BUSA 603

Module 3: Web Analytics, Segmentation, Targeting & Positioning

Brian Weikel

brian.weikel@franklin.edu

FRANKLIN UNIVERSITY
Makes It Possible

Business Analytics
Franklin University

Spring 2024

Outline[1]

1. General Marketing Data Extraction Tools and Resources
2. Social Listening Tools
3. Web Analytics
4. Content Analysis Tools
5. An Introduction to Segmentation, Targeting and Positioning
6. Segmentation, Targeting and Positioning Implementation

[1] These lecture notes map to chapters 6 and 7 of Davis (2022).

General Marketing Data Extraction Tools and Resources

Search Volume Data Tools

A **search-volume data tool** indexes the frequency of terms people use on search engines like Google or Bing.

Google Trends (GT) provides access to a largely unfiltered sample of actual search requests made to Google. The data is anonymized and aggregated. That is, no one is personally identified, search queries are categorized, and queries are aggregated. Search volume index (SVI) is a key variable returned by GT.

There are two samples of Google Trends data that can be accessed:

- Real-time data is a sample covering the last seven days.
- Non-realtime data is a separate sample from real-time data and goes as far back as 2004 and up to 72 hours before your search.

Google Trends Data Is Used in Many Fields

1. In economics it has been used to forecast other economic variables such as retail trade survey results (Robin (2018)) and for measuring the causal effect of COVID-19 state lockdowns on mental well-being attributes (Brodeur et al. (2021)).
2. In marketing it has been used to monitor the evolution of consumer tastes and market performance dynamic analyses (Du et al. (2015)).
3. For demand estimation or causal inference studies using GT data to create an instrumental variable to mitigate endogeneity issues (Barron et al. (2020)).[2]
4. In epidemiology GT data has been used to predict the activity of influenza viruses and other communicable diseases, even prior to the onset of the COVID-19 global pandemic (Ginsberg et al. (2009)).[3]

---

[2] Endogeneity is a common issue in statistics and econometrics. For a regression model specification endogeneity occurs when an explanatory variable is correlated with the error term, which causes a bias in coefficient estimates.

[3] For other uses of GT data, see Jun et al. (2018).

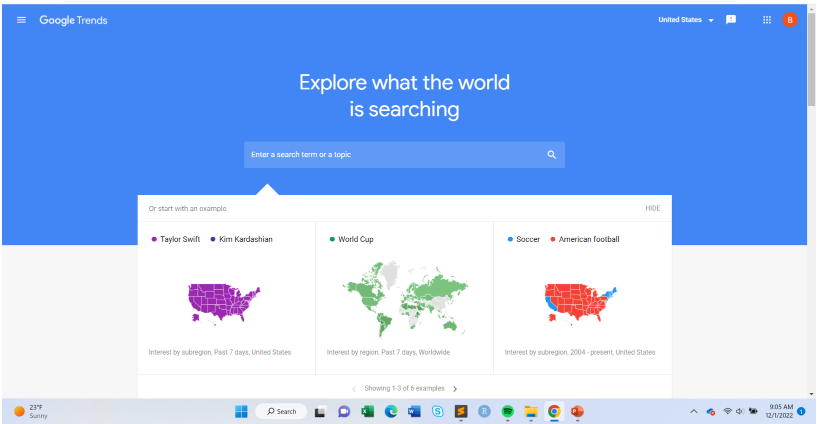

Figure 1: Google Trends Landing Page on 12/1/22

GT 12/1/22 Time Series Plot for “World Cup”

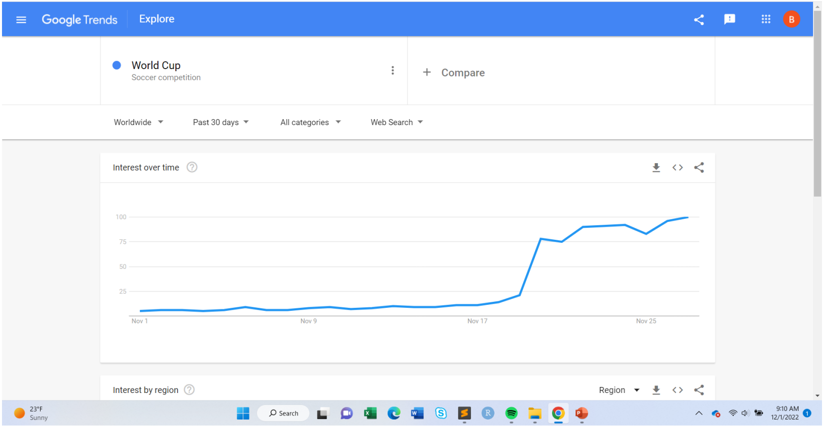

Figure 2: Google Trends 12/1/22 Time Series Plot for Keyword “World Cup”

GT 12/1/22 Country Map for “World Cup”

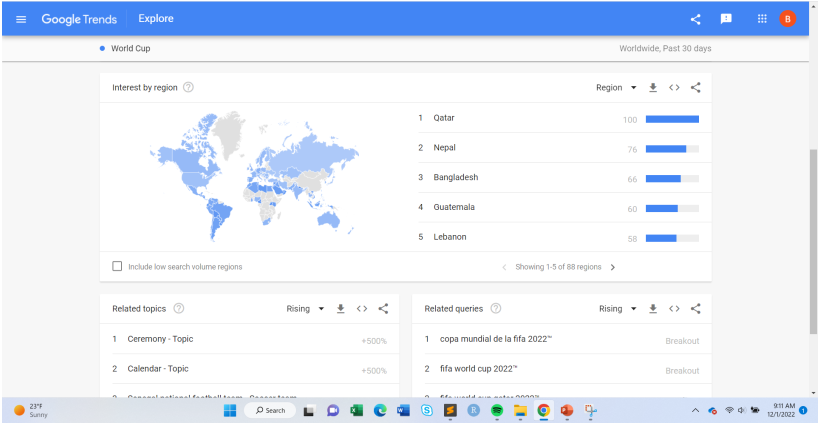

Figure 3: Google Trends 12/1/22 Country Map for Keyword “World Cup”

Is GT a Quality Data Source?

The main quality dimensions of data at the record level are accuracy, completeness, consistency and validity (Karr et al. (2006)).

1. **Accuracy** pertains to a variable measuring what it is expected to measure. Coverage biases, sampling defects, or non responses may characterize how accurate a source is.
2. A record is **complete** when it includes values for all attributes.
3. **Consistency** refers to the situation in which the relationships among the attribute values in the same record are valid. A lack of consistency is, for instance, a starting date after the end date.
4. An attribute value is **valid** when it is within a pre-established domain of acceptable values. For example, a person's age can only be a positive number. Ensuring attribute value validity is not enough for ensuring accuracy, although it is a necessary condition.

Caution is Advised When Using GT

- GT presents an issue in terms of accuracy, derived from the fact that the reports are generated from a sample of searches made by users (Choi and Varian (2012)). Since sampling methods are not disclosed by Google it is not possible to quantify the sampling error.[4]
- GT includes data for all the observations, although it does not mean that a value is provided for each time period, say, due to small search popularity. Since '0' values precisely represent low popularity, the lack of completeness is not a major issue.
- Consistency is not an issue with GT since date and SVI are always returned.
- Though there are instances where SVI is reported as a negative value, in general the GT data is valid.

[4] For an empirical analysis of the accuracy of GT data, see Cebrián and Domenech (2022).

Online Content Count Tool

An **online content count tool** calculates frequencies of word or phrases embedded within a media provider's website.

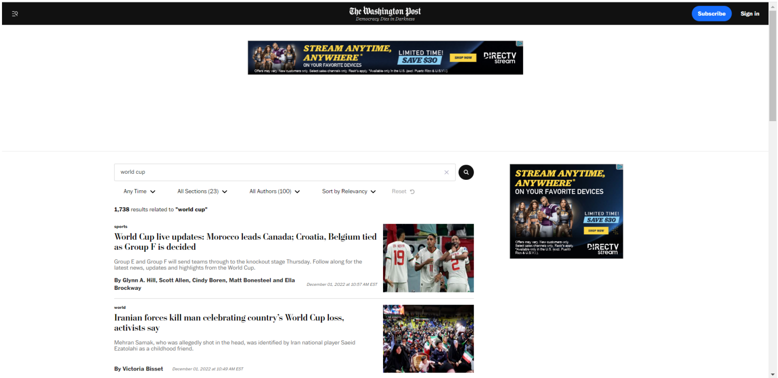

Figure 4: Returned page of Washington Post's Online Content Count Tool for 12/1/22 and Keyword "world cup"

Commercial Data

Recall from Module 2 that **third party data**, also called syndicated data, are data from another firm like second party data. However, unlike second party data the provider does not have a direct relationship or agreement with the customer. Examples include IRI, Nielsen, NielsenIQ, NPD, or Spins.

