ARIMA Sector Forecast Report

August 8, 2012

INTRODUCTION

ABOUT DESIGN & ANALYTICS

Design & Analytics is a project founded by Adam Hogan, specializing in design, modeling, and analytics support. Design and Analytics makes broken things work, puzzling things intelligible, slow things fast, and ugly things beautiful. We provide consulting and project-management services in design, analytics, and reports automation, with an emphasis in financial and economic applications.

ABOUT THE ARIMA SECTOR FORECAST REPORT SERIES

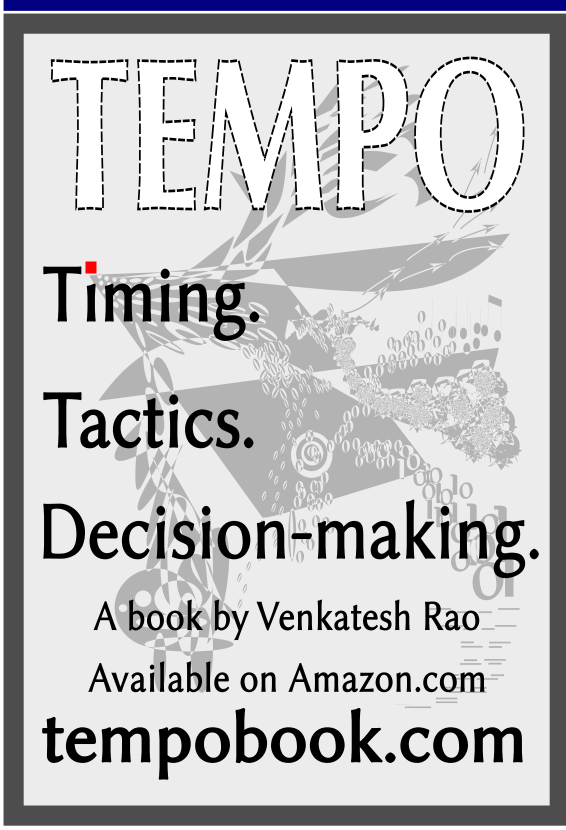
This is one of a series of daily reports created by Autoregressive Integrated Moving Average (ARIMA) models to forecast financial or economic time series that may be of wider interest. A brief description of the intuition behind this algorithm is available in the summary section. The model itself is built on an auto-fit algorithm from the Free and Open Source R statistical software package, which selects forecast parameters to use based on historical goodness-of-fit and parsimony considerations. The forecast then projects 15 business days and provides 50%- and 95%-confidence intervals. The data are graphed, saved, and reported.

