**Foundations of Marketing Analytics**

Business analytics, big data, and data science, are very hot topics today,

and for good reasons.

Company are sitting on a treasure trove of data, but usually lack the skills and

people to analyze and exploit the data efficiently.

Those companies who develop the skills, and hire the right people to analyze and

exploit the data, will have a clear competitive advantage.

It's especially true in one demand, marketing.

About 90% of the data collected by companies today are related to customer

actions in marketing activities.

Which, what pages customers visit, what products they buy, in what quantities,

and at what price, which banners customers see or which emails they open, and

how effective these actions have been to influence their behavior.

Play video starting at ::56 and follow transcript0:56

The domain of marketing analytics is absolutely huge, and

may cover fancy topics, such as, text mining, social network analysis,

sentiment analysis, real time bidding, online campaign optimization, and so on.

But at the heart of marketing, lie few basic questions,

that often remain unanswered.

One, who are my customers?

Two, which customer should I target, and spend most of my marketing budget on?

And three, what's the future value of my customers?

So I can concentrate on those who will be worth

the most to the company in the future.

That's exactly what this course will cover.

Segmentation is all about understanding your customers.

Scoring models are about targeting the right ones.

And customer lifetime value is about anticipating their future value.

These are the foundations of Marketing Analytics.

And that's what you learn to do in this course.

Play video starting at :1:54 and follow transcript1:54

This course is complementary to the course Foundations of Business Analytics,

told by Nicolas Glady.

It's not absolutely necessary, but

if you haven't done so, taking that course first would make a lot of sense.

First, we'll assume that you know how to set up and

use the statistical language R in the RStudio software,

which is covered in the course Foundations of Business Analytics.

Play video starting at :2:19 and follow transcript2:19

If you don't, you might be a bit lost, at the beginning, and

we wouldn't like to lose you.

Second, why Nicolas Glady used many different examples from different domains

in the industry.

We will use one data set and only one.

Which we will explore and analyze it extensively, and

apply all the techniques and methodologies, to that same data set.

So you have a clear view of what an in depth analysis of one data set looks like,

and how it can be performed in real life.

Play video starting at :2:53 and follow transcript2:53

Third, a big part of the job of a data scientist

is about communicating your results to non analysts, and

being able to convince a board room based on data evidence.

Play video starting at :3:4 and follow transcript3:04

That is an essential skill, and we may make the difference between an employee

who is just a geek, and someone who can become an invaluable asset to a company.

That's something we'll not cover.

With the course told by Nicolas Glady, we give you useful tips and

tricks to that effect.

Now before we dig into the topics of segmentation, targeting, and

customer lifetime value, we make sure to set up the environment properly, and

explore the data set we'll use throughout.

Play video starting at :3:34 and follow transcript3:34

Fire up our studio, I'll see you in the next video.

# Setting up the environment and exploring the data (recital)

Hello and welcome to the first Marketing analytics R tutorial.

So first of all, we will assume you know your way around with R-Studio and R.

So basically here, I've put all the dot R files

we're going to review together in sin folder,

downloaded the purchases.txt file I also set it as my

working directory, I have the latest version of R ready.

I have the code open in front of me.

And I'm going to go over the lines one by one, so

that you get yourself familiarized with the data set we're going to study here.

So the very first thing to do is to, of course, load the text file,

the purchases.txt here into a data frame which we'll call data.

Of course, here we don't need to set up the directory, it's already

here in the right directory, in my working directory Tree, and the way I'm going to

run this R code is to run line by line, so you see exactly what I'm doing.

And the best way to do that is to click here on Run.

Run the current line or selection, which will run the current

Play video starting at :1:17 and follow transcript1:17

order in R and then move to the next line.

So, I am not going to go over the details here of the code.

But basically I am reading a file delimited by tabs,

which doesn't contain any header and where the decimals are indicated by a dot.

So I run the line.

Play video starting at :1:42 and follow transcript1:42

And as you can see here, we have loaded a data frame that

contains 51,000 lines and three variables.

If we look into it, it contains three columns and

all these columns are number one,

the identifier of a customer who made a purchase at one point of time.

Number two, the amount of purchase made and

number three the date during which that purchase has been made.

Two very useful of functions are called head and summary.

Head will show you the first few lines of any data frame, any matrix,

any vector you have and then summary will compute for you for

each variables you have, minimum, median, maximum, mean, and so on.

So here what we see is the identifier of customers, the lowest number is 10,

the highest number is 264,000, number two is the amount.

The minimum amount spent is $5.

The maximum is 4,500, with an average of 62.

And here as you can see, the dates are a bit weird.

There is no minimum, no maximum, no mean.

And the reason is because R doesn't know that it's a date.

It believes it's a string.

So the first thing we are going to do is to clean up and

prepare the data for further analyses.

To make things a bit easier, remember here there is no header in the file.

So the first thing we are going to do is to give names to the columns of the data.

We are going to give them names, customer\_id, purchase\_amount and

date\_of\_purchase, by using the column names instructions here.

It's done, and now we are going tell r that's

the variable date of purchase, which we've just renamed here, it's the third column.

That the date of purchase should be considered as a date.

So we can take date of purchase transform it into a date that has the following

format year dash month dash day, and

then store it in exactly the same column name.

We are going to replace the string by its date.

Play video starting at :4:4 and follow transcript4:04

And next thing we're going to do for

ease of analysis is to extract the year of purchase.

So we're going to take the date of purchase,

just keep the year in the output.

And then store it as a numeric value, into something called

year of purchase and basically our dataset is ready.

Let's look at it again, let's rerun the head.

So as you can see we have the year of purchase which is new

column we just created, the date of purchase here and if we rerun this summary

Play video starting at :4:42 and follow transcript4:42

like this then, basically,

the date\_of\_purchase will be computed a mean, a median, a maximum and so on.

So, the data that we're going to study begins on January, 2005,

it ends on December, 2015,

and most of the purchases have been made in 2011,

we'll see a few graphs dealing with that.

Now, of course, before jumping in and analyzing the data and

applying segmentation scoring and

lifetime value and we'd like to have a peek at the data in some form.

And here we're going to cheat a little.

In real life situations you will rarely have access to

the data exactly the way you'd like it to be.

You'd probably have to extract it up from an existing database.

It will probably be messy.

You'll have to prepare the data yourself.

And the best way to extract data from a database is actually

to use SQL statements.

Now, it's a whole different purpose that Marketing analytics, so

we are not going to become an expert in SQL statements

within this course, however, I am going to use one simple instruction in

SQL which is basically extracting data from a database.

Play video starting at :6:6 and follow transcript6:06

We're not going to set up a database management system just to do that.

So what we are going to is actually to use SQL statements from within our,

as if we were managing a database.

Actually we're not, we're just simply managing a data frame

which is called Data here and we're going to explore that

using SQL statements as if it were an SQL database.

The one line of code we're going to use here and to learn from SQL and

that the only one is select.

Select is basically a statement which is used to fetch

the data from a database, which returns data in the form of a results table.

Except that here, we're not going to deal with database directly.

We're going to deal with data frames from R.

So SQLDF basically means running SQL on our data frames,

which is obviously much easier than to set up an entire MySQL or

Oracle database to analyze from.

So, what did the SELECT statement says?

It says we're going to select each year of purchase we'll find in the database.

And for each year of purchase,

see how many times with the counts different it appears.

So basically, we're going to have,

how many purchases were made each in every year?

And that count will be renamed

as counter which we simply pre-name we're going to give it to.

We extract that data from our dataframe called data.

Play video starting at :7:53 and follow transcript7:53

And basically here in SQL,

you should put the name of the table you can extract your data from.

Now, year of purchase is something that you need to group by your results.

So we're going to group by one.

Making sure that your purchase will count for each year.

And then we are going to order by one because we'd like the data to

look nice and be ordered by your purchase.

So it's basically an SQL statement that we run on our

R data frames as if they were actual SQL database.

We run it, and now we have SQL data here, which has been

created as if it were the result of an SQL select statement.

If we look at it, that's what we get.

Now we're going to draw it nicely, and we're going to create barplot.

So remember, x here is the output of the select statement, and

we created, here, a variable which is called counter.

So we're going to plot x, counter and

the name of the legend should be the year of purchase.

If you launch that, you will have a nice chart showing how

many purchases have been made every year from 2005 to 2015.

So, of course, you should make the chart much prettier if you're going to present

it to someone, but enough said.

You get pretty nice count of how many purchases have been made over a decade,

and as you can see, it goes up very quickly at the beginning, and

then it keeps growing, but at a much slower pace.

And so the next thing we'd like to know is not only how many purchases have been made

but for how much, on average, each purchase accounts for.

So exactly the same thing, year of purchase.

We're going to compute not the count but the average of purchase amount.

We'll name it avg\_amount here.

We still extract the data from data.

Play video starting at :10:7 and follow transcript10:07

GROUP BY 1, ORDER BY 1, 1 being the first variable which appears here,

we are going to replace the variable x, we don't need it anymore.

Play video starting at :10:16 and follow transcript10:16

We create the extraction and then get the data and, as you can see,

on average the purchase amounts was pretty stable to begin with and

then increased slightly to 60 something, dollar per purchase.

And if you'd like to do the exact same thing, but instead of counting

the number of purchase, or computing the average of purchase amount, you just sum

everything up to know how much has been spent per year with that specific.

Retailer you run x

you get the new data set here with how much was spent here.

And as you can see in the last year, it's close to half a million.

And if you run the chart you have very nice and positive trend.

Play video starting at :11:11 and follow transcript11:11

In terms of how much money was spent with that specific retailer.

Now, here the objective is very simple.

First, to show you how very simple SQL statements work, and how you can

run them without any database running in the background in directly from R.

And number two, to get you accustomed to what kind of data you're going to analyze.

Now here is the exact same query we're going to run,

except it includes, in one statement, the year of purchase,

the number of purchases, the average purchase amount, and

the sum purchased every year in one statement,

you just need to put a comma here, and put them all.

You could put everything in one line,

but I put them in several so it's more readable.

And if you execute everything here,

Play video starting at :12:5 and follow transcript12:05

you have a pretty nice table with year of purchase, number of purchase,

the average purchase amount and the sum of everything together.

Play video starting at :12:17 and follow transcript12:17

Okay, now we know what the data looks like and

how to query it using very simple SQL statements.

Let's dig into it and create some segmentation first.

**.R files and dataset**

Be careful : the downloading system add automatically a code before the name of the file.

**RENAME the file.**

For instance, remplace : \_dbff2c8aae767c319b2a841e9298e6df\_module0.R by module0.R

[module0](https://d3c33hcgiwev3.cloudfront.net/_dbff2c8aae767c319b2a841e9298e6df_module0.R?Expires=1703721600&Signature=C09Z~ieZvFyzcadco0aFb1kiVBtGcivkxMH7iM8bb9uw~CmxGfKfJ8NUxtBGOhz9PxXgdx5ZT8rIrULAk7gZo6CeDV28zOWP2WvggggLVV0aIIJMOU0GqtExtQ1x8Aw3fe046aGCR5ZHnSpkEa89~wuoZfK8nBBlrEMzHYcqSXk_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[R File](https://d3c33hcgiwev3.cloudfront.net/_dbff2c8aae767c319b2a841e9298e6df_module0.R?Expires=1703721600&Signature=C09Z~ieZvFyzcadco0aFb1kiVBtGcivkxMH7iM8bb9uw~CmxGfKfJ8NUxtBGOhz9PxXgdx5ZT8rIrULAk7gZo6CeDV28zOWP2WvggggLVV0aIIJMOU0GqtExtQ1x8Aw3fe046aGCR5ZHnSpkEa89~wuoZfK8nBBlrEMzHYcqSXk_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[module1](https://d3c33hcgiwev3.cloudfront.net/_32fb2283d156b7562e6d636329a5ee2f_module1.R?Expires=1703721600&Signature=kmxru52aegmmHabL-~q1AbEHUm40L6-Tlwd7hNJlwL6KCrXqCLEzRfyUYWnaIIc692PMNOwWh9zlF~IfoXcVve41Zq7Hhbd9mPo2zdqecdEJJu3tb7ZbKss4khynQaS~8I~qd-jPlCGL7QySL8~KrJV6ZKljdymC3w~FmI6iMZk_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[R File](https://d3c33hcgiwev3.cloudfront.net/_32fb2283d156b7562e6d636329a5ee2f_module1.R?Expires=1703721600&Signature=kmxru52aegmmHabL-~q1AbEHUm40L6-Tlwd7hNJlwL6KCrXqCLEzRfyUYWnaIIc692PMNOwWh9zlF~IfoXcVve41Zq7Hhbd9mPo2zdqecdEJJu3tb7ZbKss4khynQaS~8I~qd-jPlCGL7QySL8~KrJV6ZKljdymC3w~FmI6iMZk_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[module2](https://d3c33hcgiwev3.cloudfront.net/_7352e00d539520e98c1a550fb37d37a1_module2.R?Expires=1703721600&Signature=ORUCeHFw-X9W4bcNcwvGrf8IbNSBUY2LU~hSnb9m3RPrqoCPfGJekJyDywfGU1ZefxRGU3A9DcvgzJJ1fA09ycWOCptzONaldfK0-zfFEJjrBeRxNJWN6VmNzGmfhPa5Xa1CamW4KIsmNKJl6kyvbHBVqLiBql-nDQyNgfNMTrw_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[R File](https://d3c33hcgiwev3.cloudfront.net/_7352e00d539520e98c1a550fb37d37a1_module2.R?Expires=1703721600&Signature=ORUCeHFw-X9W4bcNcwvGrf8IbNSBUY2LU~hSnb9m3RPrqoCPfGJekJyDywfGU1ZefxRGU3A9DcvgzJJ1fA09ycWOCptzONaldfK0-zfFEJjrBeRxNJWN6VmNzGmfhPa5Xa1CamW4KIsmNKJl6kyvbHBVqLiBql-nDQyNgfNMTrw_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[module3](https://d3c33hcgiwev3.cloudfront.net/_43e69278e2a096e4f45b480f1b1486c4_module3.R?Expires=1703721600&Signature=EclKqT9K3RlB0W48wTW2rDpabutp7GLA1XFvhxjEs-iiSRFEKblWMNGo7MZWNsvMWnbL6Keqt9Zz4Gn~VK6gOA77oRodpGAYlukVulOpE6BIWiXlbqjXav3ZNFGPUzSBc1-7V4JNk51nmnBlELgjNYE9Fz0DNN6u6VtGr4s-~U8_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[R File](https://d3c33hcgiwev3.cloudfront.net/_43e69278e2a096e4f45b480f1b1486c4_module3.R?Expires=1703721600&Signature=EclKqT9K3RlB0W48wTW2rDpabutp7GLA1XFvhxjEs-iiSRFEKblWMNGo7MZWNsvMWnbL6Keqt9Zz4Gn~VK6gOA77oRodpGAYlukVulOpE6BIWiXlbqjXav3ZNFGPUzSBc1-7V4JNk51nmnBlELgjNYE9Fz0DNN6u6VtGr4s-~U8_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[module4](https://d3c33hcgiwev3.cloudfront.net/_db79a7d03c483402e7cf7999e494d77c_module4.R?Expires=1703721600&Signature=QgjINtP-o2~T6pCHvOCqfbVY-avEyS6h5uVhIxWQunBzgGL~nIqPxvdb5h4LJGkzosrs3URh4zOFRf~1RYVyTEhnZYdqLmZURZLiux7PEUtbkFcM3uVug7l8xq0s6br6bCVCgmzLJJUb9fBNJak9VQ9F-8VnGnAFwBN0GEtBIkA_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[R File](https://d3c33hcgiwev3.cloudfront.net/_db79a7d03c483402e7cf7999e494d77c_module4.R?Expires=1703721600&Signature=QgjINtP-o2~T6pCHvOCqfbVY-avEyS6h5uVhIxWQunBzgGL~nIqPxvdb5h4LJGkzosrs3URh4zOFRf~1RYVyTEhnZYdqLmZURZLiux7PEUtbkFcM3uVug7l8xq0s6br6bCVCgmzLJJUb9fBNJak9VQ9F-8VnGnAFwBN0GEtBIkA_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[purchases](https://d3c33hcgiwev3.cloudfront.net/_1750936296f8224de16fad133923d350_purchases.txt?Expires=1703721600&Signature=BvDETgdxumDqU~PJZXqvAiEihHH8r5s0~axx-uPgj4Nww-4Y6hCJZ58Vwwc58iHe1rZVyMqOz1KGjOpQyBd6WQb6I7JYKGEYHA1HUihZZI2~Wq1AUIoXamuwTal0dtJBEXt5sd8olOUHGwYjO99~0tr9NhBZNf-lKt8UyrMienY_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[TXT File](https://d3c33hcgiwev3.cloudfront.net/_1750936296f8224de16fad133923d350_purchases.txt?Expires=1703721600&Signature=BvDETgdxumDqU~PJZXqvAiEihHH8r5s0~axx-uPgj4Nww-4Y6hCJZ58Vwwc58iHe1rZVyMqOz1KGjOpQyBd6WQb6I7JYKGEYHA1HUihZZI2~Wq1AUIoXamuwTal0dtJBEXt5sd8olOUHGwYjO99~0tr9NhBZNf-lKt8UyrMienY_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

# Introduction

If you got to the URL indicated here,

you will see a great example of marketing segmentation.

