# R for Customer Analytics Data Preparation

Steven Ludwig

14 February 2025

# Contents

# Title Page

Customer Analytics Data Preparation

Steve Ludwig Miami, FL, USA steven.ludwig@u.northwestern.edu[1]

[1]mailto:steven.ludwig@u.northwestern.edu

# Book Update Status (-)

This book is a work in progress. Sections marked with a "-" will be receiving more updates.

# Preface

William Bernbach once said that "advertising is fundamentally persuasion and persuasion happens to be not a science, but an art (Ignatius, n.d.)." In 2008, I had a choice how I would advertising communications. I was accepted to Northwestern University and University of Texas. Northwestern represented the scientific option and Texas represented the creative option.

I pursued the science route by enrolling in the Northwestern MSIMC program.

Upon enrollment, I was overly enthusiastic about how my economics education (B.A. Univ of Michigan) would fit in to the IMC curriculum. Soon, in the first semester, I noticed IMC did not incorporate much economic theory. I also got the feeling the scientific part of the program was a little overstated.

Here are some internal conflicts I had between Northwestern's vision of IMC and the actual marketing industry.

- Northwestern stressed the science of advertising, which included digitization. However, industry leaders came in to class stressing the importance of creative in advertising and dismissing digital advertising. Back in 2008, no one saw digitization coming! Back in 2008, no one thought AI could take over creative.
- Professors "propped up" student egos and really sold IMC as the "next trend" in marketing. However, the US was in a recession and hiring had decreased. Marketing leaders would not take a risk trying new approaches to marketing.
- Most professors were adjuncts and not fully pushing the science part of advertising. Courses were taught in SPSS and Excel, not in the scientific statistical technologies.

With self-initiative, I was able to customize my grad-school experience to get the "science" that I needed. It all changed when I took a customer-lifetime value class that taught using the R programming language along with the SAS language. Finally, I saw the "science" part of the MSIMC program. I was fully

committed and spent every moment getting my taste of "customer-analytic" and "science-based" marketing. In fact, that course has motivated me to write this book!

Unfortunately I became too committed to the science part of marketing and spent the 2010's gaining experience after graduate school.

### The Experience Decade

Anders Ericsson, professor of psychology at Florida State University, concluded that it takes 10,000 hours of intensive practice to achieve mastery of complex skills (Ericsson, Krampe, and Tesch-Romer 1993). (Most people have probably heard of this by reading Outliers by Malcolm Gladwell). Assuming one can devote 3 hours per day that's about 10 years. Completing grad school was just the beginning of that journey.

In 2010, I earned my degree and searched for jobs at advertising agencies, in Chicago. I entered the job market overconfident, committed and over educated. I had tons of help and connections — Advertising agencies were eager to talk to a "Northwestern" guy. Once in interviews, they were not eager in any of my marketing thoughts or strategies. I also started to notice strong disconnects between IMC theory and application.

- Agencies still used pivot tables and Excel; Statistical-programming met nothing in the agency world.
- Agencies backed down to client requests and did not consult on strategy. (This is before consultancies got into the game)
- Salaries started very low.

None bought into my approach so I entered the workforce doing marketing communications analytics for non-profits. I entered the 'fundraising' field doing fundraising reporting and analytics. This industry is all about people talking to rich donor-prospects and solicititing funds for Universities. I found that environment to be perfect for brushing up technical skills. I continued learning "R" I focused on data and learned reporting technologies (SQL) and also invested in learning programmatic approaches to statistics instead of graphical (SPSS/EXCEL) . - companies were still using Excel and pivottables for analysis. I lacked experience in Excel - i also lacked experience working in sql databases

- Took nonprofit jobs where i could gain more experience in SQL
- Gained more analytics experience in R by working for a mid market analytics company.
- Gained critical SalesForce experience by working in the banking sector.

**Lessons Learned**

One lesson I learned is that academia is usually ahead of industry and academia is idea-driven instead of results-driven.

Ideas have little value early in the career. Value is created by doing and doing encompasses implementing the right technologies at the right time while navigating company politics. Only after mastering data could I pitch-in and apply the techniques learned in graduate school.

Second lesson -

— So it takes time for new marketing approaches and methods to diffuse into industry. A second lesson: The skills taught at Northwestern failed to compensate for career experience

And then going back to that 10,000 hours of practice thing. One part left out of Outliers is that practice is not just about the 10,000 hours. It has to be "Deliberate practice" which depends on having good teachers. Another lesson learned is that Northwestern's IMC program did successfully predict the future – they had good teachers (practitioners). I just bought into to fast and to aggressively. In other words, Academia is usually right in the approach but wrong in the timing. In fact, Northwestern's IMC program has earned a STEM certification and has doubled down in the scientific approach to advertising. Even with this scientific focus, few people know what a MSIMC degree is and that leads to another lesson learned. I just describe my experience as marketing communications analytics instead of going into the details of the classes. Using phrases like database marketing, customer valuation and .. () just confuses marketers and it is best to be vague.

I have now reconciled my marketing analytics skills with practical data management skills. Currently, I am a customer analytics strategy manager at dentsu's Merkle division. There, I focus on CDP (customer data platform) use cases and act as a CDP manager for clients. I am currently working with banking and telecom industries.

Formerly, I have worked for a variety of organizations in CDP's, sales analytics, lead management and customer retention. I've worked in streaming media, banking, non-profits, customer-data-platforms, and startup sectors. I learned that the hardest part of analytics is the data preparation, particularly launching customer data platforms (CDP's). In fact, I have experience working across electronics, fashion, retailing and insurance sectors. And, I've also gained both Treasure Data and SalesForce certifications.

# CH 1 Customer Analytics

## Landscape

Times have changed and the mar tech world is more accepting of programmatic and technical solutions. "Scientific" statistical approaches in R and Python are now commonly used in marketing. My book aims to contribute to the ever growing customer analytics landscape by focusing on customer analytics.

Specifically, my book focuses on customer analytics in the R statistical programming language. More specifically, the book focuses on transaction data preparation. In my view, data preparation is often overlooked in marketing. It is often outsourced or automated and not fully understood. Since I've went through the difficulty learning data preparation I sincerely hope this book helps aide and ease others' journeys.

### Industry definitions:

### Websites and books:

**Wikipedia:** "Customer analytics is a process by which data from customer behavior is used to help make key business decisions via market segmentation and predictive analytics. This information is used by businesses for direct marketing, site selection, and customer relationship management (Wikipedia contributors 2023)."

