Stock Prices for A.O. Smith Corporation

by Sarah Weinflash

Overview

In order to properly analyze the stock data for A.O. Smith Corporation, I think it's necessary to understand a bit about its history[1]. This is a company that has taken many forms; it has adapted with new technology and bent to the needs of its customers. Because of that malleability, A.O. Smith Corporation has been not only at the cutting edge, but also in a position of financial success for most of its existence.

Like many infants, A.O. Smith got its start in baby carriages. The company made steel tubing for carriage parts and bicycles beginning in 1874 and expanded to vehicle frames by 1899. From there, the company diversified its production with bicycle motors, oil refinery, and eventually patented the first glass-lined tank. Originally made for breweries, A.O. Smith found use for it as storage for dairy and water as well. A.O. Smith paused tank production in 1917 to focus on military manufacturing, and became the largest bomb maker in the USA during World War I. The company acquired Sawyer Electric during the Second World War and began producing electric motors once the wars were over. By the early 1960s, A.O. Smith merged those concepts with their water tanks to create the water heaters and filtration systems for which we know them today.

[1] *Integrity, innovation and customer satisfaction since 1874* (no date) *A.O. Smith Corporation History | A.O. Smith Corp.* Available at: https://www.aosmith.com/About/History/ (Accessed: 10 May 2023).

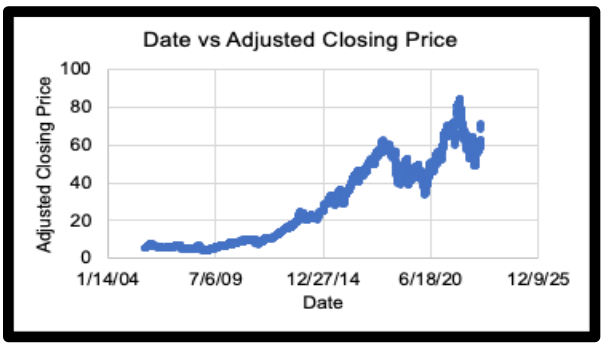
Data Sourcing, Cleaning & Transformation

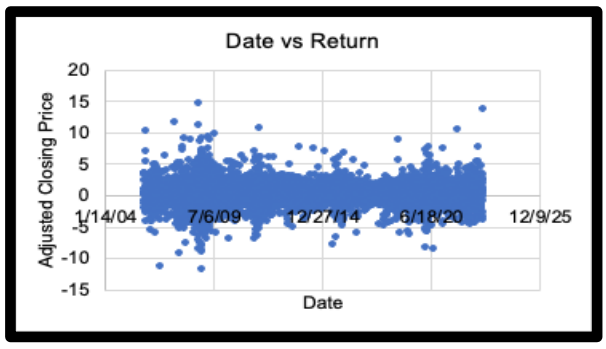
In sourcing this data from Yahoo! Finance[2], I found for each date the following information: Opening Stock Price, Closing Stock Price, Adjusted Closing Stock Price, Highest Daily Stock Price, Lowest Daily Stock Price, and Stock Volume. I have two main interests when examining this data: (1) how much the stock price changes day-to-day, and (2) how much the stock price fluctuates throughout the day. To evaluate the data on these metrics, I created two new variables: Return and Risk.

The Return variable finds the percent change from the Adjusted Close of one day to the Adjusted Close of the next day. In A.O. Smith Corporation's stocks, the Adjusted Close was 4.668 on January 3, 2006, and 4.694 on January 4, 2006; the calculated Return for January 4, 2006 was 0.557. The formula is below for reference.

$$Return = \frac{4.694 - 4.668}{4.668} = 0.557$$

This transformation allows me to measure the daily change with respect to the stock prices. A price change of $5 means a lot when the stock price grew from $1 to $6. It means less when the price shifted from $100 to $105. The Return transformation also provides a more linear evaluation of stock pricing, and one with more equal variance. This will eventually allow us to analyze the data utilizing Linear Regression techniques. The differences in the data are illustrated in the graphs below.

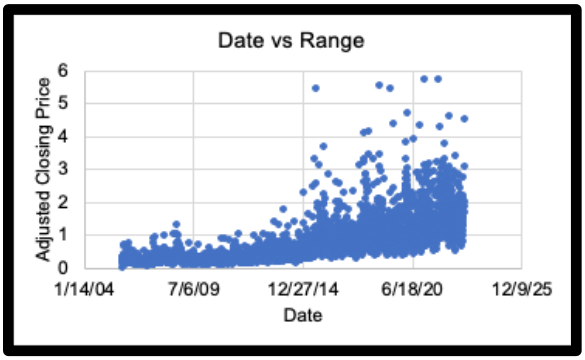


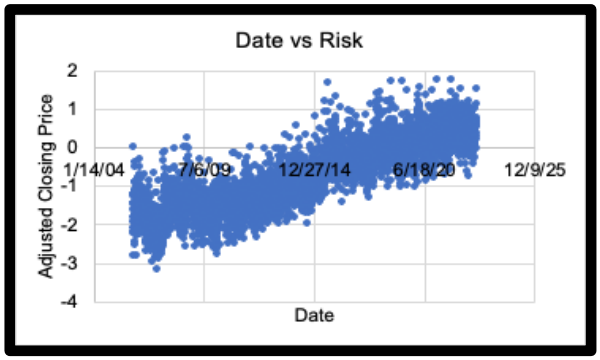


The Range is the difference between the Highest Daily Stock Price and the Lowest Daily Stock Price, Just as Return standardizes the change in stock price by dividing

[2] *A. O. Smith Corporation (AOS) stock price, news, Quote & History* (2023) *Yahoo! Finance*. Available at: https://finance.yahoo.com/quote/AOS/ (Accessed: 06 May 2023).

by the stock price, Risk standardizes the change in stock price by taking the natural logarithm. This allows it to grow very quickly with the Range, but slow its growth as the Range becomes larger. As you can see on the graphs below, the Risk is essentially a linear transformation of the Range.





When evaluating stock prices over a long period of time, dollar values may fluctuate. A.O. Smith stock prices were around $5 in 2006, and are up to $70 in 2023. Without standardization practices indicated above, one might draw the conclusion that A.O. Smith stocks are far riskier in 2023 than in 2006, because the price changes jumped from $0.20 to $1.00. However, the Return has remained fairly stable with an average of 0.