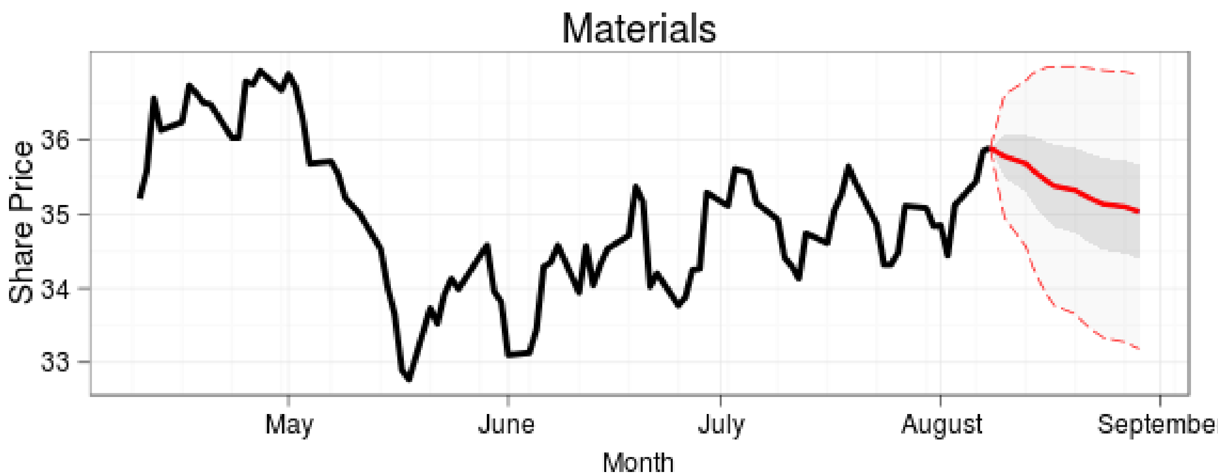
It is emphasized that the intended purpose of this forecast is strictly non-financial. Do not use this in your trading strategy, or consider it to have any bearing on buying or selling activity. It is not a recommendation of any kind, to buy, to sell, to hold, or to dream. It is simply the publication of the result of a standard and relatively naïve econometric model demonstrating how the current mechanics of the market are pricing future activity as of today. This is not equivalent to what will actually happen in the future and should be treated with skepticism. For individual stocks and sectors, there are better models. This is neither optimal nor rigorously tested. It is simply a yard stick—one of many competing yardsticks for measuring a quantity that may or may not be a yard. However, the results of the model can be interesting and useful for some purposes—depending on your assumptions—providing a point of comparison or baseline against others.

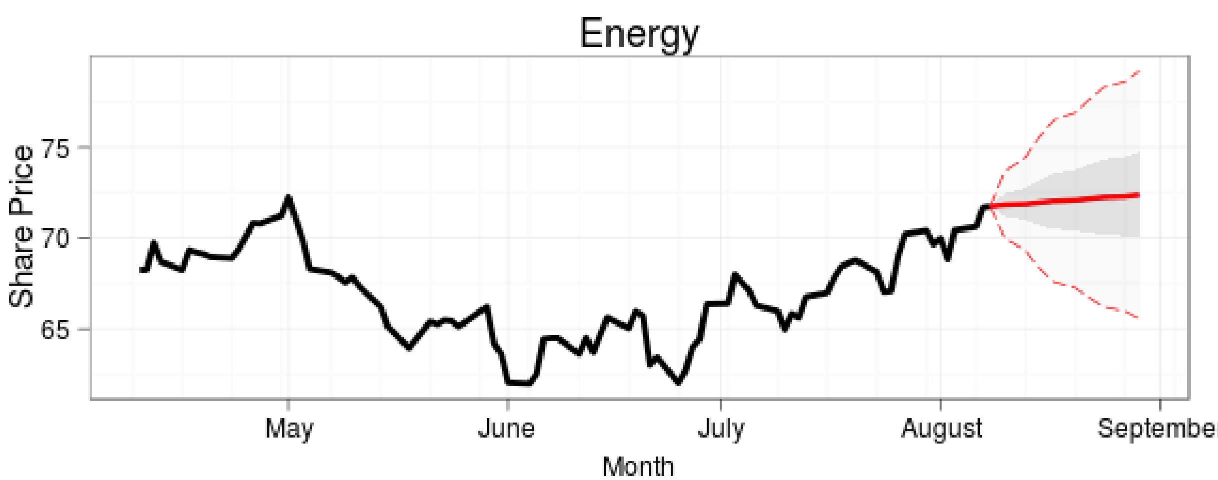
ABOUT THIS DOCUMENT

What this report showcases, in addition to the marriage of analytics and design, is a report automation infrastructure. This runs every day, and has an infrastructure devoted to it to analyze, graph, typeset, and publish automatically and without any human intervention a newsletter-quality report, something with wide applicability across fields and audiences.

