The Top 20 Business Problems

That you can solve with Data Science.

Matt Dancho

Version 1.0

Table of Contents

Section 1: Matt's Matrix of Top 20 Business Problems (that can be solved with Data Science)

Table 1 captures the **Top 20 Business Problems** that can be solved with data science and the R programming language. We will then discuss each more in depth with supporting information in Section 2.

Rank	Business Problem	Financial Impact	Algorithms Used	R Packages
1	**Churn (Customers & Employees)**	$$$	Classification (Logistic Regression)	tidymodels h2o
2	**Time Series Forecasting (Sales Prediction)**	$$$	Time Series Forecasting	modeltime
3	**Lead Scoring (Email & Sales)**	$$	Classification (Logistic Regression)	tidymodels h2o
4	**Production & Deployment**	$$$	Any	shiny plumber
5	**Customer Segmentation**	$$	Clustering (K-Means) Dimensionality Reduction (PCA, UMAP)	stats umap
6	**Fraud Detection (Insurance Fraud, Credit Card Transactions)**	$$$	Anomaly Detection	H2O
7	**Intelligent Product Recommendations**	$$	Association Rules	recommenderlab
8	**Build a Competitor Database**	$	Web Scraping	rvest
9	**A/B Testing (Website & Email Marketing)**	$$	Statistical Tests	tidyverse

Rank	Business Problem	Financial Impact	Algorithms Used	R Packages
10	**Marketing Channel Attribution**	$$	Markov Models	ChannelAttribution
11	**Financial & Portfolio Analysis**	$$	Optimization	tidyquant
12	**Sentiment Analysis**	$	Text Analysis & Natural Language Processing	tidytext
13	**Customer Network Analysis**	$$	Network Analysis	tidygraph ggraph
14	**PDF Document Text Extraction**	$	Text Analysis & Image Processing	tidytext
15	**Customer Lifetime Value (CLV & RFM)**	$$	Machine Learning Classification & Regression	tidymodels h2o
16	**Marketing Mix Modeling (MMM)**	$$	Optimization	robyn
17	**Automate Enterprise Data Pipelines with Big Data**	$$	Spark	sparklyr
18	**Risk Analysis & Business Simulation**	$$	Simulation & Markov Chain Monte Carlo (MCMC) Sampling	tidyverse
19	**Probabilistic Business Prediction with Bayesian**	$$	Bayesian	brms bayesian
20	**Business Process Optimization (Linear & Non-Linear Constraints)**	$$	Optimization	ROI ompr

Section 2: The Top 20 Business Problems Explained

Problem 1: Churn (Customer & Employees)

Churn is when a customer or employee leaves the company. Churn is a very costly problem that eats away at the organization's revenue when customers leave or employees quit.

Churn is a **binary classification problem** where we are trying to determine whether or not something will happen. Hence either a 1 or a 0. However, the 1 or 0 is actually just a label that is based on a class probability (e.g. class 1 = 0.75 and class 0 = 0.25).

To solve this problem, I use:

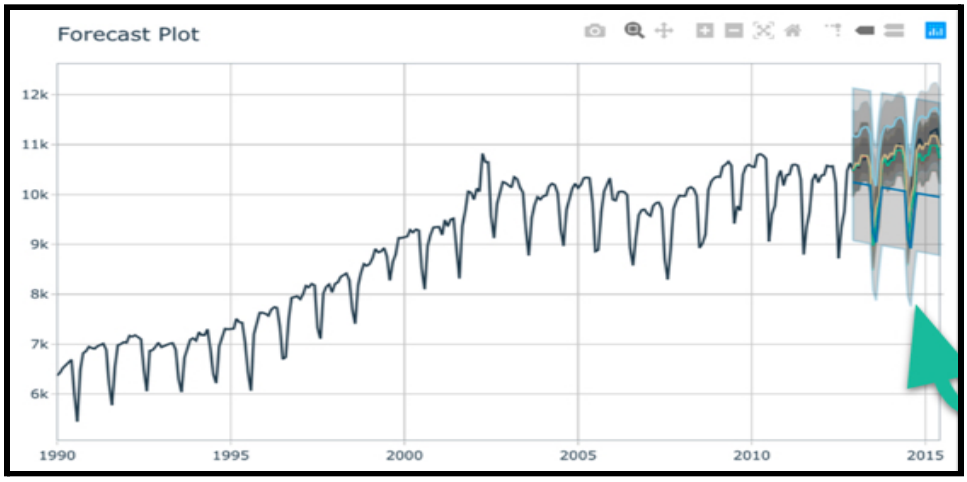
- **H2O Automatic Machine Learning** to predict the churn risk of each employee
- **Local Interpretable Model-Agnostic Explanations (LIME)** to explain why the model predicts the employee as likely to stay or leave the company.

I then make a **recommendation algorithm** to help organizations decide which actions to take. I cover how to solve this problem in-depth in my 5-Course R-Track Program.

Problem 2: Time Series Forecasting

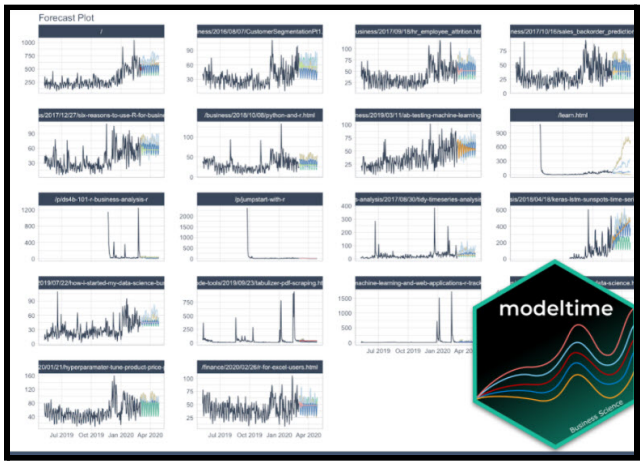
Time series forecasting is the process of using historical data (such as sales revenue over time) as inputs to predict (estimate) the next H periods (often denoted the forecast horizon). Everything from hiring new employees and purchasing raw materials via the supply chain is driven off of a demand forecast. Therefore, organizations that improve their forecasting can save millions of dollars.

A demand forecast may look something like this for a *single time series*.



Companies need to make **hundreds of thousands of forecasts** every week or month (think of forecasting for every product your company makes). This is called **high-performance forecasting.**

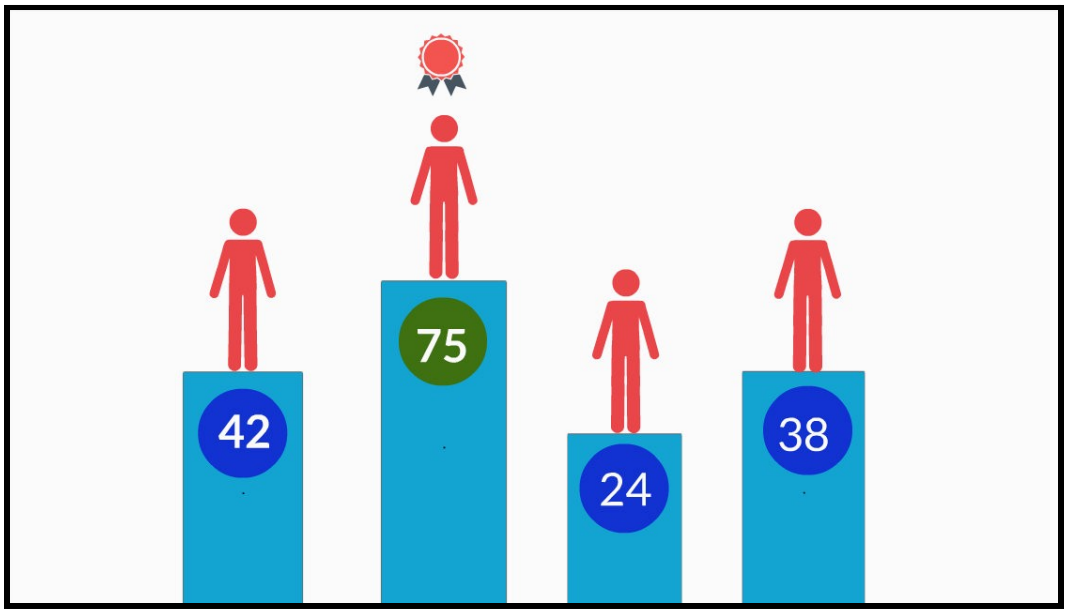
I cover high-performance forecasting using the winning strategies from 4 time series competitions using my **modeltime & timetk** R packages (forecasting R packages I created) in my 5-Course R-Track.



Problem 3: Lead Scoring (Email & Sales Marketing)

Lead scoring is a binary classification problem where you want to predict the probability of a lead turning into an order. Traditionally this is done by measuring the number of actions taken by a customer and then calculating a score by weighting each action in terms of the business's idea of how important the action is.

Data science can be used to more accurately predict the customers likelihood by combining the customer's actions taken (traditional method) and additional data like customer's time with the company, age, geographic region, and more. The result is a more accurate lead score.



