Title of the Minor Project:	Stock Market Position Optimization
Name of Team Player – 1	Ch. Chandana
Roll No of Team Player - 1	160121771076
Name of Team Player – 2	P. Kathyayini
Roll No of Team Player - 2	160121771088
Name of Team Player – 3	A. Tejas
Roll No of Team Player - 3	160121771095
Title of the Data Set	BankNifty, FinNifty, Nifty50 datasets by NSE

Description of Dataset:

- **Size of the dataset:**

```
> print(dim(data))
[1] 4090    7
> print(nrow(data))
[1] 4090
> print(ncol(df))
[1] 7
```

- **URL of the dataset:** https://finance.yahoo.com
- **Summary / Description and Domain of each attribute:**

Names of Datasets (Stocks) Considered for Analysis:

1. **BANKNIFTY:** an index comprised of the most liquid and large capitalised Indian banking stocks.
2. **NIFTY50:** an index that represents the performance of the top 50 companies listed on the National Stock Exchange (NSE) of India.
3. **FINNIFTY:** an index that includes the stock values of various companies that are part of the Indian financial sector.

Note: All prices are given in INR

Columns:

1. Date: Given in "month dd, yyyy" format (eg: Mar 28, 2024)
2. Open: The opening price of the stock for the corresponding date, given upto 2 decimal points
3. Close: The closing price of the stock for the corresponding date, given upto 2 decimal points
4. High: The highest price reached by the stock on the corresponding date, given upto 2 decimal points
5. Low: The lowest price reached by the stock on the corresponding date, given upto 2 decimal points
6. Adj. Close: The closing price after adjustments for all applicable splits and dividend distributions.
7. Volumes: Number of options, contracts bought or sold on a given trading day

Description of Task#1: Changing the data type of values in some columns

Required Pre-Processing: use as.numeric()

Attributes involved: Open, High, Low, Close, Adj.Close, Volume

R code with necessary comments:

```
file_path <- "C:\\Users\\KATHYAYINI PASUNURI\\OneDrive\\Documents\\My projects\\7th_MinorProjec
data <- read.csv(file_path)
numeric_cols <- c('Open', 'High', 'Low', 'Close', 'Adj.Close', 'Volume')
data[numeric_cols] <- lapply(data[numeric_cols], as.numeric)
str(data)
print(data)
```

Output:

```
> print(data)
          Date     Open     High      Low    Close Adj.Close Volume
1   2007-09-17  6898.00  6977.20  6843.00  6897.10  6897.020      0
2   2007-09-18  6921.15  7078.95  6883.60  7059.65  7059.568      0
3   2007-09-19  7111.00  7419.35  7111.00  7401.85  7401.764      0
4   2007-09-20  7404.95  7462.90  7343.60  7390.15  7390.064      0
5   2007-09-21  7378.30  7506.35  7367.15  7464.50  7464.413      0
6   2007-09-24  7514.40  7661.05  7514.40  7650.90  7650.811      0
7   2007-09-25  7658.50  7694.25  7490.20  7629.15  7629.061      0
8   2007-09-26  7647.10  7829.85  7591.80  7755.90  7755.810      0
9   2007-09-27  7804.55  7866.50  7747.10  7833.65  7833.559      0
10  2007-09-28  7838.25  8082.85  7836.05  8042.20  8042.107      0
11  2007-10-01  8008.55  8085.15  7913.30  7987.50  7987.407      0
12  2007-10-03  8029.80  8235.80  7820.25  8097.90  8097.806      0
13  2007-10-04  8083.30  8086.70  7828.65  8035.90  8035.807      0
14  2007-10-05  8038.10  8066.55  7789.70  7845.25  7845.159      0
15  2007-10-08  7853.15  7935.45  7516.45  7626.40  7626.311      0
16  2007-10-09  7580.90  7916.45  7535.05  7895.85  7895.758      0
17  2007-10-10  7960.65  8081.05  7907.35  8030.65  8030.557      0
18  2007-10-11  8054.30  8177.75  8005.50  8158.80  8158.705      0
19  2007-10-12  8093.65  8132.10  7889.80  7934.00  7933.908      0
20  2007-10-15  7962.55  8306.35  7962.55  8286.30  8286.203      0
```