**Customer Analytics for Dummies**: Jeff Sario defines customer analytics through the following activities (Sauro 2015):

- Gathering data
- Using mathematical models to detect patterns
- Finding the insight
- Supporting decisions
- Optimizing the customer experience
- Mapping the customer journey

He also defines the difference between customer analytics and other business metrics:

- Customer focused - data comes from customer actions or attitudes
- At the individual customer level - data is event and transaction not product or company level data
- Longitudinal - looks at customer behavior over time
- Behavioral and attitudinal - mixes customer action and customer thinking

## Software Companies

**ADOBE**: Adobe does not provide a definition of customer analytics. However, an Adobe-commissioned Forrester report provides context into their views of customer analytics. First, all of Adobe's marketing products are categorized under the Adobe Experience Platform (AEP). Second, this report talks specifically about the Real-Time Customer Data Platform, Journey Optimizer and Customer Journey Analytics (CJA). According to the report, "67% of business and technology decision-makers are in the process of adopting data capabilities to build/improve a complete view of their customer across channels." The total economic impact of these products can result in a return on investment. (McNair and Serradilla Ortiz 2023)

It's implied that Adobe places a strong emphasis on unified data and communicating with customers across channels. More specifics mentioned in the report include:

- "streamline collection, management, and action on customer data at scale."
- "real-time profiles and audiences"
- "enabling a holistic view of customers' journeys while providing flexibility and speed to transform data on the fly for analysis."
- " real-time, cross-channel customer journeys informed by customer actions, delivering personalized experiences with the support of collected intelligence and insights*

Economic benefits of the Adobe "customer analytics" solution

- Productivity analyzing customer data
- Efficiency value in campaign orchestration
- Improvement of business performance
- Efficiency in ad spending
- Avoiding technical debt

**SAS**: "Customer analytics refers to the processes and technologies that give organizations the customer insight necessary to deliver offers that are anticipated, relevant and timely (SAS 2023)."

**Salesforce**: Defines 'customer relationship management' as: A business strategy for developing and improving relations between companies and their customers (Salesforce, n.d.a).

**Treasure Data**: Treasure Data defines itself as a customer analytics solution by being named a strong performer in the The Forrester Wave™: Customer Analytics Technologies, Q2 2022 study. From their CEO, "We're proud to be the only customer analytics-focused CDP designated as a Strong Performer by Forrester, which further validates, in our opinion, that our current business model, operational strategy, and product roadmap are on track and are propelling the company to greater heights as its maturation continues in a highly-competitive industry category ("Treasure Data Named as a Strong Performer in Customer Analytics Technologies Evaluation by Independent Research Firm" 2022)."

## Consulting & Research Companies

**McKinsey**: "Customer Lifecycle Management" "By using proprietary customer data and analytics, clients can acquire, develop, and retain high-value customers more profitably and effectively—and drive the necessary organizational changes from the C-suite through to the front line." ("Customer Lifecycle Management Growth, Marketing and Sales," n.d.)

**PWC**: Customer Insights Platform "processes big data, identifies repetitive customer behavioral patterns to give you a complete view of your customer base." (Ignatius, n.d.)

**Gartner**: "Customer analytics is the use of data to understand the composition, needs and satisfaction of the customer. Also, the enabling technology used to segment buyers into groupings based on behavior, to determine general trends, or to develop targeted marketing and sales activities(Gartner, n.d.)."

**KPMG**: "KPMG Customer Advisory can help you transform the way your front office engages customers. Combining business experience with functional acumen, we provide you with deep financial analysis, robust customer insight, market intelligence, and strategic business direction to help you generate ROI from your investments in customer-centricity (KPMG, n.d.)."

## Academic definitions

**Northwestern/Kellogg**:

- According to Florian Zettelmeyer, customer analytics is the process of transforming [customer] data into knowledge [CITE].

- Kellogg offers Customer Analytics and AI course that details the scientific approach to marketing with hands-on use of technologies such as databases, analytics, machine learning, and computing systems to collect, analyze, and act on customer information [CITE].

**USC**: "Customer Analytics uses vast amounts of data to generate insights that help firms and policy makers to make data driven decisions about customers. Insights into customer behavior, attitudes, demographics and psychographics can prove to be a sustainable advantage for firms ("Program: Customer Analytics Minor - University of Southern California - Acalog ACMS™," n.d.)."

**V. Kumar (Georgia State University) J. Mack Robinson College of Business**:

V. Kumar (Regents Professor; Richard and Susan Lenny Distinguished Chair & Professor in Marketing; Executive Director, Center for Excellence in Brand and Customer Management, Georgia State University, J. Mack Robinson College of Business) defines customer analytics as a part of analytical customer relationship management. Customer analytics methods he covers in his book include (Kumar and Reinartz 2012):

- Traditional Metrics: market share, sales growth
- Customer Acquisition Metrics: acquisition rate, acquisition cost
- Customer Activity Metrics: average intra-purchase time, retention/deflection rates, survival rates, lifetime duration, probability of being active
- Customer-Based Value Metrics: size of wallet, share of category requirement, share of wallet, transition matrix, RFM, lifetime value, customer equity

Since the title of his book is customer relationship management, one can conclude that he views customer analytics as a part of customer relationship management.

## Role of Data in Marketing

According to McKinsey, data-driven marketing is the new normal [2].

Precision marketing models applied to transaction data may enable the following:

[2] https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/the-big-reset-data-driven-marketing-in-the-next-normal

- Draw inferences from behavioral patterns. An algorithm might learn, for instance, that customers who make more than two visits to a store's website within a two-week period are 30 percent more likely to make a purchase."
- Tease out salient behavioral indicators in time to act on them, marketers need continually refreshed data from a variety of sources and at a far more detailed level—looking as deeply as the city-block level in some cases
- "Companies that extend their data gathering in these ways can identify upticks in demand and where new customers are coming from, as well as assess which customers in their existing base have increased spending and where lapsed customers have gone."
- "But AI-enabled monitoring can do this in minutes, sometimes seconds."
- "By capturing new data, searching for new behavioral relationships, and enabling rapid experimentation, marketers can seize granular growth opportunities and enter the recovery with significantly greater ROI and resilience."

"To stay ahead of competitors, every year corporations invest massive amounts of money to gather information about their customers–and determine the best way to capture a larger market share."