Data Sourcing, Cleaning & Transformation *(continued)*

The dataset from Yahoo! Finance included the following variables:

- **Date**
- **Open** (price at the time the stock market opens)
- **High** (highest price of the day)
- **Low** (lowest price of the day)
- **Close** (price at the time the stock market closes)
- **Adj Close** (closing price after dividend distribution)
- **Volume** (number of shares traded)

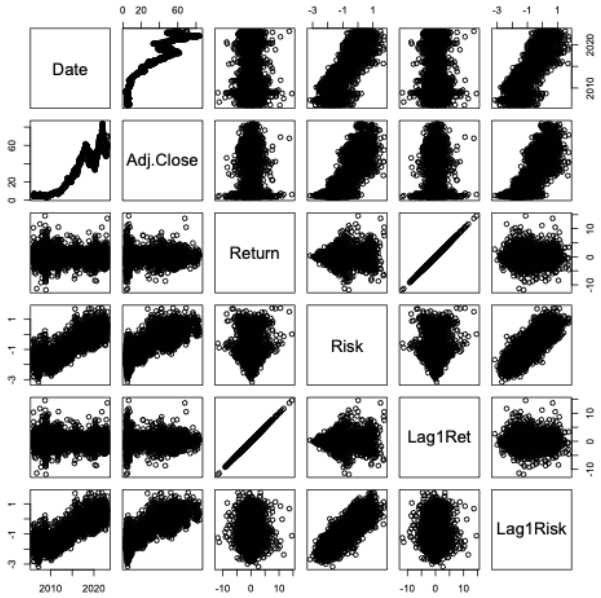
Date	Open	High	Low	Close	Adj Close	Volume
1/3/06	5.87	6.147	5.863	6.142	4.668	1702800
1/4/06	6.142	6.23	6.12	6.175	4.694	1293600
1/5/06	6.258	6.28	6.085	6.117	4.649	1296600
1/6/06	6.283	6.375	6.267	6.327	4.809	2917200

I produced the following variables:

- **Return** - *percent change in Adjusted Close price; calculated by subtracting yesterday's Adjusted Close from today's Adjusted Close, dividing by yesterday's Adjusted Close, and multiplying the formula by 100.*
- **Range** - *daily change, calculated by subtracting Low (lowest daily price) from High (highest daily price)*
- **Risk** - *natural log of Range*
- **StDevRet** - *standardized Return, compared with other Return values; subtracts the Average Return from the day's Return and dividing by the Standard Deviation of Return.*
- **StDevRisk** - *standardized Risk, compared with other Risk values; subtracts the Average Risk from the day's Risk and dividing by the Standard Deviation of Risk.*
- **Lag1Ret** - *Return the day before; "Lag 1" indicates a lag of one day*
- **RetRate** - *Rating of Return; cell outputs "LowRet" (Low Return) if the StDevRet is less than 0, and "High Ret" (High Return) otherwise.*
- **Lag1Risk** - *Risk the day before; "Lag 1" indicates a lag of one day*
- **RiskRate** - *Rating of Risk; cell outputs "LowRisk" if the StDevRisk is less than 0, and "High Risk" otherwise.*

Data Sourcing, Cleaning & Transformation *(continued)*

I reduced my dataset to just a handful of variables: Date, Adj. Close, Return, Risk, Lag1Ret, and Lag1Risk. Below is the matrix plot, which outputs pairs of scatterplots for all the aforementioned variables.



- **Date & Adj. Close**: there is a non-linear, positive trend with minimal variance; this looks like a graph called a "random walk"[3] in which a point moves randomly in one direction or another.
- **Date & Return:** there is a linear trend of zero slope with equal variance; this indicates that there may not be any relationship between Date and Return.
- **Date & Risk:** there is a linear, positive trend with equal variance; this indicates that Risk has gone up over time.
- **Adj.Close & Return:** there is a linear trend of zero slope with unequal variance. Here, the Return has a wider range of values when Adjusted Close is small. When Adjusted Close is larger, Return remains near zero. This unequal variance may become an issue during analysis.
- **Adj.Close & Risk:** there is a linear, positive trend with some unequal variance; this indicates that the Risk rises with Adjusted Close prices. There appear to be a wider range of Risk values at lower Adjusted Close prices, but may not be significant enough to cause issue during analysis.

[3] Schmidt, A. (no date) Random Walks: The Mathematics in 1 Dimension, Massachusetts Institute of Technology. Available at: https://www.mit.edu/~kardar/teaching/projects/chemotaxis(AndreaSchmidt)/random.htm (Accessed: 10 May 2023).

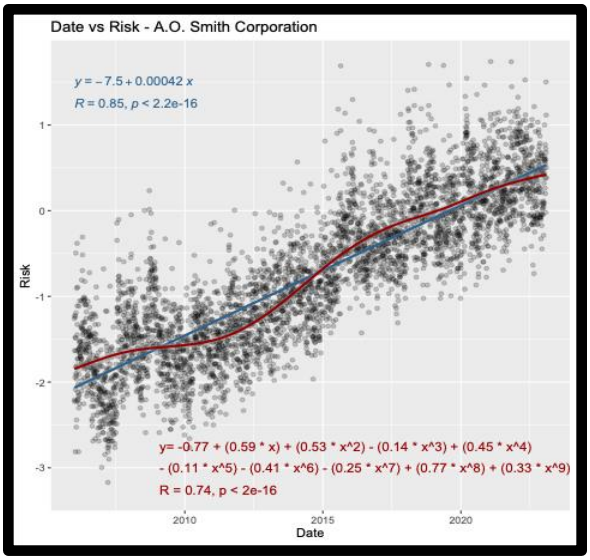
- **Return & Risk:** there is a linear trend of zero slope with unequal variance. Low values of Risk correspond with Returns around 0. The values of Return spread as Risk increases. Such unequal variance may become an issue during the analysis stage.
- **Return & Lag1Ret:** there is a strong linear trend of positive slope with equal variance; the two are almost equal.
- **Return & Lag1Risk:** there does not appear to be a trend, but the spread is mostly equal.
- **Risk & Lag1Ret:** there is a linear trend of zero slope with unequal variance.
- **Risk & Lag1Risk:** there is a linear trend of poistive slope and equal variance.
- **Lag1Ret & Lag1Risk:** there is a linear trend of zero slope and equal variance.

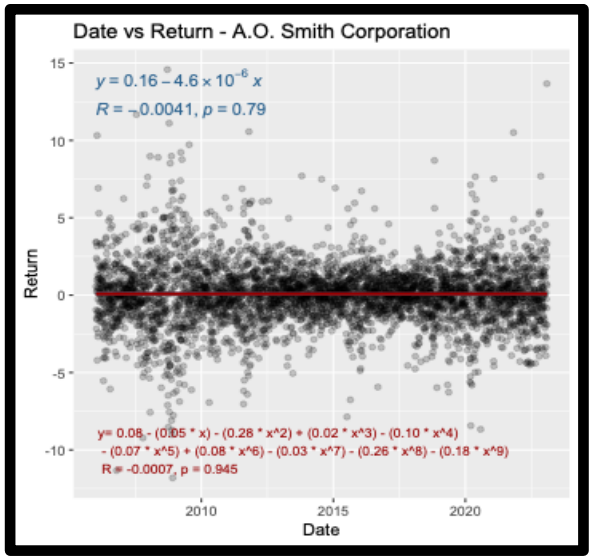
There are a few notable relationships made clear through the matrix plot:

- **Date**, **AdjClose,** and **Risk** all increase together; Adjusted Close increases with the Date due to economic inflation. With the increase of Adjusted Close, the Risk will increase. Although it does so on a logarithmic scale, the trend is still clear.
- **Return** and **Lag1Ret** increase together very strongly; this indicates that the Return of a particular day is a strong indicator for the Return value for the next day. This also means that trends seen between any variable and Return will be similar to the trends seen between those variables and Lag1Ret.
- **Risk** and **Lag1Risk** increase together less strongly than Return and Lag1Ret, but do trend together nonetheless. This indicates that the Risk of a particular day is a decent indicator for the Risk of the next day.

Linear Model

An important part of the analysis is the model or models we implement to examine the data. I've implemented two models in my dataset: the linear model ("lm", the blue line in the plots below) and the generalized additive model ("gam", the red line in the plots below).





The Linear Model finds a formula of the form "$Risk = Intercept + (Slope \times Date)$" which minimizes the sum of squared error. I've gone through the process of building a linear model step by step below.

