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

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 (GICS) 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 *Volume*s. 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**](¡'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

1.2 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
                          "grid",       ## for stacking ggplots
                          "gridExtra",  ## also for stacking ggplots
                          "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",
                    caption = "ASX Data Frame Sample with Prices
                    and Volume - 20 ASX\\_Tickers"),
              latex_options = c("striped", "hold_position"),
              position = "center",
              full_width = F,
              font_size = 10)
```

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

Table 1: ASX Data Frame Sample with Prices and Volume - 20 ASX_Tickers

ASX_Ticker	Date	Open	High	Low	Close	Volume
GGG	20190118	0.066	0.066	0.065	0.065	801690
AU8	20190107	0.245	0.245	0.240	0.240	51432
AFI	20190108	6.100	6.120	6.020	6.020	253132
Z1P	20190122	1.170	1.175	1.140	1.145	604934
TLG	20190115	0.385	0.385	0.375	0.375	141543
SGH	20190116	2.450	2.450	2.450	2.450	93
MCT	20190117	0.014	0.014	0.014	0.014	203642
VAE	20190102	60.700	60.800	59.500	59.610	2625
IEM	20190125	58.400	59.040	58.400	59.020	12991
XRF	20190103	0.140	0.140	0.140	0.140	50828
MRM	20190115	0.150	0.155	0.150	0.150	523198
BLK	20190102	0.041	0.042	0.040	0.040	6486812
TGA	20190114	0.575	0.580	0.575	0.575	38665
CCG	20190308	0.099	0.100	0.072	0.100	337347
8IH	20190115	0.085	0.085	0.085	0.085	3000
MSR	20190104	0.004	0.004	0.004	0.004	1418516
KMD	20190108	2.270	2.270	2.170	2.210	930376
TON	20190315	0.044	0.045	0.044	0.045	2017090
EML	20190408	1.785	1.785	1.755	1.775	169057
VGS	20190412	73.090	73.150	73.010	73.050	47596

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")

kable_styling(kable(sample_n(ASX_Industry_Table, size = 20),
                    caption = "ASX GICS Table - 20 ASX\\_Tickers"),
              latex_options = c("striped", "hold_position"),
              position = "center",
              full_width = F,
              font_size = 10)
```

Table 2: ASX GICS Table - 20 ASX_Tickers

Company_name	ASX_code	GICS_industry_group
KINETIKO ENERGY LTD	KKO	Energy
FAMILY INSIGHTS GROUP LIMITED	FAM	Software & Services
ECARGO HOLDINGS LIMITED	ECG	Commercial & Professional Services
BINGO INDUSTRIES LIMITED	BIN	Commercial & Professional Services
FBR LTD	FBR	Capital Goods
SCHAFFER CORPORATION LIMITED	SFC	Automobiles & Components
APPEN LIMITED	APX	Software & Services
BABY BUNTING GROUP LIMITED	BBN	Retailing
CARBINE RESOURCES LIMITED	CRB	Materials
GDI PROPERTY GROUP	GDI	Real Estate
NEW ZEALAND KING SALMON INVESTMENTS LIMITED	NZK	Food, Beverage & Tobacco
RYDER CAPITAL LIMITED	RYD	Not Applic
QMS MEDIA LIMITED	QMS	Media & Entertainment
ARGO INVESTMENTS LIMITED	ARG	Not Applic
NELSON RESOURCES LIMITED.	NES	Materials
MERCHANT HOUSE INTERNATIONAL LIMITED	MHI	Consumer Durables & Apparel
BULLETIN RESOURCES LIMITED	BNR	Materials
LATITUDE CONSOLIDATED LIMITED	LCD	Materials
TITAN MINERALS LIMITED	TTM	Materials
AMCIL LIMITED	AMH	Not Applic

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

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",
                    caption = "Sample of ASX Data Frame with
                    GICS\\_industry\\_group added - 20 ASX\\_Tickers"),
              latex_options = c("striped", "hold_position"),
              position = "center",
              full_width = F,
              font_size = 10)
```

Table 3: Sample of ASX Data Frame with GICS_industry_group added - 20 ASX_Tickers

ASX_Ticker	Date	Open	High	Low	Close	Volume	GICS_industry_group
EPD	20190214	0.470	0.470	0.470	0.470	20000	Software & Services
XAT	20190213	6024.500	6039.000	5998.200	6008.800	0	NA
FXL	20190117	1.275	1.325	1.272	1.300	804513	Diversified Financials
ANP	20190109	0.029	0.032	0.029	0.031	344448	Pharmaceuticals, Biotechnology & Life Sciences
DNA	20190319	0.066	0.067	0.066	0.066	294685	Consumer Services
ANZ	20190211	26.890	26.930	26.300	26.540	5006572	Banks
QHL	20190129	0.071	0.071	0.070	0.070	185726	Capital Goods
RIE	20190107	0.016	0.016	0.016	0.016	3321	Materials
DUI	20190319	4.130	4.140	4.120	4.140	68492	Not Applic
RDG	20190408	0.024	0.024	0.024	0.024	40000	Materials
RAC	20190214	0.086	0.086	0.081	0.081	49654	Pharmaceuticals, Biotechnology & Life Sciences
BNR	20190208	0.021	0.021	0.020	0.020	218500	Materials
TAM	20190111	0.042	0.042	0.042	0.042	200000	Materials
RUL	20190304	0.570	0.580	0.570	0.580	44408	Software & Services
FOD	20190404	0.091	0.091	0.089	0.091	1341421	Food, Beverage & Tobacco
LVT	20190304	0.375	0.390	0.375	0.385	1127275	Software & Services
RMP	20190218	0.079	0.079	0.076	0.076	2708423	Energy
CIO	20190222	0.002	0.003	0.002	0.003	316846	Technology Hardware & Equipment
SIQ	20190222	8.700	8.760	8.530	8.720	305393	Commercial & Professional Services
LSX	20190401	0.350	0.350	0.350	0.350	15302	Not Applic

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()),
                         "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",
                    caption = "Descriptives before processing"),
              full_width = F,
              latex_options = c("striped", "hold_position"),
              position = "center",
              font_size = 10)
```

