Closing Prices of Australian Stock Exchange Shares

MATH2319 - Machine Learning
Course Project

Ben Cole - s3412349

Print Date: 27/04/2019

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 and with extremely high 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 ASX Historical Data website.

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.

```
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") )
```

```r
  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)

}
```

```r
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
KFE	20190114	0.094	0.120	0.094	0.120	777570
VGS	20190115	65.490	65.790	65.400	65.770	24936
GOW	20190204	2.530	2.530	2.530	2.530	2600
DXS	20190121	11.100	11.160	11.030	11.110	2369101
TAO	20190117	0.075	0.075	0.075	0.075	15000
DTL	20190305	1.610	1.625	1.605	1.610	895520
RIO	20190110	80.250	80.770	79.460	80.130	1174681
ADH	20190130	1.840	1.910	1.800	1.905	279298
GC1	20190111	0.930	0.940	0.930	0.940	18219
TME	20190116	6.000	6.005	5.980	6.000	823154
LPD	20190214	0.017	0.018	0.017	0.018	14321658
MND	20190109	13.860	14.380	13.860	14.310	177745
NXE	20190115	0.050	0.051	0.048	0.048	25500
MCP	20190102	1.245	1.250	1.220	1.220	15459
UBN	20190103	0.040	0.040	0.040	0.040	118756
XEJ	20190115	10320.700	10465.200	10317.300	10456.200	0
ATU	20190131	0.100	0.100	0.100	0.100	365646
RAN	20190102	0.030	0.030	0.027	0.027	160333
BEM	20190118	0.079	0.080	0.075	0.075	180000
AMI	20190117	0.790	0.790	0.775	0.790	1894448

```r
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 as well. 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
RES	Energy
AIZ	Transportation
PDZ	Materials
CD2	Not Applic
ORG	Energy
PLC	Diversified Financials
CDD	Capital Goods
GGX	Energy
NAB	Banks
SOP	Commercial & Professional Services
AO1	Software & Services
NIU	Materials
TNF	Not Applic
SS6	Not Applic
KSC	Transportation
POH	Pharmaceuticals, Biotechnology & Life Sciences
JCS	Software & Services
NIO	Materials
ADX	Energy
GBT	Software & Services

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 data set 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 data set 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
WIC	20190110	1.045	1.050	1.045	1.050	52209	Not Applic
TZN	20190305	0.085	0.085	0.078	0.078	632905	Materials
SGP	20190411	3.850	3.870	3.840	3.850	3646215	Real Estate
CIN	20190118	31.400	31.400	31.400	31.400	500	Not Applic
SGC	20190307	0.025	0.025	0.025	0.025	11192	Energy
PME	20190409	16.200	16.690	16.200	16.300	85817	Health Care Equipment & Services
PEN	20190403	0.315	0.315	0.310	0.310	182753	Energy
ANZ	20190211	26.890	26.930	26.300	26.540	5006572	Banks
BSM	20190401	0.012	0.013	0.012	0.013	5414419	Materials
SSM	20190107	1.710	1.740	1.700	1.730	374306	Capital Goods
XIP	20190110	1.315	1.315	1.310	1.315	14156	Commercial & Professional Services
QVE	20190107	1.045	1.065	1.040	1.065	156713	Not Applic
MAT	20190227	0.150	0.150	0.150	0.150	98160	Materials
A40	20190130	0.180	0.185	0.170	0.175	851772	Materials
JHX	20190116	14.940	15.010	14.750	14.870	1534668	Materials
CRL	20190109	0.027	0.027	0.027	0.027	140695	Materials
PPH	20190116	3.160	3.170	3.130	3.150	535871	Software & Services
API	20190409	1.560	1.560	1.540	1.550	601309	Health Care Equipment & Services
MSV	20190111	0.041	0.041	0.040	0.040	645600	Materials
OGA	20190225	0.165	0.185	0.165	0.185	30000	Food, Beverage & Tobacco

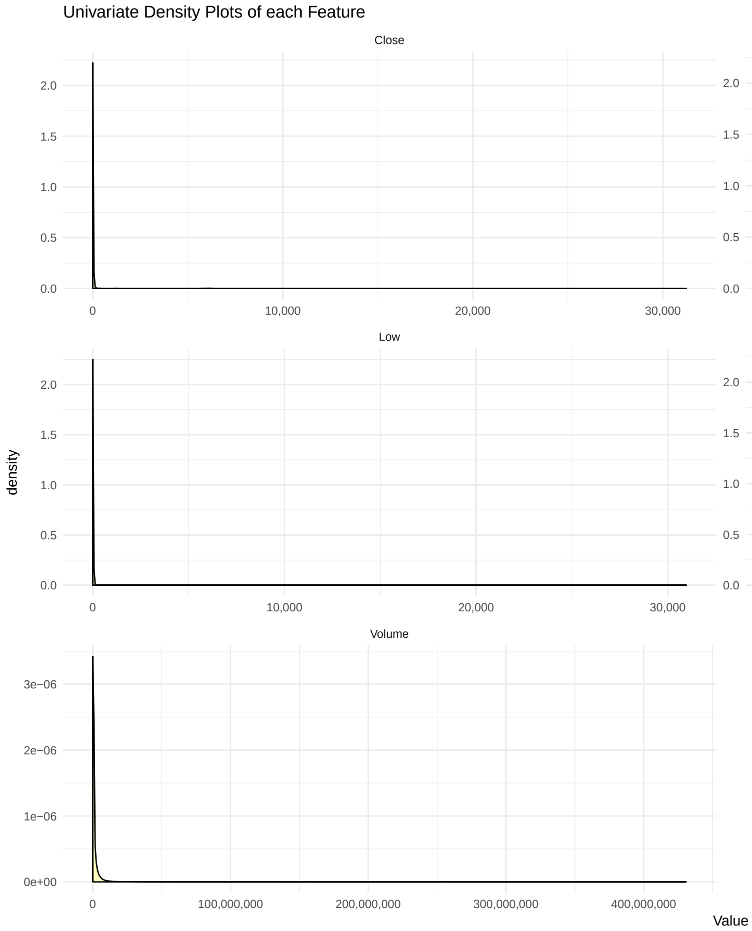
```r
                         "Min Date" = format(ymd(min(Date)), "%d-%m-%Y"),
                         "Max Date" = format(ymd(max(Date)), "%d-%m-%Y"),
                         "Minimum" = format(round(min(Value), 3),
                                            big.mark = ","),
                         "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 1/3rd 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.

