STAT394 Group Project Milestone 3: Exploratory Data Analysis

Ken MacIver, Tom Tribe, Jundi Yang, Mei Huang

2022-08-21

Contents

1 EDA summmary

```
diamonds <- read.csv("./diamonds.csv")
df <- data.frame(diamonds)
df$cut <- factor(df$cut)
df$clarity <- factor((df$clarity))
df$color <- factor(df$color)

summary(df)
```

```
##      carat               cut        color        clarity          depth
##  Min.   :0.2000   Fair      : 1610   D: 6775   SI1    :13065   Min.   :43.00
##  1st Qu.:0.4000   Good      : 4906   E: 9797   VS2    :12258   1st Qu.:61.00
##  Median :0.7000   Ideal     :21551   F: 9542   SI2    : 9194   Median :61.80
##  Mean   :0.7979   Premium   :13791   G:11292   VS1    : 8171   Mean   :61.75
##  3rd Qu.:1.0400   Very Good:12082   H: 8304   VVS2   : 5066   3rd Qu.:62.50
##  Max.   :5.0100                     I: 5422   VVS1   : 3655   Max.   :79.00
##                                     J: 2808   (Other): 2531
##      table           price             x                y
##  Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720
##  Median :57.00   Median : 2401   Median : 5.700   Median : 5.710
##  Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735
##  3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540
##  Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900
##
```

```
##        z
##  Min.   : 0.000
##  1st Qu.: 2.910
##  Median : 3.530
##  Mean   : 3.539
##  3rd Qu.: 4.040
##  Max.   :31.800
##
```

```
MySummary <- function(x){
  return(c(
    length(x),
    min(x),
    quantile(x, .25),
    median(x),
    quantile(x, .75),
    max(x),
    skewness(x),
    kurtosis(x))
  )
}

#pply(df[,-(2:4)], MySummary, MARGIN=2) -> MySummary1

#xtable(MySummary1)
```

	carat	depth	table	price	x	y	z
sample size	53940	53940	53940	53940	53940	53940	53940
minimum	0.20	43.00	43.00	326.00	0.00	0.00	0.00
first quartile	0.40	61.00	56.00	950.00	4.71	4.72	2.91
median	0.70	61.80	57.00	2401.00	5.70	5.71	3.53
third quartile	1.04	62.50	59.00	5324.25	6.54	6.54	4.04
maximum	5.01	79.00	95.00	18823.00	10.74	58.90	31.80
skewness	1.12	-0.08	0.80	1.62	0.38	2.43	1.52
kurtosis	4.26	8.74	5.80	5.18	2.38	94.21	50.08