Acxiom is a marketing company that collects tons of data

about customers through surveys, purchases,

marketing data, collection of online behavior and so on.

They might gather data about what you usually buy, where you live, what hobbies

you have, how old you are or how many of your children are still at home.

The database may be used by companies to better understand the profile

of their own customers.

Or, find people who match this PC profile and they represent good candidates for

acquisition campaigns.

Acxiom might have dozens or

even hundreds of pieces of information about any one customer in the database.

But, of course,

while extremely rich, this information is very hard to go through and use.

So, they use segmentation to summarize the information they have and make it usable.

Play video starting at :1:7 and follow transcript1:07

Each bubble you see on the chart is a segment in the Acxiom database.

In this case, the UK database.

In each segment, groups of bunch of customers who have similar

sociodemographic profiles, needs, wants, and consumption habits.

Play video starting at :1:23 and follow transcript1:23

And it becomes very easy to click on any one of those bubbles and

see what kind of customers we are talking about.

In the example below, segment 15, groups of retirees who are seldom online

with limited income and who represent about three person of the Acxiom database.

They are described clearly and in very simple terms but that simple description

maybe summarizing hundreds of indicators Acxiom has about them.

In other words segmentation transforms the magna of data

in to something that is clear and usable.

And that's really what segmentation is all about.

Play video starting at :2: and follow transcript2:00

You can't treat all your customers the same way, offer them the same product,

charge the same price, or

communicate in the same benefits, because you leave a lot of value on the table.

In 1909, Henry Ford said any customer

can have a car painted any color they want so long as it's black.

That time is long gone.

You've got to understand the customers' differences in needs, wants, desires,

habits, profiles because those differences will allow you to customize your

offerings, adapt your customer messages and optimize your marketing campaigns.

On the other hand, when you have thousands, if not millions,

of customers in your database, and

you have potentially hundreds of bits of information about each, from what they've

purchased in the past to what links they clicked on your website, or who they are,

you need to simplify, because the world as it is is too complex to apprehend.

You cannot treat each and every customer individually.

It would be too costly.

So a great segmentation is all about finding a good balance between

usability and completeness, between simplifying enough so

it remains usable and not simplifying too much, so it's still valuable.

That's what we learn to do in this module.

If you open a textbook about segmentation,

you'll find a long list of criteria to define what is a good segmentation.

Such criteria may include homogeneity within segments.

Customers within each segment should be similar enough.

Heterogeneity across segments.

Customers who are different should fall in different groups.

Play video starting at :3:41 and follow transcript3:41

Measurable, substantial, accessible or identifiable, actionable or operational.

Responsive, in the sense that each segment should react differently to different

marked in mixed elements.

And, you know what, that list is absolutely correct.

But, I'd like to simplify to its core.

A good segmentation should be statistically relevant and

managerially relevant.

That is useful to the manager using it.

In the next video segment we'll see exactly what we mean by statistically

relevant.

# Acxiom URL

[**http://www.personicx.co.uk/personicx.html**](http://www.personicx.co.uk/personicx.html)

# Hierarchical segmentation

To illustrate how you build a segmentation lets take a very simple graphical example.

Lets assume that you have only ten customers in your database and

that these ten customers are only described by two factors, or

what we call segmentation variables.

One, how often they make a purchase in one of your stores from every week

to once a year, and two, for how much they buy every time they shop.

Play video starting at ::35 and follow transcript0:35

Graphically these ten customers can be represented on a two-dimensional map,

where the first horizontal axis represents frequency of purchase,

with the most frequent shoppers being on the right.

And the second vertical axis represents the average purchase amount, with those

customers who spend the most at each trip appearing at the top of the chart.

Intuitively, how many segments do you see here?

You can see that the market neatly separates into three segments.

Each segment represents a group of customers that is distinct from the other

groups and all customers within each group are quite similar.

The purple segment contains customers who buy very often, and

for a lot of money at each purchase occasion.

The green segment groups together customers who make regular purchases, but

for lower amounts.

And the orange segment, groups together customers who shop much less frequently,

but when they do, spend a lot.

Play video starting at :1:37 and follow transcript1:37

From a natural point of view, these three segments, make a lot of sense.

The purple segment is strategic and

may generate a vast portion of the firm's profit.

The green segment constitutes the bulk of the purchases,

because they buy frequently, but may not be as profitable as one might think.

And the orange group, is a perfect target for marketing actions that might

encourage more frequent purchases, such as special store events or seasonal coupons.

But how can a statistical software find the three segments automatically?

Well, there are many methods, but

the one we'll explore is called hierarchical clustering.

The internal process the software goes through goes as follows.

First, you consider that each and every customer.

Is it's own segment,

to begin with there are as many segments as there are customers, in this case 10.

Then you ask the question which two clusters or at the very beginning, which

two customers could I group together so that I would lose the least information?

That is, they would be so similar that if I treated them as absolutely identical,

it would make no difference.

Obviously, the answer is that the two most similar customers or

segments, are those in the lower right corner, because these two customers

are almost identical in terms of frequency of purchase and purchase amount.

So if you group them together and

treat them as if they were identical you wouldn't lose too much information.

The trick is simply to continue the process.

By grouping these two very similar customers together,

we went from ten segments to nine.

Then we go from nine to eight, seven,

six, five, four, and finally three segments.

The three segments we identified earlier.

Nothing prevents us from continuing the process however.

We could keep grouping these customers into only two segments, but

of course we'd lose a lot of information.

By grouping together those customers in the top part of the chart,

we'll simply have one big segment of large spenders, but

without the ability to distinguish frequent buyers from occasional ones, and

the segmentation would be much less useful for managers.

Oh, by the way, we could even go from two to only one segment.

It's called the market as a whole.

So, as you can see, a good segmentation consists of

finding the right balance between treating each customer individually.

As if they were all their own segments.

And treating all customers the same way

as if there were only one big segment to which everybody belonged.

Play video starting at :4:33 and follow transcript4:33

But where do you draw the line?

To answer that question we need a tool called the dendogram, and

that's the topic of the next video segment.

# Selecting the "right" number of segments

In the previous video, we've shown how customers are grouped together,

step by step, from the most similar to the least similar.

At the one extreme each customer is unique, it's very own segment.

At the other extreme each customer belongs to only one big undifferentiated segment,

it's called the market.

Where do you draw the line?

How do you decide what is the best number of segments?

Well, there is no magic answer.

First and foremost it depends on how managerially relevant

your segmentation is.

If from a managerial point of view it makes sense to simplify and

reduce the number of segments to make your solution more useable,

then by all means do it.

And if it makes sense to expand your segmentation by one segment or two,

because distinguishing customers more precisely makes managerial sense,

then again, you have your answer.

But if you are unsure or you don't know where to start, there is tool

that can be used to guide your decision which is called the dendogram.

A dendogram is a graphical representation of the hierarchical clustering process,

that's why we talk about hierarchical segmentation.

All customers end up being grouped together, but there is hierarchy,

a priority, and the dendogram illustrates this nicely.

At the bottom, you have all the customers you are segmenting.

And the tree shows how quickly and

in which order these companies are grouped together into segments.

At the end of the process,

at the very top, all customers fall into the same segment.

But a dendogram will also show you how much information you lose

by grouping segments together.

If by grouping two segments together you're losing a lot of information,

then you will see a jump in the dendogram.

If you begin from the bottom and go up, and up, and

up, at one point you might identify what is called a sudden jump.

That's where the segmentation and

clustering process begins to lose to much information.

That's when you're beginning to group together

customers who are too different from one another.

And if you'd like to avoid losing that much information

you need to stop the process right before.

In this example there is a sudden jump between three and two segments.

Meaning that by going to only two segments you are losing a lot of information.

And you should stop the segmentation process before, at three.

There is no magic bullet here,

but it's a very useful tool when you don't know where to start.

(Required)

English

​

Help Us Translate

# Segmentation variables

In the previous video, we talked about how customers are grouped together into

segments based on how similar they are, and

we also talked about the ways to decide where to stop the process and

decide the best number of segments to retain.

Play video starting at ::22 and follow transcript0:22

But we avoided one very important question.

If you group customers based on how similar they are, similar on what?

Well, the correct answer is, of course, it depends.

It depends on the managerial question you're asking.

If you'd like to understand how people use your website, for

instance, maybe you should study similarities in terms of pages visited,

number of clicks, and duration of visits.

If you're trying to optimize product recommendations or

trying to personalize catalogs,

maybe you should study similarities in terms of products purchased in the past.

If you are trying to optimize marketing campaigns, you should definitely group

customers based on their similarities in terms of profit and

responsiveness to past marketing campaigns.

Play video starting at :1:8 and follow transcript1:08

The specific indicators on which you compare customers are called

segmentation variables.

The software doesn't care what kind of data you are analyzing.

If you feed the segmentation algorithm with data that is managerially irrelevant,

it will find similarities that are useless.

So whatever segmentation study you want to conduct,

the first question should always be, what managerial goal do I want to achieve?

And based on the answer, then, and only then, you should ask yourself

what segmentation variables are relevant to achieve that managerial goal.

Don't select segmentation variables simply because they are easily available to you.

(Required)

English

​

Help Us Translate

# Recency, frequency, and monetary value

In many marketing studies,

three specific marketing indicators often turn out to be invaluable.

They're called Recency, Frequency, and Monetary Value and they've been shown to

be some of the best predictors of future purchases and customer profitability.

Play video starting at ::25 and follow transcript0:25

Recency indicates when a customer made his or her last purchase.

Usually the smaller the recency, that is the more recent the last purchase,

the more likely the next purchase will happen soon.

Play video starting at ::37 and follow transcript0:37

On the other hand, if a customer has lapsed and has not made any purchase for

a long period of time, he may have lost interest or

switched to competition, which is bad news for future business.

Frequency, refers to the number of purchases made in the past.

The more purchases have been made in the past,

the more likely additional purchases will occur in the future.

Finally, monetary value refers to the amount of money spent on average at

each purchase occasion.

Obviously, the more a costumer spends on average, the more valuable he is.

Play video starting at :1:13 and follow transcript1:13

Marketing segmentation that use recency, frequency, and monetary value as

segmentation variables are often referred to by the acronym RFM segmentations.

Play video starting at :1:25 and follow transcript1:25

But these key marketing variables are usually not readily available

in the data you want to segment.

More often than not, you simply have access to transactional database

that contains a list of past purchase made by each customer.

And you need to compute recency, frequency, and monetary value yourself.

In the next video segment, we'll show you how to compute these indicators in R.

# Computing recency, frequency and monetary value with R (Recital 1)

So in this tutorial, we're going to learn how to compute recency,

frequency, and mandatory value from the data set.

We are starting from scratch, so we opened module 1.R.

Everything is clean, nothing has been uploaded or loaded in our memory.

So we going to read the data from the purchases.txt file,

as we did in the previous tutorial.

Play video starting at ::27 and follow transcript0:27

Here it is.

The data contains now 51,000 rows as usual.

Play video starting at ::33 and follow transcript0:33

As we did last time, we going to transform and

set column name so they're easier to manipulate and use.

Later on, we're going to rename the columns customer ID, purchase amount,

Play video starting at ::48 and follow transcript0:48

and date of purchase, using the column names instruction and function,

Play video starting at ::55 and follow transcript0:55

we're going to tell r that this purchase is not a string as it believes to be,

but it's a date in the format of years, months, date with a dash in between.

We'll run it.

And then we're going to compute something a bit specific.

We're going to compute the number of days that lapse

between January 1st, 2016 and the date of purchase in the data.

So we need to compute the time difference between the very last day of the database.

Actually the last day is December 31st, 2015.

You take the 1st day of 2016 and the date of purchase.

And we're going to compute the difference in days and

store that difference in a new variable called days since.

We can actually aim to use that to compute recency.

Play video starting at :1:53 and follow transcript1:53

Okay, now it's done.

Play video starting at :1:55 and follow transcript1:55

If we look at the data we have.

The days since January 1st, 2016 and the date of purchase,

in this case November 6th, 2009 and everything else.

Okay?

Now, again you can have a look at the data with the summary on average,

there's been 1,600 days between any purchase

observed in the database and January 1st, 2016.

And what we are going to do here is to compute directly recency,

frequency, and averaged initial amount.

And we can do that using an SQL like statement.

So how does this work?

Well it's a bit complex, so follow me.

We are going to select the customer ID of every customer found in the database.

The trick here is that it will only appear once for every customer.

So even though we have 51,000 purchases we'll only have as many

unique customer IDs as there are in the database with that statement.

Now, for each customer, we need to compute

the minimum number of days between all of his or

her purchases and January 1st, 2016.

Now, of course, if we take the minimum number of days, then we are going

to have the day of the last purchase, which is the very definition of recency.

Then for each customer we need to compute the frequency,

which is basically how many purchases that customer has made.

The asterisk here basically means anything in

the data that is related to that customer we could have put any known variable here,

it wouldn't change and then for the amount we going to compute the average

of purchase amount for that specific customer ID and

name that aggregate computation as amount.

Now, the trick is that we'd like to make sure that each row

only appears one for each customer.

So we going to compute that from the data and group by one, meaning that everything

here is going to be computed and grouped per customer ID.

And so if you run that statement,

the output is a new data set named customers and

as you can see, that data set only has 18,000 rows.

Meaning that the only 18,000 unique customers

in this data set of 51,000 purchases.

So 18,000 customers have made a total of 51,000 purchases.

I should like to explore the data which we just created called customers,

you can of course look at the head and as you can see for

each customer ID we have the recency the number of days between his or

her most recent purchase and January 1, 2016.

