|Selling Prices and Volumes Sold |of Australian Stock Exchange Shares

MATH2319 - Machine Learning
Course Project

Ben Cole - s3412349

Print Date: 28/04/2019

Contents

1 Phase 1 - Introduction, Cleaning, and Exploration . . . 2

1.1 Outline . . . 2

1.1.1 Nature of the Data . . . 2

1.1.2 Target Feature . . . 3

1.1.3 Descriptive Features . . . 3

1.2 Data Processing . . . 4

1.2.1 Packages . . . 4

1.2.2 Data - Price History . . . 4

1.2.3 Data - Global Industry Classification Standard . . . 8

1.2.4 Removing Company Name . . . 9

1.2.5 Descriptive Statistics . . . 9

1.2.6 Density Plots . . . 11

1.2.7 Filtering Data by Price . . . 12

1.2.8 Filtering Data by Volume . . . 15

1.2.9 Density Plots After Filtering by Price and Volume . . . 18

1.2.10 Summary Statistics of Data After Removing Extreme ASX_Tickers . . . 18

1.3 Data Exploration and Visualisation . . . 20

1.3.1 Share Price Tracking . . . 20

1.3.2 Volume of Shares Sold . . . 22

1.3.3 Number of Companies per GICS Group . . . 24

1.3.4 Mean Volumes Sold by GICS Groups . . . 26

1.3.5 Volumes Sold of each GICS per Day . . . 28

1.3.6 Pricing Features for Each GICS Group . . . 30

1.4 Summary . . . 32

1.4.1 References . . . 33

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 dataset. Data will be sourced for dates after the last date in the training and validation dataset for the following Phase 2 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 dataset to add further categorical information.

Volume of shares sold was chosen as the target feature while pricing variables and GICS grouping were chosen as descriptive features. Date was not used as a descriptive feature but retained in the dataset for future use in Phase 2 of this project.

The data was found to be heavily right-skewed for all price variables. The data was filtered to remove ASX tickers with extremely large High selling prices and with extremely large sales volumes. After filtering, the data was visualised to show that it was less skewed for all continuous descriptive features. GICS Industry Group, the only categorical descriptive feature, was also shown to be less skewed after filtering as well as somewhat similarly distributed between GICS groups.

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

The above variable names are stated on the l**ASX Historical Data website**](¡'https://www.asxhistoricaldata.com/").

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
                          "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 the file January 2019.

```
if (length(list.files(pattern = "jan")[!str_detect(
      list.files(pattern = "jan"),
        ".zip")]) == 0) {

  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)

}

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

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

if(length(Week_files_no_zip)==0) {

  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(sample_n(ASX_Data_Frame, size=20),
                    align = "rrrrrrrll"),
              latex_options = c("striped", "hold_position"),
              position = "center",
              full_width = F)
```

ASX_Ticker	Date	Open	High	Low	Close	Volume
ALY	20190111	0.016	0.016	0.016	0.016	1000000
RMP	20190131	0.085	0.086	0.084	0.084	4820742
ELS	20190207	0.520	0.545	0.510	0.535	22908
SIQ	20190306	8.500	8.560	8.430	8.500	725721
RAC	20190117	0.098	0.099	0.095	0.095	198098
THC	20190109	0.480	0.510	0.480	0.510	132080
HTA	20190314	0.135	0.135	0.135	0.135	17546
BLK	20190110	0.046	0.048	0.046	0.046	3705910
DTR	20190403	0.002	0.002	0.002	0.002	170000
PAF	20190102	0.990	1.000	0.990	1.000	21011
GML	20190115	0.015	0.016	0.015	0.015	6063621
TTT	20190109	2.300	2.350	2.250	2.280	52774
VGS	20190107	64.960	65.350	64.960	65.030	22986
NHF	20190116	5.170	5.430	5.170	5.360	544477
TGF	20190125	2.430	2.430	2.380	2.380	2129
DHR	20190115	0.004	0.004	0.004	0.004	2777750
MPL	20190207	2.760	2.770	2.730	2.740	7656056
OTW	20190118	4.600	4.640	4.590	4.600	44566
BSX	20190104	0.110	0.110	0.110	0.110	99136
CLF	20190111	1.200	1.210	1.200	1.210	87097

```
ASX_Data_Frame <- distinct(ASX_Data_Frame,
                           ASX_Ticker, Date,
                           .keep_all = T)
```

1.2.3 Data - Global Industry Classification Standard

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

```r
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(sample_n(ASX_Industry_Table, size = 20)),
              latex_options = c("striped", "hold_position"),
              position = "center",
              full_width = F)
```

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

ASX_code	GICS_industry_group
COE	Energy
NOR	Software & Services
IMU	Pharmaceuticals, Biotechnology & Life Sciences
GBZ	Materials
RDA	Not Applic
LVT	Software & Services
MYO	Software & Services
LGD	Capital Goods
MMM	Consumer Services
EXR	Energy
ORR	Materials
PEX	Materials
PAB	Pharmaceuticals, Biotechnology & Life Sciences
TOZ	Not Applic
GLV	Energy
NAM	Commercial & Professional Services
WGO	Energy
FNT	Materials
WSN	Not Applic
SOR	Diversified Financials

1.2.4 Removing Company Name

As each `ASX_ticker` is individually linked to a single `Company_name`, `Company_name` clearly does not provide any extra information to the dataset and so was removed.

```
ASX_Data_Frame$Company_name <- NULL

kable_styling(kable(sample_n(ASX_Data_Frame, 20),
                    align = "lrrrrrrl"),
              latex_options = c("striped", "hold_position"),
              position = "center",
              full_width = F)
```