```
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
PDZ	1	54	02/01/2019	12/04/2019	0.285	0.33000
EM1	1	68	02/01/2019	12/04/2019	0.006	0.00700
3DP	1	70	02/01/2019	12/04/2019	0.038	0.04300
JAN	1	56	02/01/2019	12/04/2019	0.305	0.34875
CL1	1	72	02/01/2019	12/04/2019	1.305	1.47875
IGO	1	72	02/01/2019	12/04/2019	3.720	4.21125
AEF	1	72	02/01/2019	12/04/2019	1.665	1.74000
WLL	1	62	02/01/2019	12/04/2019	4.850	5.00000
QSS	1	4	06/02/2019	14/02/2019	0.022	0.02350
HPR	1	31	11/01/2019	12/04/2019	0.056	0.06400
EML	1	72	02/01/2019	12/04/2019	1.400	1.47250
LOM	1	72	02/01/2019	12/04/2019	0.170	0.18500
CDA	1	72	02/01/2019	12/04/2019	2.890	3.10000
CLB	1	34	19/02/2019	12/04/2019	0.140	0.15500
GNX	1	71	02/01/2019	12/04/2019	0.225	0.26000
RDS	1	33	02/01/2019	11/04/2019	0.012	0.01300
EMP	1	30	02/01/2019	12/04/2019	0.002	0.00225
AOU	1	35	10/01/2019	12/04/2019	0.058	0.06200
AKP	1	72	02/01/2019	12/04/2019	18.100	20.45000
ISU	1	72	02/01/2019	12/04/2019	0.625	0.71925

```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)
```

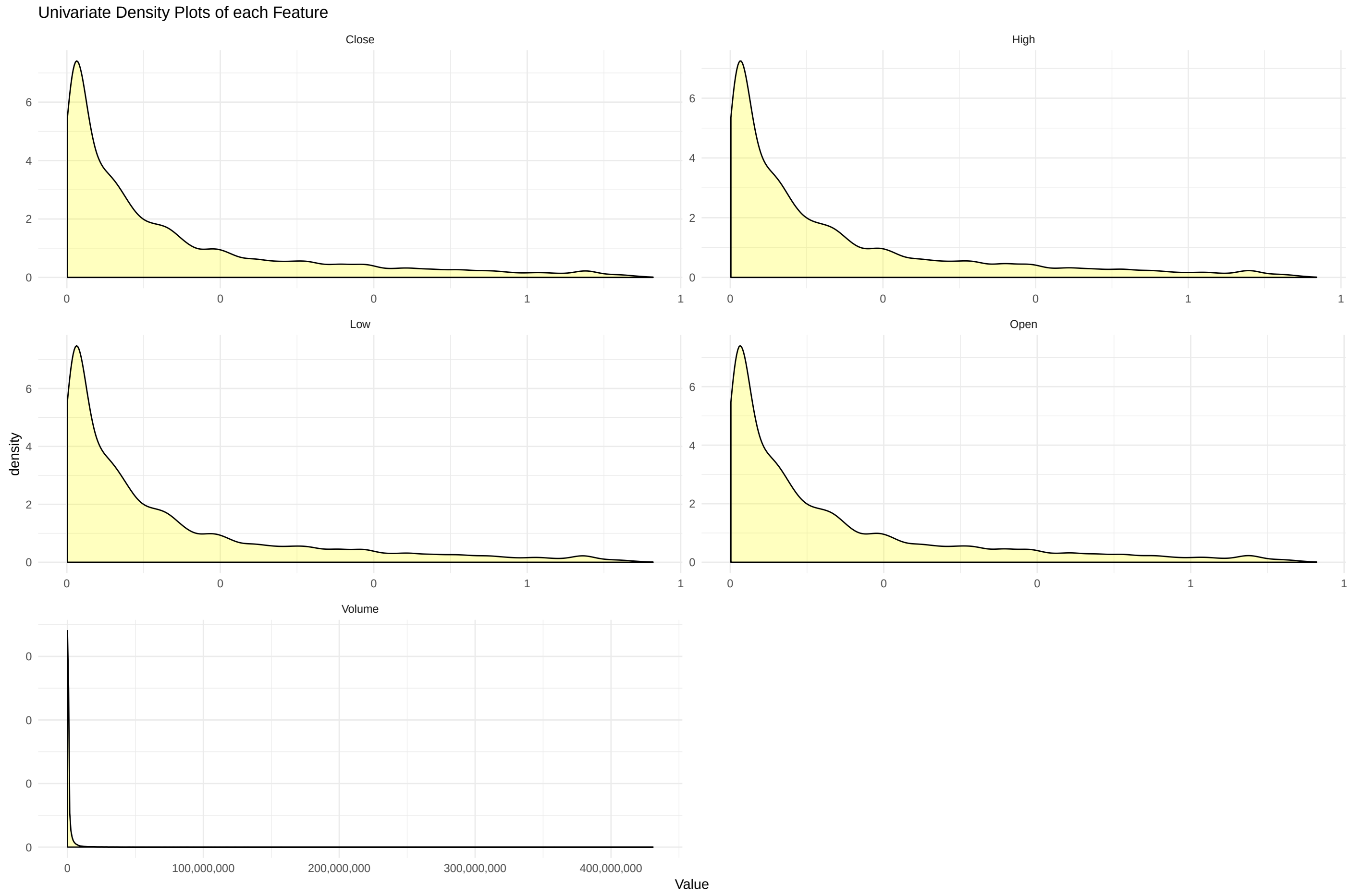
```r
ASX_Lower <- filter(ASX_Ticker_Summary_Price, Maximum < quantile(Maximum, 2/3))

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
PDZ	0.3600	0.38375	0.4000	0.41175	0.420	-0.146	-0.908
EM1	0.0175	0.03000	0.0360	0.03700	0.045	0.514	-1.182
3DP	0.0450	0.04800	0.0530	0.05555	0.062	0.848	0.416
JAN	0.3700	0.38000	0.3925	0.40000	0.410	-0.522	-0.762
CL1	1.5100	1.68000	1.7400	1.75000	1.790	0.041	-1.088
IGO	4.8125	4.93000	4.9835	5.01725	5.130	-0.820	-0.834
AEF	1.7885	2.25500	2.3990	2.53450	2.870	0.977	-0.340
WLL	5.0450	5.11750	5.2000	5.28800	5.360	0.366	-0.261
QSS	0.0240	0.02400	0.0240	0.02400	0.024	-0.750	-1.687
HPR	0.0650	0.06900	0.0700	0.07000	0.070	-0.561	0.645
EML	1.5800	1.77550	1.8195	1.84225	1.865	0.105	-1.688
LOM	0.1900	0.19500	0.2050	0.20500	0.210	-0.139	-0.365
CDA	3.1425	3.20250	3.2500	3.28450	3.350	-0.559	-0.125
CLB	0.1650	0.17750	0.2035	0.21725	0.250	1.482	2.042
GNX	0.2650	0.27500	0.2750	0.28000	0.300	-0.648	1.991
RDS	0.0150	0.01700	0.0170	0.01700	0.018	-0.226	-1.435
EMP	0.0030	0.00300	0.0030	0.00300	0.004	-0.381	-0.345
AOU	0.0650	0.07000	0.0726	0.07500	0.075	0.187	-1.335
AKP	21.3500	22.20500	23.0820	23.34500	24.750	-0.142	-0.502
ISU	0.7500	0.78000	0.7900	0.79725	0.805	-0.871	-0.156

ASX_Ticker	Date	Open	High	Low	Close	Volume	GICS_industry_group
SHM	20190111	0.580	0.580	0.560	0.565	66027	Consumer Durables & Apparel
P2P	20190312	0.410	0.435	0.400	0.420	187708	Transportation
ARE	20190214	0.026	0.026	0.026	0.026	1651840	Materials
HWK	20190311	0.019	0.019	0.019	0.019	2256	Materials
ASW	20190306	0.680	0.680	0.680	0.680	5000	Diversified Financials
MEC	20190107	0.850	0.870	0.850	0.870	43110	Diversified Financials
BOL	20190129	0.165	0.165	0.165	0.165	30000	Capital Goods
SWM	20190301	0.550	0.550	0.525	0.525	4694259	Media & Entertainment
OAR	20190201	0.023	0.023	0.023	0.023	150000	Materials
IVZ	20190110	0.040	0.046	0.040	0.044	979450	Energy
ACL	20190306	0.009	0.010	0.009	0.010	291836	Pharmaceuticals, Biotechnology & Life Sciences
MX1	20190221	0.225	0.230	0.220	0.230	135094	Health Care Equipment & Services
RNU	20190221	0.018	0.019	0.018	0.018	2735974	Energy
ABT	20190311	0.120	0.120	0.120	0.120	6000	Food, Beverage & Tobacco
TAU	20190325	0.180	0.180	0.180	0.180	50000	Consumer Services
ATX	20190117	0.140	0.140	0.135	0.135	7550	Pharmaceuticals, Biotechnology & Life Sciences
PCH	20190212	0.002	0.002	0.002	0.002	250000	Media & Entertainment
SIL	20190116	0.280	0.280	0.275	0.280	69000	Health Care Equipment & Services
EMH	20190408	0.330	0.360	0.330	0.360	86806	Materials
ABX	20190220	0.105	0.105	0.105	0.105	5431	Materials

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
APD	1	61	02/01/2019	11/04/2019	573	30942.00
JJF	1	33	02/01/2019	09/04/2019	6500	30000.00
WSI	1	69	02/01/2019	12/04/2019	6666	358571.00
TGP	1	64	02/01/2019	12/04/2019	460	18765.00
MTM	1	23	31/01/2019	10/04/2019	638	10631.00
TDO	1	43	02/01/2019	11/04/2019	2018	29723.00
ONE	1	44	03/01/2019	12/04/2019	100	1550.50
DWS	1	71	02/01/2019	12/04/2019	6600	34649.00
HCT	1	36	02/01/2019	12/04/2019	1796	20000.00
NML	1	70	02/01/2019	12/04/2019	18996	305063.50
XSJ	1	72	02/01/2019	12/04/2019	0	0.00
ABT	1	39	07/01/2019	12/04/2019	675	9850.00
QUS	1	63	03/01/2019	12/04/2019	3	665.50
ENX	1	16	07/01/2019	12/04/2019	142	4479.25
RCT	1	50	02/01/2019	12/04/2019	75	990.25
OCC	1	70	02/01/2019	12/04/2019	3356	12215.75
SHM	1	70	02/01/2019	12/04/2019	1500	27667.50
QRE	1	70	02/01/2019	12/04/2019	92	5700.25
RRS	1	15	11/01/2019	13/03/2019	250000	500000.00
FPH	1	72	02/01/2019	12/04/2019	164548	246026.25