How many purchases that guy has made.

That customer, customer id number 10 has made only one.

And the average purchase amount is 30.

If you compute some summary statistics about that new data set, called customers,

you'll see that the average recency of a customer is 1200.

So, on average customers have lapsed for about 4 years.

Some have lapsed 10 years ago,

some have lapsed only a few days ago, but the average is four years.

The minimum is barely wonder,

the maximum is pretty much the length of the entire data set.

In terms of frequency, some have made only one purchase.

Actually, many have made only one purchase.

Some have made an astounding number of 45 purchases over their lifetime,

or at least over the 10 or 11 years we are observing,

but on average, people have made 2.8 purchase.

And then in terms of amount, the average amount goes from 5 to 4,500 and

the mean is around $57 per purchase per individual.

If you like to have a better look at the data,

you can also look at the distribution of recency, frequency, and

average donation amount by creating a histogram of the distribution.

Play video starting at :6:51 and follow transcript6:51

And as you can see, here we are talking about recency.

So you have a few customers whose recency is about 4000 days, and

a bunch of customers here whose recency is much more recent.

About 100 days or so.

Play video starting at :7:8 and follow transcript7:08

If you look at the histogram in terms of frequency,

Play video starting at :7:14 and follow transcript7:14

it's even more extreme.

Many people have made only 1 or 2 purchases, and then when you go 5,

10, 20, 30 purchases, they are even rarer in the entire database.

And then finally, in terms of amount, the chart doesn't look very nice,

simply because you have a lot of customers who made purchases of $20, $50, $70.

And then extremely few who made purchases of

Play video starting at :7:50 and follow transcript7:50

much larger amounts, like $4000.

So one One thing you can do is to draw a histogram of customer amount.

But making sure it shows

many more breaks in the histogram than you'd have naturally.

So you can click Run here.

Play video starting at :8:10 and follow transcript8:10

You'll have actually a much more detailed view of the histogram,

and as you can see, most people spend around $40, $50.

A few spend over $100 and

extremely few spend above $200, on the average.

# Data transformation

In a previous video, we've showed that customers are grouped together based

on how similar or dissimilar they are to one another.

And that similarity can be seen as a measure of distance.

If two customers are very similar, they are said to be close to one another.

The issue that, sometimes, to compute how similar two customers are,

we have to compare apples and oranges.

For instance, if you'd like to group customers based on recency, frequency, and

monetary value, you are basically comparing variables that

are measured in terms of days, purchase occasions, and amounts.

These segmentation variables do not even use the same scales.

How do you compare one to another?

Play video starting at ::52 and follow transcript0:52

In segmentation studies, data transformation is essential.

You need to prepare and transform your data, so

your segmentation variables can be compared to one another.

Play video starting at :1:4 and follow transcript1:04

When your segmentation variables do not use the same scales,

you need to standardize them.

In statistical terms, to standardize means that you subtract the mean and

divide the data by its standard deviation.

We won't go into the details, but it simply means that,

regardless of the actual scales used for your segmentation variables,

they will roughly be scaled back within a range between minus two and

plus two with some external values falling outside that range.

Now, regardless of what the original scale was in days, dollars, or

purchase occasions, they can be compared to another.

Play video starting at :1:44 and follow transcript1:44

Another issue is to deal with data dispersion and

the best example is probably about purchase amounts.

When you look at the transactional database, many customers will buy for

moderate amounts and only a few will buy for very large amounts.

The distribution is said to be skewed and may look something like that.

So what's the problem?

The problem is not statistical, it's managerial.

Would you say that a customer who spends $5 on average is different

from someone who spends $15 instead?

Well, definitely the latter generates three times more money at

each purchase occasion.

For a manager,

they're definitely different, it might be best if they were in different segments.

Play video starting at :2:27 and follow transcript2:27

But what about two customers who spend $310 for the first and $320 for

the second?

From the managerial point of view, not a big different, right?

Yet, from a statistical point of view,

the difference in purchase amount Is exactly the same, $10.

When you're facing that kind of situation,

it might be worth transforming your data and taking the logarithm of the amount.

Once you take the log, the distribution would look something like this.

The same difference of $10 will have a huge impact on the left part of the chart,

those who spend the least, but a minor influence on the right part of the chart,

those who spend the most.

It's not really necessary from a statistical point of view, but from

a marginal point of view, the segmentation solution will make much more sense.

Play video starting at :3:16 and follow transcript3:16

Okay, now we are ready.

In the next two videos, we prepare our data for segmentation purpose, and

then, we'll run a segmentation and see what we got.

# Preparing and transforming your data in R (Recital 2)

In this tutorial, we are going to prepare our data for segmentation purpose.

So, if you're starting from scratch, don't forget to,

up in module1.R, rerun everything that has been covered in

the first tutorial of this module and then you are ready to go.

So, because we are going to transform our data and apply some modifications to it,

I'm going to make copy of it, which is called new\_data, so

that that new data is a perfect copy of customers with 18,000 rows.

And then, I'm going to show you something, which is a bit annoying.

If you look at the variables, one is customer ID.

And that variable, of course is useless, as far as segmentation is concerned.

So, we'd like to get rid of it.

But we don't want to lose that information either.

So what I'm going to do is to take that customer ID information and store it in

place of the row names, one, two, three, four, five, six, that you see here.

So what I'm going to do is to store as row names that column,

customer ID, and then, remove it.

Meaning, take the original customer ID Column and

set it to null, which is basically a way to remove entirely that column.

And once it's done, as you can see, the customer ID appears as row names, but

has disappeared as segmentation variables, which is exactly what we wanted to do.

Now if you look, if you remember at the histogram of average

purchase amounts, it's extremely skewed to the left.

And it's not really useful for

segmentation purpose to keep the data in that form.

So as discussed, we are going to take the log of it, and

if you plot the new histogram of that log, so we transformed and

replaced the amount with the log of the amount.

And if you take the histogram, then the distribution is much smoother,

much nicer to analyze, and we'd be much more useful for

segmentation purpose as discussed.

The very last thing to do is to remember that all these variables, recency,

frequency, monetary value, and others you might have in your segmentation

need to be compared to one another in terms of distance and differences.

But, of course, they use completely different scales.

Recency is expressed in terms of doz, frequency goes up to 10,

45 maybe at most amounts is in dollars.

So you need to scale the data.

So, each column can be compared to one another, and

that's exactly what you're going to do.

New data becomes the scale of itself.

Meaning, that each column will have an average of zero and

a standard deviation of one.

And so now, we can compare some deviation of recency to other deviations

in frequency, or deviations in amount compared to the average and

distribution of the population we are interested in.

And now, basically, the data is ready for being segmented and analyzed.

# Running a hierarchical segmentation in R (Recital 3)

If you think about it, we came a long way.

We loaded a data set.

We explored and extracted recency, frequency, and

monetary value using SQL statements.

We prepared and transformed the data to be ready for segmentation.

So I'm just going to rerun everything here, since I started fresh,

Play video starting at ::25 and follow transcript0:25

and now we are ready, and it is as if we were just prior that tutorial.

The next thing we are going to do is compute the distances among customers.

Play video starting at ::38 and follow transcript0:38

Knowing that the closer two customers are,

the sooner they will be clustered together into the same segment.

But here is an issue.

Our customer's data set contains 18,000 lines, 18,000 customers.

And so if you want to compute distances among these customers,

we'll basically ask R to compute the distances

between 18,000 customers to themselves,

which is a total combination of 340,000,000 distances in total.

So, it could work on your machine, maybe not.

Play video starting at :1:20 and follow transcript1:20

It works on mine.

But on many, that would be just too much to handle in terms of memory requirements.

So if you just compute the distance, using the dist function here,

it might generate out of memory problems.

So we are not going to do that, and we're going to take a sample of the data set.

To create a sample that is not random, always the same,

wherever it'd like to rerun,

we're going to take one customer out of every ten customers in the data set.

So we are going to create a sequence from one to the total number

of customers and only taking one every ten customers.

And that sampling mechanism will be stored in a variable that you call sample,

and which looks like this.

1, 11, 21, 31.

So we are going to take the first, the eleventh,

the twenty-first customer only, in our data set.

And we're going to use the sample suffix

to make clear that you only analyze a sub-sample.

You take the customers, and you take only the rows that match the sample mechanism.

So we take only row 1, 11, 21, 31,

and by leaving the second part of these brackets empty,

we basically mean that we take every column, every variables available.

So we take that, both for customer sample, which remember,

contains all the original data in terms of recency expressed in days,

frequency expressed in terms of number of purchases, and

monetary values expressed in terms of dollar.

And the same thing for the new data, which contains pretty much

everything but scaled, and the login for the amount.

Now, we are ready.

We're going to compute the distance among these not 18,000 customers,

but 1,800 customers which have been sampled from the larger data set.

And store the distances among all these customers,

computed on standardized data, in a variable called d.

And we're not going to look at that variable specifically,

we're just going to use it as a parameter later in the segmentation process.

Play video starting at :3:52 and follow transcript3:52

Okay, so we have a distance matrix here,

which already contains 1.7 million elements, which is quite significant.

And we need to use the distance in the hclust function.

hclust means your [INAUDIBLE] clustering,

that's the heart of the clustering method we are going to use.

There are many methods available, the one that I suggest is to use ward.D2.

You can look at the help functions and what it means exactly, but

that's a pretty robust way of using distances to create clusters together.

And the output of that cluster algorithm will be stored in a variable called c.

Play video starting at :4:38 and follow transcript4:38

And it's done already.

Now if you plot c,

you'll see a very specific plots that you probably know to dendrogram.

Play video starting at :4:49 and follow transcript4:49

Let me zoom on it.

So here you have the 1,800 customers.

Within them, obviously, its not readable.

You can make it prettier if you'd like, and

then you can see how all these individuals have been clustered together

progressively step by step up to a stage where there is only big cluster here.

So if you stop at four, you'll have that cluster over here,

that smaller cluster over there, that pretty big cluster over here,

and then a fourth cluster here to the extreme right.

If you stop at nine, then basically you cut the clustering tree,

the dendrogram, much lower, where you have nine clusters.

Where should you stop?

Well, here it becomes tricky because there are multiple criteria to use.

Both in terms of statistical fit, in terms of modular relevance,

and in terms of targeting ability.

And just for the examples, we're going to cut at nine segments.

So, you take the output of the hclust function here, which you see.

You take that c, and you cut the tree at nine.

And so the members here will contain the ID

of all the individuals you have containing each and every segment it belongs to.

If I showed the first 30 elements of that data,

you probably remember that earlier, we took here the row\_names and

replaced them with the customer\_id and removed that.

Well, the reason we did that is because now the customer\_id appear here, on top.

And so you know that individual number 10 belonged to cluster 1,

individual number 510 belonged to cluster 1 as well, and so on.

Play video starting at :6:58 and follow transcript6:58

If you run table, it will count how many individuals,

how many customers belong to each cluster.

And that's obviously a very important thing to know.

So here, cluster number 5 contains only 49 individuals,

while cluster number 7 contains 236.

Of course,

that's not very useful if you don't know what these clusters are all about, so

what you'd like to do next is to compute the average profile of each segment.

And here is the trick, we don't care about standardized variables any more.

What you care about are the averages in terms of recency, in terms of

those amount, in terms of dollars, frequency in terms of number of purchase.

So, we're going to compute the aggregate of the actual

original data set, customers or customer sample.

The first few variables, which are relevant to us,

meaning frequency, recency, monetary value.

And we are going to group them by cluster membership,

cluster membership which comes from where we decided to cut the tree, okay?

And because we are going to run some kind of segment profiling,

the function we like to use is mean.

So, that line means that tag the vitals here recency,

frequency, monetary value.

Group them by cluster membership, and for each group, compute the mean.

And if you do that, What you see, for

instance, is that if you focus on cluster number 4,

which contains 306 individuals,

cluster number 4 has an average recency of 162 days,

an average frequency of 2.4 purchase made in the past,

and an average purchase amount of $41.

While you can see the big differences,

you have cluster number 6 is much, spends much more in the shop.

Clusters number 2 spends much less.

Cluster number 3 made a huge number of purchases in the past, whereas

cluster number 8 made only one and a pretty long time ago, and so and so forth.

So if you study that more carefully, you can see that the segmentation mechanism

grouped people into clusters of Imogen's customers, all with different profiles and

all that can be characterized in terms of managerial interest.

Play video starting at :9:55 and follow transcript9:55

And that's basically the ID.

The ID is that we took 1,800 customers and

found which were alike and should be grouped into clusters.

And these nine clusters summarized pretty well the diversity of

profile you have in your database.

That's the core of segmentation.

But as we'll see in the next module,

Play video starting at :10:24 and follow transcript10:24

many companies do not use that kind of segmentation exactly.

But use what is called ad hoc segmentation, or module segmentation.

That's the next topic.

**Instructions before starting the quiz 1**

To validate Module 1 and correctly answer the questions will require that you perform the following exercise first: take the code in the file named module1.R, and modify it such that (1) the segmentation variable "frequency" is replaced by its log (before it is scaled), and (2) you select a 5-segment solution instead of a 9-segment solution.

**You'll be asked to answer these questions :**

What is the size of the largest segment?

What is the average purchase amount of the segment which contains, on average, the customers who have made their last purchase the most recently?

Of the five following criteria to determine the best number of segments, which one is the *least* relevant?

Does customer #260 belongs to the same segment than customer #5920?

Looking at the average profile of segment 1, would you say that members of this segment are typically...(you'll have to choose the right proposition)

[module1](https://d3c33hcgiwev3.cloudfront.net/_458a92824e666fdfe2f52398b0357922_module1.R?Expires=1703721600&Signature=LgluPQK8h2Y9qqtk~yIDDWzoFUOFIeePvYUshMS1GavWLyrHQ~lWOa4~px2Nwaxz6IdQ2Rudw4e0kZ7LBeFipBoNtzdi38uLeqwRhow7i4KgLlFDpfPDX5fc3NhNdCre0rsKscOBqybnUNkKl-0uULBfcDeXqy078f6AASjX-0s_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[R File](https://d3c33hcgiwev3.cloudfront.net/_458a92824e666fdfe2f52398b0357922_module1.R?Expires=1703721600&Signature=LgluPQK8h2Y9qqtk~yIDDWzoFUOFIeePvYUshMS1GavWLyrHQ~lWOa4~px2Nwaxz6IdQ2Rudw4e0kZ7LBeFipBoNtzdi38uLeqwRhow7i4KgLlFDpfPDX5fc3NhNdCre0rsKscOBqybnUNkKl-0uULBfcDeXqy078f6AASjX-0s_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[purchases](https://d3c33hcgiwev3.cloudfront.net/_e711af551fbda7ea1239aa9afd734cfd_purchases.txt?Expires=1703721600&Signature=Y8ZH~pB0gltFyq69aIzn2Iiw~FdVLXpjxtwhmR~ltAEX5KYw-syQ5nYWeDwJTgFUXX9QtYMrnmYo8Ls4p013hFZkEnND4y7C-NwBP6NIRSpWYf~mUyhn2TkaDaYDPwaoBVcm-2YnhwKMbP2Z8cpc9PmE6QVyvVC5DnzjQ-dRhoc_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[TXT File](https://d3c33hcgiwev3.cloudfront.net/_e711af551fbda7ea1239aa9afd734cfd_purchases.txt?Expires=1703721600&Signature=Y8ZH~pB0gltFyq69aIzn2Iiw~FdVLXpjxtwhmR~ltAEX5KYw-syQ5nYWeDwJTgFUXX9QtYMrnmYo8Ls4p013hFZkEnND4y7C-NwBP6NIRSpWYf~mUyhn2TkaDaYDPwaoBVcm-2YnhwKMbP2Z8cpc9PmE6QVyvVC5DnzjQ-dRhoc_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

<https://github.com/nicolasfguillaume/Marketing-Analytics-with-R/blob/master/problem_set_1.md>

# Limitations of statistical segmentation

In the previous module, we've shown you how to prepare your data,

compute recency, frequency, and monetary value, and

run a statistical segmentation to identify segments in the data set.