1.2.5 Descriptive Statistics

The dataset was heavily right-skewed, as outlined by the summary table below of each pricing feature. However, all the price features (`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()),
```

ASX_Ticker	Date	Open	High	Low	Close	Volume	GICS_industry_group
SPT	20190326	1.140	1.330	1.120	1.320	12212467	Software & Services
RLE	20190307	0.105	0.115	0.105	0.110	1486206	Energy
FLN	20190118	0.842	0.845	0.840	0.845	11655	Commercial & Professional Services
GLN	20190103	0.270	0.290	0.270	0.290	135171	Materials
KSS	20190213	0.125	0.125	0.120	0.125	87500	Commercial & Professional Services
OVN	20190214	0.300	0.300	0.280	0.280	181249	Health Care Equipment & Services
ISX	20190401	0.300	0.300	0.290	0.295	1650690	Software & Services
RUB	20190222	0.010	0.010	0.010	0.010	10000	Commercial & Professional Services
ISU	20190305	0.720	0.750	0.720	0.730	116511	Consumer Services
NHC	20190118	3.640	3.820	3.610	3.770	954842	Energy
RSG	20190116	1.150	1.150	1.115	1.140	3959848	Materials
SRN	20190305	0.003	0.003	0.003	0.003	560838	Materials
TSN	20190314	0.007	0.007	0.007	0.007	505997	Software & Services
JPR	20190305	0.030	0.030	0.030	0.030	13334	Energy
PXS	20190306	0.262	0.265	0.262	0.265	56199	Pharmaceuticals, Biotechnology & Life Sciences
NBL	20190312	3.210	3.210	3.140	3.160	49898	Retailing
CLI	20190329	0.027	0.027	0.024	0.025	7245937	Commercial & Professional Services
EQT	20190129	24.390	24.450	24.025	24.390	6155	Diversified Financials
KAI	20190205	0.022	0.022	0.021	0.021	1074295	Materials
CNU	20190220	5.030	5.050	4.980	5.010	272369	Communication Services

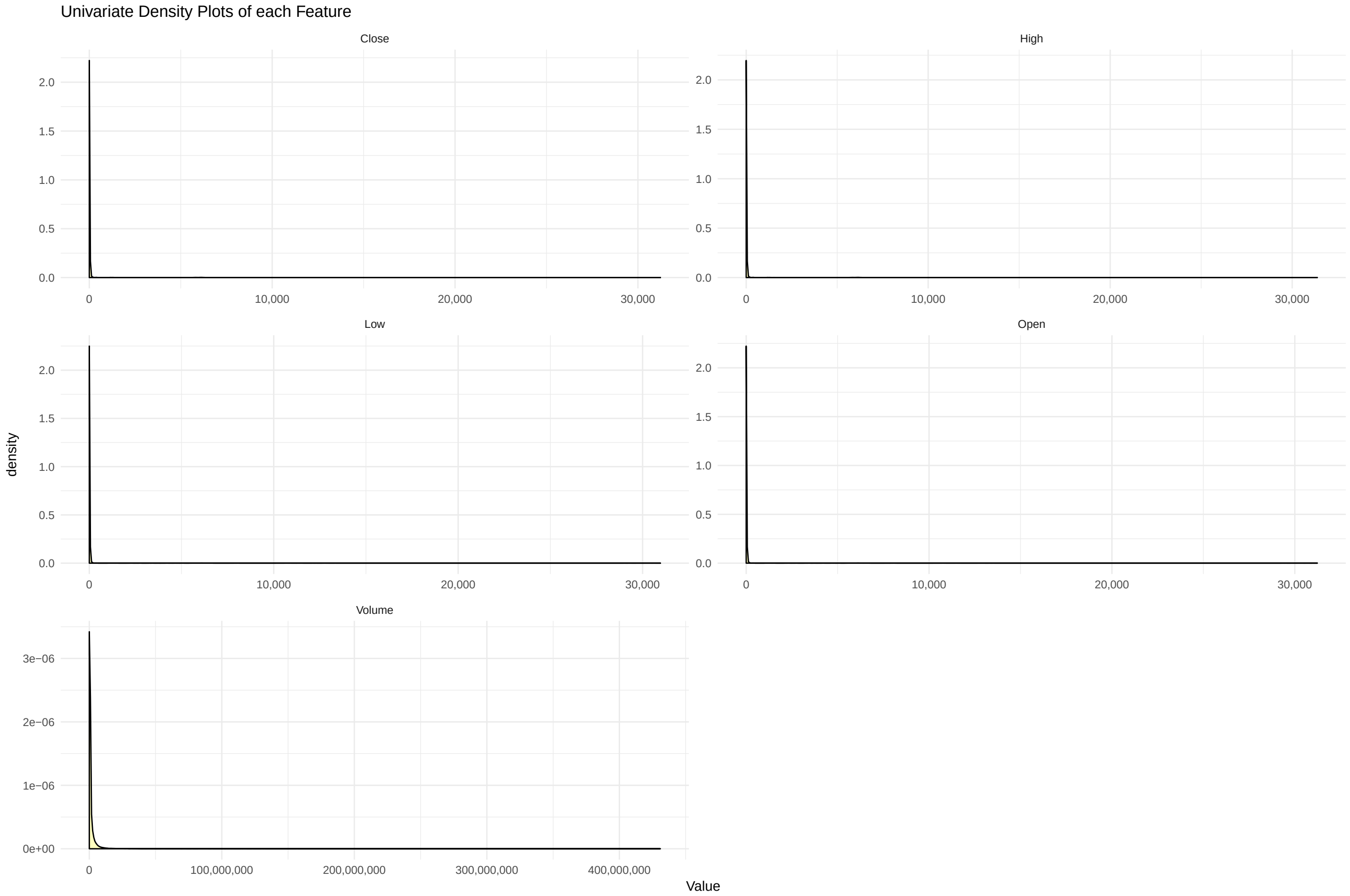
```
                         "Min Date" = format(ymd(min(Date)), "%d-%m-%Y"),
                         "Max Date" = format(ymd(max(Date)), "%d-%m-%Y"),
                         "Minimum" = format(round(min(Value), 3),
                                            big.mark = ","),
                         "Q1" = format(round(quantile(Value, 0.25), 3),
                                                  big.mark = ","),
                         "Median" = format(round(quantile(Value, 0.5), 3),
                                                  big.mark = ","),
                         "Q3" = format(round(quantile(Value, 0.75), 3),
                                                  big.mark = ","),
                         "90th Percentile" = format(round(quantile(Value, 0.9), 3),
                                                  big.mark = ","),
                         "95th Percentile" = format(round(quantile(Value, 0.95), 3),
                                                  big.mark = ","),
                         "Maximum" = format(round(max(Value), 3),
                                            big.mark = ","),
                         "Skew" = round(skewness(Value), 3),
                         "Kurtosis" = round(kurtosis(Value), 3),
                         "NA count" = format(round(sum(is.na(ASX_Data_Frame)), 3),
                                                  big.mark = ","))

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

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	0.001	0.001	0.001	0.001	0
Q1	0.062	0.064	0.061	0.062	32,000
Median	0.365	0.37	0.36	0.365	166,381
Q3	2.5	2.53	2.47	2.5	770,116
90th Percentile	13.49	13.599	13.35	13.47	2,621,474
95th Percentile	39.364	39.842	38.916	39.295	4,897,755
Maximum	31,227.1	31,376.8	30,962.3	31,227.1	430,924,497
Skew	17.065	17.091	17.053	17.075	33.523
Kurtosis	392.572	393.710	392.098	393.048	2272.266
NA count	7,717	7,717	7,717	7,717	7,717

1.2.6 Density Plots

Plotting the spread of the features only further outlined the magnitude of the skew. As such, the data was filtered to remove shares that showed high values for any feature.

```
ggplot(ASX_Long) +
  geom_density(aes(x = Value),
               fill = "yellow", alpha = 0.25) +
  scale_x_continuous(labels=comma) +
  facet_rep_wrap(~Variable, repeat.tick.labels = T,
                 scales = "free", ncol = 2) +
  ggtitle("Univariate Density Plots of each Feature") +
  theme_minimal()
```



1.2.7 Filtering Data by Price

As the data was extremely positively skewed, trimming out the top 1/3 quantile 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.

```
ASX_Ticker_Summary_Price <-
  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), 3))

ASX_kable <- sample_n(ASX_Ticker_Summary_Price, 20)
```

```r
kable_styling(kable(ASX_kable[, 1:7],
                    align = "lrrrrrr"),
              latex_options = c("striped", "hold_position"),
              position = "center",
              full_width = F,
              font_size = 10)
```

ASX_Ticker	n ASX_Tickers	n Observations	Min Date	Max Date	Minimum	Q1
ELO	1	72	02/01/2019	12/04/2019	4.750	5.35000
HSN	1	72	02/01/2019	12/04/2019	2.930	3.04000
S32	1	72	02/01/2019	12/04/2019	3.280	3.48000
BGP	1	7	07/02/2019	08/04/2019	3.100	3.20000
DYL	1	72	02/01/2019	12/04/2019	0.395	0.41500
FRN	1	28	02/01/2019	12/04/2019	0.015	0.01600
GPT	1	72	02/01/2019	12/04/2019	5.380	5.78125
UCW	1	18	10/01/2019	12/04/2019	0.110	0.12500
SGO	1	15	07/01/2019	11/04/2019	0.008	0.01100
KFE	1	66	02/01/2019	12/04/2019	0.070	0.08000
AB1	1	67	02/01/2019	12/04/2019	0.087	0.09100
SKO	1	67	02/01/2019	12/04/2019	2.530	2.97000
WLD	1	63	02/01/2019	12/04/2019	0.042	0.05250
BMN	1	72	02/01/2019	12/04/2019	0.037	0.03900
SGC	1	47	02/01/2019	10/04/2019	0.023	0.02500
ICN	1	43	02/01/2019	09/04/2019	0.017	0.01800
KGN	1	72	02/01/2019	12/04/2019	3.200	3.75500
MTO	1	71	02/01/2019	12/04/2019	1.450	1.54500
EOF	1	11	29/03/2019	12/04/2019	1.550	1.65500
SFM	1	30	03/01/2019	03/04/2019	0.078	0.08100

```r
kable_styling(kable(ASX_kable[, c(1, 8:14)],
                    align = "lrrrrrrr"),
              latex_options = c("striped", "hold_position"),
              position = "center",
              full_width = F,
              font_size = 10)
```