## Types of Marketing Data

Besides transactional data, other types of marketing data may include:

- AI / CHAT-gpt interactions
- Push-button feedback (think airport bathrooms)
- Internet product ratings
- Focus groups
- Web site browsing data a.k.a. "streaming data"
- Demographic data

| Qualitative Marketing Data | Quantitative Marketing Data |
|---|---|
| Focused on consumer behavior, surveys, consumer experience. | Focused on data, reporting and modeling. |

### Customer Communication (-)

### Customer Retention (-)

## Relating Marketing Investments to Financial Outcomes (-)

Started with: Managing Customers as Investments

*Marketing data is especially important to bridge the gap between communication investments and expenditures.*

Increasing attention from Accounting Department

** 1.) Fader/ CBLV driving interest financial valuation of companies **

Continuing with work in CBLV:

Marketing tends to attract lots of attention in finance departments because of the high resources and budgets that are requested each year. Most finance departments are data-driven and like to see the ROI of marketing efforts. The role of data in marketing is growing through the CBCV trend. This stands for customer-based corporate valuation. The idea is that the value of a company is the sum of the value of the company's customers.

CBCV Resources

- Managing Customers as Investments ("Managing Customers as Investments The Strategic Value of Customers in the Long Run [Book]," n.d.)
- Morgan Stanley https://www.morganstanley.com/im/en-us/individual-investor/insights/articles/the-economics-of-customer-businesses.html
- Theta CLV https://thetaclv.com/ ("Customer-Based Corporate Valuation" 2023a)

** Marketing department and IT department partnerships: data architecture **

Banks would not handle credit card transactions without a data architecture and data engineering department. Marketers should not either. It is no longer necessary for marketing practitioners to 'go it alone' in their data preparation and modeling. Following my approach, A long term and sustainable approach to customer data management can be achieved.

** Customer Analytics in Performance Measurement and Reporting Systems (Bonacchi and Perego 2023) **

[[NEEDS CONTENT]]

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4616541

4.) Financial effects of marketing [DeKimpe]

COVID 19 brought back interest in these types of analysis. The idea is that a shock happens (e.g. Covid-19) and then various actions are measured. This is also called VAR modeling which is a form of timeseries analysis with multiple timeseries.

## Where Companies Struggle in Customer Data Management

Most companies try to jump right into to customer analytic modeling, but can trip and fall in the details of transactional data. (If you have a customer data platform you have likely worked through these scenarios, if not these are some potential stumbling blocks that you will encounter when preparing transactional data.)

| Transactional Detail | Potential Stumbling Blocks |
|---|---|
| What transactions to include? | SAP transactional values<br><br>Excluding returns |
| Which product categories to include? | Accounting for product ownership in different product divisions. |
| How should returns be handled? | Filtering returns out of transactional data |

| Transactional Detail | Potential Stumbling Blocks |
|---|---|
| How do I connect customers to transactions? | Limitations of information gathered during the transaction.<br><br>Or using probabilistic models that do not need detailed customer information. |
| Which field in my transactional database best represents the purchase date? | E.g. shipping date or purchase date |

All Examples have nothing to do with marketing analytics and everything to do with reporting. Yet, reporting folks understand data preparation NOT marketing analytics:

- Integrating Transactional data into customer data platforms. Marketing departments have trouble communicating with SAP teams and designating which types of transactions are valid transactions. Companies can also have trouble figuring out what transactions count as purchases versus which count as returns, etc
- Matching customers to transactions: usually done by email address
- Creating data feeds that can be used for customer analytics
- Dealing with countries and different business organization units. Mapping transaction rates. Figuring out the common value of reporting and exchange rates.
- Including various transactions in various departmental reports. See potential scenarios below.

```
|              | Product Category A | Product Category B |
|--------------|--------------------|--------------------|
| US Region    |                    |                    |
| LATAM Region |                    |                    |
| Asia Region  |                    |                    |
```

Items that need worked out with reporting department.

- Getting a feed of transactional data that includes

  - only relevant transactions
  - detailed information with detailed columns
  - only relevant columns

- Having access to a table that translates product UPC codes into product categories
- Having columns that marketing analysts can use to filter data. Potentially "re-coding" ambiguous items in SAP.

## The "CADF" Approach to Customer Data Management

The premise of this book (and project) is that good customer analytics should focus on two data points to start. So what are the inputs?

First, is a customer identifier.

Second, is a transaction date.

From those inputs, insight creation around customer purchase timing and purchase behavior can begin.

Users following along in R will want to download and install the R "CADF" package that accompanies this book. This package is hosted on Github. Please enter the commands below to download my R package.

```
#library(devtools)
#install_github("stevenludwig10/CADF")
library(CADF)
```

The CADF "customer analytics data formatting" R package[3] is the bridge between raw transaction data and analytic insights. It implements successful data preparation approaches for customer analytic models. CADF is also a method for working with customer data. Users following along in this book should gain ideas how to implement this approach in their own software environment.

CADF also focuses on mathematical formulas needed for customer analytics models. CADF "Customer Analytics Data Formatting" brings the focus back to transaction data so companies can create a long-lasting framework and build

- various analytic models across contractual and non contractual situations
- simple variables like purchase strings that can be integrated into other reports

[3] https://github.com/stevenludwig10/CADF

- analytic models for various departments, cohorts and scenarios
- analytic models that has data vetted by accounting and investment companies
- analytic models that focus on what transactions actually happen (big difference in what customers say they will do versus their actual behavior)

With CADF process you can bring marketing analytics closer to managers, who are used to utilizing one-off Excel models for quick decision making. Following this process you will have cleaner transactional data and be ready for a variety of customer analytic models.

My approach separates transaction data from analysis in the following ways.

- By creating a single output that is compatible with many analytic models.
- By running data for multiple analytic scenarios.

The approaches used in the development of CADF R package can be used to implement customer analytics on multiple platforms — SAS, Microsoft Excel, Excel Online, Informatica, R, Python, etc.

## Why the Marketing Industry Needs a Customer Analytics Data Format

- Increasing trend toward customer profitability. (Wikipedia contributors 2023)
- Customer profitability is an increasing interest of investment firms, private equity and auditing firms. (I will detail these shortly.)
- Marketing now focuses on statistics around contact points and consumer views & reach. Better productivity metrics are needed to measure the cost-side of marketing.
- Introduction of cohort-based segmentation versus micro-segment.
- Combat technical challenges and bring together an approach that harnesses IT skills and marketing modeling skills.
- Enable automation of grunt work so Marketing analysts can easily try different modeling scenarios.