Although it might come as a disappointment to some of you, many companies do not

use these kind of statistical methods to segment their customer database,

but use rather a simple set of rules, and for good reasons.

First of all, customer data is added on a continuous basis.

Every day, every second.

Customers make purchase and modify their behavior.

Old customers disappear, new customers are acquired, and the segmentation

solution you've put in place becomes obsolete almost at the moment you run it.

So you need to update your segmentation very frequently.

Second, you cannot run that kind of statistical segmentation automatically

without the supervision or involvement of a data analyst or statistician.

It could become very expensive to run frequent updates

on a fast-moving database.

Finally, even if you could solve that issue of frequent costly updates,

the optimal segmentation solution will vary over time.

Maybe a bunch of new customers will be acquired and

will create a new segment solution.

Or maybe customer behavior is seasonal, and the best segmentation solution

from a strictly statistical point of view will not be the same,

whether you analyze your data during the summer season or around Christmas.

It makes managers' life very tough, because segments change all the time,

and may not be comparable from one update to the next.

How do you follow customers over time, or

put in place marketing actions specific to a segment,

if the very definition of that segment may be modified during the next update?

Play video starting at :2:5 and follow transcript2:05

For all these reasons, companies often define segmentation as a set of rules.

And these rules are fixed for some period of time, updated only after

a few quarters or years, depending on the needs of the business.

Now, don't get me wrong,

the statistical segmentation we've covered in the first module remains invaluable.

But it's more often used when your segmentation doesn't require frequent

updates, that is, when your segmentation analysis is a one-shot thing.

And even when it requires frequent updates,

it's a very useful exploratory tool to set up a managieral segmentation next.

# Developing a managerial segmentation

When you create managerial segmentation,

you need to make sure the rules you set up, are simple and relevant.

Simple means that you should not create too many segments.

If you do, your segmentation will become too complex and hard to use.

And what's the value of something if it's not used?

Play video starting at ::28 and follow transcript0:28

Relevant means that the segments you define,

need to be relevant to the managers, using the segmentation.

Play video starting at ::35 and follow transcript0:35

Why separate the segment into two subsegments

if you treat the customers in both segments the same way, it's useless.

So, let's create an example of managerial segmentation together.

Let's go back to our data set, and

think about ways managers might be using the marketing segmentation.

Let's suppose the goal is to identify, segments or

groups of customers, that should receive more or less attention.

For example, which customers should receive more catalogs, more coupons,

more emails, phone calls, or direct mail solicitations?

How should we split or segment our database?

Play video starting at :1:16 and follow transcript1:16

We'd like these decisions to be based on who the customers are,

how much they spend, and how likely they'll buy from us in the future.

And one basic criteria to predict, whether people will buy in the future,

is whether or not they've bought recently, that is recency.

Someone who bought from the company a few weeks ago is much more likely to buy

again in the future, than someone who's last purchase happened many years ago.

So, lets take our customers, and

divide them into four groups, or segments, based on their recency.

We define as active, a customer who purchased something within the last 12

months, as warm, someone whose last purchase happens a year before,

that is between 13 and 24 months.

We qualify as cold, a customer whose last purchase was between two and

three years ago.

For those who haven't purchased anything for

more than three years, we qualify them as inactive.

Let's suppose also we like to treat recently acquired customers differently.

Maybe be sending them special coupons, or a welcome packet.

Or maybe we just want to track whether we'll be able to keep those new customers

in the long run.

To that avail, we take all the customers in the active segment, and

put the new ones.

Those with history of just one purchase, so

far, into a segment we call, new active customers.

Now, let's go one step further, and

divide the remaining active customers into two subgroups.

Based on how much money they spend on average.

Based on our analysis, we'll decide to put the bar at $100, and

qualifies highest value.

Those customer's spend $100 or more, on average, at each purchase occasion.

And qualify, as lower value, those who spend less.

In total, we now have six segments.

We are not completely done yet.

To push our analysis further,

we will apply the same criteria to the wrong customers.

So we can decide on which ones we want to concentrate our marketing budget and

recovery efforts.

This segmentation is, obviously, very simple,

and you could imagine much more complex segmentations with dozens of segments,

and many more than only a few segmentations variables used.

But you get the point.

It can be already very useful.

And it's quite relevant for managers.

Play video starting at :3:42 and follow transcript3:42

In the next video segment, we'll actually create these segments in R.

# Coding a managerial segmentation in R (Recital 1)

So now we are going to create natural segmentation.

The first thing we'll do is to load the data as usual always the same one.

Purchases .60 and then as we've done before,

we're going to give the column a name that makes sense,

customer ID, purchase amount, date of purchase.

We're going to tell R that the date of purchase is actually

a debt with the ASdebt function here specifying the format

of the debt that is contained in the text file.

And, once it's done, we are going to compute two things.

We are going to extract the year of purchase,

from the date, using the function we've covered already.

And then same thing extracts, how many days have elapsed between January 1st,

2016 and each and everything purchase is made in this side.

Now it's done, we have exactly the same data as before as you can see with

the had function showing you the first few lines of the file.

As well as the summary function which shows you

a few key statistics which just mean minimum, maximum and signs of fourth.

And as you can see here in the days of purchase, we actually compute a minimum,

a mean, a maximum, and so on, which means that r has correctly identified it

as data that can be computed, averaged, and so on.

So, the next thing we are going to do again, we are going to call the library,

call SQLDF for SQL data frames, managing scripts segments

on our data frames as if they where database tables and we're again to extract

a few indicators we're going to use in these modules segmentation.

The first one of course is the customer ID of each individual and

then we're going to compute recency, which can be defined as

the number of days between January 1st, 2016 and the most recent one.

We're going to also compute first purchase here, which is exactly the same

thing except that instead of taking the most recent purchase,

we are going to take the oldest one, and so it translate into taking the maximum

number of days between any purchase made by one specific customer and January 1st.

Then we're going to compute frequency by counting the number of purchases that

customer has made, as well as the average purchase amount made by that customer.

We extract the data, the query from the data frame called data,

which we've just loaded.

And we group by one, meaning that we group by the first

variable in the query, which is customer ID.

So we'll have only one row of data per unique customer.

Play video starting at :3:9 and follow transcript3:09

And now we have our data as you can see, 18,000 customers,

5 variables, to which you can explore as we've done before.

Looking at the first few lines, off that new data,

computing summary statistics, creating histograms of recency.

As you can see here we've already covered that.

A frequency amount with a finer grain of and

sense of Now, how can you actually

create a managerial segmentation in R?

Well, the first thing to remember, is that most managerial segmentations

are nothing else than if-then-else statements.

Meaning that, if we begin with a very simple statement such as,

I'd like to call as inactive

Play video starting at :4:7 and follow transcript4:07

any customer who hasn't made any purchase for at least three years.

That if else statement will do the trick.

So if you look at that, customers 2015 is basically

the dataset we've extracted with recency, first purchase, frequency, and sum.

And you can create automatically a new variable called segment

in which we'll store a value.

That value will be either inactive or NA.

NA meaning, we haven't assigned any kind of segment yet,

and depending on the tests we do here,

meaning that if customers 2015 viable recency,

so is more than three years,

then we qualify that customer as inactive.

If not we simply don't qualify that customer at all.

If you run that, look at that.

Data five variables.

You run that, and then basically you get six variables showing up,

which is the value of the segments.

Field, out, best, on, few, if, and else test.

Either inactive for those with a recency above three years or

na for those with a recency below three years.

And of course inactive is just one segment, you can have many more,

meaning that you can put multiple if-else statements

together to create the entire imaginal segmentation.

Wait, so haven't not such a good idea.

Let me show you why.

One thing you could do is to check how many customers are in each

segment with a table, statement, and calling the segment columns.

So in the table, you'd get 9,000 customers who are inactive,

and 9,000 customers who are not.

And then you can also compute averages,

with the function aggregate, which we've seen before.

So, the data we're going to average

are pretty much everything within the customer's 2015 data set.

From the second to the fifth variable,

meaning recency, first purchase, frequency, and amount.

And we group them by a list composed of which segment they belong to.

What kind of computation do we apply to that?

The mean, meaning we going to compute the mean of

all these variables grouped by segment.

If you do that, as you can see here, the inactive segment

has an average frequency of 2,178 days,

an average first purchase of 2,500 days,

on average have made 1.8 purchases for an average amount of 48.

So as you can see, the most recent customers,

those ones, who have a recency lower than 3 years,

tend to have made more purchases and for a higher amount on average.

Now, I said before that using If-else statements for

Play video starting at :7:29 and follow transcript7:29

segmentation may not be such a good idea.

Let's see why.

Basically, if you'd like to go further, and not only have an inactive segment,

but also a called segment, what you have to do is to embed within

the if-else statement, another if-else statement.

So basically, if customers recency is above three years, then it's a yes.

We call that customer inactive.

If it's a no, then it depends.

Then we take, test the recency.

Is recency above two years?

Yes, it's gold.

No?

Then it belongs to one of the other segments yet and you could do that of

course if we run and work perfectly.

I'm going to run everything in one row.

So you have 9,000 inactive customers, as before, you have 1,900 cold customers,

you have 7,300 customers who haven't been qualified yet.

And you can see that for recency, first purchase, frequency, and sum,

you get a nice picture of which people are in which segments and how they behave.

But actually, if you have 10, 15, 20 segments,

that if else, if else, if else, could be a complete mess and

it's extremely easy to screw it up, put an if else statement

where it's not supposed to be, can you imagine for instance,

that if you have ten segments, it means you probably have eight or

nine if else statements each within one another.

It can become a real mess extremely quickly.

So usually unless you are applying an extremely simple if

else kind of structure is a really bad idea, it's very hard to read,

very hard to maintain, you are very likely to make tons of mistakes.

So what I suggest, as soon as you get to a slightly more complex

segmentation scheme, my suggest is to use the which statement.

How does it work?

Well first of all, I need to reset everything,

create a variable called segment.

We have it already and set it to everywhere.

So, everybody will begin as an NA, not applicable segment everywhere.

So we've created the variable already, and we can work with it.

That's the key.

Now, for each and every customer, we're going to look at

which match that condition,

which have a recency above three years.

And for those ones, and those ones alone we'll call them inactive and

that's the which here, which is the key.

If you do that, you get exactly the same thing that we got before, okay?

9100 inactive customers, 9200 not applicable segments yet

with the same averages and so on, so forth.

But now it's much easier to go into slightly more complex segmentation

solutions because basically each segment definition becomes it's own line and

it's much easier to follow a quite complex structure.

So here we do exactly the same thing but we'll work with a few more segments.

For instance, we're going to say that any customer with a recency

above three years is inactive.

Any customer with a recency below three years, but above two,

is called any customer with a recency of less than two years,

but more than one will be classified as warm and

then any customer with a recency of less than a year will be qualified as active.

And if you do that,

it's actually much easier to follow through a complex segmentation.

Here as you can see all customers have been qualified as either active called,

inactive or warm.

As you can see the recency matches perfectly

by definition the active customers have a recency of less than a year.

They have been quite active 4.56 purchases of the lifetime.

Quite high average purchase amount and sense of worth and

you have your four segments here.

The only trick if you use that which statement to create a segmentation

Play video starting at :12:14 and follow transcript12:14

of your own is to make sure that a,

you do not forget anyone that if you apply all these segmentation rules,

everyone would be qualified by at least one segment.

And that everyone is qualified by only segment.

Meaning that for instance here, if I removed that,

Play video starting at :12:39 and follow transcript12:39

then basically I would make a huge mistake.

Here, every customer with a recency of above three

years would be qualified as inactive.

And here,

every customer with a recency of above two years would be qualified as cold.

Well, it so happens that if you match that criteria you will match that one as well.

Meaning that none of your customers will be inactive.

They will all be cold and the inactive segment will just disappear.

So you need to make sure that your segment definition do not overlap and

that everyone is going through the segmentation process nicely.

So how does our complete Marginal segmentation look like?

Well actually, it looks like that.

And it's actually not that hard to follow

if you remember what the marginal segmentation looks like.

We need to reset everything to NA then,

we are going to qualify based on recency alone, here.

Play video starting at :13:40 and follow transcript13:40

Inactive, cold, warm, and active customers.

The trick here to make the segmentation easier to manage.

So for instance if you take all those customers who are active,

meaning they have made a purchase within the last year at least.

And you use that segment qualifier as a condition within

the next few lines, such as here for instance.

You can actually make life much easier.

So here, I'm going to select all the customers who've had segments set to

active for the time being and who've made the first purchase within the last year.

Well, if they are active and made a purchase in the last year,

they are new active.

And so I'm requalifying the segment here based on

the segment that they had before and first purchase date.

Or, here as amount.

So if you are qualified as active and you've made an average

purchase of less than $100, then you are qualified as active low value.

If it's above 100, then it's active high value.

And remember, your segments cannot overlap.

And they need to cover everybody.

So, if I did that for instance, let me remove the equals sign,

it means that some customers whose average purchase amount is exactly $100,

will fall in neither segment which is bad.

So you have your entire segmentation working here.

I'm going to apply this whole code together.

Play video starting at :15:27 and follow transcript15:27

You have the number of customers in each segment.

For instance, in terms of a new active, we have 1500 customers here with a pretty

nice explanation of the average Behavior of customers within each segment.

For instance the active high value customers have an average recency of 88,

89 days, made a first purchase 2000 days ago,

made on average 5.6 purchases of the lifetime yet

and have purchase amount on average around $240 and

that's your entire segmentation here working pretty well.

As you can see it's much more than if you had if else statements put together into

long list of Intermingled statements.

There is a few more things I'd like to do with you before

moving to the next stage of this course.

And that's here, as you can see, the name of the segments is alphabetical.

We go from active high, active low, CINNWW.

It's not really relevant from a maginal point of view, so

we're going to reorder these segments in an order that makes sense.

So it can be more easily read by managers and yourself.

So I'm going to take the segment variable here.

Say it's a factor with the exact same variable.

So I'm not changing anything.

I'm just specifying that the levels of that factor are, in that order inactive,

cold, warm, warm low, new warm, active high and so on and so forth.

It doesn't change anything except where

the values are stored internally in a specific order.

And once you rerun that, you get

the least active segments at top the most active reasons,

segments at the bottom and you can see much easier to read and to analyze.

# Describing segments

In the last video segment, we've showed how to create segments and

to assign each and every customer to the segment they belong to.

Play video starting at ::17 and follow transcript0:17

Once this is done the first thing you should do is to describe the database in

terms of segments.

How many customers you have in each segment.

What are their characteristics?

The profiles.

Play video starting at ::30 and follow transcript0:30

One nice way to do that is to compute the average profile of a customer and

use these averages to describe a typical member of each segment.

In statistics, this typical average customer is called a segment centroid.

Managers usually prefer the term segment profile.

One concept closely related to that is the concept of persona.

