Project 6 : EDA Visualization of the Diamond Dataset Using R and R Markdown

Phubordin Phanyosri

2025-07

Contents

คำแนะนำ :

ถ้าคุณดูเอกสารนี้บน Github ให้กด Ctrl+F เพื่อไปยังหัวข้อที่สนใจ (ทีบอกมาจาก My Portfolio Website)

Load Library

```
library(knitr) # ใช้สำหรับรันโค้ด R ที่ฝังในเอกสาร Markdown / LaTeX
library(tidyverse) # แพ็คเกจที่รวบรวมเครื่องใช้จัดการข้อมูล และนำเสนอข้อมูล
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Explore Data

Data Glimpse

```
diamonds |> glimpse() # โครงสร้างตาราง diamonds กี่แถว กี่คอลัมน์ ประเภทคอลัมน์ แต่ละคอลัมน์หน้าตาเป็นยัง
```

```
## Rows: 53,940
## Columns: 10
## $ carat   <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22, 0.23, 0.~
## $ cut     <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very Good, Ver~
## $ color   <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, J, J, I,~
## $ clarity <ord> SI2, SI1, VS1, VS2, SI2, VVS2, VVS1, SI1, VS2, VS1, SI1, VS1, ~
## $ depth   <dbl> 61.5, 59.8, 56.9, 62.4, 63.3, 62.8, 62.3, 61.9, 65.1, 59.4, 64~
## $ table   <dbl> 55, 61, 65, 58, 58, 57, 57, 55, 61, 61, 55, 56, 61, 54, 62, 58~
## $ price   <int> 326, 326, 327, 334, 335, 336, 336, 337, 337, 338, 339, 340, 34~
## $ x       <dbl> 3.95, 3.89, 4.05, 4.20, 4.34, 3.94, 3.95, 4.07, 3.87, 4.00, 4.~
## $ y       <dbl> 3.98, 3.84, 4.07, 4.23, 4.35, 3.96, 3.98, 4.11, 3.78, 4.05, 4.~
## $ z       <dbl> 2.43, 2.31, 2.31, 2.63, 2.75, 2.48, 2.47, 2.53, 2.49, 2.39, 2.~
```

Preview Data

ตาราง diamonds เป็นชุดข้อมูลเกี่ยวกับเพชรที่มีคอลัมน์หลากหลาย พร้อมข้อมูลคุณสมบัติที ใช้กำหนดราคาเพชร มาดูแต่ละคอลัมน์และความหมายของมัน:

Columns	Description	Type
carat	น้ำหนักของเพชร มีหน่วยเป็น "กะรัต" (carats) หน่วยวัดน้ำหนักเพชร	Numeric
cut	ระดับคุณภาพการเจียระไน (cut quality) เช่น Fair, Good, Very Good, Premium, Ideal (เรียงจากต่ำไปสูง)	Ordered Factor
color	สีของเพชร (D ถึง J โดยที D คือใสที่สุด)	Ordered Factor
clarity	ระดับความใสของเพชร โดยดูจากตำหนิหรือจุดบกพร่อง (I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF เรียงจากตำหนิชัดเจนถึงไม่มีเลย)	Ordered Factor
depth	อัตราส่วนระหว่างความลึกของเพชรกับเส้นผ่านศูนย์กลางเฉลี่ย มีหน่วยเป็นเปอร์เซ็นต์ (%): (z / mean(x, y)) * 100	Numeric
table	ความกว้างของโต๊ะเพชร (ส่วนเรียบด้านบนของเพชร) เทียบกับเส้นผ่านศูนย์กลางเฉลี่ย มีหน่วยเป็นเปอร์เซ็นต์ (%)	Numeric
price	ราคาของเพชร มีหน่วยเป็นดอลลาร์สหรัฐ	Integer
x	ความยาว (length) ของเพชรในหน่วยมิลลิเมตร	Numeric
y	ความกว้าง (width) ของเพชรในหน่วยมิลลิเมตร	Numeric
z	ความลึก (depth) ของเพชรในหน่วยมิลลิเมตร	Numeric

- ความหมาย: depth : (z / mean(x, y)) * 100
  - ค่านี้แสดงถึง **ความสมดุลของรูปร่างเพชร**:
    - **ค่า depth ต่ำเกินไป**: เพชรอาจแบนเกินไป
    - **ค่า depth สูงเกินไป**: เพชรอาจลึกหรือหนาเกินไป
  - ค่าที่เหมาะสมสำหรับ **เพชรทรงกลม (round cut)** มักอยู่ที่ประมาณ **59-62%** เพื่อให้เพชรมีประกายดีที่สุด
- เสริมข้อ 4

  จาก chart ที่แสดง boxplot ความสัมพันธ์ระหว่าง **depth** (ความลึกของเพชร) และ **cut** (คุณภาพการเจียระไนของเพชร) เราสามารถตั้งชื่อ labs ดังนี้:

  - **Title**
    - "Depth Distribution Across Diamond Cut Quality"
    - (แสดงการกระจายตัวของความลึกในแต่ละระดับคุณภาพการเจียระไน)
  - **Subtitle**:
    - "Comparing Depth Values for 'Fair', 'Good', 'Very Good', 'Premium', and 'Ideal' Cuts"
    - (เปรียบเทียบค่าความลึกในแต่ละคุณภาพการเจียระไน)
  - **Caption**:
    - "Source: ggplot2 Diamonds Dataset"
    - (ระบุแหล่งข้อมูลจาก dataset)
  - **X**:
    - "Cut Quality"
    - (แสดงคุณภาพการเจียระไน)
  - **Y**:
    - "Depth Percentage (%)"
    - (ระบุสัดส่วนความลึกของเพชร)

```
diamonds |> head() # ดู 6 แถวแรก(ไม่รวมหัวตาราง)
```

```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

```
diamonds |> tail() # ดู 6 แถวท้าย(ไม่รวมหัวตาราง)
```

```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.72 Premium   D     SI1      62.7    59  2757  5.69  5.73  3.58
## 2  0.72 Ideal     D     SI1      60.8    57  2757  5.75  5.76  3.5
## 3  0.72 Good      D     SI1      63.1    55  2757  5.69  5.75  3.61
## 4  0.7  Very Good D     SI1      62.8    60  2757  5.66  5.68  3.56
## 5  0.86 Premium   H     SI2      61      58  2757  6.15  6.12  3.74
## 6  0.75 Ideal     D     SI2      62.2    55  2757  5.83  5.87  3.64
```