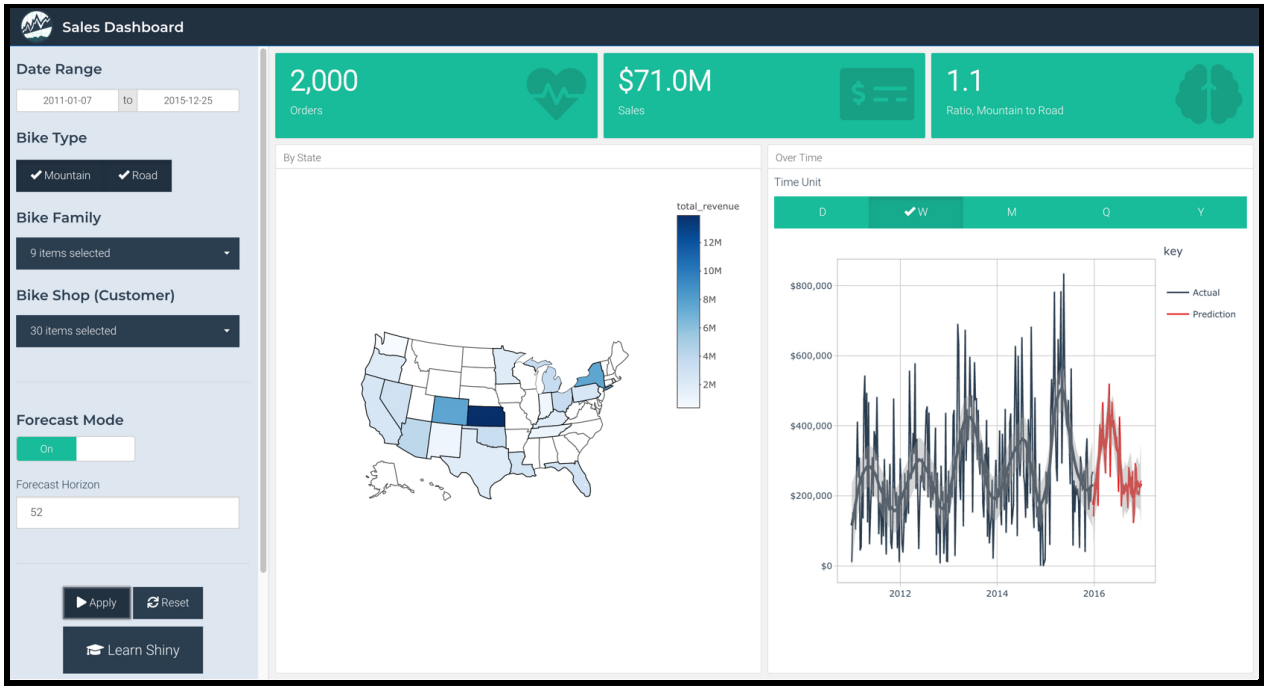
Lead scoring is a variant of Problem 1: Churn. The main difference is that instead of classifying the customer as Yes/No, we wish to **rank each of the customers** in order of most likely to purchase to least likely. Customers that are more ready to purchase can be marketed to, and those that are least likely should be nurtured.

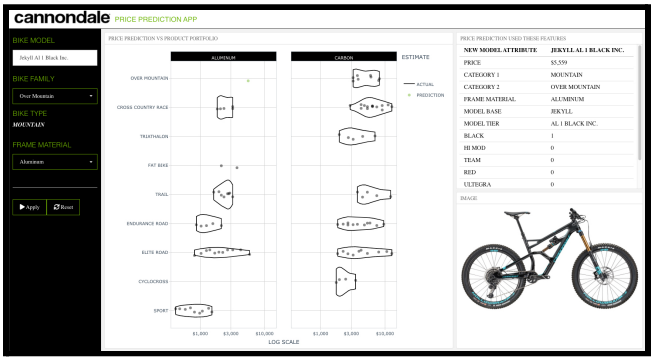
I teach how to solve the binary classification problem you will face when you do lead scoring in my 5 Course R-Track. My preference is to use **H2O Automatic Machine Learning and LIME for explainable machine learning**.

Problem 4: Production & Deployment

Production is the process of taking a machine learning algorithm or decision-making analysis code (script) and converting it into a form that is usable by the business. For R, the most common methods are to build and deploy shiny apps or plumber APIs.

The most common production app is a dashboard. The advantage of using Shiny is that a predictive analysis that is codified in R can quickly be converted into an app. Think of the app as the package for the code, and users interact with the app (using dropdowns and buttons), which then automates the process of running the code behind the scenes. Shiny apps are covered in **Shiny Web Apps Parts 1 & 2 Courses.**



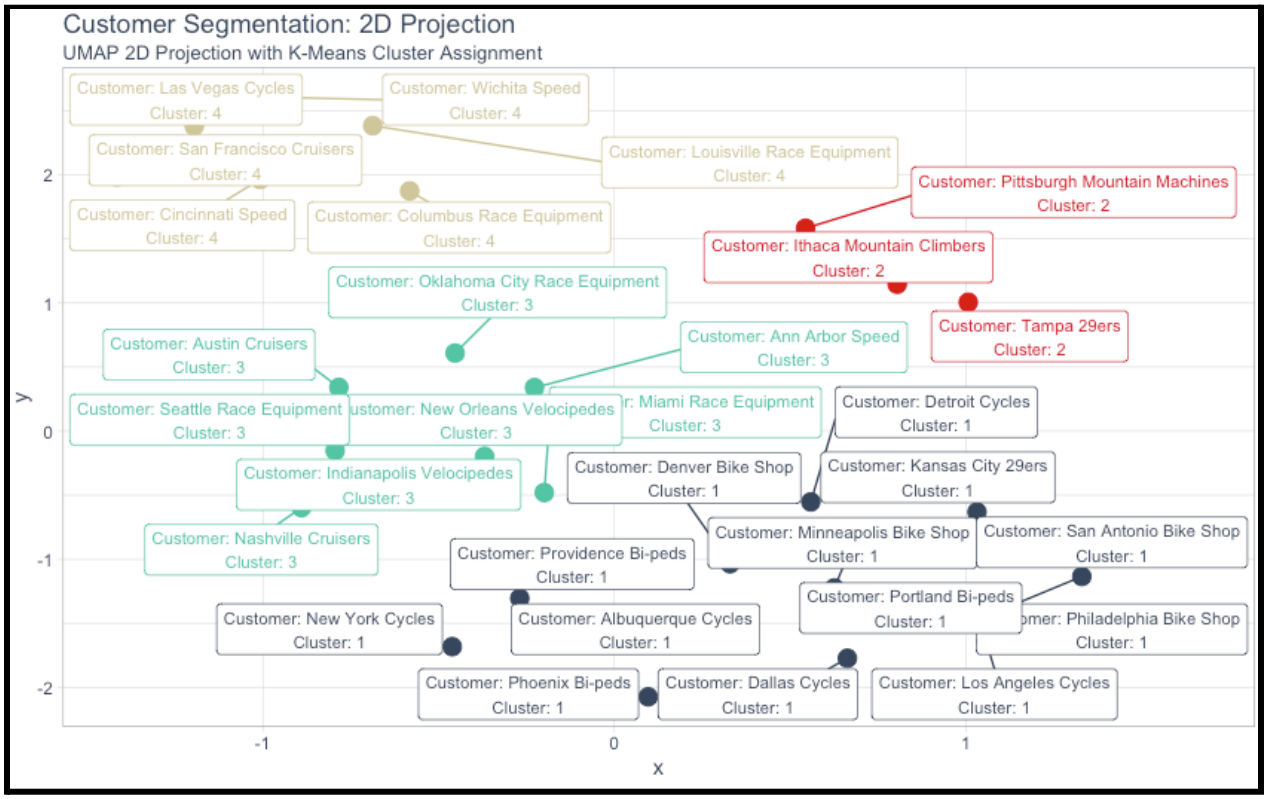


In my 5-Course R-Track, you learn how to build 4 shiny web applications (2 shown here) that cover the most common scenarios:

- Pulling data from APIs,
- integrating predictive models, and
- Displaying visualizations (images, maps, and statistical / time series plots) like those shown.