THIS REPORT'S SPONSORS




Materials
Share Price
36
35
34
33
May
June
July
August
September
Month

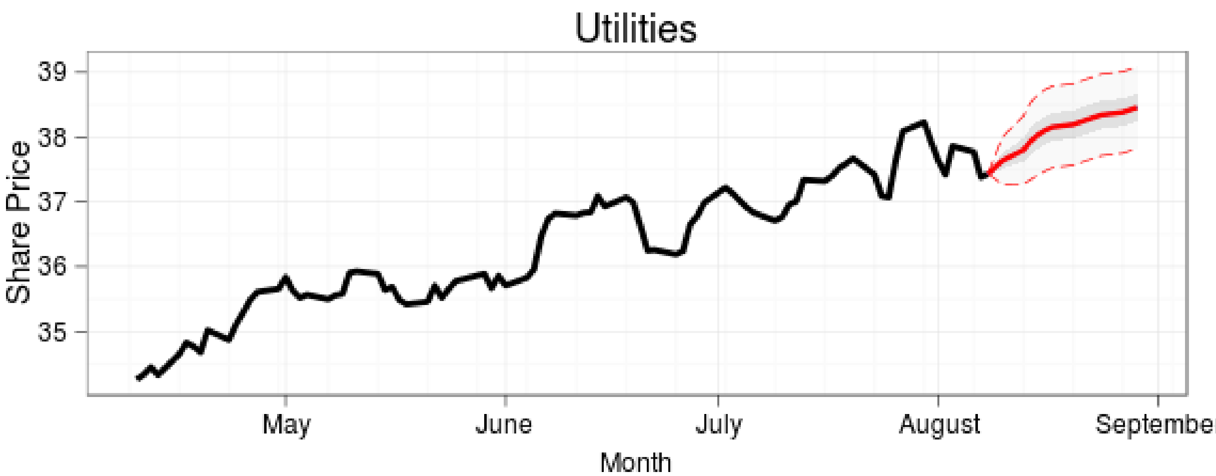

Energy
Share Price
75
70
65
May
June
July
August
September
Month

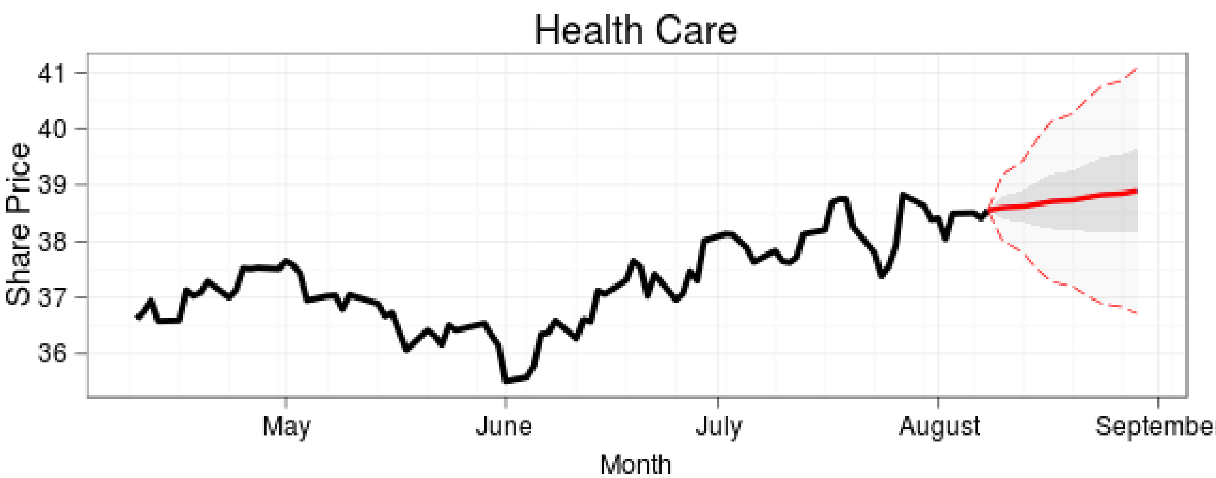

Financials
Share Price
16.0
15.5
15.0
14.5
14.0
13.5
May
June
July
August
September
Month