A stereotypical individual represents an entire segment.

You might even give him a name or invent a story, like, hi, I'm John.

I'm 32 years old and I made my first purchase three months ago for

a total of $50, and I wonder whether I'd make a new purchase in the future.

But one very important use of segmentation,

is that you can go back in the past and see how many customers were in

each segment a year ago or two years ago, and see how things have evolved.

Do you observe an increase in the number of customers in the high value segment?

What about the UK customers?

Have you added more new customers this year than the year before?

What's the trend?

Play video starting at :1:37 and follow transcript1:37

One trick to remember, however, and that's crucial, is that some customers who are in

your database today were not there a year ago, or even a few weeks ago.

Since each customer needs to belong to one, and only one segment.

If you go back in the past,

you need to create a segment of non-existent customers.

Because that's exactly what they were at the time.

Non-existent.

If you fail to do that, you will assign them, by default,

to a segment that did not belong to.

And your analysis will be flawed.

In it's a bit tricky, so let's make sure we get it right.

In the next video segment we go back in the past by a year.

And perform segmentation analysis in as if we were a year ago.

But on the data that's available to us today.

# Segmenting a database retrospectively in R (Recital 2)

How to segment a database retrospectively.

That's the point of this tutorial.

We are assuming that you ran the code up to line 100,

which is what we've covered together in the previous segment.

So I'm going to run it.

Play video starting at ::21 and follow transcript0:21

Here we are, and that's exactly where we left off this

tutorial, so how do we do that?

You probably wondered why I labeled my data Customers2015, here at that stage.

Well, simply because we're going to create the exact same file,

called customers\_2014,

which is the segmentation of the database as if we were a year ago.

So, how did it work?

Play video starting at ::54 and follow transcript0:54

Well, the first thing to do is to remember that we are a year ago.

Meaning that whatever data we take into account, anything that has happened over

the last 365 days should be discarded, as if it never happened.

When we use the where clause in the SELECT statement,

what we are doing in SQL is you say, just take into account those data points,

those lines, that match this condition.

So as long as days\_since is above 365 days, then you can take it into account.

Otherwise, just consider the data is not even there.

And everything else is almost identical.

We select the customer ID, the minimum days since which we call recency,

except that days since has been computed compared to January 1st, 2016.

We want to compute that as if we were a year ago.

So we are going to remove 365 days of that figure.

Same for first purchase.

Whatever that value is, we compute it as if we were a year ago, we remove one year.

And then we count how many purchases have been made.

We call it frequency.

Remember that any purchase that happened within the last year will be excluded.

Play video starting at :2:25 and follow transcript2:25

And here, we compute frequency a year ago.

Same for average purchase amount,

we only compute the average of purchases made before a year ago.

Play video starting at :2:36 and follow transcript2:36

And that's the statement.

Now interestingly, you should compare

2015 and 2014 we miss customers.

Some customers are not there.

Play video starting at :2:52 and follow transcript2:52

We have 16,900 customers in the data file, and

we have 18,400 in that one.

Play video starting at :3:1 and follow transcript3:01

Which is a bit weird, since we are analyzing exactly the same data.

Well it's not weird, because we've just excluded all those customers who did not

appear before January 1, 2015.

So we are excluding all those customers who have been qualified,

in a sense, as new customers within the last year.

Play video starting at :3:23 and follow transcript3:23

And if you look at that, we had 1,512

customers in the previous segmentation that qualify as new active customers.

So just compute 1,512 plus the number of customers we had a year ago,

is equal the number of customers we have today, 18,417.

Knowing that, all we need to do is apply the exact same segmentation.

And because we computed the recency, first purchase and

everything else as if everything was computed a year ago,

we don't have to change anything in the segmentation scheme here.

We just apply it, as if it were applied a year ago, on the data at the time.

Play video starting at :4:15 and follow transcript4:15

Once you have that, we do the exact same thing we've done before.

Meaning we are going to reorder the segment names in an order that makes

sense to us.

Play video starting at :4:26 and follow transcript4:26

And we look at the segmentation results.

Play video starting at :4:31 and follow transcript4:31

So here, in terms of number of customers,

a year ago we had 1,437 new customers,

we had 7,512 inactive customers, and so on so forth.

These figures have changed over the last year.

We added new customers.

New customers were considered old.

Inactive customers increased over time, and so forth.

If you look at the chart, as you can see here,

about a third of our customers are inactive.

Play video starting at :5:6 and follow transcript5:06

Probably ten persons are new active, active low value, active high value, and

so on and so forth.

That's how customers are distributed across segments in the database.

Now as you finally compute the averages of all

the variables of interest, first purchase, recency, frequency, amount, and sum for

per segment at the time, meaning a year ago, you have these updated figures.

Play video starting at :5:37 and follow transcript5:37

That's how you compute a segmentation retrospectively.

You go back in time, assume the data that has been generated over the last year,

for instance over the last period did not even exist.

Adapt how you compute recency, frequency, monetary value and accordingly.

And then you just apply everything you've applied before, same segmentation,

same transformation, same analyses, and same tables.

# Segments and revenue generation

From a managerial point of view,

it is also extremely useful to see not only to

what extent each segment contributes to today's revenues.

But also to what extent each segment today would likely contribute to tomorrow's

revenues.

For today's revenues, the analysis is quite simple.

You just need to sum up revenues by segment.

How much of your revenue is being generated by the current period

by the newly acquired customers?

The low value or the high value ones.

And so on.

You will often observe that some segments, even small ones in terms of number of

customers, generate a huge share of revenue.

Play video starting at ::46 and follow transcript0:46

What's even more interesting,

is you have a forward looking analysis of revenue generation.

That is, from what customers, today, will come tomorrow's revenue.

Play video starting at ::57 and follow transcript0:57

Will your active, high-value customers remain loyal and profitable next year?

How much revenue will your newly acquired customers generate a year from now?

And should you expect a lot of revenues from your currently inactive customers or

should they be considered lost?

Obviously, you cannot answer these questions for

sure because tomorrow has not happened yet.

Only the future will tell.

However, if you look at the recent past, you'll be able to have a pretty good idea.

Customers in one segment today are likely to act pretty much the same

as the customers in that segment a year ago.

So analyzing the past will inform us about the most likely future.

Play video starting at :1:40 and follow transcript1:40

In the next video segment, we'll see how you can do that in R.

Oh, and by the way, the very next tutorial is essential.

Because it prepares the way to scoring models,

and customer lifetime value models.

So make sure you review it carefully.

# R tutorial (Recital 3)

In this tutorial, we're going to get into really interesting and

managerially relevant issues, such as, where does money come from,

and how is revenue distributed across segments.

Play video starting at ::16 and follow transcript0:16

So the first thing I'm going to do, as always,

I'm going to execute everything we've been up to, which is line 140.

That's where we are.

Play video starting at ::27 and follow transcript0:27

We are going to run a select segment of a new kind.

Here it is.

We're going to select, for each customer,

how much purchase has [INAUDIBLE] made in 2015?

We'll do that, we name that value, revenue 2015,

it's just a quick name we're going to give it.

We extract the data from Data.

And of course, because we are only interested in purchases made in 2015.

We're only going to apply that computations on those purchase

where the year of purchase has been set to 2015.

And because we want that figure to be grouped, and

specific to each customer, we are going to group by one,

meaning we are going to group by the first variable that appears on that statement.

Let's execute that, and we have generated

a new data frames, called revenue 2015.

Now, here is the issue.

Revenue 2015 only contains

5,398 observations, or lines.

Why so?

Well simply because, and that's a regular issue with select statements.

Believe me, if you deal with SQL statements,

you'll have that issue over and over and over.

Play video starting at :2: and follow transcript2:00

Only those customers who have made at least one purchase in 2015 will

Play video starting at :2:5 and follow transcript2:05

appear in the output, and in this case,

only 5,400 customers have made a purchase in 2015.

Of course we'd like to compute revenues for everyone,

not only those active customers, but that's a different story.

We'll get to that.

If you look at the summary of these figures,

again the minimum of yearly purchase, per customer, is $5.

It should be zero if we included everyone, but

we only include those who've made a purchase in 2015.

The average is 88, and the maximum is 4,500.

So, how can we merge those customers

from 2015, to their revenue?

Well, the wrong way to do that, would be to do this.

Actual equal, you merge those customers with the revenues that they generated.

Except that the variable you've just created, actual,

has exactly 5,398 observations.

Play video starting at :3:19 and follow transcript3:19

You've just erased from your data, or at least you are not taking into account,

all those customers who haven't made any purchase.

How can we make sure that customers who haven't made

any purchase are still up there in our statistics?

Well what we are going to do is to use the exact same function,

merge, but we'll add one option here, all.X = true.

This is the data frame x, this is the data frame y.

So, when you say all x, it means in the output, any line

that appears on the left should also appear in the output, in the final output.

And by the way the reason merge does work here,

is because in customers 2015 and revenue 2015,

both data frames have Customer ID as a key.

And so they can merge both data frames, and match each and

every line from one data frame to the next.

If we use that option, all.X = true, then basically,

the actual data set here contains 18,417 observations.

We have one line for each and every individual, in the original data set.

Except that, in revenue 2015, either we have a value,

either 80, 45, or we don't have any value, and

we have N/A for not applicable.

But obviously, these not applicable revenue are actually equal to 0.

And so we are going to select from revenue 2015,

all those which are NA that are a functioning R,

and replace them with the value 0.

Once we have that, the actual data set we are going to use,

makes much more sense, for each and every customer,

such as customer number ID 10.

We have recency, first purchase, frequency, amount,

the segment to which they belong, and how much revenue they have generated in 2015.

You could compute the average revenue generated by each customer,

grouped by segments, and that's actually what you're going to do here.

We're going to compute an aggregate of the revenue in 2015,

grouped by segment to which they belong to.

And we're going to compute the mean of that revenue per segment.

And as you can see here, by definition, if you are inactive, cold,

warm, you haven't generated any revenue in 2015,

because by decision you haven't purchase anything in 2015,

otherwise, you wouldn't be qualified as warm, cold or inactive.

But if you are active, you have made some purchases in 2015,

depending on which segment you belong to.

That value might be pretty low or rather high, if you're an active high value

customer, on average, you've generated $323 of revenue,

if you're a new active customer, on average almost 80.

So, as a quick summary, we've computed

Play video starting at :6:56 and follow transcript6:56

the revenue generated in 2015, for each and every customer,

Play video starting at :7:3 and follow transcript7:03

which we saved in a data file called revenue 2015.

Then we've merged our complete list of customers with the revenue

generated by some of them, making sure that with that option,

we would keep all the customers in the original data sets, on the extreme left.

And then, we replaced all the NA values by 0.

And computed the average of revenue in 2015,

grouped by segment, and that's what we got.

Well, the statistics we've run are interesting, but they are a bit obvious.

Obviously, if you are an inactive customer, by definition,

you haven't generated any revenue in the year you were inactive.

Play video starting at :7:56 and follow transcript7:56

What we'd like to do now, is to estimate how much money you can expect.

How much revenue you can expect from an active customer today, next year.

So, of course we don't know that.

We don't' know the future, we don't know exactly what's going to happen, but

the one thing we can do, is to go back in the past.

And look at how much revenue we got from inactive

customers in 2014, going into 2015.

And that's the next step of this analysis.

So what we'll do, is to merge the revenue generated in 2015, as before.

But we're going to merge them with the customer list of 2014.

And so we're going to look into, how much revenue's been generated by each customer,

based on the segment they were in, a year ago.

Play video starting at :8:52 and follow transcript8:52

And that's why I call it forward.

Forward, as the segment in 2014 will enlight us,

about how much revenue have been generated in 2015 from these customers.

Play video starting at :9:5 and follow transcript9:05

I create the file, again, I replace all the NA by 0, and

I show and compute the aggregate revenue for each.

What does it means?

It means, for instance here,

that an inactive customer in 2014 has generated,

on average, about $3 of revenue in 2015.

Many of them have generated nothing, and remained inactive.

Some of them became active again and spent some money.

And on average, these two figures averaged to $3.

Cold goes to 6.

And look at that.

A warm, high value customer generated an average,

close to $114 in 2015.

So, a year later, which is interesting, since a warm,

high value customer is actually worth a lot more money than, for

instance, a new active customer, who has just made one purchase,

and might not remain active a year from now.

And if you reorder and look at the data in a different angle,

so what I'm going to do here is to reorder these figures, in decreasing order.

Play video starting at :10:32 and follow transcript10:32

And show you, which customers are the most profitable a year later.

Active high value customers come first with 254,

warm high value customers come second with 114, and

look at that, the new active customers only come fourth with

only an expected revenue of $31, coming in next year.

And if you plot that, as you can see, all these are segments in 2014,

and these bars represent how much revenue you could have

expected from a customer in any of those segments a year later, in 2015.

Now interestingly, losing one active high value

customer is like losing six new active customers here.

You've lost six times more money by losing the potential sales

coming from one customer in that segment, compared to one customer in that one.

And so, from a managerial perspective, it's really interesting

to understand that a customer in the active high value segment,

Play video starting at :11:49 and follow transcript11:49

going in the next 12 months, is worth six times more,

than say, a new active customer.

So, in terms of resource allocation, and how much money you want to invest,

in keeping good relationships with certain customers,

depending on which segment they belong to, that kind of information is priceless.

# Instructions before starting quiz 2

To validate Module 2 and correctly answer the questions will require that you perform the following exercise first: take the code in the file named module2.R, and modify it such that the managerial segment "new active" is divided into two sub-segments called "new active high" (with an average purchase amount of $100 or more) and "new active low" (with an average purchase amount of less than $100). Tips: make sure that you apply that modification to both the 2015 and 2014 segmentations, and that you modify accordingly the code that re-order the factor "segment".

**You'll be asked to answer these questions :**

How many "new active low" customers were there in 2015?

The number of "new active high" customers has increased between 2014 and 2015. What is the rate of that increase?

Regarding the customers who belonged to the "new warm" segment in 2014, what was there expected revenue, all things considered, in 2015?

In terms of expected revenue, which segment groups the least profitable customers?

Looking at segment description, what is the average purchase amount of a customer who belongs to the "new active high" segment?

[module2](https://d3c33hcgiwev3.cloudfront.net/_ddcfaa0519d900af2132cc52b9c363e7_module2.R?Expires=1703808000&Signature=ZCjXXkGgmO4OCnviAe~I-ngeJthZZ0tht43jmfhvFv5lT5iXfnhXt2DHz4raGv1UPkvYyWC9H2T6EFdAwqhNbH6H3pp4tAdjfnKZrvZd9mT-ARNSSByrQXRYcAY9ITGUK4RoCm4hgE9Ovl7ZuQx~2v~yzU8x~OGMCP69anY5U78_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[R File](https://d3c33hcgiwev3.cloudfront.net/_ddcfaa0519d900af2132cc52b9c363e7_module2.R?Expires=1703808000&Signature=ZCjXXkGgmO4OCnviAe~I-ngeJthZZ0tht43jmfhvFv5lT5iXfnhXt2DHz4raGv1UPkvYyWC9H2T6EFdAwqhNbH6H3pp4tAdjfnKZrvZd9mT-ARNSSByrQXRYcAY9ITGUK4RoCm4hgE9Ovl7ZuQx~2v~yzU8x~OGMCP69anY5U78_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[purchases](https://d3c33hcgiwev3.cloudfront.net/_e711af551fbda7ea1239aa9afd734cfd_purchases.txt?Expires=1703808000&Signature=E81liSN0KUb0alQScgkqT294inVKcwqimCy3V2HvdElB6kjw1lSR9igCc5YzizqayoM4XIfmtvrUSbcQr58BeAiBlS~ZEXN9~EjQnUZeWBV0~Z3RcwQUZgQtL8h6U6GDRH2OqU39BODR4eJO07a~bV~0hMjQBviN7HSTKPpZz6Y_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[TXT File](https://d3c33hcgiwev3.cloudfront.net/_e711af551fbda7ea1239aa9afd734cfd_purchases.txt?Expires=1703808000&Signature=E81liSN0KUb0alQScgkqT294inVKcwqimCy3V2HvdElB6kjw1lSR9igCc5YzizqayoM4XIfmtvrUSbcQr58BeAiBlS~ZEXN9~EjQnUZeWBV0~Z3RcwQUZgQtL8h6U6GDRH2OqU39BODR4eJO07a~bV~0hMjQBviN7HSTKPpZz6Y_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

<https://github.com/nicolasfguillaume/Marketing-Analytics-with-R/blob/master/problem_set_2.md>

# Can Target predict a customer is pregnant?

In February 2012, Charles Duhigg published a front page New York Times

magazine article entitled, How Companies Learn Your Secrets.