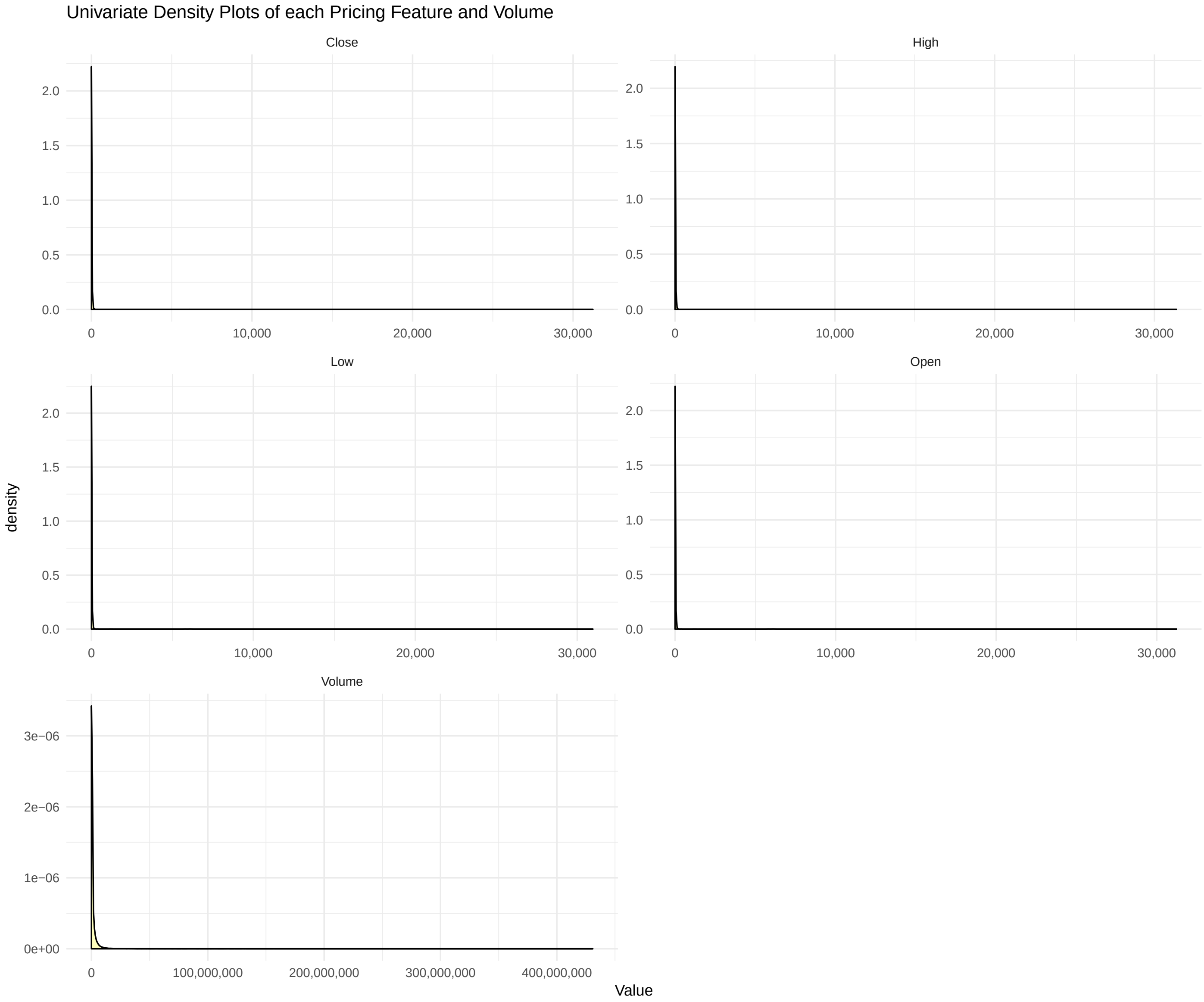
Table 4: Descriptives before processing

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

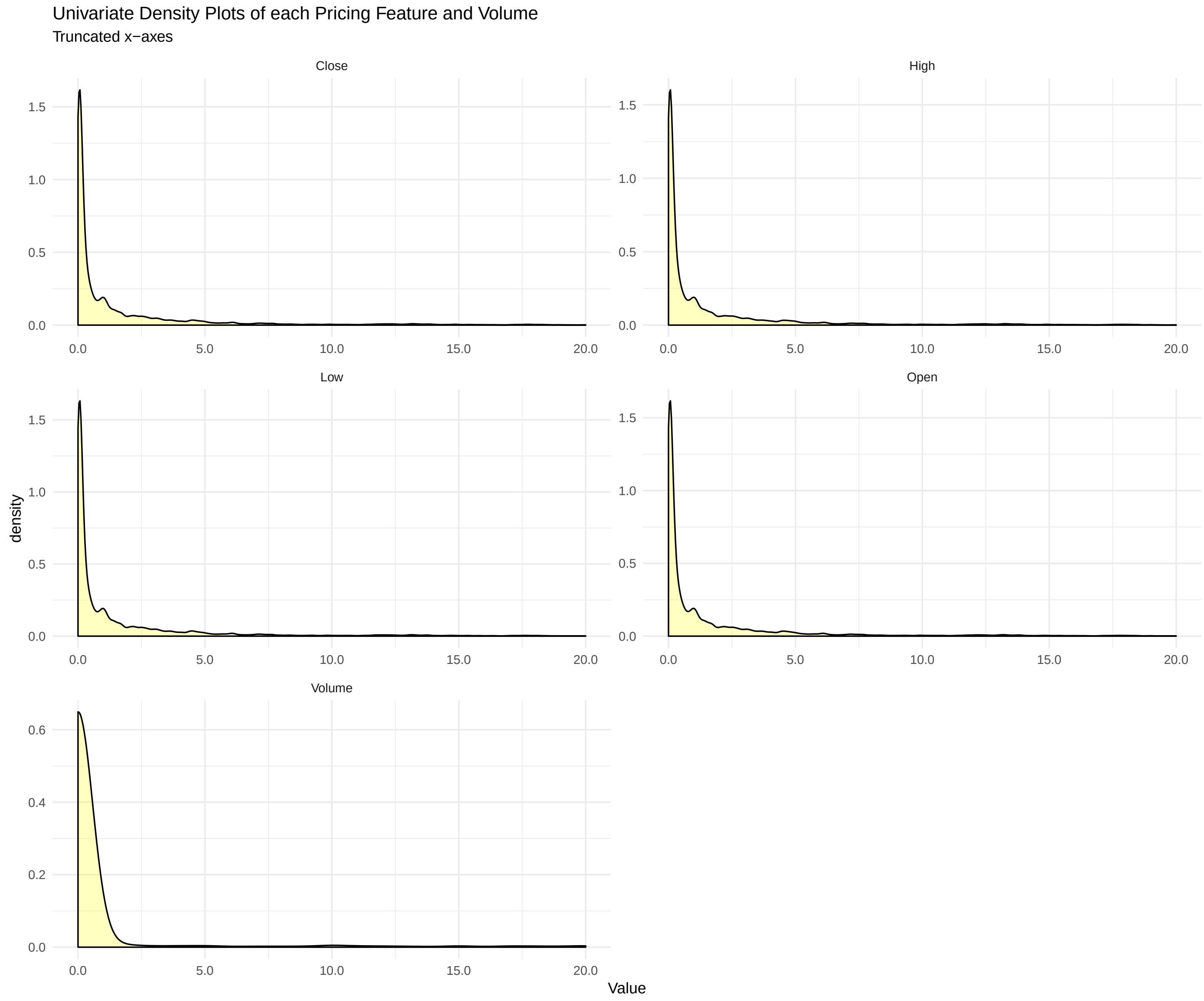
Plotting the spread of the features only further outlined the magnitude of the skew. As such, the data was filtered to remove shares that showed high values for any feature.

```
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 Pricing Feature and Volume") +
  theme_minimal()
```



To highlight he extent of the skew, the above plots were reproduced with truncated x-axes.

```
ggplot(filter(ASX_Long)) +
  geom_density(aes(x = Value),
               fill = "yellow", alpha = 0.25) +
  scale_x_continuous(labels=comma_format(accuracy = 0.1),
                     limits = c(0,20)) +
  facet_rep_wrap(~Variable, repeat.tick.labels = T,
                 scales = "free_y", ncol = 2) +
  labs(title = "Univariate Density Plots of each Pricing Feature and Volume",
       subtitle = "Truncated x-axes") +
  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 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)

kable_styling(kable(ASX_kable[, 1:7],
                    align = "lrrrrrr",
                    caption = "Descriptives for 20 ASX\\_Tickers after
                    filtering by High price"),
              latex_options = c("striped", "hold_position"),
              position = "center",
              full_width = F,
              font_size = 10)
```