```
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 = "lrrrrrl"),
              latex_options = c("striped", "hold_position"),
```

ASX_Ticker	Median	Q3	90th Percentile	95th Percentile	Maximum	Skew	Kurtosis
APD	79565.0	166581.00	331950.0	447344.00	2947380	5.498	33.253
JJF	71667.0	136045.00	193845.6	238541.40	281981	0.958	0.076
WSI	900626.0	2417055.00	4060767.8	9175767.60	19165091	3.090	9.956
TGP	37013.0	70820.25	99396.4	165415.45	693956	3.958	15.796
MTM	35000.0	263269.50	780896.2	821575.70	2207635	2.552	6.694
TDO	97982.0	160728.50	269132.4	376983.80	563972	1.674	2.597
ONE	4274.5	16991.50	40198.4	69414.25	242492	4.180	19.394
DWS	60485.0	90063.00	125344.0	161843.00	406107	2.608	8.934
HCT	47923.5	100584.00	160911.5	250027.75	304689	1.612	2.019
NML	676551.0	1075018.25	2416633.8	2633328.55	3194578	1.088	0.017
XSJ	0.0	0.00	0.0	0.00	0	NaN	NaN
ABT	22500.0	42072.00	146996.8	309023.80	1433421	4.413	20.224
QUS	1216.0	2581.00	5020.4	6680.40	42105	4.605	21.216
ENX	10447.0	16350.50	29349.0	69319.75	160612	2.961	7.907
RCT	2991.5	5491.50	10203.2	15623.45	404464	6.590	42.562
OCC	28617.0	65634.50	187817.5	274685.15	705657	3.352	13.193
SHM	45554.5	79290.50	130459.0	207063.70	294769	2.055	4.192
QRE	11020.0	20695.75	31719.6	60810.30	497021	6.869	50.307
RRS	2600000.0	4132974.50	5977632.0	6930000.00	7700000	0.530	-1.037
FPH	311269.0	438447.50	566676.6	648896.60	882294	1.278	1.300

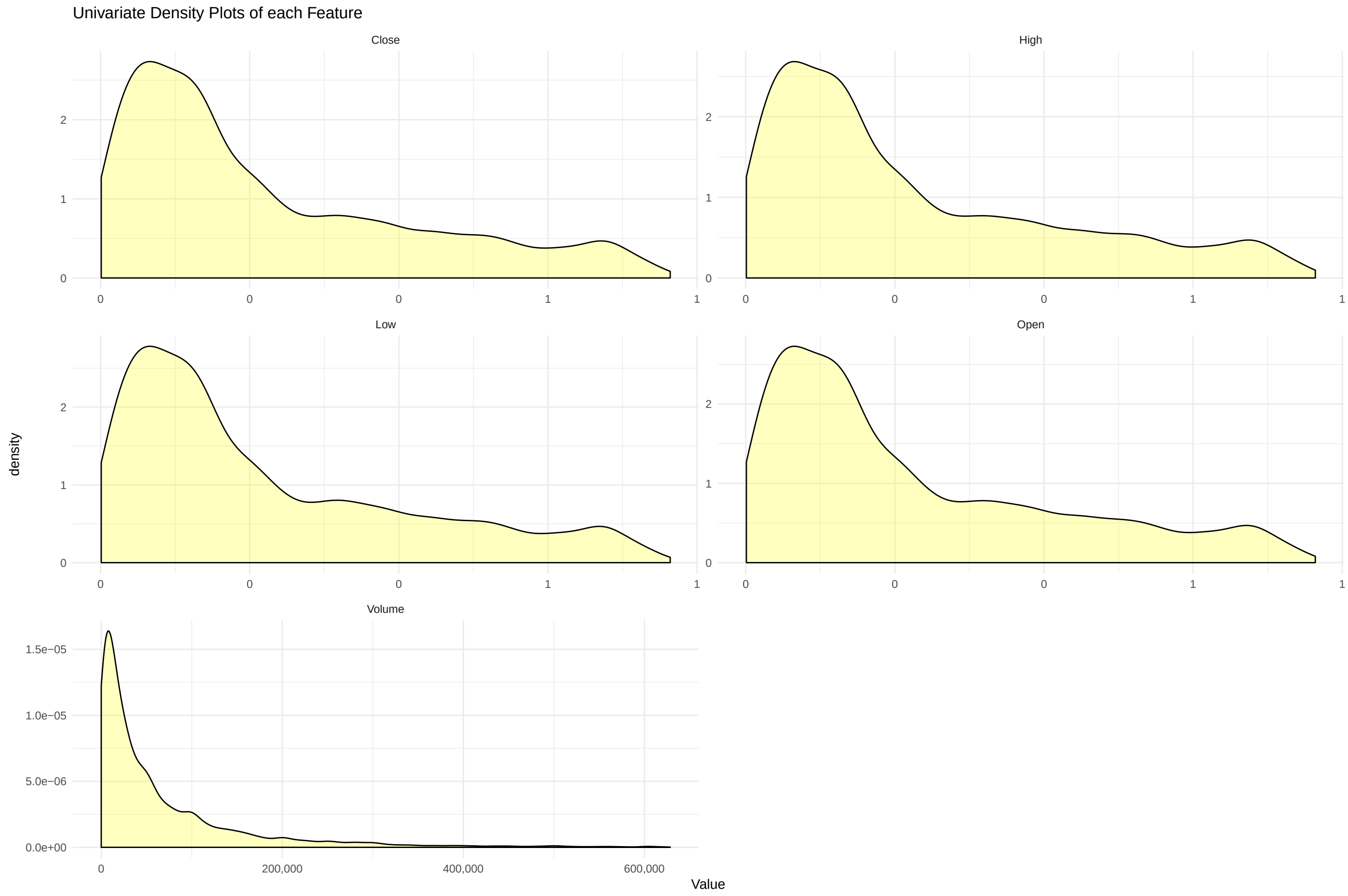
```
              position = "center",
              full_width = F)
```