### Challenges

Technical challenges in customer analytics include:

- Challenge 1: Combining customer attributes with transaction data.
- Challenge 2: Having a standard "customer analytic" format that allows data exchange between programs (Excel, R, SAS, etc)
- Challenge 3: Enabling more efficient processing of customer data in the cloud.
- Challenge 4: Proactively running datasets for multiple scenarios. Instead of specific analysis questions. (Avoids ad-hoc data sets laying around.)

## Company Approaches to Analytics

Instances where I have seen proactive analytics efforts placed behind daily reporting tasks.

- Banking company that put efforts to design a "prospect" database behind Salesforce journeys and retail promotions.
- Organization that showed no interest in mapping the growth of user adoption on a newly designed social media platform.
- Etc, [These may need generalized or cleared]

#### Approach A: Reactive

Starts with specific business division-> business division defines metrics -> analyst gathers data directly and specifically for business division and region.

#### Approach B: Proactive

Analytics planned before need exists -> Analytics is centralized -> datasets exist that can answer variety of questions -> support exists for data maintenance

#### Characteristics of Each Approach

| Characteristic | Approach A | Approach B |
|---|---|---|
| Level of investment | Low investment, utilizes existing resources | Higher investment, new resources to manage analytic inputs |

| Characteristic | Approach A | Approach B |
|---|---|---|
| IT <-> Marketing Department Coordination | Not necessary | Increasingly necessary |
| Level of centralization & control | Department has freedom to design analytical approach and gather own data. | Projects have to be within rail-guards of available system and or analytic datasets. |
| Has analytics department | No | Yes |
| VP + level leadership and potential committee collaboration | | Required on business-side, on marketing-side and on IT/technology side |
| Analysis " repeatability" for other countries and divisions | | |
| Marketing Technology Stack & Environment | Likely in earlier stages of customer database design. | Likely has invested in a centralized customer database and/or customer data platform |
| Adding data and upgrading analysis | Easily done | Will likely require entering tickets and/or work with department design team. |

## The "2 Data Point" Approach to Customer Analytics

The toughest part of any analytics effort is deciding where to start. Most analytics professionals will start with a question then prepare data for the analysis. Customer data techniques are often ignored in return for brute force coding and analysis.

Downsides to this approach:

1.) Most "questions" involve certain customer segments. Analysts cobble together Excel spreadsheets but the methods are very tied to the scenario. It is not repeatable across different segments. Marketers often want to see information for different segments.

2.) Organizations outsource analytics and the data preparation proceses are not internalized or documented. (Leaves messy SQL code that company analysts are not able to follow.)

3.) Most analytics approaches cover the data preparation and modeling for a single approach. E.g. one process for model A and one process for model B. The CADF approach focuses at preprocessing data for multiple modeling techniques.

This book teaches customer analytics by focusing on transaction data mastery first then analytics last.

# CH 2 Data Preparation Notes Before Starting

## Prerequsites

This book focuses on a very specific business use case. Preparing transactional data for customer analytic modeling. Furthermore, it focuses on R as the technology of choice to implement this.

This book assumes knowledge of the R Programming Language (or knowledge how to use R principles in other programming languages). R will be used to demonstrate implementation of a customer analytics data format — referred to as CADF in this book. A transaction dataset is provided and all examples are reproducible. Those with high technical aptitude may be able to implement some of the examples and approaches in their preferred programming language.

Interested readers can utilize one of the many resources listed below... it is assumed that the reader has general knowledge of the R statistics language. This is not an R book but R is used to build out the prototypes and approaches.

- R for Beginners ("R for Beginners," n.d.)
- The Art of R Programming: A Tour of Statistical Software Design (Matloff 2011)
- Understanding Statistics Using R! (Schumacker and Tomek 2013)

### Software Installation

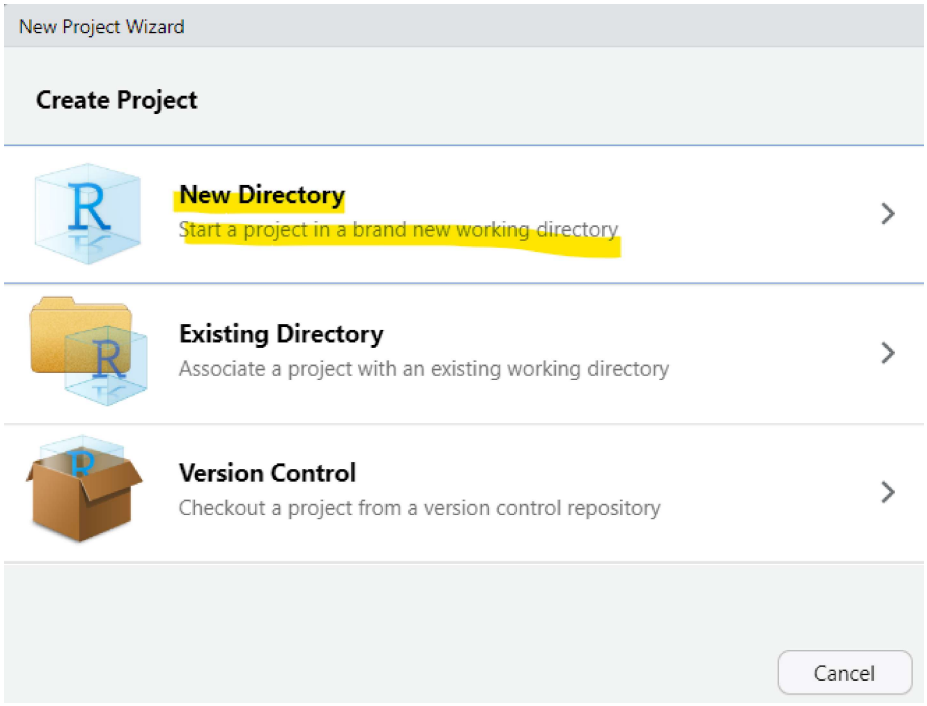
Readers who want to follow along should complete the following steps. These steps install a combination of R! and RStudio Desktop IDE which is the environment I utilized to write this book.

1.) Visit the R-project for Statistical Computing web site to download[4] and install R!

