Closing Prices of Australian Stock Exchange Shares

MATH2319 - Machine Learning

Course Project

Ben Cole - s3412349

Print Date: 25/04/2019

Contents

1 Phase 1 - Introduction, Cleaning, and Exploration

1.1 Outline

The aim of this supervised machine learning project is to predict the volume of shares sold of a large number of Australian Stock Exchange (ASX) shares in the year 2019. This phase covers the collection, cleaning, and inspection of the data. Data beginning at the 2019 calendar year through to April 2019 was sourced to use in the training and validation data set. Data will be sourced for dates after the last date in the training and validation data set for following phases of this project.

The dataset for the share prices was in a tidy and long format, with ASX ticker code, date, several price variables, and selling volume each in a separate column. A second data table was scraped from the internet that contained Global Industry Classification Standard industry groupings. This was joined to the first data set to add further categorical information.

The data was found to be heavily right-skewed for all price variables. The data was filtered to remove ASX tickers with extremely high prices, which focussed the dataset on shares that sold for under $1.00.

1.1.1 Nature of the Data

1.1.1.1 Pricing data

The data used was historical summary data of all shares available with a trading history in the ASX between 02/01/2019 through to business week (Mon - Fri) ending 12/04/2019. The data was provided by the website ASX Historical Data. The data was compressed into .zip files separated by calendar month between 02/01/2019 - 31/01/2019 and then by business week from 01/02/2019 - 12/04/2019. The raw data followed the same structure throughout all text files, and was not provided with headers. Each comma separated value followed the following headers:

- `Ticker` - the three-digit unique identifier ASX ticker code (renamed to `ASX_Ticker`)
- `Date` - date of trade information
- `Open` - price per individual share at the beginning of the day's trade
- `High` - highest price recorded per individual share during the day's trade
- `Low` - lowest price recorded per individual share during the day's trade
- `Close` - price per individual share at the end of the day's trade
- `Volume` - number of shares traded during the day

1.1.1.2 Global Industry Classification Standards Data

A second data table was scraped from the ASX website on GICS, which was spread across several pages. This contained the company name, ASX Ticker code, and GICS Industry group. Company name was not valuable to the model and discarded, wilst GICS industry group was retained. ASX Ticker code was used to join the two data frames.

1.1.2 Target Feature

The target feature selected was Volume, which is expressed only as positive integers; natural numbers.

1.1.3 Descriptive Features

Excepting `Date`[1], All other remaining variables in the data frame were used as descriptive features:

- `Ticker` - unique identifier, alphanumeric code
- `Open` - continuous positive double
- `High` - continuous positive double
- `Low` - continuous positive double
- `Close` - continuous positive double
- `Volume` - continuous positive integer
- `GICS_Industry_Group` - character factor variable

[1] Date was only retained as a means to further partition training/validation data and test data. It was not used as a descriptive feature.

1.2 Data Processing

1.2.1 Packages

The following packages were used, with brief descriptions of their uses as comments.

```
library(pacman)                          ## for loading multiple packages

suppressMessages(p_load(character.only = T,
                        install = F,
                        c("tidyverse",  ## thanks Hadley
                          "lubridate",  ## for handling dates
                          "forcats",    ## for categorial variables, not for felines
                          "zoo",        ## some data cleaning capabilities
                          "lemon",      ## add ons for ggplot
                          "rvest",      ## scraping web pages
                          "knitr",      ## knitting to RMarkdown
                          "kableExtra", ## add ons for knitr tables
                          "scales",     ## quick and easy formatting prettynums
                          "e1071",      ## for skew and kurtosis
                          "Hmisc",      ## correlation
                          "janitor")))  ## cleaning colnames
```

1.2.2 Data - Price History

The data was read making use of a nested for loop for the files that were separated by week. Just a single for loop was required for the data that was collated into January 2019.

```
##TODO: figure out why this is returning multiple observations for the same ASX_Ticker on the

Jan_file <- list.files(pattern = "jan")

unzip(Jan_file)

Jan_File_no_zip <- list.files(pattern = "jan")[!str_detect(
                                                  list.files(pattern = "jan"),
                                                    ".zip")]

ASX_Data_Week_Jan <- list()

ASX_Data_Month_Jan <- list()

for (k in 1:length(list.files(Jan_File_no_zip))) {

  ASX_Data_Week_Jan[[k]] <- read_csv( file.path(Jan_File_no_zip,
                                                list.files(Jan_File_no_zip)[k]),
                                      col_names = c("ASX_Ticker",
                                                    "Date",
                                                    "Open",
                                                    "High",
                                                    "Low",
```

```
                                                    "Close",
                                                    "Volume") )

  ASX_Data_Month_Jan[[k]] <- do.call(rbind, ASX_Data_Week_Jan)

}


h <- 1

repeat {

  unzip(list.files(pattern = "week")[h])

  h <- h+1

  if (h > length(list.files(pattern = "week"))) {
    break
  }

}

Week_files <- list.files(pattern = "week")
Zip_files <- list.files(pattern = ".zip")

Week_files_no_zip <- Week_files[!Week_files %in% Zip_files]

ASX_Data_List <- list()

ASX_Data_List_Week <- list()

for (i in 1:length(Week_files_no_zip)){

  for (j in 1:length(list.files(path=Week_files_no_zip[i]))){

    ASX_Data_List_Week[[j]] <- read_csv(file.path(Week_files_no_zip[i],
                                                  list.files(Week_files_no_zip[i])[j]),
                                        col_names=c("ASX_Ticker",
                                                    "Date",
                                                    "Open",
                                                    "High",
                                                    "Low",
                                                    "Close",
                                                    "Volume"))
  }

  ASX_Data_List[[i]] <- do.call(rbind, ASX_Data_List_Week)
```

```
}


ASX_Data_Frame_Jan <- do.call(rbind, ASX_Data_Month_Jan)

ASX_Data_Frame_Post_Jan <- do.call(rbind, ASX_Data_List)

ASX_Data_Frame <- rbind(ASX_Data_Frame_Jan,
                        ASX_Data_Frame_Post_Jan)

kable_styling(kable(head(ASX_Data_Frame, 20)),
              latex_options = c("striped", "hold_position"),
              position = "left",
              full_width = F)
```