Table 5: Descriptives for 20 ASX_Tickers after filtering by High price

ASX_Ticker	n Observations	Min Date	Max Date	Minimum	Q1	Median
TTA	5	17/01/2019	04/04/2019	0.015	0.01500	0.0160
DDT	56	02/01/2019	12/04/2019	0.002	0.00500	0.0070
MGG	72	02/01/2019	12/04/2019	1.620	1.65000	1.6725
ACP	1	28/02/2019	28/02/2019	0.008	0.00800	0.0080
CL8	55	03/01/2019	12/04/2019	0.011	0.01100	0.0120
EPM	28	07/01/2019	05/04/2019	0.002	0.00200	0.0030
OVL	41	04/01/2019	11/04/2019	0.002	0.00300	0.0030
FAR	72	02/01/2019	12/04/2019	0.054	0.05800	0.0600
SRZ	49	02/01/2019	11/04/2019	0.011	0.01300	0.0130
BSE	71	02/01/2019	12/04/2019	0.230	0.24000	0.2800
PFG	37	07/01/2019	08/04/2019	0.071	0.07500	0.0850
AGY	72	02/01/2019	12/04/2019	0.092	0.11000	0.1360
PGY	19	03/01/2019	08/04/2019	0.013	0.01550	0.0170
HUO	72	02/01/2019	12/04/2019	4.500	4.74000	4.7800
CML	47	14/01/2019	11/04/2019	0.016	0.01750	0.0180
KLH	35	02/01/2019	11/04/2019	0.002	0.00300	0.0030
HPR	31	11/01/2019	12/04/2019	0.056	0.06400	0.0650
LMW	20	02/01/2019	18/02/2019	0.410	0.44000	0.4600
BGH	8	24/01/2019	25/02/2019	0.059	0.06675	0.0695
AYK	23	07/01/2019	11/04/2019	18.500	19.00000	19.0000

```r
kable_styling(kable(ASX_kable[, c(1, 8:13)],
                    align = "lrrrrrrr",
                    caption = "Descriptives for 20 ASX\\_Tickers after
                    filtering by High price (cont)"),
              latex_options = c("striped", "hold_position"),
              position = "center",
              full_width = F,
              font_size = 10)
```

Table 6: Descriptives for 20 ASX_Tickers after filtering by High price (cont)

ASX_Ticker	Q3	90th Percentile	95th Percentile	Maximum	Skew	Kurtosis
TTA	0.01600	0.0184	0.01920	0.020	0.921	-1.100
DDT	0.00800	0.0080	0.00800	0.009	-0.769	-0.525
MGG	1.69000	1.7000	1.70000	1.705	-0.334	-1.120
ACP	0.00800	0.0080	0.00800	0.008	NaN	NaN
CL8	0.01300	0.0130	0.01400	0.015	0.958	0.769
EPM	0.00300	0.0030	0.00300	0.004	0.063	-0.973
OVL	0.00400	0.0040	0.00400	0.005	0.152	-0.238
FAR	0.06225	0.0680	0.06945	0.076	1.289	1.531
SRZ	0.01400	0.0150	0.01500	0.016	0.022	-0.520
BSE	0.29750	0.3100	0.31500	0.330	0.084	-1.316
PFG	0.09000	0.0944	0.09500	0.095	0.070	-1.723
AGY	0.14500	0.1500	0.15000	0.160	-0.569	-1.112
PGY	0.01800	0.0182	0.01910	0.020	-0.425	-0.802
HUO	4.85000	4.9405	5.03000	5.100	-0.025	0.663
CML	0.01900	0.0200	0.02000	0.021	0.505	0.186
KLH	0.00300	0.0040	0.00430	0.005	0.882	1.312
HPR	0.06900	0.0700	0.07000	0.070	-0.561	0.645
LMW	0.48000	0.4850	0.48500	0.485	-0.245	-1.343
BGH	0.07000	0.0700	0.07000	0.070	-1.048	-0.663
AYK	19.65000	20.4000	20.71600	20.750	0.992	-0.162

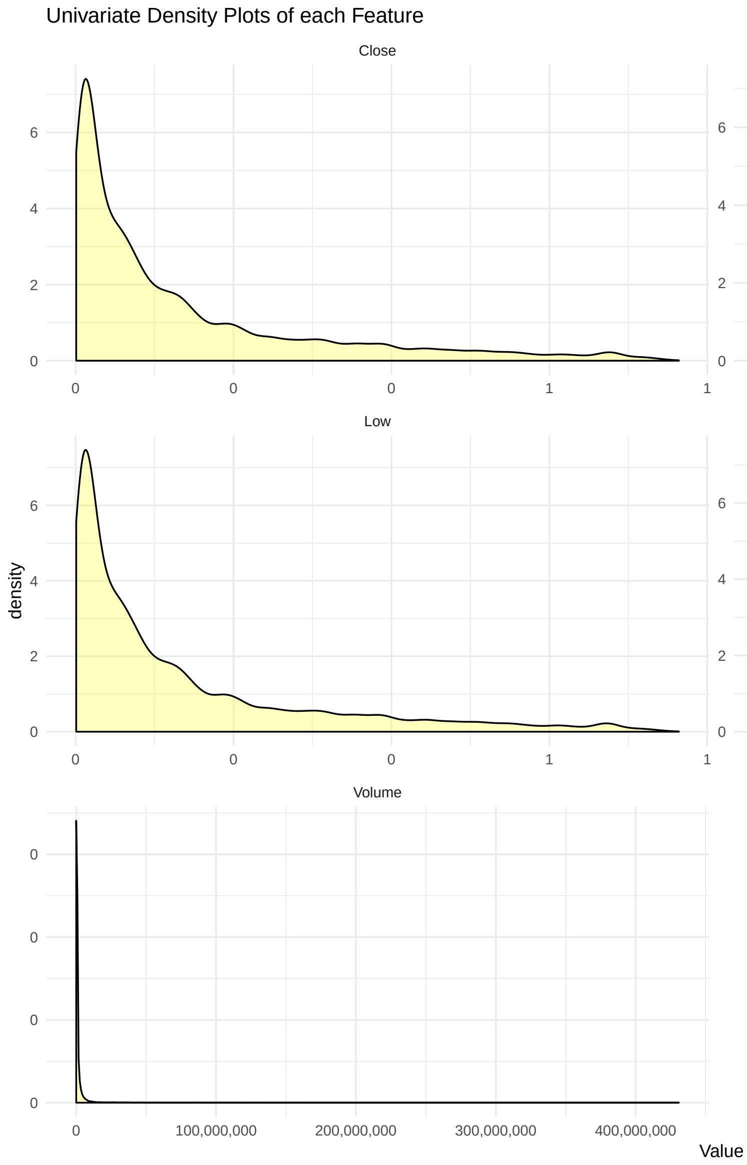
```r
ASX_Lower <- filter(ASX_Ticker_Summary_Price, Maximum < quantile(Maximum, 2/3))

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 = "lrrrrrrl",
                    caption = "Sample of ASX Data Frame after filtering
                    by High price - 20 ASX\\_Tickers"),
              latex_options = c("striped", "hold_position"),
              position = "center",
              full_width = F,
              font_size = 10)
```