Step 1: create a formula of the form below, with a unique intercept and slope.

$$Risk = Intercept + (Slope \times Date)$$
$$Risk = 1 + (0.3 \times Date)$$

Step 2: plug the date into the formula, and see what Risk value the formula predicts. The first Date value is 13151 (which I found by using as.numeric(q2a$Date[1])).

$$Risk = 1 + (0.3 \times 13151)$$
$$Risk = 3945.3$$

Step 3: find the squared difference between the formula's prediction and the actual value. The difference shows us how off the formula is from the actual data. Squaring the error allows us to compare both positive and negative error values.

$$Error = (Actual\ Risk - Predicted\ Risk)$$
$$Error = (-1.26 - 3945.3)$$
$$Squared\ Error = (-3946.56)^2$$
$$Squared\ Error = 15575335.83$$

Step 4: repeat Steps 2 & 3 until the entire dataset has been analyzed.

Step 5: add up all the error values from Step 3, this is your "sum of squared errors" for that formula.

Step 6: go back to Step 1, and repeat until you find a formula with a minimum sum of squared errors. This formula is the "best fit" for the dataset.

Luckily, we're not expected to compute this by hand; there is an algorithm in R that does this work for us, and all in a matter of seconds! The linear model created for the plots in the dataset are below.

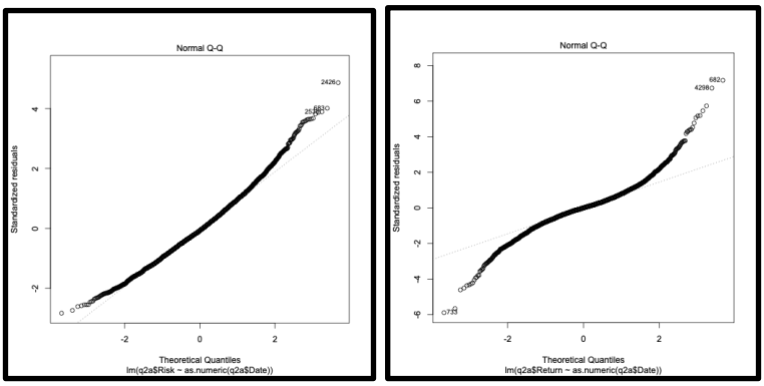
$$Risk = -7.5 + (0.00042 \times Date)$$
$$Return = 0.16 - (4.6 \times 10^{-6} \times Date)$$

There are two basic assumptions we make about the error values when we utilize the Linear Model: Equal Variance and Normal Distribution. Although error variance (the amount that the points deviate from the predicted line) is not uniform, it's not consistently above or below the line, nor does it spread out more on one end or the other. The error in other words equally varies around and across the line, and thus fulfills the Equal Variance Assumption.

A Normal Distribution looks like a bell curve. It is a shape that occurs regularly in nature, and having the error fit a Normal Distribution indicates that the item has a central value, with a handful of occurrences being slightly above and slightly below that value, and even fewer being far above and far below that value. For example, imagine that you are a berry farmer, and expect your berries to be ripe on July 1. You pick a handful just to be sure. If most of them are ripe, with a few underripe and a few overripe, then you can go ahead and pick the rest: you can expect that the majority are ripe. If the majority are overripe, you know to pick them in June next year. If a third of the berries are underripe, a third are overripe, and a third are underripe, then July 1 probably isn't the *actual* date of ripening, even if it's the *average* date of ripening. The Normal Distribution likely is not a good measurement

of the ripeness of your berries. If the errors follow a Normal Distribution, then we know that the line is a good measurement of the Risk or Return.

The Normal QQ Plot generated by R shows the Normal Distribution as a straight line; any deviations from that line are abnormal. Some deviation, which you can see in the plot of the errors of the Risk Linear Model, is to be expected. This shows that the Risk Linear Model is, more or less, a good measure of Risk. The errors in the Return Linear Model, however, are much less Normally distributed - the shape of the curve indicates that there are more values that are very high or very low than we were expecting, and fewer that are average. It shows that the line generated is more akin to an average value than a strong predictor of the actual Return value. This means that we have to be careful about drawing conclusions from this model, as the significance of the findings may be incorrect.[4]



[4] Guthrie, W. (2012) *Process Modelling*, *National Institute of Standards and Technology*. Available at: https://www.itl.nist.gov/div898/handbook/pmd/pmd.htm (Accessed: 09 May 2023).

Generalized Additive Model

The Generalized Additive Model[5] finds the formula of the form "$Risk = Intercept + (Slope \times Spline(Date))$". Unlike the Linear Model, this model may be non-linear.

Perhaps the best fit line for the data isn't

$$Risk = -7.5 + (0.00042 \times Date)$$

it could be

$$Risk = -0.2 + (0.83 \times Date^2)$$

or

$$Risk = 4 + (0.02 \times Date^2) + (0.84 \times Date^5).$$

Generalized Additive Models allow us to combine different curves (called "splines") in this way. A function with many splines will create a line that follows the data very closely. A function with few splines will be more similar to a straight line. The Generalized Additive Model aims to find the balance between a smooth curve with one that well describes the data. This tends to be found using sums of squares (just like the Linear Model) but with some penalization tactic to keep the model from simply generating a line that connects all the points together.

```
> summary.gam(Date_Risk_GAM)

Family: gaussian
Link function: identity

Formula:
q2a$Risk ~ s(as.numeric(q2a$Date))

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.765786   0.006872  -111.4   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                          edf Ref.df    F p-value
s(as.numeric(q2a$Date)) 8.445  8.913 1382  <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.741   Deviance explained = 74.2%
GCV = 0.20361  Scale est. = 0.20316   n = 4302
```

```
> summary.gam(Date_Ret_GAM)

Family: gaussian
Link function: identity

Formula:
q2a$Return ~ s(as.numeric(q2a$Date))

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.08297    0.03080   2.694  0.00709 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                          edf Ref.df     F p-value
s(as.numeric(q2a$Date)) 6.143  7.302 0.333   0.945

R-sq.(adj) =  -0.000691   Deviance explained = 0.0739%
GCV = 4.0871  Scale est. = 4.0803    n = 4301
```

One method of penalization is called a P-Spline. These "efficiently impose smoothness by directly penalizing the differences between adjacent coefficients."[6] The program I used to create the graphs generated the following formulas and

---

[5] Shafi, A. (2021) *What is a generalised additive model?*, Towards Data Science. Available at: https://towardsdatascience.com/generalised-additive-models-6dfbedf1350a (Accessed: 19 May 2023).