ASX_Ticker	Date	Open	High	Low	Close	Volume
14D	20190102	0.310	0.310	0.300	0.300	26036
3DP	20190102	0.044	0.045	0.043	0.043	854110
3PL	20190102	1.200	1.210	1.195	1.195	101947
4CE	20190102	0.015	0.015	0.014	0.014	483448
4DS	20190102	0.056	0.059	0.054	0.057	9090917
5GN	20190102	0.420	0.420	0.420	0.420	3580
88E	20190102	0.018	0.018	0.017	0.018	3749963
8EC	20190102	0.750	0.750	0.750	0.750	2000
A2B	20190102	2.220	2.220	2.100	2.100	23364
A2M	20190102	10.330	10.620	10.330	10.400	1678613
A3D	20190102	0.490	0.490	0.480	0.480	55878
A4N	20190102	0.096	0.096	0.095	0.095	64370
AAA	20190102	50.070	50.080	50.070	50.080	123005
AAC	20190102	1.105	1.120	1.090	1.100	309374
AAJ	20190102	0.005	0.005	0.005	0.005	485000
AAR	20190102	0.087	0.088	0.087	0.088	108115
AB1	20190102	0.105	0.105	0.100	0.100	41342
ABA	20190102	5.230	5.300	5.230	5.300	4653
ABC	20190102	4.240	4.345	4.240	4.240	2025847
ABL	20190102	0.004	0.004	0.004	0.004	5000

```
kable_styling(kable(sample_n(ASX_Data_Frame, size=20)),
              latex_options = c("striped", "hold_position"),
              position = "left",
              full_width = F)
```

ASX_Ticker	Date	Open	High	Low	Close	Volume
SYT	20190211	0.007	0.007	0.007	0.007	565000
AEF	20190111	1.635	1.750	1.635	1.700	46694
GRR	20190109	0.210	0.212	0.205	0.210	425974
PIA	20190213	1.120	1.120	1.115	1.120	64020
FRN	20190123	0.017	0.017	0.017	0.017	50000
CSS	20190131	1.055	1.055	1.040	1.050	38807
NAB	20190123	24.530	24.630	24.350	24.510	3970500
EVS	20190111	0.068	0.068	0.065	0.065	221067
MAQ	20190107	20.860	20.860	20.210	20.330	293
HGO	20190109	0.084	0.088	0.084	0.088	37753
SPK	20190104	3.940	3.990	3.910	3.960	936405
DCG	20190102	0.700	0.705	0.685	0.695	31602
NWS	20190108	16.860	17.100	16.860	17.030	217164
EXO	20190108	0.005	0.005	0.005	0.005	399999
BWR	20190118	1.450	1.450	1.430	1.430	29608
IMM	20190108	0.033	0.033	0.030	0.030	2587077
KAR	20190103	0.855	0.855	0.820	0.840	250242
QUS	20190103	31.790	31.790	31.790	31.790	3117
ALQ	20190108	6.880	6.880	6.795	6.860	780851
RDF	20190111	0.430	0.470	0.430	0.450	59917

```
# rm(ASX_Data_List, ASX_Data_List_Temp)

ASX_Data_Frame <- distinct(ASX_Data_Frame)
```