ASX_Ticker	Date	Open	High	Low	Close	Volume	GICS_industry_group
RMI	20190319	0.009	0.009	0.009	0.009	16500	Materials
ONE	20190208	0.450	0.450	0.450	0.450	4500	Health Care Equipment & Services
EGN	20190326	0.480	0.480	0.480	0.480	19444	Capital Goods
AMH	20190211	0.855	0.855	0.850	0.850	103275	Not Applic
ATP	20190402	0.015	0.015	0.015	0.015	5300	Consumer Durables & Apparel
CY5	20190218	0.060	0.060	0.060	0.060	39939	Materials
BIR	20190130	0.160	0.180	0.160	0.180	27493	Diversified Financials
POD	20190311	0.062	0.062	0.062	0.062	27568	Materials
AIB	20190212	0.140	0.140	0.105	0.105	5255	Not Applic
MEY	20190322	0.105	0.105	0.105	0.105	9824	Energy
CXU	20190411	0.025	0.025	0.025	0.025	44509	Energy
SMP	20190305	0.190	0.190	0.190	0.190	11456	Software & Services
NOX	20190402	0.405	0.405	0.405	0.405	2000	Pharmaceuticals, Biotechnology & Life Sciences
CMP	20190213	0.330	0.330	0.330	0.330	8919	Health Care Equipment & Services
OKR	20190118	0.220	0.220	0.220	0.220	56326	Materials
ARL	20190211	0.510	0.515	0.510	0.515	23821	Materials
VMC	20190314	0.140	0.140	0.140	0.140	2327	Materials
AHK	20190205	0.016	0.016	0.016	0.016	1334	Materials
NOV	20190325	0.270	0.275	0.245	0.245	95319	Software & Services
EME	20190213	0.105	0.105	0.105	0.105	4623	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 is 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

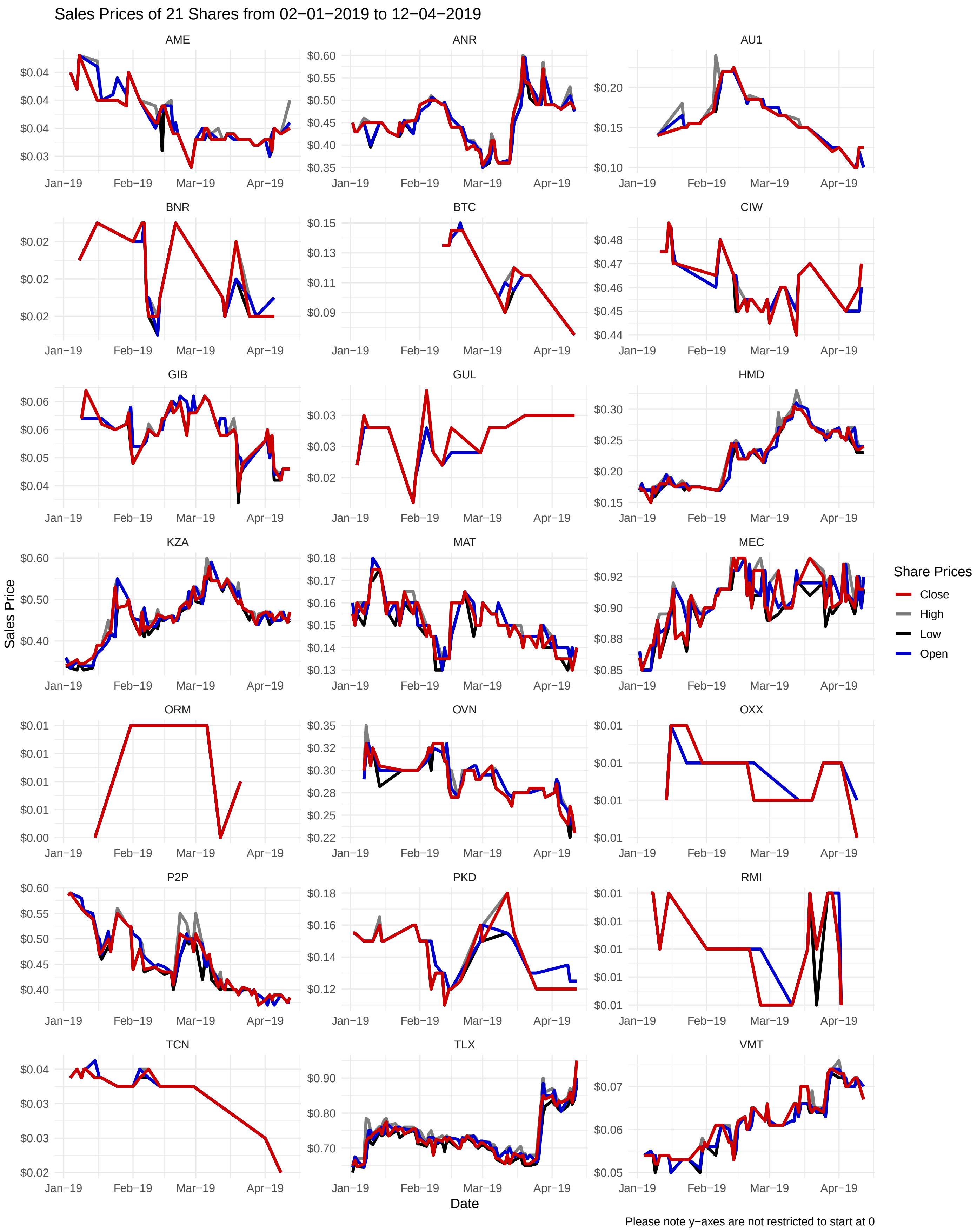
1.3 Data Exploration and Visualisation

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 appear 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))
```