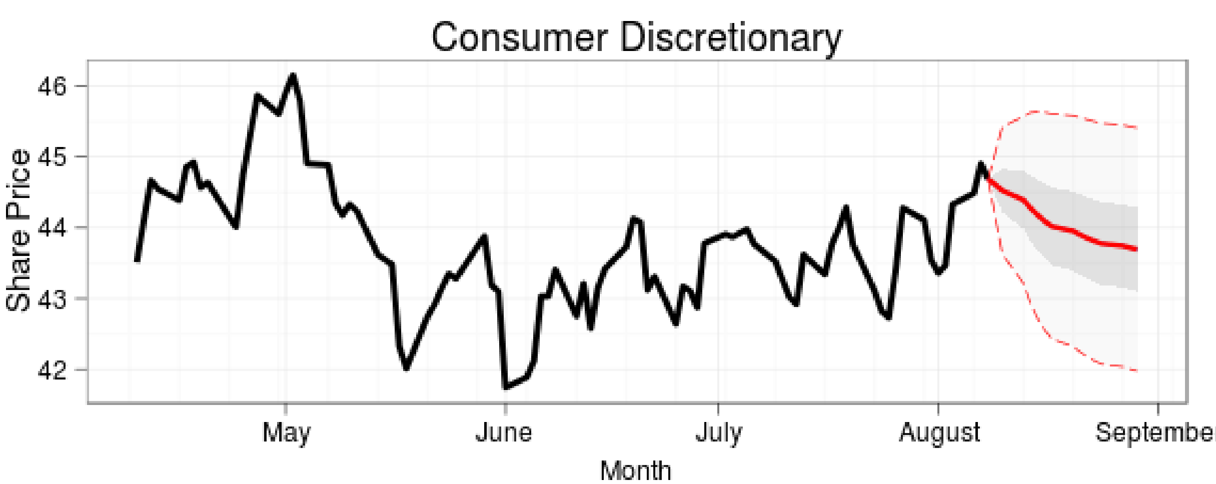

Industrials
Share Price
39
38
37
36
35
34
May
June
July
August
September
Month


Technology
Share Price
32
31
30
29
28
May
June
July
August
September
Month


Consumer Staples
Share Price
36.5
36.0
35.5
35.0
34.5
34.0
33.5
May
June
July
August
September
Month


Utilities
Share Price
39
38
37
36
35
May
June
July
August
September
Month


Health Care
Share Price
41
40
39
38
37
36
May
June
July
August
September
Month


Consumer Discretionary
Share Price
46
45
44
43
42
May
June
July
August
September
Month

SUMMARY

OVERVIEW

Growing out of an automated analysis performed for personal use for several months, this report uses a statistical, econometric model to explore potential movements in groups of US equities, evaluated at the sector level. It has been expanded for a wider audience. For limitations, please see the disclaimer section.

WHY SECTORS?

Sectors, groupings of stocks based on shared general business activity, are one way of organizing a view of the dynamics of the market. The idea is that companies doing similar things will have similar price movements. As an example from a particular industry rather than a full sector, consider how aluminum-producing companies would likely all be affected by changes in the market price of aluminum—good companies and bad companies both would suffer if prices fell for refined aluminum, or rose for mined bauxite. The more similar the business activities, the more intercorrelated their price returns will be—they will respond in related ways to the same shocks.

Some shocks will affect an entire economy, all stocks. Some shocks will affect one industry. Some shocks will affect only one firm. Analyzing sectors gives you an idea of a particular level of granularity—you can think of it as the level of resolution at which you are viewing the economy. You could certainly be more general, and look at a measure like the S&P 500, (low resolution), or less general, and look at the universe of individual stocks that compose that index (high resolution). But in terms of interpreting the information, while you would be buried trying to interpret the dynamics of all stocks at once, sectors are often an efficient level of information aggregation for many purposes.

