МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Дальневосточный федеральный университет»
(ДВФУ)

ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ
Департамент информационных и компьютерных систем

Курс «Компьютерные методы анализа больших данных»

Лабораторная работа №1
Контрольное мероприятие по рейтингу

на тему «Построение и визуализация датасета rand5»
Вариант №15

Выполнил студент Б9122-01.03.02мкт
Пелагеев Д.И.
_____

Проверил доцент Достовалов В.Н.

г. Владивосток
2024

Оглавление

Введение 2
1 Подготовка данных 3
1.1 Загрузка и первичная обработка данных 3
1.2 Описание данных 3
1.3 Формирование подмножеств 4
2 Расчёт дисперсий и средних значений 5
2.1 Создание таблицы с результатами 5
3 Визуализация данных 6
3.1 Составные гистограммы с распределением веса 6
3.2 Составные гистограммы с распределением цены 6
Заключение 8
Список использованных источников 9

Введение

Данный отчет посвящен процессу предварительного анализа данных из набора «Diamonds» в формате CSV. Нас интересуют следующие параметры:

— вес алмаза (в каратах);
— качество огранки;
— цвет алмаза;
— цена (в долларах США).

Актуальность данной работы обусловлена тем, что она служит хорошим материалом для изучения основных функций анализа данных в R.

Цель исследования — изучить возможности языка R для анализа данных и создать визуализации, которые позволят наглядно рассмотреть распределение характеристик алмазов.

Задачи работы:

а) Прочитать файл с помощью команды: `file <- read.csv(file="rand5.csv header=TRUE, sep= ")`.

б) Для двух наихудших классов `color` и `cut` построить `subset` и определить значения средней цены, дисперсию цены, значения среднего веса, дисперсию веса алмазов (таблица). Упорядоченность классов определяется функцией `unique()`.

в) Построить составные гистограммы с распределением веса, цены алмаза при фиксированных классах `color` и `cut` (см. п.2).

Основная часть

1 Подготовка данных

1.1 Загрузка и первичная обработка данных

Прежде чем мы начнем работать с данными, стоит сначала установить рабочую директорию, в которой будет находиться наш датасет “rand5.csv“.

Листинг 1 — Установка директории

```
setwd("/Users/daniil/Desktop/Education/BigData")
```

Теперь загружаем наш датасет “rand5.csv“ с помощью функции `read.csv`.

Листинг 2 — Загрузка данных

```
file ← read.csv(file="rand5.csv", header=TRUE, sep=",")
```

Проверяем размерность наших данных до очистки, затем очищаем данные и проверяем ещё раз, чтобы убедиться, были ли пустые данные.

Листинг 3 — Очистка данных

```
cat(dim(diamond), "\n")

diamond ← na.omit(diamond)

cat(dim(diamond), "\n")
```

1.2 Описание данных

Изучаем, какие данные находятся в файле.

Листинг 4 — Проверка данных

```
head(diamonds)
```

Датасет diamonds представляет собой набор данных, содержащий информацию о различных алмазах, собранную для анализа их характеристик и цен. Он включает 53 940 записей и 10 переменных, каждая из

которых описывает определенные свойства алмаза[1]. У нас модифицированный датасет, который имеет 5000 записей и 6 переменных:

— **X**: идентификатор строки;

— **carat**: масса алмаза в каратах;

— **cut**: качество огранки алмаза (Fair, Good, Very Good, Premium, Ideal);

— **color**: цвет алмаза, обозначаемый буквами от D (бесцветный) до J (с легким желтым оттенком);

— **clarity**: чистота алмаза, описываемая различными категориями (например, I1, SI1, SI2, VS1, VS2, VVS1, VVS2, IF);

— **price**: цена алмаза в долларах США.

Таблица 1 — Пример данных алмазов

X	carat	cut	color	clarity	price
1	1.20	Ideal	D	SI2	6140
2	0.37	Ideal	G	IF	1056
3	0.80	Ideal	F	VS2	3913
4	1.07	Ideal	H	SI1	4955
5	0.52	Very Good	F	VS2	1581
6	1.01	Ideal	E	SI2	4666

1.3 Формирование подмножеств

После всех манипуляций с датасетом мы должны отобрать два наихудших класса `color` и `cut`. С помощью `unique` мы упорядочим наши классы и внесём два последних элемента, то есть два наихудших элемента, в новые переменные для дальнейшей работы.

Листинг 5 — Наихудшие цвета и огранки

```
1 sort(unique(diamonds$color))
2
3 worst_colors ← sort(unique(diamonds$color))[1:2]
4
5
```

```
6 sort(unique(diamonds$cut))
7
8 worst_cuts ← sort(unique(diamonds$cut))[1:2]
```

Теперь, зная наихудшие цвета и огранки, мы можем выделить подвыборку на их основе.

