DataSciencePG STAT7022

Assignment01 Question1: Data Cleaning

Possakorn Kittipipatthanapong

June 16, 2023

1 Question 1: Data Cleaning

1.1 Read diamonds dataset

```
diamonds_df <- readRDS("C:/Users/possa/OneDrive - University of Adelaide/script_courses/c_20
head(diamonds_df) %>%
  kable(caption = "Display the first 6 rows present in diamonds dataset", booktabs = T) %>%
  kable_styling(latex_options = "scale_down", full_width = TRUE)
```

Table 1: Display the first 6 rows present in diamonds dataset

carat	c.grade	depth	table	price	x	y	volume
0.23	4B1	61.5	55	326	3.95	3.98	38.20203
0.21	3B2	59.8	61	326	3.89	3.84	34.50586
0.23	1B4	56.9	65	327	4.05	4.07	38.07688
0.31	1G1	63.3	58	335	4.34	4.35	51.91725
0.24	2G5	62.8	57	336	3.94	3.96	38.69395
0.24	2F6	62.3	57	336	3.95	3.98	38.83087

1.2 Split the data

Therefore, whether to divide the data before or after cleaning will rely on the type of cleaning you intend to perform. If the cleaning involves rectifying errors that are consistent throughout the entire data set, it may be preferable to clean the data first and then split it. However, even if the decision is to clean the data first, it is important to note that any transformations that rely on the data distribution, such as scaling, centering, or encoding based on frequency, should still be applied after splitting to prevent any influence on the test data. These transformations should be carried out separately on the training and test sets. Finally, in this case, I will split after cleaning the data.

1.3 Summarise the columns

After reviewing the data summary, we could perform the step following detail below:
- split column: c.grade to cut, colour, and clarity.
- change the column as proper type.
- review all attributes - Qualitative variables (to check with any unusual, unexpected values, or type).
- review all attributes - Quantitative variables (to check with any unusual, unexpected values, or type) then drop null.
- remove outliers if it's significant using IQR.

```
show_diamonds_df <- diamonds_df %>%
  skim_without_charts() %>%
  yank("character") %>%
  mutate(across(where(is.numeric), ~ round(., digits = 2))) %>%
  kable(caption = "Statistical summary for character variables", booktabs = T) %>%
  kable_styling(latex_options = "scale_down", full_width = TRUE, font_size = 8)
show_diamonds_df
```

Table 2: Statistical summary for character variables

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
carat	674	0.99	1	5	0	332	0
c.grade	446	0.99	2	5	0	654	0
price	80	1.00	2	5	0	11597	0

```
show_diamonds_df_2 <- diamonds_df %>%
  skim_without_charts() %>%
  yank("numeric") %>%
  mutate(across(where(is.numeric), ~ round(., digits = 2))) %>%
  kable(caption = "Statistical summary for numeric variables", booktabs = T) %>%
  kable_styling(latex_options = "scale_down", full_width = TRUE, font_size = 8)
show_diamonds_df_2
```

Table 3: Statistical summary for numeric variables

skim_variablen_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
depth 47	1	61.75	1.43	43.00	61.00	61.80	62.50	79.00
table 73	1	57.46	2.23	43.00	56.00	57.00	59.00	95.00
x 54	1	5.73	1.12	0.00	4.71	5.70	6.54	10.74
y 0	1	5.73	1.15	-7.38	4.72	5.71	6.54	58.90
volume 146	1	129.84	78.22	0.00	65.16	114.81	170.84	3840.60

1.3.1 split column: c.grade to cut, colour, and clarity

```
diamonds_df_clean_1 <- diamonds_df %>%
  mutate(
    cut = str_split_i(c.grade, "[A-Za-z]+", 1),
    colour = str_extract(c.grade, "[A-Za-z]+"),
    clarity = str_split_i(c.grade, "[A-Za-z]+", -1)
    ) %>%
  select(-c.grade)

diamonds_df_clean_1 %>%
  skim_without_charts() %>%
  yank("character") %>%
  mutate(across(where(is.numeric), ~ round(., digits = 2))) %>%
  kable(caption = "Statistical summary for character variables", booktabs = T) %>%
  kable_styling(latex_options = "scale_down", full_width = TRUE, font_size = 8)
```

Table 4: Statistical summary for character variables

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
carat	674	0.99	1	5	0	332	0
price	80	1.00	2	5	0	11597	0
cut	446	0.99	1	2	0	11	0
colour	446	0.99	1	3	0	42	0
clarity	446	0.99	0	1	74	8	0