Table 7: Sample of ASX Data Frame after filtering by High price - 20 ASX_Tickers

ASX_Ticker	Date	Open	High	Low	Close	Volume	GICS_industry_group
ADN	20190206	0.006	0.007	0.006	0.007	1283200	Materials
EM2	20190121	0.170	0.170	0.170	0.170	1220	Materials
XPE	20190212	0.002	0.002	0.001	0.001	111111	Software & Services
CXL	20190205	0.780	0.780	0.740	0.770	116140	Materials
YOW	20190313	0.088	0.090	0.084	0.085	5383287	Food, Beverage & Tobacco
BLT	20190215	0.130	0.135	0.130	0.130	45302	Pharmaceuticals, Biotechnology & Life Sciences
NXM	20190401	0.061	0.061	0.061	0.061	27026	Materials
PPG	20190213	0.195	0.200	0.190	0.200	373011	Materials
PAN	20190116	0.455	0.455	0.430	0.450	502902	Materials
CDM	20190404	0.900	0.905	0.875	0.885	545788	Not Applic
MMM	20190313	0.495	0.495	0.495	0.495	3138	Consumer Services
E25	20190201	0.160	0.160	0.150	0.160	77291	Materials
4DS	20190205	0.061	0.064	0.059	0.062	6377952	Semiconductors & Semiconductor Equipment
ODA	20190227	0.100	0.125	0.100	0.125	162700	Software & Services
IGN	20190322	0.052	0.053	0.052	0.052	68180	Commercial & Professional Services
EXU	20190208	0.160	0.160	0.160	0.160	741901	NA
EAR	20190320	0.220	0.220	0.210	0.215	835157	Materials
FAR	20190206	0.055	0.056	0.054	0.054	9397805	Energy
GMD	20190115	0.034	0.035	0.034	0.035	704079	Materials
CTP	20190213	0.140	0.145	0.140	0.145	1534046	Energy

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 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",
                    caption = "Descriptives for 20 ASX\\_Tickers after
                    filtering by High price and Volume"),
              latex_options = c("striped", "hold_position"),
              position = "center",
              full_width = F,
              font_size = 10)
```

Table 8: Descriptives for 20 ASX_Tickers after filtering by High price and Volume

ASX_Ticker	n Observations	Min Date	Max Date	Minimum	Q1	Median
EHE	72	02/01/2019	12/04/2019	332525	596908.00	814389.0
ATH	1	12/04/2019	12/04/2019	100000	100000.00	100000.0
RDV	71	02/01/2019	12/04/2019	47	2497.00	4835.0
TSC	31	08/01/2019	12/04/2019	10	53650.00	300000.0
EGL	30	04/01/2019	05/04/2019	5000	13814.50	42518.0
DTR	32	07/01/2019	12/04/2019	10000	359595.50	864175.0
EME	13	11/01/2019	09/04/2019	198	4050.00	4623.0
FOR	65	02/01/2019	12/04/2019	22	15000.00	28397.0
ESK	11	25/02/2019	29/03/2019	1000	5850.00	30723.0
EVZ	34	07/01/2019	12/04/2019	19	3974.50	15121.5
AMG	67	03/01/2019	12/04/2019	17068	175800.00	400061.0
CSS	72	02/01/2019	12/04/2019	8545	37578.25	54125.5
KGD	5	14/01/2019	19/03/2019	52	298.00	100400.0
FRX	15	10/01/2019	10/04/2019	723	31500.00	95000.0
GCR	17	02/01/2019	11/04/2019	1437	5375.00	13667.0
XMJ	72	02/01/2019	12/04/2019	0	0.00	0.0
RND	51	02/01/2019	12/04/2019	1	2500.00	5000.0
WAM	72	02/01/2019	12/04/2019	153350	392126.75	461881.0
VHY	72	02/01/2019	12/04/2019	5166	14515.00	19483.0
DNK	71	02/01/2019	12/04/2019	48	28139.50	61345.0

```
kable_styling(kable(ASX_kable[, c(1, 8:13)],
                    align = "lrrrrrrr",
                    caption = "Descriptives after filtering by
                    High price and Volume (cont)"),
              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),
```

Table 9: Descriptives after filtering by High price and Volume (cont)

ASX_Ticker	Q3	90th Percentile	95th Percentile	Maximum	Skew	Kurtosis
EHE	1429044.8	2081575.4	2289402.2	8406958	4.636	28.283
ATH	100000.0	100000.0	100000.0	100000	NaN	NaN
RDV	7984.5	12459.0	15971.5	42313	2.992	11.967
TSC	841638.5	2294082.0	3030928.5	6784463	2.762	8.347
EGL	157965.5	431634.0	470233.1	1618924	3.508	13.373
DTR	2783328.2	6079100.0	8858000.0	16368850	2.458	6.014
EME	16500.0	32000.0	69078.4	120946	2.407	4.814
FOR	51400.0	95846.0	118074.4	264125	2.483	8.451
ESK	49824.0	66970.0	70300.0	73630	0.241	-1.693
EVZ	42912.5	88270.9	127023.9	225000	2.245	5.315
AMG	886686.5	1745312.2	2223621.2	5948931	2.984	10.678
CSS	88771.0	135597.1	174739.6	489190	3.495	16.639
KGD	120777.0	347389.2	422926.6	498464	0.883	-1.145
FRX	185101.5	234405.6	273775.0	351006	0.571	-0.759
GCR	35000.0	54805.2	63132.0	82500	1.018	-0.162
XMJ	0.0	0.0	0.0	0	NaN	NaN
RND	12891.0	31601.0	43420.0	124430	3.700	16.553
WAM	619342.8	762936.4	965276.1	1324274	1.603	3.042
VHY	26389.0	47647.7	62633.3	217163	4.708	27.372
DNK	119373.0	192527.0	283296.0	727030	3.212	11.741

