Data Playbook

for the Off-Grid Pay-As-You-Go Sector

WORLD BANK GROUP

LIGHTING GLOBAL
Catalyzing markets for modern off-grid energy

GOGLA
Global Off-Grid Lighting Association

Contributing Partners

angaza

Lendable

LEGAL DISCLOSURE

ACKNOWLEDGEMENTS

This report was prepared by staff from the Information and Communication Technologies (ICT) Practice of the World Bank Group in collaboration with Lendable, and in partnership with the Global Off-Grid Lighting Association.

This report includes research and analysis funded by the Public-Private Infrastructure Advisory Facility (PPIAF), a multi-donor trust fund that provides technical assistance to facilitate private sector involvement in infrastructure, and the Trust Fund for Statistical Capacity Building (TFSCB), a multi-donor facility that aims to improve the capacity of developing countries to produce and use statistics for effective decision-making for development.

Data Playbook
for the Off-Grid Pay-As-You-Go Sector

V. 1.0

January 2018

Project Managers:

Anna Lerner, alerner@worldbank.org

Laura Sundblad, l.sundblad@gogla.org

Team Members:

Juan Andres Turner, jaturner@worldbank.org

Micah Melynk, mmelnyk@worldbank.org

Maite Lasa, maitelasagarcia@worldbank.org

Kian Behdad, kianbehdad@gwmail.gwu.edu

Partner Team Members:

Joe Brew, joe@lendable.io

Victoria Arch, victoriaa@angazadesign.com

Jennifer Sharma, jennifer@angazadesign.com

This document is part of a sector-wide initiative to enhance data use and harmonization of metrics for financial performance measurement for off-grid energy companies. For more information, please go to goo.gl/vxP25n and goo.gl/wp2oq9

This document is updated periodically.
Please check that you have the latest version from the websites above!

TABLE OF CONTENTS

Click
the colored buttons to navigate to the desired section of the toolkit

Click this button to return to table of contents

INTRODUCTION

IMPROVING THE STATE OF DATA

Section Overview: The decision regarding the data to be collected could be a challenging one. Theoretically, companies can collect and store an infinite amount of data. The variability could be vertical (the number of different events that can be logged) as well as horizontal (the number of dimensions that can be collected for each event). However, collecting every single data point that is possible could be costly, and would occupy a large part of your team - or facetime with your customers, something that might not always pay off. This raises the obvious question: what data should be collected and what data should not be collected.

I.1. What Kind of Data Can/Should You Collect?

- The principle here is very simple: collect everything you can. Every household size, median age, household income, education, latitude, longitude, and physical size of the residence, and more importantly, the power generation and power consumption trends associated with different households, hours of the day, and months of the year. Additionally, the payment habits of different households with various livelihoods in different locations will be of tremendous use to the analysts in the future. However, there are cost and time constraints discussed below that could help you prioritize.

- A lot of companies use less than 10% of their data; 90% is not even touched by analysts. Why? It is hard to know in advance when you can make use of that data in the future. As an example, what if you want to change a feature on your product that is three years old? To ensure success, you will spend time at the onset of this change process to understand the exact role of that three-year-old feature. To do this, you will need to analyze your data retrospectively. But you can only do so if you started to collect the data three years ago. This is the rationale for collecting all possible data.

I.2. What Kind of Data Should You Not Collect?

Apart from common sense in the process of designing the data collection process, there are specific limitations associated with the type and amount of data that can be collected.

- Price: Historically, the price of storing data was a problem for collecting large data sets. Today, data storage costs (in the cloud, at least) have dropped considerably.
- Developer time: The developers need to spend time to implement the tracking scripts. In order to handle complex data warehouses, companies need full-time professionals to build and maintain the data infrastructure. A rule of thumb: if the company's developers spend more time collecting data than developing the actual production process, then too much data might be collected.
- Legal questions: Consulting with local legal professionals is essential when collecting data in different countries. Some countries have strict legal restrictions about data collection and localization of stored data.

BACKGROUND

Section Overview: This section provides a brief introduction to the rapid growth of digital technologies in the off-grid Pay-As-You-Go (PAYG) sector, and the opportunities and challenges that have emerged.

Potential of Data in the PAYG Sector

Since digital technologies and sensors are an intrinsic part of PAYG business models, the industry naturally lends itself to the generation and use of large, heterogeneous, and fast-moving streams of information – otherwise known as "big data." Companies can utilize this data to improve key activities such as customer targeting, credit assessment, collections, and market research. However, many PAYG companies are not capturing the full value of the data for various reasons, including the lack of prioritization, lack of funding, and/or lack of capacity.

This Data Playbook is written to build capacity around data collection, management, and analytics in the industry at large, with an aim to facilitate data-driven decisions for PAYG companies and increase transparency of the sector. It is structured to (1) provide an introduction to data science by providing use cases for data analytics and (2) generate understanding of and capacity to make data-driven business decisions in the PAYG sector.

This Data Playbook seeks to illustrate best practices for how PAYG companies can develop and manage appropriate data infrastructures, including illustrating how the data infrastructure can be used to report on Key Performance Indicators (KPIs). It also provides an overview of the most common data collection systems, data infrastructure, and data analytics use cases in order to identify how companies may strengthen their overall data management capacities and use data for better operational insights. The work is a collaboration between GOGLA, IFC/Lighting Global, and Lendable, with input from other sector stakeholders such as Driven Data, Village Power, and several other GOGLA Members.

THE BUSINESS CASE FOR DATA ANALYTICS

Section Overview: According to Gallup's analysis, companies that apply the principles of behavioral economics outperform their peers by 85% in sales growth and more than 25% in gross margin[1]. This shows how the ability to extract information from what customers do and what they want can be strategic to the survival of companies.

A McKinsey survey of more than 700 organizations worldwide found that spending on analytics to gain competitive intelligence on future market conditions (to target customers more successfully and optimize operations and supply chains) generated an increase in operating-profits in the 6 percent range[2].

This section provides an overview of the business case for data analytics, including how it could strengthen a company's lending portfolio and financial decisions, and how PAYG companies could apply such techniques. **Specific reference will also be made to the KPI framework.**

2.1 Portfolio Health

2.1.1. Analyzing Customer Behavior and Risk Management:

Why? Light-touch analytics could offer valuable insights into customer behavior, specifically identifying common attributes of well-performing customers and increasing targeting and success of marketing materials. Additionally, questions regarding what type of products to offer each household (solar lanterns, basic home solar systems, or more advanced ones) could be better answered by analyzing the behavior of previous customers.

[1] http://www.gallup.com/services/170954/behavioral-economics.aspx

[2] https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/capturing-value-from-your-customer-data?cid=other-eml-ttn-mip-mck-oth-1801

How? A smart company can generate predictions regarding the behavior of its potential customers by correctly studying and analyzing the behavior of its existing customers. For example, studying the common attributes of a subset of existing customers with on-time payments can help the company improve targeting of new prospective customers and better inform outreach through measures such as improved design of online ads, better location of physical billboards, and more targeted direct mail marketing.

Cost/Difficulty? In its most simple form, behavioral analysis of previous customers can be carried out with only a small investment in data storage, data collection, and data analysis infrastructure. Collection can be manual or automated; storage can range from simple and small (spreadsheets) to large and complex (integrated CRMs, database schemas); analysis can be integrated into the storage and collection software in the case of spreadsheets, or be carried out in a specialty analytics software. The choice of statistical package will be further introduced in a separate section dedicated to the most widely used analytics packages, their costs, and their distinguishing features and applications. When analyzing customer behavior, key variables to collect pertain to actions, circumstance, and attitude. Actions refer to previously observed credit behavior, such as defaults or late payments; circumstance can range from household attributes, such as size, education, and income, to employment status, season, or family situation; attitude means those behavioral preferences elicited during a screening process with a credit-officer (i.e., preferred loan conditions and repayment schedule, aversion or affinity to insurance). It is worth noting that in case the capacity to perform the analytics does not exist inside the company, third party options exist which range in terms of price and complexity from the basic (a single consultant performing straightforward modeling) to the advanced (analytics firms with integrated risk profiling models, dashboards, and warning systems).

Value? The expected value of credit analytics can be approximated through retrospective counterfactual comparison exercises in which a modeler "trains" a credit scoring model on historical data and then "tests" on more recent data. By attaching costs to events such as defaults, rejected credit applications, and repayment (as well as the cost of the implementation of the analytics program itself), and by making explicit those metrics which are of most concern (such as "Portfolio at Risk"), a firm can approximate the value of implementing credit analytics.

2.2. Financial Decisions

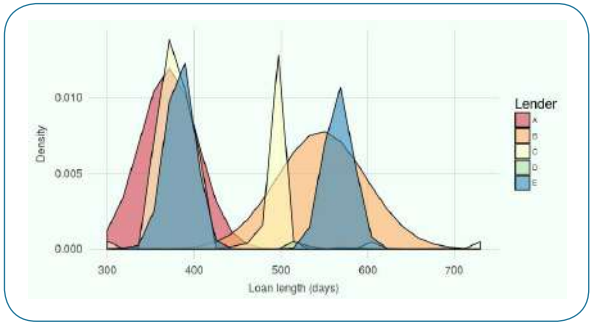
2.2.1. Loan Terms

Why? PAYG solar companies have found a way to lend to medium-to-low-income households in developing markets. "With an average daily cost of around $0.40, they can meet the household energy needs of many people — but not everyone. For a household earning $2 per day (the typical wage rate for a day of casual agricultural labor in East Africa), paying for the average PAYG product would mean allocating 20 percent of income to energy, far beyond the 10 percent threshold set for 'energy poverty.' For poor families, even 10 percent of income going to energy can be a strain because these families have so many unmet needs to cover. In addition, many of these customers encounter seasonal fluctuations in their incomes. Many companies are seeking to make their products more affordable for more people and offering payment plans that better reflect a customer's payment habits."[3]

In an effort to tailor the terms of the lending process to the needs and financial means of potential customers, companies must decide on what loan terms to offer each customer. These include:

Duration of the Loans

The most common way to improve affordability is to lengthen the tenor of the loan. The obvious upside of increasing the duration of the repayment is that it makes the daily energy payment feel more affordable to the users and therefore more attractive to a larger portion of the population. On the other hand, since it takes longer to repay the loan, the number of years the end users will own a paid-off device will be lower. In addition, since companies are more exposed to default risks in this scenario, they will have to charge higher interest which is less than desirable for low-income households.