[6] Larsen, K. (2015) *Gam: The predictive modeling silver bullet*, Multithreaded. Available at: https://multithreaded.stitchfix.com/blog/2015/07/30/gam/ (Accessed: 19 May 2023).

outputs above. They indicate that the Generalized Additive Model for Risk is highly significant (that is, it describes the data very well), while the model for Return is relatively insignificant: the intercept describes the data, but the rest of the model could just as likely have been created using totally random values. The estimated best fit lines are below.

$$Risk = -0.77 + (0.59 \times Date) + (0.53 \times Date^2) - (0.14 \times Date^3) + (0.45 \times Date^4) - (0.11 \times Date^5) - (0.41 \times Date^6) - (0.25 \times Date^7) + (0.77 \times Date^8) + (0.33 \times Date^9)$$

$$Return = 0.08 - (0.05 \times Date) - (0.28 \times Date^2) + (0.02 \times Date^3) - (0.10 \times Date^4) - (0.07 \times Date^5) + (0.08 \times Date^6) - (0.03 \times Date^7) - (0.26 \times Date^8) - (0.18 \times Date^9)$$

Graphs

```
library(ggplot2)
```

- This code brings up a package called "ggplot2" that I have downloaded, which helps me create plots that are (in my opinion) prettier and easier to format and build than those created using the built-in R functions.

```
library(ggpubr)
```

- This code brings up a package called "ggpubr", which adds some features to ggplot that would otherwise require loads of code to accomplish.
- In these graphs specifically, the package allows me to calculate the R-square value as well as the linear equation, and add both to the plot.

```
ggplot(q2a, aes(Date, Risk)) +
```

- ggplot() lets R know that I want to create a plot using the ggplot2 package.
- The first argument (or required input) is the dataset from which the data I'm plotting will be pulled. I've named the dataset here "q2a".
- The next argument is aes() which stands for aesthetics. This is where I've put in anything I want visible in my plot.
- Inside of aes(), my first argument is the x-variable, Date. Next is the y-variable, Risk.
- The + sign indicates that I have another item I want to add to my plot before it's run.

```
geom_point(color='black', alpha=0.2) +
```

- geom_point() tells ggplot that I want the data to appear as points (like a scatterplot)
- The arguments inside the parenthases are optional.
- color='black' tells ggplot that I want the dots to be black
- alpha is an argument that indicates transparency. 0 is totally transparent, and 1 is totally opaque. I chose 0.2 (or 80% transparent) because it helps make obvious the density of the data distribution.
- The + sign indicates that I have another item I want to add to my plot before it's run.

```
geom_smooth(method='lm', color='steelblue4') +
```

- geom_smooth() tells ggplot that I want the data to appear as a smoothed line.
- method='lm' indicates that the smoothing method I want to use to generate the line is a linear model, which is a straight line.

- color='steelblue' means I want the color of the line to be a grayish blue color
- The + sign indicates that I have another item I want to add to my plot before it's run.

```
geom_smooth(method='gam', color='darkred') +
```

- geom_smooth() tells ggplot that I want the data to appear as a smoothed line.
- method='gam' indicates that the smoothing method I want to use to generate the line is a general additive model.
- color='darkred' means I want the color of the line to be a brick red color.
- The + sign indicates that I have another item I want to add to my plot before it's run.

```
xlab('Date') +
```

- xlab() tells ggplot that I want to add a label to the x axis.
- The only argument inside the parenthases is the label for the x axis ("Date").
- The + sign indicates that I have another item I want to add to my plot before it's run.

```
ylab('Risk') +
```

- ylab() tells ggplot that I want to add a label to the y axis.
- The only argument inside the parenthases is the label for the y axis ("Risk").
- The + sign indicates that I have another item I want to add to my plot before it's run.

```
ggtitle('Date vs Risk - A.O. Smith Corporation') +
```

- ggtitle() tells ggplot that I want to add a title to my plot.
- The only argument inside the parenthases is the title for the plot ("Date vs Risk - A.O. Smith Corporation").
- The + sign indicates that I have another item I want to add to my plot before it's run.

```
stat_cor(label.y = 1, color='steelblue4') +
```

- stat_cor() tells ggplot that I want to calculate the correlation between Date and Risk.
- The first argument inside, label.y, indicates that I want this correlation value to be added to the graph, at the point where y = 1.
- The next argument, color='steelblue', says that I want this label to be the same color as the linear model.

- The + sign indicates that I have another item I want to add to my plot before it's run.

```
stat_regline_equation(label.y = 1.5, color='steelblue4') +
```

- stat_regline_equation tells ggplot that I want to calculate the equation that represents the linear relationship between Date and Risk.
- The first argument inside, label.y, indicates that I want this equation to be added to the graph, at the point where y = 1.5.
- The next argument, color='steelblue', says that I want this label to be the same color as the linear model.
- The + sign indicates that I have another item I want to add to my plot before it's run

```
annotate('text', x=as.Date('2009-02-01'), y=-2.75, label='y= -0.77 + (0.59 * x) + (0.53
 * x^2) - (0.14 * x^3) + (0.45 * x^4)\n - (0.11 * x^5) - (0.41 * x^6) - (0.25 * x^7) +
 (0.77 * x^8) + (0.33 * x^9)\n R = 0.74, p < 2e-16', color='darkred', size=4, hjust=0)
```

- annotate() tells ggplot that I want to add some object to the plot.
- The first argument, 'text', shows that the object I want to add is text.
- The next two arguments, indicate where I want the object to be placed. I decided on an x value near the beginning of 2009 and a y value of -2.75.
- The label argument indicates what I would like the text itself to be.
- The color argument says I want it to be the same color as the GAM line which it describes.
- The size argument signifies the size of the text.
- The hjust argument tells ggplot that I want the text to have a left justification.

The code for the next plot is below. It is nearly identical to the code above, the differences being:

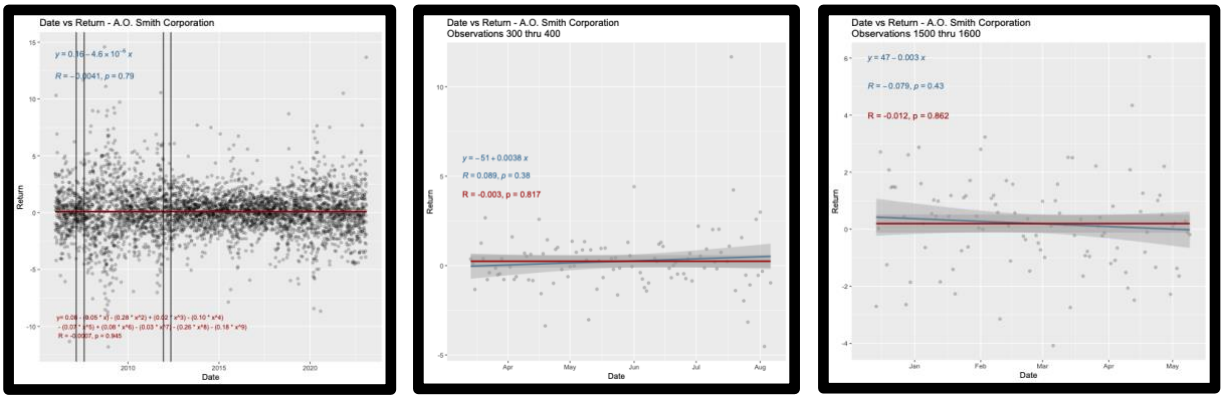
- the y-variable (Return)
- the y-label ("Return")
- the title ("Date vs Return - A.O. Smith Corporation")
- the placement of stat_cor (y = 12)
- the placement of stat_regline_equation (y = 14)
- the placement, text, and size of the annotate() object

```
ggplot(q2a, aes(Date, Return)) +
geom_point(color='black', alpha=0.2) +
geom_smooth(method='lm', color='steelblue4') +
```