The Google Cloud also has many commercial databases available for purchase.

Some Public Data

Public data are data available without purchase.[5] Sources include:

- National and local governments. For example, FRED data as maintained by the St. Louis Fed.
- Private firm public data, such as Amazon Web Services weather data.
- Google makes available various public datasets via the Google Cloud.

Academic maintained data, such as Nielsen and NielsenIQ data maintained by the University of Chicago's Kilt's Center for Marketing. This data is for academic research only.

Statista is a good place to begin to identify data sources and availability.

[5] There may limitations on who may have access to the data and how the data may be used.

Social Listening Tools

Social Listening Tools

Social listening tools are platforms that connect to various social media networks in order to extract consumer data.

Social listening tools connect to Facebook, YouTube, WhatsApp, . . . and extract data from them.

Some social listening tools can handle several platforms while others focus solely on one platform.

Social listening tools have become increasingly more important as firms are looking for more and more information to help them make better marketing decisions

Ranking of Social Media by Global Active Users

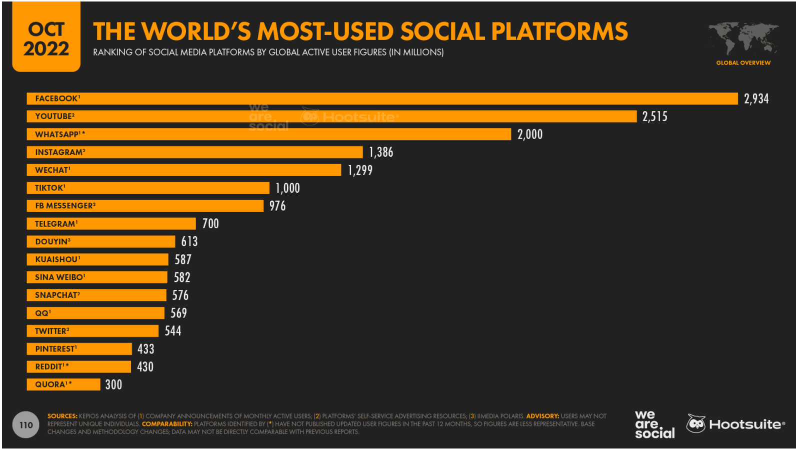

Figure 5: Ranking of Social Media Platforms by Global Active Users in Millions

Average Time Per Month For Android App Users

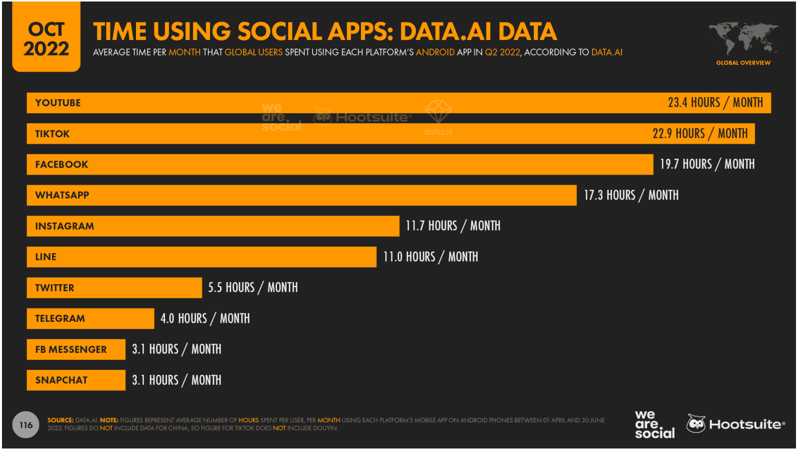

Figure 6: Average Time Per Month That Global Users Spent Using Each Platforms Android App in Q2 2022

% of Social Media Platform Users on Other Platforms

OCT 2022

SOCIAL MEDIA PLATFORM AUDIENCE OVERLAPS

PERCENTAGE OF USERS OF EACH PLATFORM AGED 16 TO 64 OUTSIDE OF CHINA WHO ALSO USE OTHER SOCIAL MEDIA PLATFORMS

GLOBAL OVERVIEW

	UNIQUE TO PLATFORM	ALSO USING FACEBOOK	ALSO USING YOUTUBE	ALSO USING WHATSAPP	ALSO USING INSTAGRAM	ALSO USING TIKTOK	ALSO USING TELEGRAM	ALSO USING SNAPCHAT	ALSO USING TWITTER	ALSO USING REDDIT	ALSO USING PINTEREST	ALSO USING LINKEDIN
FACEBOOK USERS	0.6%	100%	72.6%	72.3%	77.4%	50.2%	44.0%	33.3%	49.9%	14.1%	33.7%	31.3%
YOUTUBE USERS	0.9%	78.9%	100%	71.3%	76.5%	47.5%	46.9%	30.6%	51.2%	16.4%	35.9%	31.0%
WHATSAPP USERS	1.3%	80.5%	74.7%	100%	79.0%	48.8%	51.4%	35.2%	50.0%	13.2%	35.2%	32.3%
INSTAGRAM USERS	0.1%	82.9%	75.5%	76.0%	100%	52.2%	49.0%	37.5%	55.8%	15.3%	37.6%	31.6%
TIKTOK USERS	0.1%	83.4%	77.2%	72.8%	81.0%	100%	49.4%	40.3%	57.4%	16.7%	39.9%	30.1%
TELEGRAM USERS	0.1%	81.6%	79.8%	85.5%	84.9%	55.1%	100%	41.0%	60.8%	16.8%	39.7%	37.6%
SNAPCHAT USERS	0.1%	83.6%	77.7%	79.5%	88.1%	61.0%	55.7%	100%	62.3%	22.3%	46.2%	38.8%
TWITTER USERS	0.2%	83.4%	77.5%	75.0%	87.1%	57.8%	54.8%	41.4%	100%	21.3%	41.1%	39.5%
REDDIT USERS	0.1%	80.6%	78.0%	67.8%	81.4%	57.2%	51.7%	50.7%	72.7%	100%	57.2%	51.7%
PINTEREST USERS	0.1%	82.4%	77.5%	77.3%	85.9%	58.7%	52.4%	44.9%	60.2%	24.5%	100%	42.9%
LINKEDIN USERS	0.2%	87.6%	75.5%	81.1%	82.6%	50.6%	56.7%	43.2%	66.1%	25.3%	49.1%	100%

113

SOURCE: GWI (Q2 2022). SEE GWI.COM FOR MORE DETAILS. **NOTES:** ONLY INCLUDES USERS AGED 16 TO 64. DOES NOT INCLUDE DATA FOR CHINA. VALUES REPRESENT THE USERS OF THE PLATFORM IDENTIFIED IN THE LEFT-HAND COLUMN WHO ALSO USE THE PLATFORM IDENTIFIED IN THE ROW AT THE TOP OF EACH COLUMN. PERCENTAGES IN THE "UNIQUE TO PLATFORM" COLUMN REPRESENT USERS WHO SAY THEY DO NOT USE ANY OTHER SOCIAL NETWORK OR MESSENGER SERVICE, INCLUDING PLATFORMS NOT FEATURED IN THIS TABLE. **COMPARABILITY:** SURVEY CHANGES.

we are social · Hootsuite

Figure 7: Percentage of Users of Each Platform Aged 16 to 64 Outside of China Who Also Use Other Social Media Platforms

Social Listening Tools Purpose

A **single-attribute social listening tool** focuses on one social media characteristic such as text or image content. For example, tools that focus only on reading blog posts would be single-attribute.

Multi-platform social listening tools allow marketing analytics professionals to see data across many social media platforms, say across Snapchat, Pinterest, and Twitter.

A **native social listening tool** is a social media platform specific application that provides detailed analytics about who is interacting with the organization's platform content. The textbook mentions that Hootsuite, Union Metrics (formerly TweetReach), and Tweeple Search are examples of social listening tools.[6]

[6] The textbook does not mention that Union Metrics was acquired by Trendkite, which was subsequently acquired by Cision.