Description of Task#2: Finding number of missing values in data

Required Pre-Processing: use is.na() and count values

Attributes involved: Date, Open, High, Low, Close, Adj.Close, Volume

R code with necessary comments:

```
summary(data)
na_counts <- colSums(is.na(data))
print(na_counts)
```

Output:

```
> summary(data)
     Date                Open            High            Low             Close         Adj.Close         Volume
 Length:4090        Min.   : 3385   Min.   : 3447   Min.   : 3315   Min.   : 3340   Min.   : 3340   Min.   :0.000e+00
 Class :character   1st Qu.:10308   1st Qu.:10414   1st Qu.:10180   1st Qu.:10289   1st Qu.:10289   1st Qu.:0.000e+00
 Mode  :character   Median :18386   Median :18539   Median :18227   Median :18373   Median :18372   Median :0.000e+00
                    Mean   :20869   Mean   :21033   Mean   :20676   Mean   :20856   Mean   :20856   Mean   :6.583e+05
                    3rd Qu.:30228   3rd Qu.:30480   3rd Qu.:29960   3rd Qu.:30216   3rd Qu.:30215   3rd Qu.:4.165e+04
                    Max.   :48880   Max.   :49057   Max.   :48669   Max.   :48987   Max.   :48987   Max.   :1.798e+09
                    NA's   :303     NA's   :303     NA's   :303     NA's   :303     NA's   :303     NA's   :303
> na_counts <- colSums(is.na(data))
> print(na_counts)
     Date      Open      High       Low     Close Adj.Close    Volume
        0       303       303       303       303       303       303
```

Description of Task#3: Imputing missing values with moving averages of range=35

Required Pre-Processing: use mean() function and check the missing values with is.na=TRUE and replace using a defined function.

Attributes involved: Open, High, Low, Close, Adj.Close, Volume

R code with necessary comments:

```r
#Imputing Missing values with moving averages
# Define a function to replace NA values with local mean within a specified range
impute_local_mean <- function(x, range = 35) {
  # Create a vector to store imputed values
  imputed_values <- numeric(length(x))

  # Iterate over each element in the vector
  for (i in seq_along(x)) {
    if (is.na(x[i])) {
      # Calculate the local mean within the specified range
      lower_bound <- max(1, i - range)
      upper_bound <- min(length(x), i + range)
      local_values <- x[lower_bound:upper_bound]
      imputed_values[i] <- mean(local_values, na.rm = TRUE)
    } else {
      # Keep the original value if it's not NA
      imputed_values[i] <- x[i]
    }
  }

  return(imputed_values)
}
# Apply the custom imputation function to each column of the dataframe
clean_data <- as.data.frame(lapply(data, impute_local_mean))
# Note: Replace 'data' with the name of your dataframe containing NA values
clean_data
na_counts <- colSums(is.na(clean_data))
print(na_counts)
```

Output:

```
> clean_data
          Date     Open     High      Low    Close Adj.Close Volume
1   2007-09-17  6898.00  6977.20  6843.00  6897.10  6897.020      0
2   2007-09-18  6921.15  7078.95  6883.60  7059.65  7059.568      0
3   2007-09-19  7111.00  7419.35  7111.00  7401.85  7401.764      0
4   2007-09-20  7404.95  7462.90  7343.60  7390.15  7390.064      0
5   2007-09-21  7378.30  7506.35  7367.15  7464.50  7464.413      0
6   2007-09-24  7514.40  7661.05  7514.40  7650.90  7650.811      0
7   2007-09-25  7658.50  7694.25  7490.20  7629.15  7629.061      0
8   2007-09-26  7647.10  7829.85  7591.80  7755.90  7755.810      0
9   2007-09-27  7804.55  7866.50  7747.10  7833.65  7833.559      0
10  2007-09-28  7838.25  8082.85  7836.05  8042.20  8042.107      0
11  2007-10-01  8008.55  8085.15  7913.30  7987.50  7987.407      0
12  2007-10-03  8029.80  8235.80  7820.25  8097.90  8097.806      0
13  2007-10-04  8083.30  8086.70  7828.65  8035.90  8035.807      0
14  2007-10-05  8038.10  8066.55  7789.70  7845.25  7845.159      0
15  2007-10-08  7853.15  7935.45  7516.45  7626.40  7626.311      0
16  2007-10-09  7580.90  7916.45  7535.05  7895.85  7895.758      0
17  2007-10-10  7960.65  8081.05  7907.35  8030.65  8030.557      0
18  2007-10-11  8054.30  8177.75  8005.50  8158.80  8158.705      0
19  2007-10-12  8093.65  8132.10  7889.80  7934.00  7933.908      0
20  2007-10-15  7962.55  8306.35  7962.55  8286.30  8286.203      0
21  2007-10-16  8361.40  8491.65  8240.30  8452.20  8452.102      0
22  2007-10-17  8071.55  8218.10  7641.50  8099.90  8099.806      0
23  2007-10-18  8055.15  8192.45  7519.70  7608.75  7608.662      0
24  2007-10-19  7637.40  7718.90  7279.30  7423.80  7423.713      0
25  2007-10-22  7374.35  7621.65  7289.20  7568.00  7567.912      0
26  2007-10-23  7621.05  8118.75  7621.05  8101.10  8101.006      0

> na_counts <- colSums(is.na(clean_data))
> print(na_counts)
     Date      Open      High       Low     Close Adj.Close    Volume
        0         0         0         0         0         0         0
```