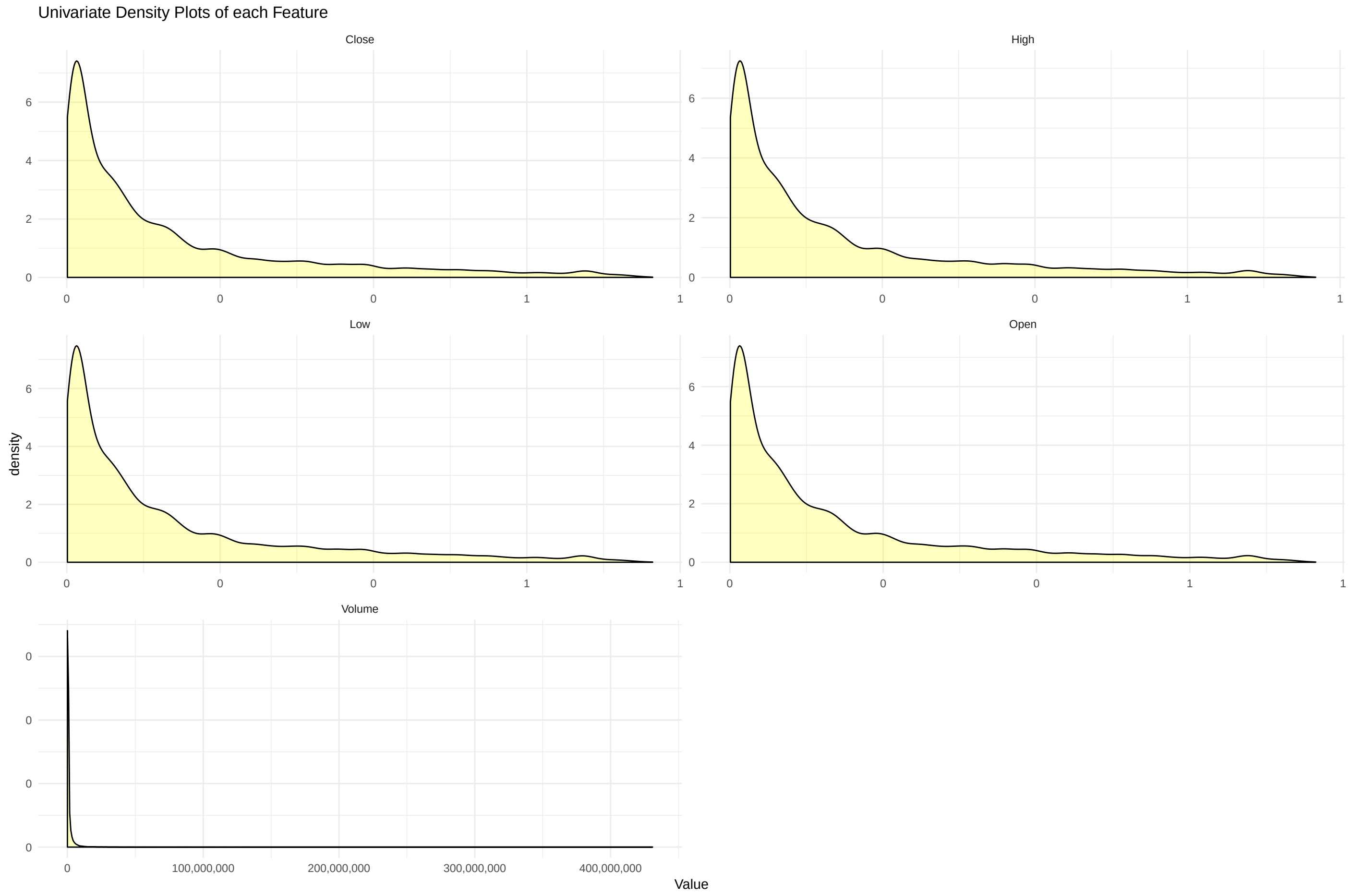
```r
ASX_Lower <- filter(ASX_Ticker_Summary_Price, Maximum < quantile(Maximum, 2/3))

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),
                    align = "lrrrrrl"),
              latex_options = c("striped", "hold_position"),
              position = "center",
              full_width = F)
```

ASX_Ticker	Median	Q3	90th Percentile	95th Percentile	Maximum	Skew	Kurtosis
ELO	5.6750	5.79250	6.0290	6.28450	6.600	0.187	-0.160
HSN	3.3700	3.50000	3.5980	3.63450	3.650	-0.013	-1.729
S32	3.7800	3.89000	3.9200	3.96000	3.990	-0.452	-1.360
BGP	3.2500	3.30000	3.3200	3.33500	3.350	-0.433	-1.510
DYL	0.4425	0.45625	0.5045	0.51225	0.530	0.595	-0.630
FRN	0.0170	0.01725	0.0190	0.01900	0.019	0.058	-0.951
GPT	6.0100	6.16625	6.2545	6.28725	6.320	-0.649	-0.804
UCW	0.1300	0.13875	0.1495	0.16150	0.170	0.891	0.208
SGO	0.0120	0.01400	0.0198	0.02100	0.021	0.911	-0.474
KFE	0.0875	0.09775	0.1100	0.12000	0.125	0.839	0.078
AB1	0.0980	0.10000	0.1120	0.12500	0.155	2.377	6.035
SKO	3.0400	3.17500	3.2620	3.33100	3.500	-0.381	0.081
WLD	0.0540	0.05600	0.0580	0.05990	0.070	0.150	2.165
BMN	0.0410	0.04625	0.0499	0.05100	0.055	0.930	-0.275
SGC	0.0260	0.02700	0.0270	0.02800	0.030	0.481	0.438
ICN	0.0190	0.01900	0.0200	0.02000	0.020	-0.009	-0.869
KGN	4.1625	4.33250	4.4960	4.59450	4.880	-0.336	-0.792
MTO	1.6500	1.85250	2.0350	2.07750	2.100	0.479	-0.947
EOF	1.8800	2.33000	2.4600	2.55000	2.640	0.403	-1.601
SFM	0.0860	0.09000	0.0930	0.09355	0.099	0.384	-0.748

ASX_Ticker	Date	Open	High	Low	Close	Volume	GICS_industry_group
SYA	20190305	0.023	0.023	0.022	0.022	378729	Materials
IAB	20190129	0.595	0.595	0.590	0.590	120000	Communication Services
KOR	20190214	0.027	0.027	0.027	0.027	223434	Materials
CDY	20190226	0.265	0.265	0.265	0.265	10000	Pharmaceuticals, Biotechnology & Life Sciences
GPX	20190228	0.250	0.250	0.240	0.240	30462	Materials
DXN	20190206	0.080	0.080	0.077	0.077	333546	Software & Services
DCC	20190104	0.050	0.050	0.050	0.050	473073	Software & Services
TOP	20190226	0.635	0.635	0.635	0.635	90156	Not Applic
BNO	20190319	0.180	0.180	0.165	0.165	679957	Pharmaceuticals, Biotechnology & Life Sciences
SVD	20190228	0.012	0.012	0.012	0.012	562500	Materials
NWE	20190228	0.003	0.003	0.003	0.003	950000	Energy
MYX	20190205	0.785	0.805	0.780	0.785	4546871	Pharmaceuticals, Biotechnology & Life Sciences
MRQ	20190320	0.005	0.005	0.005	0.005	275006	Materials
TIA	20190410	0.450	0.450	0.420	0.420	2110	Real Estate
CXL	20190307	0.805	0.805	0.800	0.800	96067	Materials
KNM	20190111	0.030	0.030	0.030	0.030	6623	Media & Entertainment
PLS	20190122	0.705	0.710	0.695	0.695	3257813	Materials
STM	20190201	0.033	0.038	0.033	0.036	5571161	Materials
RDM	20190408	0.100	0.100	0.099	0.100	209954	Materials
MPW	20190107	0.150	0.150	0.150	0.150	65544	Software & Services

Univariate density plots of the spread of the data after filtering still showed that the pricing features were skewed, albeit much less. The spread of data for `Volume` was still highly skewed, and so the same method for filtering the pricing features also needed to be applied to `Volume`.

```r
ggplot(ASX_Long_Lower) +
  geom_density(aes(x=Value),
               fill = "yellow",
               alpha = 0.25) +
  scale_x_continuous(labels = comma) +
  scale_y_continuous(labels = comma) +
  facet_rep_wrap(~Variable, repeat.tick.labels = T,
                 scales = "free", ncol = 2) +
  ggtitle("Univariate Density Plots of each Feature") +
  theme_minimal()
```



1.2.8 Filtering Data by Volume

The data was filtered by `ASX_Ticker` to remove the top 1/3 quantile of `Volume`.

```
ASX_Ticker_Summary_Volume <-
  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(Volume),
            "Q1" = quantile(Volume, 0.25),
            "Median" = quantile(Volume, 0.5),
            "Q3" = quantile(Volume, 0.75),
            "90th Percentile" = quantile(Volume, 0.9),
            "95th Percentile" = quantile(Volume, 0.95),
            "Maximum" = max(Volume),
            "Skew" = round(skewness(Volume), 3),
            "Kurtosis" = round(kurtosis(Volume), 3))

ASX_kable <- sample_n(ASX_Ticker_Summary_Volume, 20)

kable_styling(kable(ASX_kable[, 1:7],
                    align = "lrrrrrr"),
              latex_options = c("striped", "hold_position"),
              position = "center",
              full_width = F,
              font_size = 10)
```