```
                    align = "lrrrrrrl",
                    caption = "Sample of ASX Data Frame After filtering by
                    High price and Volume - 20 ASX\\_Tickers"),
              latex_options = c("striped", "hold_position"),
              position = "center",
              full_width = F,
              font_size = 10)
```

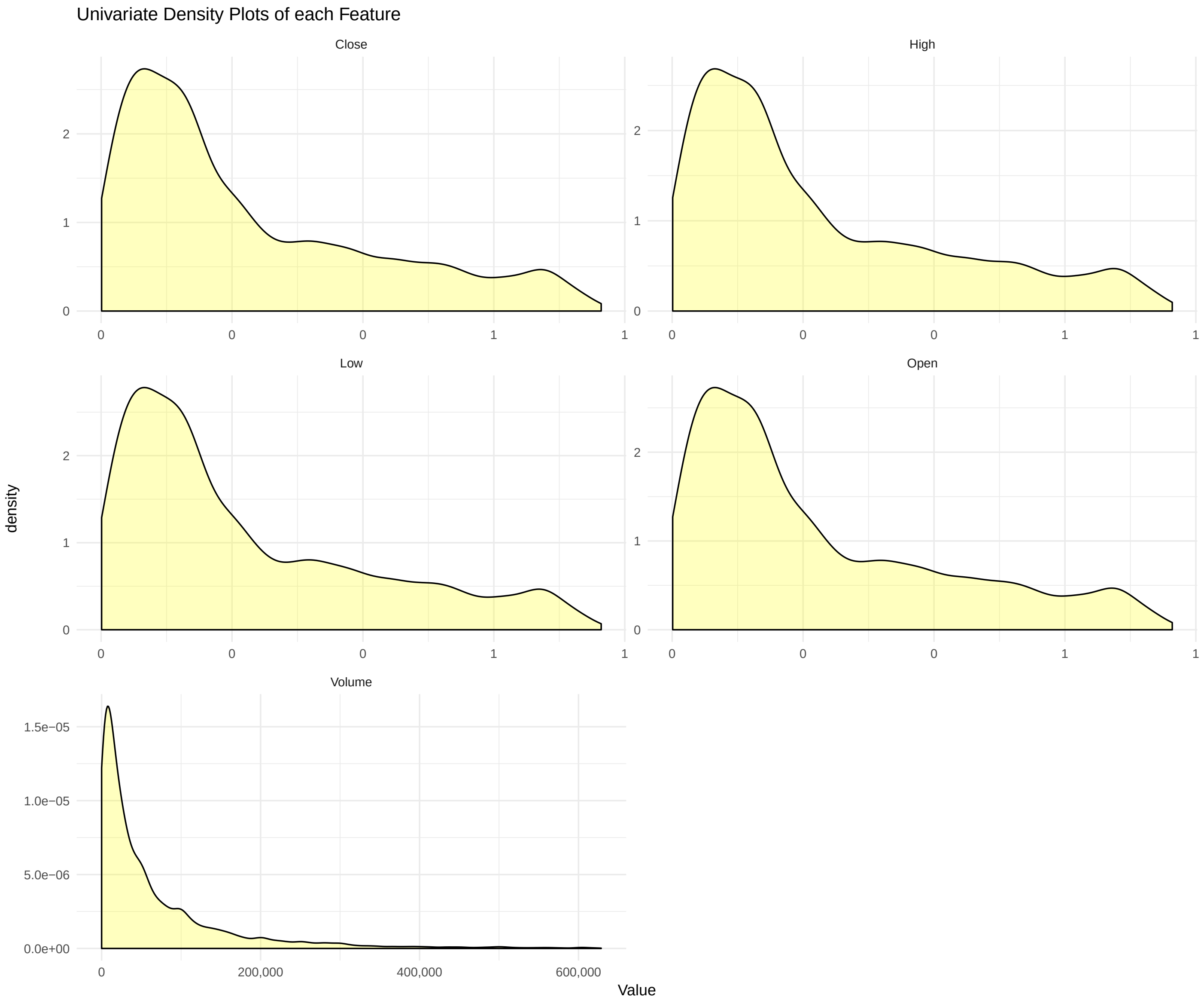
Table 10: Sample of ASX Data Frame After filtering by High price and Volume - 20 ASX_Tickers

ASX_Ticker	Date	Open	High	Low	Close	Volume	GICS_industry_group
VMC	20190111	0.170	0.170	0.170	0.170	5770	Materials
PLX	20190220	0.540	0.540	0.540	0.540	1361	Materials
KLO	20190322	0.095	0.095	0.095	0.095	4000	Real Estate
EMH	20190122	0.390	0.390	0.390	0.390	2000	Materials
SVS	20190320	0.250	0.250	0.250	0.250	1500	Not Applic
FPC	20190327	0.840	0.845	0.840	0.845	31834	Not Applic
S66	20190121	0.555	0.555	0.555	0.555	18983	Household & Personal Products
PKD	20190111	0.150	0.150	0.150	0.150	108779	Commercial & Professional Services
CCG	20190117	0.080	0.081	0.080	0.081	25617	Software & Services
MQR	20190307	0.060	0.060	0.060	0.060	15000	Materials
CBY	20190315	0.310	0.310	0.310	0.310	3500	Materials
IMC	20190301	0.255	0.255	0.230	0.235	165720	Pharmaceuticals, Biotechnology & Life Sciences
PNW	20190313	0.320	0.350	0.320	0.350	610000	Media & Entertainment
ID8	20190327	0.445	0.520	0.445	0.520	62000	Software & Services
AXE	20190110	0.073	0.076	0.073	0.076	71750	Materials
AMB	20190117	0.075	0.075	0.075	0.075	167000	Commercial & Professional Services
KZR	20190319	0.130	0.130	0.125	0.125	95000	Materials
PTB	20190322	0.700	0.700	0.690	0.690	257894	Capital Goods
MKG	20190411	0.120	0.120	0.120	0.120	18130	Materials
BUG	20190408	0.220	0.220	0.220	0.220	15000	Food, Beverage & Tobacco

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.

```r
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.

```r
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",
                    caption = "Descriptives for ASX Data Frame after
                    filtering by High Price and Volume"),
              latex_options = c("striped", "hold_position"),
              position = "center",
              full_width = F,
              font_size = 10)
```

Table 11: Descriptives for ASX Data Frame after filtering by High Price and Volume