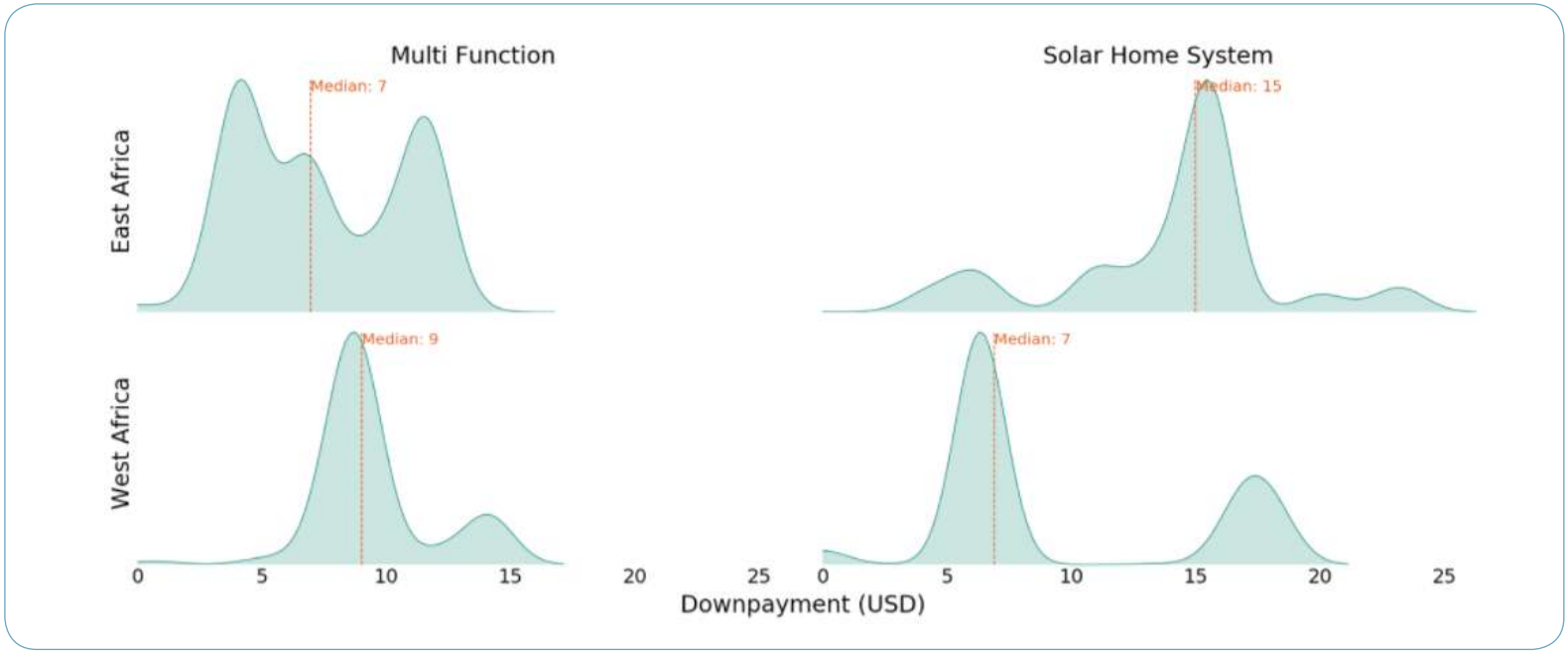


[Figure: Distribution of loan terms among 5 East African PAYG solar lenders. All provide loans of variable length, many clustered around 12, 18 and 24 months.]

[1] http://www.cgap.org/blog/how-can-pay-you-go-solar-work-poor-people

Deposit Size

PAYG companies typically use a deposit as both a sales strategy and an initial credit assessment. In theory, households who have already paid a portion of the costs will have an incentive to continue to pay until they fully own the device. For instance, in East Africa, companies typically request around 10-15% deposit for solar home systems, with the median down payment being $15. In addition, as many PAYG customers have no credit histories, companies use deposits to separate good-paying clients from those who will have a higher chance of default later on. However, using deposits as credit screening mechanisms is more complicated. In fact, from our own research with Angaza, the size of the down payment does not have a correlation to loans being paid off on time. So, while the deposit may indicate a customer's ability to pay at a point in time, it may not be a strong indicator of that future loan's performance and how that customer will pay over time. Therefore, the key is to strike a balance between all of these factors and decide what is the ideal amount of deposit in any particular geographic location or even for each individual household based on historical data and economic theory.



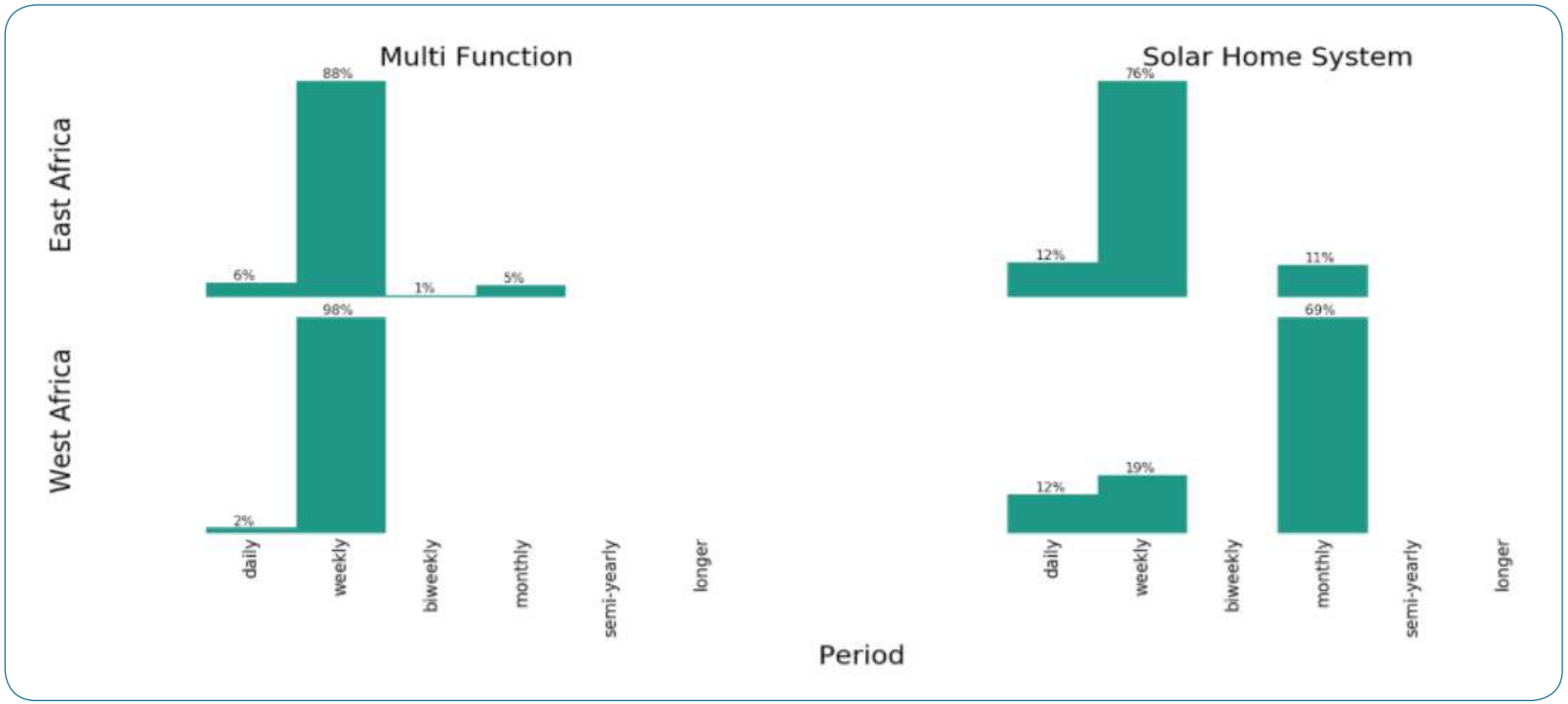
[Figure: Distribution of down payments for both MF and SHS in East and West Africa. Source: Trends Analysis for the PAYG Solar Industry, Angaza, IFC, GOGLA, and World Bank]

Flexibility of Loan Terms

The flexibility of the loan payment is clearly one of most attractive attributes of PAYG solar. It allows customers to adjust their payments without fearing repossession according to their cyclical and, at times, unpredictable income streams, which is very common for farmers and cattle owners. Companies structure payment terms differently but there seems to be a strong preference for asking customers for weekly payments. In our research over 75% of companies ask for weekly payments. However, there may be certain loan terms that may lead to better payment outcomes, or loan performance, than others. For instance, our research shows that as the time between each payment decreases, the likelihood that the account will be paid off in time increases. This may indicate why companies tend to prefer to establish weekly payment terms with customers. On the other hand, because PAYG companies in turn need to finance their operations through external funds, they might have to face higher rates since the money is being channeled to the actual lenders erratically.

How? This can be done in the context of splitting potential customers into "will buy"/"will not buy", or "will pay back"/"will default" or more appropriately "high-risk/low-risk". The result could, for example, be a classification of potential customers as high-risk (those with a probability of default higher than, say, 30%) or low-risk (those with a probability of default lower than, say, 30%). It is worth noting that the classification concept, along with any of the techniques mentioned, is not limited to a binary choice and can easily be extended to cases where there are more than two classes. As an example, a company might decide to classify their potential customers into three groups of high-risk, medium-risk, and low-risk.

Cost/Difficulty? Having a probabilistic idea about how a potential customer will react to a new/existing product and whether/when a client will pay back a loan, can be achieved through the use of regression analysis, classification trees, or clustering algorithms. These are all routine predictive algorithms performed by almost all of the popular packages introduced in this playbook.



[Figure: Breakdown of Typical Contract Payment Periods for PAYG companies in East and West Africa Source: Trends Analysis for the PAYG Solar Industry, Angaza, IFC, GOGLA and World Bank]

Value? Overly-strict policies can have the detrimental effect of forcing many households to default, only after missing a payment by a few days. The costs associated with repossession of the device coupled with the opportunity cost of the lost client could be huge, not to mention the bad optics and negative publicity of the action. As a result, in many cases, it is financially more viable to allow for a small cost associated with one or few late payments than to accept the much greater costs of a default. Some PAYG companies are even taking the extra step to monitor farming conditions in rural areas (temperature, sunlight, rain, and storms) to predict farming and harvesting complications and offer preemptive payment flexibility to their clients.

A side note on foreign exchange:

Many PAYG companies rely on foreign investment during their growth phase. It is important to keep in mind that, depending on local exchange rates, extending loan tenor may not always be the most profitable strategy, particularly in contexts with high inflation. The below chart shows a basket of African currencies against the US dollar during the three year period from mid 2013 through mid 2017. The average depreciation was approximately 20%, meaning that late term repayment may not always be in a lender's best interest, even if allowing it reduces the nominal default rate.