ASX_Ticker	n ASX_Tickers	n Observations	Min Date	Max Date	Minimum	Q1
MRR	1	25	04/01/2019	10/04/2019	10600	29705.00
SL1	1	65	02/01/2019	12/04/2019	10	130690.00
TSL	1	61	04/01/2019	12/04/2019	1	200000.00
CD2	1	41	09/01/2019	12/04/2019	1507	5676.00
STN	1	23	24/01/2019	12/04/2019	400	16544.00
ASW	1	18	05/02/2019	11/04/2019	250	1466.25
BPL	1	46	02/01/2019	12/04/2019	3500	20336.50
CBY	1	25	07/03/2019	12/04/2019	1875	11581.00
EML	1	72	02/01/2019	12/04/2019	9973	80078.75
VGL	1	59	04/01/2019	12/04/2019	12	493.00
DDT	1	56	02/01/2019	12/04/2019	10000	200000.00
VGS	1	72	02/01/2019	12/04/2019	11859	26883.00
ACB	1	31	04/01/2019	11/04/2019	410	10217.00
WSA	1	72	02/01/2019	12/04/2019	860658	1188627.00
TTA	1	5	17/01/2019	04/04/2019	12005	12666.00
UTR	1	65	02/01/2019	12/04/2019	4896	272483.00
KNL	1	62	02/01/2019	12/04/2019	2352	24944.00
NWM	1	25	02/01/2019	12/04/2019	115	6667.00
WND	1	62	02/01/2019	12/04/2019	22	1646.00
MTL	1	1	20/02/2019	20/02/2019	1000000	1000000.00

```
kable_styling(kable(ASX_kable[, c(1, 8:14)],
                    align = "lrrrrrrr"),
              latex_options = c("striped", "hold_position"),
              position = "center",
              full_width = F,
              font_size = 10)
```

```
ASX_Lower_Volume <- filter(ASX_Ticker_Summary_Volume,
                           Maximum < quantile(Maximum, 1/3))

ASX_Long_Lower <- filter(ASX_Long_Lower, ASX_Ticker %in% ASX_Lower_Volume$ASX_Ticker)

ASX_Data_Lower <- filter(ASX_Data_Lower, ASX_Ticker %in% ASX_Lower_Volume$ASX_Ticker)

kable_styling(kable(sample_n(ASX_Data_Lower, 20),
                    align = "lrrrrrrl"),
              latex_options = c("striped", "hold_position"),
```

ASX_Ticker	Median	Q3	90th Percentile	95th Percentile	Maximum	Skew	Kurtosis
MRR	51000.0	70000.00	233105.2	277710.00	302630	1.601	1.144
SL1	520000.0	1048665.00	1633864.4	3244205.40	20793003	5.986	39.503
TSL	550000.0	1215592.00	1800000.0	2201263.00	6031793	2.661	10.507
CD2	9926.0	18604.00	29050.0	32646.00	63804	1.741	3.624
STN	34832.0	63347.00	97668.6	136979.20	143763	1.021	0.051
ASW	5495.5	11050.00	23171.0	25537.10	33947	1.227	0.446
BPL	36881.0	127456.25	259638.5	298837.50	517599	1.712	2.464
CBY	35800.0	94200.00	149492.6	285729.80	437216	2.273	4.752
EML	172587.0	514570.50	1130583.6	2594737.45	5381282	3.323	12.329
VGL	3000.0	5412.00	11438.2	21874.00	41208	2.754	8.415
DDT	960152.0	3349024.00	6571764.5	10075649.75	13441443	1.778	2.652
VGS	37977.5	54594.75	82255.2	91716.35	116133	1.164	0.824
ACB	22000.0	71765.50	151800.0	466387.00	1489706	3.941	15.973
WSA	1528275.5	1846020.75	2455506.3	2836346.45	5045356	2.223	6.226
TTA	15006.0	20000.00	26000.0	28000.00	30000	0.673	-1.479
UTR	962514.0	2325507.00	4124242.0	4873468.80	16438885	3.601	17.563
KNL	80417.0	149872.75	247437.0	362703.65	708703	2.370	6.778
NWM	15000.0	25000.00	42237.0	49541.00	72800	1.404	1.611
WND	11009.0	28936.25	122887.3	237550.40	1686647	4.707	22.757
MTL	1000000.0	1000000.00	1000000.0	1000000.00	1000000	NaN	NaN

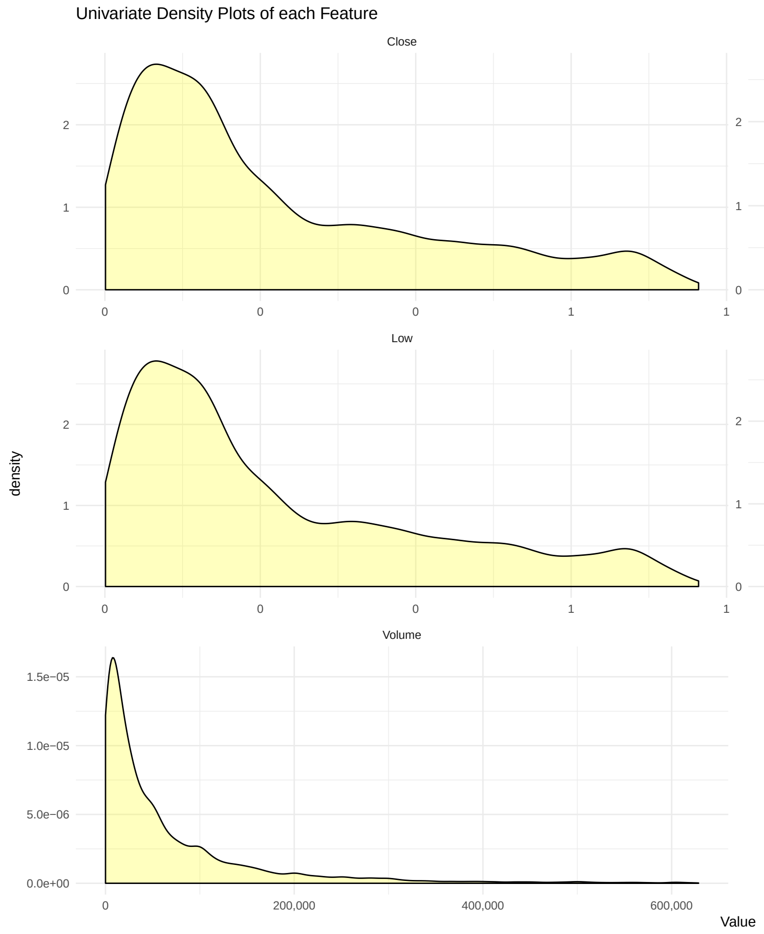
```
             position = "center",
             full_width = F)
```

ASX_Ticker	Date	Open	High	Low	Close	Volume	GICS_industry_group
MPP	20190304	0.545	0.550	0.545	0.550	2000	Capital Goods
FRX	20190405	0.039	0.039	0.039	0.039	723	Communication Services
BPL	20190220	0.020	0.020	0.019	0.019	263040	Materials
RNO	20190308	0.190	0.190	0.190	0.190	256550	Pharmaceuticals, Biotechnology & Life Sciences
MRC	20190409	0.190	0.190	0.190	0.190	5250	Materials
CAM	20190121	0.840	0.850	0.840	0.850	6200	Not Applic
VP7	20190305	0.035	0.035	0.035	0.035	142857	Real Estate
BWF	20190405	0.830	0.855	0.825	0.855	12890	Diversified Financials
CWL	20190218	0.055	0.064	0.055	0.064	16207	Software & Services
S66	20190326	0.595	0.595	0.580	0.580	38000	Household & Personal Products
VMC	20190206	0.145	0.150	0.140	0.140	147535	Materials
IS3	20190411	0.135	0.135	0.135	0.135	3192	Media & Entertainment
NVU	20190329	0.082	0.082	0.080	0.082	160000	Technology Hardware & Equipment
AOU	20190129	0.070	0.075	0.070	0.075	71054	Materials
OVN	20190408	0.255	0.255	0.240	0.240	64502	Health Care Equipment & Services
CAM	20190221	0.835	0.835	0.835	0.835	24904	Not Applic
RTE	20190322	0.430	0.430	0.430	0.430	1163	Consumer Services
TZL	20190328	0.170	0.170	0.165	0.165	8384	Technology Hardware & Equipment
NWF	20190315	0.155	0.160	0.150	0.150	108126	Materials
GEV	20190219	0.160	0.165	0.155	0.155	96833	Energy

1.2.9 Density Plots After Filtering by Price and Volume

After removing extreme values in the `High` and `Volume` feature, univariate density plots were still right skewed but much less extreme.

```
ggplot(ASX_Long_Lower) +
  geom_density(aes(x=Value),
               fill = "yellow",
               alpha = 0.25) +
  scale_x_continuous(labels=comma) +
  facet_rep_wrap(~Variable, repeat.tick.labels = T,
                 scales = "free", ncol = 2) +
  ggtitle("Univariate Density Plots of each Feature") +
  theme_minimal()
```