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

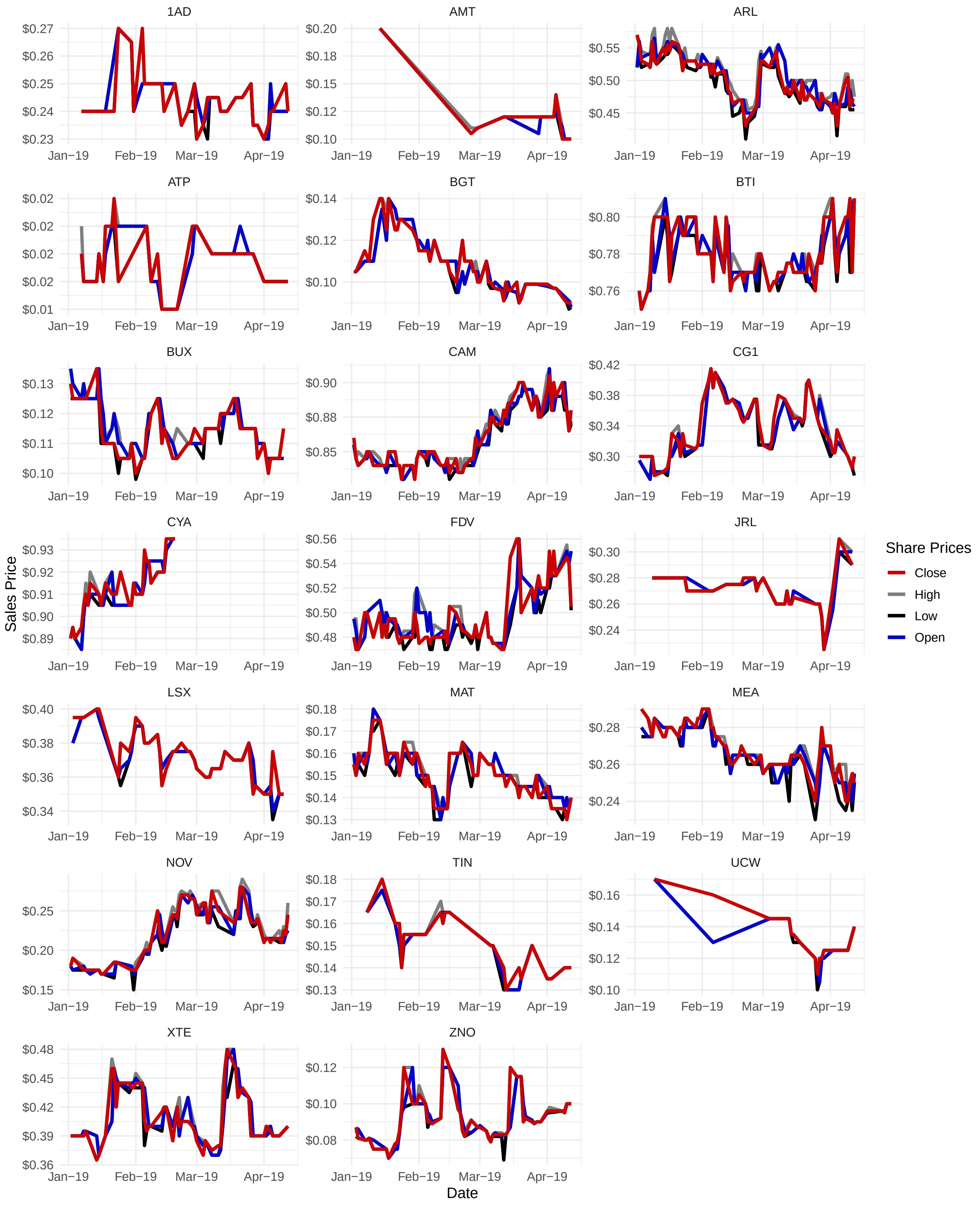
1.3 Data Exploration and Visualisation

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))
```

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

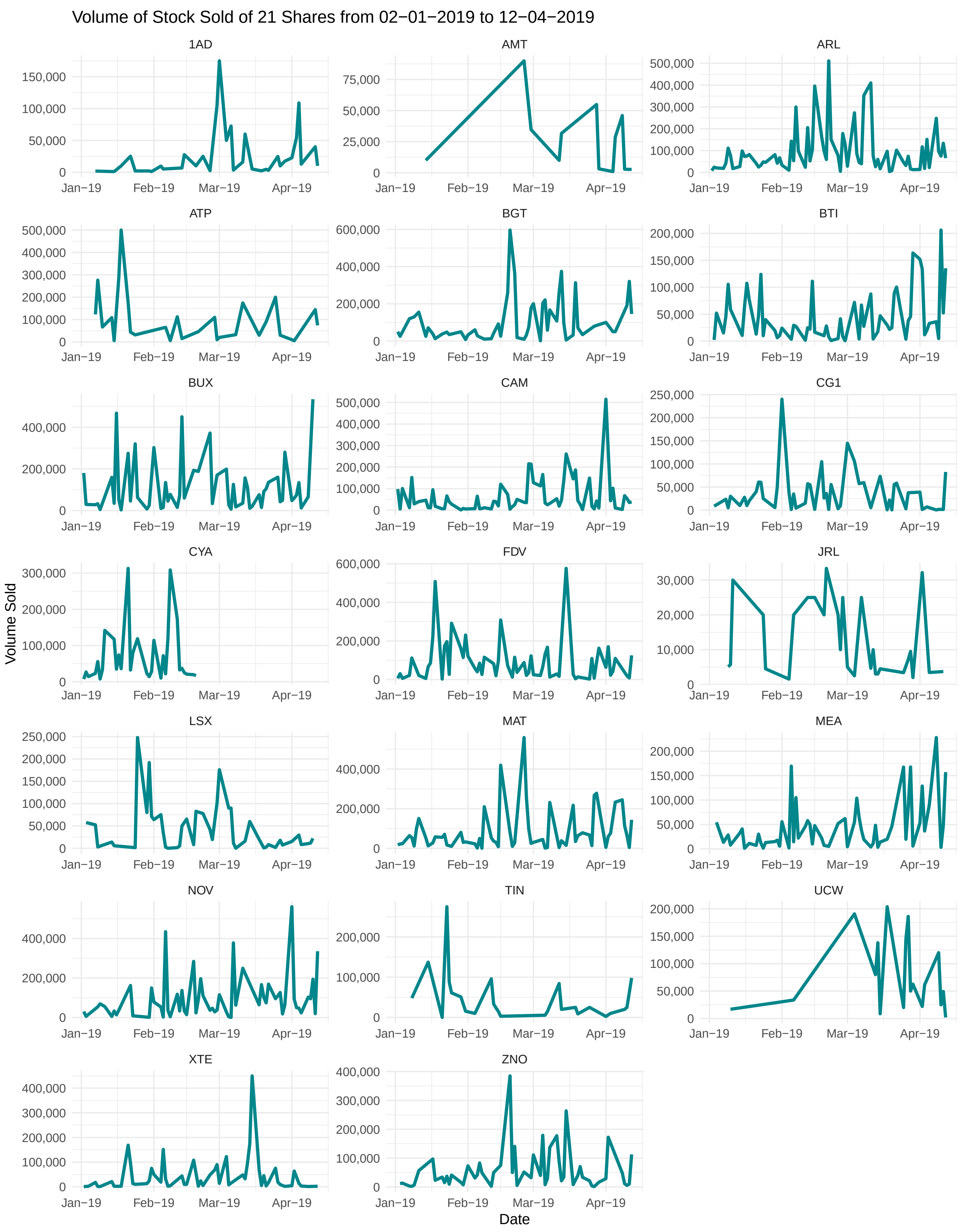


Please note y−axes are not restricted to start at 0

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))
```