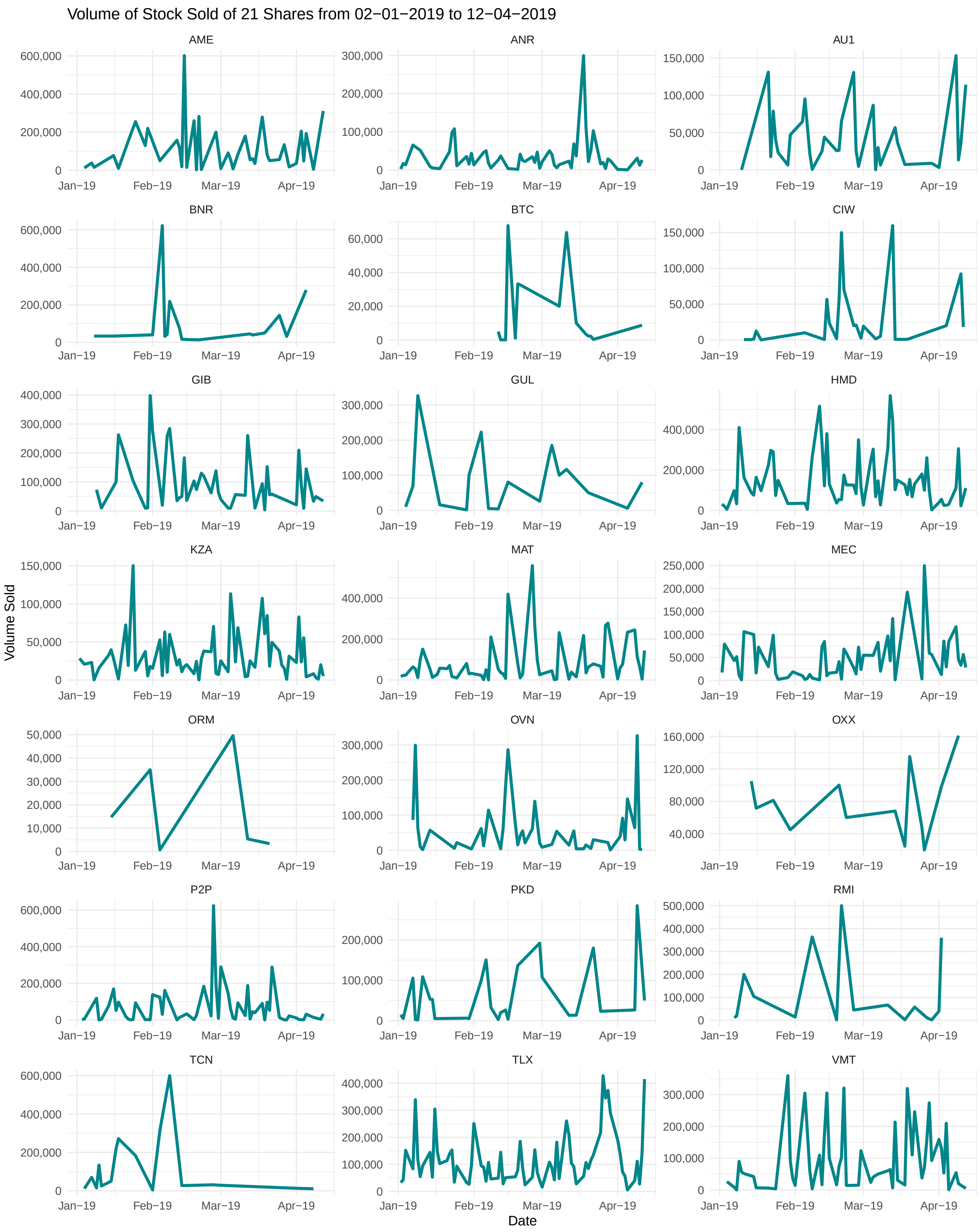
[2] pseudo-random; from a uniform distribution and not a truly random selection.


