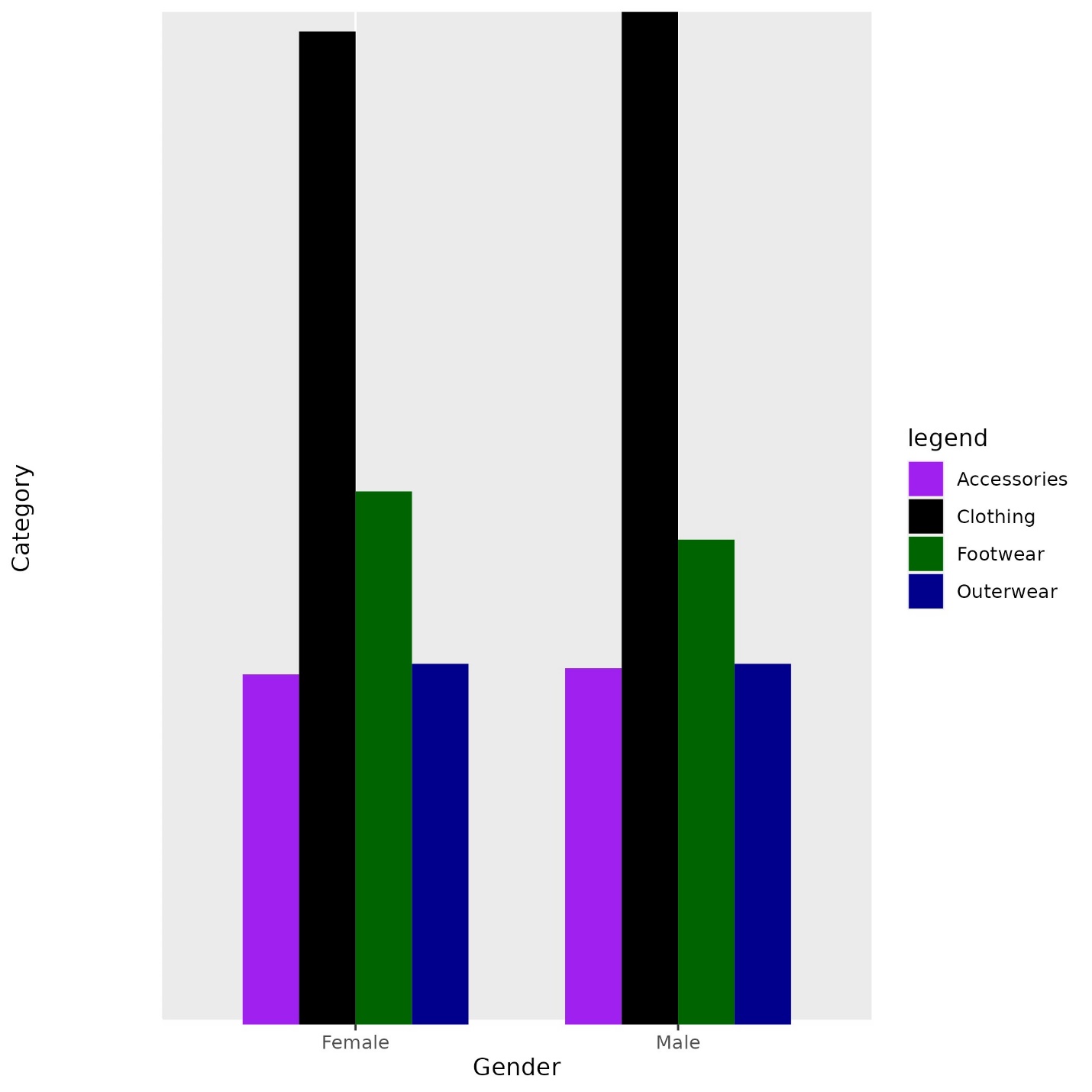
Кроме суммы, потраченной за один поход в магазин, важно знать как часто покупатель будет возвращаться.



**Рисунок 3. Частота покупок**

Вывод: девушки покупают одежду и аксессуары чаще, чем мужчины. У женщин преобладают еженедельные покупки, а у мужчин – раз в квартал.