1.2.10 Summary Statistics of Data After Removing Extreme ASX_Tickers

After filtering by price (`High`) and `Volume`, each of the price features were much less skewed; all below 1.0. `Volume` was still somewhat skewed, but further filtering the data based on this feature might risk the accuracy of the model in Phase 2. The skew for `Volume` before filtering was 33.523, whereas after filtering was 2.658.

```
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" = format(round(min(Value), 2),
                                                  big.mark = ","),
                               "Q1" = format(round(quantile(Value, 0.25), 3),
                                                  big.mark = ","),
                               "Median" = format(round(quantile(Value, 0.5), 3),
                                                  big.mark = ","),
                               "Q3" = format(round(quantile(Value, 0.75), 3),
                                                  big.mark = ","),
                               "90th Percentile" = format(round(quantile(Value, 0.9), 3),
                                                          big.mark = ","),
                               "95th Percentile" = format(round(quantile(Value, 0.95), 3),
                                                          big.mark = ","),
                               "Maximum" = format(round(max(Value), 3),
                                                  big.mark = ","),
                               "Skew" = round(skewness(Value), 3),
                               "Kurtosis" = round(kurtosis(Value), 2))

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

Variable	Close	High	Low	Open	Volume
n ASX_Tickers	393	393	393	393	393
n Observations	12,400	12,400	12,400	12,400	12,400
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	0.093	0.095	0.091	0.092	10,000
Median	0.19	0.19	0.185	0.19	31,466
Q3	0.435	0.44	0.43	0.435	84,516.5
90th Percentile	0.68	0.685	0.675	0.68	172,462.5
95th Percentile	0.81	0.815	0.805	0.81	256,482.5
Maximum	0.955	0.955	0.955	0.955	628,543
Skew	0.982	0.969	0.993	0.979	2.658
Kurtosis	-0.11	-0.15	-0.09	-0.12	8.61

1.3 Data Exploration and Visualisation

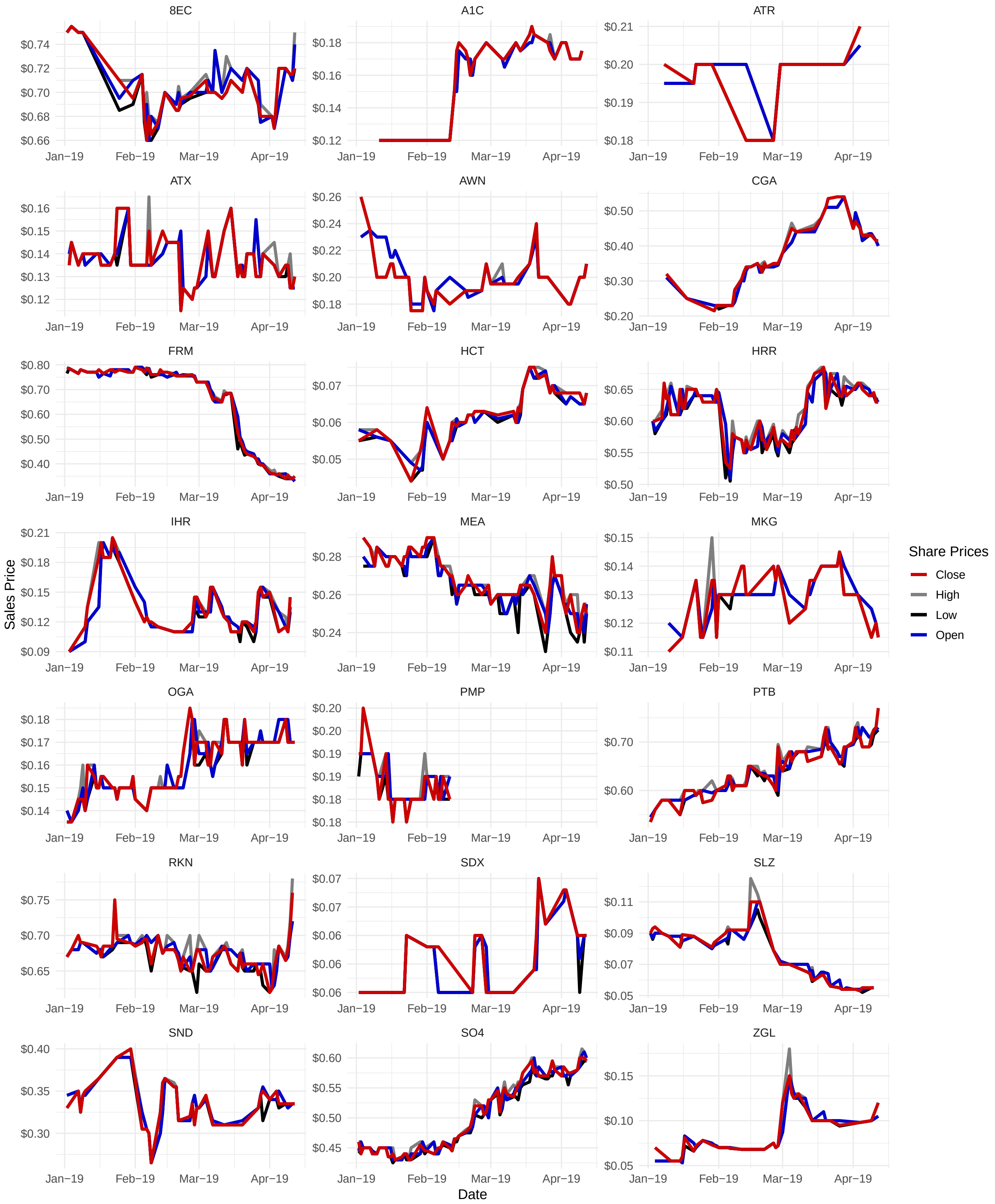
1.3.1 Share Price Tracking

The visualisations below of share prices for 21 randomly[2] selected stocks did not reveal any consistent trends or abnormalities. Each of the below stocks appeared to resemble normal pricing behaviour for share prices. All four pricing variables (`Open`, `Low`, `High`, `Close`) all appeared to be very highly correlated, but with an estimated correlation of $r \neq 1$.

```r
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 = 21)

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_y_continuous("Sales Price",
                     labels = dollar) +
  scale_color_manual(name = "Share Prices",
                     values = c("Open"="blue3",
                                "High"="grey50",
                                "Low"="black",
                                "Close"="red3")) +
  labs(title = "Sales Prices of 21 Shares from 02-01-2019 to 12-04-2019",
       caption = "Please note y-axes are not restricted to start at 0") +
  facet_rep_wrap(~ASX_Ticker, repeat.tick.labels = T,
                 scales = "free_y", ncol = 3) +
  theme_minimal() +
  theme(text = element_text(size = 12))
```

[2] pseudo-random; from a uniform distribution and not a truly random selection.

Sales Prices of 21 Shares from 02−01−2019 to 12−04−2019



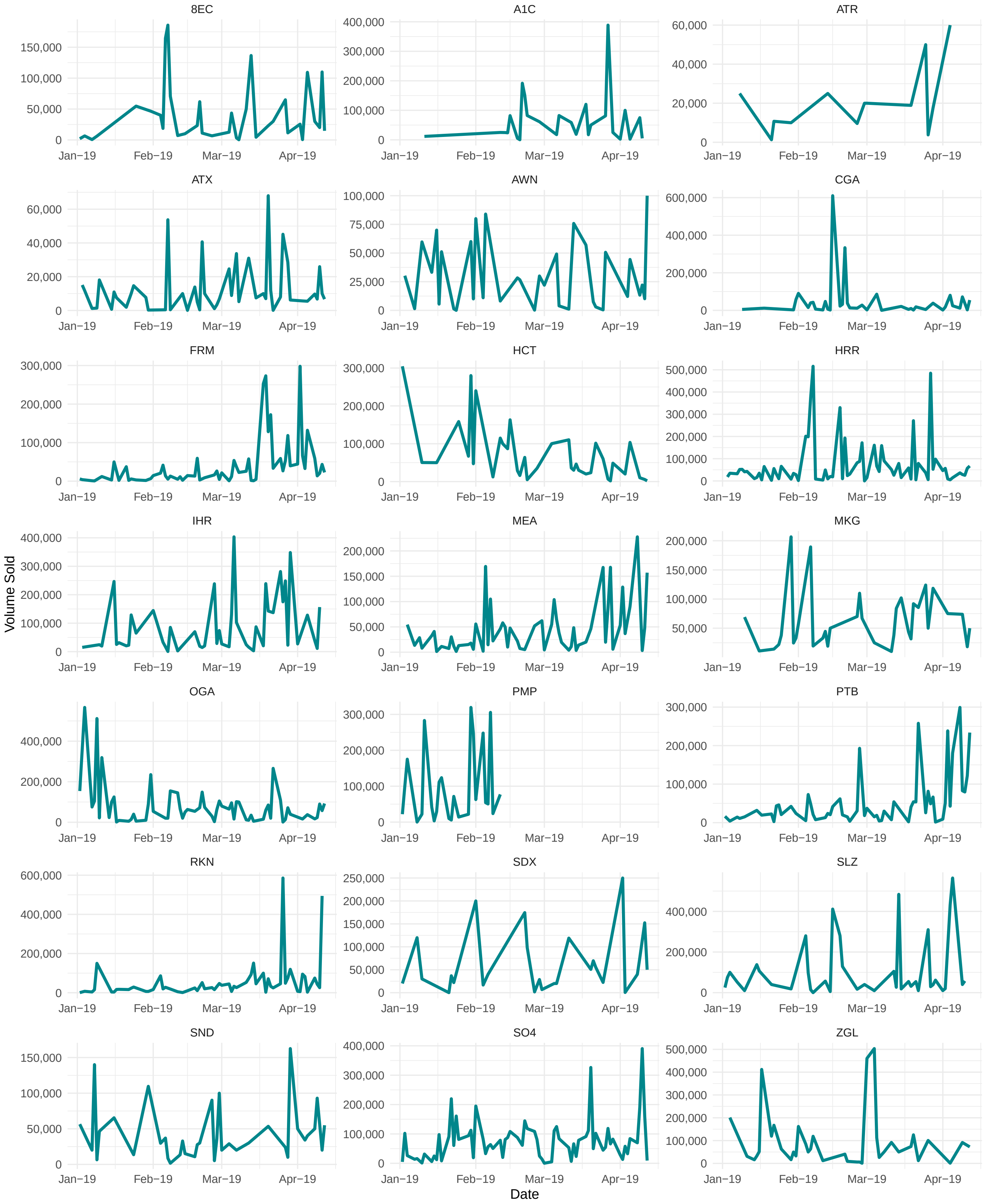
Please note y−axes are not restricted to start at 0

1.3.2 Volume of Shares Sold

The below visualisation of the volume of stocks sold from same 21 shares was quite different to the price features. The volumes of stocks sold appeared to be highly variable and erratic, with large spikes breaking up long periods of low selling days to weeks. This seems to suggest that the buying and selling nature of stocks does not have a strong correlation with any of the pricing variables.