In that article, the reporter explains how to target American retail store.

Use preanalytics to identify which of its customers were most likely

to be pregnant, based on their recent purchase history in browsing behavior.

Now, being pregnant is usually not the things companies try to predict.

They usually prefer to focus on more mundane customer behavior,

like the probability of buying a product, if it's recommended to you,

or the likelihood that the customer will use a coupon, or click on a link.

Think about how Amazon recommends products to you,

or how Google decides to show you certain ads, and not others.

There are mathematical models behind all those decisions.

Companies may also try to predict how much a specific customer is likely to

buy of the next quarter, or her likelihood of terminating a contract, and

switching to the competition, even before she actually decides to do so.

So the company can act in that knowledge and try to prevent her from leaving.

Play video starting at :1:22 and follow transcript1:22

But if you think about it,

to predict customer pregnancy makes perfect business sense.

Given that there is a tremendous sales opportunity, when a family prepares for

a new born, you can see the marketing potential.

Now to be fair, the story about pregnancy prediction has been

inflated by the media, as they usually do.

The analytics target doesn't have superpowers, and any creative model,

can only assess that some people are more likely to be pregnant than others,

sometimes with great accuracy, but not always.

Sure enough, if you are a 27 year old woman who suddenly changes her buying

habits, begins to buy prenatal vitamins, doesn't buy feminine hygiene

products anymore, and cuts down on the amount of bacon she eats every week,

chances are she's more likely to be pregnant than me.

Play video starting at :2:17 and follow transcript2:17

The article is interesting, nonetheless, because it shed light

on the ability of companies to exploit the data they have about the customers, and

predict things we did not even think were possible.

Play video starting at :2:31 and follow transcript2:31

Interest companies and banks have done that forever,

trying to predict the likelihood you'll default on your debt, have a car accident,

or even die, indicates of life insurance, and

they adapt their codes bears on these predictions.

It's disturbing but we're used to it.

Now, predictive analytics are becoming mainstream in marketing, as well.

In this module, you learn how to do it yourself.

# What you need to develop a scoring model

Let's think about the example we just covered in the previous video segment, and

try to think about how Target could predict whether a customer is

pregnant or not.

Target is storing data about its customers.

As soon as you fill in a form,

use your store loyalty card when making a purchase, or login to the website.

They can track, store, and retrieve information about you.

They know who you are, where you live, what you buy, and how often you buy it.

Play video starting at ::36 and follow transcript0:36

Thinking about the pregnancy prediction example there is a bunch of indicators

that could work pretty well to identify who might be pregnant.

Well the first one is gender.

It's probably safe to assign zero probability of being pregnant to

all male customers in the database.

Another predictor is age.

You are definitely more likely to be pregnant if you are in your twenties to

thirties than if you are in your fifties,

although we can't be absolutely sure about that.

But there are also a lot of subtle indicators which in and by themselves

are not saying much, but when combined together tell an interesting story.

If a woman begins to buy more comfortable and slightly larger clothes than usual.

Suddenly decides to switch soap brand and buys only hypoallergenic products.

Doesn't buy tobacco products anymore, and changes her eating habits.

All these indicators put together might give us a clue

about whether she's pregnant or not.

But understanding that intuitively is not enough.

What you need to do is to take these predictors, all the potential predictors,

and quantify the impact on the likelihood you are trying to asses.

Whether it's the chance of being pregnant, or the risk of switching to a competitor.

Play video starting at :1:55 and follow transcript1:55

Suddenly buying a different brand of soap may be a very bad predictor.

And it doesn't predict anything.

Maybe buying prenatal vitamins does.

How do you know?

You need two things.

You need a calibration data and you need a statistical model to analyze it.

That's what we'll talk about in the next video.

# Calibration data and statistical model

Calibration data is a dataset in which you observe both the predictors, and

the variable you are trying to predict.

And one of the most common way to get that kind of information

is to go back in the past.

How does a bank predict the transfer default on a loan?

Well, they will look at people who are dressed like you a year ago,

or five years ago, and see how many of those defaulted.

How did Target build a scoring model to predict customer pregnancy?

Because some customers went to Target, willingly, and

disclosed they were pregnant.

They did so by creating a baby registry.

And once Target had that information about which customers were pregnant, and when,

and what they bought or did not buy, they could mine the database to find patterns

that could have predicted what they already knew for a sample of customers.

Play video starting at :1: and follow transcript1:00

After building, and calibrating, the predictive model, they could apply it to

everyone in the database, and assess everyone's probability of being pregnant.

Play video starting at :1:11 and follow transcript1:11

Now, let's forget about the pregnancy example for now, and

focus on what we'll do with our datasets.

In this course, you are going to predict how much money

customers are going to spend over the next 12 month.

Play video starting at :1:25 and follow transcript1:25

And to do so, we are going to create a calibration data.

We are going back time, 12 month ago, and

extract from that data, two separate types of variables.

The first type of variables are called predictors.

These are the bits of information we had about each customer at the time, as if

we went back in time and had no clue about what they did over the next 12 months.

For instance, we compute what was their recency, frequency, and

monetary value, for each customer, a year ago.

Actually, we've done it already in the previous module,

when we ran our segmentation model retrospectively.

So we have that already.

The second type of viable is what happened after that.

It's the variable we try to predict, it's sometimes called the target variable.

In our case, we'll try to predict whether a customer will remain active and

make at least one purchase.

And if so, how much money each customer is likely to spend.

In other words, we predict both the probability and the likely amount.

And because we'll build our calibration data by going back in time,

we already know the answer.

Because we observed it for each and every customer.

Once we have built our calibration data, containing both the predictors and

the target variable, the next step is to link the two through a statistical model.

The model will predict customers probabilities of purchase as well as

the most likely amount.

And in the process, reveal the importance of each predictor.

The importance of each predictor will be revealed by what we call weights and

their statistical significance.

If the weights are large and statistically significant,

it means they are good predictors.

If not, it means they contribute very little to the predictions.

The best way to explain this, is to dig into our R code and

see how it's done in practice.

And that's exactly what we are going to do in a few minutes.

But before we get there, let's remember something.

What we are going to do is very typical in marketing applications.

We are going to predict the probability that the customer will buy something.

And if they do, predict how much money they'll spend.

So, we are actually going to model two separate processes.

The first one is to predict the probability.

The second one is to predict an amount.

To do so we will build two separate models that we will

combine together later into one big scoring model.

But the calibration data will not be exactly the same.

To predict the probability of buying, we can use the entire of customers

because we can observe for all of them, whether they purchase something or not.

If you bought something, we'll code it as a one.

If not, we'll code it as a zero.

We can observe that behavior or lack of behavior for everyone and

use it in our calibration data.

For the amount spent, however, it's a different story.

If a customer hasn't purchased anything over the last 12 month,

we have no way of knowing how much she would have spent, had she made a purchase.

So, for the first model to predict the probability of purchase,

we'll use the entire database to calibrate our predictive model.

For the second model, the one that predicts how much money you'll spend if

you buy something, we will only include, in our calibration data, those customers

who purchased something over the last 12 month, and ignore the others.

How to do that will be explained in detail in the next video.

# Building a predictive model in R (Recital)

In this third module we're going to cover a complete a creative modelling and

scoring application from A to Z, so bear with me.

The first step is to again, as usual, load the data,

rename the columns correctly, compute date of purchase,

year of purchase, the number of [INAUDIBLE].

January 1st 2016, which will later be used to compute recency.

I'm going pretty quickly over all these things.

And the first step is basically to extract all the predictors

that we're going to use in our predictive model.

Remember that as we said the predictors are variables computed a year ago,

and these data will be used to predict what happened over the last 12 months.

So, exactly as we've done in the previous tutorial we're going

to compute everything we know about customers at the end of 2014.

These are the predictors.

And then we are going to look at what they did in 2015.

These are the target variables.

We are going to predict.

The next step is to merge these two together.

And again, exactly as in the previous tutorial for

module number two, we're going to merge customers from

2014 with the revenue they generated in 2015,

making sure that all the customers in the first data sets will remain in the data.

We are going to call the data sets the in sample,

meaning that we're going to run in sample predictions, and

then later on we're going to run out of sample predictions on customers in 2015.

Again, we merge everything, we transform the not applicable

values into zero, and we have our revenue in 2015,

which is how much money they've spent in 2015.

And many of them spent zero.

We're going to create a new variable.

Whether they spend anything, we're going to call that variable active 2015.

And, basically, we look at revenue in 2015.

If it's above zero, it's a yes.

If it's zero, it's a no.

And we'll store that value as numeric.

So instead of storing true faults, we're going to store zeros and ones.

Pretty much pretty standard here.

We execute everything, and the next step, just to look at what we've created.

Play video starting at :2:52 and follow transcript2:52

This is the data we're going to calibrate our creative models.

We have recency, first purchase, frequency, average donation amount,

maximum amount spent, which is a new variable we're going to use [INAUDIBLE]

and then how much they've spent, and whether or not they've spent anything.

So obviously, if there is a zero here, you have a zero there.

Either is a positive value.

You have a one over here.

Play video starting at :3:22 and follow transcript3:22

That's our calibration data.

Play video starting at :3:25 and follow transcript3:25

Now, we are going to calibrate the first model, which is the probability model.

The likelihood that a customer we'd be active in 2015 or not.

We are going to use the NNET library,

which contains the function Multinom.

Multinom stands for Multinomial Model.

It's an extremely useful model to predict outcomes that can either be zero or

one and nothing else, which is exactly what we'd like to use.

So, we're not going to use traditional linear models as we could use later on.

We use the binary model where the output can either be zero or one.

Play video starting at :4:8 and follow transcript4:08

And here is how is works, the output of the model which I call prob

dot model is the output of the Multinom function with formula

that states that active 2015 is a function of recency

first purchase as frequency, average amount and maximum amount.

Which is five of we have introduced just for fun.

And the data is the data we created the calibration data called In sample.

Next, these will fit, calibrate the entire model on the data set.

And then we can going to extract the coefficients and the standard

deviations of these coefficients, and output not only the coefficients,

the standard deviations, but the ratio of those as well.

Let me execute these first, so it converged,

the model converged with, if you look at the sign of recency.

The recency parameter, for instance, it's negative.

Which makes perfect sense, right?

The larger the recency, meaning the more days have lapsed between the last purchase

and today, the less likely you're going to make any other purchase in the future.

So, if your last purchase was three four ten years ago,

it's extremely unlikely that you'll make any purchase very soon.

Meaning that the sign of the primary is negative, the higher the recency,

the lower the probability.

However, if you look at frequency, that primary is positive.

Meaning that the more purchases you've made in the past, the more likely

you'll make additional purchases in the future, which makes perfect sense.

Now, these two parameter values are the most interesting,

simply because if you look at standard deviations and

at the ratio between coefficients and standard deviation, which usually indicate

to what extent each parameter value is significant, or not.

If it's above two, or below minus two, usually it's a good sign, and

as you can see here, recently is huge, minus 32.

Way, way below minus two, so It's highly significant.

Frequency is also the ratio of the frequency coefficient and

it's standard deviation is extremely high as well, close to 15, but all the others

are pretty close to zero or at least not as good and not as large as the other.

So the impact of first purchase, average amount, and

maximum amount on the predictions is actually pretty limited.

So now we have created our model, the probability model.

And we have stored everything we need in that

variable to later on make additional predictions.

What we'd like to do now is to predict if you're going to be active,

how much are you going to spend with that specific retailer over the year 2015?

The issue here, as we've discussed In the previous video.

That model can only be calibrated on those customers who actually purchased

something, and so we need to sub sample to only take those customers who

were active in 2015, so we can calibrate an estimate how much they spend.

Play video starting at :7:51 and follow transcript7:51

And feet the model.

So what we are going to do is take the sample, the viable 2015.

Look which ones are equal to one and we'll store the index

of those customers in a variable that we'll call Z for the time being being.

Z would be a vector indicating which customers have been active in 2015, and

only on those customers will we actually calibrate the second monetary model.

So we're going to run that.

If you look at the head of the data with only the index Z we've retained,

as you can see, all these customers have active 2015 at one,

which is exactly what we wanted.

And all have spent something in 2015.

And, finally, if you look at the value active 2015, everything is equal to one.

Play video starting at :8:52 and follow transcript8:52

We only have active customers in there.

In terms of revenue, customers have spent anything between $5 and

$4500 with that retailer of the year 2015.

Now, what you are going to do is calibrate the monetary model.

Meaning we begin to predict how much they spend in 2015 based on

only two things here, the average amount they spend usually,

and the maximum amount they spend.

So we have two different predictors.

Here we're not going to use the Multinom function,

because the output is not something either zero or one, it can be anything.

LM, which stands for Linear Model,

will fit a linear model to match as closely as we can revenue 2015 based

on the predictor's average amount and maximum amount.

And the data here is not the entire sample, but

only those customers who can be found in the index, Z here.

So only those customers who actually spend something.

Play video starting at :10:2 and follow transcript10:02

Say if we run that model,

it should be extremely quick, and show a summary of that model.

As you can see, all the statistics here

estimate parameter standard errors t values, everything is highly significant.

Play video starting at :10:19 and follow transcript10:19

Our square value is point 60,

which is basically a signal of the fit of the model.

But, we have a slight issue here.

And the issue is that, let's plot on one hand how much has been spent by customers.

And on the other,

we take the amount model we've just created here through linear regression,

and look at the fitted values, which are the values predicted by the model.

If we plot that, Well,

the chart will look terribly ugly.

And the reason is that most customers have spent pretty small amounts.

50, 60, 70, 100, even $200, and

a few outliers have spent huge amounts, up to three four thousands.

And so basically the model is trying to fit a line through this cloud

of points where actually no line is clearly a good fit.

So what we're going to do, very much like what we did in the segmentation model,

instead of creating a model with

Play video starting at :11:37 and follow transcript11:37

the amount here we're going to create an amount with log.

So instead of predicting revenue 2015 based on average amount and

maximum amount.

We're going to predict the log of revenue 2015 based on

the log on average I purchase amount and maximum purchase amount.

Doing that.

Play video starting at :12:2 and follow transcript12:02

Actually, as you can see, the r square here has improved,

meaning that we fit the data much better.

And if you look at the plot, then it makes much more sense.