1.3.2 Reviewing column type

- carat, price, depth, table, x, y, volume : change to numeric because of currency
- cut: following the reference with limited collections, we should change to factor (ordering later)
- colour: following the reference with limited collections, we should change to factor but later
- clarity: following the reference with limited collections, we should change to factor (ordering later)

```
diamonds_df_clean_2 <- diamonds_df_clean_1 %>%
  mutate(
    carat = as.numeric(carat),
    price = as.numeric(price),
    cut = as.factor(cut),
    clarity = as.factor(clarity),
    depth = as.numeric(depth),
    table = as.numeric(table),
    x = as.numeric(x),
    y = as.numeric(y),
    volume = as.numeric(volume)
    )
```

```
diamonds_df_clean_2 %>%
  skim_without_charts() %>%
  yank(c("factor")) %>%
  # mutate(across(where(is.numeric), ~ round(., digits = 2))) %>%
  kable(caption = "Statistical summary for qualitative variables", booktabs = T) %>%
  kable_styling(latex_options = "scale_down", full_width = TRUE, font_size = 8)
```

Table 5: Statistical summary for qualitative variables

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
cut	446	0.9917316	FALSE	11	4: 21305, 3: 13630, 2: 11924, 1: 4842
clarity	446	0.9917316	FALSE	8	2: 12942, 3: 12141, 1: 9114, 4: 8089

```
diamonds_df_clean_2 %>%
  skim_without_charts() %>%
  yank("numeric") %>%
```

```
  mutate(across(where(is.numeric), ~ round(., digits = 2))) %>%
  kable(caption = "Statistical summary for quantitative variables", booktabs = T) %>%
  kable_styling(latex_options = "scale_down", full_width = TRUE, font_size = 8)
```

Table 6: Statistical summary for quantitative variables

skim_variablen_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
carat 675	0.99	0.79	0.48	-2.50	0.40	0.70	1.04	5.01
depth 47	1.00	61.75	1.43	43.00	61.00	61.80	62.50	79.00
table 73	1.00	57.46	2.23	43.00	56.00	57.00	59.00	95.00
price 83	1.00	3932.34	3989.23	-1.00	949.00	2401.00	5324.00	18823.00
x 54	1.00	5.73	1.12	0.00	4.71	5.70	6.54	10.74
y 0	1.00	5.73	1.15	-7.38	4.72	5.71	6.54	58.90
volume 146	1.00	129.84	78.22	0.00	65.16	114.81	170.84	3840.60

1.3.3 review all attributes: qualitative varaibles

1.3.3.1 review all attributes: cut

- For double alphabet, we will group to the single alphabet
- Finally, apply factor with order following the meaning in reference

```
# Review the data first
diamonds_df_clean_2 %>%
  count(cut) %>%
  kable(caption = "Display the summary of the total number for each cut before cleaning", bo
  kable_styling(latex_options = "scale_down", full_width = TRUE, font_size = 8)
```

Table 7: Display the summary of the total number for each cut before cleaning

cut	n
0	1586
00	10
1	4842
11	10
2	11924
22	25
3	13630
33	34
4	21305
44	48
5	80
NA	446

```
# mutate column
diamonds_df_clean_3 <- diamonds_df_clean_2 %>%
  mutate(cut = substr(cut,1,1),
         cut = case_when(
           cut == '0' ~ 'Fair',
           cut == '1' ~ 'Good',
           cut == '2' ~ 'Very Good',
```

```
            cut == '3' ~ 'Premium',
            cut == '4' ~ 'Ideal',
            .default = NULL
            ),
          cut = ordered(cut, levels = c('Fair', 'Good', 'Very Good', 'Premium', 'Ideal'))
          )

# Review the data after manipulated
diamonds_df_clean_3 %>%
  count(cut) %>%
  kable(caption = "Display the summary of the total number for each cut after cleaning", boo
  kable_styling(latex_options = "scale_down", full_width = TRUE, font_size = 8)
```

Table 8: Display the summary of the total number for each cut after cleaning

cut	n
Fair	1596
Good	4852
Very Good	11949
Premium	13664
Ideal	21353
NA	526

1.3.3.2 review all attributes: colour

- For multi-alphabet, if it have single unique alphabet, I will apply to single digit alphabet. Others will be removed.
- Change type to factor

```
# Review the data first
diamonds_df_clean_3 %>%
  count(colour) %>%
  head(15) %>%
  kable(caption = "Display top 15 of the summary of the total number for each colour before
  kable_styling(latex_options = "scale_down", full_width = TRUE, font_size = 8)
```

