Data Transformation - Prepare Data

Mobile Vendor Market Share Thailand 2018 - 2022

source : https://gs.statcounter.com/vendor-market-share/mobile

```
# download library for prepare data
library(tidyverse)
```

```
# change woking directory
setwd("/data/notebook_files/phone_th/")
```

```
# read csv file
df <-
  list.files(path = "/data/notebook_files/phone_th", pattern = "*.csv") %>%
  map_df(~read_csv(.))
head(df)
```

A tibble: 6 × 47

Date	Samsung	Apple	Oppo	Huawei	Mobicel	Unknown	Lava	Wiko	BBK	…	Other	Opp	Re
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	…	<dbl>	<dbl>	<(
2018-01	32.64	16.72	10.61	5.76	4.53	3.94	4.47	2.45	3.60	…	0.07	NA	N.
2018-02	32.59	16.30	10.67	6.10	4.98	4.11	4.40	2.45	3.51	…	0.07	NA	N.
2018-03	32.32	17.45	10.67	6.49	5.31	4.11	3.89	2.34	3.45	…	0.07	NA	N.
2018-04	31.72	19.36	11.39	6.85	5.54	3.21	3.09	2.62	3.35	…	0.08	NA	N.
2018-05	33.42	17.94	11.82	7.68	5.63	2.84	2.78	2.85	3.13	…	0.07	NA	N.
2018-06	32.56	17.29	12.68	8.08	6.08	2.89	2.73	3.00	3.07	…	0.07	NA	N.

```
Rows: 12 Columns: 38

── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr  (1): Date
```

```
dbl (37): Samsung, Apple, Oppo, Huawei, Mobicel, Unknown, Lava, Wiko, BBK,

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this mes

Rows: 12 Columns: 38

── Column specification ────────────────────────────────────────────────────
Delimiter: ","
chr  (1): Date
dbl (37): Samsung, Apple, Oppo, Huawei, Mobicel, Unknown, Wiko, Xiaomi, BBK
```

```
# change column name to lower
names(df) <- tolower(names(df))
head(df)
```

A tibble: 6 × 47

date	samsung	apple	oppo	huawei	mobicel	unknown	lava	wiko	bbk	…	other	opp	rea
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	…	<dbl>	<dbl>	<d
2018-01	32.64	16.72	10.61	5.76	4.53	3.94	4.47	2.45	3.60	…	0.07	NA	NA
2018-02	32.59	16.30	10.67	6.10	4.98	4.11	4.40	2.45	3.51	…	0.07	NA	NA
2018-03	32.32	17.45	10.67	6.49	5.31	4.11	3.89	2.34	3.45	…	0.07	NA	NA
2018-04	31.72	19.36	11.39	6.85	5.54	3.21	3.09	2.62	3.35	…	0.08	NA	NA
2018-05	33.42	17.94	11.82	7.68	5.63	2.84	2.78	2.85	3.13	…	0.07	NA	NA
2018-06	32.56	17.29	12.68	8.08	6.08	2.89	2.73	3.00	3.07	…	0.07	NA	NA

```
# change wide format(raw) to long format
df <- df %>%
  pivot_longer(-date,
               names_to = 'brand',
               values_to = 'market_share')
head(df)
```

A tibble: 6 × 3

date	brand	market_share
<chr>	<chr>	<dbl>
2018-01	samsung	32.64
2018-01	apple	16.72
2018-01	oppo	10.61
2018-01	huawei	5.76
2018-01	mobicel	4.53
2018-01	unknown	3.94

```
# check missing value
df %>%
  filter(!complete.cases(.)) %>%
  head(10)
```

A tibble: 10 × 3

date	brand	market_share
<chr>	<chr>	<dbl>
2018-01	opp	NA
2018-01	realme	NA
2018-01	tecno	NA
2018-01	vivo	NA
2018-01	honor	NA
2018-01	razer	NA
2018-01	infinex	NA
2018-01	itel	NA
2018-01	tcl	NA
2018-02	opp	NA

```
# if we delete all NA, it isn't affect any value. So I will delete all NA rows
df <- df %>%
  na.omit()
head(df)
```

A tibble: 6 × 3


```
# export csv file
write.csv(df, "phone_th_clean.csv", row.names = F)
```