Hootsuite Marketing Collateral and Information

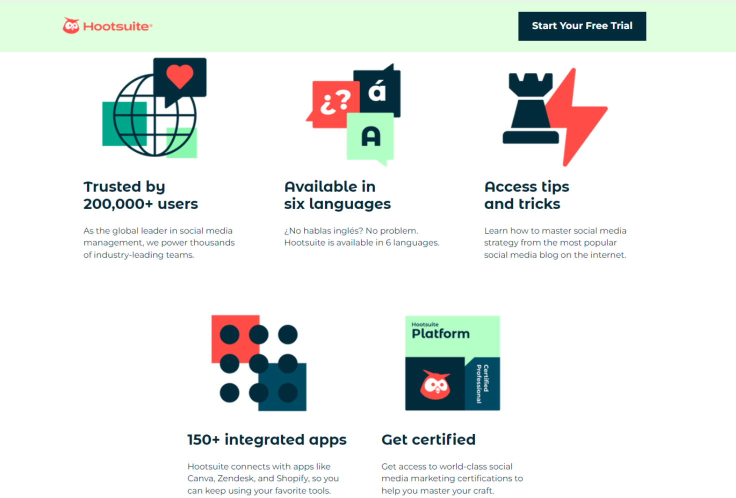

Figure 8: Hootsuite Marketing Collateral and Information

Audience Insight Marketing Collateral and Information

FRANKLIN UNIVERSITY
Makes It Possible

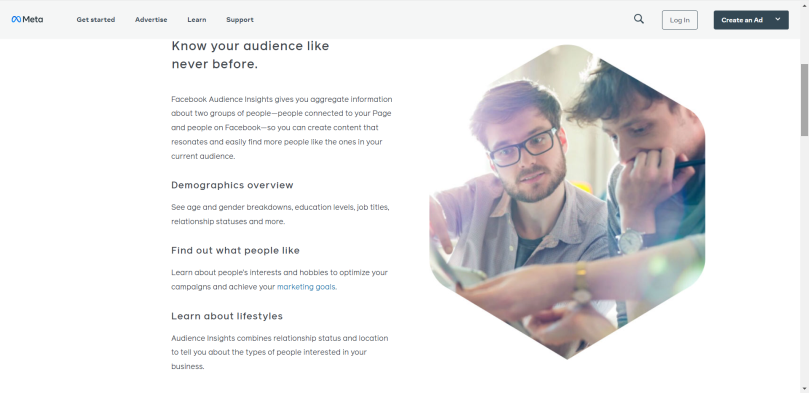

Figure 9: Facebook Audience Insight Marketing Collateral and Information

Web Analytics

Defining Web Analytics

Gartner defines **web analytics** as specialized analytic applications used to understand and improve online channel user experience, visitor acquisition and actions, and to optimize digital marketing and advertising campaigns.

- Commercial products offer reporting, segmentation, analytical and performance management, historical storage and integration with other data sources and processes.
- The tools are used by marketing professionals, advertisers, content developers and the website's operations team, and increasingly provide input to automated tools that target improved customer experience.

Distinguishing Google Analytics and Google Ads

FRANKLIN UNIVERSITY
Makes It Possible

While Google Analytics (GA) has a large share of the web analytics market, competitors include Leadfeeder, Piwik Pro Analytics Suite, Smartlook, and Woopra.

- Challenges with GA include the non-negligible effort required to ensure the tool is GDPR compliant, and it is not the simplest tool to learn and use.
- To overcome these challenges, Google offers many courses on how to set-up and use the service.

Google Analytics is not Google Ads.

1. Google Analytics is a web analytics service offered by Google that tracks and reports website traffic, currently as a tool inside the Google Marketing Platform.

Specifying Google Ads

2. Google Ads, previously Google AdWords and Google AdWords Express, is an online advertising solution that businesses use to promote their products and services on Google Search, YouTube, and other sites across the web.
    - Google Ads also allows advertisers to choose specific goals for their ads, like driving phone calls or website visits.
    - With a Google Ads account, advertisers can customize their budgets and targeting, and start or stop their ads at any time. Google ads typically appear on two networks.
        - Google Search Network, which includes Google search result pages, other Google sites like Maps and Shopping, and partnering search sites.
        - Google Display Network, which includes Google sites like YouTube, Blogger, and thousands of partnering websites.

Google Analytics 4 is the Next Version of GA

On July 1, 2023, the pre-existing and standard GA default property **Universal Analytics** would no longer process data.[7]

Data will only flow into **Google Analytics 4** (GA4), the next version of GA (i.e., replacement of Universal Analytics).

This page lists the steps required to use GA4 for an organization's website or app, where one step is pasting a Google tag (i.e., a JavaScript snippet for a website) immediately after the `<head>` on each page of the organization's website or using a Firebase software development kit (SDK) for an app.

[7] Via a one-time processing extension, one will be able to see Universal Analytics reports until July 1, 2024. Some refer to Universal Analytics as "Analytics Classic."

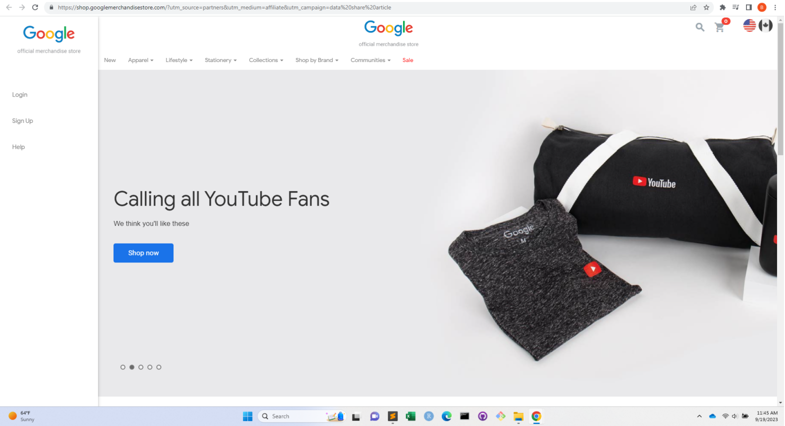

Figure 10: Google Merchandise Store

The Google-provided GA4 demo uses data from either the Google Merchandise Store or the app game Flood-It!

Flood-It!

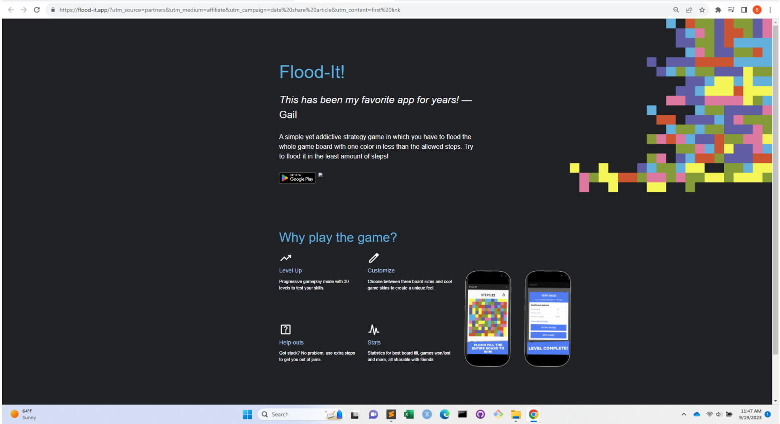

Figure 11: Flood-It!

Defining Account, Property and Data Stream

The GA4 `Admin` pages allows one to configure the account, properties, and data streams.

- An *account* is the access point for GA4.
- A *property* is a website, mobile application, or device (e.g., a kiosk or point-of-sale device). A property lives within an account.[8] Properties are the containers for GA4 end-user reports based on the data collected from apps and sites. It is the level at which GA4 processes data and where GA4 can connect with other Google products, like Google Ads.
- A *data stream* lives within a property and is the source of data from the app or website. A property can have one or many data streams.

[8] An account can include multiple properties and property types, but a property can belong to only one GA4 account.

GA4 Admin

The remainder of the screenshots in this section are for the Google-provided properties “GA4 - Google Merch Shop” or “GA4 - Flood-It!”.

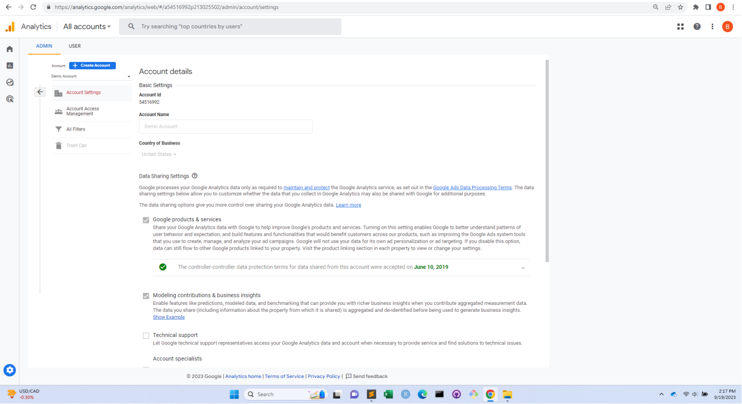

Figure 12: GA4 Admin

GA4 `Admin` Configuration Options

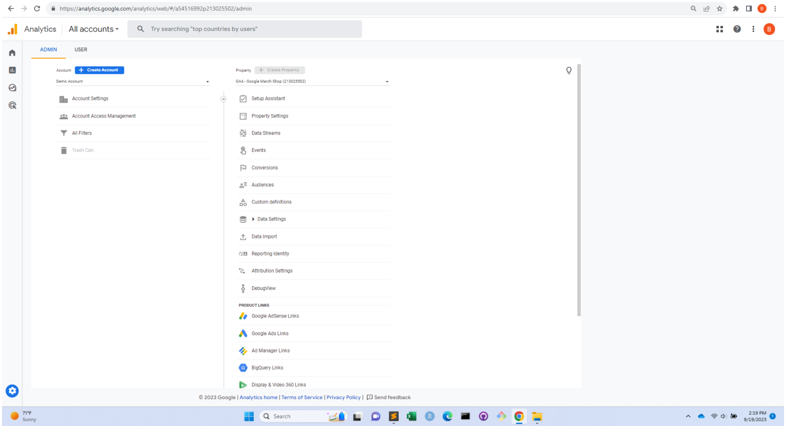

Figure 13: GA4 `Admin` Configuration Options

`Admin` has many configurable options, where the `Setup Assistant` may be used.