1.2.3 Data - Global Industry Classification Standard

The sales data of ASX shares were enriched by adding Global Industry Classification Standard (GICS) information as well. A new table was scraped containing all companies listed on the ASX.

```
ASX_Html_Pages <- list()

for (i in 1:length(letters)) {

  ASX_Html_Pages[[i]] <- paste0(
    "https://www.asx.com.au/asx/research/listedCompanies.do?coName=",
                                toupper(letters[i]))

}

ASX_Html_Pages[length(ASX_Html_Pages)+1] <-
  "https://www.asx.com.au/asx/research/listedCompanies.do?coName=0-9"

ASX_Html_Read_list <- list()

for (i in 1:length(ASX_Html_Pages)) {

  ASX_Html_Read_list[i] <- html_table(
    html_nodes(
      read_html(x=ASX_Html_Pages[[i]]),
      "table"),
    fill = T)

  if (i > length(ASX_Html_Pages)) {
    break
  }

}


ASX_Industry_Table <- do.call(rbind, ASX_Html_Read_list)

ASX_Industry_Table <- clean_names(ASX_Industry_Table, "parsed")

ASX_Industry_Table <- select(ASX_Industry_Table,
                             -Company_name)

kable_styling(kable(head(ASX_Industry_Table, 20)),
              latex_options = c("striped", "hold_position"),
              position = "left",
              full_width = F)
```

ASX_code	GICS_industry_group
NA	NA
ACB	Energy
AYI	Diversified Financials
A2B	Commercial & Professional Services
ABP	Real Estate
ABL	Energy
AEG	Not Applic
ABT	Food, Beverage & Tobacco
AJC	Energy
AKG	Consumer Services
AX8	Materials
AX1	Retailing
ACS	Materials
ACQ	Not Applic
ACF	Capital Goods
ACR	Pharmaceuticals, Biotechnology & Life Sciences
ACW	Pharmaceuticals, Biotechnology & Life Sciences
AIV	Materials
ADA	Software & Services
ADH	Retailing

```
kable_styling(kable(sample_n(ASX_Industry_Table, size = 20)),
              latex_options = c("striped", "hold_position"),
              position = "left",
              full_width = F)
```

ASX_code	GICS_industry_group
AGL	Utilities
RVS	Semiconductors & Semiconductor Equipment
BKM	Commercial & Professional Services
MBT	Energy
LMG	Materials
KKL	Technology Hardware & Equipment
ASG	Retailing
XRF	Capital Goods
NIU	Materials
NCK	Retailing
SIL	Health Care Equipment & Services
RTR	Materials
WAX	Not Applic
SGL	Food, Beverage & Tobacco
SCL	Software & Services
LVE	Software & Services
CUV	Pharmaceuticals, Biotechnology & Life Sciences
BMP	Real Estate
DAE	Not Applic
TRA	Retailing

```
ASX_Data_Frame <- left_join(x = ASX_Data_Frame,
                            y = ASX_Industry_Table,
                            by = c("ASX_Ticker" = "ASX_code"))
```

1.2.4 Descriptive Statistics

The data set was incredibly right-skewed, as outlined by the summary table below of each variable. However, all the price variables (Close, High, Low, Open) appeared to have similar measures of skew, kurtosis, and IQR.