Description of Task#4: Changing the 0 values in Volume column to a number

Required Pre-Processing: use a function to insert a value between 200000 to 300000

Attributes involved: Volume

R code with necessary comments:

```r
# Set seed for reproducibility (optional)
set.seed(123)

# Identify zero values in the volume column
zero_indices <- which(clean_data$Volume == 0)

# Calculate the number of zero values
num_zeros <- length(zero_indices)

# Generate random integers between 200,000 and 300,000
random_numbers <- sample(200000:500000, size = num_zeros, replace = TRUE)

# Replace zero values in the volume column with random numbers
clean_data$Volume[zero_indices] <- random_numbers

# Convert the volume column to integer type
clean_data$Volume <- as.integer(clean_data$Volume)


# Display the updated dataset
print(clean_data)
```

Output:

```
> print(clean_data)
          Date     Open     High      Low    Close Adj.Close Volume
1   2007-09-17  6898.00  6977.20  6843.00  6897.10  6897.020 388941
2   2007-09-18  6921.15  7078.95  6883.60  7059.65  7059.568 334057
3   2007-09-19  7111.00  7419.35  7111.00  7401.85  7401.764 324021
4   2007-09-20  7404.95  7462.90  7343.60  7390.15  7390.064 360996
5   2007-09-21  7378.30  7506.35  7367.15  7464.50  7464.413 426317
6   2007-09-24  7514.40  7661.05  7514.40  7650.90  7650.811 324506
7   2007-09-25  7658.50  7694.25  7490.20  7629.15  7629.061 393626
8   2007-09-26  7647.10  7829.85  7591.80  7755.90  7755.810 245403
9   2007-09-27  7804.55  7866.50  7747.10  7833.65  7833.559 265160
10  2007-09-28  7838.25  8082.85  7836.05  8042.20  8042.107 259133
11  2007-10-01  8008.55  8085.15  7913.30  7987.50  7987.407 383203
12  2007-10-03  8029.80  8235.80  7820.25  8097.90  8097.806 345254
13  2007-10-04  8083.30  8086.70  7828.65  8035.90  8035.807 477323
14  2007-10-05  8038.10  8066.55  7789.70  7845.25  7845.159 289708
15  2007-10-08  7853.15  7935.45  7516.45  7626.40  7626.311 230537
16  2007-10-09  7580.90  7916.45  7535.05  7895.85  7895.758 425588
17  2007-10-10  7960.65  8081.05  7907.35  8030.65  8030.557 344607
18  2007-10-11  8054.30  8177.75  8005.50  8158.80  8158.705 290076
19  2007-10-12  8093.65  8132.10  7889.80  7934.00  7933.908 468359
20  2007-10-15  7962.55  8306.35  7962.55  8286.30  8286.203 414590
```

Description of Task#5: Adding a new column “Today_point_difference” to calculate difference between that day’s opening and closing price

