Dow Jones Index Case Study

Eva Beyebach, Pablo Chacon, & Danya Saed

I. Executive Summary

This Case study focused on predicting weekly stock returns using historical Dow Jones Index Data. After cleaning and exploring the data, we performed three different models (linear regression, SVR, and Decision Trees) and the aim was to get the best model in predicting `percent_change_next_weeks_price` based on the independent variables.

The Linear Regression model exhibited the most precise predictions with an RMSE of 3.29, marginally outperforming the SVM and Decision Tree models, which recorded RMSEs of 3.38 and 3.47, respectively. The Linear Model also helped eliminate collinearity issues between independent variables and helped us determine the inputs for generating the final three models (Week + high + percent_change_price + percent_change_volume_over_last_wk + volume).

Even though the RMSE for the linear model was the lowest, we decided to choose the SVM model. The model itself does a better job of considering the categorical and continuous variables and the computing power of SVMs is much greater. Additionally, the SVM boasts the ability to employ different kernels that allow for greater computing power for non-linear data (stock data is unpredictable) and this model could be an advantage for that.

CAPM was also utilized in this case study to assess the given stocks' risk compared to the market and get the best return for each stock. After getting all beta and return results, we chose KO, MCD, and MRK to be the best stocks. They all had a beta lower than 1 (less risky than the market) and better returns than the other stocks.

II. The Problem

The research at hand addresses the challenge of financial forecasting by focusing on the prediction of weekly stock returns using historical data. The central aim is to build a predictive model that can accurately forecast the highest rate of return for the following week, emphasizing the metric percent_change_next_weeks_price. To establish the model's efficacy, the available data is segmented into two quarters: the first quarter's data is utilized for model training, while the second quarter's data is employed to validate the model's predictive prowess. Among the techniques under consideration are Linear Regression, Decision Trees, and Support Vector Regression, with a further exploration of the Capital Asset Pricing Model for evaluating stock risks. An intriguing aspect of the dataset is the uniform stock price drop during the week ending May 27, 2011, which may be integral to the model's risk assessment capabilities.

The goal of this report extends beyond the creation of an accurate predictive model; it includes a thorough examination of various modeling techniques and the impact of lagged variables on financial forecasting. This includes a methodical analysis of whether incorporating lagged variables can significantly improve the model's predictive accuracy. The project aims to not only achieve high predictive accuracy but also to contribute valuable insights into the mechanisms of financial prediction and the potential implications of temporal data patterns in stock market analysis.

III. Review of Related Literature

In the realm of financial analytics, recent explorations into the "Dow Jones Index case" dataset have marked a significant leap forward in our understanding of stock market behaviors and predictive modeling techniques. The pioneering analysis by Brown, Pelosi, and Dirska in 2013 utilized this dataset's weekly stock performance records, including pivotal variables like price changes, volume shifts, and dividend yields, to enhance algorithmic forecasting of stock returns. This foundational work has inspired a breadth of studies, leveraging cutting-edge methodologies such as genetic algorithms to refine financial forecasting models, as highlighted in seminal contributions to the UCI Machine Learning Repository.

Further investigations, including those disseminated via GitHub, have delved deep into the dataset to unearth correlations between stock characteristics and market outcomes, pioneering the development of predictive models that probe the undercurrents of market dynamics. Notably, the dataset's record of the universal stock downturn during the week concluding on May 27, 2011, provides a compelling case for risk assessment and insights into market responses to stress factors. These analyses underscore the role of the transaction volume as a significant predictor, revealing its substantial influence on stock performance and its potential to guide strategic investment decisions.

This body of work underscores the Dow Jones Index dataset's invaluable contribution to financial market research, algorithmic trading, and the advancement of predictive analytics. It equips researchers and practitioners with a rich dataset to inform and refine investment strategies, embodying a crucial step forward in the quest to demystify the complexities of stock market fluctuations and investor behaviors.

IV. Methodology

In our exploration of the "Dow Jones Index case," we employ a methodological approach that leans heavily on predictive analytics, using a blend of Linear Models, Linear Regression, Support Vector Regression (SVR), Decision Trees, and the Capital Asset Pricing Model (CAPM). Each of these methods brings a unique lens to the study of financial time-series data.

Linear models stand as the cornerstone of statistical modeling in finance, underpinning our efforts with their simplicity and interpretability. They work on the

assumption that there is a linear relationship between the dependent variables (in our case, percent_change_next_weeks_price) and one or more independent variables (such as volume and dividend returns). We leverage linear regression, a parametric approach within the linear models family, to quantify the strength of the influence that the independent variables have on the target variable. The goal is to fit a line, known as the regression line, through a scatter plot of data points that minimizes the sum of the squared residuals, providing a baseline model for stock price returns.