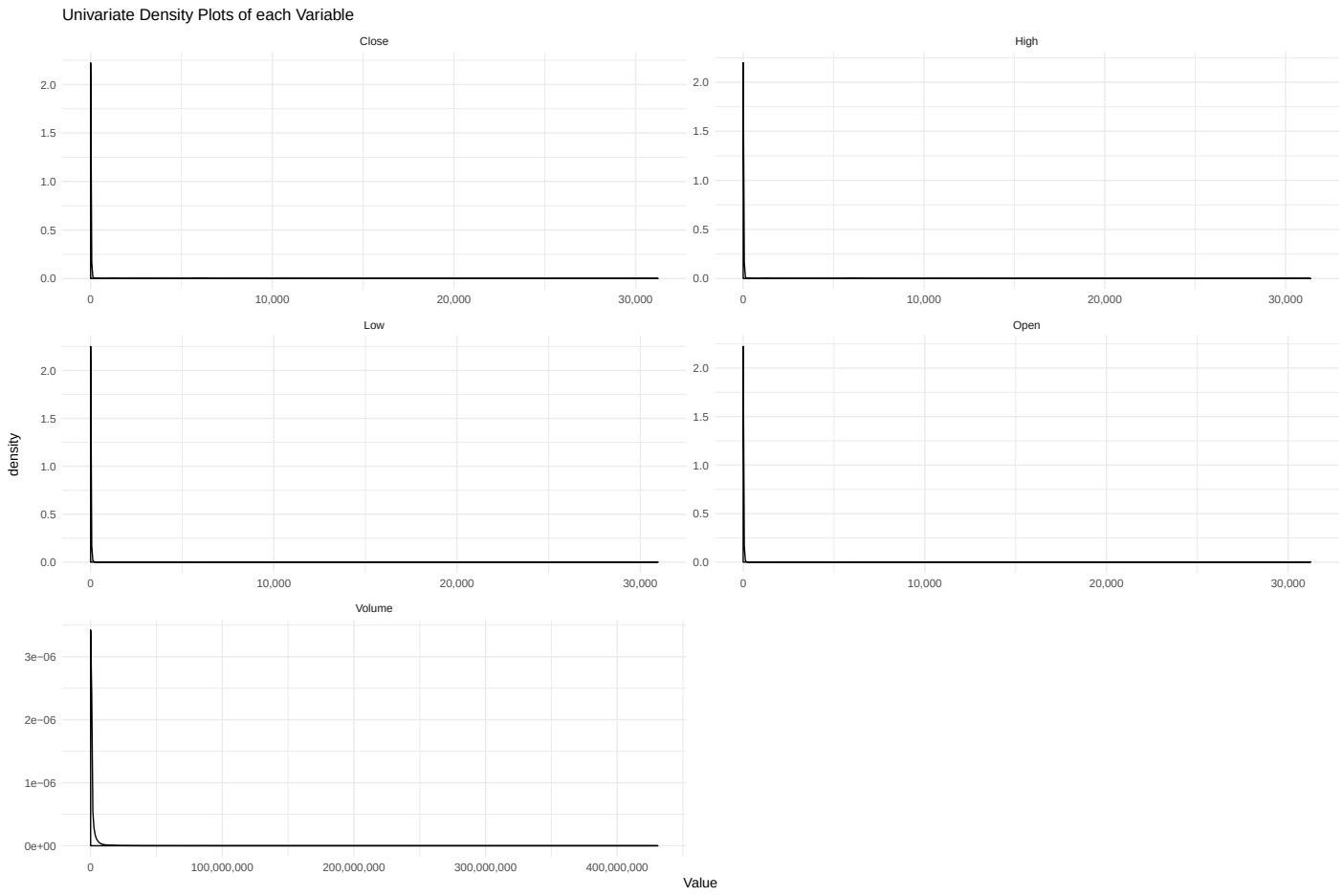
```
ASX_Long <- gather(ASX_Data_Frame,
                   Open:Volume,
                   key="Variable",
                   value="Value")

ASX_Summary <- summarise(group_by(ASX_Long,
                                  Variable),
                         "n ASX_Tickers" = comma(length(unique(ASX_Ticker))),
                         "n Observations" = comma(n()),
                         "Min Date" = format(ymd(min(Date)), "%d-%m-%Y"),
                         "Max Date" = format(ymd(max(Date)), "%d-%m-%Y"),
                         "Minimum" = comma(min(Value)),
                         "Q1" = comma(quantile(Value, 0.25)),
                         "Median" = comma(quantile(Value, 0.5)),
                         "Q3" = comma(quantile(Value, 0.75)),
                         "90th Percentile" = comma(quantile(Value, 0.9)),
                         "95th Percentile" = comma(quantile(Value, 0.95)),
                         "Maximum" = comma(max(Value)),
                         "Skew" = round(skewness(Value), 3),
                         "Kurtosis" = round(kurtosis(Value), 2),
                         "NA count" = comma(sum(is.na(ASX_Data_Frame))))

kable_styling(kable(t(ASX_Summary)),
              full_width = F,
              latex_options = c("striped", "hold_position"),
              position = "left")
```

Variable	Close	High	Low	Open	Volume
n ASX_Tickers	2,048	2,048	2,048	2,048	2,048
n Observations	109,452	109,452	109,452	109,452	109,452
Min Date	02-01-2019	02-01-2019	02-01-2019	02-01-2019	02-01-2019
Max Date	12-04-2019	12-04-2019	12-04-2019	12-04-2019	12-04-2019
Minimum	0	0	0	0	0
Q1	0	0	0	0	32,000
Median	0	0	0	0	166,381
Q3	2	3	2	2	770,116
90th Percentile	13	14	13	13	2,621,474
95th Percentile	39	40	39	39	4,897,755
Maximum	31,227	31,377	30,962	31,227	430,924,497
Skew	17.065	17.091	17.053	17.075	33.523
Kurtosis	392.57	393.71	392.10	393.05	2272.27
NA count	7,661	7,661	7,661	7,661	7,661

1.2.5 Density Plots

Plotting the spread of the variables only further outlines the extremity of the skew. As such, it makes sense to trim upper extreme outliers from the data set.

```
ggplot(ASX_Long) +
  geom_density(aes(x=Value)) +
  scale_x_continuous(labels=comma) +
  facet_rep_wrap(~Variable, repeat.tick.labels = T,
                 scales = "free", ncol = 2) +
  ggtitle("Univariate Density Plots of each Variable") +
  theme_minimal()
```