So we've put more weight to the smaller values.

Less weight to the very large values, and here you can see,

you can imagine a pretty nice line going through that cloud of points predicting

Play video starting at :12:33 and follow transcript12:33

the revenue of 2015 best on the model we've just created.

To summarize we have just calibrated two models the first one over here,

to predict the likelihood that someone will be active.

And the second one over here to predict how

much they will spend if they will be active in 2015.

Now the end game of this exercise is actually to

apply the models to predict the future.

So what we are going to do is to look at today's behavior,

today's data, and extract exactly the same information

about today's customers as we used in terms of predictors about a year ago.

So we're going to extract recency, first purchase, average amount, maximum amount,

and everything else for the 2015 customers about whom of course

we have no idea who will be active next year, and how much they'll spend in 2016.

But we can try to predict that, with our model, so

once you create that and that what is usually called the out of sample data set.

Play video starting at :13:47 and follow transcript13:47

You have all your customers at the end of 2015, 18,417 people.

And you are going to predict their probabilities to be active in 2015,

based on an object which we've created that's our probability model.

The new data is customers 2015, and the type of predictions we are going

to make are the actual probabilities that's the primary of the model.

The probabilities that someone will be active or not.

So we add we create a new value the prob predicted,

the probability predicted, to the data set we have and

that column will actually contain the predictions from the probability model.

We're going to create another column called revenue predicted,

where the predictions we come from the amount model.

In here we're going to apply that to the entire data set.

However, remember that the amount model

is actually predicting the log of the revenue.

So you should like to the actual revenue,

you need to take the exponential of the predictions,

since the log is the inverse of the exponential, and vice versa.

So, you predict the log of the amount using the amount model,

and then you exponentiate that to get the actual revenue predicted, and

the score the actual score of your customers is the conjunction

of probability predicted multiplied by amount.

So if you have a 10% chance of buying for $100,

your score will be 10% of 100, which is 10.

So we run these three lines

Play video starting at :15:47 and follow transcript15:47

And then we summarized the results.

If we look at predicted probabilities on average, people,

customers in the database have a 22.5% chance of being active.

Some customers are predicted to be be absolutely certain to be

active close to one.

Other customers are predicted to be absolutely certain to be non active

close to zero.

And many are in between.

If you look at predicted revenue, so if they are active how

much they are going to spend next year the average is $65.

It goes between six and 38,000 here.

And of course it's six and

not zero because it assumes that you will spend something.

So the overall score Is actually a function of both and

probability and revenue together.

And the score has a mean of 18,8.

What does it mean?

From a managerial point of view, that value is extremely important.

It means that, on the average, every customer in this database,

will spend 18 point eight dollar next year.

Some will spend zero, a lot of them will spend zero.

Some will spend maybe $333.

Some will spend 50, and so and so forth, but on average it will be 18.8.

Some have a score very near zero.

We don't expect anything coming from them.

Some have a score extremely high,

meaning they're potentially extremely profitable for the firm.

And if you look at the histogram, of course, most people are around here

you could actually create a histogram of the log or look into more details.

But what we do is a slightly different exercise.

Play video starting at :17:52 and follow transcript17:52

We'll take the customers here.

Look at their predict score and only retain those with the score.

Above $50, and if you look at that,

it will create a vector of people with a score above $50,

which contains a total of 1323 customers.

So in the list of 18,000 customers here about

1,300 have a predicted score of $50 or

more and you can see which ones they are.

Here you have the index of all the customers that have

a predicted score above 50.

And if you like to apply targeting applications, if you'd like to identify

the customers with which you should spend the most marketing dollar.

Those customers are the ones with the high score, obviously.

# Instructions before starting quiz 3

To validate Module 3 and correctly answer the questions will require that you perform the following exercise first: take the code in the file named module3.R, and modify the probability model such as the predictors taken into account exclusively include (1) recency, (2) the log of recency, (3) frequency, (4) the log of frequency. Also, list the detailed predictions made for the first 10 customers.

**You'll be asked to answer these questions :**

In the modified model, which of the following predictors does not seem to bear a significant influence on probability predictions?

Suppose a customer has a 20% probability of being active, and a predicted revenue of $200 if he is. Which of the following statements is not correct?

Looking at the predictions made for the first 10 customers in the database, which of the following statements is not correct?

Among the following customers, which one is expected to be the most profitable for the firm?

In the probability model, why is the sign of the recency parameter expected to be negative?

[module3](https://d3c33hcgiwev3.cloudfront.net/_95c88f77b8100bcf4cf6afed5a3aa03c_module3.R?Expires=1703808000&Signature=NNweBt1Se5TUJE5za42imChC45r7Y~qfFcLCB397JpDhVOPtH8VJ4T65qMQggI-vOBcpvmHFDNfmuh76RoP0pj9uS6HXiKLuZl~oKWMjVR5y6UAOXFVojDJn0iPQmc3-7QkqFMsx4hnEJkfDBGcCXBRMolVtSfuAPJjZRhUxa74_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[R File](https://d3c33hcgiwev3.cloudfront.net/_95c88f77b8100bcf4cf6afed5a3aa03c_module3.R?Expires=1703808000&Signature=NNweBt1Se5TUJE5za42imChC45r7Y~qfFcLCB397JpDhVOPtH8VJ4T65qMQggI-vOBcpvmHFDNfmuh76RoP0pj9uS6HXiKLuZl~oKWMjVR5y6UAOXFVojDJn0iPQmc3-7QkqFMsx4hnEJkfDBGcCXBRMolVtSfuAPJjZRhUxa74_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[purchases](https://d3c33hcgiwev3.cloudfront.net/_755d8b708343dc96024c0cfa4523c532_purchases.txt?Expires=1703808000&Signature=eRLO5yi~kKz4-8OKXEfJWqOobubsgzYQjLA4PBwOlRG9lth7oprtgYGYQZd0QEv~VYO727WZx9gw2BSbPrw6ip4RyeJ-SapiyGaAv0hDQvvkBMOn-yVBLbTpiK7b8kNvaLPZxZ-4zyd2qGMT1BwRBS0bZNnF2A3-yF8NPYG06bk_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[TXT File](https://d3c33hcgiwev3.cloudfront.net/_755d8b708343dc96024c0cfa4523c532_purchases.txt?Expires=1703808000&Signature=eRLO5yi~kKz4-8OKXEfJWqOobubsgzYQjLA4PBwOlRG9lth7oprtgYGYQZd0QEv~VYO727WZx9gw2BSbPrw6ip4RyeJ-SapiyGaAv0hDQvvkBMOn-yVBLbTpiK7b8kNvaLPZxZ-4zyd2qGMT1BwRBS0bZNnF2A3-yF8NPYG06bk_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

<https://github.com/nicolasfguillaume/Marketing-Analytics-with-R/blob/master/problem_set_3.md>

# What is customer lifetime value and why it matters

Imagine the following situation,

your company launches a large scale customer acquisition campaign.

Your firm invests in public relations, advertising, email campaigns.

It buys a host of key words online and set up a promotion and

offer a hard discount for customers who purchase for the first time.

Play video starting at ::28 and follow transcript0:28

At the end of the campaign, because it's so hard to get through to the clever and

to reach, convince and acquire new customers.

Well, the acquisition campaigns turn out to be only a minor success.

Sure enough, the company has acquired new customers and generating new sales,

but the sales generated do not even cover the cost of the campaign.

Play video starting at ::52 and follow transcript0:52

The company has lost money.

Now what?

Was it a success?

But of course, the manager in charge can always

argue that these newly acquired customers will remain loyal, purchase again and

generate additional revenue in the future.

In a sense, the acquisition campaign is not a loss, it's an investment and

the company will reap its benefits over the next few weeks, month or even years.

Well, maybe it's true or maybe it's just wishful thinking and the major is trying

to find a execute to justify a campaign that was not successful, but

wouldn't it be nice to be able to actually compute the value of a new customer

over a given period of time and compare that figure to the cost of acquiring

such a new customer to prove or disprove it was a good investment.

That's the whole point of customer lifetime value models or CLV.

The goal of such methods is analyze what is happening today and

what has happened in the recent past in order to predict the revenues customers

will generate in the future, that's what you learn to do in this module.

Customer lifetime value models have many other applications in practice.

For instance, you could compare one acquisition campaign to another.

Suppose you compare two campaigns, one with a very deep price discount and

one with a moderate discount.

It's likely that the promotional campaign that offers the larger price discount will

attract more customers, but

these customers will be less profitable in the short-term and

be much less loyal in the long run once the price discount disappears.

Because the only reason they came in the first place was for the price promotion.

At the end of the day, offering a moderate price discount and

attracting fewer customers might be the best strategy.

If they generate more revenue and remain loyal longer, but

how can you know for sure?

If you cannot put a dollar figure on the long-term value of your customers, but

customer lifetime value is not only useful when applied to new customers.

It's also very useful when applied to existing ones.

One of the best applications of customer lifetime value is to identify,

which customers or which segments of customers are strategic for

the future success of the firm.

Play video starting at :3:17 and follow transcript3:17

Suppose you have two segments, customers in segment number one are generating and

average $100 of revenue per quarter.

Customers in segment two are generating $150 on average.

At first sight, you should invest a bit more effort to satisfy and

retain those customers who belong to the second segment.

Maybe allocate more resources.

Make sure they have access to the best customer service.

Closely monitor the loyalty and so on.

But if you analyze their lifetime value,

you realize that customers in the first segment are more loyal stay longer and

spend money more consistently with the firm.

While the customers in segment two are only seasonal, not loyal at all.

Switching to competition as soon as they find a better offer or

promotion and maybe your one time shoppers.

When comparing the lifetime values,

you realize customers in segment one have a lifetime average of $2,000 and

those in segment two barely spends 400 in average.

Wouldn't it be nice to know how to compute this figures in the first place and

concentrate your marketing dollars on the most valuable customers in the long-term?

And what do you think would happen if you didn't?

Let's see how we can do that.

# Transition probabilities and transition matrix

There are many different techniques to compute customer lifetime value.

Some are extremely simplistic.

Play video starting at ::16 and follow transcript0:16

Probably too simple to be of value.

Other techniques are overly complex.

We'll chose a middle ground approach and learn to compute customer lifetime value

using a concept we've already covered in previous modules, segmentation.

In one of the modules we've already covered, we've talked about managerial

segmentation and how you can develop rules to assign each and

every customer to a segment based on his or her most recent behaviors,

such as recency, frequency, monetary value, and SAN.

Remember that we've also discussed how you could, not only run

a segmentation study on the most recent data you have available, but also how

you could go back in the past and run a segmentation analysis retrospectively.

Play video starting at :1:3 and follow transcript1:03

It gives you interesting insights about how the firm is doing,

which segments are growing,

are you adding more new customers today than you were a year ago, and so on.

In a sense, each segmentation study is a picture of your customer database.

You can easily build a series of snapshots.

But you can do even better, and

analyze how you went from one snapshot to the next.

Let me explain.

Let's suppose you have a very simple segmentation in place.

Actually, we'll take an even simplified version of the segmentation

we did together, and keep only five segments.

Active high value customers, active low value customers, warm, cold and

inactive customers.

How we define these segments doesn't really matter for this illustration.

The most active and most profitable customers are on top and

the least valuable customers are in the bottom.

Play video starting at :1:56 and follow transcript1:56

Using this segmentation,

you can make a snapshot of your customer database to date.

Now how many customers you have and how many fall into each segment.

And of course, you can have the same snapshot taken say,

a year ago with the same information.

What's interesting is that you can also check how customers went from

one segment to another.

If you take the active high value customers, maybe some kept making high

volume purchases and remained in the same segment, which is good news for the third.

Others remained active but spent less and moved to the lower value segments.

Yet other customers might have not purchased anything at all over the last 12

month, and because of that, have now fallen into different segments.

The trick is to look at these figures and transform them into probabilities,

and consider that these probabilities will likely remain stable over time.

In other words, if half of your high value customers a year ago

remain in the same segment this year, then it's likely that about 50%

of your high value customers today will remain high value customers next year.

And the same logic will apply to the year after, and the year after, and so on.

Basically, by analyzing what happened of the recent past,

you will predict what will likely happened in the near future.

Of course, you need to compute these transition probabilities from

any segment to any segment.

And in most realistic context, with say 10 or 15 segments or

more, it can become a real mess to represent these transitions graphically.

So, the tool we'll use instead is called a Transition Matrix.

The leftmost column represent the segments in which customers were said a year ago

and the topmost row represent the segments in which customers are today.

Each value in that matrix, each probability,

represent the likelihood of going from one segment to another, and, of course,

because all customers need to go somewhere, each row sends 200 persons.

Please note that by the very definition of your segments,

some sales may always be equal to zero, meaning there is

a zero probability of going from one specific segment to another.

For instance,

you may never go from active to inactive in just one period, but that's fine.

That's usually the case.

This transition matrix is essential because this is the key element we use to

compute customer lifetime value.

So before we go any further,

let me show you in the next video how you compute the transition matrix in R.

# How to compute a transition matrix in R (Recital 1)

In this tutorial I'm going to show you how to compute a transition matrix,

which is basically a matrix showing the likelihood of going from

one segment to the next over one specific period of time.

So you know the dataset, I'm not going to go over the details here.

I'm loading it, preparing it as we've done so far

Play video starting at ::23 and follow transcript0:23

and then I'm going to compute two things

using code we've already covered in the second module.

I'm going to compute customer segments in 2015, exactly as we've done previously.

Play video starting at ::44 and follow transcript0:44

Here we are.

And then I'm going to compute customer segments in 2014,

also exactly as we've done previously.

Play video starting at ::57 and follow transcript0:57

So we had 16,900 customers in 2014, and

we'd like to know what they became in 2015.

Just one comment here.

We are ordering the segments in a very specific order.

It's extremely important to do that, because later on,

we need these segments to be in a very specific order, and

to be in the same order, whatever the data set we're going to look at.

Play video starting at :1:27 and follow transcript1:27

Now, we have everything we need.

We have the customers in 2014, and the segment they belong to, and

we have exactly the same things for customers in 2015,

and the segments they belong to.

Now, we need to merge these two data sets,

meaning we need to put in to one new data variable all the information

we had about cusiners in 2014 and in 2015.

There is a trick though.

The columns in this data frame and

the columns in this data frame have exactly the same names.

So if you merge them simply, they will disappear and

be replaced by one another so

we need to specify we are going to merge them by customer ID,

keep all the customers on the left as we've done previously.

And look at what we have now.

We have a pretty big data set, with recency, first purchase,

frequency, amount, and segment with a .x.

Meaning that these variables come from the x the customers in 2014,

and we have the exact same list with recency, first purchase,

frequency, and so on, with a .y appendix.

The two vitals we really care about are segment.x which is the segment in 2014 and

segment.y which is the segment in 2015.

So, I'm going to create an occurrence table.

Meaning I'm going to cross segments in 2014, segments in 2015 and count

Play video starting at :3:21 and follow transcript3:21

how many people meet these two criteria.

Play video starting at :3:27 and follow transcript3:27

And actually we are almost there.

If you look at this block, that's basically

to which segments people belong in 2014.

This top row here is which segment they belong to in 2015.

And any value shows how customers went from one to the other.

So among all the inactive customers in 2014,

7,227 remained inactive in the next year,

35 became active high value,

250 became active low value.

And, if you look for instance, at the new active customers, so

the newly acquired customers.

Play video starting at :4:19 and follow transcript4:19

They cannot remain new, right?