Currently, this report makes use of indices that track these sectors as the data upon which to base forecasts. However, we will continuously expand our methodology to incorporate custom indices and broader measures of sector activity and health.

METHODOLOGY

What follows is a description of the intuition behind the ARIMA models used to construct this forecast. For a more technical introduction you can find great pieces on ARIMA models by searching online. What follows is intended to be more of a practitioner's introduction.

An ARIMA model can be broken down into two components, an autoregression (AR) component and a moving average (MA) component. (The (I) portion of the name is less interesting—it comes from the model potentially being "integrated," which means you may need to model the rate of change or the rate of the rate of change, rather than the price series itself.)

Now, let's consider each of the parts: "Autoregression" can be thought of as the component of the model that picks up trends. If the price has gone up over the last three periods, the autoregression optimally fits a line through those last price observations and concludes the price will keep going up, following this newly established trend.

Autoregression models are characterized by their "lags." This refers to how many periods back the model takes into account. In the example above, in which we drew a line based on the last three observations, we were considering 3 lags, and we would write this as an "AR(3) model." But you don't have to stop at 3—a model using an auto-fit function often takes higher order lags. In a general form, you would write AR(p), where p represents an integer chosen to be some number smaller than the length of your data.

Let's consider the second component of the joint model, the Moving Average. Speaking in Wall Street adages and mixed metaphors, while sometimes "the trend is your friend," we also often observe that prices are not unlike gravity in that "what goes up must come down." The moving average component provides the mathematical representation of that force, which is often referred to as "mean reversion." Where if the last three prices were 1.00, 2.00, and 3.00, an AR model might suggest the next price will be 4.00. However, the MA side, ever the contrarian, might instead suggest going back to 2.00, that both 1.00 and 3.00 were far deviant from the mean, or the normal resting state of the price.

Which one is right? Neither is right all the time, and there are even some cases in which both can be wrong. The strength of an ARIMA model, however, is that it lets you represent both kinds of motion and, based on historical data, to calibrate their relative strengths. This calibration occurs by giving some of the parameters more weight than others. Additionally, by having the number of lags under the AR(p) side be different than the MA(q) side, you can represent a wide variety of different dynamics for the same price series. For example, for a particular price series, you may find that the most recent trend dominates in the short term, but that a longer-term moving average is also predictive.

So, how many lags should we consider? This depends most importantly on the series—some phenomena at some times are better described by parameterizing one side more heavily than the other. When running the ARIMA model, we seek to find an optimal fit based on historical data, so this balance changes as the data change.

But more philosophically, the choice of lag also depends on your objectives. If you're an economist looking to understand the phenomenon, you'll often choose a criterion that takes very few. Economist Milton Friedman advised never using more than 3 terms in any regression, because you'd never be able to correctly interpret the cross-terms of anything higher. Economists (as well as real scientists and philosophers, too) value parsimony, how much you're able to explain while minimizing the amount of data or number of assumptions you need to make. Statisticians have devised a number of different ways of measuring parsimony, but most popular for model fitting in econometrics are the "information criteria" called AIC and BIC (the Akaike Information Criterion and Bayesian Information Criterion, respectively). These roughly say, "Please explain as much as you can, but your explanation is penalized by each additional bit of model scaffolding, such as each additional lag, that you need to add." Considered this way, a model that predicts the future less accurately, but requires fewer components of scaffolding can be preferable to a model that predicts the future more accurately, but requires many more inputs.