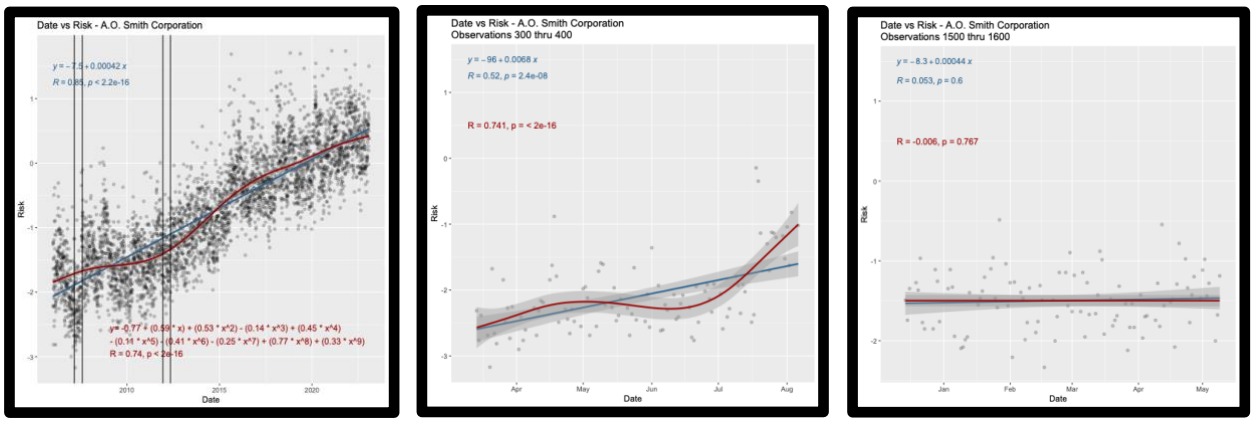
```
geom_smooth(method='gam', color='darkred') +
xlab('Date') +
ylab('Return') +
ggtitle('Date vs Return - A.O. Smith Corporation') +
stat_cor(label.y = 12, color='steelblue4') +
stat_regline_equation(label.y = 14, color='steelblue4') +
annotate('text', x=as.Date('2006-02-01'), y=-10, label='y= 0.08 - (0.05 * x) - (0.28 *
 x^2) + (0.02 * x^3) - (0.10 * x^4)\n - (0.07 * x^5) + (0.08 * x^6) - (0.03 * x^7) - (0.26
 * x^8) - (0.18 * x^9)\n R = -0.0007, p = 0.945', color='darkred', size=3, hjust=0)
```

Analysis

On both graphs, I found that though there are outlying values, variations from the central tendency are fairly equal. In other words, spikes are balanced by dips. This is particularly true in the graph of Date versus Return, which is evidenced by the fact that even the non-linear graph built by the Generalized Additive Model appears linear. The Date versus Risk plot shows some consistently abnormal high risk (in 2006 - 2009 and 2015 - 2020) as well as abnormally low risk (in 2010 - 2015). These diverging Risk levels are mostly captured in the Generalized Additive Model's curves. One exception in the dip in Risk level around 2007.



I captured two segments of the total Return sample in the graphs above: observations 300 thru 400 (March 14 thru August 6 2007) and 1500 thru 1600 (December 14 2011 thru May 9 2012). These show Linear Models with very different intercepts. The slopes are slightly different - the slope for the full sample is nearly zero; the first partial sample has a slope of positive 0.004, and the second partial sample has a slope of negative 0.003. Although the R-square values for both the Linear and the Generalized Additive Models are statistically insignificant for both the full and partial samples, these modeled changes show just how varied the actual data are in the short-term.



I performed the same analysis on Risk, and placed vertical lines on the graph of the full sample dataset identifying the locations of the smaller sample sets. The first sample comes from a moment in which the data strongly deviates from both models - while the data around it has very high risk, that sample was captured during a period of very low risk. The model for the full sample splits the difference, walking a path between the mountanous risk in 2006 and 2008, and the valley of low risk in 2007.

For the first sample of observations, the Linear Model shows a lower intercept and higher slope than the full sample; the R-square value was also lower, but the p-value showed statistical significance. In the Generalized Additive Model, the R-square value was nearly the same as for the full sample, and just as statistically significant. The second sample of observations, on the other hand, has a very similar Linear Model to the full sample. However, its R-square value was nearly 0, and the p-value showed no statistical significance. The Generalized Additive Model had an R-square value of nearly zero, with a p-value showing no statistical significance.

The interesting and surprising discovery to me was the difference in p-value by sample. The outcome from the Return data made sense to me: statistical insignificance begets statistical insignificance. I could reconcile if both samples were statistically signficicant, or statistically insignificant. This result, however, required a deeper understanding of the material, which led me to research on Simpson's Paradox[7].

[7] Sprenger, J. and Weinberger, N. (2021) Simpson's paradox, Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/entries/paradox-simpson/#:~:text=Simpson's%20Paradox%20is%20a%20statistical,population%20is%20divided%20into%20subpopulations. (Accessed: 09 May 2023).

Although Simpson's Paradox is often used to describe results by group, it can also explain non-categorical data, such as the shift in time I'm examining in this dataset, so long as you treat the different times as different groups. Simpson's Paradox traditionally explains how subgroups with similar results might coalesce to produce a different result in the full sample. While that's not entirely relevant to this case, reading the literature on those occurences helped me conceptualize the issue I was facing here. Namely, that subgroups of small sizes may not follow the same patterns as the overall sample.

The first sub-sample captured a moment of significant change: the Dow Jones hit a peak of 14,000 on July 19, 2007[8] and by mid-August, Bear Sterns had liquidated two hedge funds, American Home Mortgage had filed for bankruptcy, and the credit rating of mortgage lender Countrywide Financial had plummetted.[9] It's not surprising, then, to see Risk creeping up, and to have that pattern be captured with statistical significance.

On the other hand, the second sample was taken during a moment of stagnancy. Although "The Great Recession" might be considered over and federal stimuli were limited in scope,[10] the economy was only slowly crawling back; it wouldn't reach its pre-recession levels for another few months.[11] This stagnancy in the economy is apparent in the sample distribution, which shows a very slight increase in slope, but generally an apparently random smattering of Risk values. This randomness is reflected in the p-value, which shows that the values show no pattern whatsoever in the sub-sample. I now understand that this lack of pattern is itself a pattern: the centrality of Risk around -1.5 has meaning only when compared to other Risk values.

---

[8] Dow Jones industrial average charts (2023) Yahoo! Finance. Available at: https://finance.yahoo.com/quote/%5EDJI/chart/ (Accessed: 10 May 2023).

[9] Usatoday (2013) *Timeline: Key events in financial crisis*, *USA Today*. Available at: https://www.usatoday.com/story/money/business/2013/09/08/chronology-2008-financial-crisis-lehman/2779515/ (Accessed: 11 May 2023).