```r
ggplot(ASX_Data_Samples) +
  geom_line(aes(x=Date, y=Volume),
            size=1.25, col = "turquoise4") +
  scale_x_date(date_breaks = "month", date_labels = "%b-%y") +
  scale_y_continuous("Volume Sold",
                     labels = comma)+
  ggtitle("Volume of Stock Sold of 21 Shares from 02-01-2019 to 12-04-2019") +
  facet_rep_wrap(~ASX_Ticker, repeat.tick.labels = T,
                 scales = "free_y", ncol = 3) +
  theme_minimal() +
  theme(text = element_text(size = 12))
```

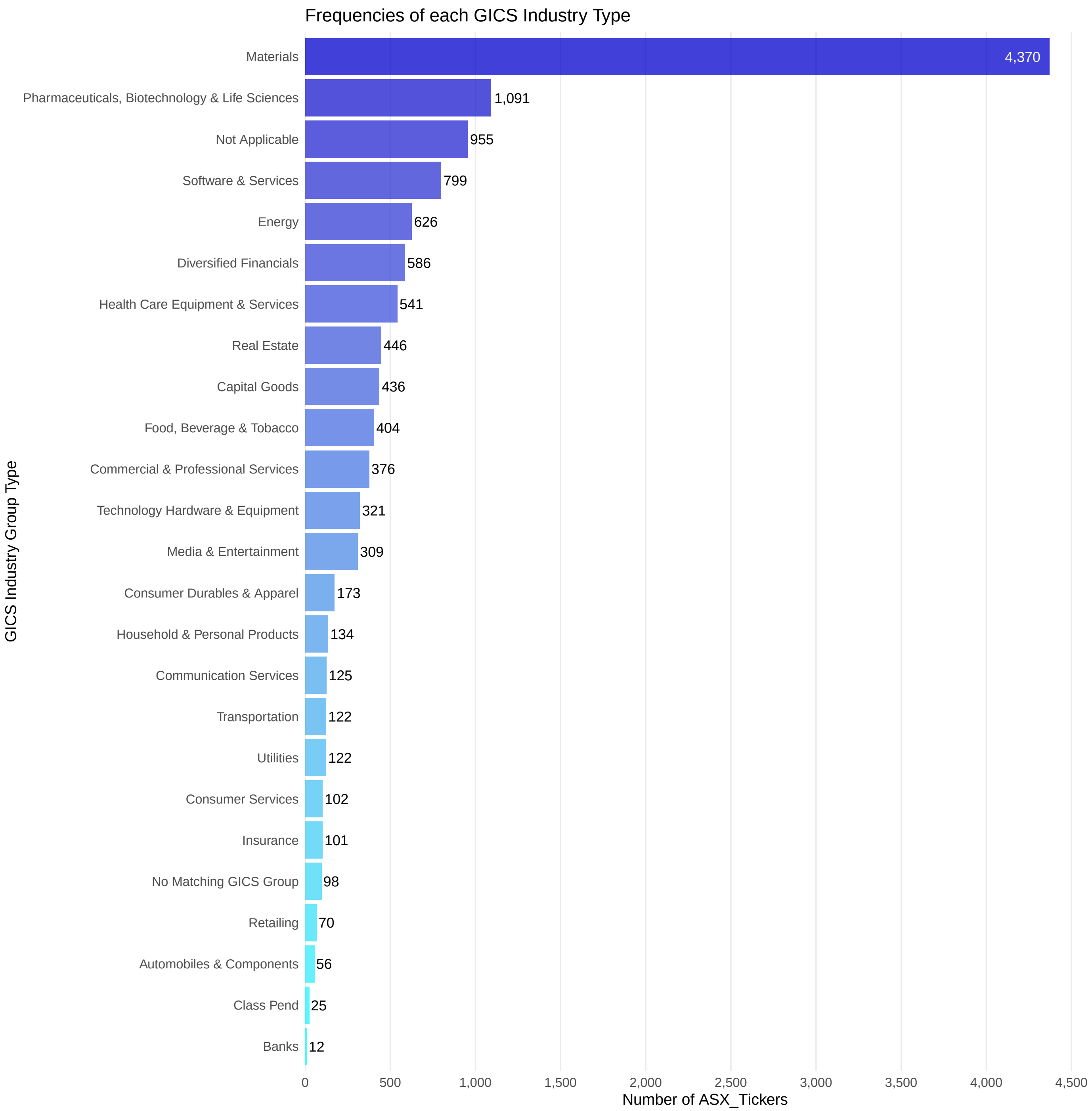
Volume of Stock Sold of 21 Shares from 02−01−2019 to 12−04−2019



1.3.3 Number of Companies per GICS Group

The Materials industry group was the most frequently occurring GICS grouping in the dataset with 4,370 different ASX_Tickers. This was nearly four-times the size of the second-most frequently occurring GICS grouping; Pharmaceuticals, Biotechnology & Life Sciences with 1,091 different ASX_Tickers.

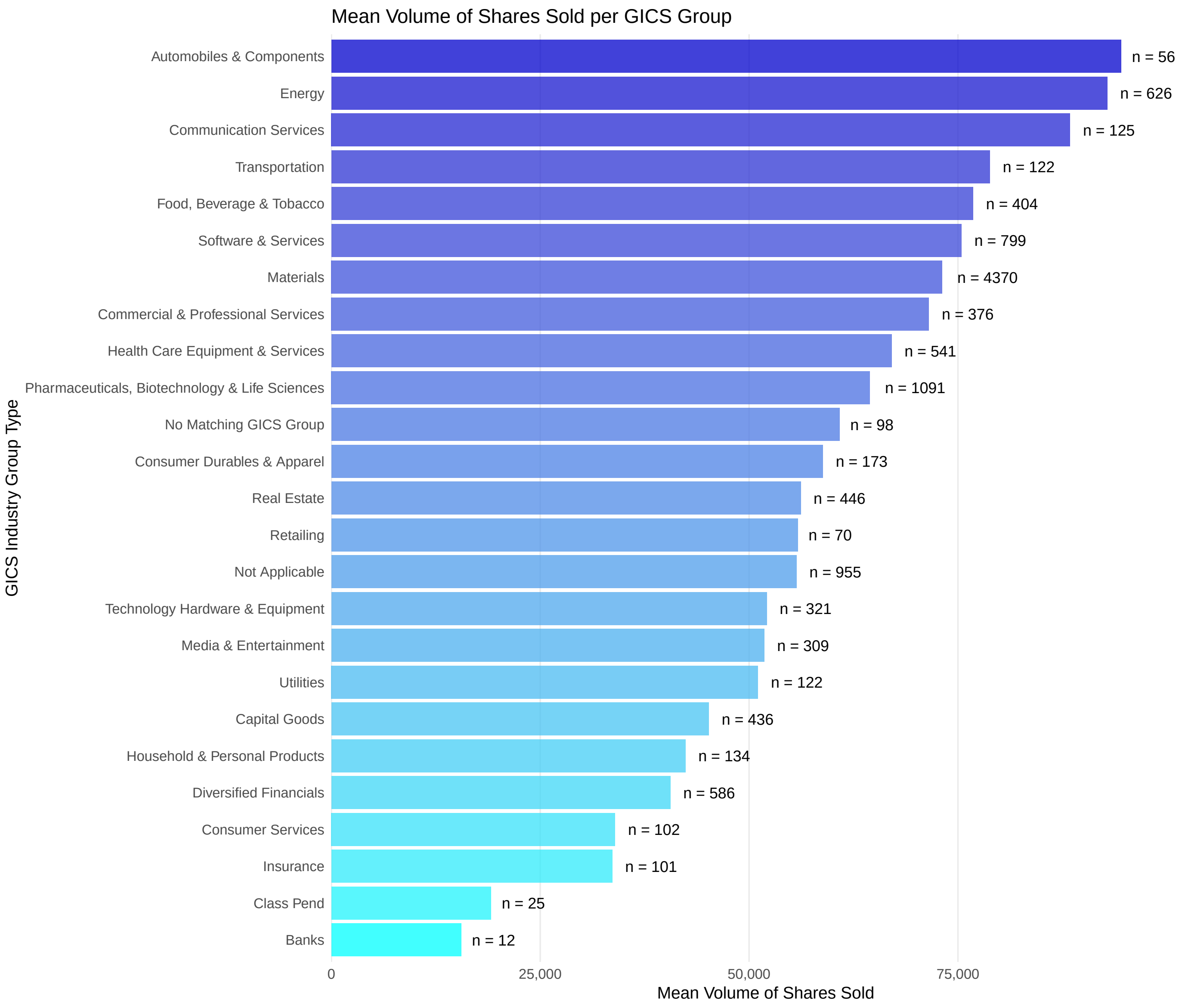
```r
ASX_Data_Lower$GICS_industry_group <- recode(ASX_Data_Lower$GICS_industry_group,
                                             "Not Applic"="Not Applicable")