Events, Parameters, User Properties and Conversions

Events are user interactions with a website or app that can be measured, like a video view.

Event parameters are additional pieces of information sent with events that can further specify an action the user took or add further context to the event, like the name of the video or how long the user watched it.

User properties are attributes about who is using the organization's app or website that can help end-users better understand segments of the organization's user base, like geographic location or device used.

A *conversion* is an event that one considers important to the organization, such as a purchase or a download. One may want to share conversions with marketing platforms, like Google Ads.

Dimensions, Metric and Their Values

Dimensions answer the question, "Who, what, or where?" While *metrics* answer the question, "How many?" For example, dimensions answer the question, "What device is most commonly used?" While metrics answer questions like, "How many users visited the site yesterday?"

- A *dimension* is a text-based label that represents the data.
- *Dimension values* are the text values for the dimensions that have been selected.
- A *metric* is a numeric value that represents the data.
- *Metric values* are the the numeric values that are being calculated for the dimension values selected.

Managing Events

Recall GA4 collects and stores user interactions with the website or app as events. Events provide insight into what is happening on the website or app, such as:

- Clicks and page views on the website.
- Installs and opens on the app.
- User engagement and conversions on either platform.

Also recall event parameters give additional information with each event. For example, for a collected page view event, GA4 includes the page's URL as an event parameter value.

<u>Automatically collected</u> events are triggered by basic interactions with the app and/or site (as indicated under the event name in a GA4 table). As long as one uses the GA4 Firebase SDK or gtag.js additional code to collect these events is unneeded.

GA4 Admin Event Definitions

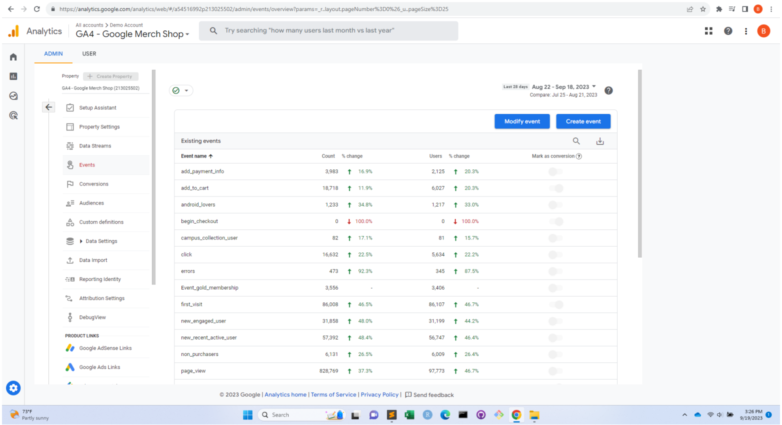

Figure 14: GA4 Admin Events

Enhanced Measurement

Enhanced measurement lets one measure interactions with a website or app by enabling new events in GA4. No code changes are required. When enabled for a web data stream the GA4 tag automatically starts sending event information right away.

- *Recommended events* are those the organization implements, but which have predefined names and parameters already recognized by GA4. Adding these events to the website or mobile app helps one measure additional behavior, as well as generate more useful reports.
- A *custom event* is one whose name and parameters select GA4 end-users define. A custom event lets one collect <u>specific</u> business data that GA4 does not otherwise automatically collect.

When one implements new and custom events and their associated parameters on the website or app, one starts sending this new data to GA4. In these cases, one <u>must</u> create custom dimensions and metrics that correspond to the data so they will be available in the reports.

GA4 Admin Custom Dimensions

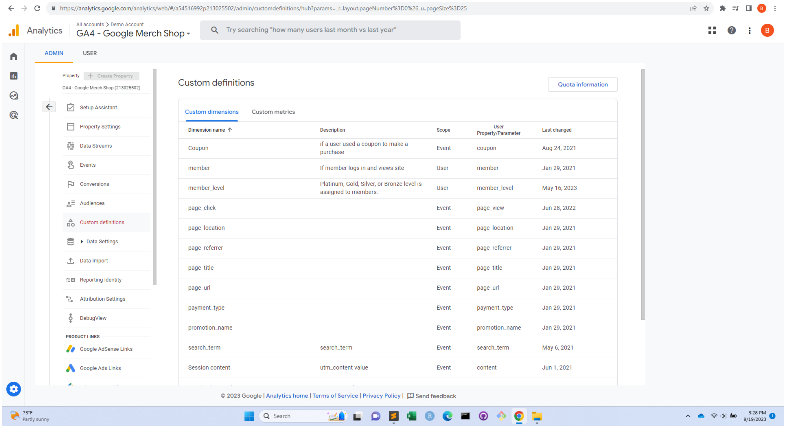

Figure 15: GA4 Admin Custom Dimensions

Note a create custom dimension button is not illustrated in the screenshot since we are not one of the "select GA4 end-users" of the "GA4 - Google Merch Shop" property.

GA4 Admin Custom Metrics

FRANKLIN UNIVERSITY
Makes It Possible

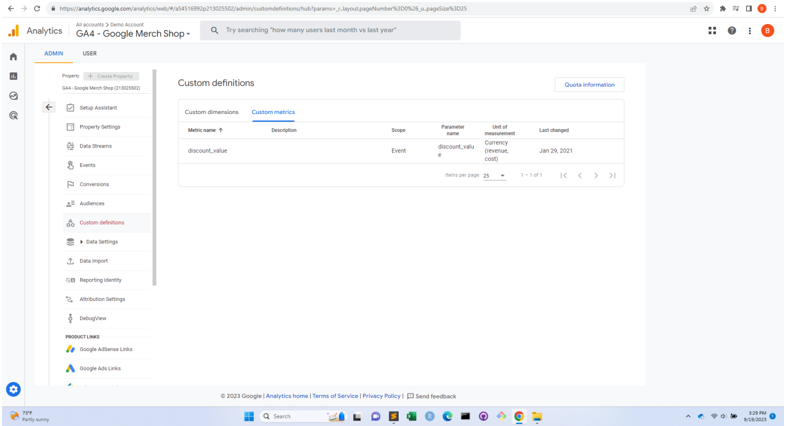

Figure 16: GA4 Admin Custom Metrics

Note a create custom metric button is <u>not</u> illustrated in the screenshot since we are <u>not</u> one of the "select GA4 end-users" of the property.

Managing Account Access and Settings

- *Administrators* have full control of the GA4 account. They can manage users (add or delete users, assign any role or data restriction) and grant full permissions to any user.
- *Editors* have full control of the settings of the account and its properties. Editors cannot manage users.
- *Marketers* can create, edit, and delete *audiences*, conversions, *attribution models*, events, and *lookback windows*.[9]

[9] (1) A GA4 audience is a group of site and/or app users who have generated similar behavioral data or who share demographic or other descriptive data (e.g., same age group, were acquired by the same campaign). (2) Attribution is the act of assigning credit for conversions to different ads, clicks, and factors along a user's conversion path. A GA4 attribution model can be a rule, a set of rules, or a data-driven algorithm that determines how conversion credit is assigned to conversion path touchpoints. (3) Google Ad Manager only records conversions for users who have previously seen or clicked on a Google Ad Manager ad within a period of time that one specifies, called a lookback window (lbw). There are two lbws, one for clicks and one for impressions. The lbws are set for each activity group, and can be from 1 to 30 days.

Managing Account Access and Settings Continued

- *Analysts* can create, edit, and delete certain property assets, like *explorations*. They can also collaborate on shared assets.[10]
- *Viewers* can see settings and data, and they can change the data they see in reports, like adding comparisons or adding a secondary dimension. Viewers can see new reports and collections in the left navigation, but they can't make changes to the navigation.
- Those who are classified as *None* are people who have no role for the account or property, but may have a role for a related account or property.

For the remainder of this section we will review a few GA4 reports that are frequently used. As you will discover during the tutorials, these reports are just a small subset of those provided by GA4.

[10] Explorations are private by default. The creator of an exploration can share it with others.

GA4 `Home Page` Insights

The `Home Page` surfaces information that is relevant to an organization.