Table 9: Display top 15 of the summary of the total number for each colour before cleaning

colour	n
A	6642
AA	63
AAA	19
B	9591
BA	1
BB	94
BBB	38
BBE	1
BE	1
C	9314
CA	1
CB	1
CC	108
CCC	21
CE	1

```
# mutate column
diamonds_df_clean_3 <- diamonds_df_clean_3 %>%
  mutate(
    colour = case_when(
      substr(colour,1,1) == substr(colour,nchar(colour),nchar(colour)) ~ substr(colour,1,1),
      .default = NULL
      ),
    colour = as.factor(colour)
  )

# Review the data after manipulated
diamonds_df_clean_3 %>%
  count(colour) %>%
  kable(caption = "Display the summary of the total number for each colour after cleaning",
  kable_styling(latex_options = "scale_down", full_width = TRUE, font_size = 8)
```

Table 10: Display the summary of the total number for each colour after cleaning

colour	n
A	6724
B	9723
C	9443
D	11178
E	8239
F	5378
G	2788
NA	467

1.3.4 review all attributes: clarity.

- The scope of clarity starting from 0 to 7, so make unrelated to Null
- Match with refers to clarity in reference
- Change type to factor with order

```
# Review the data first
diamonds_df_clean_3 %>%
  count(clarity) %>%
  kable(caption = "Display the summary of the total number for each clarity before cleaning'
  kable_styling(latex_options = "scale_down", full_width = TRUE, font_size = 8)
```

Table 11: Display the summary of the total number for each clarity before cleaning

clarity	n
	74
1	9114
2	12942
3	12141
4	8089
5	5015
6	3616
7	2503
NA	446

```
# mutate column
diamonds_df_clean_3 <- diamonds_df_clean_3 %>%
  mutate(
    clarity = case_when(
      clarity == '1' ~ 'SI2',
      clarity == '2' ~ 'SI1',
      clarity == '3' ~ 'VS2',
      clarity == '4' ~ 'VS1',
      clarity == '5' ~ 'VVS2',
      clarity == '6' ~ 'VVS1',
      clarity == '7' ~ 'IF',
      .default = NULL
      ),
    clarity = ordered(clarity, levels = c('SI2', 'SI1', 'VS2', 'VS1', 'VVS2', 'VVS1', 'IF'))
    )

# Review the data after manipulated
diamonds_df_clean_3 %>%
  count(clarity) %>%
  kable(caption = "Display the summary of the total number for each clarity after cleaning",
  kable_styling(latex_options = "scale_down", full_width = TRUE, font_size = 8)
```

Table 12: Display the summary of the total number for each clarity after cleaning

clarity	n
SI2	9114
SI1	12942
VS2	12141
VS1	8089
VVS2	5015
VVS1	3616
IF	2503
NA	520

1.3.5 review all attributes: quantitative varaibles

```
diamonds_df_clean_3 %>%
  skim_without_charts() %>%
  yank("numeric") %>%
  mutate(across(where(is.numeric), ~ round(., digits = 2))) %>%
  kable(caption = "Statistical summary for numeric variables", booktabs = T) %>%
  kable_styling(latex_options = "scale_down", full_width = TRUE, font_size = 8)
```

Table 13: Statistical summary for numeric variables

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
carat	675	0.99	0.79	0.48	-2.50	0.40	0.70	1.04	5.01
depth	47	1.00	61.75	1.43	43.00	61.00	61.80	62.50	79.00
table	73	1.00	57.46	2.23	43.00	56.00	57.00	59.00	95.00
price	83	1.00	3932.34	3989.23	-1.00	949.00	2401.00	5324.00	18823.00
x	54	1.00	5.73	1.12	0.00	4.71	5.70	6.54	10.74
y	0	1.00	5.73	1.15	-7.38	4.72	5.71	6.54	58.90
volume	146	1.00	129.84	78.22	0.00	65.16	114.81	170.84	3840.60

```
diamonds_df_clean_3 %>%
  stack() %>%
  ggplot(aes(x = values)) +
  geom_boxplot() +
  facet_wrap(vars(ind), scales = "free")
```