ASX_Data_Lower$GICS_industry_group[is.na(
  ASX_Data_Lower$GICS_industry_group)] <-
  "No Matching GICS Group"

ASX_Data_Lower$GICS_industry_group[ASX_Data_Lower$GICS_industry_group == "NA"] <-
  "No Matching GICS Group"

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

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, max(ASX_Data_Count$Count)*1.075,
                                  by = 500),
                     limits = c(0, max(ASX_Data_Count$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)) +
  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())
```



1.3.4 Mean Volumes Sold by GICS Groups

The below plot shows that, after some filtering, the mean volume of shares sold is very similar between GICS industry groups.

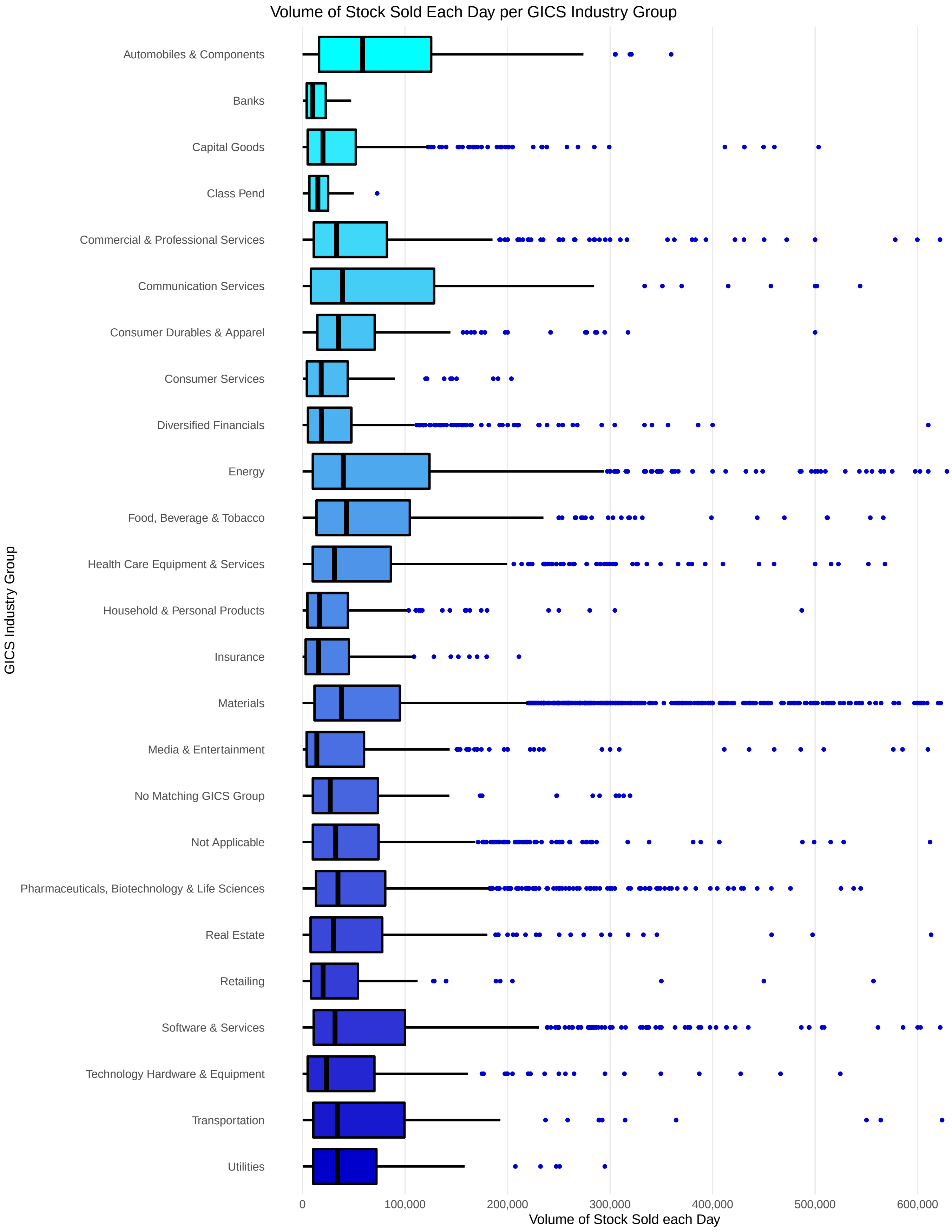
```r
ASX_Lower_Vol <- summarise(group_by(ASX_Data_Lower,
                                    GICS_industry_group),
                           Mean_Vol = mean(Volume),
                           n_Companies = n())

ASX_Lower_Vol$GICS_industry_group <- factor(ASX_Lower_Vol$GICS_industry_group,
                                            levels = ASX_Lower_Vol$GICS_industry_group[
                                              order(ASX_Lower_Vol$Mean_Vol)])