2.2.2. Fraud Detection

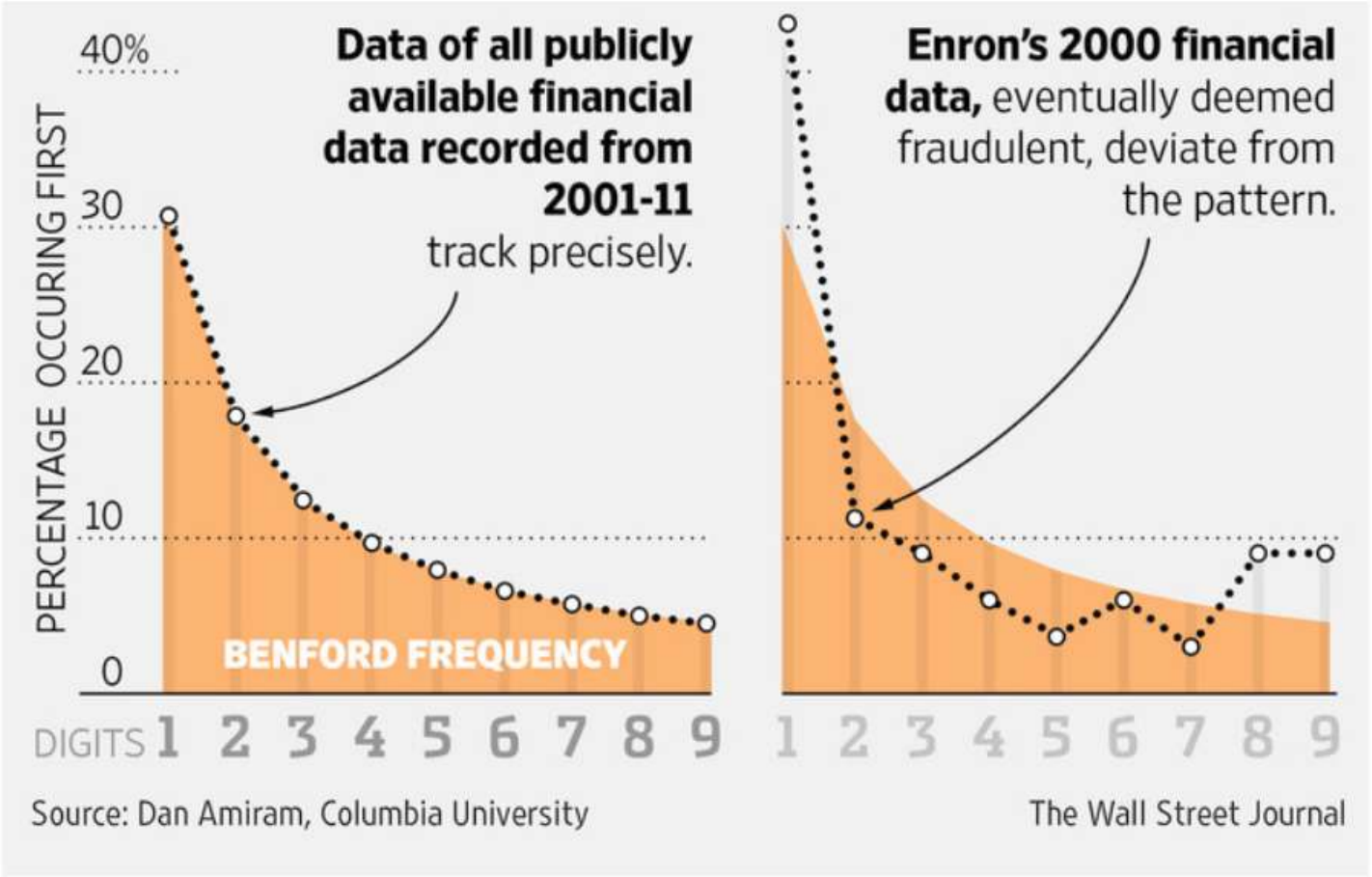
Why? Whenever there is human involvement in information gathering or reporting processes, there is a possibility for intentional or unintentional misrepresentation. When the person in charge of reporting data can personally benefit from misrepresenting the numbers (such as amount of cash received from customers, number of units sold to customers, percentage of broken units, number of working hours per day), there is even a higher chance of fraud.

How? There is extensive literature concerning the techniques that raise red flags at the prospect of a high probability of fraud in a specific set of data, especially financial reports. The techniques make use of the mathematical relationship between numbers (e.g., Benford's Law), or merely test for a logical trend across time at a specific location, or a trend across different locations or people at a specific point in time. Using some probabilistic models and advanced analytics, fraud detection algorithms raise red flags when they conceive of a high probability of fraud, which can then be further investigated by auditors or other decision makers.

Cost/Difficulty? Fraud detection can be done by writing simple algorithms designed to detect irregularities based on probabilities that mostly require mathematical knowledge rather than programming skills (done in Excel or R). More professionally, specialized fraud detection software can be used, such as Similarity Fraud Detection, Tipalti, Membercheck, SAS Anti-Money Laundering, etc.[4]

Value? Like all other companies, PAYG companies receive numerous periodic reports from different sources, such as their subsidiaries, field agents, and contractors. These reports can be financial (such as sales, expenses, accounts receivable, accounts payable), or operational (such as number of households visited per day, the distance travelled by field agents per day, the number of equipments fixed by the maintenance team). These are a handful of scenarios where use of fairly elementary analytics tools can raise flags as to the potential for fraudulent activities.

Illustration: The following is an illustration of applying Benford's Law to detect doctored numbers in financial reports. According to Benford's Law, approximately 30.1% of numbers in a list of financial transactions begin with "1". Each successive digit should represent a progressively smaller proportion. Below, orange indicates the expected Benford frequencies. When digits stray from the pattern, fraud may be to blame. This is clearly the case for Enron, where one of the biggest financial frauds in history happened, versus all publicly available financial data, which tracks the frequencies precisely.



2.2.3. Markdown Strategy

Why? Sometimes a company has huge inventories of products that near the end of their technological lifetime. This means that, although the products work perfectly, the next generation of that product is in development High holding costs coupled with low value of unsold units makes it attractive for the company to know the appeal of the aging product as it relates to category of customer and mark-down strategy. Given the quick pace of technology development in the off-grid solar sector, having a clear technology markdown strategy could be a smart way to manage the technology risk.

How? Innovative pricing and revenue management techniques exist that measure the economic elasticity of products and the interrelationship between sales of competing or complementary products to predict the buying behavior of different classes of customers with respect to different portfolios of products and their prices. Then, optimization tools can be deployed to identify the best markdown strategy to maximize net profit, which is the difference between sales (dollar amount, not number of items sold) and holding costs, plus any salvage value from unsold products.

Cost/Difficulty? All of the calculations and techniques mentioned can be done with the most basic tools, such as Excel, and its free add-on tools, such as Solver (for the optimization step). Larger problems, consisting of hundreds or thousands of products and scenarios could become difficult to handle with Excel. In these cases, more advanced mathematical programming tools (such as AMPL) can be utilized, which requires larger investment and demonstratively more skilled analysts.

Value? The ability to drive revenue by optimally changing prices without using additional resources or renegotiating holding costs or costs of goods sold is very valuable for companies that have to deal with rapidly evolving technologies and short horizons for product relevance. This technique helps companies make the most out of their current products, while minimizing the number of unsold products given away for their salvage value.

2.2.4. Selling Data

Why? There is always the possibility of information on past clients that cannot necessarily be turned into additional profit for the company, but that could be of great value to other companies up or down the chain. For example, a cellular phone company, an ISP, a distributor of electrical heaters, electric ovens, or many other electronic appliance could be interested in the data that the company has collected over the years.

How? Through the consistent collection of relevant data and utilizing data warehousing software, companies can ensure the availability of data in the future, even if it is not clear at the moment of collection how that might some day turn into tangible profit.

Cost/Difficulty? Since no additional technique or step is required, the cost of collecting more data is virtually zero, with the exception of the cost of storing huge amounts of data, which is fortunately becoming less expensive by the day.

Value? The data can be either sold or leveraged to receive other useful data from other companies in the future reciprocation versus payment, which means lower costs.

2.3. Strategic and Operational Decisions

2.3.1. What Regions to Target

Why? When making strategic decisions on which regions to target next, it is critical to identify the regions with highest potential for profit.

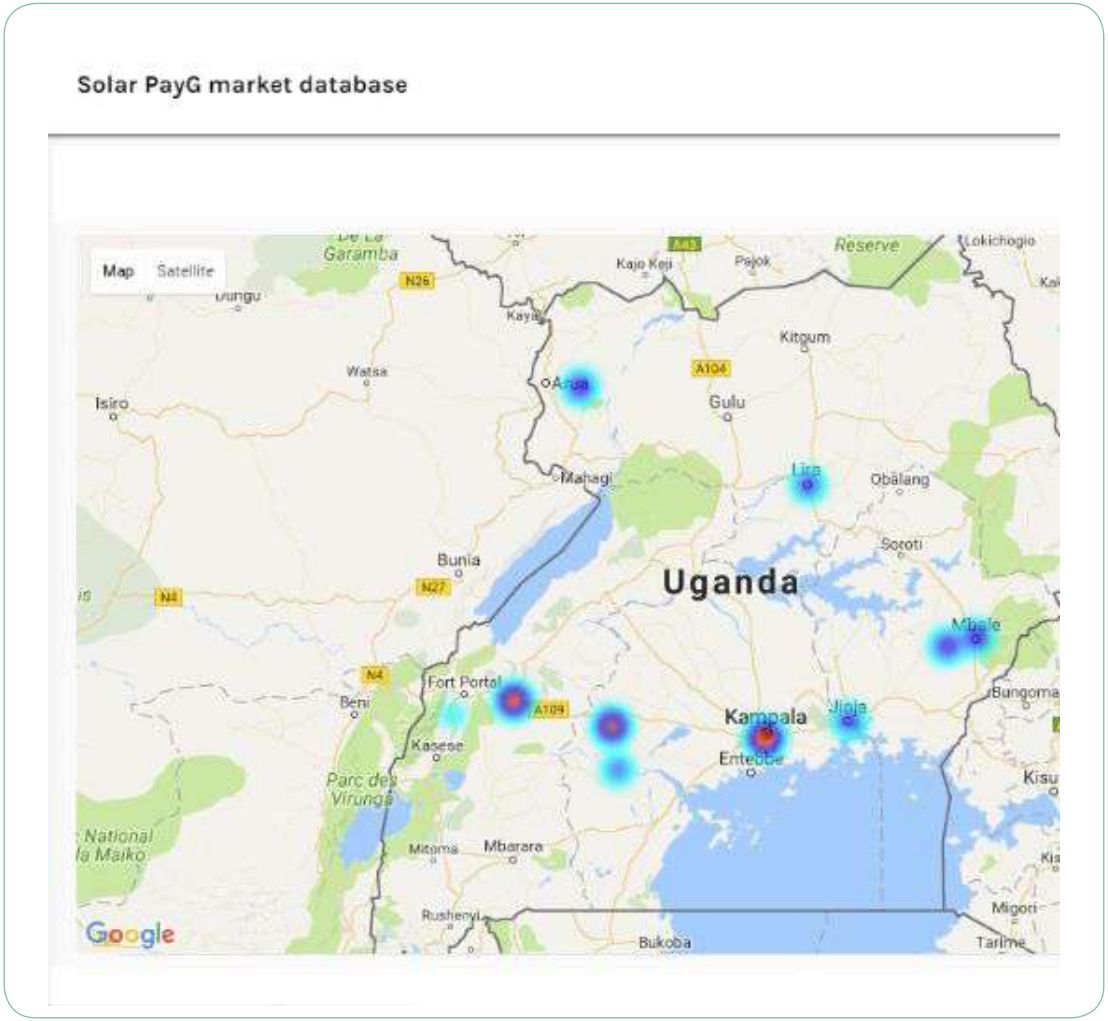
How? The profit can come from different sources: higher unit sales, lower transportation and maintenance costs, distance from traditional power plants, distance from power grids, solar energy intake throughout the year, other renewable energy alternatives (an area could be better suitable for wind power generators rather than solar), number of other competitors active (to identify already congested areas), and so on. Data science techniques along with advanced optimization methods can be deployed to identify the regions with highest potential for profit.

Cost/Difficulty? These decisions are typically made by observing uncongested areas on a population heatmap superimposed onto a rough map of existing power plants and energy grids. In order to maximize profits, these simplistic techniques must be replaced by more analytical methods involving mathematical formulation of many more of the mentioned factors along with any restrictions faced by the company (such as geographical limitations, number of employees, budget, etc.) in order to find the optimal region or regions to enter. Depending on the number of variables and number of constraints faced by the company, the problem can be solved easily or could require a great deal of time, effort, and investment.