Sales Prices of 21 Shares from 02−01−2019 to 12−04−2019
AME
$0.04
$0.04
$0.04
$0.03
ANR
$0.60
$0.55
$0.50
$0.45
$0.40
$0.35
AU1
$0.20
$0.15
$0.10
BNR
$0.02
$0.02
$0.02
BTC
$0.15
$0.13
$0.11
$0.09
CIW
$0.48
$0.47
$0.46
$0.45
$0.44
GIB
$0.06
$0.06
$0.05
$0.04
GUL
$0.03
$0.03
$0.02
HMD
$0.30
$0.25
$0.20
$0.15
KZA
$0.60
$0.50
$0.40
MAT
$0.18
$0.17
$0.16
$0.15
$0.14
$0.13
MEC
$0.92
$0.90
$0.88
$0.85
ORM
$0.01
$0.01
$0.01
$0.01
$0.00
OVN
$0.35
$0.32
$0.30
$0.28
$0.25
$0.22
OXX
$0.01
$0.01
$0.01
$0.01
P2P
$0.60
$0.55
$0.50
$0.45
$0.40
PKD
$0.18
$0.16
$0.14
$0.12
RMI
$0.01
$0.01
$0.01
$0.01
$0.01
TCN
$0.04
$0.03
$0.03
$0.02
TLX
$0.90
$0.80
$0.70
VMT
$0.07
$0.06
$0.05
Jan−19
Feb−19
Mar−19
Apr−19
Sales Price
Date
Share Prices
Close
High
Low
Open
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.