Problem 5: Customer Segmentation

Customer segmentation is the process of separating customers into groups based on common purchasing behavior (e.g. this group tends to buy more shoes vs this group tends to buy more handbags). Companies can then target these segments and tailor marketing messages most commonly via email. **Segmentation increases revenue and improves customer satisfaction. Win-win.**



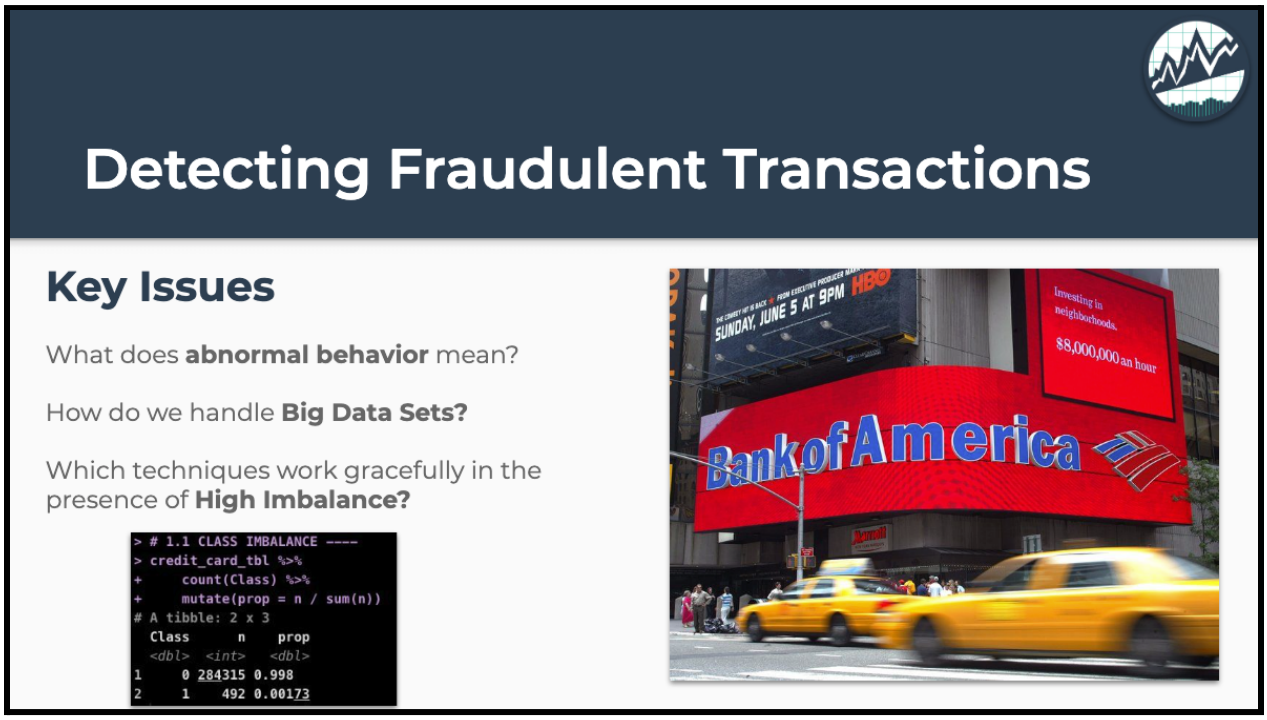
In my 5-Course R-Track, you use clustering and dimensionality reduction to visualize segments of customers. I show you how to apply:

- **K-Means:** A common clustering algorithm that does well at identifying segments within business data
- **UMAP:** A dimensionality reduction technique that allows us to plot the variance of the data in two dimensions

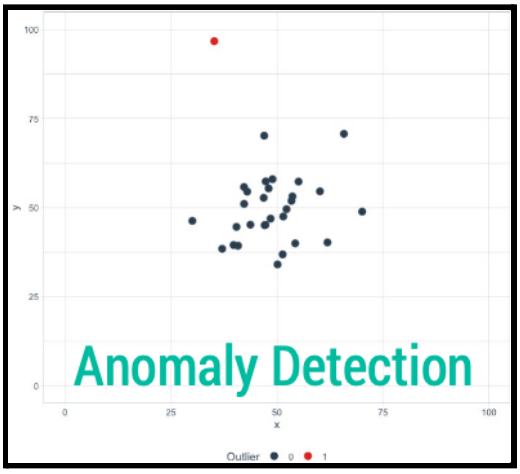
The result is a customer segmentation that can be used to target similar customers with tailored marketing messages.

Problem 6: Fraud Detection with Anomaly Identification

Fraud detection is identifying abnormal transactions or behavior from customers. Often the abnormal behavior can indicate that something isn't right about the transaction or insurance claim, and therefore requires a hold until further evaluation can take place.



A common method is **anomaly detection** which is used to identify abnormal events. This can be interpreted as suspicious behavior.



I use **H2O Isolation Forests** for anomaly detection using the h2o package in R.

For **time series anomaly detection**, I created the anomalize R package. I have since incorporated most of the time series anomaly detection functionality into my **timetk R package**.

Problem 7: Market Basket Analysis (Recommendations)

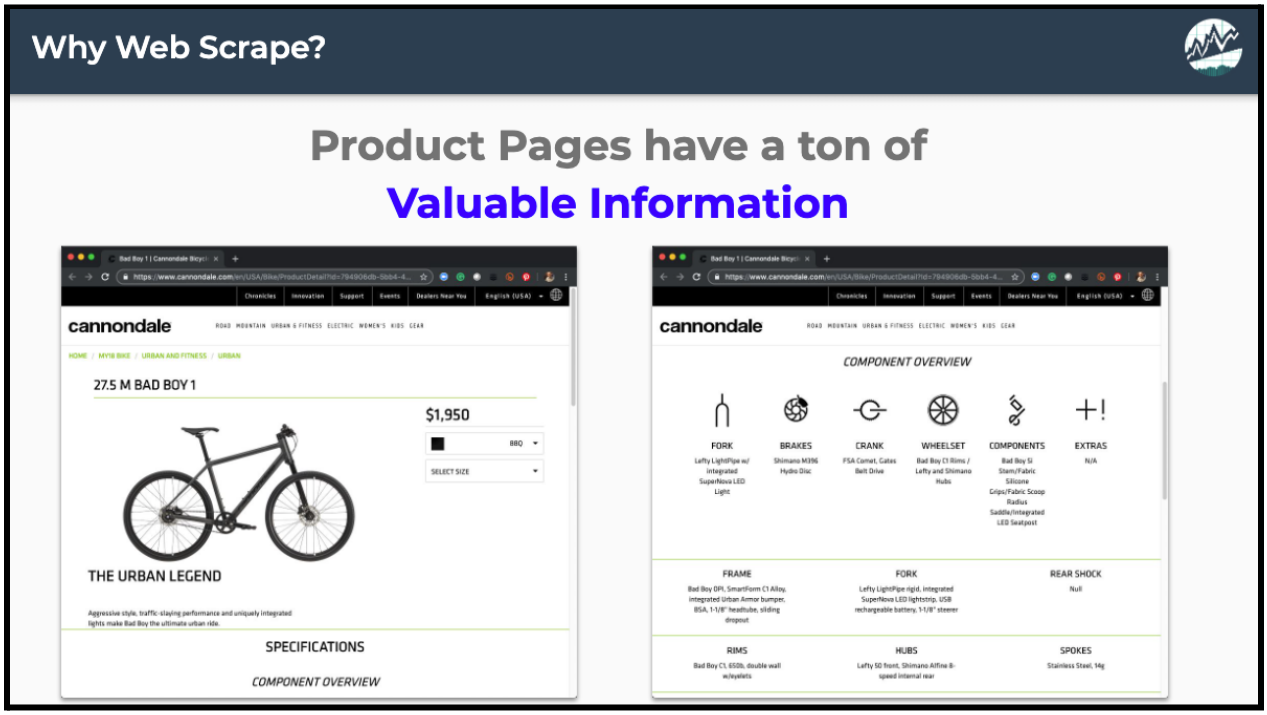
Market Basket Analysis is a technique used to identify which products get purchased together. We can then use this information to recommend products based on a customer's cart (prior to checkout). This tactic increases the Average Order Value (AOV).

One of the most common techniques I use to mine customer cart's and make recommendations is **association rules.** I use the arules R package to implement various association rules mining.

Problem 8: Web-Scraping to Build a Competitor Database

Building and tracking your competitors can help companies better understand competitor behavior and look for areas of improvement to seek as a competitive advantage in the marketplace.

The internet has a massive amount of product and service information, and **many of your company's competitors list this information online**. Your organization can take advantage of the free data by building a strategic database. The question is then, how best to collect the competitor data? That's where web-scraping comes in.



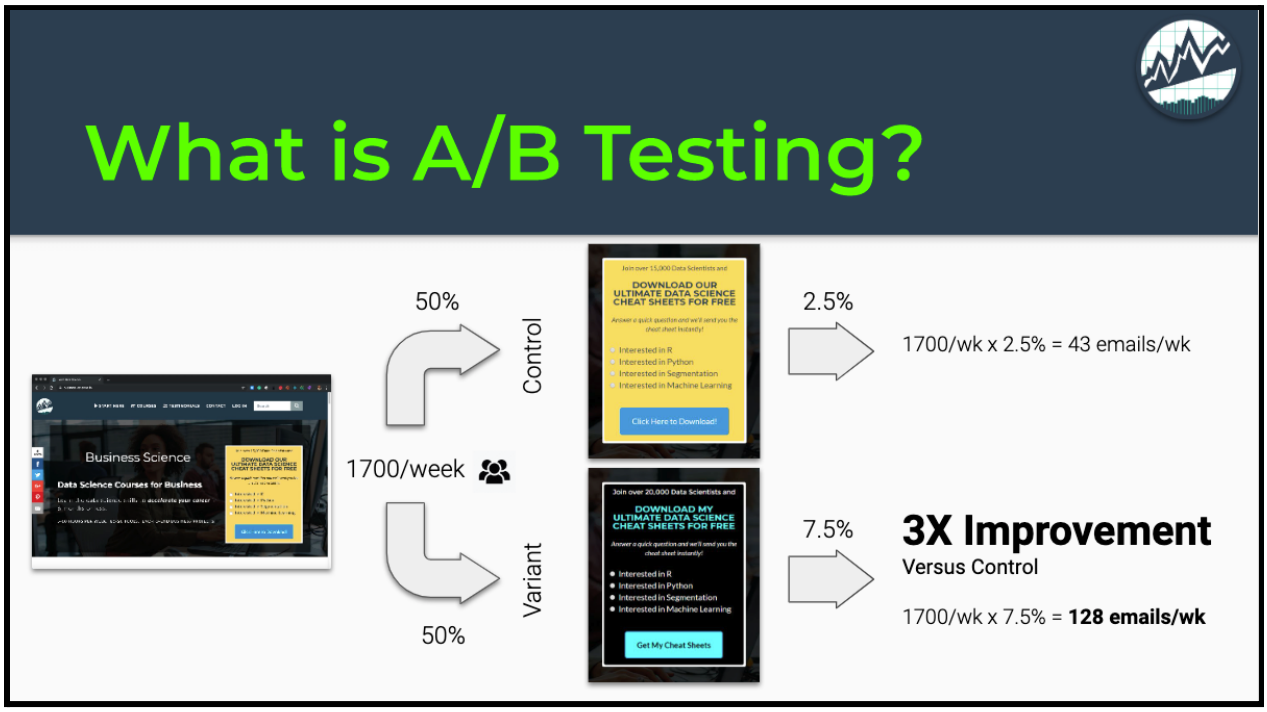
Web-scraping is the most common method for developing a database of competitive information. In R, I commonly use the:

- rvest R package to extract HTML data into R
- tidyverse to convert the raw HTML into tabular data

For those that need to learn R, I cover the tidyverse in my 5-Course R-Track.