Value? The value stems from the difference between trial and error using human judgments and the guaranteed best option possible among many.

Illustration

Village Power Heat Map Visualisation: Sample location of PAYG units



2.3.2. How to Assign Field Agents or Maintenance Crew to Different Locations

Why? To strike a balance between offering good service to customers and saving costs associated with hiring more field agents, the company needs to maximize the utility of its agents and minimize time wasted by standing idle or spending too much time on the road commuting between different sites and customers.

How? By keeping track of number of units sold, their locations, usages, probability of failure of a unit, and probability of late payments based on historical data, a company can plan for the optimized number of agents to have on-call in any given region. After receiving site visit requests from customers, the company can dispatch its agents in an optimal fashion to maximize visits while minimizing commute time with the minimal number of agents on call.

Cost/Difficulty? Depending on the number of customers, agents, and distances, this could be a very difficult and expensive problem to solve. In fact, typically only multi-billion dollar companies such as Uber, specializing in transportation while dealing with a huge number of random variables (traffic, driving habits, hourly demand), have the knowledge and computing power to solve this family of problems with high accuracy. For smaller companies with core business models other than transportation, it is usually more financially viable to follow simpler rules. By using dashboards that keep track of visit requests and the location of their agents, small companies can decide where each agent should go next in order to avoid having too many agents in the same locality.

Value? Although not optimal, this is still more efficient than not keeping track of the said information and dispatching random agents as soon as a request is made for a site visit.

2.3.3. Delivery Logistics

Why? Giant logistics companies such as UPS, FedEx, or DHL have extensively been using data science techniques coupled with network optimization techniques to improve operational efficiency. Discovering the best routes to ship for different combinations of products has been essential in boosting operational efficiency, and therefore saving costs.

How? Collected and stored data on number and location of agents, potential customers, warehouses, and the costs and distances associated with different routes using different vehicles (bikes, motorcycles, trucks), along with the capacities of each of these options, will later be used as inputs to optimization packages to produce the optimal way to design the delivery logistics of the business.

Cost/Difficulty? For relatively small problems with a couple of agents, warehouses, routes, and means of transportation, a simple Excel sheet can be used to feed the free Excel Solver add-in to calculate the optimal design. More complicated problems with hundreds or thousands of items will require the use of more specialized mathematical programming tools (such as AMPL) that can handle large problems more efficiently.

Value? Despite the absence of great variety in products and services, PAYG companies can still benefit from logistics efficiency through finding the best locations for renting warehouses, the minimum number of salespeople and technicians required to fulfill a certain level of customer satisfaction, the optimal number of customer service personnel to manage a call center, and best routes for sending technicians to deliver, install, or fix a unit from designated headquarters.

2.3.4. Market Basket Analysis

Why? For most grocery stores and ecommerce websites, it would be extremely valuable to know what combinations of products tend to be bought at the same time. This way, the company can recommend a list of products for its customers based on what they have already bought or are about to buy. Another familiar example of this technique is the movie or TV-show recommendation algorithm used by Netflix, Amazon Prime, and Hulu.

How? One of the ways to find this out is to use an algorithm called association rules or market basket analysis. The algorithm compares the standalone probability of any given customer buying a particular item with the probability of buying the same item given that the customer has already bought different items or a family of items. For example, if the standalone probability of a random customer buying milk is 10%, but of those who buy cereals, 80% also buy milk, the strong lift in the associated probability (80%/10% = 8) might be a reason to either bundle the two products together, put the two products close to each other at the grocery store, or recommend to a customer who has bought cereals to also consider buying milk, in case they have forgotten to do so.

Cost/Difficulty? Most of the popular analytics tools in the market can handle association problems with relatively little effort or cost. The difficulty might though arise when collecting the relevant information and storing it in the correct format that can then be fed to the software.

Value? This technique could be deployed depending on the variability of products offered to PAYG customers (solar panels, electric bulbs, batteries, internet modems, and other electronic appliances). As a rule of thumb, the technique becomes more relevant as more products are added to the basket. This means that there is a possibility of synergy in combining the collected data by PAYG companies with the data collected by other companies in the region through sharing/selling the relevant data. For instance, sharing information about the probability of a cell phone purchase along with the number of cell phones required and the creditworthiness of the household to a cell phone distributor could be of great value to PAYG companies.

DATA MANAGEMENT PLATFORMS

Section Overview: This section provides a brief introduction to the types and capabilities of data storage software, followed by a brief introduction to the data analysis tools most often used in the industry. The section will contain examples of both free and paid packages, along with their applications.

3.1 Types of Software for Collecting/Storing Data

- Microsoft SQL Server Reporting and Analysis Services
- Oracle Database DBMS
- Teradata Data Warehouse
- Cloudera Distribution Hadoop
- Amazon Redshift
- SAP BW, SAP Hana
- IBM Smart Analytics System, PureData System for Analytics (Netezza)

3.1.2. Required Capabilities

A data warehouse maintains a copy of information from the source transaction systems. This architectural complexity provides the opportunity to:

- Integrate data from multiple sources into a single database and data model.
- Integrate data from multiple source systems, enabling a central view across the enterprise. This benefit is always valuable, but particularly so when the organization has grown by merger.
- Improve data quality by providing consistent codes and descriptions, flagging, or even fixing bad data.
- Present the organization's information consistently.
- Provide a single common data model for all data of interest regardless of the data's source.

[5] https://www.gartner.com/reviews/market/data-warehouse-solutions

- Restructure the data so that it makes sense to the business users.
- Restructure the data so that it delivers excellent query performance, even for complex analytic queries, without impacting the operational systems.
- Add value to operational business applications, notably customer relationship management (CRM) systems.
- Make decision–support queries easier to write.
- Optimized data warehouse architectures allow data scientists to organize and disambiguate repetitive data.

3.2 Analytics Tools and Packages

3.2.1. FREE EXAMPLES

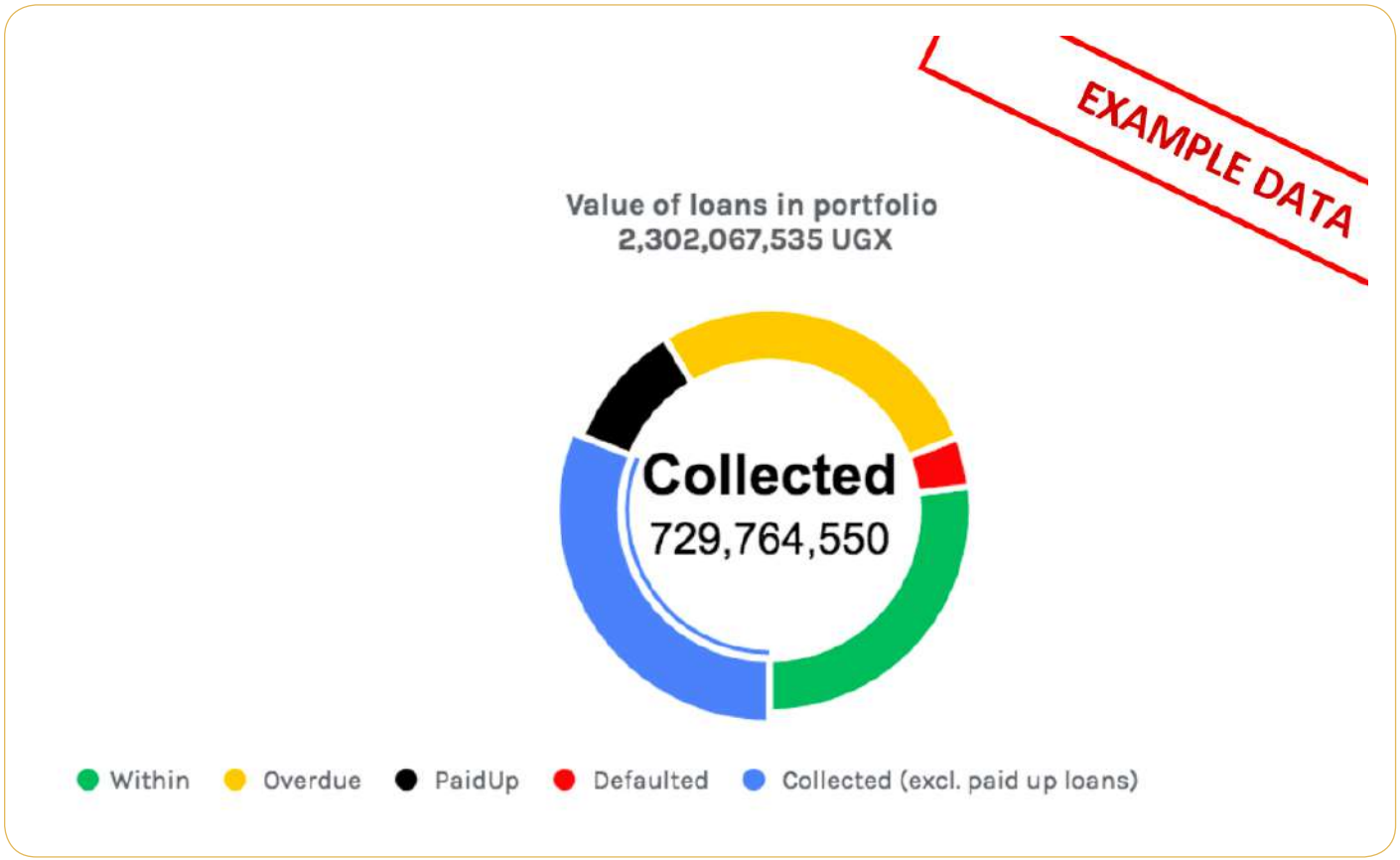
Village Power Data Infrastructure Tools

Village Power Data Infrastructure Tools, developed with the support of the UNCDF CleanStart Programme, are three open-source software applications that allow PAYG companies to manage key aspects of their business and share company data in a secure, anonymized way with investors and key partners. The tool is comprised of three main areas including a (i) Loan Portal, (ii) Market Database, and (iii) Data Collection Tool.

The Loan Portal (illustrated below) provides companies with the ability to share the performance of an anonymized subset of accounts from a company's lending portfolio with investors or other trusted partners. The company administers their PAYG Loan Portal and controls access to the portal dashboard independently. It provides companies and investors with a (i) current snapshot of each batch of loans (number of loans, status, value of loans), (ii) portfolio development over time (number of loans, collections, overdue break out), and (iii) loan repayment curve (view of average customer payment behaviours in each batch versus expected payment behavior).