Листинг 6 — Подмножество

```
1     subset ← diamond[(diamond$cut %in% worst_cuts) &
        (diamond$color %in% worst_colors), ]
```

2 Расчёт дисперсий и средних значений

Среднее значение — это сумма всех значений в выборке или совокупности, делённая на их количество.

Дисперсия — это статистическая мера, которая показывает степень разброса значений в выборке или генеральной совокупности относительно их среднего значения.

2.1 Создание таблицы с результатами

Чтобы найти нужные величины, воспользуемся встроенными методами в R [2] [3]. Для поиска дисперсий воспользуемся функцией `var`, а для поиска среднего значения — `mean`.

Листинг 7 — Создание таблицы

```
1     summary ← data.frame(
2       Mean_price = mean(subset$price),
3       Var_price = var(subset$price),
4       Mean_weight = mean(subset$carat),
5       Var_weight = var(subset$carat)
6     )
```

И выведем это всё с помощью следующей строки:

Листинг 8 — Вывод

```
1     summary
```

Получим следующую таблицу:

Таблица 2 — Пример данных алмазов

Mean_price	Var_price	Mean_weight	Var_weight
3099.337	9558683	0.726413	0.1957291

3 Визуализация данных

Визуализацию модифицированного набора diamonds будем производить с помощью библиотеки ggplot2 [4], которая позволяет создавать наглядные графики на языке R.

3.1 Составные гистограммы с распределением веса

Строим гистограмму распределения веса алмазов при фиксированных классах color и cut, равных D, E и GOOD, Fair соответственно.

Листинг 9 — Код для первой гистограммы

```
ggplot(subset, aes(x = carat, fill = interaction(color, cut))) +
  geom_histogram(binwidth = 0.1, position = "stack") +
  labs(title = "Distribution of Diamond Carat by Color and Cut",
       x = "Carat",
       y = "Count") +
  scale_fill_discrete(name = "Color and Cut") +
  theme_minimal()
```

3.2 Составные гистограммы с распределением цены

Строим гистограмму распределения цены алмазов при фиксированных классах color и cut, равных D, E и GOOD, Fair соответственно.

Листинг 10 — Код для второй гистограммы

```
ggplot(subset, aes(x = price, fill = interaction(color, cut))) +
  geom_histogram(binwidth = 500, position = "stack") +
  labs(title = "Distribution of Diamond Price by Color and Cut",
       x = "Price",
       y = "Count") +
  scale_fill_discrete(name = "Color and Cut") +
  theme_minimal()
```

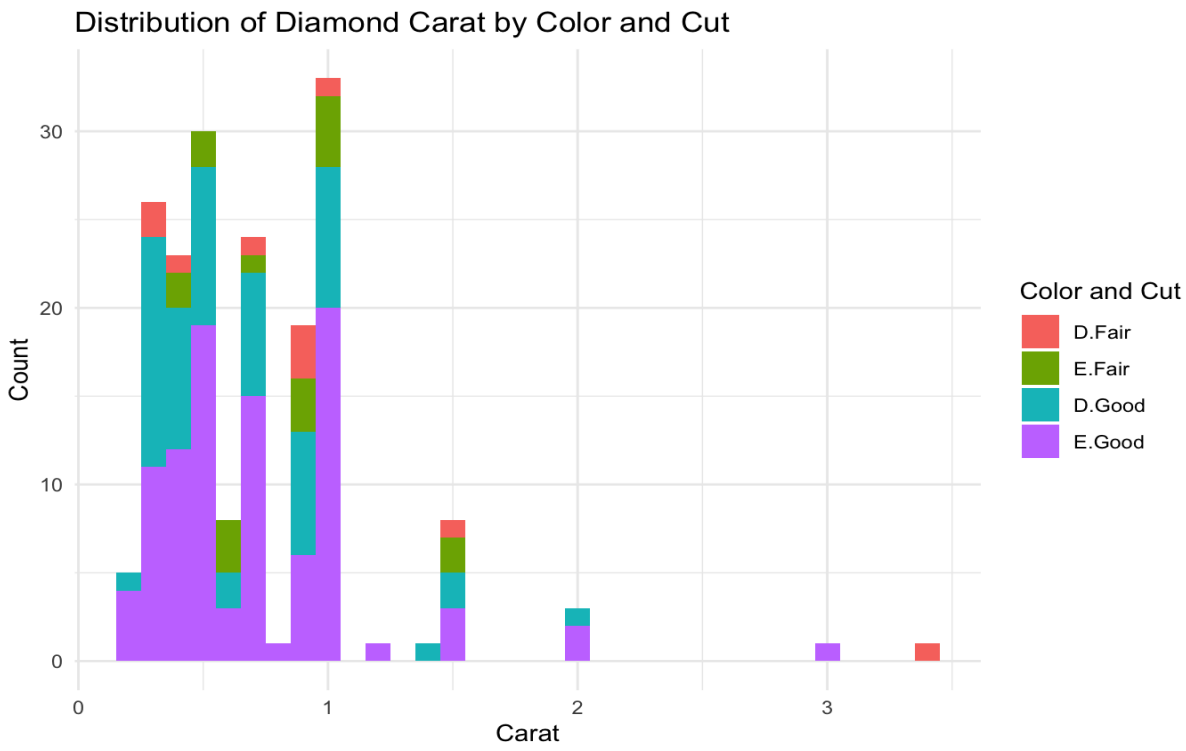

Рисунок 1 — Гистограмма распределения веса алмазов при фиксированных классах color и cut