Problem 9: A/B Testing (Website & Emails)

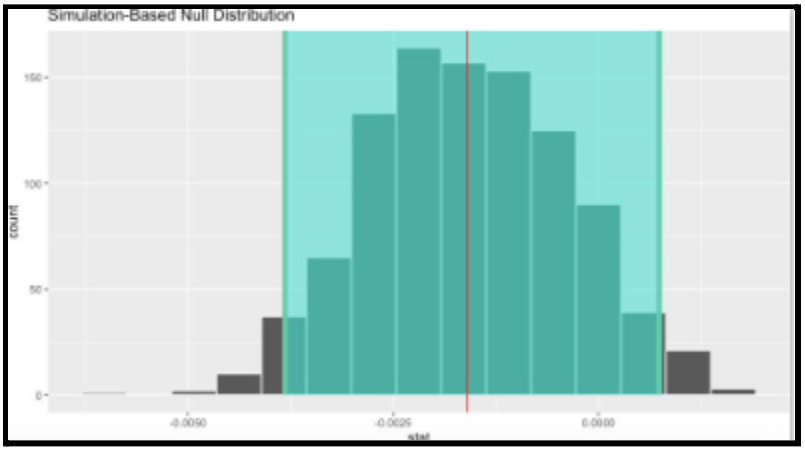
A/B testing is the process of testing changes to websites and emails by dividing the traffic into two and randomly assigning to two variants. It's most common to make small changes to each variant, and then to test the impact using statistical analysis.



In R, I use the **tidyverse** to analyze web-traffic and email split tests.

Then I add **confidence intervals** to my assessments to determine if we should accept or reject the null hypothesis.

The tidyverse is covered in my 5-Course R-Track.



Problem 10: Marketing Channel Attribution (Budget)

Marketing Attribution (Channel Attribution) is a common problem that marketers face when allocating the correct amount of money to each marketing advertisement channel (e.g. Facebook, Instagram, Paid Search) to optimize for conversions.

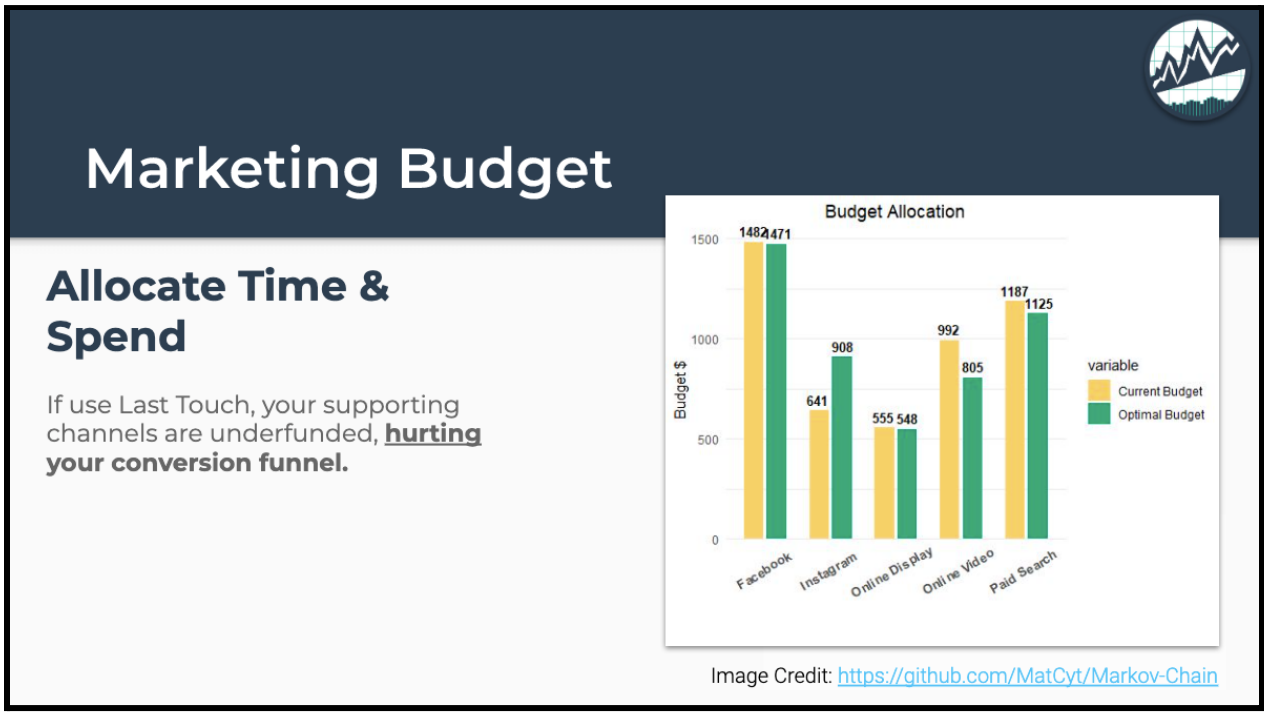


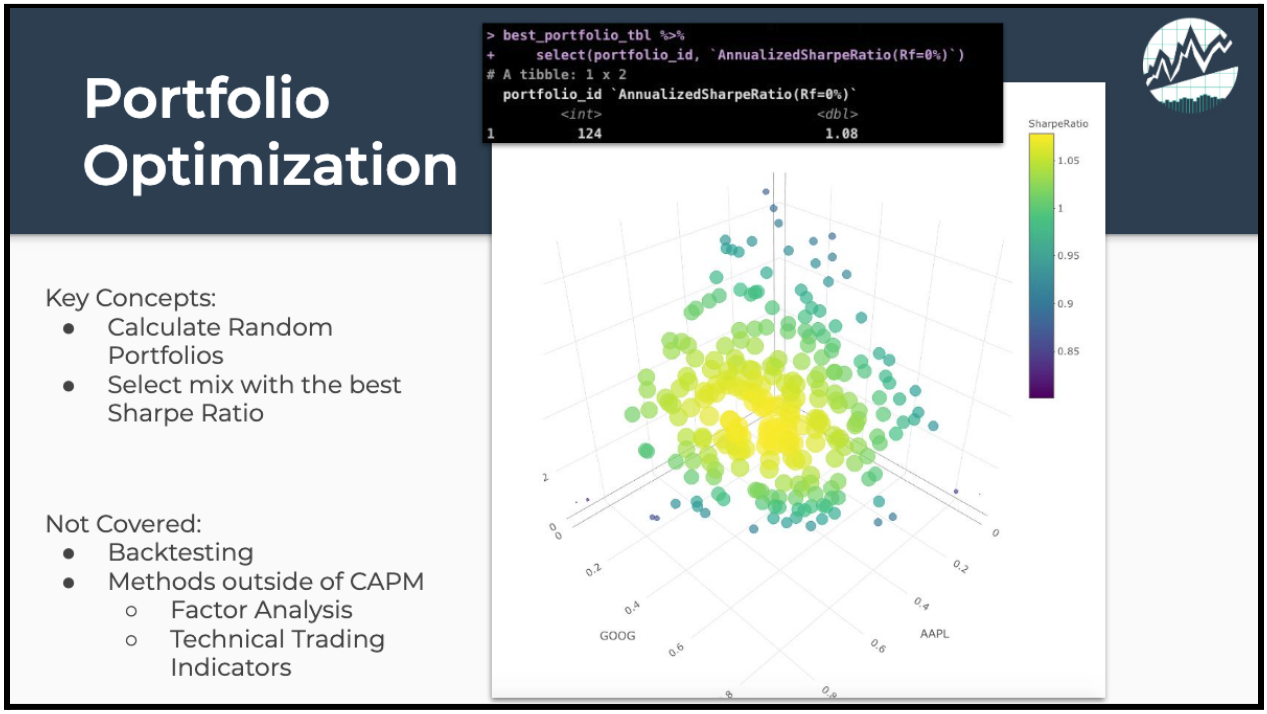
Image Credit: https://github.com/MatCyt/Markov-Chain

I use the **ChannelAttribution** R package in R to help solve the Marketing Attribution Problem.