Forecast junkies and computer scientists, on the other hand, tend to place less of a value emphasis on parsimony, occasionally adding as much data as possible to improve model fit (measured in R^2) aiming at forecast performance (measured in Root Mean Square Error). Occasionally, this can have downright laughable results. As an example, any set of information you add as an explanatory variable that is not perfectly correlated with your data will improve the R^2 measure of goodness-of-fit. By construction, this is a measure of explained variance, and it always increases even when you offer it bad explanations—it takes no account of quality or reasonableness. So every additional variable, especially independent ones, will increase your model's ability to "explain" historical variance. That means even adding, for example, a time series of the runs scored in the Detroit Tigers' 2006 season will improve your R^2. What it will not do, however, is improve your model and forecast performance—which has only explained the past variance better by tilting contingent noise, while saying nothing useful about or related to the actual phenomenon, or how it may rhyme with the future.

ARIMA FIT

If you have explored auto-ARIMA models on stocks before, you'll notice that this model has higher-order dynamics than default parameter optimizations. While we use an AIC and BIC optimization, we're not prioritizing explanatory parsimony quite as highly as the defaults, and include more terms than prescribed by these methods.

This is because we're aiming at a small forecast horizon, in which we believe that overfitting at the same period, while it obviously misses lower resolution shocks, may be beneficial in estimating shocks of the same or higher resolution—ie, for the same timeframe, while you will miss the the bigger, macro movements, you will more accurately capture movements of the same or smaller timeframe. As we perform more tests, additional model fit information will be added to this report.

AUTOMATION

OVERVIEW

Instead of solving the problem once, solve it for all-time as data change—such a system can be viewed mathematically as the construction of a function that maps data input into useful, productive output. In a business sense, it's about producing reusable infrastructure and reducing mental cycles wasted on tedium. But this is not trivial. To automate a report, generally one has to:

- Collect data
- Clean data
- Perform analysis
- Organize and format results for presentation
- Publish report

Each of these steps has its own hurdles and, depending on circumstances, some can be very difficult to overcome. However, viewing an analysis not as a single point, but as a function that is applied and re-executed over time is a step that lets you grow your knowledge: rather than doing, undoing, and redoing time and again as time passes and facts change, designing for all-time lets you set the machine in motion and move on to solving new and bigger problems at the same level of resource. It has also been my experience that solving the meta-problem of an automated report produces as a side-effect a better analysis—errors reveals themselves, and the overall thought behind the analysis is forced to become more structured.

THIS PROJECT

With this initial infrastructure up, the ARIMA Forecast Report will be able to evolve in content and complexity as time goes by as an illustration of how labor resources are better allocated by automatic reports. Please check back daily to see what other components have been added. Other reports on topics of related interest are also forthcoming.

DISCLAIMER

GENERAL DISCLAIMER

This is not investment advice. These forecasts are not true. There is a technical reason, a legal reason, and a philosophical reason for that.

The technical reason—the markets we're investigating are relatively efficient. What this means is that if the price history contained information sufficient to systematically predict even the direction of the price series, it would be possible for investors to trade on that information—and the moment they did, that new information would update the price. This happens until the price contains so much of investors' thoughts about the future that it corresponds perfectly with them—then there's nothing to predict. The information now is identical to the price, and we're left again knowing nothing about the future. Furthermore, the data I base this on are about a day old—millions of algo trades and countless clicks from the dark hearts of Bloomberg terminals have happened on this price information before it ever works its way into this model. In short, not only is the information economically insufficient for you to do anything with, you also can't even hope to be the first one to act on it, even if it were. The price is stale.

The legal reason—your trading and forecasting is your own responsibility. The service I provide is only the posting of the outcome of a single, naïve, and not particularly informative, rather standard econometric model. Understanding the outcome of that model may be interesting and may help you think about the past and future, but I am not responsible for your trading losses any more than I am responsible for your trading gains—I do not accept either. By consulting this document, you agree to use this information responsibly or not at all, and that should you incur loss, you alone are responsible. Consider this for all purposes a document of mathematical, econometric, or possibly political interest, but not a financial one.