Required Pre-Processing: use subtraction operator

Attributes involved: Open, Close

R code with necessary comments:

```
# Add a new column 'difference' to calculate the price difference
clean_data$Today_point_difference <- clean_data$Close - clean_data$Open

# Display the updated dataset with the new 'difference' column
print(clean_data)
```

Output:

```
> print(clean_data)
          Date     Open     High      Low    Close Adj.Close Volume Today_point_difference
1   2007-09-17  6898.00  6977.20  6843.00  6897.10  6897.020 388941              -0.899902
2   2007-09-18  6921.15  7078.95  6883.60  7059.65  7059.568 334057             138.500000
3   2007-09-19  7111.00  7419.35  7111.00  7401.85  7401.764 324021             290.850098
4   2007-09-20  7404.95  7462.90  7343.60  7390.15  7390.064 360996             -14.800293
5   2007-09-21  7378.30  7506.35  7367.15  7464.50  7464.413 426317              86.200195
6   2007-09-24  7514.40  7661.05  7514.40  7650.90  7650.811 324506             136.500000
7   2007-09-25  7658.50  7694.25  7490.20  7629.15  7629.061 393626             -29.350098
8   2007-09-26  7647.10  7829.85  7591.80  7755.90  7755.810 245403             108.799804
9   2007-09-27  7804.55  7866.50  7747.10  7833.65  7833.559 265160              29.100097
10  2007-09-28  7838.25  8082.85  7836.05  8042.20  8042.107 259133             203.950195
11  2007-10-01  8008.55  8085.15  7913.30  7987.50  7987.407 383203             -21.049805
12  2007-10-03  8029.80  8235.80  7820.25  8097.90  8097.806 345254              68.100097
13  2007-10-04  8083.30  8086.70  7828.65  8035.90  8035.807 477323             -47.399903
14  2007-10-05  8038.10  8066.55  7789.70  7845.25  7845.159 289708            -192.850098
15  2007-10-08  7853.15  7935.45  7516.45  7626.40  7626.311 230537            -226.750000
16  2007-10-09  7580.90  7916.45  7535.05  7895.85  7895.758 425588             314.950196
17  2007-10-10  7960.65  8081.05  7907.35  8030.65  8030.557 344607              70.000000
18  2007-10-11  8054.30  8177.75  8005.50  8158.80  8158.705 290076             104.500000
19  2007-10-12  8093.65  8132.10  7889.80  7934.00  7933.908 468359            -159.649902
20  2007-10-15  7962.55  8306.35  7962.55  8286.30  8286.203 414590             323.750000
```

Description of Task#6: Adding a new column “closing_opening_difference” to calculate difference between today’s opening and yesterday’s closing price

Required Pre-Processing: use mutate and lag functions to calculate the new column

Attributes involved: Open, Close

R code with necessary comments:

```
install.packages("dplyr")

library(dplyr)

# Sort the dataframe by date (if not already sorted)
clean_data <- clean_data[order(clean_data$Date), ]

# Calculate the difference between yesterday's close and today's open
clean_data <- clean_data %>%
  mutate(yesterday_close = lag(Close, default = first(Close)),  # Get yesterday's closing price
         today_open = Open,  # Today's opening price
         price_difference =  yesterday_close - today_open )  # Calculate the price difference

# Rename the new column for clarity
colnames(clean_data)[which(names(clean_data) == "price_difference")] <- "closing_opening_difference"

# Display the updated dataframe
print(clean_data)
```