A newer technique is the **FB Robyn algorithm (robyn package)** is available for Marketing Mix Modeling (MMM) (See Problem 16), a similar and related business problem.

Problem 11: Financial Investment & Portfolio Optimization

Financial investment and portfolio analysis is the process of aggregating stock data, using technical indicators, making portfolios and optimizing to reduce key metrics like the Sharpe Ratio in order to improve investment decision-making.



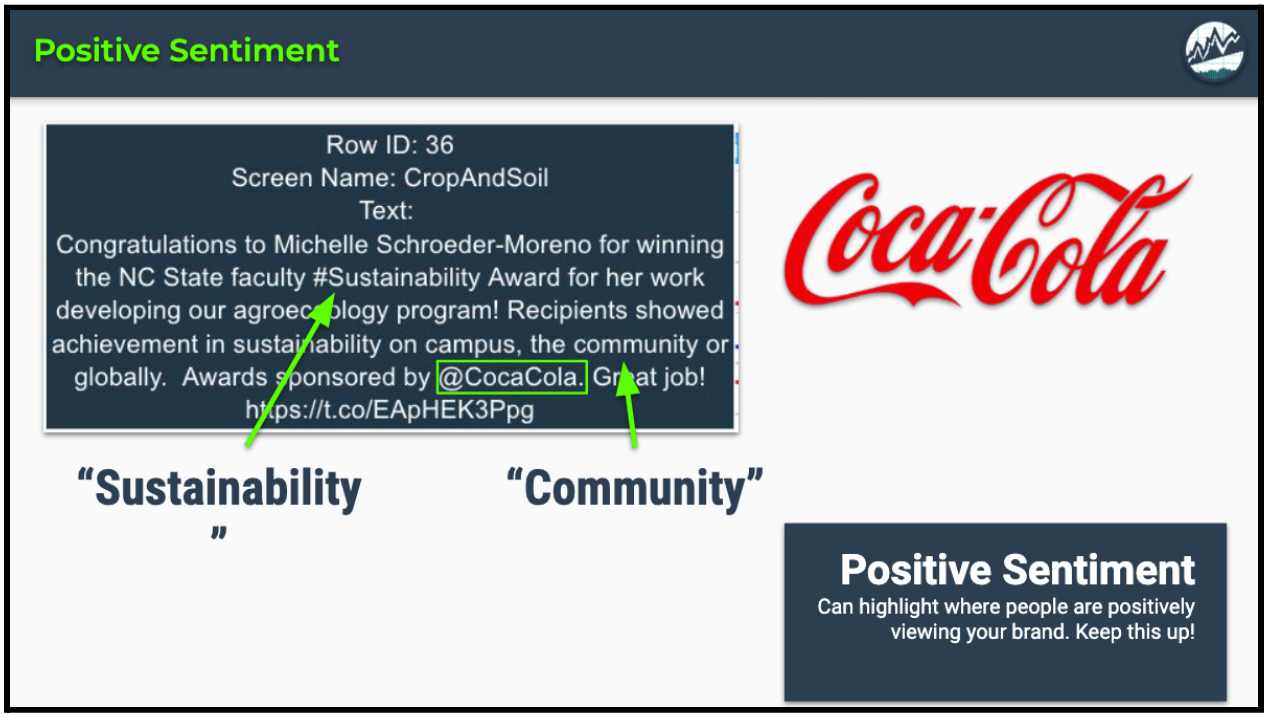
I created the **tidyquant R package**, which provides a streamlined workflow for financial investment and portfolio analysis.

Tidyquant has:

- An API to pull financial data into R
- A full suite of financial and technical analysis tools
- Integrated portfolio analysis for combining financial assets and analyzing a portfolio

Problem 12: Social Sentiment Analysis for Brands

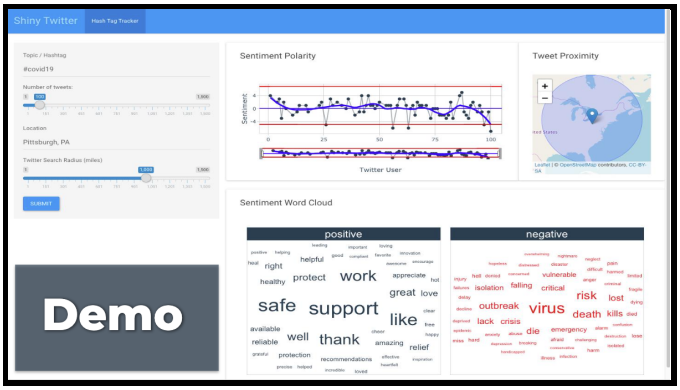
Sentiment Analysis is used to assess the positive or negative score of text. One key application is in the use of **Brand Sentiment**, which can be used for companies like Coca-Cola to analyze social media sentiment as a gauge for how well their brand is being viewed to the public. Positive sentiment indicates your customers are viewing your brand positively, and negative sentiment could be cause for concern.



This shiny app I made combines:

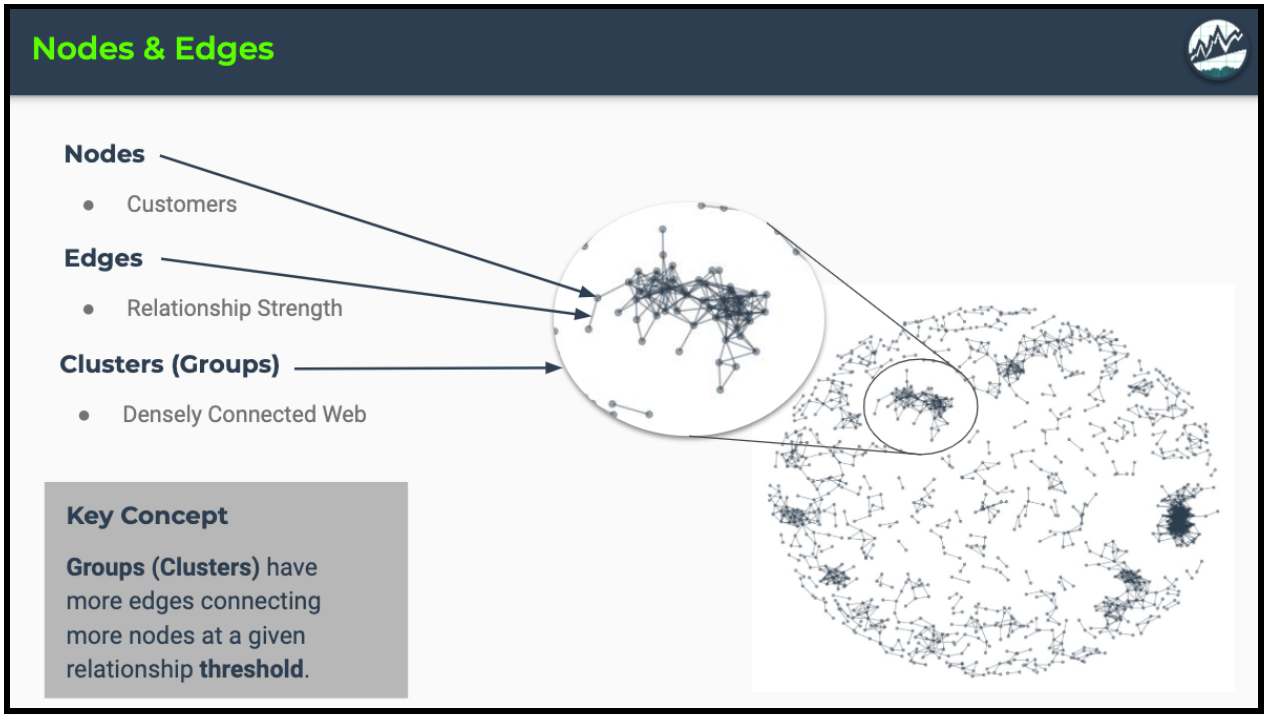
- Rtweets Twitter API
- Tidytext Text Analysis
- Tidyverse Data wrangling & Viz
- Shiny Web Apps

To make a Shiny App for Sentiment analysis for Brand Assessment.



Problem 13: Network Analysis for Influential Customers

Network analysis can be used to identify influential customers. The key difference between network analysis and clustering is that in clustering you just have a group assignment label, but in network analysis you gain the ability to understand which customers are related AND which customers are the most influential. Once you know which customers are most influential, you can design marketing messages to the "influencers", which will have a higher effect on the entire group.