Support Vector Regression (SVR) is employed in our study to address the complexities of financial forecasting. It excels in deciphering non-linear patterns within the stock market data, efficiently handling outliers and market noise by mapping inputs into higher-dimensional spaces for linear analysis. This capability, coupled with SVR's use of kernel functions and its inherent robustness against financial volatility, positions it as a powerful tool in our analytical arsenal, providing depth and a global perspective to our predictive endeavors.

Decision Trees emerge as a methodological choice in our study due to their unique capability to segment the Dow Jones dataset into decision-based subsets. This non-parametric model thrives on classifying data by the features' intrinsic values, creating a tree-like structure that helps to elucidate the conditional relationships between market indicators and future stock performance. Despite their insightful delineation of variable interactions, Decision Trees demand judicious calibration, like pruning and cross-validation, to mitigate their tendency toward overfitting and ensure a more generalizable model.

Lastly, the Capital Asset Pricing Model (CAPM) brings a theoretical framework to assess the risk and expected return of the stocks. While not a predictive model in the conventional sense, CAPM provides a perspective on how stock prices should behave, given the risk-free rate, the stock's beta (a measure of its volatility in relation to the market), and the expected market return. This allows us to juxtapose our predictive insights with theoretical expectations, yielding a holistic understanding of the stocks in the Dow Jones Index.

Together, these methods form a multifaceted approach to dissecting and predicting the nuances of stock price movements, each bringing its strengths and considerations to the fore. The integration of their insights will be instrumental in crafting a robust predictive model for the "Dow Jones Index case" study.