Output:

```
> print(clean_data)
         Date     Open     High      Low    Close Adj.Close Volume Today_point_difference yesterday_close today_open closing_opening_difference
1  2007-09-17  6898.00  6977.20  6843.00  6897.10  6897.020 388941              -0.899902         6897.10    6898.00                 -0.899902
2  2007-09-18  6921.15  7078.95  6883.60  7059.65  7059.568 334057             138.500000         6897.10    6921.15                -24.049804
3  2007-09-19  7111.00  7419.35  7111.00  7401.85  7401.764 324021             290.850098         7059.65    7111.00                -51.350098
4  2007-09-20  7404.95  7462.90  7343.60  7390.15  7390.064 360996             -14.800293         7401.85    7404.95                 -3.100097
5  2007-09-21  7378.30  7506.35  7367.15  7464.50  7464.413 426317              86.200195         7390.15    7378.30                 11.850097
6  2007-09-24  7514.40  7661.05  7514.40  7650.90  7650.811 324506             136.500000         7464.50    7514.40                -49.899902
7  2007-09-25  7658.50  7694.25  7490.20  7629.15  7629.061 393626             -29.350098         7650.90    7658.50                 -7.600098
8  2007-09-26  7647.10  7829.85  7591.80  7755.90  7755.810 245403             108.799804         7629.15    7647.10                -17.950196
9  2007-09-27  7804.55  7866.50  7747.10  7833.65  7833.559 265160              29.100097         7755.90    7804.55                -48.649903
10 2007-09-28  7838.25  8082.85  7836.05  8042.20  8042.107 259133             203.950195         7833.65    7838.25                 -4.600098
11 2007-10-01  8008.55  8085.15  7913.30  7987.50  7987.407 383203             -21.049805         8042.20    8008.55                 33.650390
12 2007-10-03  8029.80  8235.80  7820.25  8097.90  8097.806 345254              68.100097         7987.50    8029.80                -42.299805
13 2007-10-04  8083.30  8086.70  7828.65  8035.90  8035.807 477323             -47.399903         8097.90    8083.30                 14.600097
14 2007-10-05  8038.10  8066.55  7789.70  7845.25  7845.159 289708            -192.850098         8035.90    8038.10                 -2.200196
15 2007-10-08  7853.15  7935.45  7516.45  7626.40  7626.311 230537            -226.750000         7845.25    7853.15                 -7.899902
16 2007-10-09  7580.90  7916.45  7535.05  7895.85  7895.758 425588             314.950196         7626.40    7580.90                 45.500000
17 2007-10-10  7960.65  8081.05  7907.35  8030.65  8030.557 344607              70.000000         7895.85    7960.65                -64.799804
18 2007-10-11  8054.30  8177.75  8005.50  8158.80  8158.705 290076             104.500000         8030.65    8054.30                -23.649903
19 2007-10-12  8093.65  8132.10  7889.80  7934.00  7933.908 468359            -159.649902         8158.80    8093.65                 65.149903
20 2007-10-15  7962.55  8306.35  7962.55  8286.30  8286.203 414590             323.750000         7934.00    7962.55                -28.549805
```

Description of Task#7: Barplot between closing price and volumes

Required Pre-Processing: use barplot function

Attributes involved: Close, Volume

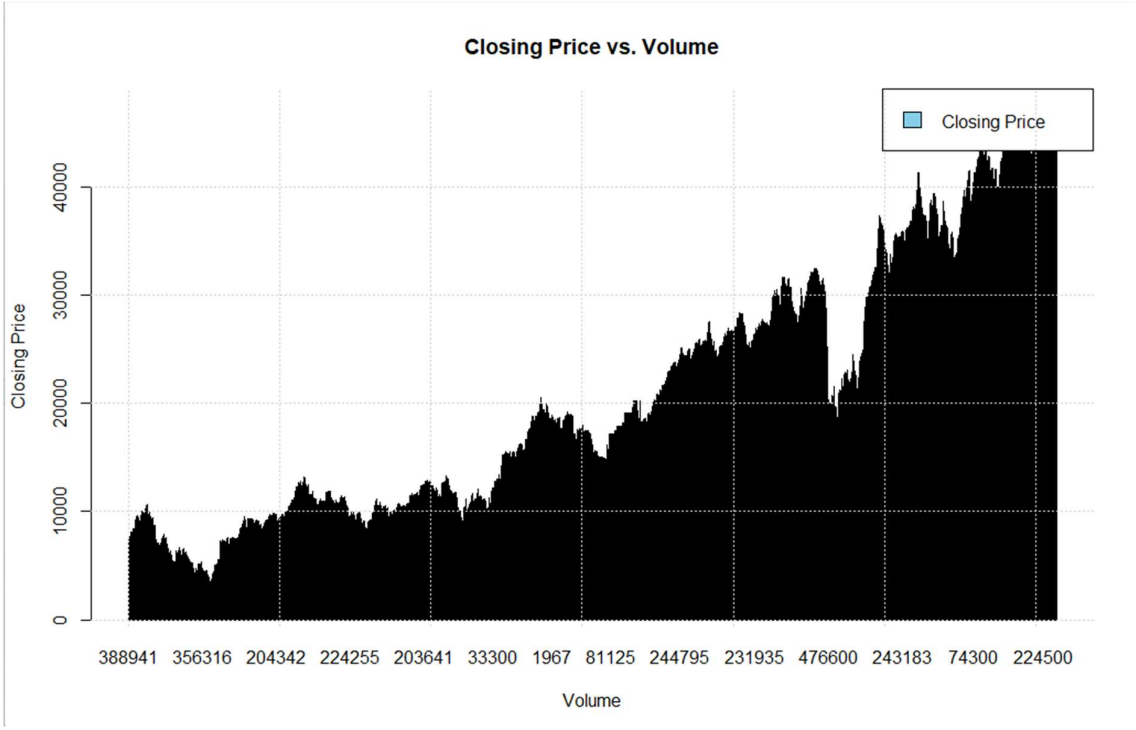
R code with necessary comments:

```r
# Plotting a bar graph of Closing Price and Volume
barplot(clean_data$Close, names.arg = clean_data$Volume, xlab = "Volume", ylab = "Closing Price",
        main = "Closing Price vs. Volume", col = "skyblue", border = "black",
        space = 0.5)

# Adding a legend
legend("topright", legend = "Closing Price", fill = "skyblue", border = "black")

# Adding gridlines for clarity (optional)
grid()
```