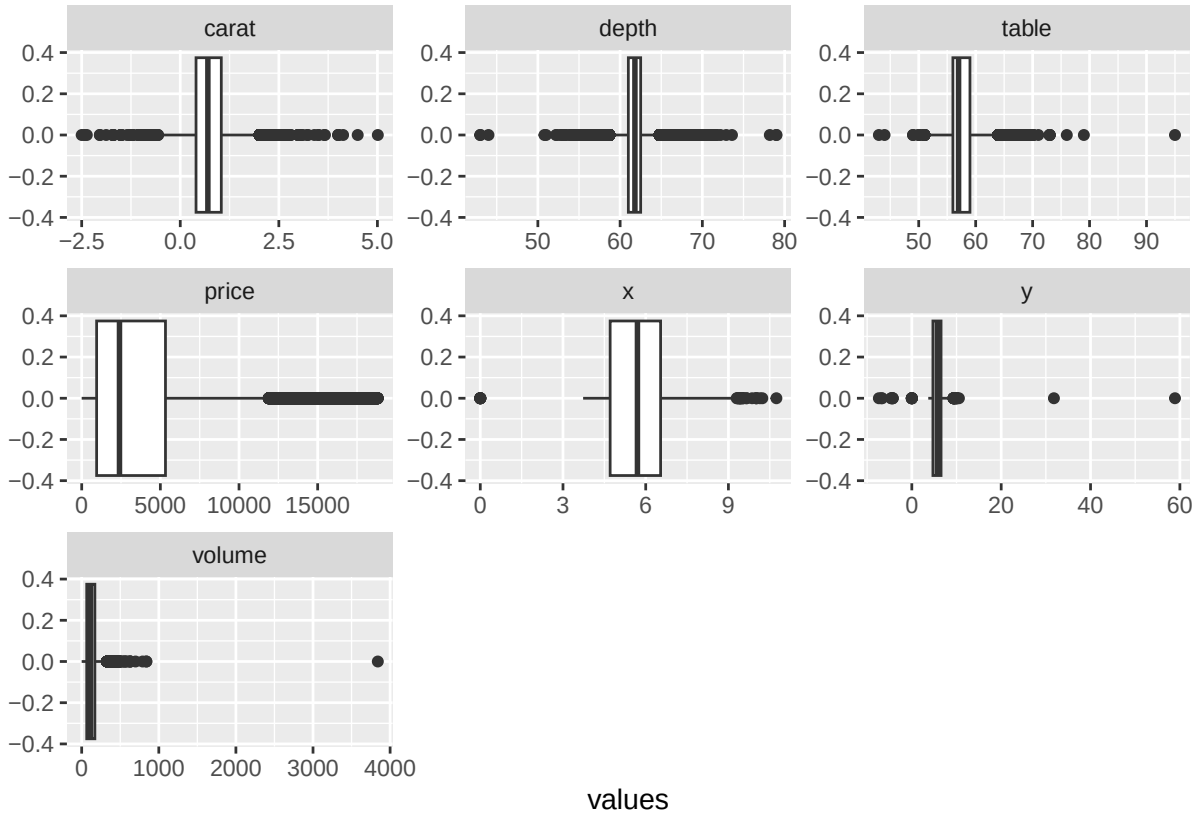

Figure 1: Summary of the distribution on each quantitative variables by Box plot

Following the summary table and boxplot for each column:
- For carat, price, they should be only logical positive values. Further, depth and table need to be positive number following the criteria in reference.
- For x, y, and volume, they also should be positive and non-zero values.
- Filtering null value from previous processing

```
diamonds_df_clean_4 <- diamonds_df_clean_3 %>%
  filter(
    (carat >= 0) & (price >= 0) & (depth >= 0) & (table >= 0) &
      (x > 0) & (y > 0) & (volume > 0)
  ) %>%
  na.omit()
diamonds_df_clean_4 %>%
  skim_without_charts() %>%
  yank("numeric") %>%
  mutate(across(where(is.numeric), ~ round(., digits = 2))) %>%
  kable(caption = "Statistical summary for numeric variables aftering filtering the abnormal
  kable_styling(latex_options = "scale_down", full_width = TRUE, font_size = 8)
```

Table 14: Statistical summary for numeric variables aftering filtering the abnormal cases

skim_variablen_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
carat 0	1	0.80	0.47	0.20	0.40	0.70	1.04	5.01
depth 0	1	61.75	1.43	43.00	61.00	61.80	62.50	79.00
table 0	1	57.46	2.24	43.00	56.00	57.00	59.00	95.00
price 0	1	3931.90	3988.42	326.00	949.00	2401.00	5323.00	18823.00
x 0	1	5.73	1.12	3.73	4.71	5.70	6.54	10.74
y 0	1	5.74	1.14	3.68	4.72	5.71	6.54	58.90
volume 0	1	129.94	78.31	31.71	65.20	114.86	170.85	3840.60