One can use the page to monitor traffic, navigate around GA4, and get insights about the organization's websites and mobile apps.

The `Overview` *card* shows metrics that are relevant with a trendline for each metric.

- Cards focus on a specific objective.
- These cards are typically previews of a detail report, which lets one dig deeper into the data set.

GA4 Home Page

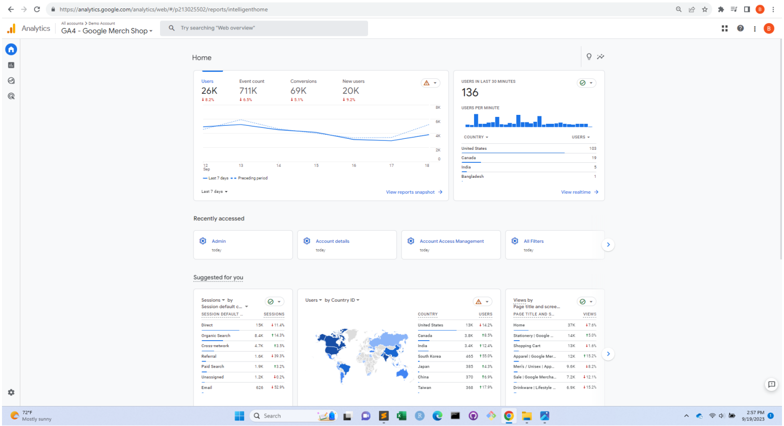

Figure 17: GA4 Home Page

Card Controls and Search Box

Card controls include:

- Dimension/metric picker: If a card has dimension or metric pickers across the top, one can use them to change which data is displayed in the card.
- Date range selector: One can use the date range selector in the upper right to select the time frame for reporting.
- Link to associated report: To open the associated report, select the link at the bottom right of the card. This will open a more in-depth report on the card topic.

To find a specific report or insight use the *search box* at the top of the GA4 account to ask a question, such as, "How many people came to the website last month?" When one selects the search box, they will see recent searches and reports opened, as well as suggested queries.

FRANKLIN UNIVERSITY
Makes It Possible

Realtime lets one monitor activity on the website or app as it happens.

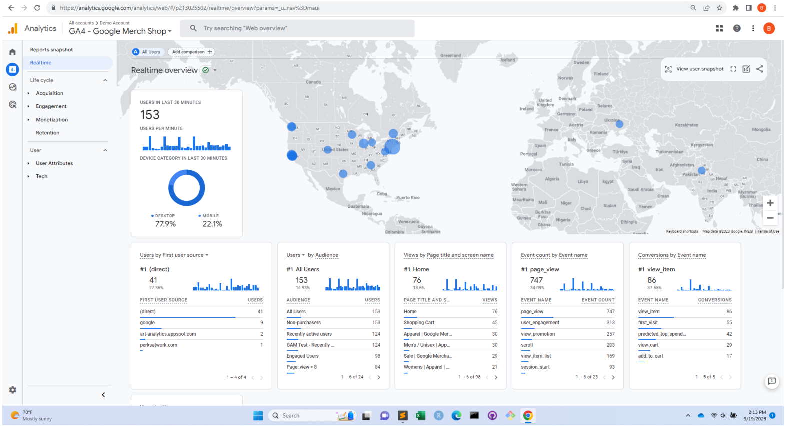

Figure 18: GA4 Realtime Report

GA4 `Realtime` Report Insights

`Realtime` lets one see:

- 'Users in Last 30 Minutes', which shows all users on the organization's site or app.
- 'New users', which shows users on the organization's site or app for the first time.
- 'Users', which shows returning users.

With `Realtime`, one can immediately and continuously monitor the effects that new campaigns and site changes have on traffic.

- See whether a one-day promotion is driving traffic to the site or app.
- Monitor the immediate effects on traffic from a blog/social network post or tweet.
- Monitor whether new and changed content on the site is affecting traffic. May also be used to verify that the measurement code is working on the site or app.

GA4 Reports snapshot

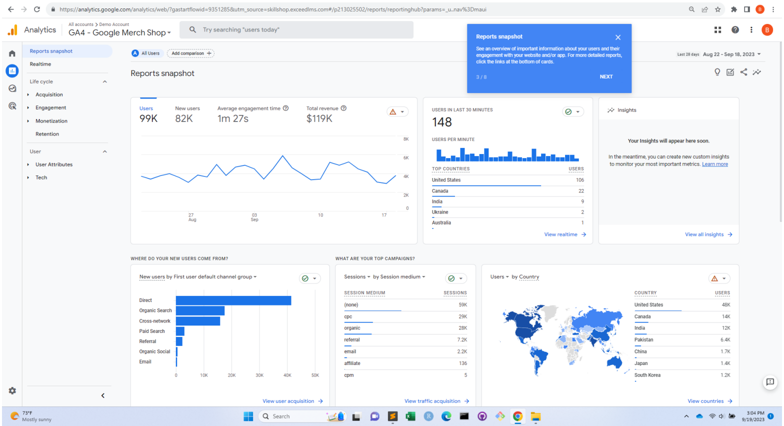

Figure 19: GA4 Reports snapshot

The Reports snapshot is the overview report displayed when one clicks Reports in the left navigation. Any overview report can be set as the Reports snapshot.

FRANKLIN UNIVERSITY
Makes It Possible

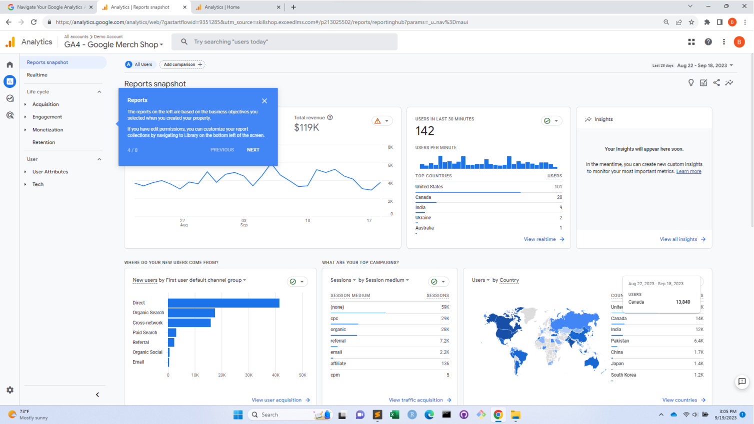

Figure 20: GA4 Reports Sections

The *Reports* section allows one to view ready-made reports that answer common questions about how users are interacting with the app or website.

GA4 Acquisition Overview Report

FRANKLIN UNIVERSITY
Makes It Possible

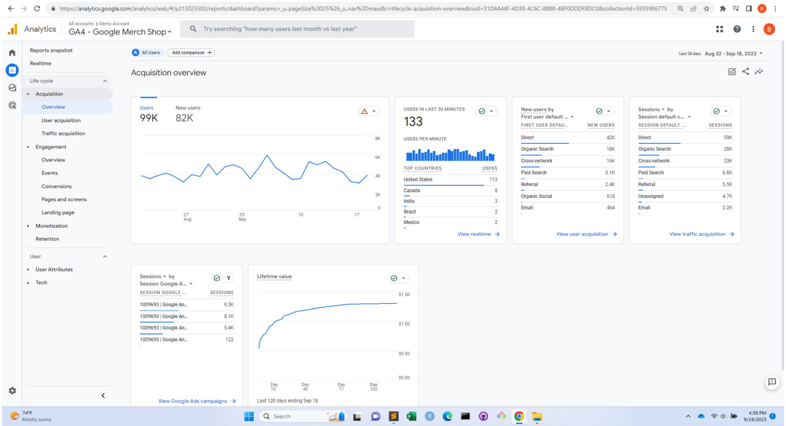

Figure 21: GA4 Acquisition Overview Report

This pre-made report can help determine if marketing is attracting new website or app users, if re-engagement campaigns are bringing users back, and if marketing strategies should continue or be adjusted.

FRANKLIN UNIVERSITY
Makes It Possible

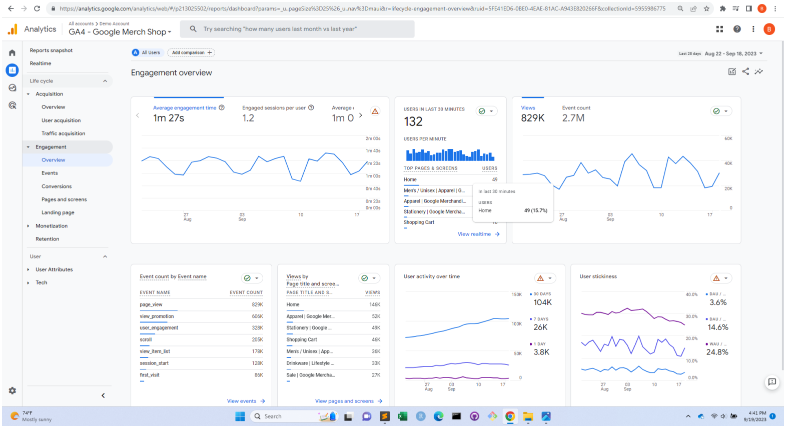

Figure 22: GA4 Engagement Overview Report

This pre-made report can be used to compare key engagement metrics over time, understand which pages and screens users are visiting, and identify the features with which they are interacting.