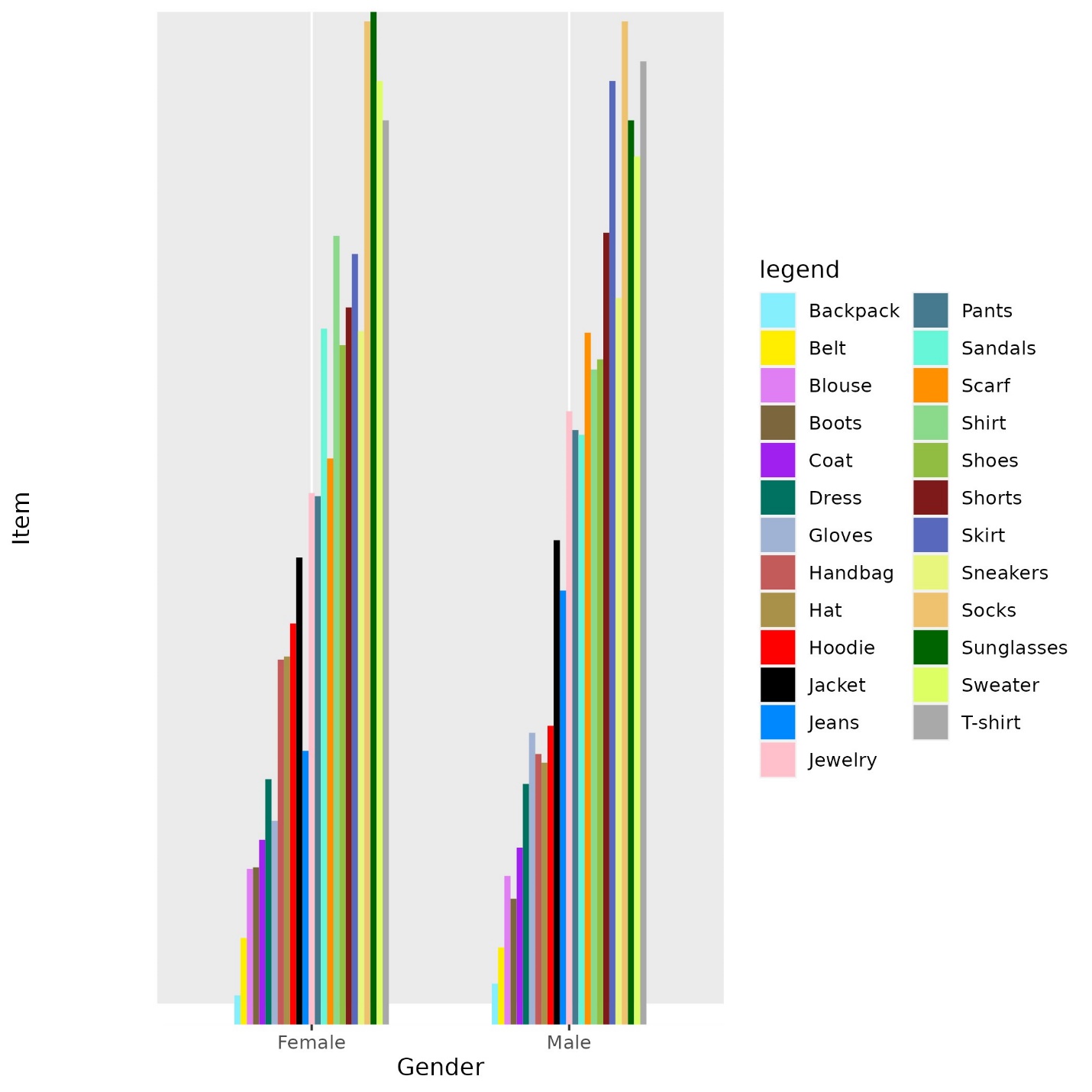
Также важно знать, какой тип товара интересен конкретному полу.



**Рисунок 4. Распределение по типу товара**

Вывод: девушки чаще покупают обувь, а мужчины – одежду.

Кроме типа одежды можно посмотреть, какие именно вещи популярны у каждого пола.



**Рисунок 5. Распределение вещей**

Вывод: у девушек наиболее частый товар – солнечные очки и носки, а у мужчин – носки и футболки.

1. ПРЕДСТАВЛЕНИЕ МОДЕЛИ

Для создания модели использовался язык программирования R. Данные в датасете были урезаны для выравнивания данных по мужчинам и женщинам.

Сначала переменные были переделаны в факториальные значения  
**names\_to\_factorize <- c(3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 16, 17)**

**data[, names\_to\_factorize] <- lapply(data[, names\_to\_factorize], factor)**

Удалены ненужные колонки.

**data <- data[, -c(1)]**

Следующим шагом стало разделение данных на тренировочные и тестовые в соотношении 60 на 40 процентов.

**split <- sample.split(data$Gender, SplitRatio = 0.6)**

**train <- subset(data, split == TRUE)**

**test <- subset(data, split == FALSE)**

С помощью функции glm тренируется модель.

**model\_glm <- glm(Gender ~ ., family = "binomial", data = train)**

Создание матрицы ошибок.

**predict\_test <- predict(model\_glm, newdata = test, type = "response")**

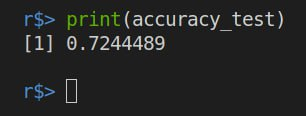
**confusion\_matrix\_test <- table(test$Gender, predict\_test >= 0.5)**

Подсчет точности:

**accuracy\_test <- sum(diag(confusion\_matrix\_test)) / nrow(test)**

**print(accuracy\_test)**

Результат:



**Рисунок 6. Результат работы модели**

ВЫВОДЫ

Проделанная работа дала результат – графики показывают определенные тенденции, а построенная модель с точностью 72 процента выдает верный результат.

Визуализацию можно использовать на практике для фокуса товаров под определенных клиентов для увеличения прибыли, например, расширив ассортимент женских моделей солнечных очков или наоборот потратить средства на продвижение непопулярных товаров.

Увеличить процент точности можно с большим количеством данных. На данный момент разброс среди товаров и людей недостаточно велик, чтобы находить ярко выраженные тенденции.

БИБЛИОГРАФИЯ

1. <https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset>
2. <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm>
3. <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>
4. <https://www.datacamp.com/tutorial/generalized-linear-models>