Output:



Description of Task#8: Scatterplot of Closing Pricce vs Opening Price

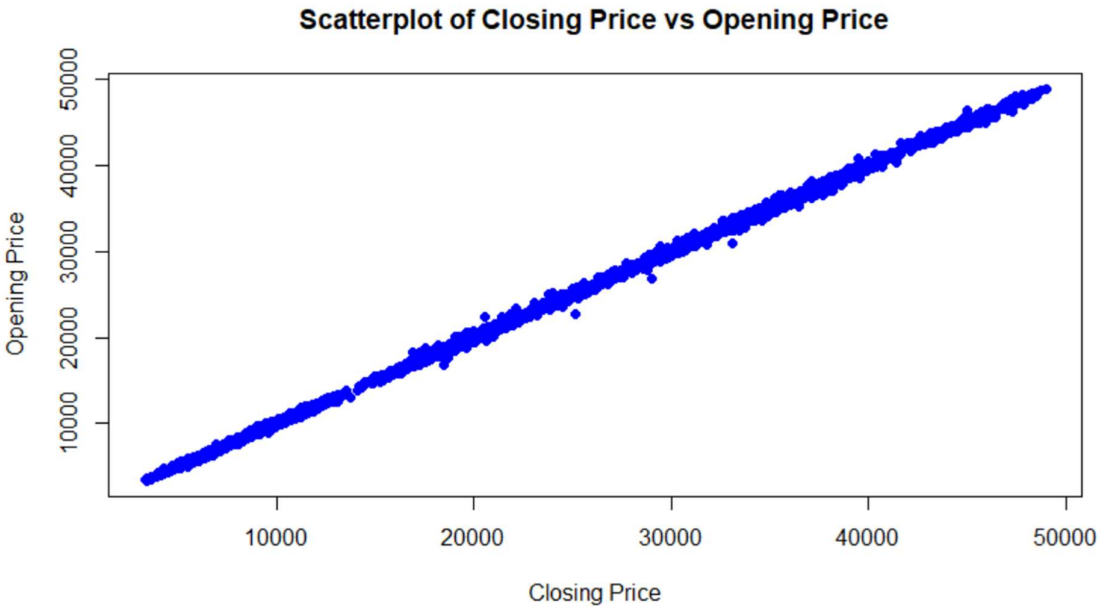
Required Pre-Processing: use plot function

Attributes involved: Open, Close

R code with necessary comments:

```r
# Customized scatterplot
plot(clean_data$Close, clean_data$Open,
     col = "blue",  # Change point color
     pch = 16,      # Use solid circles for points
     xlab = "Closing Price",
     ylab = "Opening Price",
     main = "Scatterplot of Closing Price vs Opening Price")
```

Output:


Scatterplot of Closing Price vs Opening Price
Opening Price
10000
20000
30000
40000
50000
Closing Price
10000
20000
30000
40000
50000