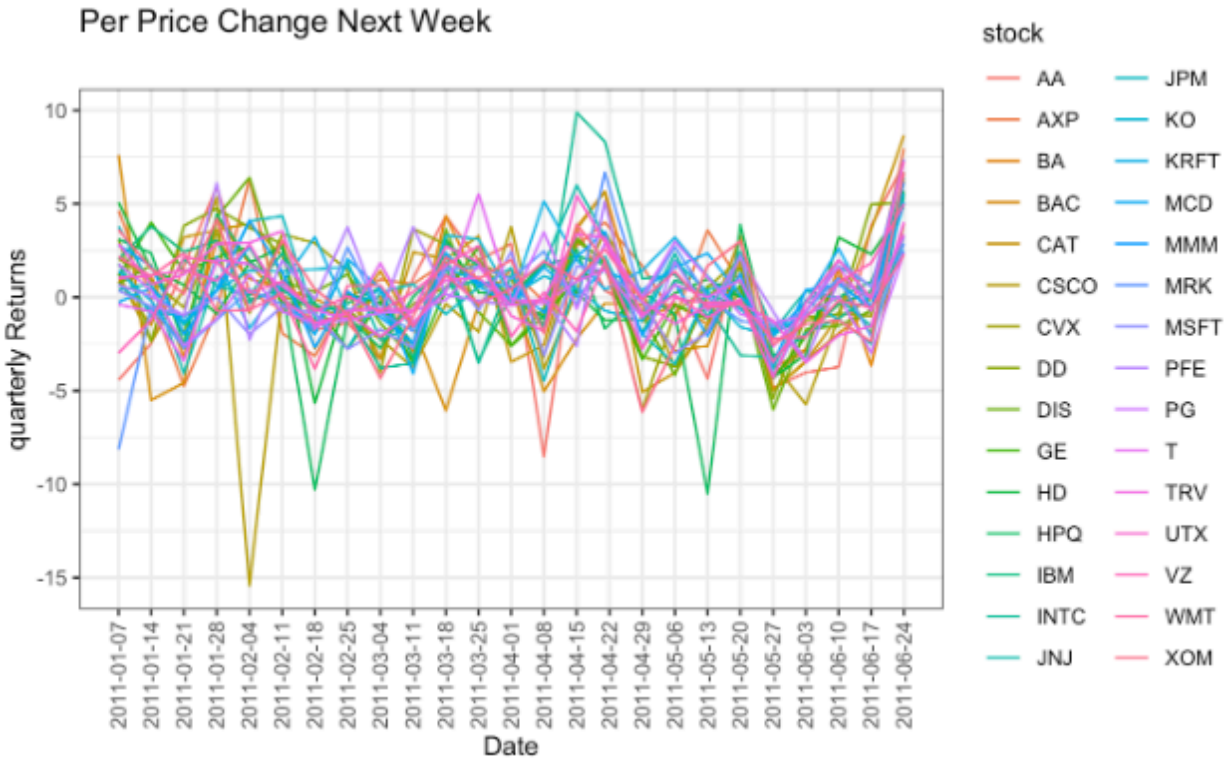
V. Data

The data set consists of 30 stocks in the Dow Jones Industrial Index. There are a total of 750 observations with 16 individual variables. Below is a summary of the data set and the variables:

```
   quarter        stock              date              open              high              low               close
Min.   :1.00   Length:750         Length:750         Length:750        Length:750        Length:750        Length:750
1st Qu.:1.00   Class :character   Class :character   Class :character  Class :character  Class :character  Class :character
Median :2.00   Mode  :character   Mode  :character   Mode  :character  Mode  :character  Mode  :character  Mode  :character
Mean   :1.52
3rd Qu.:2.00
Max.   :2.00

     volume          percent_change_price percent_change_volume_over_last_wk previous_weeks_volume next_weeks_open    next_weeks_close
Min.   :9.719e+06   Min.   :-15.42290    Min.   :-61.4332                   Min.   :9.719e+06     Length:750         Length:750
1st Qu.:3.087e+07   1st Qu.: -1.28805    1st Qu.:-19.8043                   1st Qu.:3.068e+07     Class :character   Class :character
Median :5.306e+07   Median :  0.00000    Median :  0.5126                   Median :5.295e+07     Mode  :character   Mode  :character
Mean   :1.175e+08   Mean   :  0.05026    Mean   :  5.5936                   Mean   :1.174e+08
3rd Qu.:1.327e+08   3rd Qu.:  1.65089    3rd Qu.: 21.8006                   3rd Qu.:1.333e+08
Max.   :1.453e+09   Max.   :  9.88223    Max.   :327.4089                   Max.   :1.453e+09
                                         NA's   :30                         NA's   :30
percent_change_next_weeks_price days_to_next_dividend percent_return_next_dividend
Min.   :-15.4229                Min.   :  0.00        Min.   :0.06557
1st Qu.: -1.2221                1st Qu.: 24.00        1st Qu.:0.53455
Median :  0.1012                Median : 47.00        Median :0.68107
Mean   :  0.2385                Mean   : 52.53        Mean   :0.69183
3rd Qu.:  1.8456                3rd Qu.: 69.00        3rd Qu.:0.85429
Max.   :  9.8822                Max.   :336.00        Max.   :1.56421
```

The first visualization created as an overview of the Per Price Change Next week variable:



Overall, all the stocks exhibit similar movements between 2011-01-07 and 2011-06-24. There are rises and declines in each stock that may not completely imitate the others, but there is a large number of stocks that ended higher in the 5% - 10% change in the next week's price.

The majority of the variables are numeric since the data mostly explores numeric changes in the stock's prices. The stock variable is text since it is the identifiable ticker of each stock. One of the first steps in the data-cleaning process was to look at any NA values. According to the data, there appeared to be 30 NA values in the percent_change_volume_over_last_wk and previous_weeks_volume variables, as seen below:

```
                quarter                   stock                    date                    open 
                      0                       0                       0                       0 
                   high                     low                   close                  volume 
                      0                       0                       0                       0 
  percent_change_price percent_change_volume_over_last_wk previous_weeks_volume      next_weeks_open 
                      0                      30                      30                       0 
     next_weeks_close    percent_change_next_weeks_price days_to_next_dividend percent_return_next_dividend 
                      0                       0                       0                       0 
```

Upon further analysis of these NA values, it was clear that they only appeared during the first for each ticker. This makes sense considering that there cannot be a change from the previous week for the first week of the study. However, these NA values must still be fixed, since there was a risk of the models performing inaccurately if the NA values were kept in the data. To fix this, the NA values were converted to 0.

```
```{r}
na_values <- dji[rowSums(is.na(dji)) >0,]
```
```

Another important step was changing character variables to numeric if needed. There were several variables in the data that were supposed to be numeric but were coded as characters. This would affect the model and results, so they needed to be converted:

```
quarter     stock         date                open             high             low              close            volume        
1:360   AA      : 25   Min.   :2011-01-07   Min.   : 10.59   Min.   : 10.94   Min.   : 10.40   Min.   : 10.52   Min.   :9.719e+06  
2:390   AXP     : 25   1st Qu.:2011-02-18   1st Qu.: 29.83   1st Qu.: 30.63   1st Qu.: 28.72   1st Qu.: 30.36   1st Qu.:3.087e+07  
        BA      : 25   Median :2011-04-01   Median : 45.97   Median : 46.88   Median : 44.80   Median : 45.93   Median :5.306e+07  
        BAC     : 25   Mean   :2011-03-31   Mean   : 53.65   Mean   : 54.67   Mean   : 52.64   Mean   : 53.73   Mean   :1.175e+08  
        CAT     : 25   3rd Qu.:2011-05-13   3rd Qu.: 72.72   3rd Qu.: 74.29   3rd Qu.: 71.04   3rd Qu.: 72.67   3rd Qu.:1.327e+08  
        CSCO    : 25   Max.   :2011-06-24   Max.   :172.11   Max.   :173.54   Max.   :167.82   Max.   :170.58   Max.   :1.453e+09  
        (Other):600                                                                                                                 
percent_change_price percent_change_volume_over_last_wk previous_weeks_volume next_weeks_open  next_weeks_close 
Min.   :-15.42290    Min.   :-61.4332                   Min.   :9.719e+06     Min.   : 10.52   Min.   : 10.52  
1st Qu.: -1.28805    1st Qu.:-19.8043                   1st Qu.:3.068e+07     1st Qu.: 30.32   1st Qu.: 30.46  
Median :  0.00000    Median :  0.5126                   Median :5.295e+07     Median : 46.02   Median : 46.12  
Mean   :  0.05026    Mean   :  5.5936                   Mean   :1.174e+08     Mean   : 53.70   Mean   : 53.89  
3rd Qu.:  1.65089    3rd Qu.: 21.8006                   3rd Qu.:1.333e+08     3rd Qu.: 72.72   3rd Qu.: 72.92  
Max.   :  9.88223    Max.   :327.4089                   Max.   :1.453e+09     Max.   :172.11   Max.   :174.54  
                     NA's   :30                         NA's   :30                                             
percent_change_next_weeks_price days_to_next_dividend percent_return_next_dividend
Min.   :-15.4229                Min.   :  0.00         Min.   :0.06557            
1st Qu.: -1.2221                1st Qu.: 24.00         1st Qu.:0.53455            
Median :  0.1012                Median : 47.00         Median :0.68107            
Mean   :  0.2385                Mean   : 52.53         Mean   :0.69183            
3rd Qu.:  1.8456                3rd Qu.: 69.00         3rd Qu.:0.85429            
Max.   :  9.8822                Max.   :336.00         Max.   :1.56421            
```