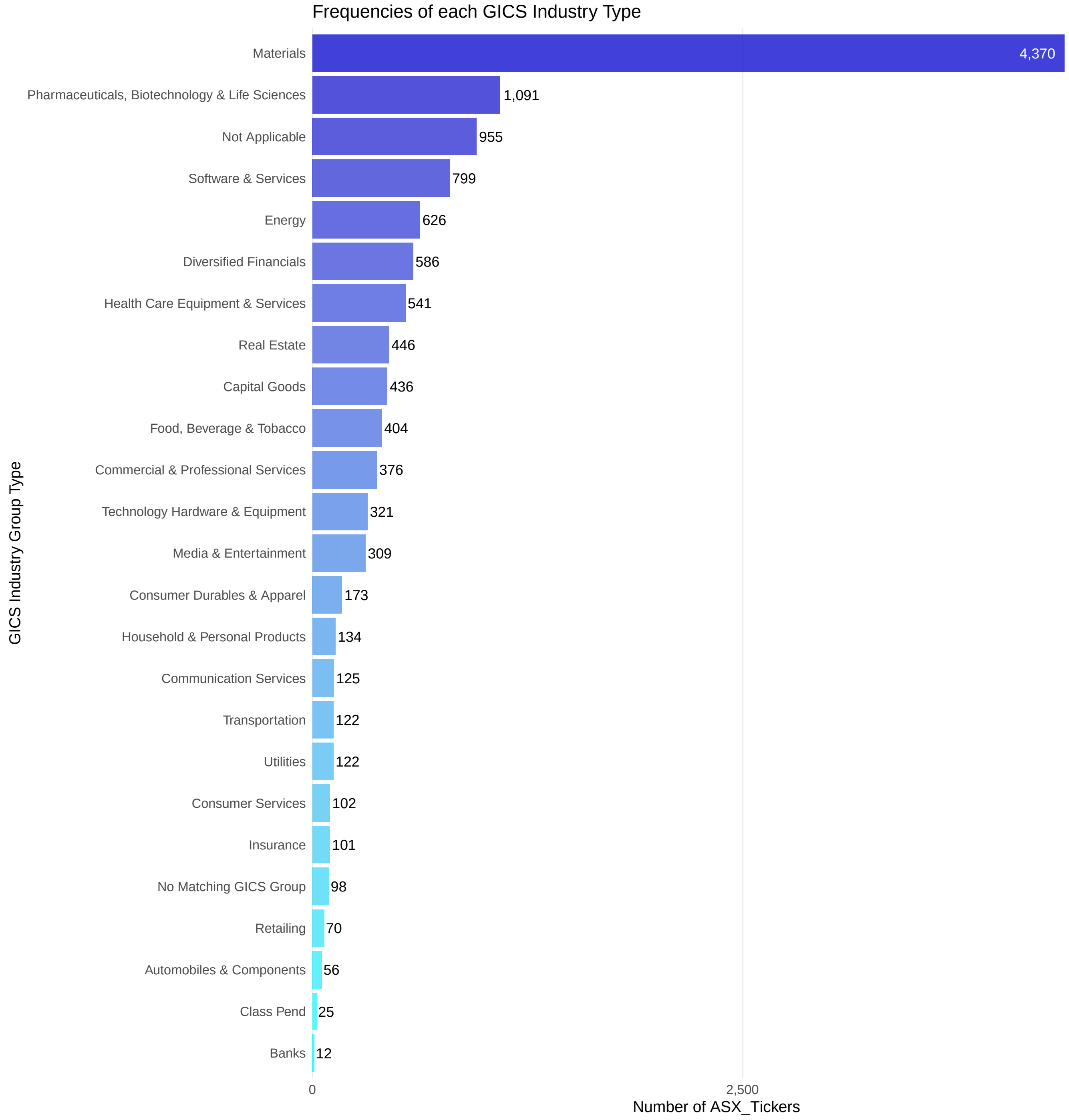
```
ggplot(ASX_Data_Samples) +
  geom_line(aes(x=Date, y=Volume),
            size=1.25, col = "turquoise4") +
  scale_x_date(date_breaks = "month", date_labels = "%b-%y") +
  scale_y_continuous("Volume Sold",
                     labels = comma)+
  ggtitle("Volume of Stock Sold of 21 Shares from 02-01-2019 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
AME
ANR
AU1
BNR
BTC
CIW
GIB
GUL
HMD
KZA
MAT
MEC
ORM
OVN
OXX
P2P
PKD
RMI
TCN
TLX
VMT
Volume Sold
Date
Jan−19
Feb−19
Mar−19
Apr−19
0
50,000
100,000
150,000
200,000
250,000
300,000
400,000
600,000
10,000
20,000
30,000
40,000
60,000
80,000
120,000
160,000
500,000

1.3.3 Number of Companies per GICS Group

The Materials industry group was the most frequently occurring GICS grouping in the data set 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.

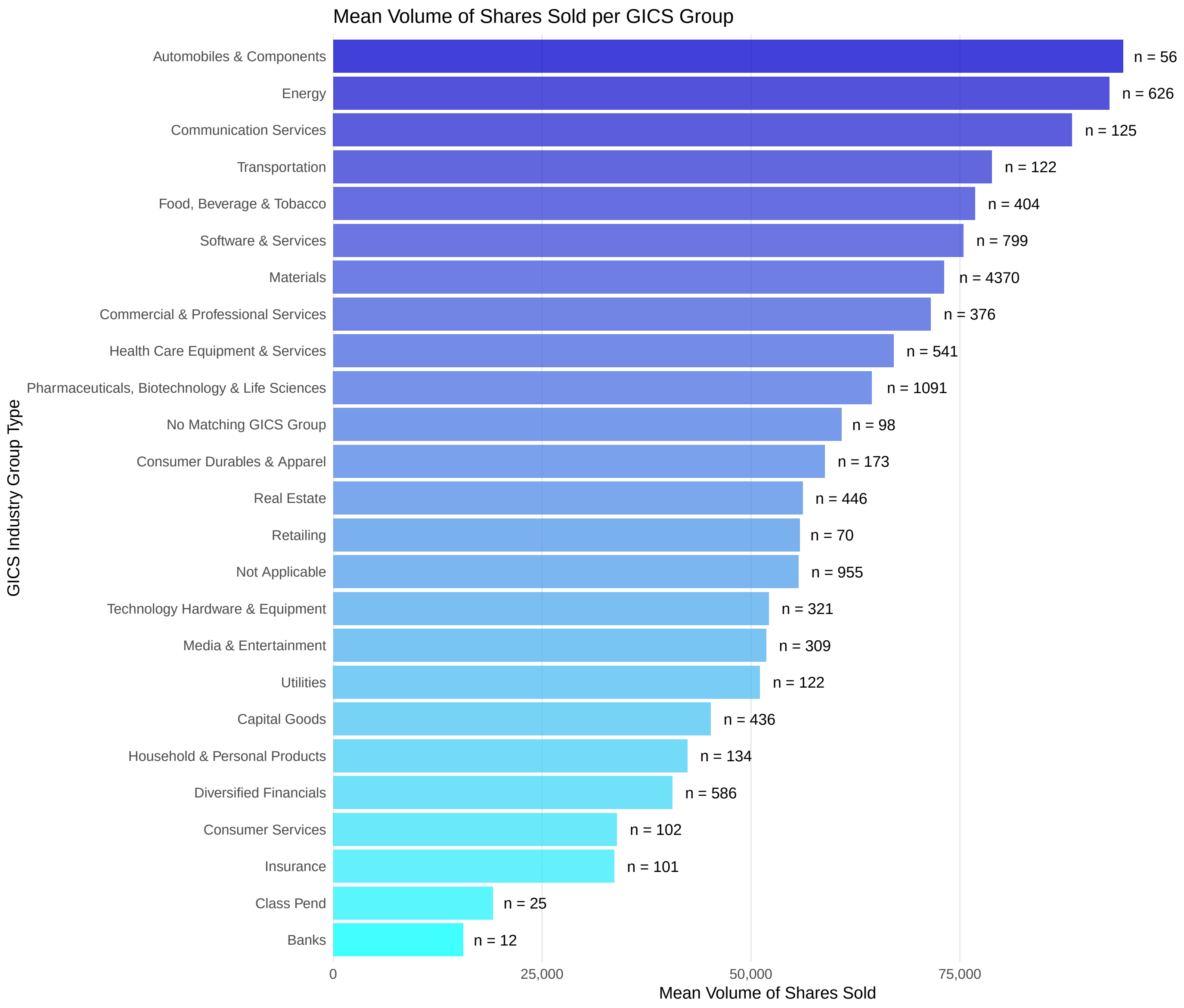
```
ASX_Data_Lower$GICS_industry_group <- recode(ASX_Data_Lower$GICS_industry_group,
                                             "Not Applic"="Not Applicable")