They need to move to a different segment immediately by definition.

So among all the new active customers in 2014,

an astounding 938 didn't

purchase anything the next year and became a new warm.

That's about two out of three.

Play video starting at :4:43 and follow transcript4:43

Among the others 89 became active high value,

and 410 became active low value.

But of course, we don't really care about the absolute numbers.

What we care about are proportions,

likelihood, probabilities of going from one segment to the next.

So instead of working with the transition here immediately, what we do is

divide each row by the row sum.

Row sum is a function that computes the sum of each row,

one by one, and so, by doing that, what we obtain in

return is something where the sum of each row is equal to one.

So here, for instance, if you were an inactive customer in 2014,

you had a 96.2% chance of remaining inactive.

And if you were an active high value customer,

you had a 74.5% chance of remaining in that segment and

25% chance of becoming a warm high value customer.

If you were a newly acquired customer in 2014,

you had a 28% chance of becoming an active low value customer,

6% chance only of becoming an active high value customer and

an outstanding 65% chance of doing absolutely nothing and

switching immediately to the new warm customer segments.

Play video starting at :6:24 and follow transcript6:24

That's actually all we need to have.

We have our entire transition matrix,

computing the likelihood of going from any segment to any segment.

Play video starting at :6:37 and follow transcript6:37

The only thing we need to do now is to use that transition matrix

to make predictions.

# Using the transition matrix to estimate how customers will evolve

In the last video, we've shown how to compute a transition matrix.

Once you have that, everything else becomes pretty simple.

The next step of the process is to estimate how many customers you have

in each segment next year, and the year after, and the year after.

And so on, for as long as you'd like.

Although many firms will stop after three, five, or ten years at most.

To estimate these figures, you need to know two things.

One, the transition matrix, and

two how many customers you have in each segment to date.

Remember that the transition matrix summarizes how customers got where they

are today, and

we'll use that exact same information to predict where they will be tomorrow.

Play video starting at ::53 and follow transcript0:53

Now without going too deep into the mathematical details,

this operation consists of multiplying a matrix by a vector.

That's it.

We multiply the transition matrix by a vector that's towards

the number of customers in each segment to date.

The output of that operation will be another vector,

that will contain the estimated number of customers in each segment in a year.

If you take that new vector you just obtained, and

multiply by the transition matrix one more time, you get an estimate of the number of

customers in each segment two years from now, and so on.

Play video starting at :1:29 and follow transcript1:29

Let's then practice how to do that in r.

# Using the transition matrix to make predictions in R (Recital 2)

So we continue this tutorial where we left it.

Play video starting at ::5 and follow transcript0:05

We have a transition matrix showing the likelihood of going from any segment

to any segment over one year.

And what we're going to do is to prepare a matrix,

a placeholder to store all the information we are going

to predict based on this transition matrix.

The number of row is the number of segments we have.

And the number of columns is the number of years we are going to predict

the evolution of segment membership.

We are going to predict over ten periods in the future,

and we'll include one more as the present.

So we create that matrix and the beginning it's an empty one.

Play video starting at ::57 and follow transcript0:57

Okay, only 11 columns, eight rows.

Nothing in there.

The first thing to do is to copy that first

Play video starting at :1:9 and follow transcript1:09

column with how many customers we have in each

segment at the end of 2015, number one.

Then, we're going to give names to all of these columns.

And here, we're going to use a sequence.

from 2015 to 2025 these are the 11 years we want to predict segment membership.

Play video starting at :1:36 and follow transcript1:36

And finally, we're going to give to each row the name of its corresponding segment.

Play video starting at :1:44 and follow transcript1:44

And if you look at the variable we've just created,

you have all the years we're going to make predictions for,

all the segments here appearing as row names,

and the number of customers we have In each segment in 2015.

We'd like to fill out these columns

with predictions made using the transition matrix.

So, for each column, the segments membership over that year

would be a function of segment membership of the year prior,

Play video starting at :2:24 and follow transcript2:24

multiplied by the transition matrix.

And that's it.

We simply took the data from a year prior, multiplied it by the transition matrix,

to know how many customers we had in the next year.

And then through the loop of the four statement here,

we take that result, plug it here, and compute the next year in terms of fourth,

until we reach the end of the loop, which will be the year 2025.

If you look at the metrics we just created, it contains the segment names,

the years and the number of customers in each segment of a time.

Now of course because it's probabilistic, we don't have rounded values.

It doesn't make sense to predict that we'll have 143.5 customers in a segment.

We're going to run these values later on.

But if you look at predictions, for instance, we're going to predoct segment

membership for segment number one over the next ten years.

Play video starting at :3:29 and follow transcript3:29

These are our predictions.

We begin at nine thousand something or 9,158, and

then we expect the number of inactive customers to grow quickly and

then stabilize over something like 13,000.

If you look at the predictions for segment number two,

which is the cold customers, It's actually going the other way around.

We have a lot of cold customers, but actually many of those cold customers will

become inactive and leave the cold segment.

Many active customers will become warm and cold and

Play video starting at :4:10 and follow transcript4:10

appear in 2016 and '17, but eventually they'll become inactive as well, and

then we settle at much lower value as well.

And if you just show these figures over time, I have rounded the results so

that we don't have any decimal points which would make the results hard to read.

As you can see, the number of inactive customers, for instance,

is expected to go from 9,100 to

Play video starting at :4:37 and follow transcript4:37

13,800 by the year 2025.

Interestingly, if you look at the number of new,

active customers here, we have 1,512 in 2015.

We do not expect any new active customers here.

Why?

Well, simply because if you are already in the database, you cannot become new again.

The only way to become a new active customers is to be acquired.

And here, we are measuring only the customer lifetime value of those customers

who are already in the database.

So, by definition, the new active customers will become something else.

Maybe warm, active, cold, inactive customers of a time.

But that segment will remain empty,

unless we run acquisition campaigns and add customers.

Now as you can see we have the predicted number of customers per segment of

the specific number of years, you could go to 20-30, if you'd like or just three.

Play video starting at :5:45 and follow transcript5:45

But we don't have any revenue yet.

That's the next step.

# Assigning and discounting revenue

We are almost there.

We know how to computer a transition matrix and

how to use the transition matrix to guess how many customers will be in each segment

in a year, two years or even 20 years.

If we wanted to.

Now we just need to transform these figures into Dollars or Euro or Yen.

The first step is to assume that the revenue generated by a customer can be

fully explained and predicted by the segment to which he or she belongs.

Whether today or ten years from now.

Now if an average customer in a high value segment generates an average $100,

we'll simply assume that this figure will not change over the years.

In reality it might go up or down, but without additional information

our best guess is to assume that this figure will remain stable over time.

But of course all $100 are not created equally, for

one customer will spend $100 today another will spend the same amount.

But only five years from now, the first customer is more valuable.

Why?

Because future revenues are uncertain, distant and not immediate.

Therefore, they need to be discounted.

When we discount revenues in [INAUDIBLE] computations,

we are not simply taking about inflation.

We are talking about risk.

And uncertainty.

And even if there were zero uncertainty about future revenues,

most firms put more weight on short-term revenues.

So, a dollar today is worth more than the same dollar tomorrow.

For that reason, we discount future revenues in our computations.

It's like a reverse interest rate.

The longer you have to wait to generate revenues,

the less these revenues will be worth for the firm today.

There's no clear guideline about the kind of discount rate you should apply.

But simply remember this, the higher the discount rate,

the more short term focused you will be.

On the other hand, with a discount rate closer to zero, future revenues would not

be discounted as much and computations would be much more long-term focused.

Okay, time to wrap this up and put everything together.

Let's see how you can actually compute customer lifetime value in r.

# Computing customer lifetime value in R (Recital 3)

Okay, here we are for our very last tutorial on customer lifetime value.

I going to rerun all the codes from module four that

are up to the point we left it, which is around line 109.

Play video starting at ::22 and follow transcript0:22

Here we are.

That's where we were.

And now we're going to use these predictions we've

just made to compute the value of the database.

How do we do that?

Well, we need to assign a revenue to each segment.

And we're not going to compute it again, we've done it already.

If you look at the Module 2 code on lines 160 and

161, that's where we got these figures.

Customers were inactive, cold, warm, by definition and generate no revenue at all.

The active high value customers on average generate $323 of revenue per year.

And then for the active low value and new active, it goes down to 52 and 79.

So we're going to just steal the data from that analysis and

store it, in a variable called yearly revenue.

The next step is simply to take the variable called segments,

which contains predictions of segment membership over the next 11 years.

And multiply the segment membership by how much revenue revenue

each segment generates, which is exactly what we've inputted here.

And so if you print it,

as you can see here the active high value customers will generate,

actually have generated in 2015, $185,000.

Then the number of customers in that segment we increase, decrease, stabilize,

and every time for every customer in that segment they will generate $323.

And you see the values here going up and

down as a function of predictions.

Play video starting at :2:18 and follow transcript2:18

The next thing we'd like to do, because that's segment per segment, year by year.

What we'd like to do is to compute the sum of each column.

So we have the yearly revenue of the database for that year.

We are going to use the function call sum.

So we're going to sum each column to itself of the revenue per segment variable

we just create and store the output as a new variable called yearly revenue.

That's what we do, we print it, and so in terms of revenue predictions,

without any discount we've generated $478,000 in 2015.

And because people will become slowly inactive over time

by the year 2025 we'll have only $307,000 in revenue.

So, it will slowly decrease over time as you can see here.

And you can create a bar plot of revenue prediction going pretty high in 2015 and

then, decreasing quite quickly up to 2025,

which is the last year we made predictions for.

If you'd like to compute the cumulative revenue.

So how much money will we have made in 2025 correlated over 2015,

16, 17, 18, and so on and so forth.

You can compute that using what is called the cum sum,

cumulative sum of yearly revenue.

You compute it, print it, and

that's how much revenue will have been generated over the years.

And as you can see of course,

it can only increase since every area you add new revenue to that.

But you probably see that the slope of the curve is slowly deteriorating

because every year a different customer hence we lose a different revenue.

Play video starting at :4:23 and follow transcript4:23

But a dollar ten years from now is not worth a dollar to them.

It needs to be discounted.

I'm going to set the discount rate at 10% and

compute the discount rate for years one to 11.

So basically from year 2015 to 2025, knowing that for

2015 the discount rate cannot be applied since it's today.

So I remove one year here as if it were immediately now.

And the function is 1 / ((1 + discount\_rate)

to the power of how many years you need to wait to get the money.

And, as you can see.

In the first year you're not going to discount anything.

That's today, today's money worth in today's dollars.

Then after a year, a dollar would be worth $0.90.

Then only $0.82 75 and so on.

And as you can see, after 10 years it will only be worth $0.38.

How am I going to use that information?

Well, simply by taking all the values I've generated so far.

Such as the yearly revenue, and multiply these figures by

the discount rate to get something that is worth $2,000.

That's how much money in two days worth is gonna be generated in 2008, 19, and so on.

And so by the year 2025 we generate the equivalent of what

would be worse to the $118,000 for two reasons.

Number one, it's in ten years, so it needs to be discounted.

Number two, it's in ten years.

Meaning many customers will have left and will not be active anymore.

And if I plot that here,

here you have the undiscounted revenue generated of the years.

And here you have the discounted ones.

As you can see it, drops dramatically.

Since the further away in the future you get the revenue,

the less worth it is in today's dollars.

In terms of cumulative revenue, exactly the same thing.

You take the yearly revenue discounted and

compute the sum, and you get every everything we've had already,

except it's now represented in discounted revenue.

Play video starting at :7:1 and follow transcript7:01

The last question we could ask is, off the next ten years,

how much is my database worth?

What's the true value, the discounted cumulative value

of my database in terms of expected revenue of the next ten years.

What you can do is to look at how much it would be worth accumulated at

the end of the period you are analyzing, in this case the 11th period.

That would be 225, of which you remove the revenue from to date,

2015, which already happened.

And if you run and print that, the answer is 2,150,000.

So, if you had to value your database in today's dollars in terms of how much

revenue it will generate over the next ten years, the answer is $2.15 million.

# Instructions before starting the quiz 4

To validate Module 4 and correctly answer the questions will require that you perform the following exercise first: take the code in the file named module4.R, and modify it such as, in the simulations, from 2016 to 2025, 1,000 new customers are acquired every year and added to the database. Tip: it will require that you add a line of code between the lines #99 and #100 to specify that segment 8 has an influx of 1,000 new customers every year.

**You'll be asked to answer these questions :**

How many "inactive" customers does the model predict there will be in 2025?

What would the database be worth by 2025 (cumulated revenues, discounted) if a constant influx of 1,000 new customers was added every year till then?

Looking at the transition matrix, what is the likelihood of a "new warm" customer to become "cold"?

Looking at the transition matrix, the likelihood of transitioning from "new active" to "cold" customer is 0%. Why so?

After the modification to the code, how many customers will there be in the database by 2022?

[module4](https://d3c33hcgiwev3.cloudfront.net/_0b18c557e754c29f09eeb881a8e6255d_module4.R?Expires=1703894400&Signature=b4ZQDscggW59w0~nE48AGe~GOfEvyCsvbPfRkAl8F8HRDE2~zW~am-jkZ2xt02tmWxdKO~D-4KaaFySYUeSOepMpWutPfhC-s7we562JFrhfQ8ymHLWKhOzqohiyKpWsq~sfm8MqyIHtFQRRooi8Nhaff9s4eQZknxzfJ-QPans_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[R File](https://d3c33hcgiwev3.cloudfront.net/_0b18c557e754c29f09eeb881a8e6255d_module4.R?Expires=1703894400&Signature=b4ZQDscggW59w0~nE48AGe~GOfEvyCsvbPfRkAl8F8HRDE2~zW~am-jkZ2xt02tmWxdKO~D-4KaaFySYUeSOepMpWutPfhC-s7we562JFrhfQ8ymHLWKhOzqohiyKpWsq~sfm8MqyIHtFQRRooi8Nhaff9s4eQZknxzfJ-QPans_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[purchases](https://d3c33hcgiwev3.cloudfront.net/_a03b30bcfcb5145b2141d5b2c042e103_purchases.txt?Expires=1703894400&Signature=Q2Gyklrqa9pmrtX-mrM~T43SturslPzuYt-n6eIMXMmj1jHCrb0CYTuqgJkC-~QPwWSB8TqahKQiR8LZ32Zsldbw7gqq2PlpdK-cqD5scnG~KPisy09xyMAI0NijuWpHQ7dv4OnB1L4XaAR3JH-f8AP1KRGwG1G1IT~IyQY9RjI_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[TXT File](https://d3c33hcgiwev3.cloudfront.net/_a03b30bcfcb5145b2141d5b2c042e103_purchases.txt?Expires=1703894400&Signature=Q2Gyklrqa9pmrtX-mrM~T43SturslPzuYt-n6eIMXMmj1jHCrb0CYTuqgJkC-~QPwWSB8TqahKQiR8LZ32Zsldbw7gqq2PlpdK-cqD5scnG~KPisy09xyMAI0NijuWpHQ7dv4OnB1L4XaAR3JH-f8AP1KRGwG1G1IT~IyQY9RjI_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

<https://github.com/nicolasfguillaume/Marketing-Analytics-with-R/blob/master/problem_set_4.md#:~:text=3%2F%20Looking%20at%20the%20transition,customer%20to%20become%20%22cold%22%3F&text=The%20likelihood%20of%20a%20%22new,It's%20extremely%20likely>.