As can be seen above, the character variables in the previous summary were converted to numeric if the change was warranted. This will allow for visualizations and correlation plots to be created later on to reveal assumptions and findings on the data.

After cleaning the data, the next step was creating visualizations and tables to reveal important findings in the data. The first analyses were computing average dividend yield and average change in price comparing the 30 stocks to the average and exploring which stocks outperformed and underperformed these averages. This is important for any investor to understand depending on their investment approach. Passive investors may be satisfied with matching the index's performance and earning the average dividend yield for the average. If so, investing in a Dow Jones ETF may be an appropriate strategy. However, some investors may prefer a riskier and more active investment strategy, and this would require picking individual stocks that may have the potential to deliver higher returns than the index. For this reason, it would be important to look at which stocks could fit this approach:

```
avg_div <- aggregate(dji$percent_return_next_dividend, list(dji$stock), FUN=mean)
avg_div <- setNames(avg_div,
                    c("Stock","Return % on Next Divident"))
summary(avg_div)
```

```
     Stock    Return % on Next Divident
 AA      : 1   Min.   :0.07754
 AXP     : 1   1st Qu.:0.55817
 BA      : 1   Median :0.68573
 BAC     : 1   Mean   :0.69183
 CAT     : 1   3rd Qu.:0.86945
 CSCO    : 1   Max.   :1.45439
```

As seen in the image above, the average dividend yield was computed and then the average of those averages for each stock was computed. The average dividend yield for all the stocks across the period of the data was 0.69%. This will serve as the base result to compare against the other stocks. After this, a simple script was created to sort the stocks based on whether their average dividend return was higher or lower than 0.69%:

```
```{r}
above_avg <- avg_div[avg_div$`Return % on Next Divident` >0.69,]
print(paste(above_avg$Stock))
```
```

```
 [1] "CVX"  "DD"   "DIS"  "GE"   "INTC" "JNJ"  "KO"   "KRFT" "MCD"  "MRK"  "PFE"  "PG"   "T"    "VZ"
```

```{r}
below_avg <- avg_div[avg_div$`Return % on Next Divident`<=0.69,]
print(paste((below_avg$Stock)))
```

```
 [1] "AA"   "AXP"  "BA"   "BAC"  "CAT"  "CSCO" "HD"   "HPQ"  "IBM"  "JPM"  "MMM"  "MSFT" "TRV"  "UTX"  "WMT"  "XOM"
```

The first image contains the stocks that had a higher dividend return than the index, while the second image contains those stocks that yielded a lower dividend than the average. Investors would now be able to explore the overperformers and assess whether these could be attractive investments due to their higher dividend yield.

The same analysis was run but for the percent change in the price. The same logic was applied: compile the average percent change in price for each stock and compute the average of those averages. The average is shown below along with the stocks that outperformed the average.

```
     Stock    % Change in Price During Week
 AA      : 1   Min.   :-1.17830
 AXP     : 1   1st Qu.:-0.05772
 BA      : 1   Median : 0.13168
 BAC     : 1   Mean   : 0.05026
 CAT     : 1   3rd Qu.: 0.26621
 CSCO    : 1   Max.   : 0.59169
```

```
above_avg <- avg_chg[avg_chg$`% Change in Price During Week` >0.05026,]
print(paste(above_avg$Stock))
```