1.2.6 Filtering data

As the data was extremely positively skewed, trimming out the top 1/3rd of the data allowed for concentration on the shares with similar prices. The data was trimmed by `ASX_Ticker` to remove shares that sold for `High` prices in the top 1/3 quantile at any date during the time considered. Summary statistics on the variables showed that this filtered data focussed on shares that sold for between $0.02 and $0.96 on any date. However, this data is still for a total of 1,365 unique `ASX_Tickers` spread across nearly 65,000 observations, so the data set is somewhat sizeable.

```
ASX_Ticker_Summary <- summarise(group_by(ASX_Data_Frame, ASX_Ticker),
                                "n ASX_Tickers" = comma(length(unique(ASX_Ticker))),
                                "n Observations" = comma(n()),
                                "Min Date" = format(ymd(min(Date)), "%d/%m/%Y"),
                                "Max Date" = format(ymd(max(Date)), "%d/%m/%Y"),
                                "Minimum" = min(High),
                                "Q1" = quantile(High, 0.25),
                                "Median" = quantile(High, 0.5),
                                "Q3" = quantile(High, 0.75),
                                "90th Percentile" = quantile(High, 0.9),
                                "95th Percentile" = quantile(High, 0.95),
                                "Maximum" = max(High),
                                "Skew" = round(skewness(High), 3),
                                "Kurtosis" = round(kurtosis(High), 2),
                                "NA count" = comma(sum(is.na(ASX_Data_Frame))))
```

```
kable_styling(kable(sample_n(ASX_Ticker_Summary, 20)),
              latex_options = c("striped", "hold_position"),
              position = "left",
              full_width = F)
```

ASX_Ticker	n ASX_Tickers	n Observations	Min Date	Max Date	Minimum	Q1	Median	Q
CIO	1	45	03/01/2019	12/04/2019	0.002	0.00300	0.0030	0.003
EMP	1	30	02/01/2019	12/04/2019	0.002	0.00225	0.0030	0.003
PSZ	1	58	02/01/2019	12/04/2019	0.054	0.07200	0.0915	0.125
FCC	1	64	02/01/2019	12/04/2019	0.155	0.17500	0.1850	0.195
DCG	1	72	02/01/2019	12/04/2019	0.695	0.72500	0.8450	0.880
ASH	1	46	03/01/2019	12/04/2019	0.240	0.25000	0.2535	0.263
LBL	1	67	02/01/2019	12/04/2019	0.315	0.38000	0.4000	0.420
D2O	1	72	02/01/2019	12/04/2019	1.375	1.47375	1.5050	1.580
CGS	1	64	02/01/2019	12/04/2019	0.195	0.26750	0.4500	0.547
NTI	1	43	02/01/2019	12/04/2019	0.007	0.01800	0.0260	0.030
LNK	1	72	02/01/2019	12/04/2019	6.750	7.20500	7.3500	7.618
XFL	1	72	02/01/2019	12/04/2019	5560.100	5814.22500	6055.4500	6110.775
GNC	1	72	02/01/2019	12/04/2019	8.990	9.22500	9.3950	9.600
PWH	1	72	02/01/2019	12/04/2019	3.150	3.48750	3.5900	3.650
E25	1	52	02/01/2019	11/04/2019	0.155	0.17000	0.1750	0.190
WEL	1	34	15/01/2019	12/04/2019	0.019	0.02000	0.0215	0.036
CHN	1	71	02/01/2019	12/04/2019	0.120	0.13250	0.1400	0.155
CAZ	1	46	04/01/2019	12/04/2019	0.018	0.02400	0.0265	0.029
XNJ	1	72	02/01/2019	12/04/2019	5652.700	5884.15000	6185.0000	6268.100
AMC	1	72	02/01/2019	12/04/2019	13.230	13.65250	14.8525	15.055

```
ASX_Lower <- filter(ASX_Ticker_Summary, Maximum < quantile(Maximum, 2/3))

nrow(ASX_Lower)
```

```
## [1] 1365
```

```
ASX_Long_Lower <- filter(ASX_Long, ASX_Ticker %in% ASX_Lower$ASX_Ticker)

ASX_Data_Lower <- filter(ASX_Data_Frame, ASX_Ticker %in% ASX_Lower$ASX_Ticker)

kable_styling(kable(sample_n(ASX_Data_Lower, 20)),
              latex_options = c("striped", "hold_position"),
              position = "left",
              full_width = F)
```

ASX_Ticker	Date	Open	High	Low	Close	Volume	GICS_industry_group
AFR	20190222	0.017	0.017	0.016	0.017	195959	Energy
FZO	20190329	0.200	0.220	0.200	0.200	124479	Software & Services
SE1	20190308	0.115	0.130	0.115	0.125	366765	Semiconductors & Semiconductor Equipment
FOD	20190205	0.145	0.150	0.140	0.150	698448	Food, Beverage & Tobacco
YPB	20190122	0.016	0.019	0.016	0.019	1222637	Commercial & Professional Services
BRK	20190111	0.012	0.013	0.012	0.013	945864	Energy
MML	20190214	0.370	0.370	0.355	0.360	171182	Materials
ANP	20190111	0.030	0.032	0.030	0.030	65482	Pharmaceuticals, Biotechnology & Life Sciences
AUR	20190125	0.031	0.033	0.030	0.032	894668	Materials
NWF	20190328	0.155	0.155	0.155	0.155	30354	Materials
LOM	20190218	0.210	0.210	0.190	0.190	475315	Materials
CAM	20190221	0.835	0.835	0.835	0.835	24904	Not Applic
NC6	20190404	0.069	0.069	0.065	0.065	35392	Pharmaceuticals, Biotechnology & Life Sciences
SIL	20190110	0.270	0.270	0.265	0.265	49200	Health Care Equipment & Services
SIS	20190116	0.072	0.072	0.070	0.070	93173	Software & Services
CRL	20190222	0.036	0.036	0.036	0.036	100000	Materials
KMT	20190108	0.014	0.014	0.013	0.013	120666	Materials
UTR	20190201	0.007	0.007	0.005	0.005	16438885	Capital Goods
ALK	20190114	0.190	0.195	0.190	0.195	40695	Materials
QFY	20190222	0.007	0.007	0.007	0.007	1200206	Technology Hardware & Equipment