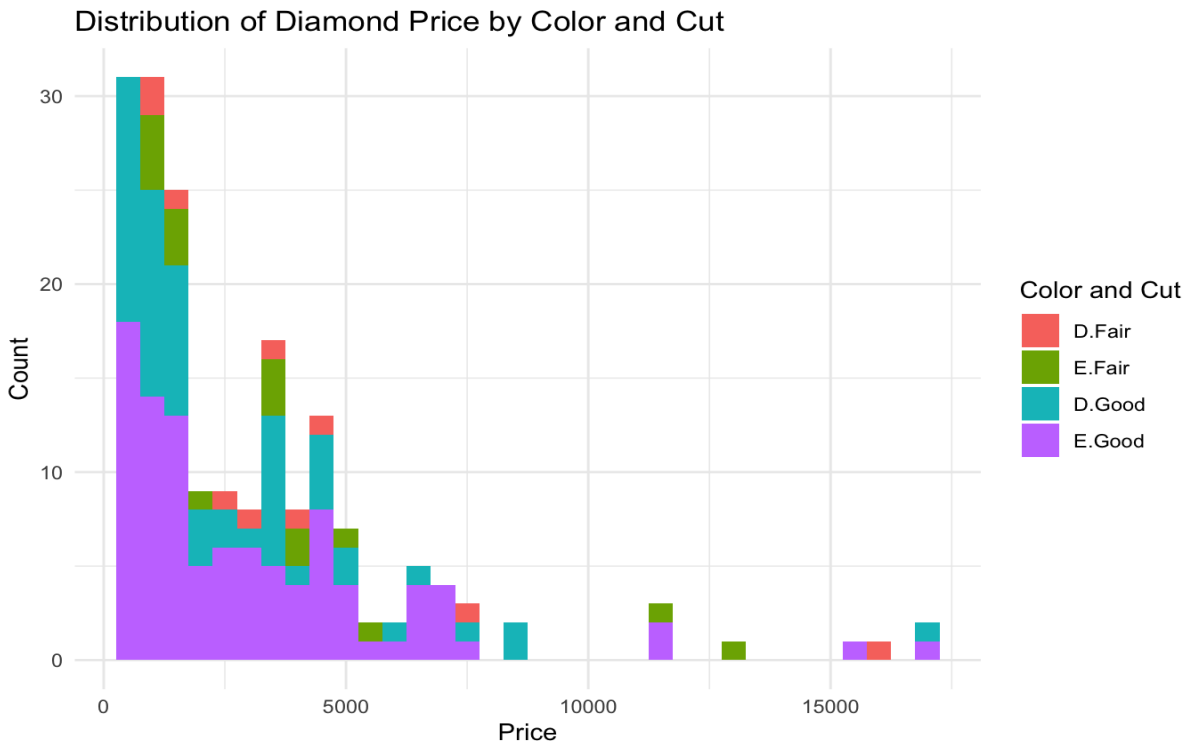

Рисунок 2 — Гистограмма распределения цены алмазов при фиксированных классах color и cut

Заключение

В данном отчёте была проведена предварительная обработка и анализ данных из набора «Diamonds».

Визуализация данных с использованием библиотеки `ggplot2` позволила проанализировать распределения веса и цены алмазов в зависимости от их характеристик.

Проведённый анализ демонстрирует возможности языка R и его библиотек для проведения предварительной обработки данных, вычисления ключевых статистических показателей и создания информативных визуализаций.

Список использованных источников

1. *R Foundation for Statistical Computing*. Diamonds dataset. — 2023. — Режим доступа: Diamonds dataset (дата обращения: 3.11.2024).

2. Официальная страница среды статистического моделирования R. — 2024. — Режим доступа: R: The R Project for Statistical Computing (дата обращения: 3.11.2024).

3. Официальная страница интегрированной среды разработки RStudio. — 2024. — Режим доступа: RStudio | Open source & professional software for data science teams - RStudio дата обращения: 3.11.2024).

4. *Wickham, Hadley*. ggplot2: Elegant Graphics for Data Analysis (Use R!) / Hadley Wickham. — New York: Springer, 2009. — Режим доступа: ggplot2 (дата обращения: 3.11.2024).