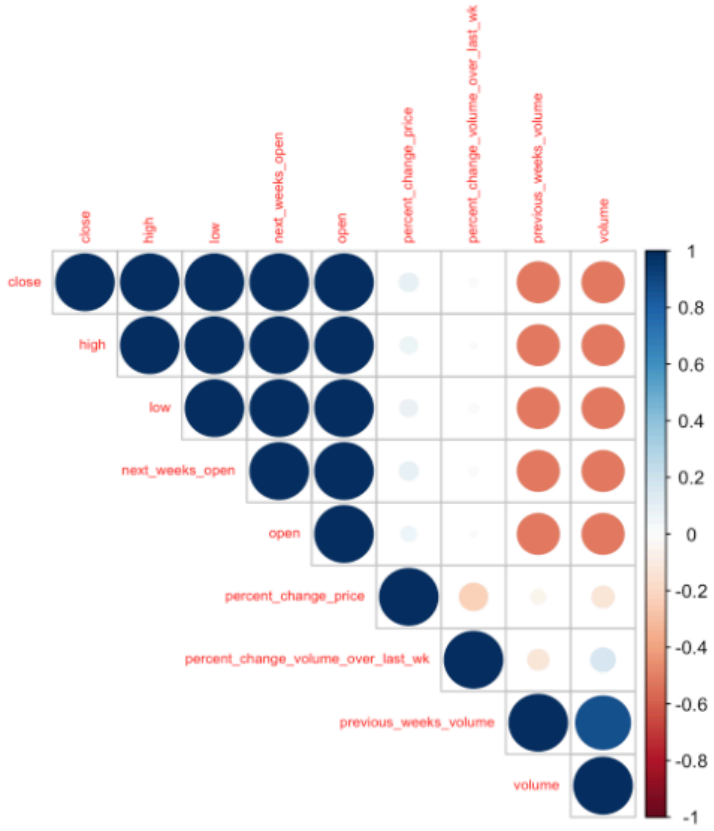
```
 [1] "AXP"  "BA"   "CAT"  "CVX"  "DD"   "DIS"  "HD"   "IBM"  "INTC" "JNJ"  "KO"   "KRFT" "MCD"  "MMM"  "PFE"  "T"    "TRV"  "UTX"  "WMT"
[20] "XOM"
```

```
below_avg <- avg_chg[avg_chg$`% Change in Price During Week`<=0.05026,]
print(paste((below_avg$Stock)))
```

```
 [1] "AA"   "BAC"  "CSCO" "GE"   "HPQ"  "JPM"  "MRK"  "MSFT" "PG"   "VZ"
```

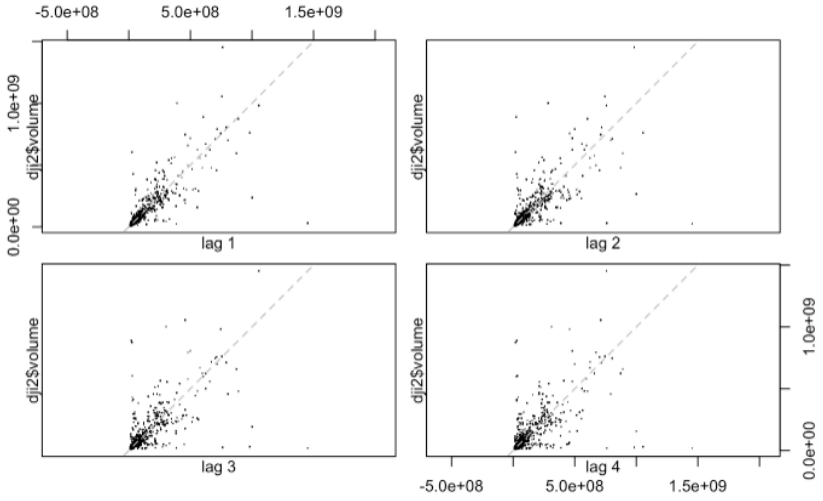
The stocks in the first image have seen a greater average rise in price during the week. Investors seeking a more active approach may choose to include these stocks in their portfolio.