I use these two R packages for network analysis:

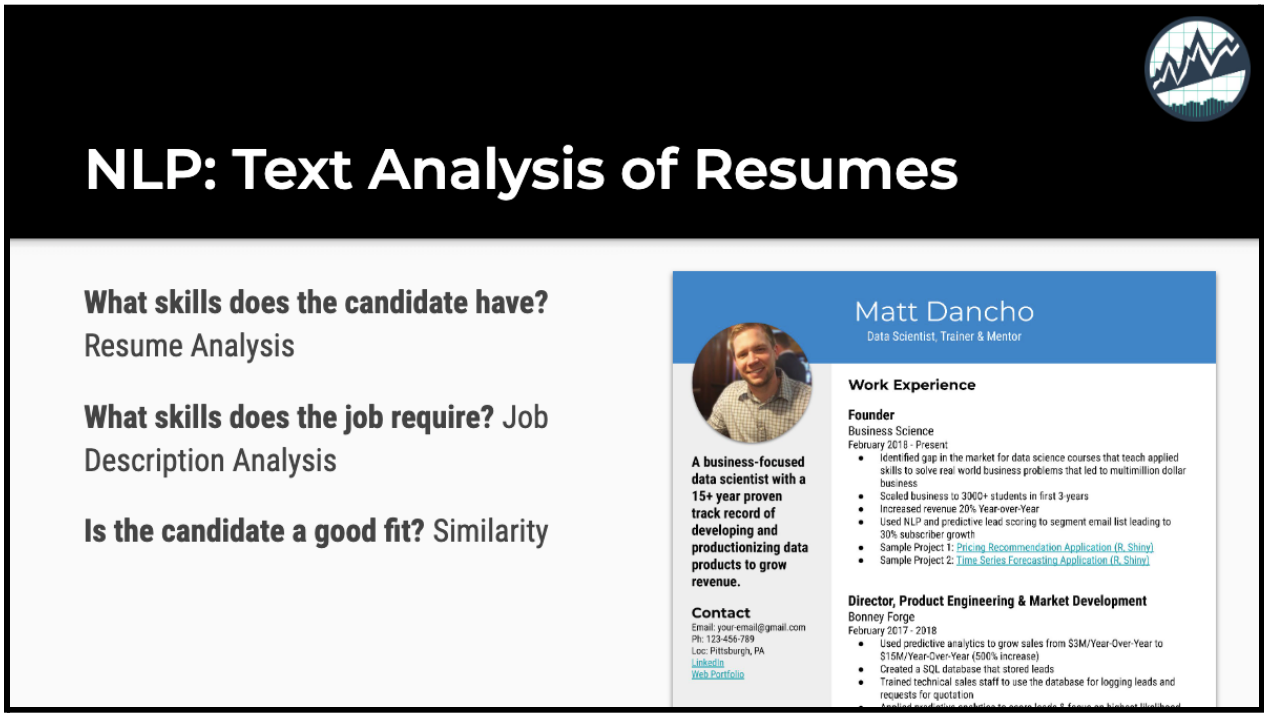
- **tidygraph** for network analysis
- **ggraph** for network visualization

Both of these packages follow the **tidyverse-centric concepts** learned in my 5-Course R-Track. Once you learn the tidyverse, these network analysis R packages are easy to pick up.

Problem 14: PDF Text Scraping and Document Analysis

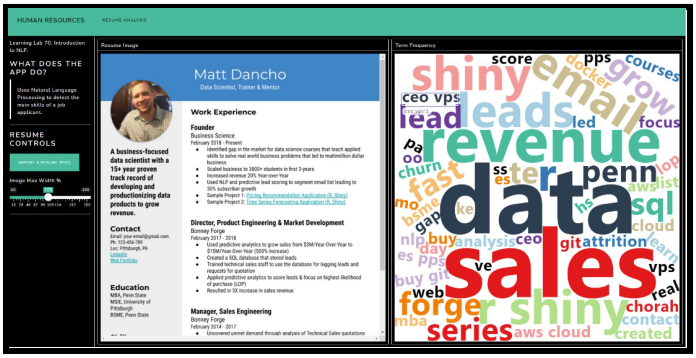
PDF and Document Text-Scraping is becoming increasingly popular. Businesses have mounds of unstructured data stored as PDF, Word, and Excel documents. This data can be mined for important information.

One example is **Resume Analysis.** I use a technique called Natural Language Processing (NLP) to evaluate the similarity of a job applicant's resume to the company's job description. This allows me to filter out unqualified candidates, without wasting time reviewing thousands of resumes.



This shiny app I made combines:

- **shiny** for web applications
- **tidytext** for text analysis
- **magick** for images
- **tidyverse** for similarity

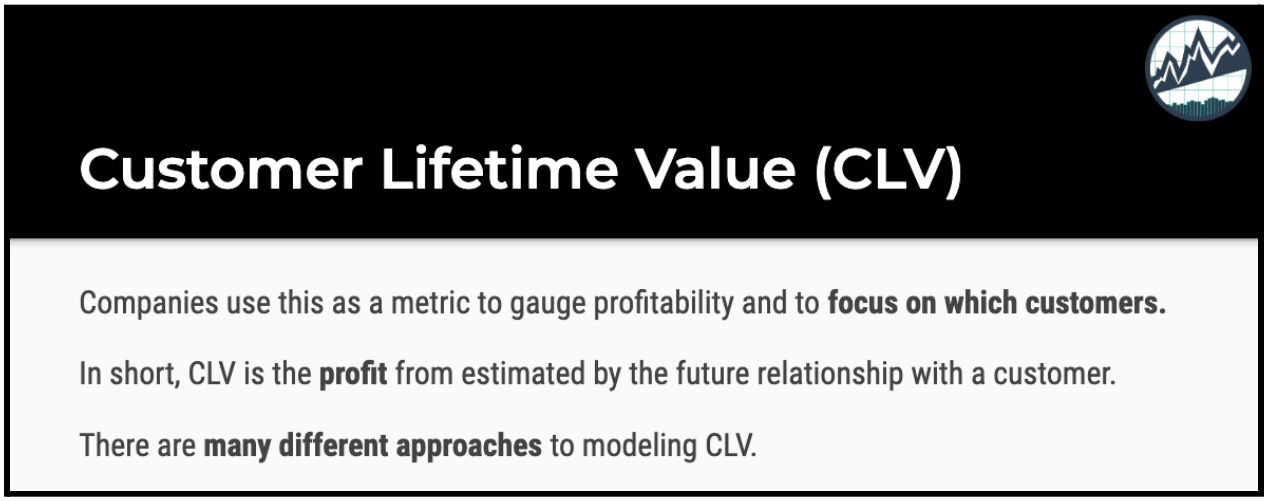


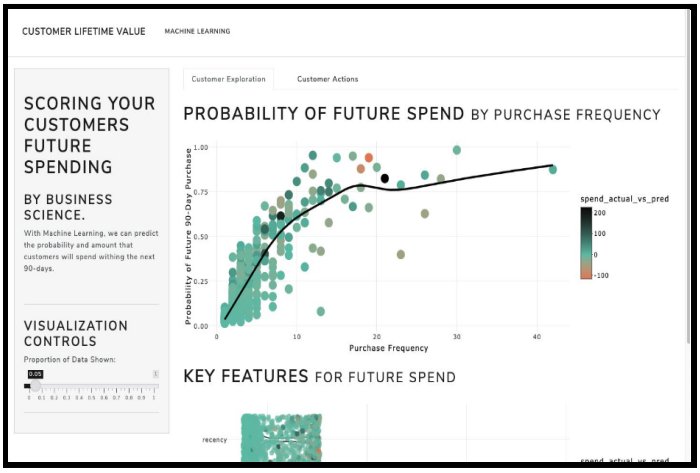
To make a **resume analyzer** web application.

Problem 15: Customer Lifetime Value (CLV & RFM)

Customer Lifetime Value (CLV) is used to highlight which customers to focus on. Companies use CLV to **estimate the profitability** of the future relationship with a customer. Most algorithms use **RFM (Recency-Frequency-Monetary)** attributes for customers to help estimate customer lifetime value.

There are two approaches to CLV including (1) traditional economic and (2) machine learning approaches. I favor machine learning as I normally get higher accuracy and better predictive insights when I model using the newer technology (machine learning).





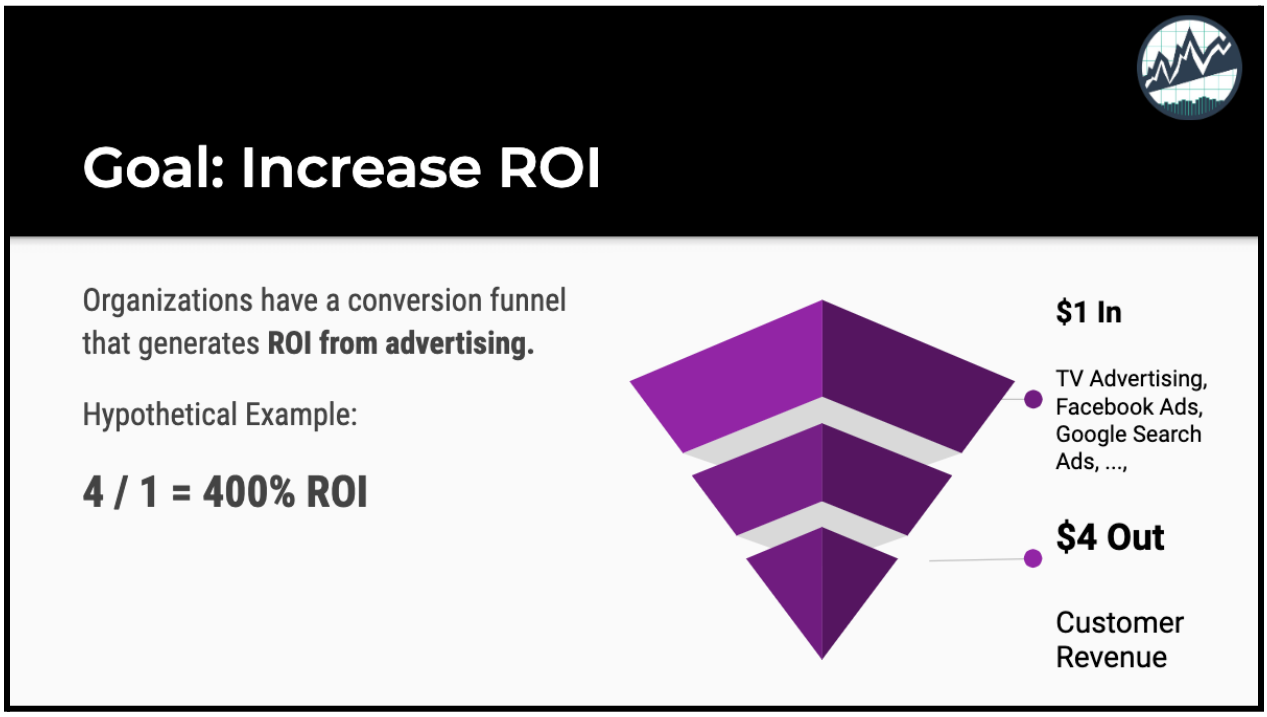
This shiny CLV app I made combines:

- **shiny** for Web Apps
- **tidymodels** for CLV estimation using Machine Learning

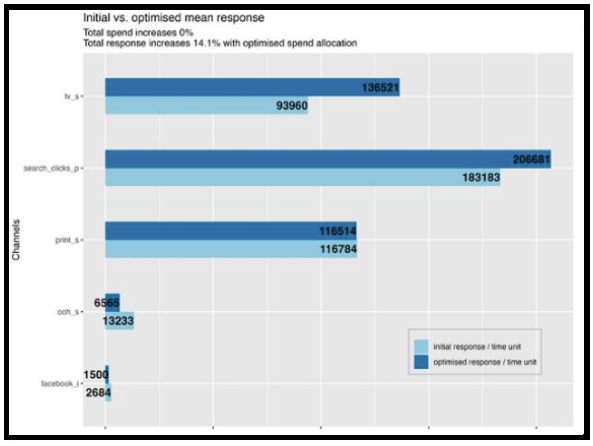
The customer lifetime value app **scores the customers** by estimated future purchasing in the next 90-days.

Problem 16: Marketing Mix Modeling (MMM)

Marketing Mix Modeling (MMM) is a technique that helps marketers quantify the impact of advertising and marketing actions on the customer revenue that's generated. MMM is often used in the **marketing budget allocation process** to account for causation in the key marketing variables such as paid media variables, organic variables, contextual variables, and time-based and seasonal variables.

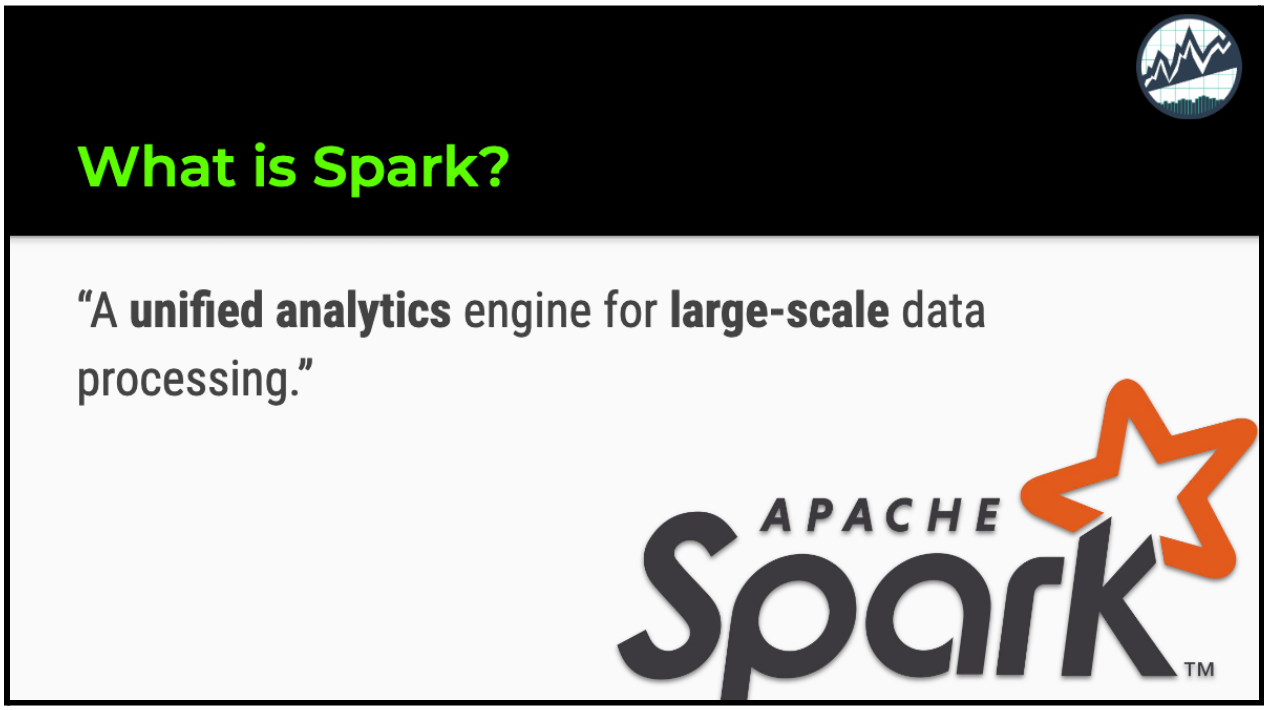


Here I'm using the **Robyn R package** developed by Facebook to help *automate* the MMM process, thus reducing the human bias in budget allocation decisions.



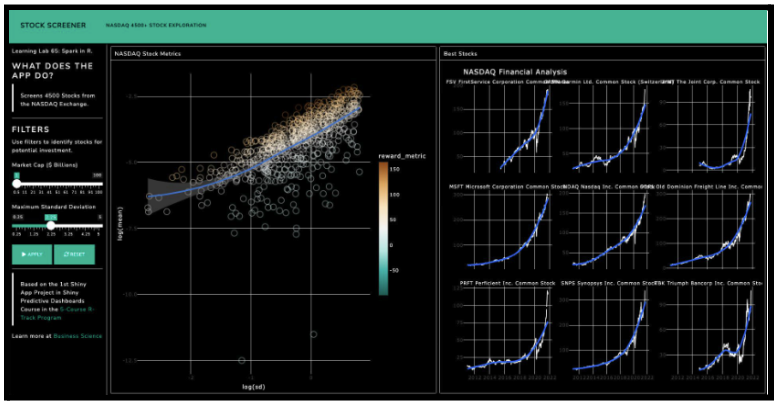
Problem 17: Automate Enterprise Data Pipelines with Big Data

Data is getting bigger to the point where it becomes difficult for people to process enterprise data on their computers. To meet these big data demands, organizations are adopting tools like Spark, which is a unified analytics engine for large-scale data processing.



This shiny app I made combines:

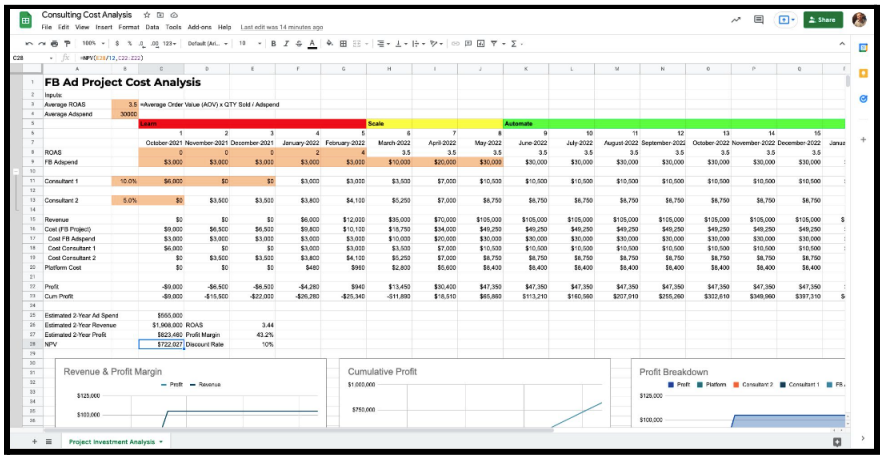
- **shiny** for Web Apps
- **Spark** for Big Data (**sparklyr**)



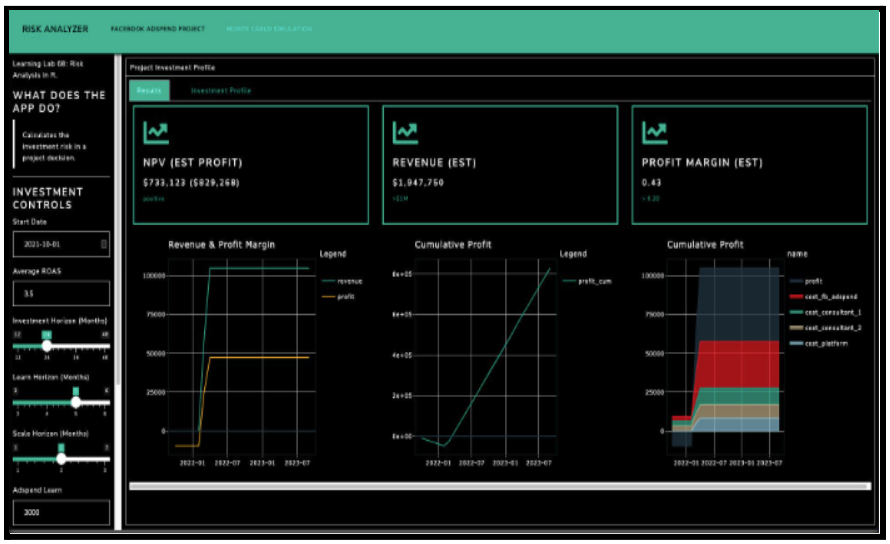
To automate the processing of 4,500 stocks I'm evaluating on the NASDAQ stock exchange.