1.3.6 Remove outliers if it's significant using IQR

```
outliers <- function(x) {
  Q1 <- quantile(x, probs=.25)
  Q3 <- quantile(x, probs=.75)
  iqr = Q3-Q1

  upper_limit = Q3 + (iqr*1.5)
  lower_limit = Q1 - (iqr*1.5)
  x > upper_limit | x < lower_limit
}

remove_outliers <- function(df, cols = names(df)) {
  for (col in cols) {
    df <- df[!outliers(df[[col]]),]
  }
  df
}
```

1.3.6.1 remove_outlier function

```
# set the numeric column to remove outlier
col_numeric <- colnames(diamonds_df_clean_4[,sapply(diamonds_df_clean_4,is.numeric)])

diamonds_df_clean_final <- remove_outliers(diamonds_df_clean_4, col_numeric)

diamonds_df_clean_final %>%
  stack() %>%
  ggplot(aes(x = values)) +
  geom_boxplot() +
  facet_wrap(vars(ind), scales = "free")
```

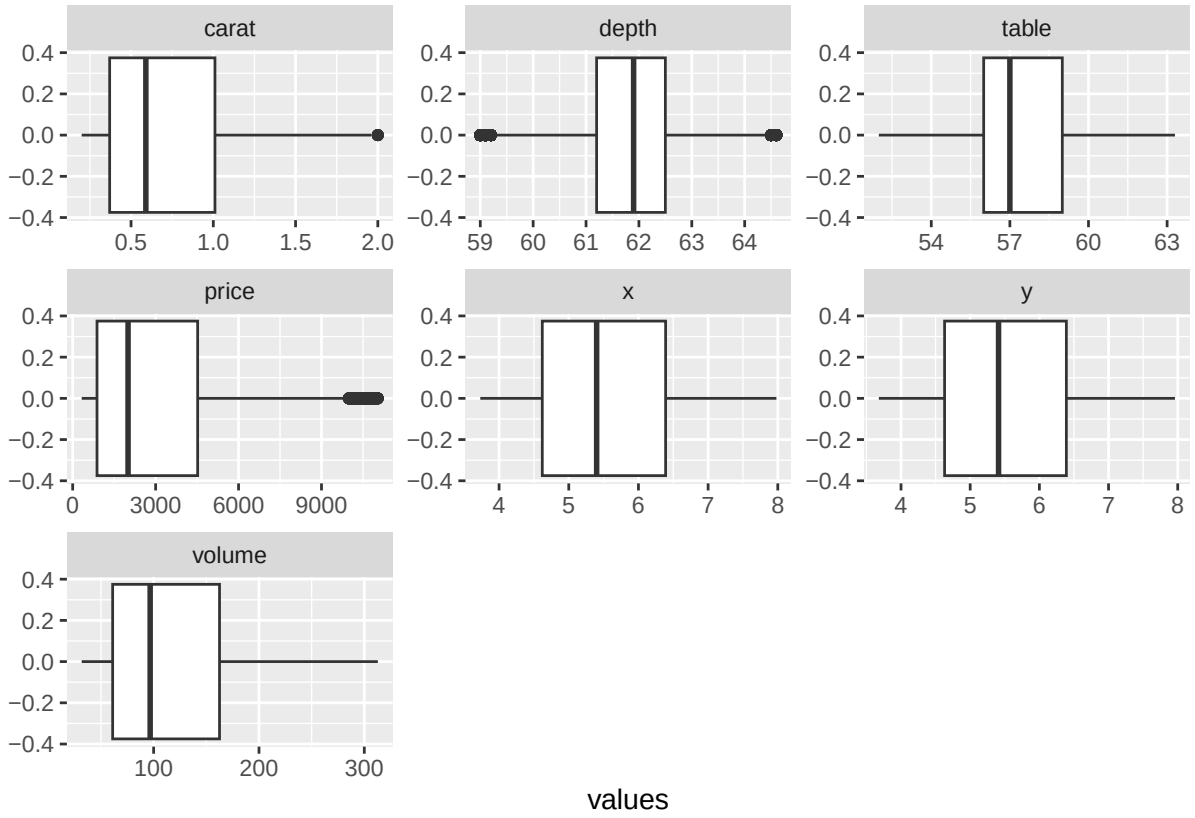

Figure 2: Summary of the distribution on each quantitative variables by Box plot after removing the outlier