1.2.6.1 Summary Statistics of Data After Removing Extreme ASX_Tickers

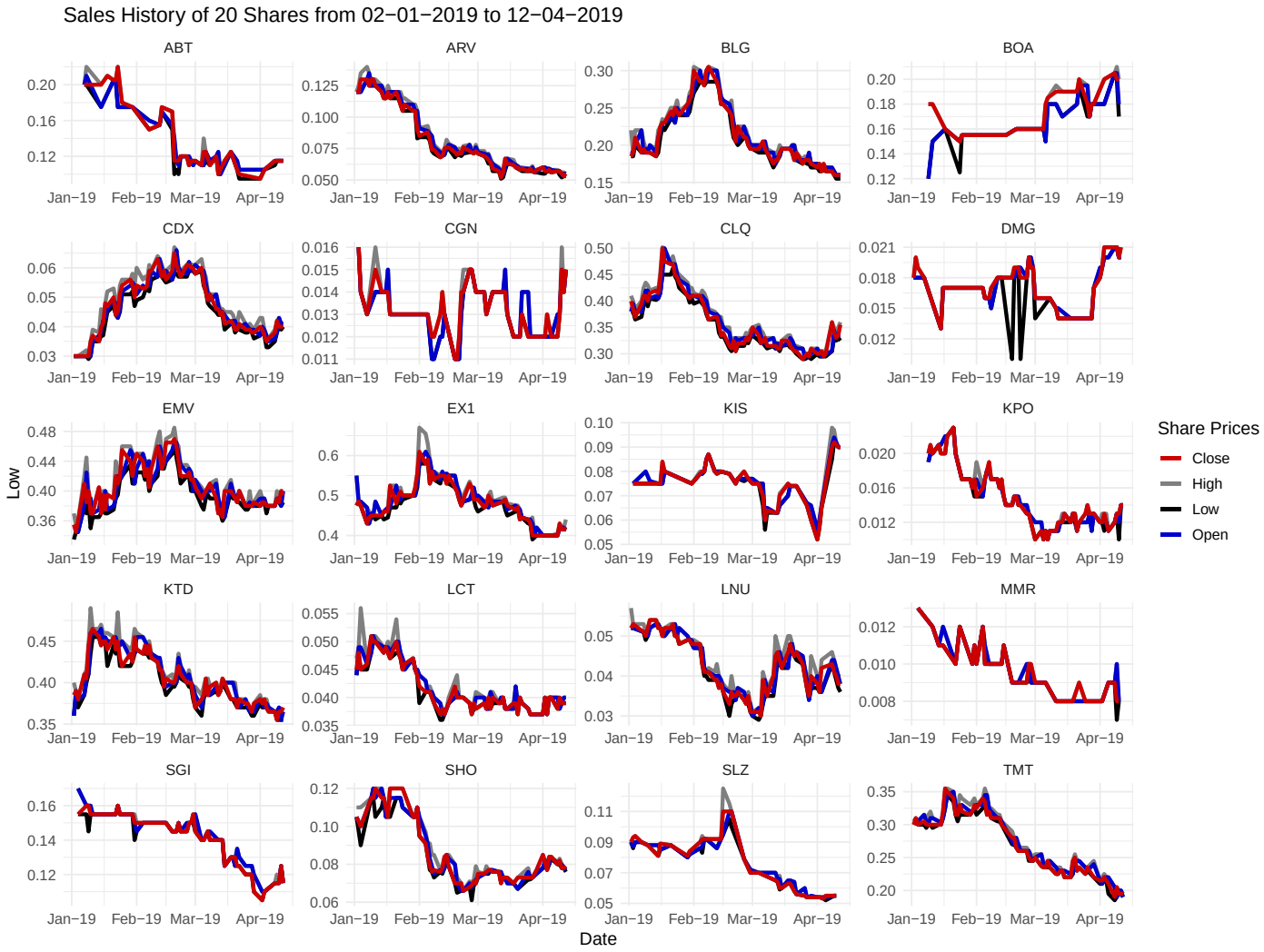
```
ASX_Summary_Lower <- summarise(group_by(ASX_Long_Lower,
                                        Variable),
                               "n ASX_Tickers" = comma(length(unique(ASX_Ticker))),
                               "n Observations" = comma(n()),
                               "Min Date" = format(ymd(min(Date)), "%d/%m/%Y"),
                               "Max Date" = format(ymd(max(Date)), "%d/%m/%Y"),
                               "Minimum" = round(min(Value), 2),
                               "Q1" = round(quantile(Value, 0.25), 2),
                               "Median" = round(quantile(Value, 0.5), 2),
                               "Q3" = round(quantile(Value, 0.75), 2),
                               "90th Percentile" = round(quantile(Value, 0.9), 2),
                               "95th Percentile" = round(quantile(Value, 0.95), 2),
                               "Maximum" = round(max(Value), 2),
                               "Skew" = round(skewness(Value), 3),
                               "Kurtosis" = round(kurtosis(Value), 2),
                               "NA count" = comma(sum(is.na(ASX_Data_Frame))))

kable_styling(kable(t(ASX_Summary_Lower)),
              latex_options = c("striped", "hold_position"),
              position = "left",
              full_width = F)
```