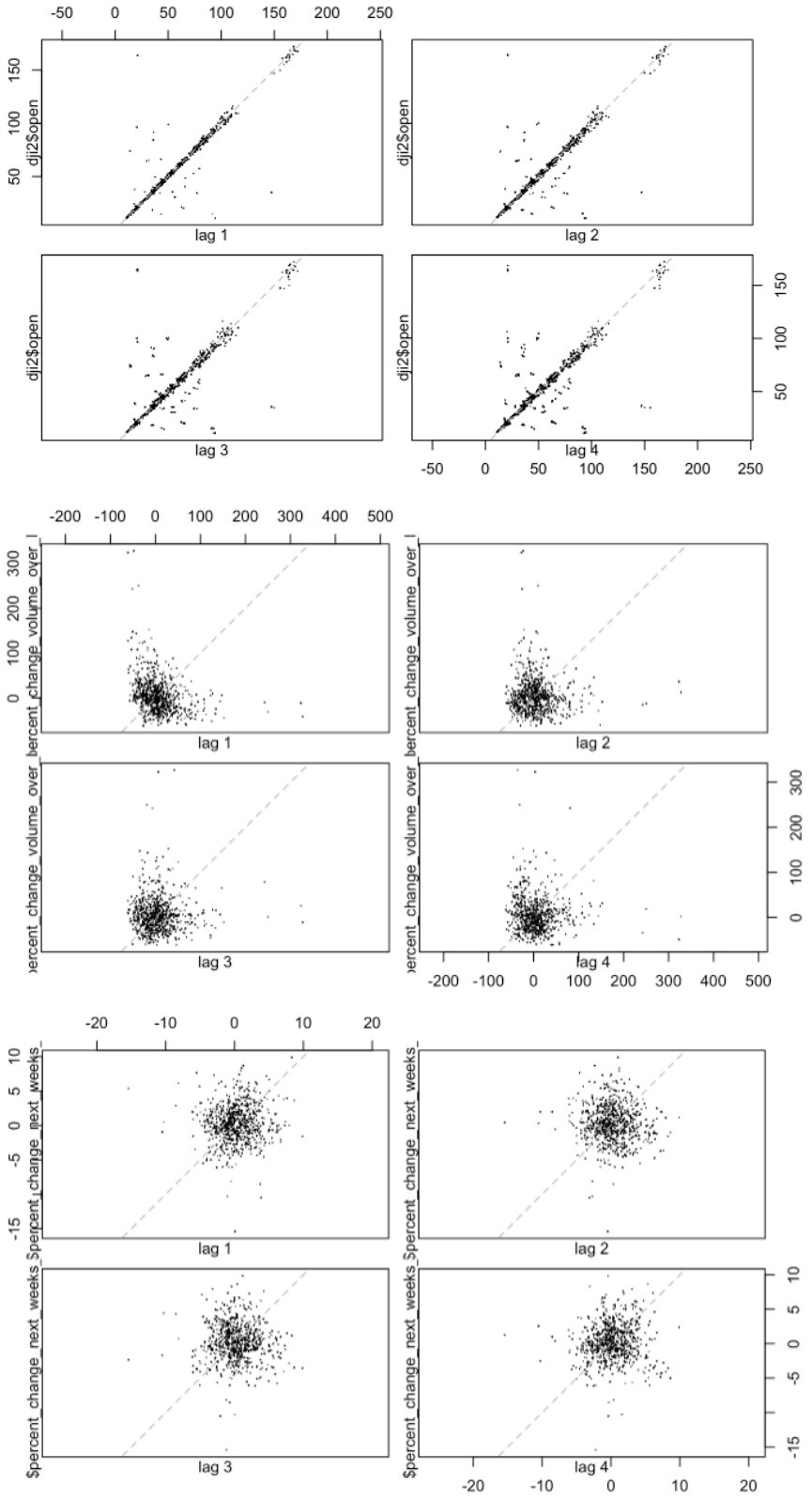
After this, a correlation plot between the numeric variables was created to reveal relationships between the variables and seek out any potential multi-collinearity concerns in the data:



There are a lot of very highly correlated variables in the data set. It looks like `close`, `high`, `open`, `low`, and `next_weeks_open`have a near-perfect correlation with each other. This makes sense considering that stocks will usually increase or decrease close to each of these values. For this reason, multi-collinearity must be considered as a potential issue in the models.

Lagged variables were also explored as part of this analysis, with open, volume, percent_change_volume_over_last_wk, and percent_change_next_weeks_price being analyzed:




dji2$open
lag 1
lag 2
lag 3
lag 4
percent_change_volume_over
percent_change_next_weeks

The best-autocorrelated variable in the analysis was open since it is the most linear of the graphs. Other variables don't appear to have a high autocorrelation, so there is no need to add lagged variables.

Models:

3 models were utilized to predict the variable percent_change_next_weeks_price: Linear Regression, Support Vector Machine, and a Decision Tree.

VI. Findings

The evaluation of three predictive models—Linear Regression, Support Vector Machine (SVM), and Decision Tree—has demonstrated a closely matched level of accuracy, as evidenced by the Root Mean Square Error (RMSE) metrics for each. Linear Regression leads the trio with the most accurate predictions, posting an RMSE of 3.29. This is closely followed by the SVM and Decision Tree models, which report RMSEs of 3.38 and 3.47, respectively. These marginal differences in RMSE values suggest that while Linear Regression has a slight advantage, all models exhibit comparable efficacy in this study's context.

Below are the RMSEs of each model:

- Linear Regression RMSE: 3.29
  - `[1] "The RMSE value for the lm model is: 3.29"`
- SVM RMSE: 3.38
  - `[1] "The RMSE value for the SVM model is: 3.38"`
- Decision Tree RMSE: 3.47
  - `[1] "The RMSE value for the Tree model is: 3.47"`

The Decision Tree model, with the highest RMSE of 3.47, still demonstrates a reasonable degree of accuracy, affirming its applicability to specific predictive scenarios. The dataset's high quality—free from significant anomalies or sample issues—lends credibility to these findings.