Problem 18: Risk Analysis & Business Simulation

Business analysis and simulation is done in Excel using expensive add-ons like the Palisade Decision Tools suite. But, R is a free replacement, which can save hundreds of thousands of dollars in expensive software, reducing errors, and automating the risk analysis process.



R can be used to convert this Excel Capital Expenditure Investment Project from a single Net Present Value (NPV) Calculation into a full **Risk Simulation with confidence intervals.**



Here I'm using:

- The **tidyverse** for simulation
- **shiny** for web applications

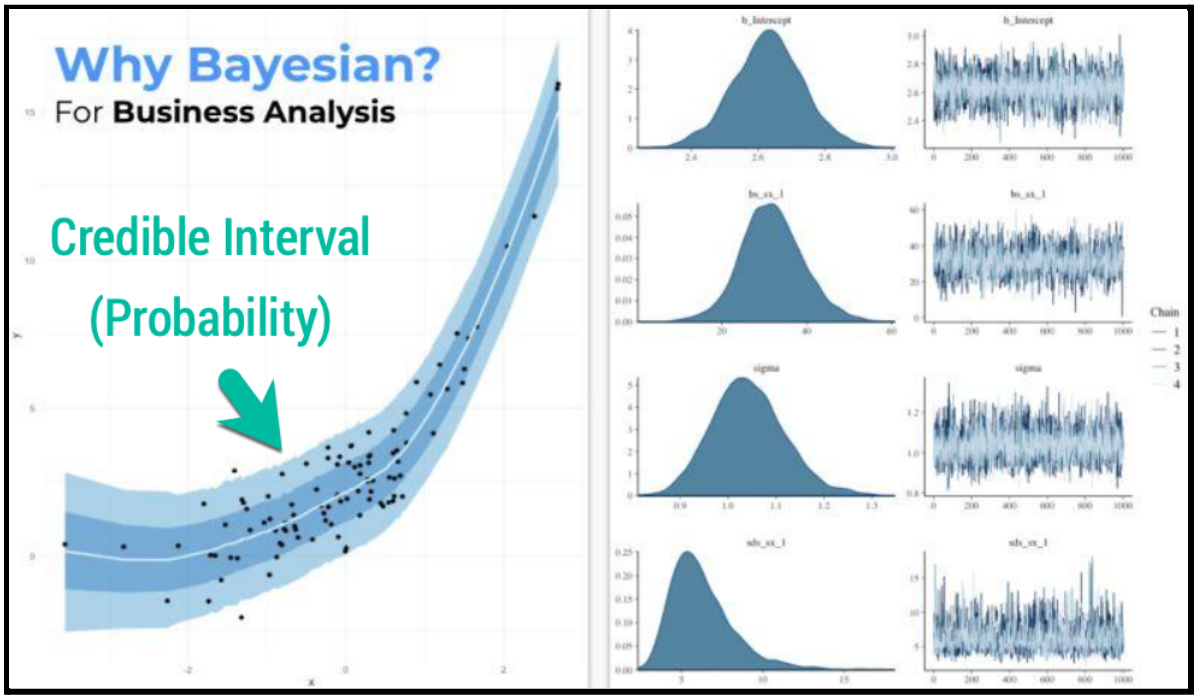
The result is a shiny web application that users can use to **automate business simulation and risk analysis.**

Problem 19: Probabilistic Business Prediction with Bayesian

In prediction, many times it's more important to **understand your confidence** than simply make a prediction. Understanding confidence allows organizations to make better decisions that reduce their *downside risk*, a key factor in determining the profitability of a model.

This is where **Bayesian methods** can help us to predict with a more accurate probability estimate. Bayesian uses a unique approach called Markov Chain Monte Carlo (MCMC) resampling to estimate a Credible Interval.

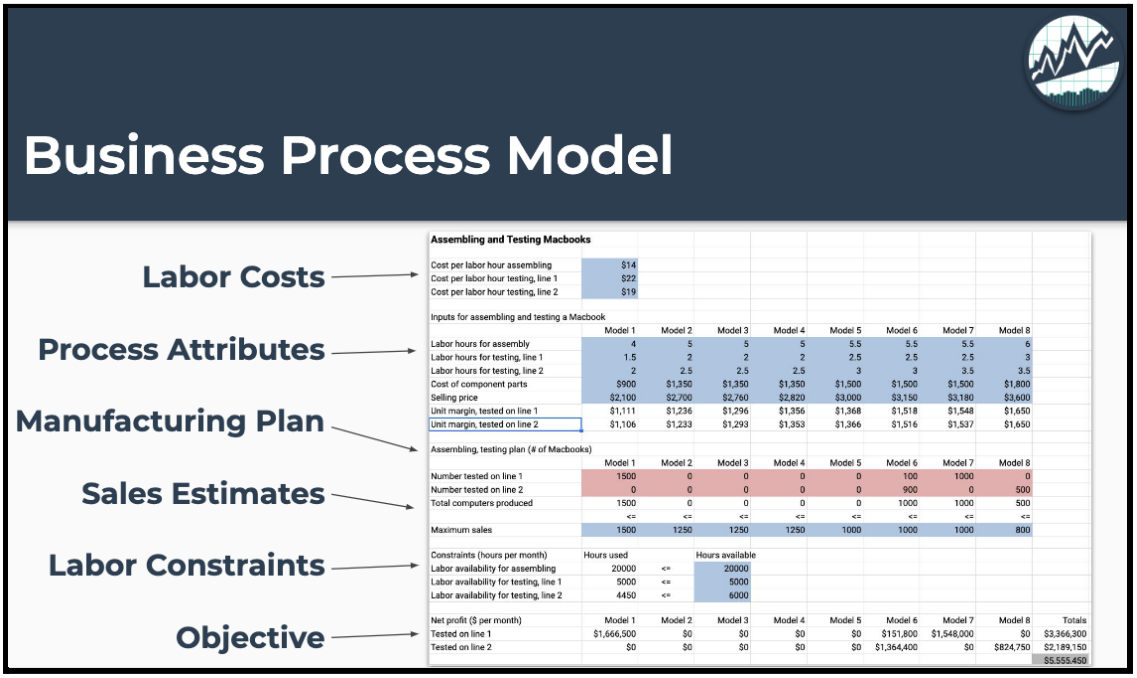
The Bayesian Credible Interval is a probability that surrounds the prediction. This differs from the traditional approach by adjusting to the variance of the data rather than assuming a constant variance (which is incorrect in most cases).



For bayesian analysis and probabilistic prediction, I use the **brms R library and a tidymodels extension named bayesian.** Once you learn the tidymodels framework, which I cover in my 5-Course R-Track Program, extending your analysis to Bayesian is easy to do.

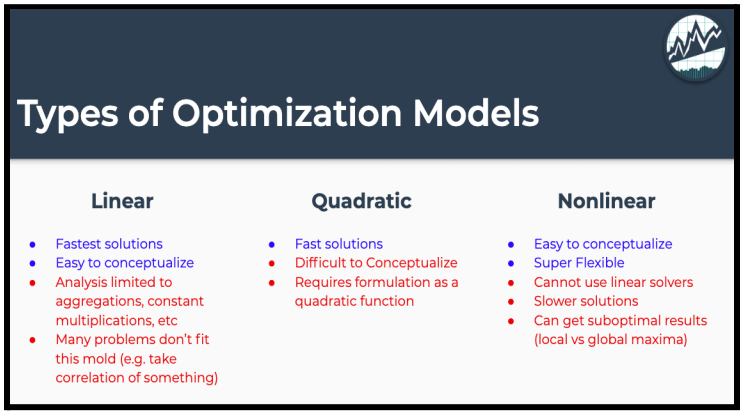
Problem 20: Business Process Optimization (Linear & Nonlinear Constraints)

Business process optimization is used to optimize an objective (e.g. maximize profit) given that a company has limited resources (e.g. production line constraints, labor constraints, supply constraints).



The two R packages that I use are:

- **ompr:** linear constraints (y = mx)
- **ROI:** non-linear constraints (y = mx^2)



Implementing these two methodologies, I've optimized problems in R including:

- **Linear:** Business processes to determine the optimal number of products to maximize profit
- **Nonlinear:** The optimal stock portfolio mix to maximize the Sharpe ratio

Section 3: What are your next steps?

The next question is, *“How do you learn to solve business problems with data science?”*

My name is Matt Dancho, and I turn business people, data enthusiasts, and people that have an obsession with data science into elite data scientists.

I’d love to show you your next steps.

Join the program that has helped over 3,000 students launch their data science career.

If you want to see your next steps then click the button below…

Show Me