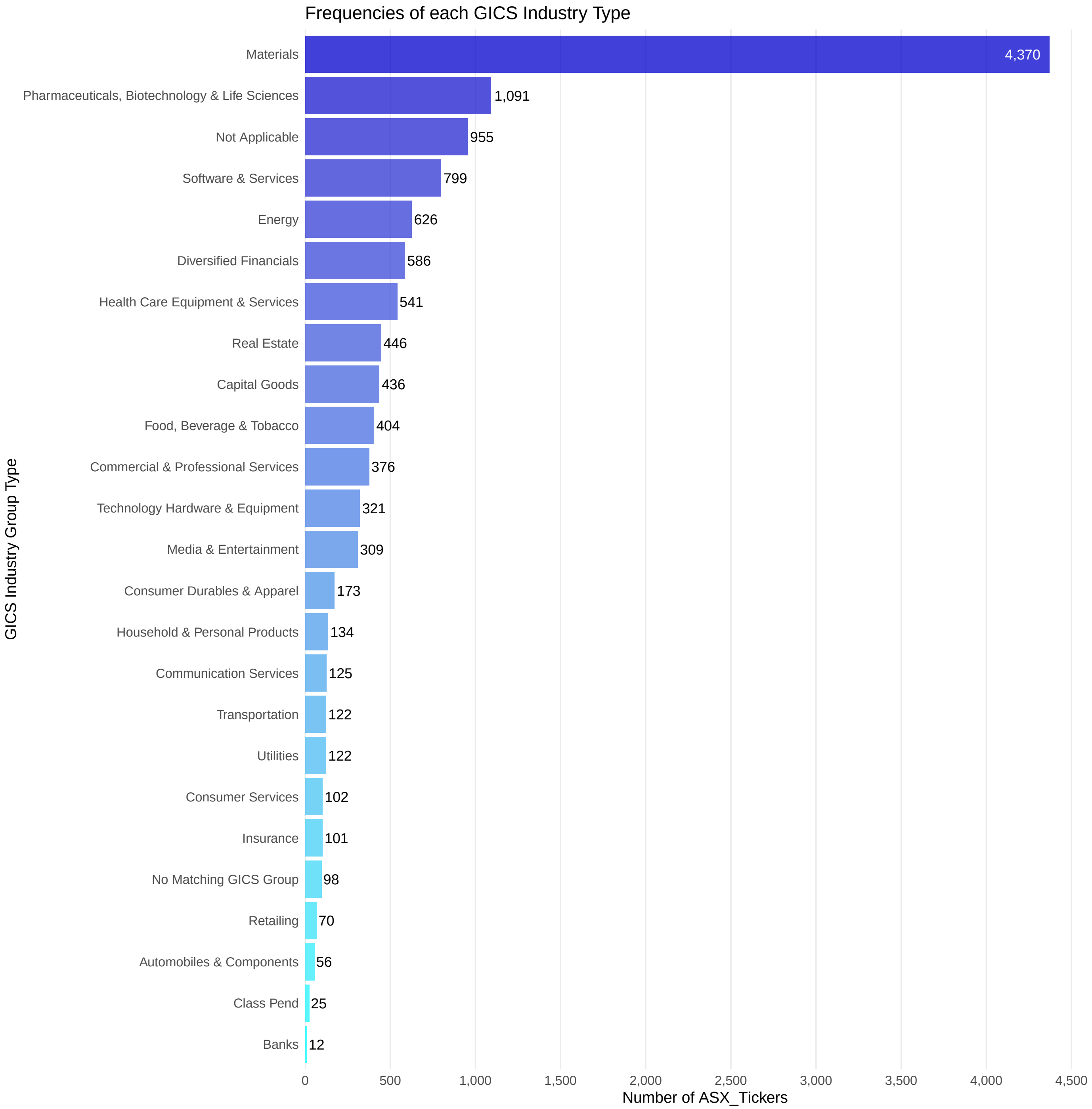


Volume of Stock Sold of 21 Shares from 02−01−2019 to 12−04−2019

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.

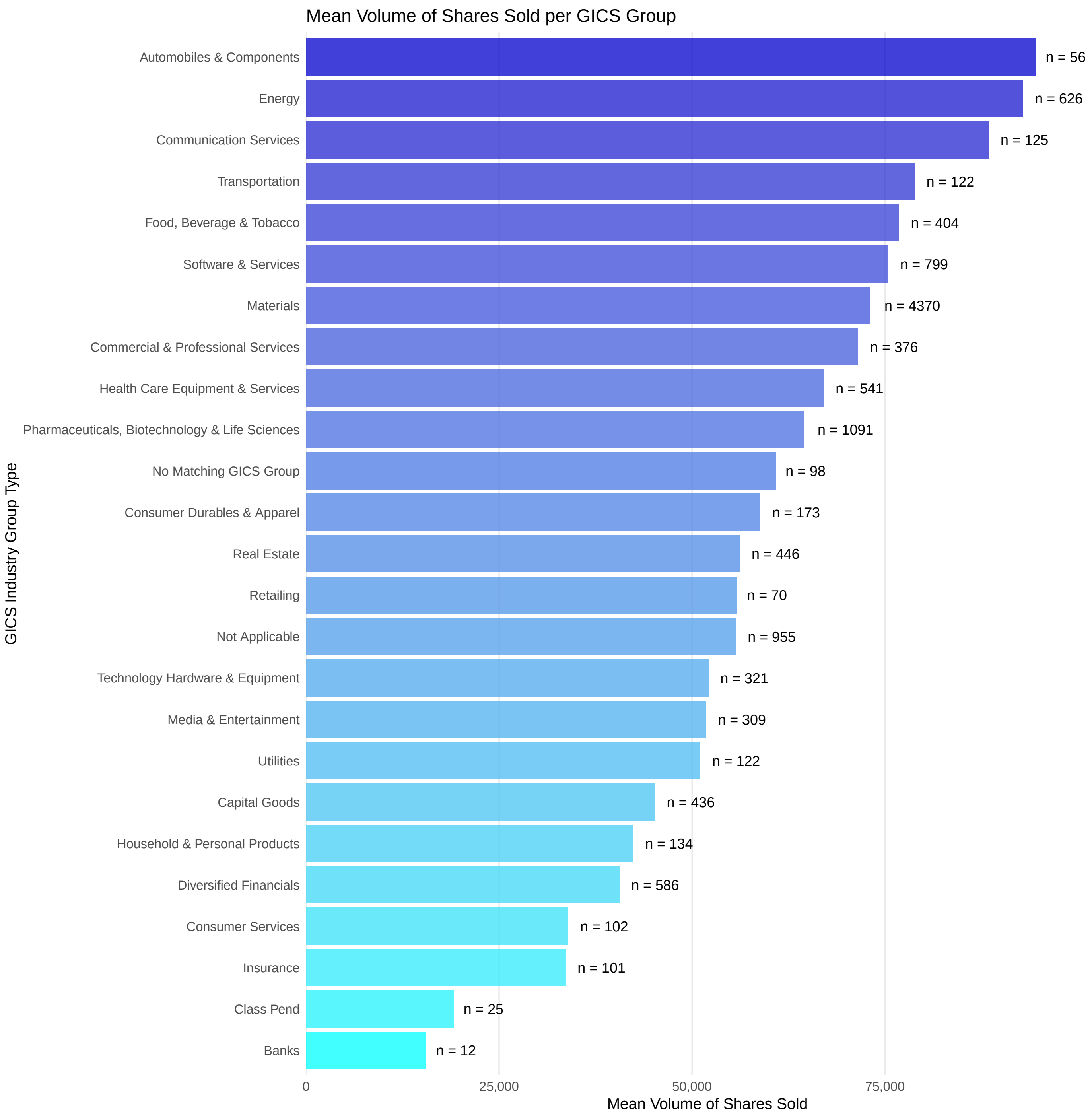
```r
ASX_Data_Lower$GICS_industry_group <- recode(ASX_Data_Lower$GICS_industry_group,
                                             "Not Applic"="Not Applicable")

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.

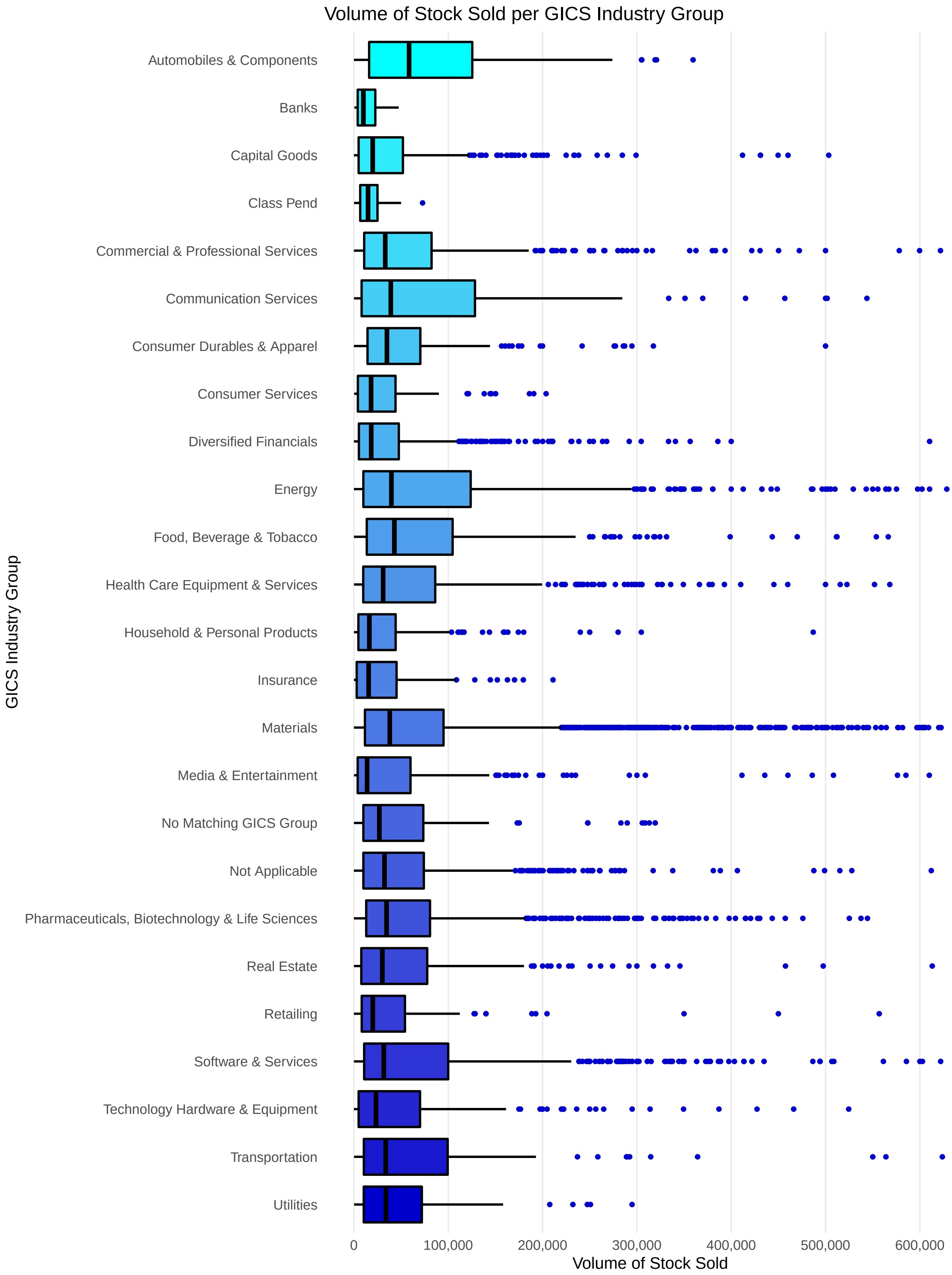
```
ASX_Lower_Vol <- summarise(group_by(ASX_Data_Lower,
                                    GICS_industry_group),
                           Mean_Vol = mean(Volume),
                           n_Companies = n())