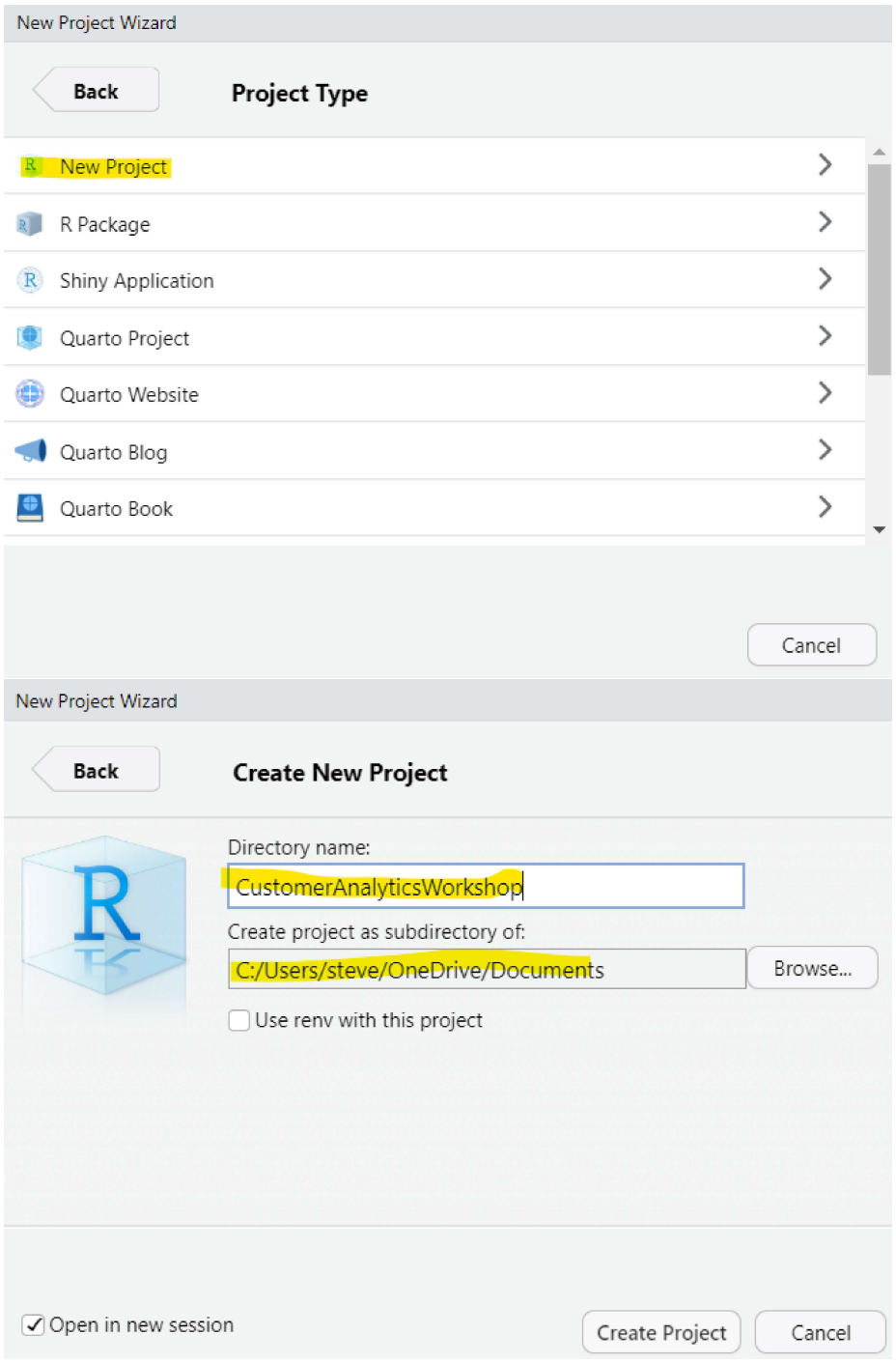
1a.) Mac users will use this link[5] to download R!

2.) After R! installation is completed head to the posit Web site to download RStudio[6]. RStudio Desktop free studio is recommended. Download it and install.

3.) Launch RStudio and click File -> New Project ...

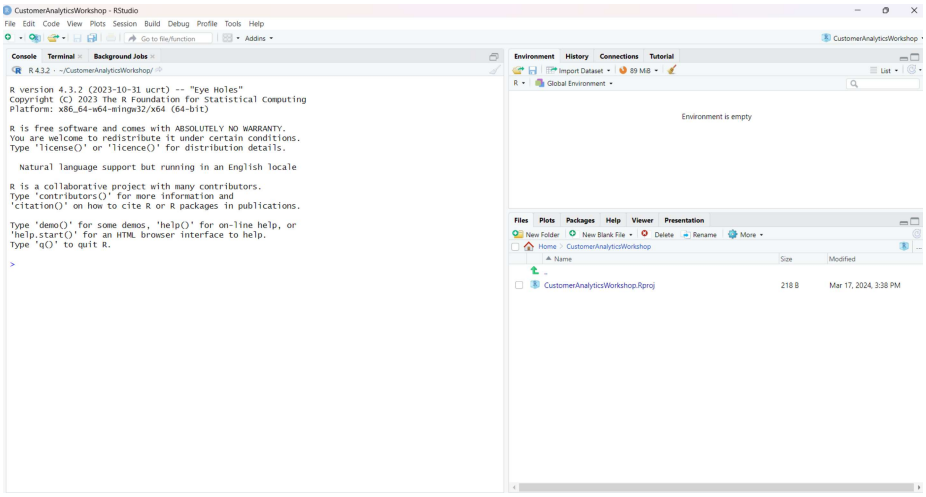



[4]https://cran.r-project.org/bin/windows/base/
[5]https://cran.r-project.org/bin/macosx/
[6]https://posit.co/downloads/




4.) Call the project CustomerAnalyticsWorkshop and save to your preferred

directory. This keeps all your learning under one project and saves R data while working.




5.) This book comes with R code. To obtain the R code, visit the following link[7] and download the following file _____.

6.) To install the R code in RStudio, click on Tools -> Install Packages ...