It is, however, essential to contextualize these results within the realm of stock price prediction, where pinpoint accuracy is elusive. While these models yield insightful forecasts, they should be integrated into investment strategies with prudence. Such strategies ought to reflect an investor's risk appetite and investment goals, emphasizing the need for thorough research and judicious decision-making in the volatile landscape of the financial markets.

	Stock	Beta	Return
1	AA	0.8375749	-0.0107689733
2	AXP	1.3515404	0.0055907185
3	BA	1.9688908	-0.0027825797
4	BAC	0.3556423	-0.0195779879
5	CAT	1.8548586	-0.0097110414
6	CSCO	1.2697805	-0.0106090129
7	CVX	1.0423751	-0.0081685116
8	DD	1.1236230	-0.0047131451
9	DIS	1.1507706	-0.0106754761
10	GE	1.1337849	-0.0101498833
11	HD	0.4938798	-0.0054648135
12	HPQ	0.4632382	-0.0127242932
13	IBM	0.6665337	0.0004597155
14	INTC	2.3830177	0.0067095476
15	JNJ	1.1060306	0.0076926646
16	JPM	0.8413239	-0.0130777421
17	KRFT	0.1814769	-0.0028199519
18	KO	0.3665143	0.0077276390
19	MCD	0.4454153	0.0062959624
20	MMM	1.1827522	-0.0018855042
21	MRK	0.9452724	0.0038983268
22	MSFT	0.9303859	-0.0037401081
23	PFE	1.0526662	-0.0009370483
24	PG	0.4899781	0.0008692695
25	T	0.5926711	-0.0004450134
26	TRV	0.9669550	-0.0041672918
27	UTX	1.4443650	-0.0007517248

The subsequent investment analysis, focused on identifying stocks with a strong potential for returns, underscores Coca-Cola (KO), McDonald's (MCD), and Merck & Co. (MRK) as the top contenders. Characterized by beta values below 1, these stocks signify a lower risk profile relative to the market, combined with positive returns indicative of profitability. Coca-Cola (KO) is distinguished by its particularly low beta of 0.3656 and a significant return of 0.7727%, positioning it as a stable investment opportunity. McDonald's (MCD) also presents as a resilient option with a beta of 0.4454 and a return of 0.6296%. Merck & Co. (MRK), with a beta just shy of 1 at 0.9452 and a return of 0.3840%, shows robust performance in the pharmaceutical industry.

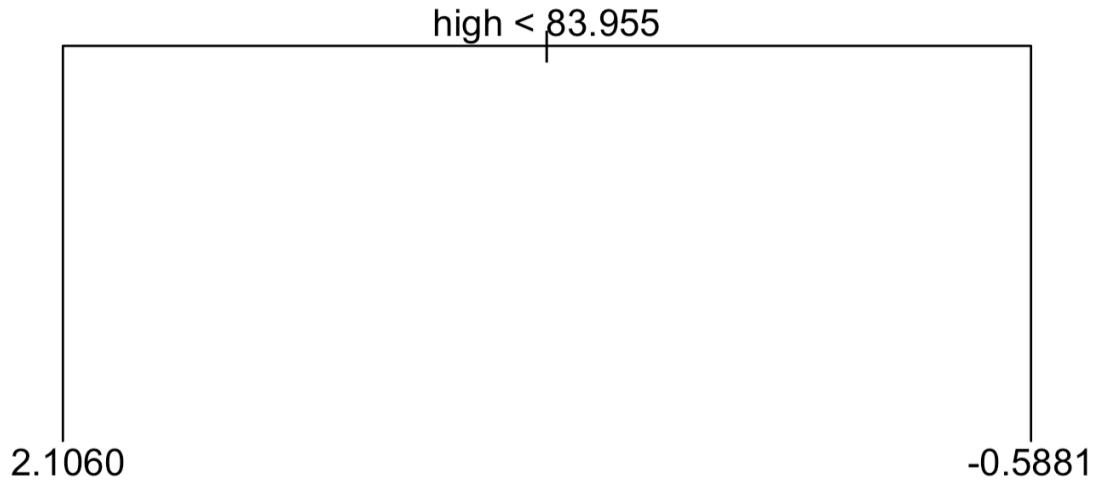
In conclusion, Coca-Cola (KO), McDonald's (MCD), and Merck & Co. (MRK) represent the most advantageous investment choices among the stocks evaluated,

delivering a balance of minimal market risk and promising returns. Investors seeking to diversify their portfolio with reliable and potentially lucrative assets would do well to consider these stocks, especially in the pursuit of mitigating risk and achieving financial success.

