STATS 7022 - Data Science PG

Assignment 2

Possakorn A1873765

July 11, 2023

Read Dataset

Read the data set from providing sources and preview data summary using **skimr**.

```
diamonds_df <- read_rds('diamonds_clean.rds')

diamonds_df %>%
  skim_without_charts()
```

Table 1: Data summary

Name	Piped data
Number of rows	51513
Number of columns	8
_____	
Column type frequency:	
character	1
numeric	7
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
c.grade	0	1	3	3	0	245	0

Variable type: numeric

skim_variable	n_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	3934.72	3990.54	326.00	950.00	2402.00	5329.00	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	130.00	78.38	31.71	65.21	114.89	170.89	3840.60

```
head(diamonds_df) %>%
  knitr::kable(caption = "The first 6 rows of the dataset, diamonds_clean")
```

Table 4: The first 6 rows of the dataset, diamonds_clean

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

Data Pre-processing

Pre-processing step00: Overview data understanding

Following the c.grade, there are three related details including cut, color, and clarity. After check the column length, c.grade have only 3 characters following **Table05**.

```
diamonds_df %>%
  mutate(c.grade.length = str_length(c.grade)) %>%
  group_by(c.grade.length) %>%
  summarize(n()) %>%
  knitr::kable(caption = "The summarize c.grade column length, diamonds_clean")
```

Table 5: The summarize c.grade column length, diamonds_clean

c.grade.length	n()
3	51513

Then, we preview c.grade column to match with pre-processing method as shown as on **Table06**

```
diamonds_df %>%
  select(c.grade) %>%
  head() %>%
  knitr::kable(caption = "The first 6 rows of the dataset, diamonds_clean")
```

Table 6: The first 6 rows of the dataset, diamonds_clean

c.grade
4B1
3B2
1B4
1G1
2G5
2F6

Pre-processing step00: c.grade splitting columns

we start to split columns to expected feature using **substr** as shown in **Table07**.

```
diamonds_prep_df <- diamonds_df %>%
  mutate(
    cut = substr(c.grade, 1, 1), # cut = first character
    colour = substr(c.grade, 2, 2), # colour = second character
    clarity = substr(c.grade, 3, 3) # clarity = third character
    )

diamonds_prep_df %>%
  select(c.grade, cut, colour, clarity) %>%
  head() %>%
  knitr::kable(caption = "The first 6 rows of the dataset after splitting columns")
```

Table 7: The first 6 rows of the dataset after splitting columns

c.grade	cut	colour	clarity
4B1	4	B	1
3B2	3	B	2
1B4	1	B	4
1G1	1	G	1
2G5	2	G	5
2F6	2	F	6

Pre-processing step01: Derive cut from the first character of c.grade

preview data following Table08

```
diamonds_prep_df %>%
  group_by(cut) %>%
  summarize(n()) %>%
  knitr::kable(caption = "The summary amount of transactions by cut")
```

Table 8: The summary amount of transactions by cut

cut	n()
0	1536
1	4683
2	11527
3	13169
4	20598

preprocessing data

cut column is pre-processed by matching with naming and reordering to ordinal data.

```
diamonds_prep_df <- diamonds_prep_df %>%
  mutate(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'))
    )
```

Pre-processing step02: Derive colour from the first character of c.grade

preview data following Table09

```
diamonds_prep_df %>%
  group_by(colour) %>%
  summarize(n()) %>%
  knitr::kable(caption = "The summary amount of transactions by colour")
```

Table 9: The summary amount of transactions by colour

colour	n()
A	6492
B	9337
C	9090
D	10771
E	7950
F	5181
G	2692

preprocessing data

colour column is change the column type to factor.

```
diamonds_prep_df <- diamonds_prep_df %>%
  mutate(colour = as.factor(colour)) # change data type to factor for further analysis
```

Pre-processing step03: Derive clarity from the first character of c.grade

preview data following Table10

```
diamonds_prep_df %>%
  group_by(clarity) %>%
  summarize(n()) %>%
  knitr::kable(caption = "The summary amount of transactions by clarity")
```

Table 10: The summary amount of transactions by clarity

clarity	n()
1	8793
2	12508
3	11722
4	7780
5	4818
6	3480
7	2412

preprocessing data

clarity column is pre-processed by matching with naming and reordering to ordinal data.

```
diamonds_prep_df <- diamonds_prep_df %>%
  mutate(clarity = case_when(
    clarity == '0' ~ 'I2',
    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('I2','SI2', 'SI1', 'VS2', 'VS1', 'VVS2', 'VVS1', 'IF'))
    )
```

Pre-processing step04: Derive Z from volume, x, and y.

```
diamonds_prep_df <- diamonds_prep_df %>%
  mutate(
    z = volume / (x * y) # create new variable following potential formula
    )
```

Pre-processing step05: Once cut, colour, clarity, and z have been derived, drop c.grade and volume.

After preprocess step, there are new generated columns and also need to drop the old column to remove the redundant information.

```
diamonds_prep_df <- diamonds_prep_df %>%
  select(-c.grade, -volume)
```

summary dataset after preprocessing

Using **skimr** to summarise the statistical information based on pre-processed data set, the output derive new variables and drop initial column as shown as in **Table11** and show the first 6 rows to check output after changing in **Table14**.

```
diamonds_prep_df %>%
  skim_without_charts()
```

Table 11: Data summary

Name	Piped data
Number of rows	51513
Number of columns	10
Column type frequency:	
factor	3
numeric	7
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
cut	0	1	TRUE	5	ide: 20598, pre: 13169, ver: 11527, goo: 4683
colour	0	1	FALSE	7	D: 10771, B: 9337, C: 9090, E: 7950
clarity	0	1	TRUE	7	SI1: 12508, VS2: 11722, SI2: 8793, VS1: 7780

Variable type: numeric

skim_variable	n_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	3934.72	3990.54	326.00	950.00	2402.00	5329.00	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
z	0	1	3.54	0.70	1.07	2.91	3.53	4.04	31.80

```
head(diamonds_prep_df) %>%
  knitr::kable(caption = "The first 6 rows of the dataset after pre-processing")
```

Table 14: The first 6 rows of the dataset after pre-processing

carat	depth	table	price	x	y	cut	colour	clarity	z
0.23	61.5	55	326	3.95	3.98	ideal	B	SI2	2.43
0.21	59.8	61	326	3.89	3.84	premium	B	SI1	2.31
0.23	56.9	65	327	4.05	4.07	good	B	VS1	2.31
0.31	63.3	58	335	4.34	4.35	good	G	SI2	2.75
0.24	62.8	57	336	3.94	3.96	very good	G	VVS2	2.48
0.24	62.3	57	336	3.95	3.98	very good	F	VVS1	2.47

export dataset

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