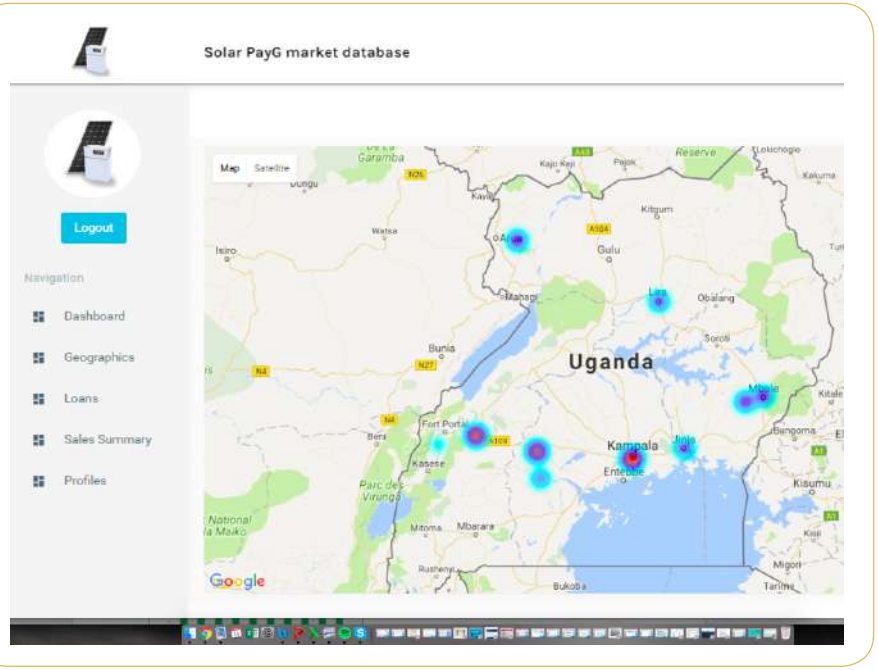
[5] https://www.gartner.com/reviews/market/data-warehouse-solutions

Loan Portal Screenshot



The Market Database allows companies to pool anonymized PAYG data to create a market view of key metrics. The tool can provide companies and investors with (i) geographical summaries and heatmap visualisation of key metrics such as the number of units, installed capacity and loan repayment statistics.

The Data Collection Tool facilitates the collection of data based on harmonized definitions and market KPIs at key customer touchpoints along the solar PAYG customer journey. Based on the Open Data Kit (ODK), it provides a starting point for solar PAYG operators for the customer data collection at relevant customer touchpoints, for example, at point of sale/contract, kit collection, installation, troubleshooting, and repossession.



R[6] is an open source programming language and software environment for statistical computing and graphics that is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years. R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows, and MacOS.

[6] https://www.r-project.org/about.html

Illustration

The following is a screenshot of a regression analysis done in RStudio, which is a free and open-source integrated development environment (IDE) for R. The code is written on the top-left pane, run in the lower left pane, and the resulting plots are shown in the lower right pane. In this illustration, a regression is performed to explore the relationship between a movie's length and its expected gross sales.

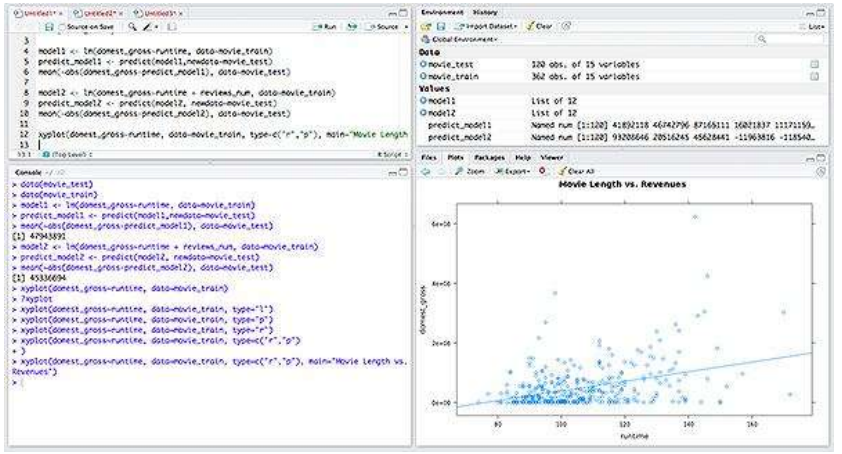


Figure: RStudio Screenshot[7]

Python is a widely used high-level programming language for general-purpose programming, which since 2003, has consistently ranked in the top ten most popular programming languages. As of March 2017, it is the fifth most popular language[8]. Like R, Python is open source and free. Though R is more common among statisticians and data analysts, Python's more general orientation makes it a favorite among data scientists who want to integrate analysis and modeling into a larger framework, such as an API or web service.

Illustration

The following is a screenshot of a Bayesian ridge regression analysis done in Spyder, which is an open source IDE for scientific programming in the Python language. Very similar to R, there is a pane for the codes, a pane for exploring variables, and another pane to illustrate the graphs and visualizations.

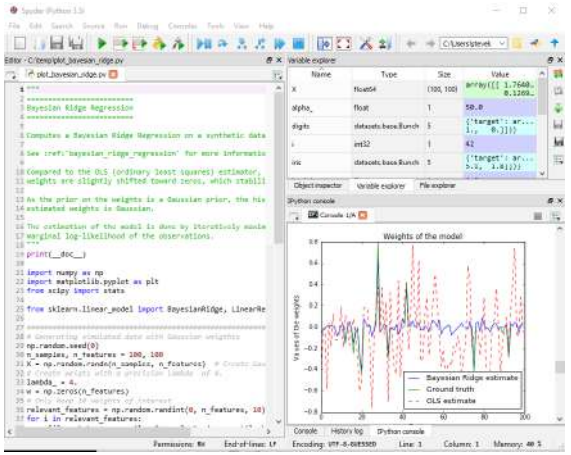


Figure: Spyder Screenshot[9]

[7] https://www.kcet.org/history-society/introduction-to-data-science-an-authentic-approach-to-21st-century-learning
[8] https://mybroadband.co.za/news/software/203762-most-popular-programming-languages-in-the-world-2.html
[9] http://www.virtustate.com/scikit-learn-machine-learning-python

A typical classification output from the software looks like the following, where three different very common techniques (Linear Discriminant Analysis, Support Vector Machine, and Logistic Regression) are used and the results are compared to classify a point as either red or blue after observing a large number of red and blue points.

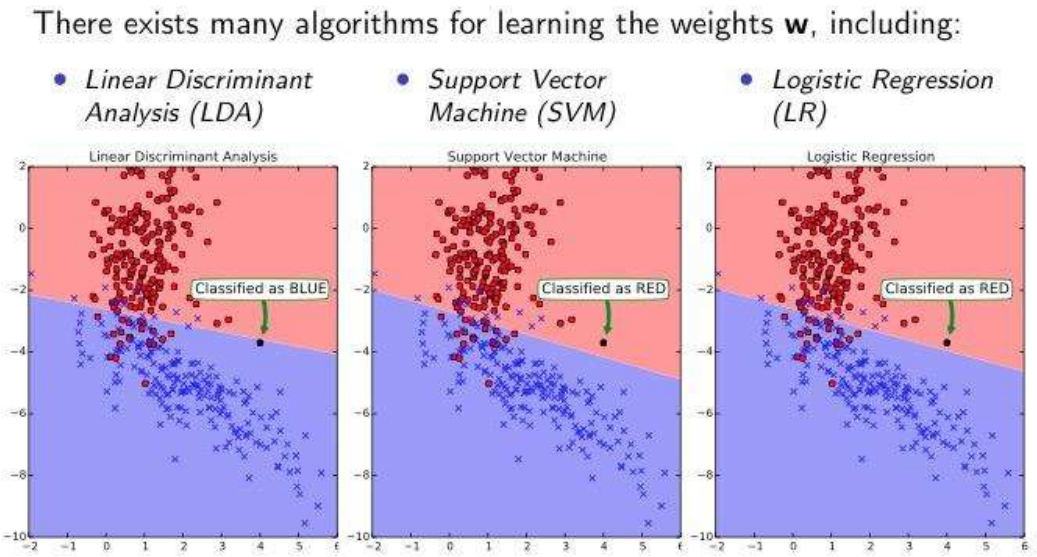


Figure: A classification example done in Python[10]

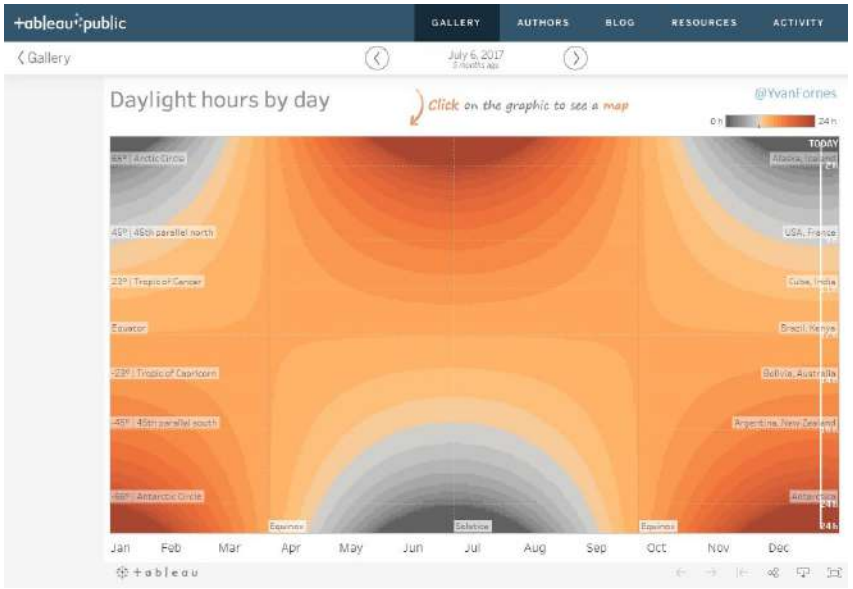
Tableau Public is a free service that lets anyone publish interactive data visualizations to the web[11]. Visualizations that have been published to Tableau Public ("vizzes") can be embedded into web-pages, blogs, and social media. Since no programming skills are required, Tableau Public is for anyone interested in understanding data and sharing those findings as data visualizations with the public. An important aspect of using the service is that a workbook published to Tableau Public will be accessible by anyone on the internet.

[10] http://www.virtustate.com/scikit-learn-machine-learning-python
[11] https://community.tableau.com/docs/DOC-9135

Illustration

The following are some examples of the visualizations available on the website and available to everyone. They cover different ideas from “An overview of the United Nations Development Group”, to “Daylight Hours by Day”, to “Word Usage in Sacred Texts”.