GA4 Monetization Overview Report

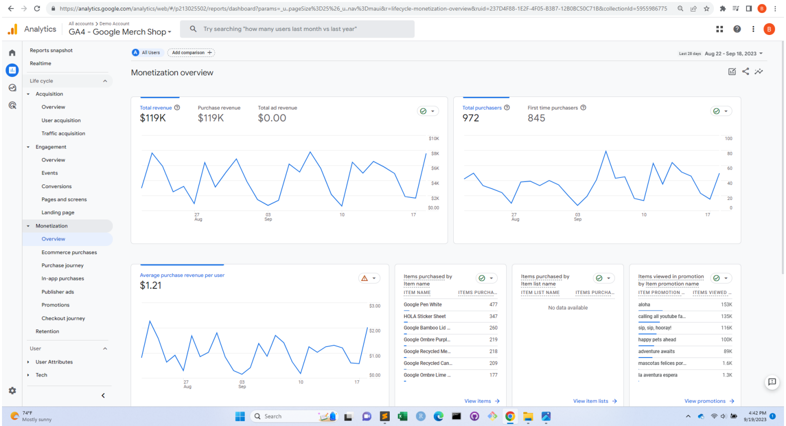

Figure 23: GA4 Monetization Overview Report

This pre-made report can help determine which products are selling, if promotions and coupons are successfully bringing in new users, and if ads displayed on the mobile app yield revenue.

GA4 Explorations Section

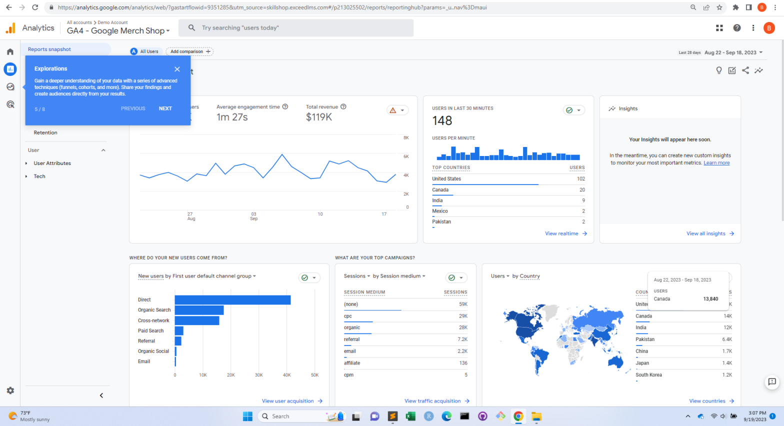

Figure 24: GA4 Explorations Section

Explorations is a collection of advanced techniques that go beyond standard reports to help an organization uncover deeper insights about user behavior.

GA4 Advertising Section

FRANKLIN UNIVERSITY
Makes It Possible

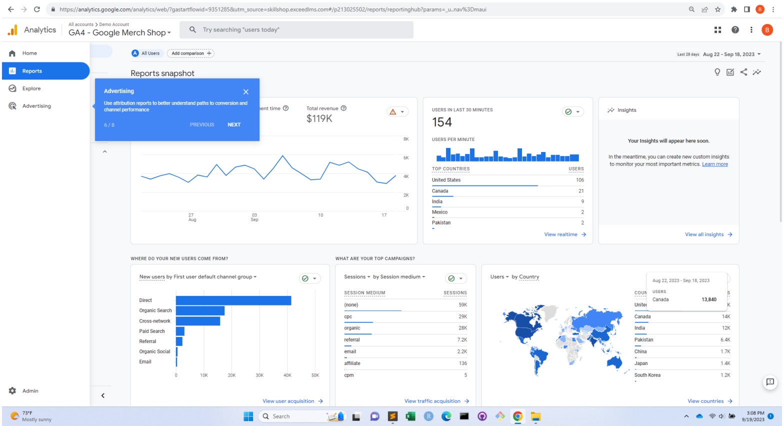

Figure 25: GA4 Advertising Section

The reports in this section help one better understand the media ROI across all channels, make informed decisions about budget allocation, and evaluate attribution models.

Content Analysis Tools

Comparing Content Analysis & Social Listening Tools

Content analysis tools allow marketing analytics professionals to analyze digital text, photos, audio, or visual formats of communication in greater detail than social listening tools.

- Social listening tools generally provide broad analyses to monitor conversations on Instagram, Twitter, and Facebook in real time.
- Firms use content analysis tools to provide deep analyses on words or phrases that are associated with specific images posted on social media. Content analysis tools typically include both (1) **sentiment analysis**, which focuses on translating words into customers' attitudes, and (2) **text analysis**, which focuses more on general themes.

Commercial Content Analysis Tools

Linguistic Inquiry and Word Count (LIWC), is one of several content analysis commercial tools.

It interprets text to reveal the thoughts, attitudes, feelings, personality, and motivations of the author.

- It has decades-worth of built-in dictionaries that translate words and phrases into psychological states (e.g., irony, sarcasm, metaphor, ...). More specifically, LIWC uses statistical supervised learners to attribute words to psychological states.
- It also allows the user to define dictionary entries, relating words to a particular attitude.
- Essentially, LIWC allows language to be translated into useful insights about the likely sentiment behind a writer's words.

Data Scraping

Data scraping is a computer-programmed extraction of information from individual computer screens, websites, or reports. The legal and legitimate uses focus primarily on scraping from public websites.

A **web crawler**, also known as a **web bot** or a **web spider**, uses web scraping to read HTML files. Search firms, like Google and Microsoft, use web crawlers to index the state of the web periodically in an effort to make searching quicker.[11]

The text mentions the following commercial data scrapers: Datastreamer (previously Spinn3r), Dexi.io, and Octoparse.

[11] Google's web crawler is Googlebot. Microsoft Bing's standard web crawler is Bingbot.

An Introduction to Segmentation, Targeting and Positioning

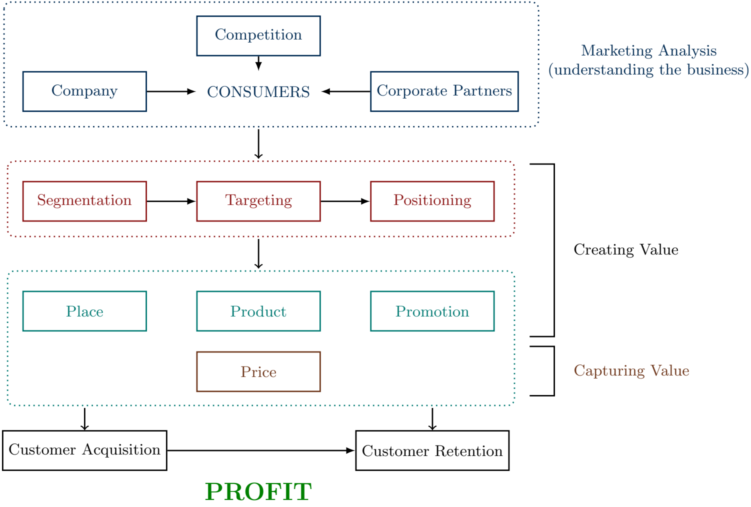

Figure 26: The Big Picture

Defining Segmentation

Segmentation, targeting, and positioning, often shortened by marketing analytics professionals to "STP", is a core principle of marketing. **Segmentation** means dividing the total market of customers into smaller groups that are alike or **homogeneous**.

There are three classifications of marketing segmentation data: demographic, psychographic and behavioral.

1. Demographic Segmentation
   - Demographic data include customer characteristics like gender, age, income, ethnicity, residence, political affiliation, education level, and customer languages spoken.
   - The textbook lists 10 external sources of demographic data (e.g., U.S. Census Bureau, Canada Open Data, European Open Data . . .).

Psychographic Segmentation

2. Psychographic Segmentation

   - Psychographic data are the psychological characteristics, values, life-stages, and lifestyles of customers.
   - While more difficult to find, depending on the marketing objective, psychographics are more powerful than demographics because they frequently improve identifying people in their core beliefs.
   - Claritas and Ersi are vendors of zip code level psychographic data.
   - Equifax and Experian are vendors of household/person level psychographic data.
   - A University of North Carolina page lists several other sources of psychographic data.

Demographic and Psychographic Segmentation View

Consumers organized on the basis of lifestyle and values.