The philosophical reason—even if the ARIMA model were able to perfectly encapsulate the dynamics, the competing forces, the buffeting tempest between reversion and trend, the wrestling match between bulls and bears, and capture mathematically perfectly how past events will play out in the future—what it does not and cannot ever hope to capture, just like any mathematical model—is the simple fact that the future has not happened yet, and there remain unknown unknowns lurking in Fortuna's wings capable of turning the world upside down. Even if the model were excellent, its excellence is extruded through the contingent die of the past. To the extent the future may be different from the past, and to the extent that, in fact, the way the future differs from the past may very well be different from the way in which the future has historically ever differed from the past (!), this and all models based on the past are wrong. Please keep that in mind always when forecasting, or acting on information in general.

NON-FINANCIAL DISCLAIMER

Design & Analytics is not affiliated with any company reported on and does not at the time of this report have any financial ownership position in the indices above. However, even if we were to do so, we do reiterate, this is not a financial report, and not a recommendation to buy or sell.

SPONSORSHIP DISCLAIMER

The sponsors of this report should likewise not be construed as holding the views presented by the models or their creator. Rather, they support this report for the purpose of reaching out to a mathematically and financially literate audience interested in topics related to the subject matter. They are not endorsing these models or results or making statements of any kind, and explicitly not of a prospective or financial nature to be interpreted as commentary.

CONFIDENTIALITY & TERMS OF USE

By reviewing this document, you agree not to disseminate it without the consent of Design and Analytics. You are more than welcome to link to our site and point as many people as you would like to the page where you found this document (thanks!), but you may not transmit this document directly without licensing authorization.

Part of the revenue that makes this posting possible is derived from a web-based advertising network. If you send this PDF through email to your friend rather than directing her to our site, you directly rob us of advertising revenue, both in that you did not click and in that you did not link, reducing our rank and the overall value of our ads. While there are ways we could have presented this report that would have ensured you were stuck on our site to view this content, we chose not to use these methods, and that's largely for your benefit. We ask that you respect our freedom to compose this document in our chosen format, especially as it grants you the capability of offline viewing, and that you are considerate of our property and our requirements upon its usage.

If you do not agree to abide by this policy, please do not use (look at) this document or any part of it. This document remains exclusive property of Design & Analytics. While you are able to view it yourself for personal purposes or professional interests, please ensure that all dissemination is done by directing viewers to the following link: www.designandanalytics.com/reports

LEGAL

Published newsletters are exempt from SEC registration as financial advisors under Section 401(f) of the Uniform Securities Act.

APPENDIX

CREDITS

This document is created using tools from

- The R project
- The LaTeX typesetting system
- The XeTeX engine
- The Quantmod package
- The ARIMA package
- The ggplot package

MOTIVATION AND OTHER RESOURCES

If you are interested in creating a platform resembling this, or utilizing similar infrastructure, please let me know. I believe very strongly in helping develop infrastructure to make reports happen frequently, easily, repeatably, continuously, and without human intervention in business, academia, and government—any place where you prefer or would value things to be true, timely, and efficient.

I also believe that finished, publication-quality analyses for reporting purposes should be done in statistical or programming languages, such as Matlab, R, or Python, and typeset with tools like LaTeX, rather than word processors—a report in spreadsheets like Excel, is a report still in Beta. Even artists and designers can gain from a programmatic and repeatable approach to their craft—and even better when we integrate the analysts and designers together.

Design & Analytics can help you develop your analyses instead of repeating the same ones mechanically, wasting both labor and thought, freeing you to do more with less. As a rule of thumb, if you ever do the same action more than once on a computer, you're not using the computer, the computer is using you. If you're being used, you're welcome to contact us for pointers.

QUESTIONS, PROBLEMS, ETC?

If you'd like to get in touch for support, thanks, *errata*, or other, please contact us at the following address: contact@designandanalytics.com

VERSION INFORMATION

Forecasting module:	version 1.0.0
Design module:	version 1.2.0
Policy and text module:	version 1.1.0
Production and dissemination module:	version 1.0.0