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

ggplot(ASX_Lower_Vol) +
  geom_bar(aes(x = GICS_industry_group, y = Mean_Vol,
               fill = GICS_industry_group),
           stat = "identity", show.legend = F,
           alpha = 0.75) +
  geom_text(aes(x = GICS_industry_group,
                y = Mean_Vol,
                label = paste("n =",
                              n_Companies)),
            hjust=-0.25) +
  scale_y_continuous(breaks = seq(0,max(ASX_Lower_Vol$Mean_Vol), 25000),
                     limits = c(0,max(ASX_Lower_Vol$Mean_Vol)*1.1),
                     expand = c(0,0),
                     labels = comma,
                     "Mean Volume of Shares Sold") +
  scale_x_discrete("GICS Industry Group Type") +
  ggtitle("Mean Volume of Shares Sold per GICS Group") +
  scale_fill_manual(values = fill_grad) +
  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())
```


Mean Volume of Shares Sold per GICS Group
GICS Industry Group Type
Automobiles & Components
n = 56
Energy
n = 626
Communication Services
n = 125
Transportation
n = 122
Food, Beverage & Tobacco
n = 404
Software & Services
n = 799
Materials
n = 4370
Commercial & Professional Services
n = 376
Health Care Equipment & Services
n = 541
Pharmaceuticals, Biotechnology & Life Sciences
n = 1091
No Matching GICS Group
n = 98
Consumer Durables & Apparel
n = 173
Real Estate
n = 446
Retailing
n = 70
Not Applicable
n = 955
Technology Hardware & Equipment
n = 321
Media & Entertainment
n = 309
Utilities
n = 122
Capital Goods
n = 436
Household & Personal Products
n = 134
Diversified Financials
n = 586
Consumer Services
n = 102
Insurance
n = 101
Class Pend
n = 25
Banks
n = 12
0
25,000
50,000
75,000
Mean Volume of Shares Sold

1.3.5 Volumes Sold of each GICS per Day

To further explore the spread of the data, the volumes sold of shares within each GICS was visualised as boxplots for the total time period in the dataset. These boxplots below showed that, despite the dataset being right-skewed, that the skew is present across most GICS groups.

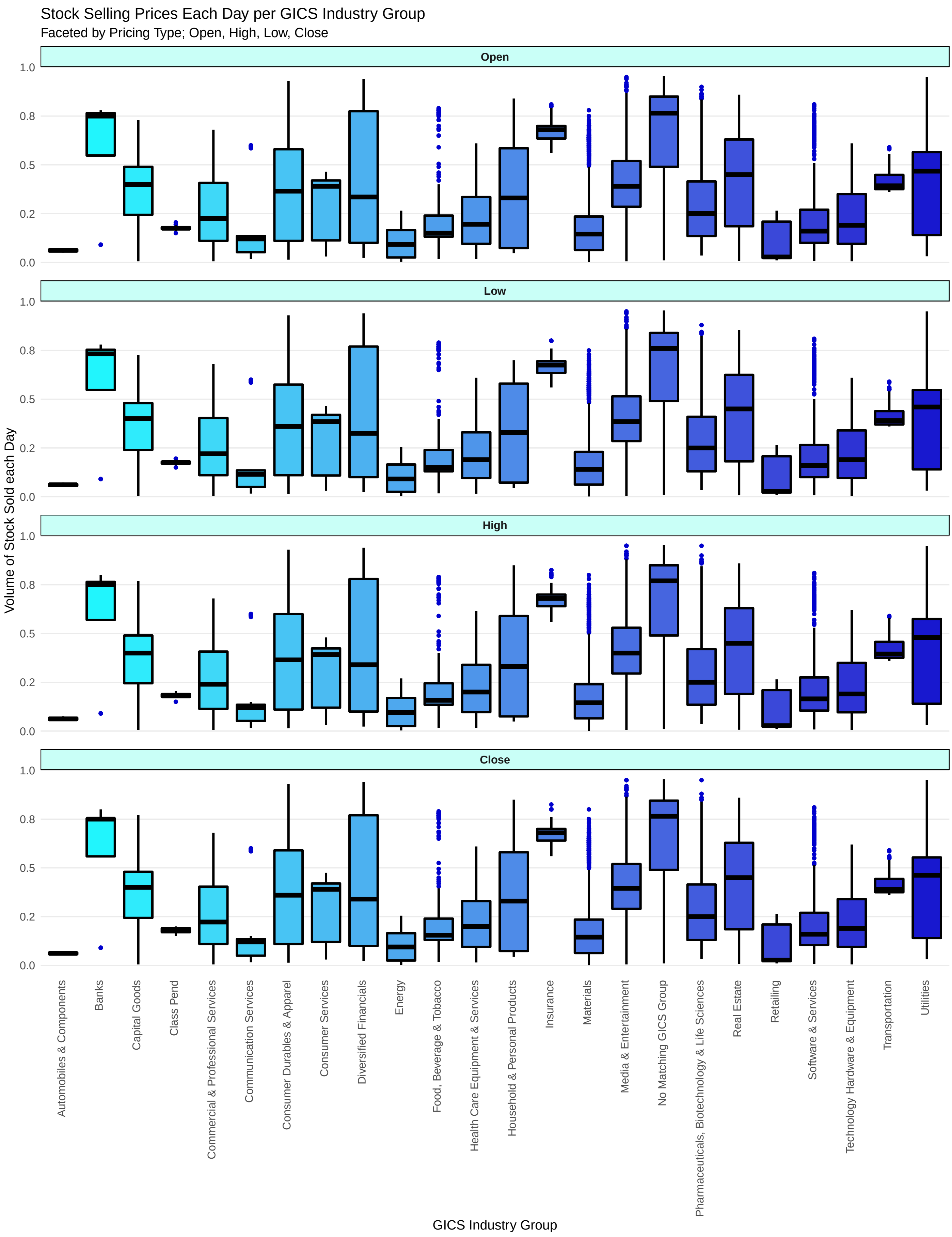
```
ggplot(ASX_Data_Lower) +
  geom_boxplot(aes(x = fct_rev(GICS_industry_group), y = Volume,
                   fill = GICS_industry_group),
               show.legend = F, col = "black",
               size = 1,
               outlier.size = 1.25,
               outlier.colour = "blue3") +
  scale_x_discrete("GICS Industry Group") +
  scale_y_continuous("Volume of Stock Sold each Day",
                     labels = comma,
                     breaks = seq(0, max(ASX_Data_Lower$Volume),
                                  100000)) +
  scale_fill_manual(values = fill_grad) +
  labs(title = "Volume of Stock Sold Each Day per GICS Industry Group") +
  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())
```


Volume of Stock Sold Each Day per GICS Industry Group
GICS Industry Group
Automobiles & Components
Banks
Capital Goods
Class Pend
Commercial & Professional Services
Communication Services
Consumer Durables & Apparel
Consumer Services
Diversified Financials
Energy
Food, Beverage & Tobacco
Health Care Equipment & Services
Household & Personal Products
Insurance
Materials
Media & Entertainment
No Matching GICS Group
Not Applicable
Pharmaceuticals, Biotechnology & Life Sciences
Real Estate
Retailing
Software & Services
Technology Hardware & Equipment
Transportation
Utilities
0
100,000
200,000
300,000
400,000
500,000
600,000
Volume of Stock Sold each Day

1.3.6 Pricing Features for Each GICS Group

Boxplots were generated for each Pricing Feature for each GICS group. Just like with the boxplots for `Volume` above, this visualisation showed the spread of each of the Pricing descriptive features over the total time period collected. Unlike the `Volume` boxplots above, the Pricing features showed less skew within GICS group and less similarity between groups.

```r
ASX_Long_Lower$GICS_industry_group[is.na(ASX_Long_Lower$GICS_industry_group)] <-
  "No Matching GICS Group"

ASX_Long_Lower$GICS_industry_group[ASX_Long_Lower$GICS_industry_group ==
                                     "Not Applic"] <- "No Matching GICS Group"

ggplot(filter(ASX_Long_Lower, Variable != "Volume")) +
  geom_boxplot(aes(x = GICS_industry_group, y = Value,
                   fill = GICS_industry_group),
               show.legend = F, col = "black",
               size = 1,
               outlier.size = 1.25,
               outlier.colour = "blue3") +
  facet_rep_wrap(~fct_rev(Variable), scales = "free_y",
                 ncol = 1, repeat.tick.labels = "y") +
  scale_x_discrete("GICS Industry Group") +
  scale_y_continuous("Volume of Stock Sold each Day",
                     labels = comma_format(accuracy = 0.1)) +
  scale_fill_manual(values = fill_grad) +
  labs(title = "Stock Selling Prices Each Day per GICS Industry Group",
       subtitle = "Faceted by Pricing Type; Open, High, Low, Close") +
  theme_minimal() +
  theme(panel.grid.minor.x = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.y = element_blank(),
        axis.text.x = element_text(angle = 90,
                                   hjust = 1, vjust = 0.25),
        text = element_text(size = 12),
        panel.border = element_blank(),
        strip.background = element_rect(fill = "#c9fff7"),
        strip.text = element_text(face = "bold"))
```

Stock Selling Prices Each Day per GICS Industry Group
Faceted by Pricing Type; Open, High, Low, Close
Open
Low
High
Close
1.0
0.8
0.5
0.2
0.0
Volume of Stock Sold each Day
Automobiles & Components
Banks
Capital Goods
Class Pend
Commercial & Professional Services
Communication Services
Consumer Durables & Apparel
Consumer Services
Diversified Financials
Energy
Food, Beverage & Tobacco
Health Care Equipment & Services
Household & Personal Products
Insurance
Materials
Media & Entertainment
No Matching GICS Group
Pharmaceuticals, Biotechnology & Life Sciences
Real Estate
Retailing
Software & Services
Technology Hardware & Equipment
Transportation
Utilities
GICS Industry Group

1.4 Summary

After compiling the data, it was observed to be heavily skewed for all continuous descriptive features. Price and Volume features were used to filter ASX Tickers to remove extreme values that were causing the right-skew. The dataset remaining was still right-skewed, but to a much lesser extent.

GICS Industry Group was added to the dataset, which included a descriptive feature `Company_name`. Company name was deemed to provide no information gain as each `ASX_Ticker` was linked to a unique Company name, and so Company Name was removed.

Several visualisations, both univariate and multivariate, were produced that explored the nature of the data. Univariate density plots were produced to show the spread of the descriptive features before and after filtering extreme values. Time series line plots were also produced to investigate the behaviour of pricing features and the sales volume feature. GICS was also explored by frequency of each group and mean volume sold per group. The spread of the data was also explored by GICS group for all continuous descriptive features and for the target feature Volume.

1.4.1 References

1. *ASX Historical Data*, ASXHistoricalData.com, viewed 19 April 2019, <https://www.asxhistoricaldata.com>

2. Australian Securities Exchange (ASX), *GICS*, viewed 22 April, 2019, <https://www.asx.com.au/products/gics.htm>