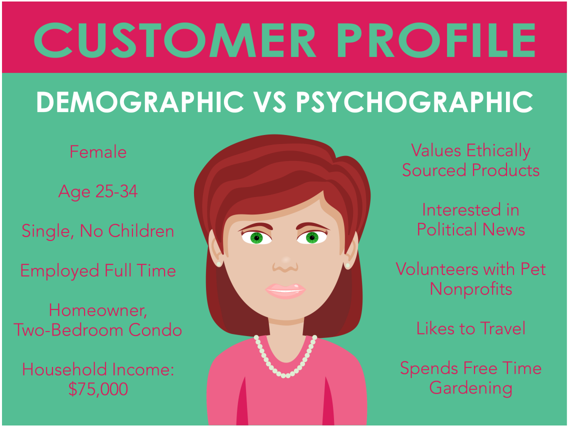

Figure 27: Demographic and Psychographic Segmentation Views

Behavioral Segmentation

3. Behavioral Segmentation
   - Behavioral data is data generated by, or in response to, a customer's engagement with a business.
     - This can include things like page views, email sign-ups, or other important user actions.
     - Common sources of behavioral data include websites, mobile apps, CRM systems, marketing automation systems, call centers, help desks, and billing systems.
   - Similar to demographic data, first-party data is often the best source of behavioral data.
   - The text book lists 5 external sources of behavioral data (e.g., Nielsen, IRI, . . .)

Behavioral Segmentation Option Examples

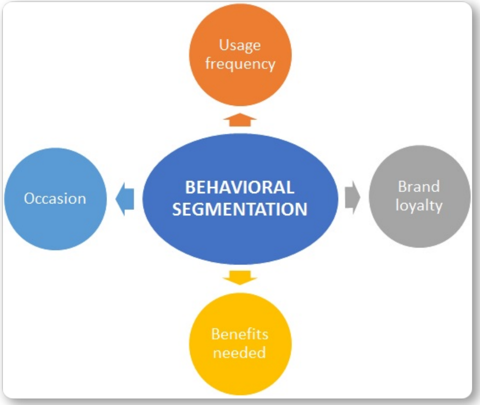

Figure 28: Behavioral Segmentation Options

There are Several Other Segmentation Methods

Segmentation	Sample Segments
Geographic	Continent: North America, Asia, Europe South America, Africa Within the United States: Pacific, Mountain, Central, South, Mid-Atlantic, Northeast.
Demographic	Age, gender, income
Psychographic	Lifestyle, life-stage, self-concept, self-values
Benefits	Convenience, economy, prestige
Behavioral	Occasion, loyalty, usage, benefits

Table 1: Segmentation Methods

Targeting

Targeting means appealing to particular segments of customers. There are three main strategies for targeting: differentiated, concentrated, and undifferentiated targeting.

- **Differentiated targeting** is when a firm targets each of the potential segments it discovers, each with a different marketing mix. For example, Petsmart's organic dog food category manager is looking to target a specific type of person - a health conscious, animal loving and eco-friendly individual; thus this segment will receive different (say) emails than those of other segments.
- **Concentrated targeting** is when a firm focuses on a single market segment. For example, GMC Hummer focus is on those who favor rugged country and off-roading.
- **Undifferentiated targeting**, which is not recommended, means firms ignore the data on segmentation and offer the same marketing effort to everyone in the population.

Positioning

Positioning is a marketing strategy that establishes the way a customer perceives a product or firm relative to the rest of the marketplace.

This is accomplished by changing the marketing mix – what the product is like, what its price is, where customers purchase it, and how it is promoted.

Positioning is distinct from targeting. Targeting is identifying which segment(s) to pursue, and positioning is the way a business tries to affect the way people in the targeted segments perceive it by changing its marketing mix.

Segmentation, Targeting and Positioning Implementation

The 6 Steps for STP Implementation

There are six steps to implementing STP using data.

The steps can be accomplished using a number of data analytics techniques.

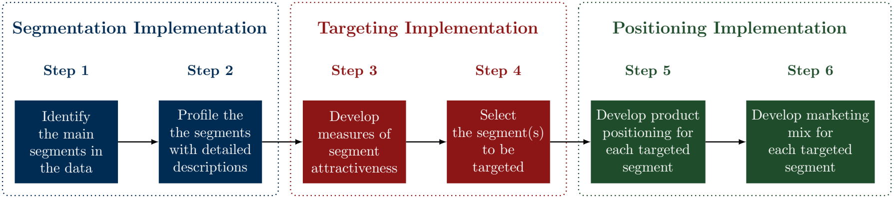

Figure 29: STP Implementation Steps

Identifying Segments by Statistical Learning

Step 1

Statistical learning refers to a large set of supervised and unsupervised tools for understanding data.

1. **Supervised statistical learning** involves building a statistical model for predicting, or estimating, an output based on one or more inputs.
   - Problems of this nature occur in fields as diverse as astrophysics, business, economics, medicine, and psychology.
   - Examples of methods include:
     1.1 Linear regression
     1.2 Logistic regression
     1.3 Naive Bayes classifiers
     1.4 Kernel regression

Supervised Learning Methods Continued

1.5 Local polynomial regression

1.6 Regression splines

1.7 Smoothing splines

1.8 Generalized additive models[12]

1.8 Penalized regression[13]

1.10 Decision trees

1.11 Random forests and other ensemble model approaches

[12] Kernel regression, local polynomial regression, regression splines, smoothing splines and generalized additive models are **semiparametric regression** approaches. For additional information about semiparametric regression, see **?**.

[13] Examples include ridge regression, lasso regression, and elastic net regression. For a discussion of these methods see Zou and Hastie (2005).

SVM and Neural Networks

1.12 Support vector machines (SVM)

1.13 Neural networks, also known as artificial neural networks or simulated neural networks[14].

[14] The four basic neural network specifications are multilayer perceptrons, convolutional neural networks, radial basis functional neural networks, and recurrent neural networks. A neural network with multiple hidden layers and multiple nodes in each hidden layer is known as a **deep learning system** or a **deep neural network**. The word “deep” in Deep Learning refers to the number of hidden layers – depth of the neural network. For an introduction to the most basic network structures, see James et al. (2021). For a review of deep learning concepts, architectures, challenges, applications, and future directions see Alzubaidi et al. (2021).

Unsupervised Learning

2. With **unsupervised statistical learning**, there are inputs but no supervising output; nevertheless we can learn relationships and structure from such data.

   - Unsupervised learning is a set of statistical tools intended for the setting in which we have only a set of **features** $X_1, X_2, \ldots, X_p$ measured on n observations.[15]
   - We are not interested in prediction, because we do not have an associated **response variable** Y. Rather, the goal is to discover interesting things about the measurements on $X_1, X_2, \ldots, X_p$. Two unsupervised learning types are:

     2.1 **Principal components analysis** (PCA) is a tool used for data visualization or data pre-processing before supervised techniques are applied.

     2.2 **Clustering** is a broad class of methods for discovering unknown subgroups in data, such as in STP.

[15] A feature may also be referred to as an attribute, characteristic or variable.

Fundamentals of Clustering

When we cluster the observations of a data set, we seek to partition them into distinct groups so that the observations within each group are quite similar to each other, while observations in different groups are quite different from each other.

To make this concrete, we must define what it means for two or more observations to be similar or different. The method chosen to conduct clustering yields specificity on similarity.

There are many cluster analysis algorithms. The two best-known clustering approaches are K-**means clustering**, where we choose the number of clusters, and **hierarchical clustering**.

In hierarchical clustering, we do not know in advance how many clusters we want. Thus a tree-like visual representation of the combining of observations into clusters is produced. This visualization is called a **dendrogram**.

A Dendrogram

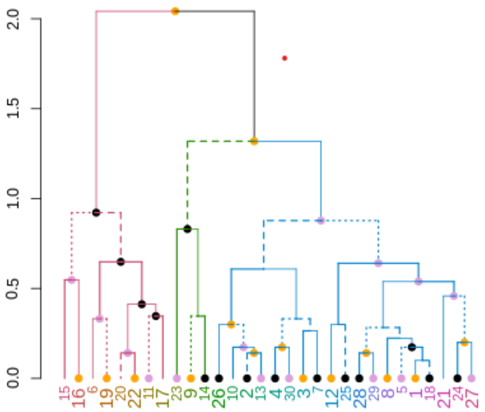

Figure 30: A Dendrogram

The distance between clusters is on the ordinate and the observations on the abscissa.

K-means clustering is a simple and elegant approach for partitioning a data set into K distinct, non-overlapping clusters.[16] To perform K-means clustering, we must first specify the desired number of clusters K; then the K-means algorithm will assign each observation to exactly one of the K clusters.