Figure: An overview of the United Nations Development Group, summarizing issues, regions, agencies, and support type[12]



[Figure: Visualizing daylight hours based on geographic position and time of year[13]

[12] https://public.tableau.com/en-us/s/gallery/united-nations-development-group?gallery=featured

[13] https://public.tableau.com/en-us/s/gallery/daylight-duration?gallery=featured

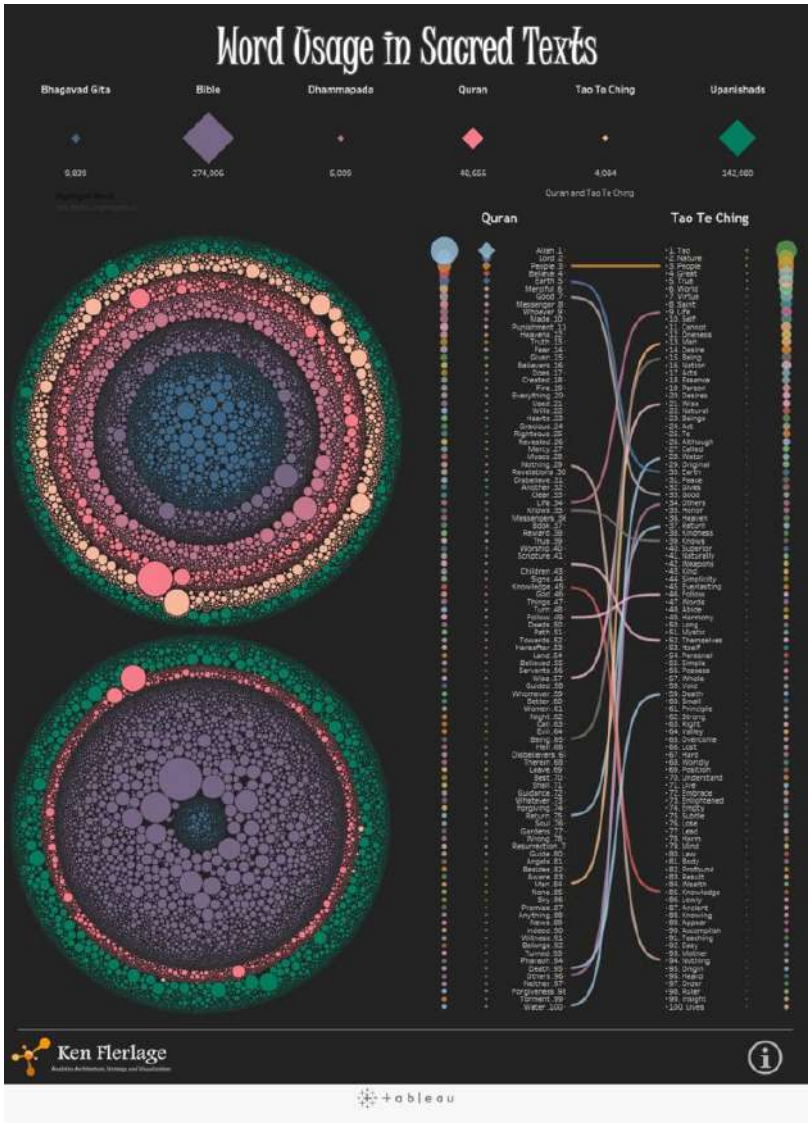


Figure: Visualizing a comparison of word usage in six sacred texts[14]

[14] https://public.tableau.com/en-us/s/gallery/word-usage-sacred-texts

3.2.2. PAID EXAMPLES

3.2.2.1 Data Visualization: Visual analytics is an outgrowth of the fields of information visualization and scientific visualization that focuses on analytical reasoning facilitated by interactive visual interfaces.

Tableau is a business intelligence tool that focuses on data visualization and dashboarding. Tableau's corporate slogan, "Answer questions as fast as you can think them[15]," highlights the company's insistence on providing users with a quick way to solve problems. Tableau is very user-friendly and intuitive, and it does not require programming skills. It also has a mapping functionality, and can plot latitude and longitude coordinates.

Illustration

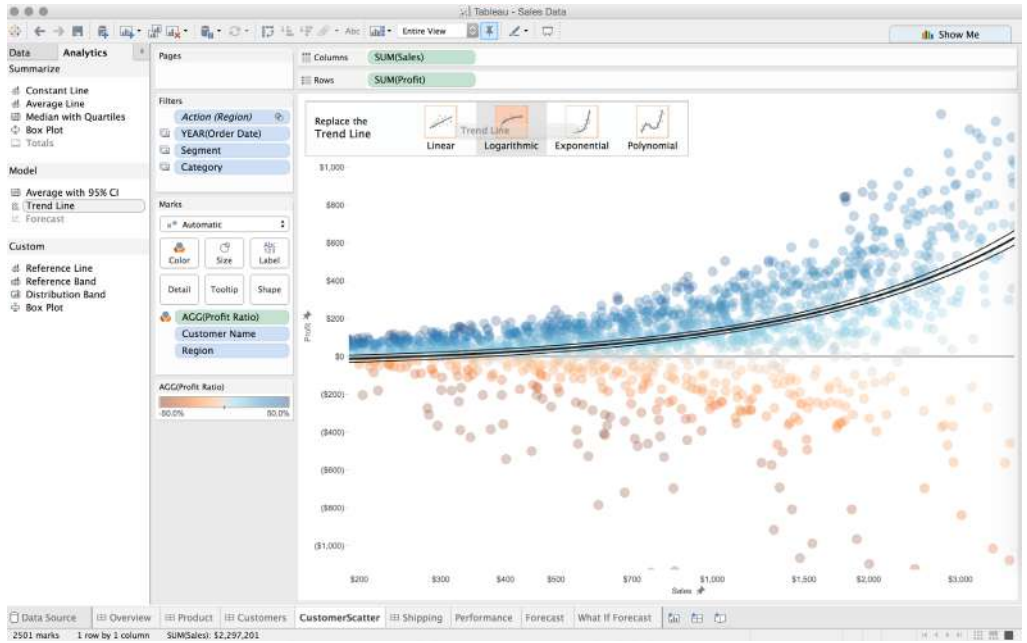
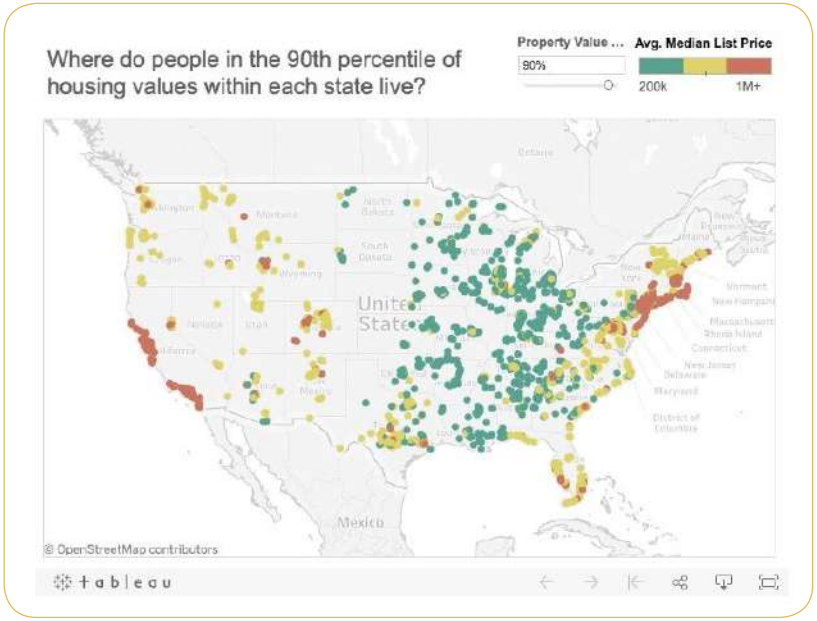


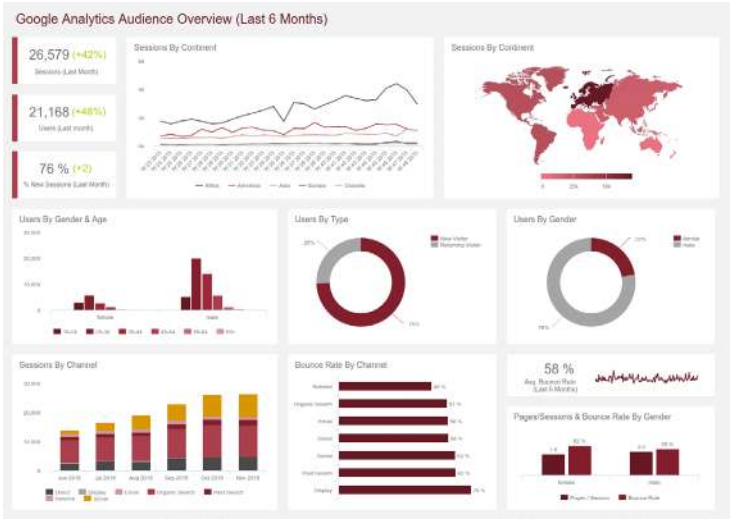
Figure: A logarithmic trendline between sales and profitability using Tableau[16]

[15] https://www.tableau.com/trial/p3group
[16] https://www.saimgs.com/imglib/other_pages/BI/1-tableau-drag-and-drop-analysis.png

Another interesting feature in Tableau is the ease with which one can create interactive dashboards and post them on their website with easy-to-use drop-down menus that add extra layers of applicability to the data visualization.



[Figure: A visualization of where the top 10 percent of richest people per state live using Tableau[17]]



[Figure: Google Analytics audience overview dashboard with Tableau[18]]

Illustration

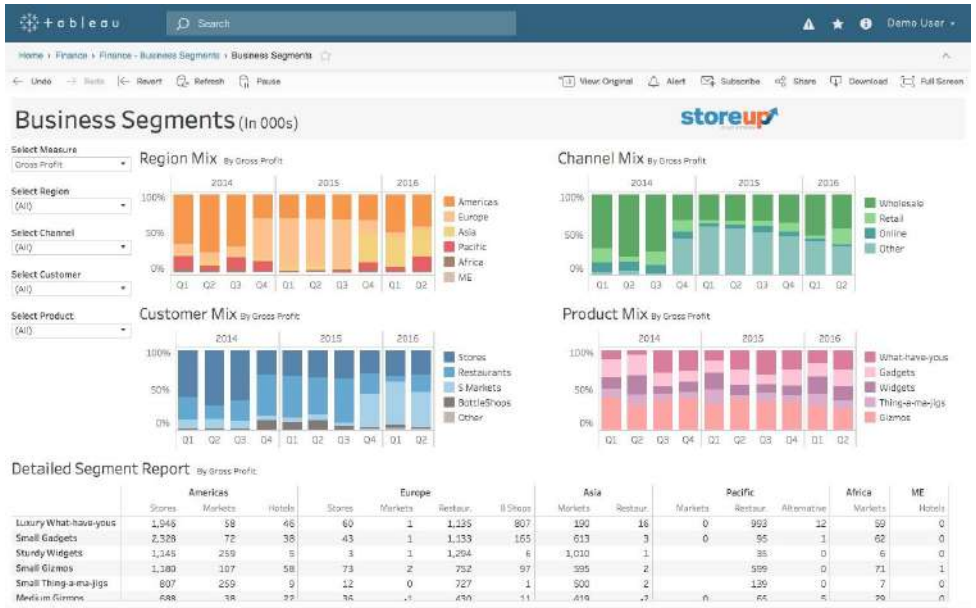
The following is a Google Analytics KPI dashboard that provides a general overview of a website's audience so that a company or individual can gain a better understanding of who actually is visiting the website. The example includes data on gender, the origin of the website's visitors per continent and the volume of sessions associated to each, the progression of visitors (returning or new), as well as more in-depth information about the site content which can include top landing pages broken down by gender (not featured).

[17] https://www.tableau.com/about/blog/2015/1/analytics-flow-36387
[18] https://www.datapine.com/blog/custom-google-analytics-dashboards-examples/

There is also a Tableau Server option for large companies that requires a secure online connection between the visualization dashboards and secure databases (maybe in the cloud) across many users inside the company. Tableau Server is also more scalable, and permissions for data sources and content are also manageable.

Illustration

The following is a visualization of the gross profit of the sales department of a company broken down by region mix, channel mix, customer mix, and product mix, along with a detailed segment report produced by connecting to the company's databases using Tableau Server.



[Figure: Visualization of gross profit with Tableau Server[19]]

As one of the analytics industry leaders, SAS's solution to the huge demand for a visualization package is the release of SAS Visual Analytics. The tool has many features, such as visual data exploration, text sentiment analysis, forecasting, scenario analysis, geographical maps, goal seeking, path seeking, decision trees, interactive reports, and dashboards.