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$Count)*1.075,
                                  by = 2500),
                     limits = c(0, max(ASX_Data_Count$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)) +
  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.

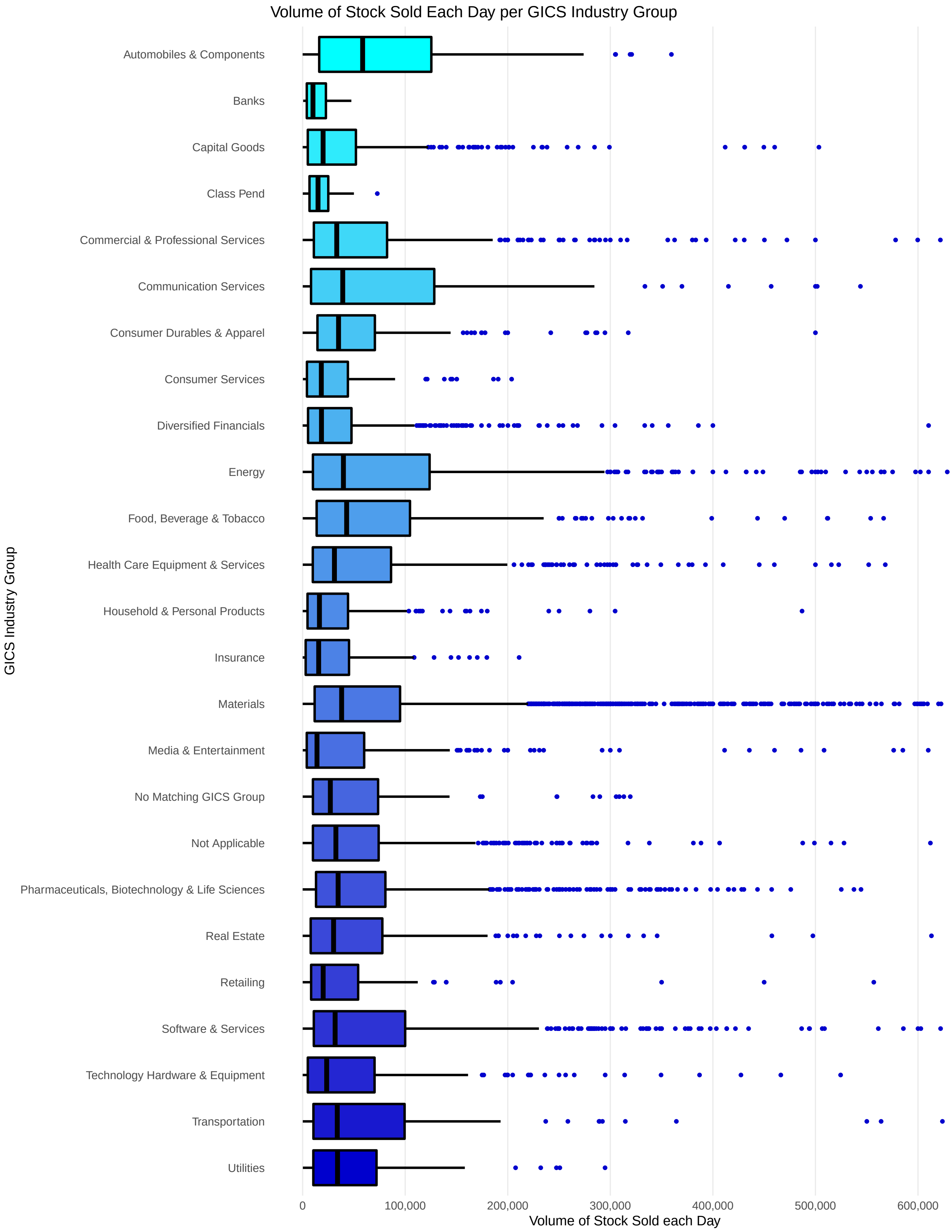
```
ASX_Lower_Vol <- summarise(group_by(ASX_Data_Lower,
                                    GICS_industry_group),
                           Mean_Vol = mean(Volume),
                           n_Companies = n())

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 data set. These boxplots below showed that, despite the data set being right-skewed, that the skew is present across most GICS groups.

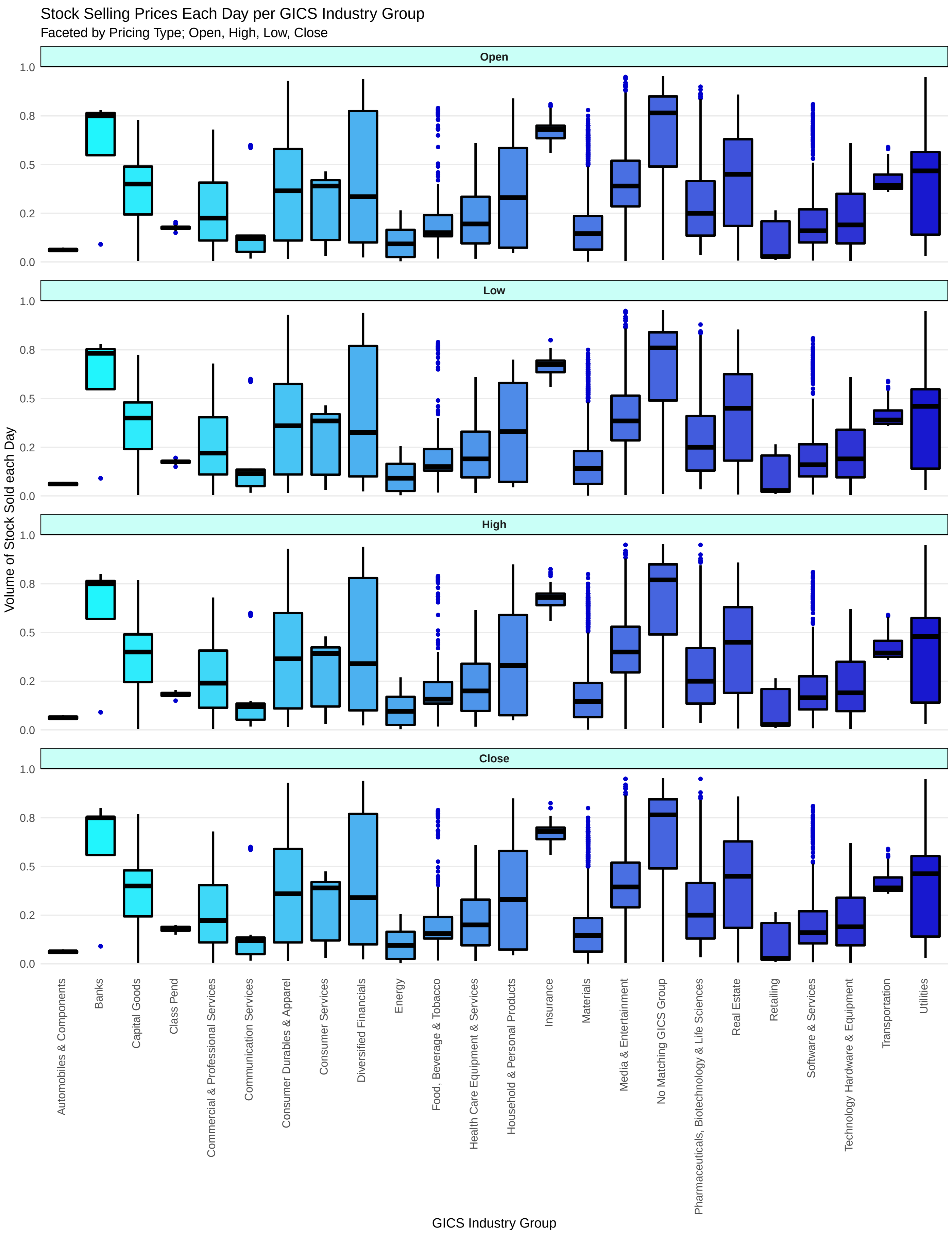
```r
ggplot(ASX_Data_Lower) +
  geom_boxplot(aes(x = fct_rev(GICS_industry_group), y = Volume,
                   fill = GICS_industry_group),
               show.legend = F, col = "black",
               size = 1,
               outlier.size = 1.25,
               outlier.colour = "blue3") +
  scale_x_discrete("GICS 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.

```
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
Volume of Stock Sold each Day
1.0
0.8
0.5
0.2
0.0
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 data set, 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