Let $C_1, C_2 \ldots, C_K$ denote sets containing the indices of the observations in each cluster. These sets satisfy 2 properties:

1. $C_1 \cup C_2 \ldots \cup C_K = 1, 2, \ldots, n$. That is, each observation belongs to at least one of the K clusters.
2. $C_i \cap C_j = \emptyset, \ \forall i \neq j$. Thus the clusters are non-overlapping (i.e., no observation belongs to more than one cluster).

[16] The source for much of the information that follows on clustering was sourced from Hastie et al. (2009). Another excellent introduction to clustering is Everitt et al. (2011).

Determining "Good" Clustering

The idea behind K-means clustering is that a "good "clustering is one for which the **within-cluster variation** is as small as possible.[17]

The within-cluster variation for cluster C_k is a measure $W(C_k)$ of the amount by which the observations within a cluster differ from each other.

Therefore we want to solve the problem,

$$\underset{C_1, C_2 \ldots, C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}. \tag{1}$$

Let x_{ij} be the jth feature of observation (e.g., person) i. The features are used to determine the clusters.

[17] As a secondary objective, we would like to have large **between-cluster variation**.

Specifying the Clustering Objective

To make Equation (1) operational, we need to choose a specification for W. While there are many possible specifications, the most common choice uses **squared Euclidean distance**.

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2, \tag{2}$$

where $|C_k|$ is the number of observations in the k^{th} cluster. Combining Equations (1) and (2), we have,

$$\underset{C_1, C_2 \ldots, C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}. \tag{3}$$

Finding a solution for Equation (3) is a very difficult problem to solve precisely since there are almost K^n ways to partition n observations into K clusters.

Solving the Clustering Objective

The following algorithm is guaranteed to decrease the value of the objective at each step.

1. Randomly assign a number, from 1 to K, to each of the n observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   2.1 For each of the K clusters, compute the cluster **centroid**. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.[18]
   2.2 Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

[18] More generally, a centroid is the center of a geometric object's internal mass, assuming uniform density.

Choosing the Number of K-Means Clusters

Determining the number of clusters is a hard and important problem. There is no efficient and universal method for identifying the initial partitions and the number of clusters K.[19]

Managerial Approach to Choosing the Number of Clusters

1. Determine the maximum number clusters that could be targeted. This is a function of company resources.
2. After conducting a K-means cluster analysis, consider the proportion or number of people per cluster. If the proportion or number is too small to target, than reduce the number of clusters.
3. Ensure there is sufficient differentiation between segments. Thus measures of between segment variation are important to review.

[19] For a survey of a few statistical methods used to determine the number of clusters, see Chiang and Mirkin (2010).

Sidebar About PCA and Clustering

PCA plots are frequently used to find potential clusters by simplifying the complexity in high-dimensional data while retaining trends and patterns.

This reduction in the dimensionality of a data set is accomplished by linearly transforming the data into a new coordinate system where (most of) the variation in the data can be described with fewer dimensions than the initial data. That is, PCA reduces data dimensionality by geometrically projecting it onto lower dimensions called *principal components* (PCs).

- All PCs, linear combinations of the original variables, beyond the first are required to be uncorrelated with all previous PCs.
- PCs are very useful for clustering in the presence of the **curse of dimensionality**.[20]

[20] For additional information on using PCA to detect the number of clusters see Lever et al. (2017).

The Curse of Dimensionality

High dimensional data refers to data with a large number of features, variables or dimensions often represented by the columns in a data set, where each row is an instance or observation. Frequently, the number of variables (columns) can exceed the number of observations (rows).

The term "curse of dimensionality" refers to the difficulty of dynamic optimization with many variables.[21] Broadly, the following issues are faced when working with high dimensional data.

1. Working with large dimensional data is computational challenging. The processing and storing of high dimensional data require substantially more computational resources.

[21] The curse of dimensionality may be restated as, "As the number of features or dimensions grows, the amount of data we need to generalize accurately grows exponentially." The term "curse of dimensionality" was coined in Bellman (1957).

Accuracy Degradation & the Curse of Dimensionality

2. Working with large dimensional data is computational challenging. The processing and storing of high dimensional data require substantially more computational resources.

3. Large dimensional data is very hard to visualize and interpret.[22]

4. Generally as number of features increases redundancy also increases – more noise is added to data than signal. This results in degradation of performance of all analytic methods.

[22] PCA assists on overcoming this issue.

A Curse of Dimensionality Example

Suppose during the last 2 years a B-to-B construction equipment and material supplier serviced 1,024 customers.

The company's marketing department decides to partition customers based on purchase recency in days, say by creating a variable defined as 1 divided by the number of days since last purchase, R, and 2 year total monetary value, M, using quartiles. Thus 25% of customers will have a recency between $Q_{0,R}$ (i.e., the minimum of R) and $Q_{1,R}$, 25% will have a recency between $Q_{1,R}$ and $Q_{2,R}$, ..., and 25% will have a a monetary value between $Q_{3,M}$ and $Q_{4,M}$. There are $4^2 = 16$ R and M combinations.

If the 1,024 customers are randomly distributed among each R and M combination, then on average there are 64 customers within each possible combination, which is a good enough sample to draw conclusions. If for example customer i, $R_i \in [Q_{0,R}, Q_{1,R}]$ and $M_i \in [Q_{0,M}, Q_{1,M}]$, we may infer the customer has attrited.

The Curse of Dimensionality Example Continued

Now suppose the marketing department decides to expand the analysis to make it a Recency-Frequency-Monetary Value (RFM) analysis supplemented by the number of items purchased by the customer. If frequency F and number of items I are partitioned into quartiles, then there are $4^4 = 256$ combinations.

Assuming again customers are randomly distributed across the 4 combinations, the average number of customers per combination is 4.

If the analysis is expanded from 4 variables to 8 variables and quartiles are still used to define partitions, then there are $4^8 = 65,536$ combinations with an average number of customers per combination equal to 0.0156. This means that almost all possible combinations are never observed, and the company has realized the curse of dimensionality.

Profiling Segments

Step 2

Once segments have been determined, we will want identify which people are in the segments by **profiling** them.

A simple profiling method is calculate average characteristics (e.g., age, income, . . .) and proportions (e.g., gender, lifestyle segments, life-stage segments, . . .) for all people assigned to a particular segment.

Once this has been done, it is customary to name the segments. Such as *Fashionista*, *Deal Shopper*, *Premium Buyer*, . . .

Segment Attractiveness and Selecting Segments

Step 3

Next, we would like to know which segments are most attractive according to criteria such as past sales. We can also compare means on marketing outcome variables like customer satisfaction, loyalty, revenue or profits.

The process is similar to that of profiling the segments.

Step 4

The marketing analytics professional work with management to decide which segment(s) to target based on business objectives.

Product Positioning and Marketing Mix

FRANKLIN UNIVERSITY
Makes It Possible

Step 5

Once we have identified the targeted segments, we will need to determine how to position the company differently across segments.

A simple means profiling analyses on marketing mix variables could be useful to learn about what each segment might value in the company's positioning.

Step 6

The final step in the STP is to use the data on product positioning for each targeted segment to develop a marketing mix plan for each target segment using the 4 or 7 P's.

References

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaria, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(53):1–74.

Barron, K., Kung, E., and Proserpio, D. (2020). The effect of home-sharing on house prices and rents: Evidence from Airbnb. *Marketing Science*, 40(1):23–47.

Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ.

Brodeur, A., Clark, A. E., Fleche, S., and Powdthavee, N. (2021). COVID-19, lockdowns and well-being: Evidence from Google Trends. *Journal of Public Economics*, 193:104346.

Cebrián, E. and Domenech, J. (2022). Is Google Trends a quality data source? *Applied Economics Letters*, 0(0):1–5.

Chiang, M. M.-T. and Mirkin, B. G. (2010). Intelligent choice of the number of clusters in k-means clustering: An experimental study with different cluster spreads. *Journal of Classification*, 27(1):3–40.

Choi, H. and Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, 88(s1):2–9.

Davis, B. (2022). *Marketing Analytics*. Edify Publ., ISBN: 978-1-7346888-4-9.

Du, R. Y. D., Hu, Y., and Damangir, S. (2015). Leveraging trends in online searches for product features in market response modeling. *Journal of Marketing*, 79(1):29–43.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley, West Sussex, UK.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, New York, NY.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, NY, 2nd edition.

Jun, S.-P., Yoo, H. S., and Choi, S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological Forecasting and Social Change*, 130:69–87.

Karr, A. F., Sanil, A. P., and Banks, D. L. (2006). Data quality: A statistical perspective. *Statistical Methodology*, 3(2):137–173.

Lever, J., Krzywinski, M., and Altman, N. (2017). Points of significance: Principal component analysis. *Nature Methods*, 14(7):641–642.

Robin, F. R. (2018). Use of Google Trends data in Banque de France monthly retail trade surveys. *Economie et Statistique / Economics and Statistics*, (505–506):35–63.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320.