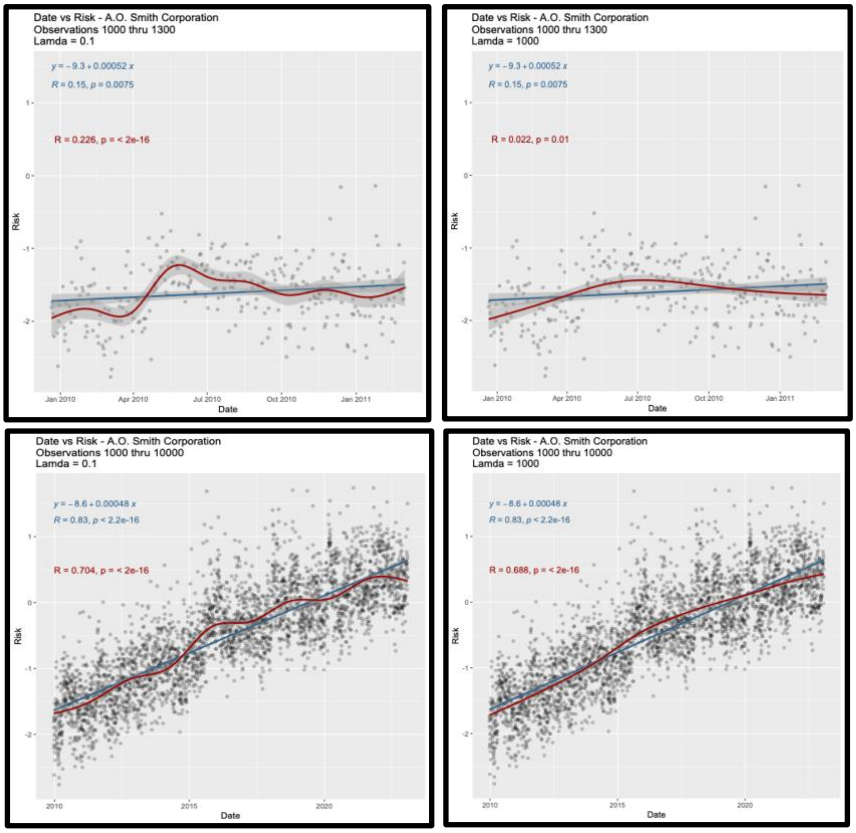
[10] Applebaum, B. (2011) 'Stimulus by Fed Is Disappointing, Economists Say', New York Times, 24 April. Available at: https://www.nytimes.com/2011/04/24/business/economy/24fed.html (Accessed: 09 May 2023).

[11] Real Gross Domestic Product (2023) St. Louis Fed. Available at: https://fred.stlouisfed.org/series/GDPC1 (Accessed: 10 May 2023).

Analysis *(continued)*

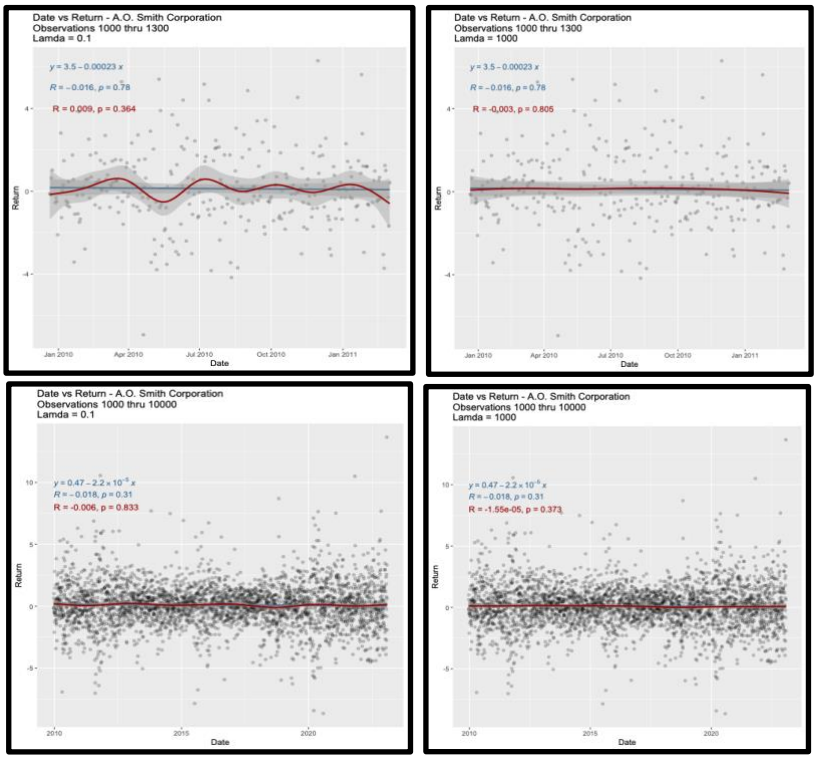
Below, you will find four charts for Risk and four charts for Return. Each shows two sample sizes: 300 and 9000. Each sample size is shown with two values for lambda: 0.1 and 1000.

Date vs. Risk



In the Generalized Additive Model, the lambda value penalizes the number of splines in the curve. Thus, increasing the value of lamda will force the model into a smoother line. You can see this in the sample graphs above, with lambda equal to 0.1 and 1000. A higher lambda will also result in a lower R-square value, because the smoother line does not fit the data as well as one with more curvature; this is evident in the outputs above and below.

Date vs Return



When there is a strong relationship between two variables, a higher sample size leads to a lower p-value.[12] This is because the p-value measures the likelihood that the relationship might be due to randomness. While a small sample might randomly display a relationship, it is less likely that a large sample would do the same. While the p-value for the Risk data was at a minimim when lambda was set to 0.1, the higher lambda with the smaller sample had a p-value of 0.01. Increasing the sample size but holding lambda steady shifted that p-value to a minimum.

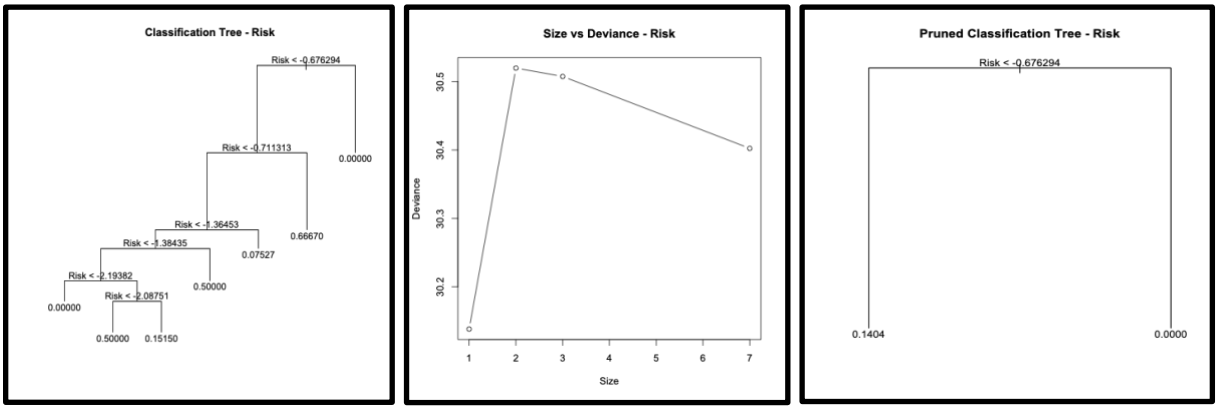
However, the Return data did not exhibit similar effects. When lambda was set to 0.1, the p-value increased when the sample size increased. The model was able to pick up on some patterns in the data (though not with statistical significance) when there was a small amount of data, because each data point was more important to the model. When the sample size increased, the model recognized the patterns for what they were: randomness. It reported as such through the higher p-value.

[12] Thiese MS, Ronna B, Ott U. P value interpretations and considerations. J Thorac Dis. 2016 Sep;8(9):E928–E931. doi: 10.21037/jtd.2016.08.16. PMID: 27747028; PMCID: PMC5059270.

Classification and Regression Trees (CARTs)

One notable economic recession that intersects with the stock data occurred from December 2007 thru June 2009; I created a new variable in the dataset which indicates whether or not the value occurred during this time period.