1.3.6.2 Show the the distribution after removing the outlier

1.3.7 Comparison the statistic information before and after data cleaning process

```
diamonds_df %>%
  skim_without_charts() %>%
  yank("character") %>%
  mutate(across(where(is.numeric), ~ round(., digits = 2))) %>%
  kable(caption = "Statistical summary for qualitative variables of initial diamonds dataset
  kable_styling(latex_options = "scale_down", full_width = TRUE, font_size = 8)
```

Table 15: Statistical summary for qualitative variables of initial diamonds dataset

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
carat	674	0.99	1	5	0	332	0
c.grade	446	0.99	2	5	0	654	0
price	80	1.00	2	5	0	11597	0

```
diamonds_df_clean_final %>%
  skim_without_charts() %>%
  yank("factor") %>%
  mutate(across(where(is.numeric), ~ round(., digits = 2))) %>%
  kable(caption = "Statistical summary for qualitative variables of cleaned diamonds dataset
  kable_styling(latex_options = "scale_down", full_width = TRUE, font_size = 8)
```

Table 16: Statistical summary for qualitative variables of cleaned diamonds dataset

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
cut	0	1	TRUE	5	Ide: 19451, Pre: 11392, Ver: 10397, Goo: 3550
colour	0	1	FALSE	7	D: 9524, B: 8460, C: 8024, E: 6780
clarity	0	1	TRUE	7	SI1: 10940, VS2: 10379, SI2: 7022, VS1: 7014

1.3.7.1 Comparison the statistic - Qualitative variables

```
show_diamonds_df_2
```

Table 17: Statistical summary for numeric variables

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
depth	47	1	61.75	1.43	43.00	61.00	61.80	62.50	79.00
table	73	1	57.46	2.23	43.00	56.00	57.00	59.00	95.00
x	54	1	5.73	1.12	0.00	4.71	5.70	6.54	10.74
y	0	1	5.73	1.15	-7.38	4.72	5.71	6.54	58.90
volume	146	1	129.84	78.22	0.00	65.16	114.81	170.84	3840.60

```
diamonds_df_clean_final %>%
  skim_without_charts() %>%
  yank("numeric") %>%
  mutate(across(where(is.numeric), ~ round(., digits = 2))) %>%
  kable(caption = "Statistical summary for qualitative variables of cleaned diamonds dataset
  kable_styling(latex_options = "scale_down", full_width = TRUE, font_size = 8)
```

Table 18: Statistical summary for qualitative variables of cleaned diamonds dataset

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
carat	0	1	0.70	0.36	0.20	0.37	0.59	1.01	2.00
depth	0	1	61.80	1.08	59.00	61.20	61.90	62.50	64.60
table	0	1	57.24	2.02	52.00	56.00	57.00	59.00	63.30
price	0	1	2996.89	2595.58	326.00	880.00	2003.00	4521.00	11037.00
x	0	1	5.52	0.96	3.73	4.62	5.40	6.39	7.98
y	0	1	5.53	0.96	3.68	4.63	5.41	6.39	7.96
volume	0	1	113.70	58.31	31.71	61.11	96.71	162.55	312.82

1.3.7.2 Comparison the statistic - Qualitative variables

```
diamonds_df %>%
  skim_without_charts() %>%
  summary() %>%
  print()
```

1.3.7.3 Comparison the statistic summary

```
## -- Data Summary ------------------------
##                            Values
## Name                       Piped data
## Number of rows             53940
## Number of columns          8
## _________________________
## Column type frequency:
##   character                3
##   numeric                  5
## _________________________
## Group variables            None
```

```
diamonds_df_clean_final %>%
  skim() %>%
  summary() %>%
  print()
```

```
## -- Data Summary ------------------------
##                            Values
## Name                       Piped data
## Number of rows             45052
## Number of columns          10
## _________________________
## Column type frequency:
##   factor                   3
##   numeric                  7
## _________________________
## Group variables            None
```

After removed the outlier out, there represent the significant reduction of amount of information from 53,490 rows to 45,052 around 15.8%.

1.4 Export the clean data

```
filenamepath <- glue::glue("{lubridate::today()}_diamonds.rds")
write_rds(diamonds_df_clean_final,filenamepath)
```