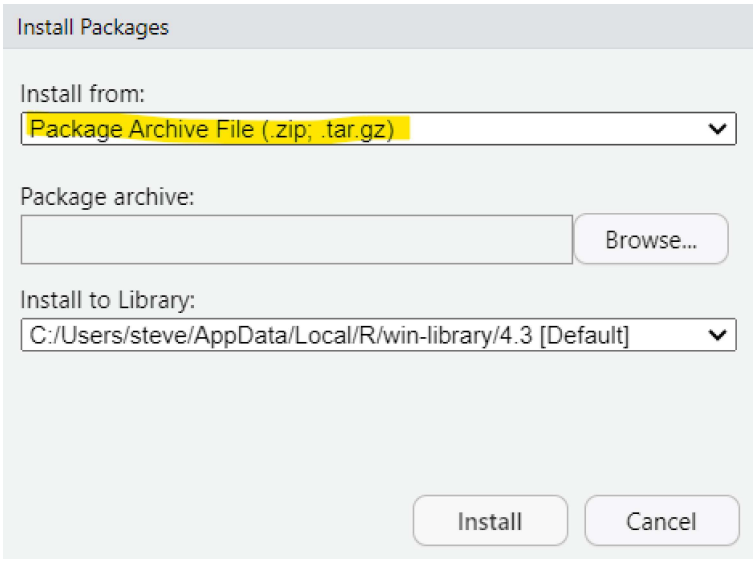



Figure 1: Installing CADF Package through Downloaded File

[7] github.com

# Data preparation concepts

Once the steps above are completed you are ready to begin the journey of data preparation in customer analytics. This book will teach data preparation through statistical programming techniques, focusing on R software.

But first, I will want to differentiate between technical and conceptual data preparation concepts. Data preparations are either technical having specific code and settings to follow. Sometimes, however, data preparation concepts are conceptual in nature. These conceptual methods are best practices based on years of working with customer data.

I define technical data concepts as programming techniques that are specific to data analysis. What makes these concepts technical is that they are implemented through code, instead of point and click. For example, doing pivot tables in Excel might not be technical. But, doing data aggregation in R would be technical. Why? Because, statistical programming languages use programming instead of graphical interfaces like Excel.

Conceptual data preparation concepts are those approach that I have developed over the years that provide efficiency in customer analytics projects. This book will focus on conceptual data preparation techniques. Those techniques are skills picked up based on experimentation and experience. I assume the reader is able to use AI & search tools to learn technical data concepts.

## Technical

- Working with dates and calculating time between purchases.
- Accessing / modifying / updating data otherwise known as data indexing in R
- Data frames
- Data aggregation
- Combining data frames using joins.

## Conceptual

- Separating analytics from complex transactional queries. Systems such as SAP have complex rules for accounting for customer returns. Also for matching transactions to customers and matching transactions to purchase categories. To overcome these complexities, focus on 2 data points. Customer ID and transaction date. Fast-forward and experiment with customer analytics and then work with reporting teams who can help refine your input transactional data.

- Use of reproducible analytics. This involves breaking analytics problems into separate tasks using functions.
- Combining aggregate transaction data with customer demographic data.
- Knowing how much transactional data to collect going back X year.
- Knowing when to use event-based cohorts versus time-based cohorts.

## Introduction to Cohorts

Officially, according to Merriam Webster a cohort is defined as "a group of individuals having a statistical factor (such as age or class membership) in common in a demographic study ("Definition of COHORT" 2023)." You may hear the term cohort to describe a class of incoming graduates. Or, potentially to describe a group participating in a medical study.

In marketing, cohorts are defined as a group of consumers that complete a marketing action around a similar time. If subscription services are analyzed it will be a group of consumers who subscribe to a service around a similar time. These cohorts may share similar characteristics such as geographic location or similar product purchase.

Readers should understand cohorts before reading this book. Cohorts are simply those customers in the target segment who are included in the analytics projects. In fact, the word segment may be exchanged for cohort if that is easier for marketing managers to understand.

Another assumption that needs relaxed when analyzing cohorts is the concept of first purchase. The term can be a little misleading based on how marketing departments retain and utilize data. Consider this example: you start to run cohort data for a analytics project in the banking sector. You look at the customer database and see transactions all the way back to 1970. However, your manager says limit transactions to 8 years ago. So, if customer ABC is in your database and makes their first purchase in 1970 and then makes another purchase in 2020 — the 2020 purchase will effectively count as the first purchase.

## Types of Cohorts in Marketing

The term cohort can have various meanings in marketing. We will cover the following types

| Cohort Characteristic | Date-based Cohort | Event-based Cohort |
|---|---|---|
| Cohort inclusion triggered by: | Customer id exists in queried transactions based on date range in query. E.g. all customers making a purchase in January 2020. | Customer id exists in segment that has completed action associated with event. |
| Relevant transactional data to pull. | Pull data based on relevant list of customers. Filter transactional data between X and Y date. | Pull list of customers who have completed the event in question. |
| Relevant date. How T is measured. | Time between first purchase date and purchase date | Time between marketing event and purchase date |
| Time - constant or relative | Rule for first purchase is constant based on first purchase within time frame of relevant transactions selected. | Different customers will have different first purchase dates. |

– Date-based cohort: The statistical factor in common is the first purchase date – Event-based cohort: Consumers in this type of cohort were all exposed to a similar event at the same time. This could have been a targeted marketing message. Or it could have been an internal event. – All-in one large cohort: Treats everyone as one.

## Date-based Cohorts

What is a date-based cohort?

1. Focus on data pull is on a transactional period x to y.
2. Tracks group behavior over time
3. Time is relatively consistent for everyone in the study

How to recognize this request: Manager says — show me all customers who signed up in January 2020 and look at their behavior up to the current year 2023.

- Date of first purchase in the same month (time-based cohort)

- Date of first purchase based on a promotion ran on a SPECIFIC date

1. Your marketing manager has decided to include all customers who made their first purchase in month X and year Z. (The date is predetermined up front.)
2. Query your customer database and get a list of all customer ID's where first purchase date = XX_ZZZZ (YY_MMMM). The query should include customer id and purchase date.
3. Pull all transactions associated with #2. So you will need customer id and transaction date.
4. Group #3 by customer id and sort by purchase date, ascending
5. You now have one dataset for each customer

Date Based Cohort Method 2: Most likely used with startups or new product initiatives

## Event based Cohorts

What is an event-based cohort? Event-based cohorts are usually used in analyzing purchase behavior that occurs more infrequently and occurs with higher-value purchases. You may also see event-based cohorts with cheep items although this is less common and less relevant in analysis. For example, does receiving a free birthday drink at your local coffee shop result in higher customer lifetime-value?

More examples of event based cohorts:

1. Consumer completes a specific action - sometimes called a trigger event
2. Consumer attends or responds to a specific marketing event or message.
3. Reaching a specific purchase milestone. All newly designated "silver club" members this month.
4. Date when consumer has first child

Consumers form an event-based cohorts when they share a common event that happens in a designated time frame. Making a second purchase is one example. Upon triggering this event, the consumer enters this event based cohort with other consumers making their second purchase. In this sense, the consumers share a common purchase behavior. These consumers can be pooled together and analyzed to see what happens after making that second purchase. As mentioned, you can see how this type of analysis would be relevant for large-ticket purchases.