[19] https://10az.online.tableau.com/#/site/demodepot/workbooks/1111799/views

Illustration

The following is a screenshot of a visualization performed on customer retail data from an LA county, demonstrated on a map and broken up by the products purchased.

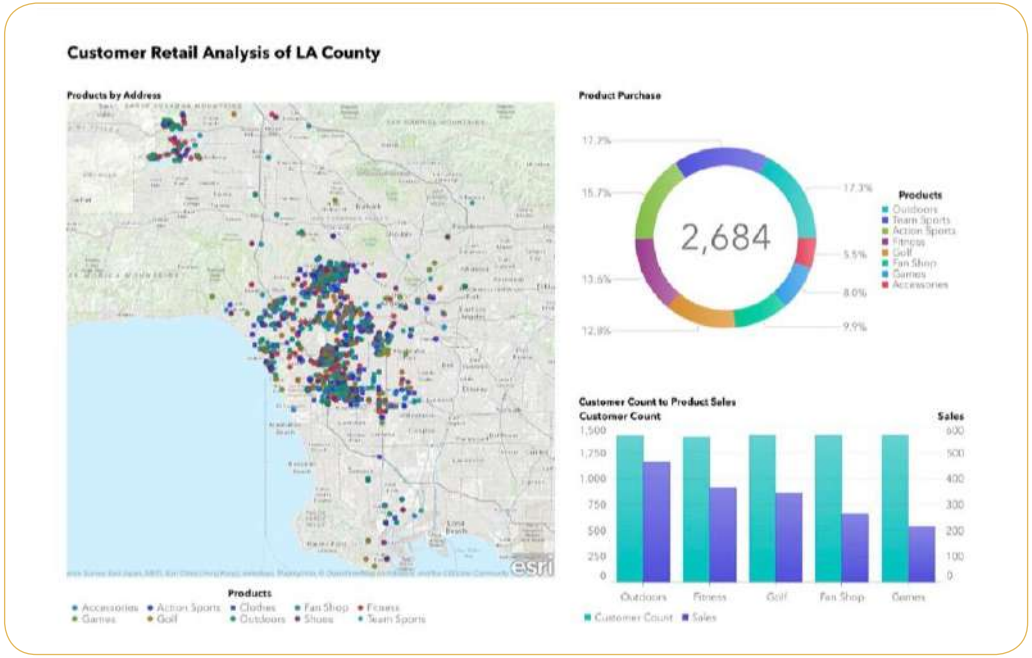


Figure: Visualization of customer retail analysis of LA county by SAS Visual Analytics[20]

3.2.2.2 Data Mining: These packages help execute actual learning, or exploitation of information from large data sets using methods that involve regression analysis, neural networks, cluster analysis, decision trees, and many more techniques.

STATA is a general-purpose statistical software package. Most of its users work in research, especially in the fields of economics, sociology, political science, biomedicine, and epidemiology. STATA's capabilities include data management, statistical analysis, graphics, simulations, regression, and custom programming. It also has a system to disseminate user-written programs that lets it grow continuously.

[20] https://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-visual-analytics-on-sas-viya-108779.pdf

Illustration

The following is a screenshot the STATA interface, followed by an example of a kernel density function plot of January temperatures in cities broken up by region - North Central, Northeast, South, and West.

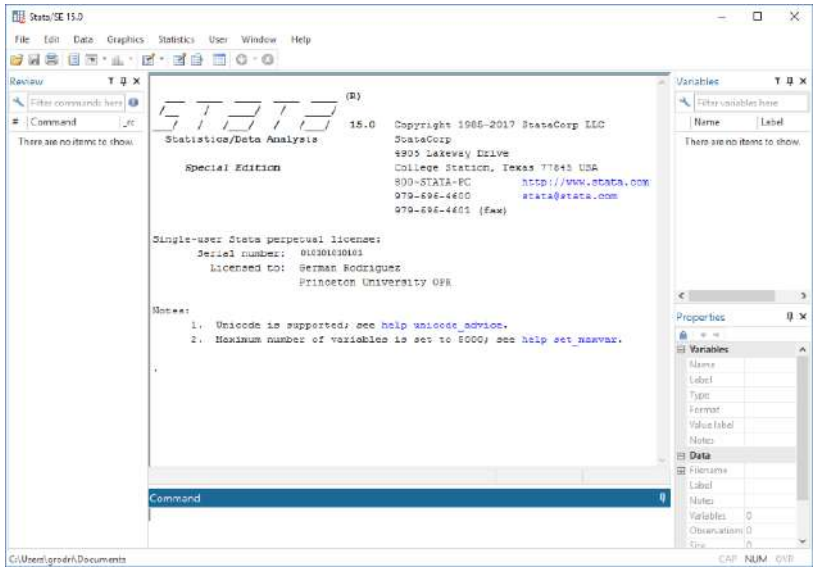


Figure: The STATA interface[21]

```
. twoway rarea d1 zero x1, color("blue%50") ///
>    || rarea d2 zero x2, color("purple%50") ///
>    || rarea d3 zero x3, color("orange%50")  ///
>    || rarea d4 zero x4, color("red%50") ///
>       title(January Temperatures by Region) ///
>       ytitle("Smoothed density") ///
>       legend(ring(0) pos(2) col(1) order(2 "NC" 1 "NE" 3 "S" 4 "W"))

. graph export kernel.png, width(500) replace
(file kernel.png written in PNG format)
```

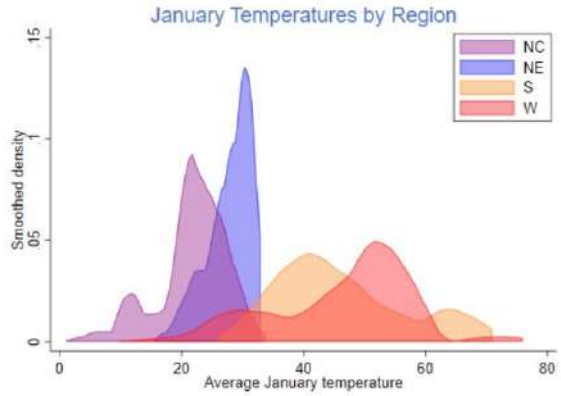


Figure: Plotting density estimates of January temperatures by region using STATA[22]

[21] http://data.princeton.edu/stata/
[22] http://data.princeton.edu/stata/graphics.html

SAS Enterprise Miner is a solution to create accurate predictive and descriptive models on large volumes of data across different sources in the organization. SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market.

Illustration

The following is a screenshot of the SAS Enterprise Miner interface, followed by an example output from the software, which is, in this case, the use of association rules (or market basket analysis) as an attempt to discover which items are purchased together with high confidence, and which items are not purchased with others – in order to help a retailer decide on a marketing strategy to become more profitable.

Where 1 is the Toolbar Shortcut Buttons, 2 is the Project Panel, 3 is the Properties Panel, 4 is the Property Help Panel, 5 is the Toolbar, 6 is the Diagram Workspace, and 7 is the Diagram Navigation Toolbar. The power of Enterprise Miner is the way the data scientist can design, build, edit, run, compare, and save the techniques and processing through graphically building, ordering, sequencing, and connecting nodes in the Diagram Workspace, and then very efficiently view and edit the settings of data sources, diagrams, nodes, and users in the Properties Panel.

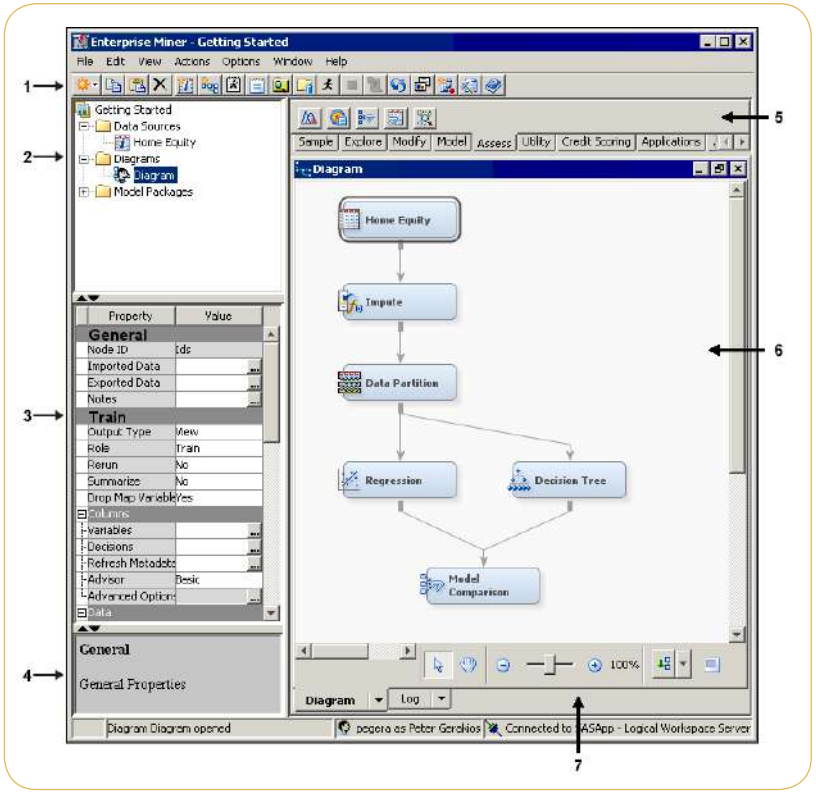


Figure: A screenshot of SAS Enterprise Miner's interface[23]

[23] http://support.sas.com/documentation/cdl/en/emgsj/66375/HTML/default/viewer.htm#p0avnz8kd00zq2n1itr7rgshh5w3.htm

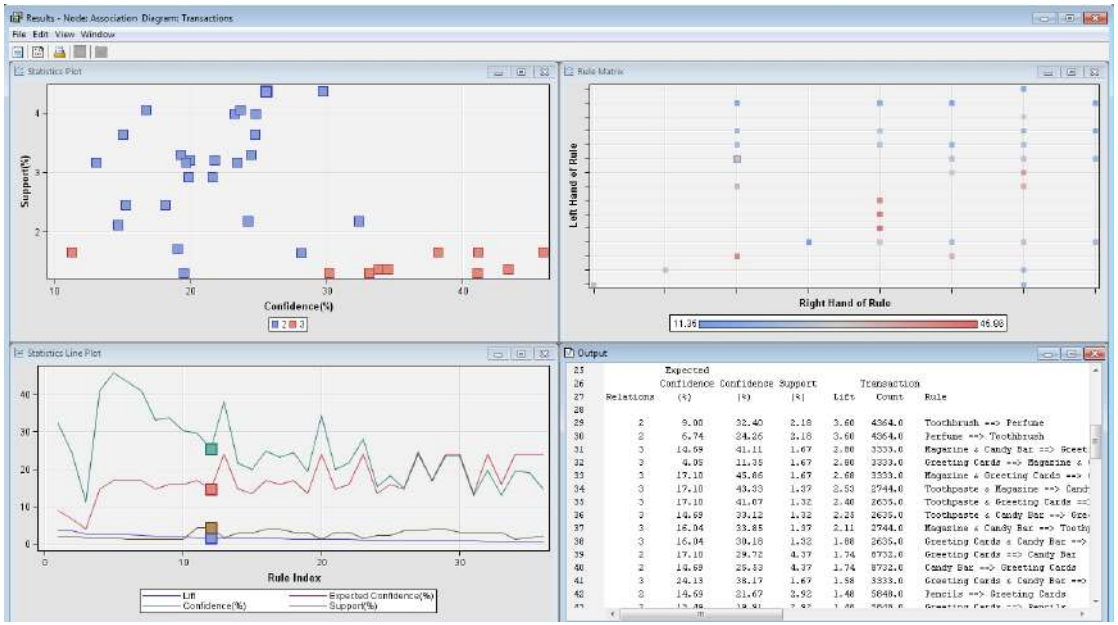


Figure: Market Basket Analysis example by SAS Enterprise Miner[24]

STS JMP (pronounced "jump") is a suite of computer programs for statistical analysis developed by the JMP business unit of SAS Institute. JMP is used in applications such as Six Sigma, quality control, and engineering, design of experiments, and research in science, engineering, and social sciences. The software can be purchased in any of five capability configurations: JMP, JMP Pro, JMP Clinical, JMP Genomics, and the JMP Graph Builder App for the iPad. The software is focused on exploratory visual analytics, where users investigate and explore data. These explorations can also be verified by hypothesis testing, data mining, or other analytic methods.