Next, I created a Classification and Regression Tree (CART), which identifies the output of the Recession variable given a certain level of Risk.[13] Since the Recession variable was binary, this meant that the model indicated the probability of the data occurring during the Recession. I utilized cross validation to find the most efficient tree that maintains accuracy. The lowest deviance occurred when the tree was only one value. The model does not output single-branched trees, so I plotted the pruned classification tree with two values, one branch of which has a probability of zero. The original classification tree, the cross validation plot, and the pruned classification tree are below.

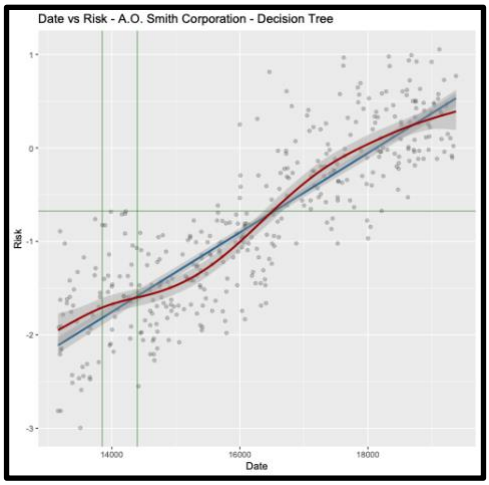


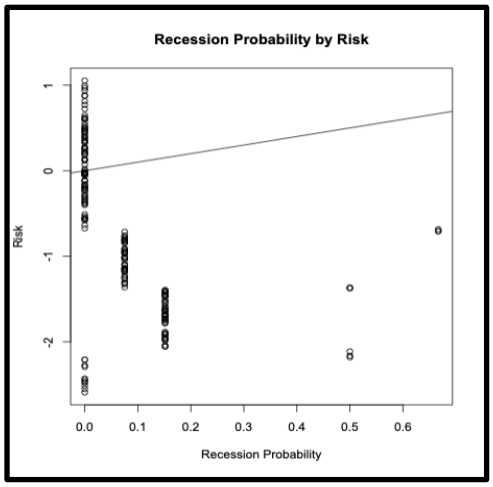
For my own edification, I plotted the data points in the set along with lines indicating the dates of the Recession and the branches of the pruned classification tree. The classification tree indicated that if the risk is below -0.67, then there is a 14% chance that the stock data came from the period of Recession. This seems to fit with what the graph below shows.

I assessed the predictive quality of this tree by plotting the Recession Probability by Risk, as well as using a table to estimate whether the data came from this period of Recession. The tree was able to predict whether or not the data came from the

[13] Le, J. (2018, June 19). R decision trees tutorial: Examples & code in R for Regression & Classification. DataCamp. https://www.datacamp.com/tutorial/decision-trees-R

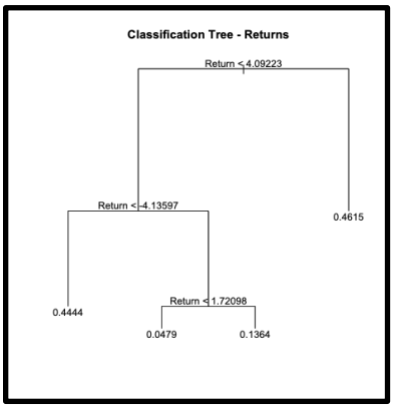
period of Recession with 91% accuracy. The Risk values during the Recession are fairly clustered together, so the Mean Squared Error value of 1.72 is not shocking.

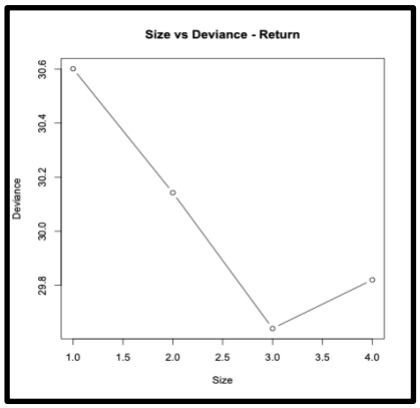


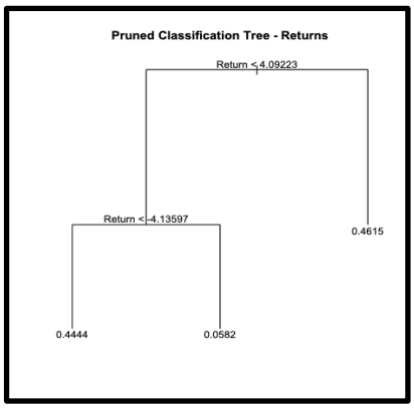


	Recession	
	0	1
0	178	17
1	1	4

I then created a CART using Return as the predictive measure. As with Risk, I utilized cross validation to find the most efficient tree. The lowest deviance occurred when the tree had three branches. The original classification tree, the cross validation plot, and the pruned classification tree are all below.



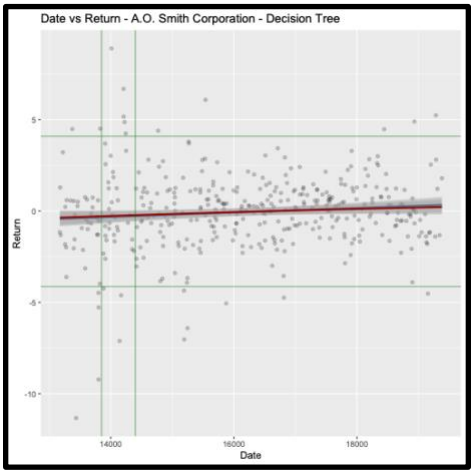


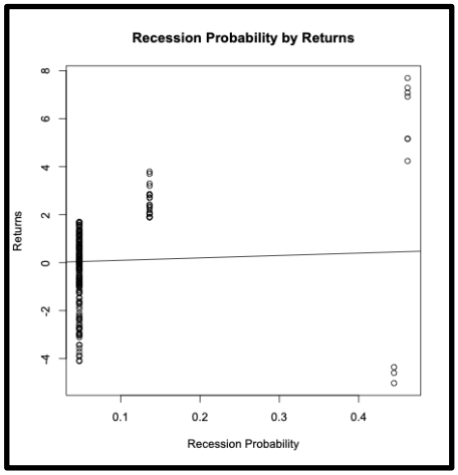


The output of the pruned tree was surprising to me: why and how could a Return of greater than 4.1 indicate that the value was 46% likely to come from the Recession? I admit that I re-ran my data a few times, spent a while assuming I was reading the table wrong, and tried to figure out what I was missing. Finally, I created a plot of the data with the values from the tree and dates of the Recession highlighted.

My discovery was both surprising and relieving: the Returns indeed were quite high during the Recession dates. (They were also quite low, but that wasn't as astonishing to me.) Utilizing a table, I found that the tree was able to predict the

likelihood of the data coming from the period of Recession with 90% accuracy! The Mean Squared Error was higher, around 4.412. Given the wide and even spread of the data - particularly during the Recession - this higher value is expected.