Event-based cohorts are one of the reasons why it is best to analyze "time since activity" as a relative measure versus absolute measures of time. It is a mistake for the analyst to generate date ranges and then bucket consumers into those

date ranges. It is more efficient to compute time, T as number of days then group and count customers by T. In practice cohorts don't fall on neatly organized days. And, especially, event-based cohorts do not fall on neatly organized days.

Event-based cohorts demonstrate the need to measure time as a relative factor. This approach allows more flexibility for event-based cohorts and all-in-one cohorts. Why?. Time never lines up perfect, even with date-based cohorts. Consider this scenario. You launch a promotion on January 1, 2023 the promotion lasts one month. Some customers may sign up on January 1. Some may sign up on January 15. You cannot measure time since January 1st because consumers will sign up at different times. Months since purchase needs to be measured from date of first purchase. Another note is that organizations usually do not know the true first purchase date anyways. Sometimes, organizations only keep 8 years of data. So, first purchase will always be first purchase in the transactional dataset. (Left censoring)

## One large cohort

The one large cohort method is when you pull all of your transactional data and when you compute values of T for all customers. This method can be used to determine general purchase patterns. This method is also most compatible with AI and machine learning. For instance, you could generate purchase strings for each customer. Then, run models to predict which purchase strings correlate with variables of interest.

1. Pulls a list of customer transactions between dates x and y.
2. Group the transactions by customer id
3. Within each group:
    - Sort by purchase date, decending

- Add a column T, computed as (PURCHASE_DATE - min(PURCHASE_DATE) /

  30.

- Create a numeric sequence using values in b. min(T):max(T)
- Create a purchase string from c. If the value of b is in c then Y else N.

4. Create a table in your database with the purchase string.

Later in the book we will cover how to create purchase strings in more detail.

# The Role of "T" in Customer Analytics

One cannot mention cohorts without mentioning time, T. In customer analytics, T represents "time since" activity. T can be measured in months, days or years. The correct way to measure T is to start with days. To get months, divide by 30. To get years, divide by 365.25.

## The wrong way to compute T

When doing customer analytics, most analysts will start by defining time ranges. They will assemble a list of dates then figure out which customers go in which date bucket. This process of creating time buckets and mapping purchases into those time buckets is extremely difficult and prone to errors. In fact, most organizations may have predefined categories for specific duration of customer relationships. These categories will often have fancy names and complex reporting requirements.

Another trap that analysts may fall in involves working with event-based cohorts. Inexperienced analysis may try to mark down the event date then calculate consumer behavior relative to that event date.

The trap specifically is associating an event with a specific date range. So, the analyst will structure a query based on dates instead of based on customer ID's and event flags. Consumers can hit certain milestones on different dates, therefore, it is very important to keep time as a relative measurement. Even if companies have events on specific days, there could be events with different dates that different consumers attend. (E.g.For example, marketing company xyz may run an event in January and one in February. )

## The correct way to compute T

In date-based cohorts, T will represent the first purchase date for a consumer. This first purchase date will be the minimum date value for each customer in the dataset. You can see how the time range that you use to pull transaction date will impact the first purchase date for each consumer. When deploying customer data platforms, most companies will include the past 3-5 years of transactional data. Therefore, in essence, the first purchase is first purchase within that 3-5 year time period.

Value of T = 1 in date-based cohorts: T=1 will represent a consumer's first purchase within the scope of the transactional data pulled for analysis. The maximum value of T will be when the consumer pauses their relationship with your firm.

Right way:

| Month Start | Month End | # Customers |
|---|---|---|
| 1/1 | 1/31 | 200 |
| 2/1 | 2/28 | 100 |
| 3/1 | 3/31 | 50 |

First purchase date will be T=1. T is computed as days between purchase date and first purchase date. Once you have days divide accordingly to get T measured in months (or years, if necessary). T is a relative measure and it is computed for each transaction.

Benefits of this approach:

- Processing is done at the customer level and only for relevant customer id's.
- T is a relative measure (to first purchase) so behavior can be compared along different initatives
- Consumers hit milestone events at different times. So this approach is most consistent with any "Event-based" cohorting.
- Even in "date based" cohort analysis, consumer activity does not fall into neatly organized buckets.

The problem here is that the time is relative to action

# CH 3 - Data Preparation R Techniques

Gartner defines data preparation as "an iterative-agile process for exploring, combining, cleaning and transforming raw data into curated data sets for self-service data integration, data science, data discovery, and BI/analytics ("Gartner Glossary - Data Preparation," n.d.)."

CADF prepares transactional data for customer analytics projects. The input for CADF requires two columns *customer id* and *purchase amount*. There is some minimal data preparation involved to prepare the data for input into CADF.

Note that CADF does not:

- Build transactional data-marts
- Filter what transactions to include
- Complete identity resolution
- Link transactional data to customers
- Select and place consumers into cohorts. (That is done by your transaction data query, described below)

Most organizations will have very clean transactional data due to financial and accounting rules. This data will likely sit in SAP or Anaplan. There will be one row for each transaction id and each item purchased. Probably no customer id but you may have email address or credit card number or some type of identifier.

Key Challenge- finding a transactional dataset with linked customers

The main challenge is linking it to customers. This data will have an unique ID for each transactions. The data will likely have table for each transaction basket (option 2 below) and then a separate table for each transaction item (option 1 below). Another challenge is determining which types of transactions to include in your customer analytic model. This may be done by transaction types or linking transactions to product categories. (Remember to account for abnormalities in the transactional data such as returns.)

| CustomerID | Purchase Date | Customer Email Address | Order Number | Item Number | Item Pri |
|---|---|---|---|---|---|
| 12345678 | 4/1/2023 | customer@company.com | 00000001 | ABC-123 | 100 |
| 12345678 | 4/1/2023 | customer@company.com | 00000001 | DEF-45 | 50 |

| CustomerID | Purchase Date | Order Number | Order Total |
|---|---|---|---|
| 12345678 | 4/1/2023 | 00000001 | 150 |

Option 1 - One row for each customer transaction basket

Simplified Option 2 - order level transactional table

If you do not have these tables it is best to network within your organization and find out who manages the transactional data. Here is my suggested process

1.) If your organization has a customer data platform (CDP) you will likely already have a good data to use that contains customer id's and transactions. Usually organizations will have a CDP system. Most CDP systems have an approach for handling customer data.