Variable	Close	High	Low	Open	Volume
n ASX_Tickers	1,365	1,365	1,365	1,365	1,365
n Observations	64,873	64,873	64,873	64,873	64,873
Min Date	02/01/2019	02/01/2019	02/01/2019	02/01/2019	02/01/2019
Max Date	12/04/2019	12/04/2019	12/04/2019	12/04/2019	12/04/2019
Minimum	0	0	0	0	1
Q1	0.02	0.02	0.02	0.02	45873.00
Median	0.08	0.09	0.08	0.08	182490.00
Q3	0.23	0.24	0.23	0.23	712515.00
90th Percentile	0.48	0.49	0.48	0.48	2353937.80
95th Percentile	0.64	0.65	0.63	0.64	4523797.80
Maximum	0.96	0.96	0.96	0.96	430924497.00
Skew	1.675	1.659	1.689	1.672	33.654
Kurtosis	2.20	2.12	2.27	2.19	2005.06
NA count	7,661	7,661	7,661	7,661	7,661

1.3 Data Exploration and Visualisation

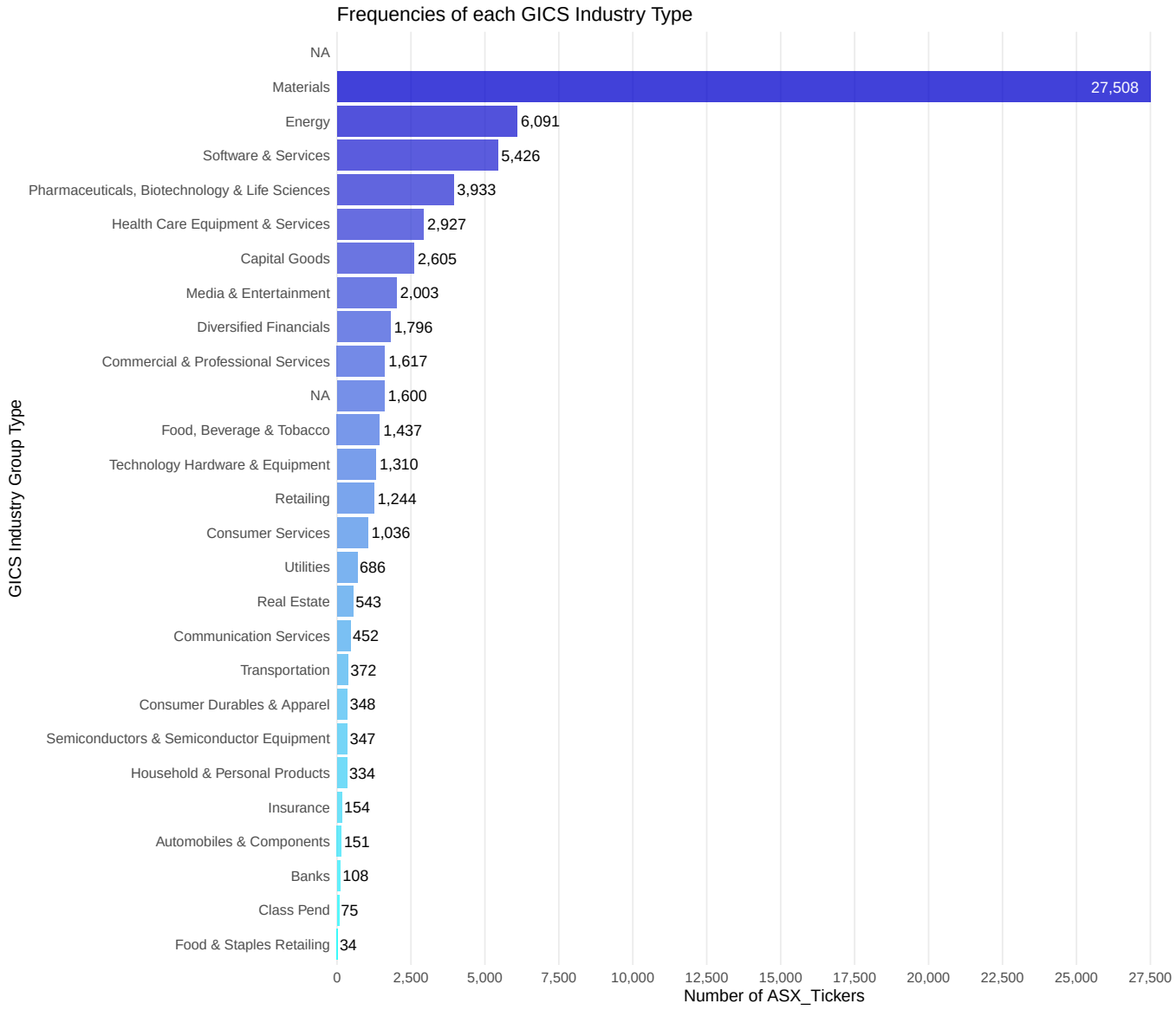
1.3.1 Univariate Visualisations

```
ASX_Data_Lower$Date <- ymd(ASX_Data_Lower$Date)

ASX_Data_Lower <- arrange(ASX_Data_Lower, ASX_Ticker, Date)

Sample_Tickers <- sample(ASX_Data_Lower$ASX_Ticker, size = 20)

ASX_Data_Samples <- arrange(filter(ASX_Data_Lower, ASX_Ticker %in% Sample_Tickers),
                            ASX_Ticker, Date)

ggplot(ASX_Data_Samples) +
  geom_line(aes(x=Date, y=Low, col="Low"), size=1.25) +
  geom_line(aes(x=Date, y=High, col="High"), size=1.25) +
  geom_line(aes(x=Date, y=Open, col="Open"), size=1.25) +
  geom_line(aes(x=Date, y=Close, col="Close"), size=1.25) +
  scale_x_date(date_breaks = "month", date_labels = "%b-%y") +
  scale_color_manual(name = "Share Prices",
                     values = c("Open"="blue3",
                                "High"="grey50",
                                "Low"="black",
                                "Close"="red3")) +
  ggtitle("Sales History of 20 Shares from 02-01-2019 to 12-04-2019") +
  facet_rep_wrap(~ASX_Ticker, repeat.tick.labels = T,
                 scales = "free_y", ncol = 4) +
  theme_minimal() +
  theme(text = element_text(size = 12))
```




```
ASX_Data_Lower$GICS_industry_group <- recode(ASX_Data_Lower$GICS_industry_group, "Not Applic"

fill_grad <-
  seq_gradient_pal("blue3",
                   "cyan")(seq(0,1,
                               length.out = length(
                                 unique(ASX_Data_Lower$GICS_industry_group))-1))

ASX_Data_Count <- summarise(group_by(ASX_Data_Lower,