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.

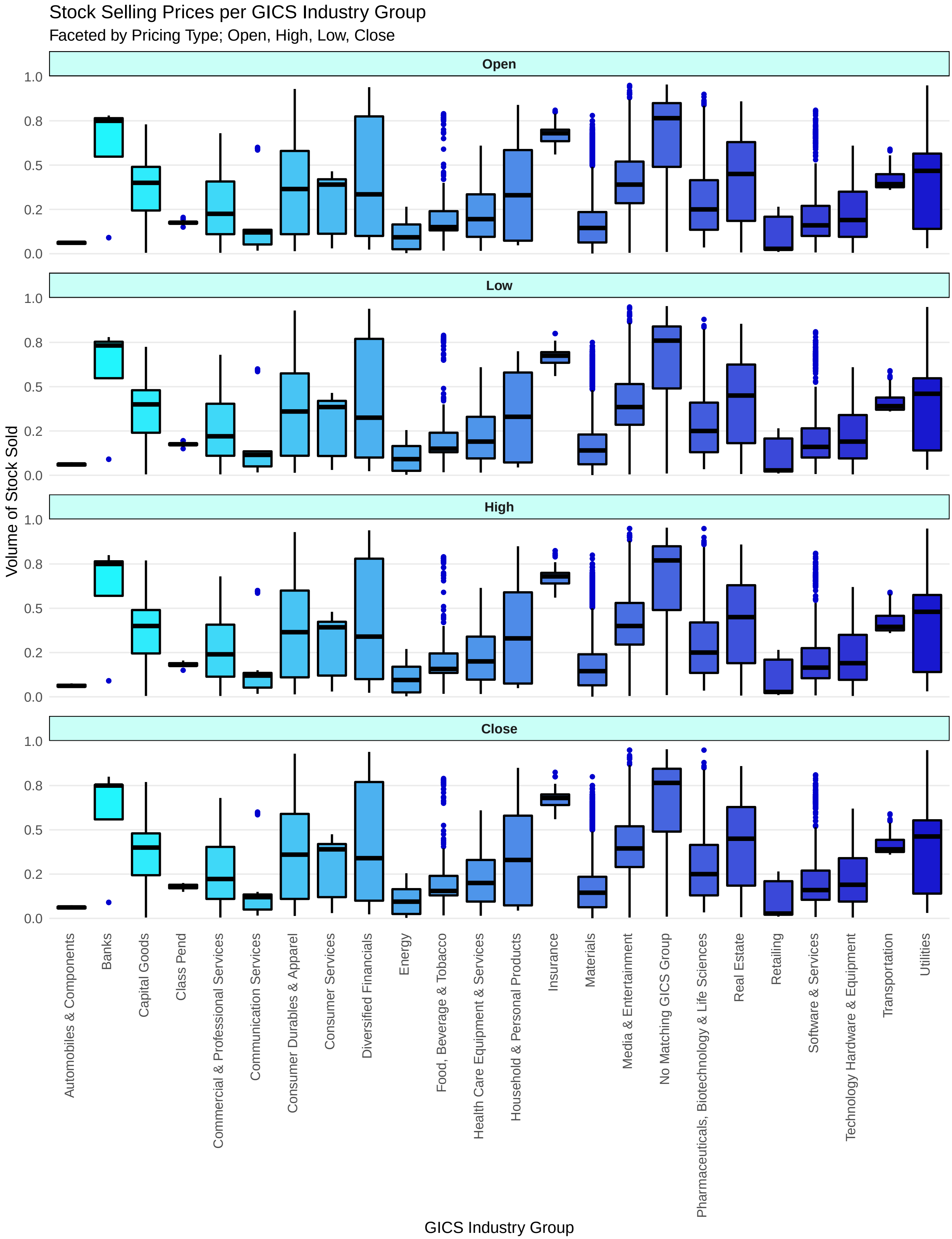
```
ggplot(ASX_Data_Lower) +
  geom_boxplot(aes(x = fct_rev(GICS_industry_group), y = Volume,
                   fill = GICS_industry_group),
               show.legend = F, col = "black",
               size = 0.8,
               outlier.size = 1.25,
               outlier.colour = "blue3") +
  scale_x_discrete("GICS Industry Group") +
  scale_y_continuous("Volume of Stock Sold",
                     labels = comma,
                     breaks = seq(0, max(ASX_Data_Lower$Volume),
                                  100000)) +
  scale_fill_manual(values = fill_grad) +
  labs(title = "Volume of Stock Sold 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 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

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 = 0.8,
               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",
                     labels = comma_format(accuracy = 0.1)) +
  scale_fill_manual(values = fill_grad) +
  labs(title = "Stock Selling Prices 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 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
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>