CDP Resources

- Customer Data Platform Institute ("(CDPI) Customer Data Platform Institute" 2020)
- Treasure Data ("Product Documentation Home - Product Documentation - Treasure Data Product Documentation" n.d.a)
- Acquia ("Examples of Standard Transaction & Transaction Item Feeds — Acquia Docs," n.d.a)
- SalesForce Data Cloud Model (Salesforce, n.d.b)

2.) If you do not have a CDP you will need to create a input transactional dataset for customer analytics modeling. CADF requires just customer id and purchase date.

2a.) Conduct a discovery process with the IT teams that manage accounting and transactional systems (like SAP). Share project details with IT department and work with data engineering and data architecture departments as early as possible in the planning process. Your goal is to build a relationship with the IT department so they can help you build a raw transaction data that sits between accounting data and customer data. Ask questions to determine what kind of customer information is gathered when a transaction takes place.

Transactional Data Resources:

- Definition of Transactional Data ("Transactional Data - an Overview | ScienceDirect Topics," n.d.a)
- Snowflake Definition of Transactional Data ("What Is a Transactional Database?" n.d.a)

2b.) Also try to learn about any processes that link transactions to customers. For this you may need to involve current loyalty and marketing departments.

2c.) Ask your IT team to build a transactional dataset like the one in step 2 shown above. Option 2 is what CADF expects as input.

3.) For marketing analytics departments that have SQL analyst or technical people ask for Option 1. That's the most flexible then your team can roll up a database for option 2 using option 1 as input.

Once you have access to transactional data you will be well positioned to begin customer analytics.

You do not need transactional data to use this book. I have included the famous CDNOW dataset in the CADF package that can be used for the exercises in this book. Chapter 2 dives into data preparation techniques that are important for any customer analytic projects. Projects may include ad-hoc analytic requests. Or, data departments may be looking to implement a full CADF framework.

## Typing commands vs. point and click - The R Language

This book teaches customer analytic data preparation utilizing R Programming Language. R is a little different than Excel or Tableau because users issue text commands to run data analysis versus point and click. Running text commands versus 'point-and-click' is a trend in analytics. SAS has launched the cloud based Viya programming language. Python is growing in popularity. Even Microsoft PowerBI resorts to text commands for the DAX language and custom calculations.

I admit, most organizations do not use R for enterprise implementations. R is used to **demonstrate** all the techniques with the hopes that readers can tailor and customize the approaches for their environments. For instance, I am using my work in R to create a CADF implementation for Microsoft .NET C# environments. Chapter 7 will cover enterprise implementations which often occur in SQL or Snowflake.

Here are some resources for those who are new to R. I encourage new R users to learn Base R[8] functionality first before attempting to learn specialized variants like tidyverse [9].

- The "Use R" series ("Use R!" n.d.)
- An Overview of the R Language by Chris Chapman and Elea McDonnell Feit, 2019 (C. Chapman and Feit 2015)
- Stackoverflow.com
- R Package Vintages: Example: (Dziurzynski, Wadsworth, and McCarthy 2020)

## Accessing/Modifying Rows and Columns

Most users update spreadsheets through the point and click interface. Those familiar with Excel know that rows are numbered and columns are lettered. Its relatively simple to click on column A to access the entire column. Same reasoning goes for accessing an entire row. Advanced users know that Excel has array and matrix functionality. Excel is even able to handle matrices. This is all nice for ease of use but bad for automation.

It's best to think of spreadsheets when navigating R vectors and matricies. Accessing data within R is done through row and column notation (instead of point and click). I'll use a tic-tac-toe example to demonstrate how to access data within R.

[8] https://www.r-bloggers.com/2022/08/base-r-is-alive-and-well/
[9] https://www.tidyverse.org/

Below, I am generating a character matrix to resemble a tic-tac-toe board. We'll call this ttt.

```
ttt <- matrix(c('x', 'o', 'x',
                'x', 'x', 'o',
                'x', 'x', 'o' ), nrow = 3, byrow=TRUE)
knitr::kable(ttt)
```

| | | |
|---|---|---|
| x | o | x |
| x | x | o |
| x | x | o |

For those that like Excel, I have generated matrix with A-B-C labels. Cells with a A represent column A. Cells with a 1 represent row 1.

```
excel <- matrix(c('A1', 'B1', 'C1',
                'A2', 'B2', 'C2',
                'A1', 'B3', 'C3' ), nrow = 3, byrow=TRUE)
```

| | | |
|---|---|---|
| A1 | B1 | C1 |
| A2 | B2 | C2 |
| A1 | B3 | C3 |

Here is the notation in R to access various cells in the 'ttt' matrix. Note that ttt is the name of our matrix. Row and column indexes are placed in brackets. The structure is variablename[R, C]. Leave C blank to access entire row. Leave R blank to access entire column.
If I type ttt[1, ] that is accessing the first row. If you type ttt, your are accessing the entire matrix.

```
ttt[1,]
```

```
## [1] "x" "o" "x"
```

Accessing the first row, in the Excel example. This would be like selecting row "A" in Excel. (Note that A1, B1 and C1 are returned. This represents row 1.)

```
excel[1,]
```

```
## [1] "A1" "B1" "C1"
```

Now let's move to accessing columns:

First column, tic-tac-toe example. Note the syntax I am using ttt[R , C]. Blank before the comma means all columns.

```
ttt[, 1]
```

```
## [1] "x" "x" "x"
```

Same idea, expressed with the Excel example. Let's return all the column A cells. Note that R prints all of the A results in a row. You are in fact accessing the A column even though it looks like the data has been converted into a row.

```
excel[ , 1]
```

```
## [1] "A1" "A2" "A1"
```

To see this, let's change all the A cells to AA. (Note we're accessing rows 1:3 and column 1)

```
excel[ , 1] <- "AAA"
```

Now, lets replace the excel variable with the original

```
excel <- matrix(c("A1", "A2", "A3", "B1", "B2", "B3",
                  "C1", "C2", "C3"), nrow=3)
```

## Data Frames

Now lets demonstrate the same idea with data frames. Even though R is based on matrices and vectors the data frame is the most common way analysts will work with data. Most analysts will load data from a CSV file into an R dataframe.

```
data("exceldata")
excel_df <- exceldata
```

Access the data frame. Show first few rows using the head function

```
head(excel_df)
```