                                     GICS_industry_group),
                            "Count" = n())

ggplot(ASX_Data_Lower, aes(x = fct_rev(fct_infreq(GICS_industry_group)),
                           fill = fct_infreq(GICS_industry_group))) +
  geom_bar(show.legend = F, alpha = 0.75) +
  geom_text(data = filter(ASX_Data_Count,
                          GICS_industry_group != "Materials"),
            aes(x = GICS_industry_group,
                y = Count,
                label = comma(Count)),
            hjust = -0.1) +
```

```
  geom_text(data = filter(ASX_Data_Count,
                          GICS_industry_group == "Materials"),
            aes(x = GICS_industry_group,
                y = Count,
                label = comma(Count)),
            hjust = 1.25, col="white") +
ggtitle("Frequencies of each GICS Industry Type") +
scale_y_continuous(breaks = seq(0,30000, 2500),
                   limits = c(0,max(ASX_Data_Count$Count)*1.075),
                   expand = c(0,0),
                   labels = comma,
                   "Number of ASX_Tickers") +
scale_x_discrete("GICS Industry Group Type") +
scale_fill_manual(values = c(fill_grad, "grey60")) +
theme_minimal() +
coord_flip() +
theme(panel.grid.minor.x = element_blank(),
      panel.grid.major.y = element_blank(),
      panel.grid.minor.y = element_blank(),
      text = element_text(size = 12),
      panel.border = element_blank())
```


Frequencies of each GICS Industry Type
GICS Industry Group Type
NA
Materials
Energy
Software & Services
Pharmaceuticals, Biotechnology & Life Sciences
Health Care Equipment & Services
Capital Goods
Media & Entertainment
Diversified Financials
Commercial & Professional Services
NA
Food, Beverage & Tobacco
Technology Hardware & Equipment
Retailing
Consumer Services
Utilities
Real Estate
Communication Services
Transportation
Consumer Durables & Apparel
Semiconductors & Semiconductor Equipment
Household & Personal Products
Insurance
Automobiles & Components
Banks
Class Pend
Food & Staples Retailing
27,508
6,091
5,426
3,933
2,927
2,605
2,003
1,796
1,617
1,600
1,437
1,310
1,244
1,036
686
543
452
372
348
347
334
154
151
108
75
34
0
2,500
5,000
7,500
10,000
12,500
15,000
17,500
20,000
22,500
25,000
27,500
Number of ASX_Tickers