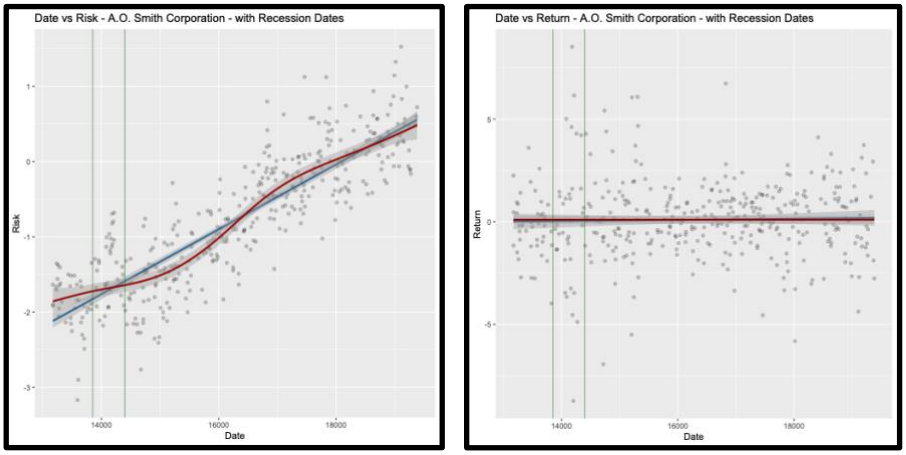


	Recession	
	0	1
0	177	19
1	1	3

Conclusion

I chose to create a smaller dataset using only stock data that occurred during The Great Recession, from December 2007 thru June 2009. In the scatterplots below, this segmentation looks similar to the "chimney" concept, coined by Sir Francis Galton. The chimney, as explained by Professor Lawrence Tatum, selects a portion of the data for deeper examination. Specifically, it's used to estimate the Expected Value of the data, conditional on that data having specific predictor values. Here, I will be utilizing this concept to estimate the Expected Value of Risk and Return, conditional on those values occurring during The Great Recession.

As you can see in the plots below, the Risk data during this time period falls between approximately -0.75 and -2.25, and does not quite fall in line with the trends in the rest of the data. The Return data does not indicate a strong shift in centrality, and has a much wider range, from around -7.5 to 9. This range is unusual in this otherwise fairly dense dataset.



There are a number of prediction models we can use to estimate the Expected Values of Risk and Return, but examining them in conjunction with one another will uncover information about the data as well as the models themselves. I will utilize: Mean, Median, Linear Model, Generalized Additive Model (GAM), and Categorical and Regression Trees (CART) for this analysis. They are described below.

- The Mean, or average, value sums all the values and divides them by the number of data points. This is a great way to estimate the middle of the data, so long as there aren't significant outliers.
- The Median value orders the values from lowest to highest, and chooses the middle value. That is, if there are 5 values in the datset (3, 8, 9, 10,15), it would choose the third value (9). Median is a strong estimator of centrality when there is plenty of data and some outliers that might skew the mean.
- The Linear and Generalized Additive Models were described earlier (*see pages 11-17*). I used each to find the value of Risk or Return estimated for a date during the Recession (May 22, 2008) and used the median values of the entire lines to estimate Risk or Return for the whole dataset.
- Categorical and Regression Trees (CART) estimate the value of the response variable for a given value of the predictor variable. In this case, CART estimated the response variable for two categories: inside of the Recession and outside of the Recession. I reversed the CART I had previously utilized, and predicted Risk and Return using the Recession variable. Note that the tree for the Return dataset was single-node (meaning, the outcomes were the same in and out of the Recession).

Below, I've recorded the prediction models and their outputs, which estimate the Risk and Return values for the full dataset, as well as for the portion of the dataset that occurred during the Recession.

	Risk		Return	
	Full	Partial	Full	Partial
Mean	-0.7844	-1.441	0.1009	0.0522
Median	-0.7853	-1.474	0.0883	-0.1585
Linear Model	-0.7650*	-1.703	0.0829*	0.0934
GAM	-0.8044*	-1.627	0.0776*	0.0370
CART	-0.7006**	-1.441	0.0829**	0.0829

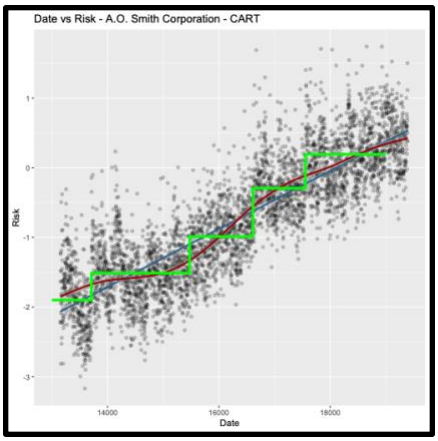
** median value from full dataset*
*** estimated value for the remaining data, not in the partial dataset*

In the Risk dataset, the values for the full model are much higher than those from the partial model. This is not so for the Return dataset, which shows values that are remarkably similar. In fact, the range of the Return partial dataset (0.21) is almost the same as the range of both Return datasets combined (0.24)! In comparison, the range of the Risk data's partial dataset (0.26) is almost four times smaller than total range (1.00).

What I gather from this is that there is not much difference between estimated Return values during the Recession and outside of the Recession, while the Recession had a large impact on the Risk values of the stocks. This is not to say it had no impact on Return - in fact, the extreme variation indicates that it *did* have significant impact, but these models are measuring Expected Value, rather than Variance. For this reason, I will focus the most of my analysis on the efficacy of these models on predicting Risk during and outside of the Recession. Because the Mean and Median values for Risk are similar during the Recession, I can assume there isn't much skew in the data, and can use them as a basis for comparison in the success of the models' predictions for the partial dataset.

During the Recession, the Linear Model predicts Risk to be quite a bit lower than the other predictors; looking at the graph, it's clear that the line sits below the majority of the data. The GAM hovers a bit higher, but still is not an accurate reflection of the specific data we're evaluating. The CART, however, does very well capture the approximate value of Risk during the Recession. This is, of course, because the model was built to do just that: estimate Risk during these dates.

How effective are these models at estimating Risk outside of the Recession? One measure of this is Mean Squared Error (MSE), which finds the difference between each predicted value and the actual value it was trying to predict, squares the difference, sums up all those differences, and divides them by the number of values in the dataset. For the Linear Model, this value was 0.22. For the GAM, it was 0.20: even lower! Unsurprisingly, the MSE for the CART was significantly higher than the other models (0.74). This is because the Linear and Generalized Additive Models were built to estimate the Risk of the entire dataset, and both therefore do this fairly well. The CART was built to estimate the Risk in two areas: during the Recession and not during the Recession. Because the Risk values during the Recession are quite similar to one another, it did this relatively well. However, there is less consistency in the Risk values outside of the Recession. Therefore, its estimate for Risk outside of the Recession is not as accurate.



In order to better understand the capabilities of the CART model, I decided to perform a similar analysis but *not* restrict segmentation to inside and outside of the Recession. The plot is above; the the green lines represent the values predicted by CART. As in the previous plots, the red line represents the GAM and the blue line represents the Linear Model. The MSE was 0.1995 - lower than either the Linear Model *or* the GAM.

This surprised me, because I expected that the curvature of the GAM to follow the data more closely and evenly. However, I noticed that there were more steps in the CART plot than there were curves in the GAM plot, and figured this was likely the cause. You see, there are different overfitting prevention techniques in each model: the Linear Model is restricted by the fact that it must be a straight line; the GAM is restricted by Lambda, which penalizes additional splines using cross validation; and CART is restricted by alpha, which penalizes additional branches using cross validation.[14] The cross validation techniques are meant to retain a certain number of degrees of freedom, but the benefits of those additional degrees of freedom are not captured when using the MSE as a comparison tool.

[14] James, G., Witten, D., Hastie, T., Tibshirani, R. (2022). Tree-Based Methods in An introduction to statistical learning: With applications in R (pp. 303 - 308). essay, Springer.