[24] https://dataminingandvisualisation.wordpress.com/2013/11/03/market-basket-in-sas/

Illustration

The following is a snapshot of the JMP interface showing the toolbar, data table, column options, and row options, followed by a screenshot of partitioning panel which can be done easily in JMP, and finally an output example from JMP that shows a decision tree.

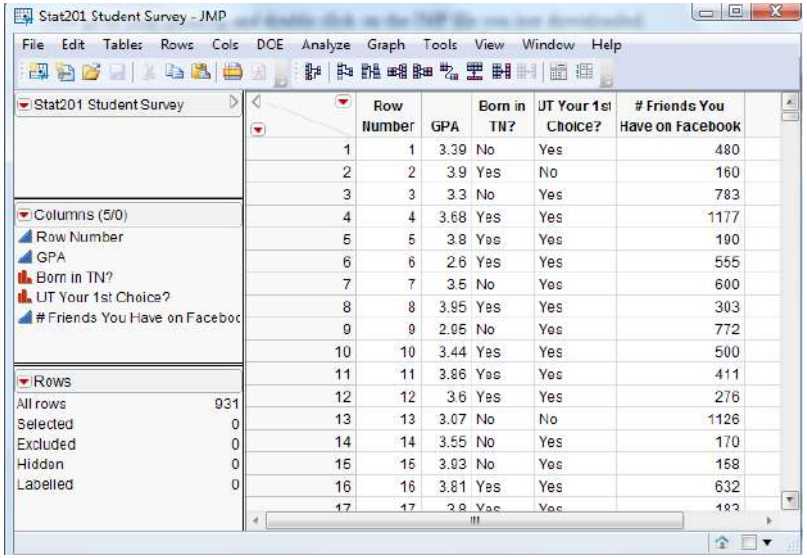


Figure: A screenshot of SAS JMP's interface[25]

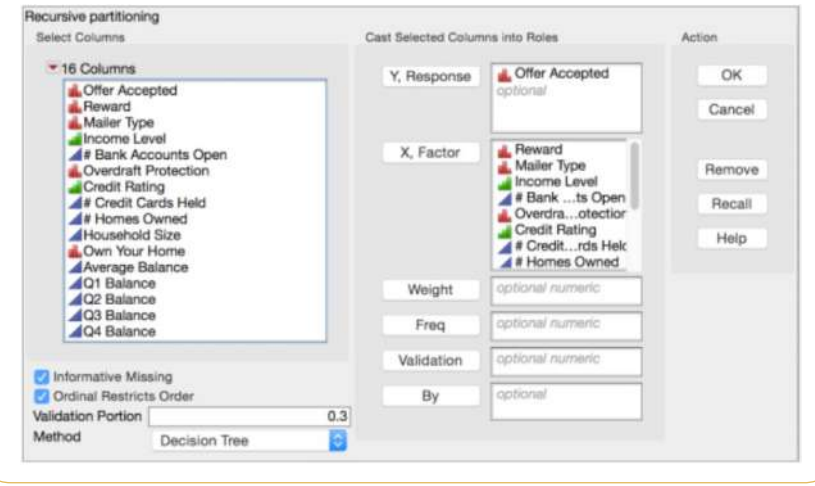


Figure: A screenshot of partitioning panel in JMP[26]

[25] http://web.utk.edu/~cwiek/201Tutorials/RandomSample/

[26] https://www.jmp.com/content/dam/jmp/documents/en/academic/case-study-library/case-study-library-12/analytics-cases/ct-creditcardmarketing.pdf

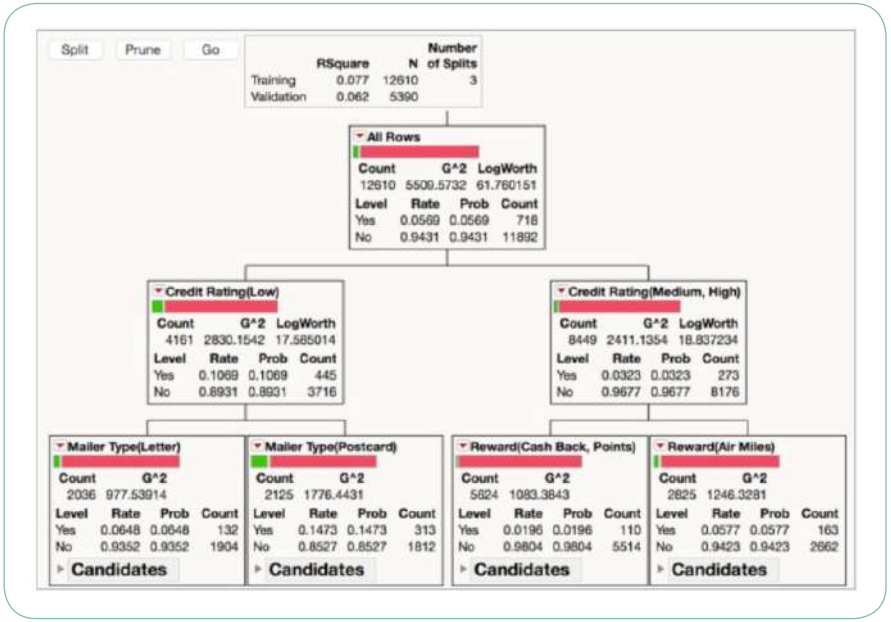
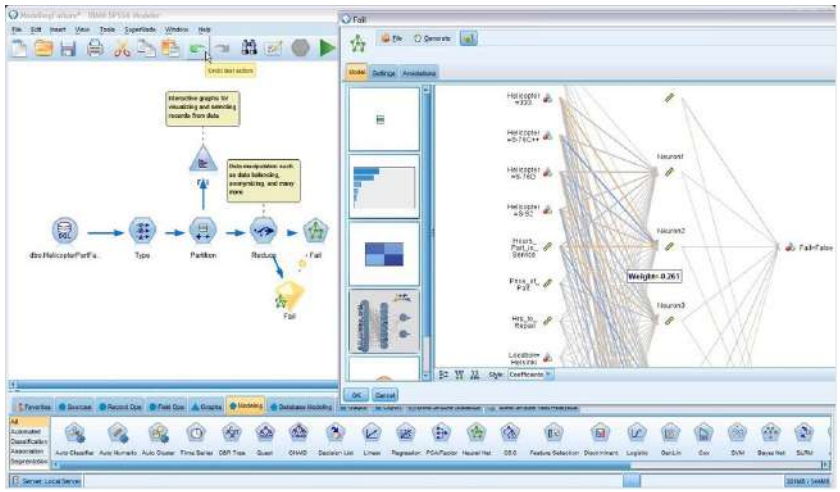


Figure: Decision Tree example by SAS JMP[27]

IBM SPSS Modeler is a data mining and text analytics software application from IBM. It is used to build predictive models and conduct other analytic tasks. It has a visual interface which allows users to leverage statistical and data mining algorithms without programming. One of its main aims from the outset was to get rid of unnecessary complexity in data transformations, and to make complex predictive models easy to use.



[Figure: Artificial Neural Networks Example in SPSS Modeler[28]*]*

[27] https://www.jmp.com/content/dam/jmp/documents/en/academic/case-study-library/case-study-library-12/analytics-cases/ct-creditcardmarketing.pdf

[28] http://www.spss.com.hk/software/modeler/index.htm?tab=1

ABOUT

The World Bank's engagement in the energy sector is designed to help client countries secure the affordable, reliable, and sustainable energy supply needed to end extreme poverty and promote shared prosperity. Its strategy mirrors the objectives of the Sustainable Energy for All Initiative and the Sustainable Development Goal (SDG) on energy, or SDG7: achieving universal access, accelerating improvements in energy efficiency, and doubling the global share of renewable energy by 2030.
Learn more at worldbank.org/energy.

Lighting Global is the World Bank Group's platform to support sustainable growth of the international off-grid solar market as a means of rapidly increasing energy access to the 1.2 billion people without grid electricity. Through Lighting Global, the International Finance Corporation (IFC) and the World Bank work with the Global Off-Grid Lighting Association (GOGLA), manufacturers, distributors, and other development partners to develop the modern off-grid energy market.
Learn more at https://www.lightingglobal.org/

GOGLA is a neutral, independent, not-for-profit industry association which acts as a sector enabler and advocate. GOGLA supports the growth and strengthens the market for clean, quality off-grid lighting and electrical systems for households, SMEs and communities in developing countries.
Learn more at gogla.org.

Making credit accessible to those that deserve it most, Lendable is passionate about making African consumer and SME credit a competitive asset class. Lendable builds technology and financial products to bring finance to 100 million borrowers that deserve it and created the first debt platform designed specifically for African Alternative Lenders. This includes non-banking, asset backed finance providers operating in microfinance, asset financing, asset leasing and a range of pay-as-you-go services.
Learn more at lendablemarketplace.com

Angaza removes the upfront price barrier of clean energy products, like solar home systems, by enabling off-grid customers in emerging markets to pre-pay for energy in affordable amounts spread over time. Through their B2B technology licensing model, Angaza provides embedded data transfer technologies and a cloud-based analytics software platform that allows manufacturers and distributors to finance clean energy products with a Pay-As-You-Go (PAYG) pricing model. Additionally, Angaza leverages usage and diagnostic data collected from every PAYG unit sold to lower after-sales support costs and present customer credit risk ratings. Based in San Francisco and Nairobi, Angaza's platform is operating in India, Kenya, Malawi, Nicaragua, Pakistan, Sierra Leone, South Africa, Uganda, Benin, Côte d'Ivoire, Madagascar, Myanmar, Rwanda, Senegal, and Tanzania.
Learn more at https://www.angaza.com/

Data Playbook

for the Off-Grid Pay-As-You-Go Sector

January 2018

Project Managers:

Anna Lerner, **alerner@worldbank.org**

Laura Sundblad, **l.sundblad@gogla.org**

Team Members:

Juan Andres Turner, **jaturner@worldbank.org**

Micah Melynk, **mmelnyk@worldbank.org**

Maite Lasa, **maitelasagarcia@worldbank.org**

Kian Behdad, **kianbehdad@gwmail.gwu.edu**

Partner Team Members:

Joe Brew, **joe@lendable.io**

Victoria Arch, **victoriaa@angazadesign.com**

Jennifer Sharma, **jennifer@angazadesign.com**

Contributing Partners