VII. Best Predictors

In determining the optimal predictors for our models, we crafted the final input formula as follows: percent_change_next_weeks_price ~ Week + high + percent_change_price + percent_change_volume_over_last_wk + volume. Initially, we excluded certain variables that reflected future outcomes to avoid artificially inflating the predictive power of our models. Subsequently, we conducted a series of Variance Inflation Factor (VIF) analyses and linear models (lm), refining our variable selection until all VIF values were below the threshold of 5, thereby eliminating multicollinearity concerns.

Upon running the preliminary model incorporating all independent variables, we discovered that 'volume' was the sole statistically significant predictor. However, in our refined model, which comprised a reduced set of variables, none exhibited statistical significance. Further exploration using a decision tree revealed that the 'high' variable most significantly influenced the percent change in the following week's price.



VIII. Assumptions & Limitations

Let's take a look at the overall assumptions and limitations of each model to determine which model might be the most appropriate under certain circumstances:

Support Vector Machine (SVM) models are highly versatile in classification tasks, with the following advantages:

- Effectiveness in datasets with a high dimensionality.

- Capability to outperform in situations where features outnumber observations.
- Adaptability through the use of various kernel functions to address non-linear data.

Nevertheless, SVMs also present certain limitations:

- Interpretation challenges arise from its inability to provide probability outputs.
- Optimal performance on smaller datasets is a result of the extensive training time required.
- Computational demands increase significantly with the size and complexity of the dataset.

Linear regression is a fundamental statistical approach for modeling the relationship between a dependent variable and one or more independent variables. Its advantages are as follows:

- Its simplicity and interpretability facilitate a clear understanding of variable relationships.
- Computationally efficient, linear regression processes large datasets with speed.
- Historical data can be used effectively to predict outcomes and discern trends.

However, linear regression is not without limitations:

- A presupposed linear relationship between variables may not exist in reality.
- Outlier sensitivity could disproportionately influence the regression line's slope.
- Complex or non-linear data sets might lead to underfitting and subpar predictive outcomes.

Decision trees are favored for both classification and regression due to their clear representation of decisions. Their advantages include:

- Ease of understanding and interpretation lends itself well to decision analysis.
- Flexibility in handling both numerical and categorical data types.
- Preprocessing demands, such as normalization or scaling, are minimal.

Yet, decision trees have their disadvantages:

- A susceptibility to overfitting, particularly with complex structures, may diminish the model's applicability.
- Instability can arise from minor data alterations, leading to vastly different trees.
- A predisposition towards classes with a greater number of instances may result in biased decision-making if one class is predominant.

IX. Best Model For Case

For the purposes of this case report, the Support Vector Machine (SVM) model emerges as the most suitable choice based on the results obtained. Despite the RMSE being slightly higher compared to Linear Regression, the SVM model offers superior handling of both categorical and continuous variables, along with enhanced computational capabilities. The potential of SVMs to be effectively optimized for larger datasets means they could process and manage the data more efficiently than both Linear Regression and Decision Trees.

A significant advantage of SVM is its ability to utilize various kernels, which is particularly beneficial for modeling the non-linear patterns often found in stock data. The flexibility to adjust to such irregularities renders the SVM model a robust tool for scenarios where data unpredictability is a concern. This adaptability also underscores a limitation of Linear Regression for this case; its fundamental assumption of linearity may not be appropriate for the complexities inherent in stock data. Moreover, Linear Regression may lack the requisite computing power for more in-depth analysis that such data could demand.

While Decision Trees are indeed powerful and can effectively handle classification and regression tasks, their interpretability becomes challenging when the trees become overly complex. A Decision Tree may generate an extensive array of nodes and branches, which can complicate the analyst's ability to understand and evaluate the model's outcomes.

X. Conclusion

After cleaning the data and performing various visualizations and tables, 3 models were created to predict the variable *percent_change_next_weeks_price*: Linear Regression, SVM, and Decision Tree. Following the computation and evaluation of these models, the Capital Asset Pricing Model (CAPM) was employed to assign each stock a corresponding beta and return. This enables a comparative assessment of the risk and potential reward of each stock against the Dow Jones Industrial Average benchmark.

The SVM model was identified as the superior predictive tool, courtesy of its comparatively lower RMSE and inherent advantages when applied to the dataset in question. According to the CAPM analysis, Johnson & Johnson (JNJ) and the Coca-Cola Company (KO) both recorded the highest return rate at 0.08; however, KO demonstrated

the least risk with the lowest beta value of 0.367, which suggests it offers a substantial reward relative to the index. Conversely, Bank of America Corp. (BA) showed the lowest return at -0.020, albeit with a modest beta of 0.356.

These insights allow investors to scrutinize the individual betas and returns of each stock to evaluate their attractiveness as potential investments. Such analysis can inform a more active investment strategy, contrasting with passive investors who might prefer the broader exposure of investing in the overall Dow Jones index.